JUMS

Volume 10, Issue 2, June 2025

JUNIOR MANAGEMENT SCIENCE

Published by Junior Management Science e.V.

ISSN: 2942-1861

Junior Management Science

www.jums.academy
ISSN: 2942-1861

The Impact of Profitability on Scope 1, 2 and 3 GHG Emissions in Europe

Yannick Hohenstein

University of St.Gallen

Abstract

This thesis examines the effect of corporate profitability on the levels of greenhouse gas (GHG) emissions, specifically analyzing Scope 1, 2, and 3 emissions for European companies listed on the STOXX Europe 600 index from 2017 to 2023. Given increasing regulatory pressures, inconclusive evidence on whether profitability drives sustainability, and potential bidirectional causality, researching this relationship is highly relevant. Using a systematic literature review (SLR) and fixed-effects regressions, this thesis investigates this relationship. Results show profitability, measured by return on assets (ROA), negatively correlates with Scope 3 emissions, suggesting higher profits may promote sustainability. However, no significant correlation exists for Scope 1 and 2 emissions, except for a positive link with Scope 2 emissions in low-emission sectors. High-emission industries show stronger model explanatory power, indicating a closer profitability-emissions link. Findings are robust against outliers but vary with changing profitability metrics. This research contributes to the profitability-sustainability debate, offering insights for policymakers, scholars, and managers, while emphasizing the need to consider industry and Scope-specific dynamics to combat climate change.

Keywords: GHG emissions; profitability; sustainability reporting

1. Introduction

In the last decades, the accelerating pace of climate change has brought the issue of greenhouse gas (GHG) emissions to the forefront of global discussions (Manabe, 2019; Solomon et al., 2009; van Vuuren & Riahi, 2008). The impact of corporate activities on the environment, mainly through GHG emissions, has become a critical area of concern. Companies worldwide are still making substantial profits based on business practices detrimental to the environment (Trucost, 2013). The United Nations' (UN) Sustainable Development Goals (SDGs) and the Paris Climate Agreement of 2015 underscore the need for a global effort to reduce GHG emissions and combat climate change (United Nations, 2015a). These international frameworks have set the stage for more stringent regulations and reporting requirements, particularly in the European Union (EU), which is recognised as a leader in sustainability reporting (Barbu et al., 2022). The EU has taken significant steps to integrate sustainability into corporate reporting, primarily through the Non-Financial Reporting Directive (NFRD) and its successor, the Corporate Sustainability Reporting Directive (CSRD) (European Union, 2022). These directives mandate large companies to disclose their environmental social, and governance (ESG) performance, with a specific emphasis on GHG emissions categorised under Scope 1, Scope 2 and Scope 3 as per the GHG Protocol (WRI & WBCSD, 2004). Scope 1 encompasses direct emissions from sources owned or controlled by the company. Scope 2 refers to indirect emissions resulting from the production of electricity, steam, heating and cooling that the company purchases. Scope 3 covers all other indirect emissions associated with the company's value chain. (WRI & WBCSD, 2004)

Amidst the increasing public and regulatory attention on GHG emissions, both scholars and business professionals have questioned whether "it pays to be green" (see, e.g., Busch and Hoffmann, 2011; Cote, 2021; Hoang et al., 2020; Lewandowski, 2017). This inquiry suggests that companies achieving lower GHG emissions may experience enhanced profitability or increased firm value. This perspective aligns with Porter's Hypothesis, which posits a "win-win" scenario

DOI: https://doi.org/10.5282/jums/v10i2pp292-333

where stricter regulations foster innovation, improve competitive advantage and ultimately enhance financial performance (Porter, 1980; Porter & van der Linde, 1995; Waddock & Graves, 1997). However, existing literature presents mixed findings on this relationship (Galama & Scholtens, 2021; Iwata & Okada, 2011; J. Wang et al., 2021). Some studies even suggest that "it pays not to be green", implying that higher GHG emissions may be associated with greater profitability (Rokhmawati et al., 2015; L. Wang et al., 2014). These conflicting results highlight the complexity of this research area, with some scholars proposing the existence of reverse causality or bidirectionality between financial performance and GHG emissions, which could significantly influence the observed outcomes (Endrikat et al., 2014; Testa & D'Amato, 2017; Waddock & Graves, 1997). Nonetheless, limited research addresses this potential reverse relationship, encapsulated in the question: "Does profitability drive sustainability?" (Hassan & Romilly, 2018; Meng et al., 2023; Shahgholian, 2019). This potential relationship, grounded in the Slack Resource Theory and aspects of Stakeholder and Legitimacy Theory, suggests that more profitable companies may naturally invest more in GHG reduction efforts to achieve legitimacy and manage stakeholder relations (see, e.g., Cyert and March, 1963; Dowling and Pfeffer, 1975; Freeman, 1984; Waddock and Graves, 1997). The impact of profitability on GHG emissions represents a critical yet underexplored area of study, which could contribute to a deeper understanding of the profitability-sustainability nexus.

This thesis addresses this research gap by comprehensively analysing Scope 1, 2 and 3 GHG emissions reported by European companies from 2017 to 2023 and empirically examining profitability's impact on these emissions. Given Europe's robust reporting framework, high data quality and availability are anticipated. Consequently, the study will focus on companies listed on the STOXX Europe 600 index, including some of the region's largest firms. The central research questions of this thesis are twofold:

(1) *What are the Scope 1, 2 and 3 GHG emissions levels for European companies from 2017 to 2023?*

(2) *How does firm profitability impact total and individual Scope 1, 2 and 3 GHG emissions?*

By answering these questions, this thesis aims to contribute to the current literature on corporate sustainability reporting, CO2-Footprints, and the relation between financial performance and GHG emissions to provide valuable insights for policymakers, corporate managers, and other stakeholders. This work will be structured as follows: The second chapter provides a detailed overview of the fundamentals of sustainability reporting, including the regulatory landscape and the specific requirements of the GHG Protocol. The third chapter presents a systematic literature review (SLR), highlighting the academic relevance of the research questions and identifying gaps in the existing literature. The fourth chapter explains the theoretical framework, drawing on Slack Resources, Legitimacy and Stakeholder Theory, to explain the potential impact of profitability on GHG emissions. The fifth chapter develops and discusses the hypotheses for this regression. The sixth chapter outlines the methodology used to collect and analyse data, followed by a presentation of the results in the seventh chapter. The concluding chapter discusses the implications of the findings, their limitations and provides concluding remarks.

This research is particularly timely as companies prepare to comply with the new CSRD requirements, which will make the disclosure of all three Scopes of GHG emissions mandatory for approximately 50,000 companies starting in 2024 (European Parliament, 2022; European Union, 2022). The findings of this thesis will not only shed light on the current state of GHG emissions reporting in Europe but also guide future research and policies. Furthermore, by exploring the relationship between profitability and GHG emissions, this study aims to inform the ongoing debate on whether and how economic performance is aligned with environmental sustainability. Before proceeding with the literature review and the analysis of GHG emissions, it is essential to understand the basics of sustainability reporting, specifically the GHG Protocol, which will be discussed in the following chapters.

2. Fundamentals of Sustainability Reporting

Broad publications of GHG emissions by companies occurred relatively recently and has been largely influenced by recent advancements in non-financial reporting practices. A basic understanding of the non-financial or sustainability reporting landscape is necessary to analyse the countervailing trends in GHG emissions and understand the factors influencing them. Therefore, this thesis first briefly introduces sustainability reporting and the sustainability reporting landscape.

2.1. Introduction to Sustainability Reporting

The introduction to sustainability reporting begins with a basic definition of the term and then briefly discusses its importance, benefits, and challenges.

2.1.1. Definition

At first, the meaning of sustainability reporting might seem easy to grasp; it focuses primarily on Environmental, Social, and Governance (ESG) topics and is also described as non-financial information (NFI). However, according to Erkens et al. (2015), who analysed 787 articles published in 53 journals from 1973 to 2013, non-financial information seems to need a more precise definition. They attribute this to the ambiguity of the concept of NFI and try to define the topic on their own.

Before we move on to the definition of NFI, it is helpful to first define financial reporting to distinguish between the two topics and highlight the differences. Traditional financial reporting has become highly standardised and is based on generally accepted accounting principles (Ampofo & Sellani, 2005). In Europe, for example, these are published by international associations such as the International Accounting Standards Board (IASB) and form the basis of today's financial reporting (Van Greuning et al., 2011). This type of reporting aims to inform investors about a company's financial performance. The IFRS Framework states that the objective is to "*provide financial information about the reporting entity that is useful to existing and potential investors, lenders and other creditors in making decisions about providing resources to the entity*" (IFRS Foundation, 2018, Conceptual Framework, §1.2). Per definition, the disclosure of financial information provides the correct information for investors, lenders, and other creditors, but in the last decades, calls from investors and other stakeholders for non-financial reporting on crucial ESG issues have increased (KPMG, 2022).

According to Erkens et al. (2015, p. 25), NFI can be defined as a disclosure "*on dimensions of performance other than the traditional assessment of financial performance*", including, but not limited to, topics related to ESG. Tarquinio and Posadas (2020) conducted a literature review on the term "non-financial information" and found that there is still no consensus on the exact definition of this term. In addition to the NFI, the term "sustainability reporting" is employed almost synonymously, and increased use of it can be observed (Baumüller & Grbenic, 2021; Eccles et al., 2020). The change from the Non-Financial Reporting Directive to the Corporate Sustainability Reporting Directive is an example of the shift to the term "sustainability reporting", which, as the name suggests, consciously emphasises the importance of a more integrated way of thinking about global issues and a tool to fight climate change (Baumüller & Sopp, 2022). The term "sustainability reporting" has now established itself and, to some extent, replaces and expands the term "non-financial information" (Baumüller & Grbenic, 2021). For this thesis, these definitions are sufficient since we limit ourselves to the information on GHG emissions included in the sustainability or annual reports and do not engage with the documents in their entirety. Having established an understanding of the definition of sustainability reporting, the next step is to delve into its relevance and importance for the business landscape.

2.1.2. Importance and Relevance

The topic of sustainability reporting has become omnipresent for companies, and an increase in research concerning sustainability reporting can be observed (Erkens et al., 2015). New regulations primarily drive the trend, as around 11,700 public-interest entities have been obliged to report by the EU NFRD starting in 2017, and about 50,000 will be, under the new CSRD (European Broadcasting Union, 2023). This reporting regulation is needed because past efforts to fight climate change have not been enough, and governments have committed themselves, albeit not legally binding, to achieving the SDGs (United Nations, 2015b). Conversely, this means they must encourage the achievement of the climate goals and monitor progress through national or international regulation. The reporting of non-financial information has made significant progress over the last years and comes with great benefits for various stakeholders (Bualllay, 2019; James, 2015), but still has significant challenges to overcome, particularly concerning its alignment with the attainment of the UN SDGs (Tsalis et al., 2020). Both benefits and challenges will be discussed in the following two chapters.

2.1.3. Benefits and Advantages

Various research on the benefits of sustainability reporting was published, and the positive effects can be observed for companies and the common good (Bellantuono et al., 2016; Ioannou & Serafeim, 2017; Tomar, 2022). Research conducted by Tomar (2022) analysed the effects of the U.S. GHG Reporting Program on the GHG amounts emitted by facilities and found that the disclosure alone led to a 7,9% reduction of their respective GHG emissions. Benchmarking and reporting GHG emissions alone seem to encourage reduction and is, therefore, a welcome positive effect of sustainability reporting (Tomar, 2022). Another benefit is the increased transparency and disclosures firms make on sustainability issues (Ioannou & Serafeim, 2017). The stakeholders are, on the one hand, pushing firms to increase disclosures and, on the other hand, benefit from it because mandatory but also voluntary reporting on environmental, social, and governance matters provides the stakeholders with insights into companies that would not be common before this trend (Bellantuono et al., 2016; Fernandez-Feijoo et al., 2014; Herremans et al., 2016; Manetti & Toccafondi, 2012). In 2015, the Chief Executive Officer of the Global Reporting Initiative (GRI), a global standard-setter for sustainability reporting, proposed another view of sustainability reporting during an interview (Kiron & Kruschwitz, 2015). According to him, the reports can highlight material and relevant sustainability issues for the companies (Kiron & Kruschwitz, 2015) and, therefore, be used as a strategic tool for decision-making and risk management, which was already researched by C. A. Adams and Frost (2008). Furthermore, sustainability reporting and, therefore, the combination of higher transparency, better risk assessment and decision-making seems to have a positive impact on firm valuations (Kuzey & Uyar, 2017; Loh et al., 2017).

Nevertheless, most research observing the benefits of sustainability reporting was conducted before it became mandatory for most major European companies. The current regulatory developments, namely the NFRD and upcoming CSRD, could lead to a situation where it is no longer reporting per se, which brings advantages for the companies but rather relative performance towards sustainability goals. After having reviewed the potential benefits, we will look at the current challenges sustainability reporting faces.

2.1.4. Challenges and Obstacles

Although the beginnings of sustainability reporting go back several decades, many challenges can still be observed. Despite a significant number of companies using the GRI standards for their reporting, a considerable challenge is the lack of comparability between their current reports and past ones, as well as with the reports of other companies and industries (Zsóka & Vajkai, 2018). Another study by Cardoni et al. (2019) analysed the comparability of 41 GRI reports of listed oil and gas companies and noted the low comparability between the reports. Poor comparability is still a problem that will hopefully improve with more regulation and requirements on crucial aspects like the key performance indicators and the format of sustainability reports.

Reporting standards such as GRI seem to have increased the quality of sustainability reports, as Diouf and Boiral (2017) analysed through stakeholder interviews. However, the quality of the sustainability reports still lacks behind financial reporting and is highly influenced by the specific application and interpretation, e.g., the GRI principles (Boiral et al., 2019; Diouf & Boiral, 2017). Next to quality issues, the materiality is challenging to assess due to the subjective nature of specific information (Wu et al., 2018). A solution would be assurance statements, as we see them for financial statements and annual reports (Wallage, 2000). However, a study by O'Dwyer and Owen (2005) and a newer one by Boiral and Heras-Saizarbitoria (2020) question the usefulness of this practice and show the lack of reliability of assurance statements. A significant issue Boiral and Heras-Saizarbitoria (2020) criticises in the assurance procedures is the seeming disconnection "*from real sustainability issues and reporting requirements*" (Boiral & Heras-Saizarbitoria, 2020, p. 12). Time will tell how and whether mandatory audits on sustainability reporting will prevail. As for now, the new EU CSRD will require limited assurance of sustainability information (European Union, 2022). The low quality, low comparability and lack of transparency of sustainability reports contradict their actual goal, namely, providing transparent information on the sustainability performance of companies. In a study of 21 GRI reports rated A and A+, Boiral (2013) found that 90% of the relevant sustainability events were not correctly presented in the reports. Furthermore, greenwashing is still a problem, making it difficult for sustainability reports to build credibility in the fight against climate change (de Freitas Netto et al., 2020). This undermines the transparency and credibility of the reports (Boiral, 2013; de Freitas Netto et al., 2020) and raises the question of whether they are conducive to achieving climate goals.

In summary, despite standards such as the GRI and efforts by companies, sustainability reports remain difficult to compare and can lack transparency. Regulators and independent initiatives have been trying to establish standards for several years and have already greatly improved reporting, but a multitude of diverse standards and frameworks have emerged, leading to complexity and challenges in comprehension and implementation.

2.2. The Sustainability Reporting Landscape

Building on the introduction to sustainability reporting, the following chapter will explore the landscape of regulations, standards, and frameworks around sustainability reporting, with a focus on the global goals and principles and the regulations in Europe.

2.2.1. Introduction to Sustainability Regulations and Frameworks

Sustainability reports have been an integral part of corporate reporting for several years. In contrast to financial reporting, the regulatory environment was and still is much more fragmented (Young, 2023). This section analyses the landscape around sustainability reports, and the latest developments in the field are discussed. Inspired by the publication of Helbing (2022) on the reporting landscape, this work opted for a pyramid-shaped structure, displayed in Figure 1, which represents the various sub-areas of sustainability reporting effectively. The SDGs of the UN and the Paris Climate Agreement are the overarching goals for sustainability reporting, and the governmental regulations to achieve them will be examined in the following. The focus is on the European standards NFRD and CSRD, which have already been published and cover the companies in our study. Not to be forgotten are the China ESG Disclosure Standards, the upcoming SEC Climate Disclosures from the USA, and other country-specific regulations, which we will not examine further in the context of this work. The cornerstones of sustainability reports are the various frameworks and standards that have been established in recent years. These include the newly founded International Sustainability Standards Board (ISSB), which aims to consolidate multiple standards and frameworks under the IFRS Foundation to establish itself as a global standard (IFRS Foundation, 2024). In addition, the GRI, the GHG Protocol, the Task Force on Climate-related Financial Disclosures (TCFD), the Science Based Targets Initiative (SBTi), and the Carbon Disclosure Project (CDP), have also established themselves in the sustainability reporting landscape.

The two sub-areas, Global Goals & Principles and Governmental Regulations, will be covered in more detail below. The GHG Protocol and the three different Scopes are discussed in a separate chapter due to their importance for our analysis of corporate GHG emissions.

2.2.2. Global Goals and Principles

The 17 SDGs adopted by the UN in September 2015 mark a milestone for global goals and have also influenced sustainability reports (United Nations, 2015b). For the first time in history, the UN shifted to "*one sustainable development agenda*", setting the goals on a global scale (Biermann et al., 2017, p. 26). The new approach adopted by the UN is "governance by goals", which is not legally binding (Kim, 2016) but based on shared objectives from the UN Member states (Biermann et al., 2017). An additional unique characteristic of the SDGs is the focus on all relevant actors, including companies and social organisations, rather than only focusing on the states (United Nations, 2015b). The 17 SDGs combine 169 defined targets with specific deadlines, but some remain qualitative, leaving room for interpretation (United

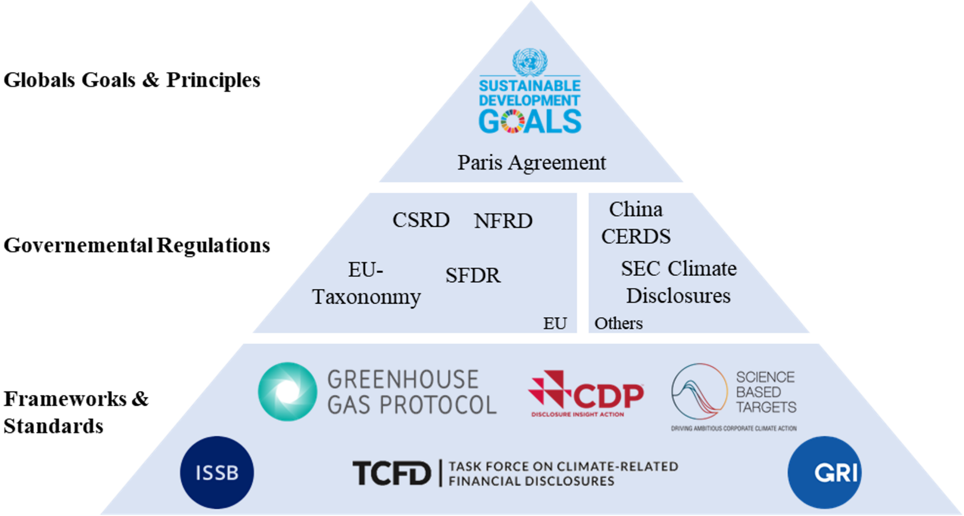


Figure 1: Sustainability Reporting Landscape Pyramid, based on Helbing (2022)

Nations, 2015b). In recent years, standards setters, institutions, and companies have sought to include SDGs in their corporate reporting, an essential step towards achieving the goals (Elalfy et al., 2021; Subramaniam et al., 2023). For example, GRI links the GRI standards to SDGs, thus allowing companies to firmly establish the SDGs in their reporting (GRI, 2022). However, the qualitative nature of some SDGs, together with the challenges of sustainability reporting discussed in Chapter 2.1.4, lead to shortcomings such as intangibility, low standardisation, omission of negative impacts, and lack of comparability (Diaz-Sarachaga, 2021). Despite their voluntary character and some shortcomings in the disclosures, the SDGs have found their way into sustainability reports. They can be seen as the global goals and principles that businesses, governments and other parts of society aim to achieve.

Another global goal alongside the SDGs is the limitation of the global average temperature increase to well below 2°C, as agreed on by the UN in the Paris Agreement, the first in time legally binding global climate change agreement (United Nations, 2015a). This agreement explicitly limits the rise in temperature and the global emission levels, which is linked to the GHG emissions in sustainability reports. The main mitigation objectives are to limit the global average temperature increase to well below 2°C above pre-industrial level and strive to the more ambitious 1,5°C target. These targets require global emissions to peak as soon as possible and subsequently reduced quickly. In addition, it was agreed in the Paris Agreement to track the progress of the commitments and to rely on a transparent system for this purpose. (United Nations, 2015a) Subsequently, the limitation of GHG emissions and transparent measurement of targets requires countries and companies to clearly disclose and reduce GHG emissions.

The two UN conventions require, although only the Paris Agreement is legally binding, governments to incorporate the goals into their legislation (United Nations, 2015a, 2015b). To meet these requirements, countries and country unions such as the EU have published laws and requirements for sustainability reporting, which we will discuss in the following chapter.

2.2.3. NFRD and CSRD in Europe

Regulations shape today's financial reporting and have contributed significantly to the standardisation and comparability of financial reports (Van Greuning et al., 2011). Similarly, new regulations on non-financial reporting have developed in recent years and already characterise a significant proportion of sustainability reports. Based on global principles, this text will now focus on the European scope only.

In the European Union, the first relevant regulation on non-financial reporting was published on 5 December 2014, under the name NFRD (European Union, 2014). Directive 2014/95/EU on disclosure of non-financial and diversity information requires large public-interest entities with more than 500 employees, which amounts to approximately 11'700 companies in the European Union, to disclose relevant non-financial information to investors and other stakeholder (European Broadcasting Union, 2023). To quote the official summary of the law:

> *"Such companies are required to give a review of their business model, policies, outcomes, principal risks and key performance indicators, including on: environmental matters; social and employee aspects; respect for human rights; anti-corruption and bribery issues."* (European Union, 2019, p. 1)

The NFRD required companies to comply with the directive for the first time in the 2017 financial year reports published in 2018, raising the sustainability reporting requirements in Europe. Although the disclosure of GHG emissions by Scopes only becomes mandatory with the CSRD, a large share of European companies already reporting their Scope 1, 2 and 3 GHG emissions is expected. Therefore, the GHG emission numbers of FY 2017 mark the ideal starting period for our analysis period from 2017 to 2023. Nevertheless, the NFRD gave the reporting companies substantial freedom in the choice of how to report and did not require a specific standard or framework, which led to difficulties in comparability, relevance, and reliability of the different non-financial disclosures (Hahnkamper-Vandenbulcke, 2021).

As part of the European Green Deal, it was decided on 11 December 2019 to review the NFRD and solve the associated problems and shortcomings (Hahnkamper-Vandenbulcke, 2021). The main issues and needs identified during the public consultation were the lack of comparability, reliability and relevance, overlaps with other regulations, the lack of a mandatory reporting standard, stricter audit requirements, a digitalisation of non-financial reporting, the disclosure of the materiality assessment procedures used by companies and last but not least the extension of mandatory non-financial reporting to other listed and incorporated companies active in the EU (Hahnkamper-Vandenbulcke, 2021). The EU's solution to these problems was to come into force on January, 5, 2023 under the CSRD (European Union, 2022). The CSRD applies to companies with two out of the three following characteristics: >500 employees and/or, > € 40mio turnover and/or, > € 20mio total assets and for all listed companies (European Union, 2022), which enlarges the number of companies required to report under CSRD to approximately 50,000 (European Parliament, 2022; European Union, 2022). In addition to the supplementary companies covered by the new directive, the reporting requirements of the NFRD remain in effect, next to the additional requirements introduced by the CSRD (European Union, 2022). Companies must report in accordance with the CSRD from the 2024 financial year onwards, following the new European Sustainability Reporting Standards (ESRS) developed by the European Financial Reporting Advisory Group (EFRAG). Since compliance with new standards involves significant direct and indirect costs, and organisational effort, as EFRAG's cost analysis points out (EFRAG, 2023a), there will be simplified reporting for small and medium-sized enterprises. With the ESRS, the European Union is responding to the demand of Stakeholders for a uniform standard for sustainability reports, which should lead to greater comparability (European Commission, 2023).

An essential principle introduced with the CSRD is the double materiality, which states that companies must first document the impact of sustainability issues on their company's financial and corporate situation and, secondly, the impact the company has on sustainability issues. In contrast to the regulations of the NFRD, this requires companies to report on topics that impact the environment but not their economic situation, thus preventing one-sided reporting (envoria, 2022). Furthermore, the CSRD requires companies to report additional information on intangibles, including forward-looking targets, and link them to the relevant targets of the Paris Agreement and UN SDGs.

Another objective of the new directive is a standardised reporting design. Most of the relevant information from CSRD-compliant reporting will have to be digitised and machine-readable in the European Single Electronic Format (ESEF/XHTML), which should facilitate comparability and information search within the sustainability reports (ESMA, n.d.). Finally, the new CSRD introduces a mandatory limited external assurance of the published sustainability information (European Commission, n.d.).

It is still too early to observe the effects of the CSRD on sustainability reporting, but the NFRD has already led to interesting developments. A study by Cuomo et al. (2022) analysed the effects of the NFRD on corporate social responsibility and found an increase in performance and transparency. Another study linked the NFRD to better environmental and social performance on ESG scores but could not find a significant effect on the governance dimension (Aluchna et al., 2023).

In summary, significant developments in the regulatory environment of the European Union are observed. The new CSRD addresses many of the problems of the NFRD, which will hopefully lead to the desired effects, such as increased transparency, comparability, GHG reduction and usefulness of sustainability reporting. A single standard standing out when it comes to the definition and calculation of GHG emissions is the GHG Protocol. Therefore, getting an overview of this standard and understanding the individual Scope 1, 2 and 3 GHG emission Scopes is worthwhile. Accordingly, the GHG Protocol will be discussed in the next chapter.

2.3. The GHG Protocol: Scope 1, 2 and 3 GHG Emissions

The GHG emissions of European companies are a key focus of this study, and the GHG Protocol has established itself as a standard for their definition and calculations. Therefore, this chapter will provide a brief introduction to the GHG Protocol and the individual Scope 1, 2 and 3 emissions.

2.3.1. Introduction and Relevance of the GHG Protocol

The GHG Protocol Initiative was launched in 1998 by a partnership of NGOs, governments, businesses and institutions. The first edition of the GHG Protocol Corporate Standards was published in 2001, with a revised edition in 2004, and was well received by the stakeholders (Green, 2010). The protocol provides a standard and recommendations for companies, as well as other organisations, to quantify their GHG emissions, and includes accounting and reporting guidelines for the seven GHG defined by the Kyoto Protocol: carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), sulphur hexafluoride (SF6), and Nitrogen trifluoride (NF3). (WRI & WBCSD, 2004) After its introduction, the GHG Protocol has gained acceptance as a standard in recent years and is explicitly recommended or required by the GRI, CDP, SBTi, and ESRS, among others, to calculate GHG emissions (CDP, 2023; EFRAG, 2023b; Green, 2010; GRI, 2024; SBTi, 2024).

An introduction to the GHG protocol is essential for this work, as the GHG Protocol has become the standard in reporting and accordingly, most of the calculation of GHG emissions by companies are calculated with the GHG Protocol Corporate Standard (Green, 2010). The GHG Protocol provides a comprehensive framework consisting of five steps to identify and quantify GHG emissions. These emissions are categorised into three distinct Scopes—Scope 1, Scope 2, and Scope 3—each of which has unique implications and methodologies for calculation. Consequently, a short insight into the three different Scopes, explained by the example of Thyssenkrupp AG, will be given. The first step is to identify the sources of GHG emissions, which typically occur from stationary combustion, mobile combustion, process emissions, and fugitive emissions. After the identification comes the selection of a calculation approach; the most accurate way would be to measure the emissions directly at the point of origin, which can hardly be guaranteed in reality and would often cause too high costs. Therefore, emission factors for specific processes or fuel quantities are often used, allowing a cost-effective and relatively accurate measurement. However, companies are always encouraged to use the most accurate and appropriate method. Next comes the collection of data across the three Scopes and the application of calculation tools, like the GHG Protocol Initiative publishes on their website. The calculation tools can be divided into two categories: the cross-sector tools for GHG emissions that apply to multiple sectors equally, like stationary combustion and mobile combustion, and the sector-specific tools for specific sectors like cement, steel, aluminium, or offices. Finally, the collected information must be aggregated at the corporate level. This can be done with the centralised and decentralised approaches; a centralised approach requests activity or fuel use data from the reporting units, and the emissions are calculated by the central based on this information; a decentralised approach requires reporting units to calculate GHG emission themselves, which leads to additional work for the strategic business units but creates more understanding for the emissions. (WRI & WBCSD, 2004, p. 41–46) Next follows a short description of Scope 1 to 3 and examples of the respective emissions.

2.3.2. The Three Emission Scopes

Scope 1 GHG emissions refer to direct emissions of GHG from sources owned or controlled by an organisation. These emissions result from activities or processes that occur within an organisation's operational boundaries. Common sources of Scope 1 emissions include on-site combustion of fossil fuels, such as those used in heating, industrial processes, and transportation, as well as emissions from chemical reactions or other on-site activities. (WRI & WBCSD, 2004) According to the definition, Scope 1 emissions will be high for companies burning fossil fuels during their production. ThyssenKrupp AG (TK) seems to be a good example as they recorded comparably high emissions for Scope 1 and provided further information on their methodology in their CDP Response Report – Climate Change 2023 (Thyssenkrupp, 2024). The company records all its emissions according to the Corporate GHG Protocol and chose October 1, 2017, to September 30, 2018, as a base year for all three emission Scopes. Scope 1 emissions for the base year were 24.2 Mio. t. of CO_2equivalents (CO2e) and 21.4 Mio. t. of CO_2e for the year 2023, which is relatively high due to their direct emissions from coal and coke usage in their steel business (Thyssenkrupp, 2024). According to TK, the steel division is responsible for 95% of their GHG emissions, and blast furnaces and electric arc furnaces cause the most significant volume.

Scope 2 GHG emissions encompass indirect emissions associated with consuming purchased or outsourced energy, such as electricity, steam, or heat. These emissions occur outside an organisation's operational boundaries but result from the generation of energy the organisation uses. Common sources of Scope 2 emissions include electricity purchased from the grid, district heating or cooling systems. (WRI & WBCSD, 2004) TK's Scope 2 GHG emissions are calculated using a location- and market-based approach. The location-based approach defines a specific CO2e per kWh number for everyone using the same power grid. The market-based approach allows the company to calculate its emissions based on specific energy purchase agreements, with an energy mix varying from the grid average (brightest, n.d.). The location-based Scope 2 emissions of TK for 2023 are 0.8 Mio. t. of CO_2e and 1.1 Mio. t. of CO_2e for the market-based approach, indicating that TK sources energy from specific supply contracts with higher than grid average CO_2e emissions per kWh (Thyssenkrupp, 2024).

Scope 3 GHG emissions encompass all other indirect emissions that occur due to an organisation's activities but are beyond its direct control and operational boundaries. Typical sources of Scope 3 emissions involve emissions associated with the entire supply chain of a product or service, including the life cycle, purchased goods and services, transportation and distribution, employee commuting, and the disposal or end-of-life treatment of products and services. These emissions can be much larger than a company's Scope 1 and 2 emissions and often account for the most significant portion of an organisation's total carbon footprint. (WRI & WBCSD, 2004) Continuing with the TK example, it becomes clear that measuring Scope 3 emissions is a major challenge for companies. The Scope 3 calculation from TK is based on the Corporate Value Chain Accounting and Reporting Standard of the GHG Protocol and is distributed across 17 emission categories (WBCSD, 2011). The most important in the case of TK appears to be *Purchased goods and services* with 27.2 Mio. t. of CO_2e, *Fuel-and-energy-related activities (not included in Scope 1 or 2)* with 4 Mio. t. of CO_2e and *Upstream transportation and distribution* with 5.3 Mio. t. of CO_2e. Other categories that are of minor relevance to TK and cause no or only minor emissions are *the use of sold products, employee commuting, business travel, capital goods, investments, or franchises* (Thyssenkrupp, 2024). In total, the Scope 3 GHG emissions of TK are assumed to be about 37 Mio. t. of CO_2e, making Scope 3 emissions the most significant part of total emissions (Thyssenkrupp, 2024). As shown in the TK example, Scope 3 emissions are challenging to assess as they come from many sources that a company cannot always directly influence. The accuracy and completeness have been criticised in current literature (Downie & Stubbs, 2013; Ducoulombier, 2021), and according to the research of Hertwich and Wood (2018), the Scope 3 emissions percentage of total emissions is highly variating across industries.

After this brief introduction to sustainability reporting, the reporting landscape and, in particular, the GHG Protocol, the core question of this thesis will be addressed. In the forthcoming chapter, a systematic literature review will be conducted to provide an overview of the existing research, thus establishing the relevance and validity of the research questions.

3. Systematic Literature Review and Academic Relevance

It is essential to contextualise the topic within current and past research to assess the relevance of this thesis. This is achieved with a systematic literature review examining the impact of financial performance on GHG emissions. The first section offers an overview of the systematic literature review methodology utilised, while the second section discusses the SLR findings and academic relevance of this thesis.

3.1. Systemic Literature Review

This chapter begins with a brief introduction to SLRs, followed by an explanation of the five-step methodological approach used to perform this SLR by Khan et al. (2003)

3.1.1. SLR Methodology

A systemic literature review is a "*clearly formulated question, identifies relevant studies, appraises their quality and summarises the evidence using explicit methodology*" (Khan et al., 2003, p. 118). The SLR's advantages are the transparency and reproducibility of research findings (Snyder, 2019), and it can help to systematically identify current studies, research approaches, trends and findings about the topic of this work: the impact of profitability on GHG emission levels. A five-step approach by Khan et al. (2003) is used to conduct the SLR, as it provides a clear structure to this research.

Step 1: Framing Questions for a Review

The research questions remain the same as presented in the introduction and are divided into two parts:

(1) *What are the Scope 1, 2 and 3 GHG emissions levels for European companies from 2017-2023?*

(2) *How does firm profitability impact total and individual Scope 1, 2 and 3 GHG emissions?*

The goal of the SLR is to systematically identify current research on these or similar topics, find research gaps, and assess the academic relevance of the research questions.

Step 2: Identifying Relevant Work

Relevant work is identified with a proper research strategy, including carefully selecting databases, defining key search terms, and systematically documenting the entire research process. Web of Science and Scopus were selected for the databases due to their wide range of academic articles and size. The search terms derived from the research questions above were organised into different blocks, summarising related terms in English to achieve optimal accuracy. Only English keywords were utilised in the search process, as prior analysis indicated that the most pertinent literature is predominantly available in this language. Albeit this thesis focuses on the European scope, the SLR will look for worldwide studies, to understand the global state of research. Table 1 provides an overview of the search terminology across the three identified blocks, effectively representing the research questions.

Consequently, each database of interest, Web of Science and Scopus, is subject to a query search, and all matching results exported to Endnote for a title and abstract screening in the next step. The exact search query can be found in Appendix 1.

Step 3: Assessing the Quality of Studies

This step involves critically evaluating the quality and relevance of the studies identified in the previous research step, using predetermined criteria for including or excluding studies. The inclusion criteria are outlined as follows:

Availability and Access: The papers must be accessible and available through the University of St. Gallen libraries.

Language: Papers must be written in English.

Date of Publication: The studies should be published between 1997 and 2024, aligning with the Kyoto Protocol's resolution, which marked a significant milestone in the global effort to combat GHG emissions (United Nations, 1998).

Relevance: The papers must be relevant to the research questions and align with analysing the relation between financial performance and GHG emissions, as determined by reviewing the titles and abstracts.

Publication Status: Only papers that are published and peer-reviewed in renowned journals will be considered.

The exclusion criteria automatically apply to papers not meeting the above inclusion criteria. This approach ensures

Table 1: SLR Search Term Matrix

	Block 1	Block 2	Block 3
Primary Search Terms	Financial Performance	GHG Emission	Firm
Related Terms	Profitability, Economic Performance, Financial Returns	Greenhouse Gas Emission, Carbon Emission, Carbon Footprint, Carbon Dioxide Equivalent (CO2e), CO2 Emissions, Carbon Dioxide Emission, Carbon Output, Scope 1, Scope 2, Scope 3	Companies, Company, Corporation, Corporate, Organisation, Organization

that only the most pertinent and credible studies are considered in the research. All studies identified in step 3 are classified with a four stars scoring system described below, allowing to classify the studies by direction of analysis (Profitability <–> GHG Emissions)

> * = **Excluded**, if one of the inclusion criteria is not met.
>
> ** = **Excluded**, gives a good overview of the relation of GHG emissions and financial performance but does not address the research question thoroughly enough.
>
> *** = **Included**, addresses the research question indirectly, analysing the link of corporate GHG emissions on the financial performance or more broadly, the relation between both variables.
>
> **** = **Included**, addresses the research questions directly, analysing the impact of corporate financial performance on corporate GHG emissions.

Step 4: Summarising the Evidence & Step 5: Interpreting the Findings

Upon reviewing the relevant academic literature, a comprehensive summary will be presented. This summary will highlight the topics previously explored by scholars, accompanied by a PRISMA Flow Diagram, based on Page et al. (2021), to illustrate the selection process from identified studies to those included in this literature review. Additionally, the studies gathered through the SLR will be leveraged to address the research questions and pinpoint existing gaps in the current research, thereby justifying the necessity of this study.

3.2. Findings and Academic Relevance

In the following chapter, the results of the literature review are presented. First, a brief introduction is provided, followed by an analysis of the number of studies found and their geographical and industry focus. Additionally, the used variables and the direction of analysis between the variables of financial performance and GHG emissions is discussed. After examining the theoretical frameworks employed, the findings of the studies are analysed and the chapter concludes with a summary.

3.2.1. Introduction

The systematic literature review approach ultimately identified 1,320 records, of which 69 were included in the final review sample. Figure 2, a PRISMA-Diagram, represents the SLR screening process and the number of records excluded during each step. The 69 studies included in the final sample were all reviewed in-depth. The following elements were collected, if applicable, in an Excel table: Geographical focus, industry focus, indices, research questions, hypotheses, study design, methodology, dependent variables, independent variables, moderating variables, control variables, Scopes of GHG emissions included, relation direction of GHG emissions to financial performance, findings, published year, time period of sample, sample size (in firms) and the theories used in the theoretical background.

3.2.2. Sample Overview

The systematic literature review identified 69 relevant papers from various industries, regions, and years. This chapter will first provide an overview of the sample of 69 studies before analysing their geographical and industry focus. As shown in Figure 3, the first relevant study was published in 2010, although the research was aimed at studies starting in 1997. The exact reason for this is unknown but probably attributable to the lack of relevant data before 2010 or the limited scope of the search of this SLR in terms of Databases or search terms. Nevertheless, a positive trend in the number of publications per year can be observed, and there is an increasing amount of research on the relationship between GHG emissions and financial performance.

The climate crisis is a global problem for all industries, and research in this field is expected across various regions and sectors. Figure 4 shows that most studies analysed samples of global companies without a specific regional focus. The studies with a country or regional focus are well diversified across high- and low-income regions, indicating that the topic is relevant to a global audience. For the industry focus of the studies, also shown in Figure 4, we identified the four categories: Multi-Sector, Manufacturing, CO2 Intensive, and Others. Most studies did not focus on a specific industry; ten focused on manufacturing, and five focused on what we classified as CO2-intensive sectors, which can include energy, oil and gas, and similar. As CO2-intensive sectors are, per definition, responsible for the majority of GHG emissions, a focus on these high-polluting sectors makes sense (Ritchie et al., 2020).

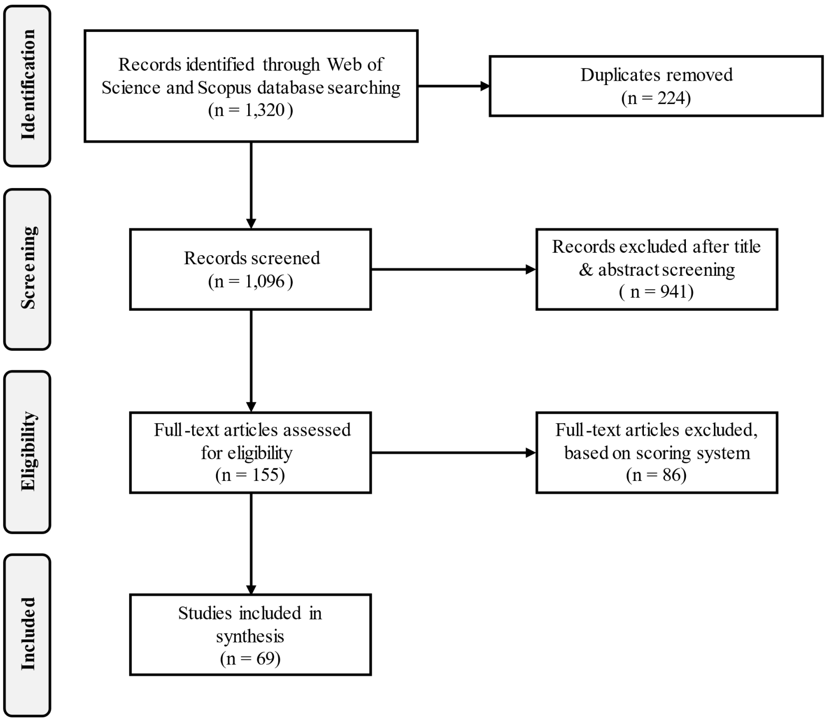

Identification
Records identified through Web of Science and Scopus database searching (n = 1,320)
Duplicates removed (n = 224)
Screening
Records screened (n = 1,096)
Records excluded after title & abstract screening (n = 941)
Eligibility
Full-text articles assessed for eligibility (n = 155)
Full-text articles excluded, based on scoring system (n = 86)
Included
Studies included in synthesis (n = 69)

Figure 2: PRISMA-Diagram of SLR, based on Page et al. (2021)

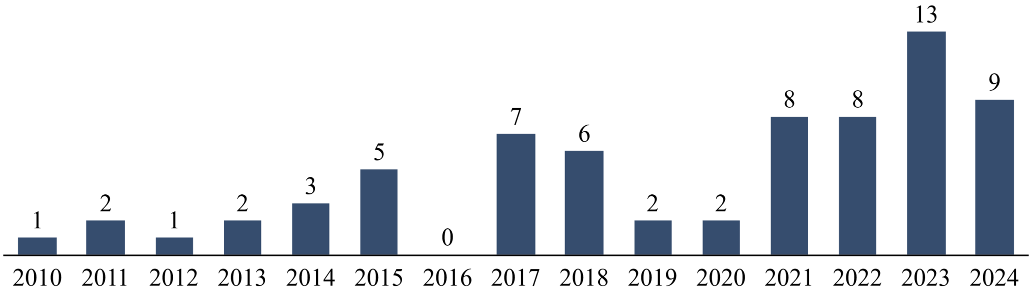


Figure 3: Distribution of Identified Studies between 2010-2023

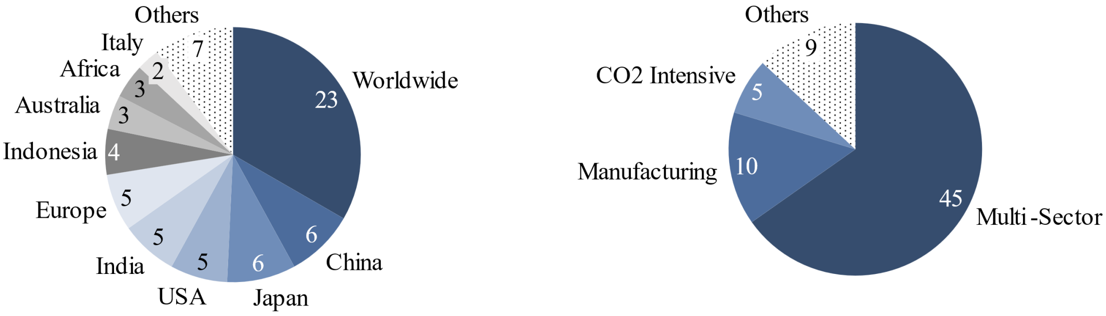


Figure 4: Number of Studies by Geographical and Industry Focus

In the systematic literature review sample comprising 69 studies, 62 are quantitative, while seven are qualitative. Focusing exclusively on the quantitative studies, the average sample size includes 523 firms, and the average data collection period spans 9.2 years. This calculation excludes the 11 studies with a data sample covering only one year. Concluding the broad overview of the study sample, the next chapter will focus on the most relevant types of variables used to analyse the relationship between the GHG emissions of companies and their financial performance.

3.2.3. Variables and Directions of Analysis

This thesis investigates the impact of profitability on GHG emissions across all Scopes. The primary objective of this SLR was to identify relevant studies examining this relationship and the broader correlation between GHG emissions and financial performance. Notably, 64 or over 90% of the studies reviewed focused on the unilateral impact of GHG emissions on financial performance rather than examining how factors such as profitability influence emission levels, therefore, classified as ***, as mentioned in the methodology. Similarly to our findings, Meng et al. (2023) noted the lack of literature on the impact of financial performance on carbon performance. Regardless of the direction of variable analysis, we will initially provide an overview of the metrics used to measure GHG emissions and financial performance. Additionally, we will briefly discuss the most commonly employed control variables in these studies to ensure that our analysis is grounded in scientifically recognised methodologies.

For most studies, GHG emissions represented the independent variable, and a variation of terms was used for similar or different metrics of GHG emissions, which can create some confusion at first glance. The commonly used terms to describe this variable are GHG emissions, which is used in this thesis, and environmental or carbon performance. Carbon performance is a widely used term, but the exact meaning or calculation can vary extensively, as it can represent total emissions (Busch et al., 2022) or emission intensity (Ganda, 2022; Meng et al., 2023). The metrics can be classified into two categories, mainly GHG emissions in absolute amounts (Ababneh, 2019; Ganda & Milondzo, 2018) or GHG emissions in relative amounts in relation to firm size metrics like sales or assets (Benkraiem et al., 2023; Di Pillo et al., 2017; Fujii et al., 2013). The absolute amount of GHG is an essential metric because the Kyoto Protocol and the Paris Agreement aim to reduce GHG emissions to fight climate change (United Nations, 2015a). Studies like Busch et al. (2022), Ganda (2022), and Homroy (2023) have been using the absolute amount of total GHG emissions or the change in GHG emissions between years as a variable to assess carbon performance. However, the relative amount introduces a notion of productivity in the metric, which is equally essential to achieving climate goals. Two different metrics commonly used to assess the relative carbon performance of a company are carbon intensity, defined as the total amount of GHG emissions divided by net sales or net assets, and carbon productivity, defined as net sales divided by Total GHG emissions (Benkraiem et al., 2023; Di Pillo et al., 2017; Ghose et al., 2023; P. Kumar & Firoz, 2018). Furthermore, some studies used the natural logarithm of the amounts above to counter the non-normal distribution of GHG emissions across firms (Busch & Hoffmann, 2011; Delmas et al., 2015; Desai et al., 2021; Houqe et al., 2022; Mahapatra et al., 2021; Raval et al., 2021; L. Wang et al., 2014). Unfortunately, as illustrated in Figure 5, 41 studies, representing approximately 59%, fail to specify what they mean by total GHG emissions, omitting any reference to the Scopes of GHG emissions considered in their analysis.

This lack of differentiation represents a significant research gap. Consequently, this study aims to address this and focus on all three Scopes of GHG emissions, recognising their distinct real-world implications. Each Scope is critically essential and necessitates different interventions from management and policymakers.

Financial performance or profitability is commonly approximated by two categories of metrics in the analysed studies: market-based and accounting-based metrics (Busch & Lewandowski, 2018). The most represented market-based metric is Tobin's Q (Busch & Lewandowski, 2018; L. Wang et al., 2014), a ratio comparing the market value of a firm to the replacement cost of its assets. This metric reflects a forward-looking perspective on profitability (Tobin, 1969). Tobin's Q incorporates future earnings expectations, making it a crucial indicator for assessing a firm's potential financial performance. However, for the scope of this research, we want to focus on actual and not expected earnings, which makes accounting-based profitability metrics better suited. The most represented accounting-based metric is Return on Assets (ROA) representing the operating performance of a firm (H. B. Chen & Manu, 2022; Gallego-Alvarez et al., 2015), followed by Return on Equity (ROE) representing the financial performance of a firm, Return on Sales (ROS), Return on investment (ROI) and similar (Galama & Scholtens, 2021; Velte, 2023; Q. Wang, 2023). However, financial performance is commonly used for all kinds of financial metrics and does not only refer to ROE.

In order to isolate the effects of the independent variable on the dependent variable, reduce biases, and improve the validity, the implementation of control variables is crucial for viable results in a regression model (Bernerth & Aguinis, 2016; Shibata, 1981). An excessive number of control variables can increase the complexity of the model (Gordon, 1968). Therefore, a focus on the most important ones observed in the SLR sample is taken. The first noteworthy control variable is related to firm size, usually approximated by the natural logarithm of total assets or sales (Busch et al., 2022; Delmas et al., 2015; Hassan & Romilly, 2018). The second is revenue growth from year to year (Raval et al., 2021; Rokhmawati & Gunardi, 2017). The third one is leverage, calculated in different ways but generally representing the amount of debt compared to total equity, assets, or profit measures (Raval et al., 2021; Rokhmawati & Gunardi, 2017). The fourth widely used control variable is capital intensity, calculated as capital expenditures divided by sales (Rokhmawati et al., 2017; Tarmizi & Brahmana, 2023). Last but not least, the fifth variable often represents the industry of a firm (Hassan & Romilly, 2018; Mahapatra et al., 2021; Shahgholian, 2019). As one of the few studies to incorporate variables from other domains, Velte (2023) analysed metrics related to corporate governance and their impact on GHG emissions. His findings suggest that board diversity may be an indicator of lower GHG emissions, making board diversity a potentially valuable control variable for our analysis.

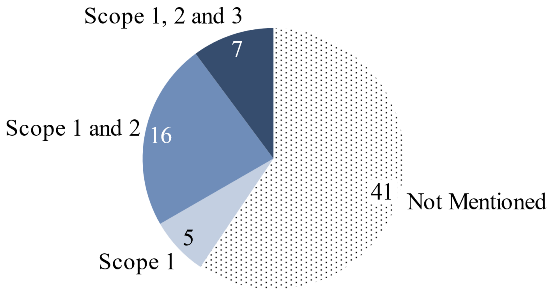


Figure 5: Number of Studies Distinguishing between the Scope 1, 2 and 3 GHG Emission Levels

3.2.4. Common Theoretical Frameworks in the Studies

This chapter will focus on the most common theories used to explain the relationship between GHG emissions and a company's financial performance. Two theories leading to potentially contractionary outcomes are the Neo-Classical Theory or Traditionalist View, and Porter's Hypothesis (Porter, 1980) or, similarly, the Revisionist View (Porter & van der Linde, 1995). Porter (1980) laid the foundation of this theory by stating that stringent environmental regulation can lead to better innovation. Further, he elaborated on the relationship between environment and competitiveness in Porter and van der Linde (1995). This theory argues for a win-win situation where reducing GHG emissions by regulation requires innovation, improving efficiency, and financial performance. The Traditionalist or Neo-Classical Theory argues for a win-lose relationship, as stringent environmental regulation leads to additional costs for firms, having a negative impact on profitability and competitiveness (Palmer et al., 1995). To build upon the win-win perspective, the Resource-Based View (RBV) (Barney, 1991; Wernerfelt, 1984) and Natural Resource-Based View (NRBV) (Hart, 1995) are also commonly used in many studies of our sample. Both theories argue that a company can achieve a competitive advantage by effectively using its internal resources, and the NRBV by Hart (1995) adds environmental sustainability to the framework. In this thesis context, this theory emphasises that focusing on better environmental sustainability and reducing GHG emissions creates a competitive advantage for a firm, potentially leading to higher financial performance.

Shifting from the economic-based to social-based theories, another critical theory to explain this relation is the Stakeholder Theory introduced by Freeman (1984), arguing for a shift from shareholders profit maximisation to the needs of a broader scope of stakeholders. In the context of the analysed relationship, this theory argues, particularly for firms with strong financial performance, that there is an increased societal expectation to engage all stakeholders. This encompasses, notably, reducing the environmental footprint and implementing environmentally beneficial initiatives associated with fewer GHG emissions. This argument is also supported by the Legitimacy Theory from Dowling and Pfeffer (1975), which discusses the relationship between organisations, stakeholders, and society. The theory posits that organisations strive to achieve and maintain legitimacy by aligning with societal norms, values, and expectations, an effect that is further strengthened by better financial performance (Akhter et al., 2023; Kuruppu et al., 2019). Given that climate change is unequivocally recognised as a global challenge, Legitimacy Theory suggests that firms are likely to voluntarily adhere to norms to reduce their GHG emissions to address this pressing issue.

Lastly, the Slack Resource Theory provides a robust theoretical foundation for understanding the anticipated impact of financial performance on a company's GHG emissions, which has been mentioned several times in the studies of this SLR. This theory, introduced by Cyert and March (1963), posits that firms with excess resources are better positioned to enhance performance across various dimensions, including environmental responsibility and GHG emission levels. Empirical research has established a positive association between slack resources and corporate social performance, suggesting that financial performance can facilitate improved environmental outcomes through the availability of slack resources (Waddock & Graves, 1997). This linkage underscores the integral role of resource availability in enabling firms to meet social and environmental objectives, thereby aligning financial success with sustainability goals. Having established the theoretical framework, the subsequent chapter will be devoted to a detailed examination of the study findings, underpinning the formulation of the research hypotheses for this thesis.

3.2.5. Findings of the Studies

The 69 studies were classified into 5 categories, indicating the main findings. To facilitate the comparability of study results, the categories were delineated as follows:

> **Negative:** There is a negative correlation between GHG emissions and financial performance, indicating that higher emissions are associated with poorer financial performance.
>
> **Positive:** There is a positive correlation between GHG emissions and financial performance, indicating that higher emissions are associated with better financial performance.

Mixed: A relationship between GHG emissions and financial performance is observed, but the direction of this relationship is unclear.

Non-linear: A non-linear relationship exists between GHG emissions and financial performance.

Not Significant: No relationship is found between GHG emissions and a company's financial performance.

This categorisation, shown in Figure 6, ensures a systematic and consistent approach to analysing and comparing the findings of various studies.

Out of 69 studies, 38 or 55% indicate a negative relationship between the amount of GHG emissions and the financial performance of companies. Desai et al. (2021) examined the impact of GHG emissions on financial performance utilising emissions data from the Carbon Disclosure Project, covering the period from 2013 to 2019 in India. Using Scope 1 GHG emissions to approximate the carbon footprint and both market- and accounting-based metrics for financial performance, the study indicates a significant negative relationship of GHG emissions on both measures. Another study by Gallego-Alvarez et al. (2015) analysed emission data from 89 companies between 2006 and 2009 and found a positive impact of GHG emission reduction on financial (defined by ROE) but not operational performance (defined by ROA).

Five studies associated higher GHG emissions positively with financial performance, favouring the neo-classical win-lose view mentioned in Chapter 3.2.4. Busch et al. (2022) analysed 4873 companies between 2005-2014 and used ROA as short-term and Tobin's Q as long-term metrics for financial performance as the dependent variable and total (direct and indirect) GHG emissions as the independent variable. The study found strong evidence of a positive relationship between GHG emissions and short-term financial performance (ROA) and long-term financial performance (Tobin's Q), indicating that higher emissions are associated with better financial performance. Similarly, L. Wang et al. (2014) examined 69 Australian companies and found a positive correlation between GHG emissions and financial performance across all industry sectors, stating that Australia's dominant mining industry could explain this finding. It is also important to emphasise that four studies reporting positive findings utilised data samples from single years. This methodological approach may compromise the comparability and validity of the results.

Several studies did not make clear, one-sided conclusions and found mixed results for the abovementioned relationship. A study from Bouaddi et al. (2023) found a difference in the effect depending on the size of a company, where the carbon emissions negatively affected small-size firms. However, the effect became positive with the increased size of the company. Another factor that seems to influence this relationship is a firm's industry. While a reduction in GHG emissions does lead to higher financial returns for firms in a sample of privately owned Australian firms, it does not appear to pay off for firms in environmentally sensitive sectors (Qian & Xing, 2018). Furthermore, two studies found a difference in the effect, depending on the metric used for financial performance. Delmas et al. (2015) found a detrimental impact of environmental performance on the present-time oriented ROA but a beneficial impact on the long-term oriented Tobin's Q, indicating a difference between short- and long-term effects of GHG emissions on financial performance. In contrast, van Emous et al. (2021) found lower GHG emissions to improve ROA, ROE, and ROS but no significant effect on Tobin's Q and a firm's current ratio. In conclusion, these mixed results indicate a further need to analyse the relationship between GHG and financial performance, considering the metrics used to approximate the variables, their time horizon, and the respective industries of the firms.

The last interesting finding not mentioned in previous studies is the potential non-linear relationship between GHG emissions and financial performance. Several studies have observed non-linear relationships both of U-shaped and inverted U-shaped form (Fujii et al., 2013; Misani & Pogutz, 2015; Ogunrinde et al., 2022; Tatsuo, 2010). Misani and Pogutz (2015) approximated financial performance by Tobin's Q and found that firms achieve the highest financial performance with neither too high nor too low carbon performance, indicating an inverted U-shaped relation. Similarly, a study involving Japanese manufacturing firms demonstrates an inverted U-shaped relationship between environmental and economic performance, signifying economic benefits from GHG reduction only up to a certain trade-off point. Adding to the mixed results and the influence of the firm industry discussed above, differences between non-linear relationships can also be observed depending on the industry emissions levels (Ogunrinde et al., 2022). For firms in the low-carbon sectors, an inverted U-shaped relationship between financial performance and carbon intensity seems to exist, whereas, for firms in high-carbon sectors, a U-shaped relationship seems to be the case (Ogunrinde et al., 2022). These studies highlight the complexity of the discussed relationships and the strong impact of influencing factors like a firm industry. However, the limited number of studies shows the necessity for further research on non-linear relationships.

Although the results sometimes appear contradictory, and many studies have mixed findings, a consensus seems to emerge, that GHG emissions negatively correlate with financial performance. However, the analysis predominantly focuses on the impact of GHG emissions on financial performance, with insufficient attention given to examining the potential reverse impact, reverse causality, or bidirectional relationship — how financial performance affects GHG emissions, or if profitability drives sustainability.

3.2.6. Impact of Financial Performance on GHG Emissions

As previously noted, there is a lack of research examining the impact of financial performance on GHG emissions. Six studies have addressed this relationship within the sample. Therefore, they are classified as ****, and their findings are outlined subsequently. Hassan and Romilly (2018) analysed

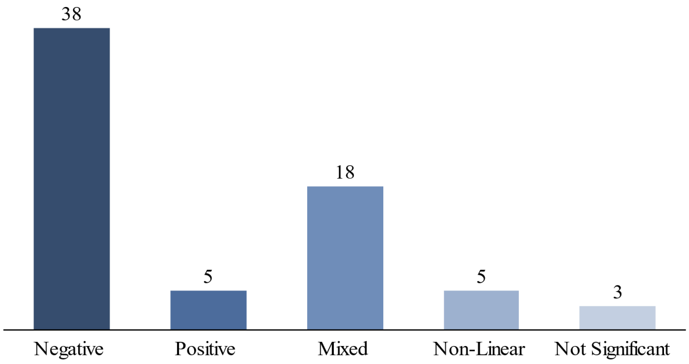


Figure 6: Relationship Types of Identified Studies

the relationship between corporate economic performance, environmental disclosure, and GHG emissions in different directions. Although a highly significant negative impact of GHG emissions on financial performance was found, the impact on the reverse relationship of financial performance on GHG emissions does not seem significant. Another study by Meng et al. (2023) with 352 Chinese companies, found that higher financial performance is linked to lower GHG emissions, which is contradictory to the previous study from Hassan and Romilly (2018). The impact of financial performance on the GHG emissions of companies was also analysed in combination with the R&D expenditures by Vaitiekuniene et al. (2024), which found a significant negative correlation between ROA and a relative measure of GHG emission. R&D expenditures were also negatively associated with GHG emissions (Vaitiekuniene et al., 2024), which is consistent with the Resource-based View Theory, according to which a company with more financial flexibility invests more in R&D, which can also lead to a reduction in GHG emissions (Hart, 1995). These findings indicate a potential effect of financial performance on GHG emissions. However, given the limited number of studies examining this specific relation, it is impossible to draw definitive conclusions. Albeit not analysed by many studies, studies analysing the impact of GHG emission on financial performance also mention this relationship and possible reverse or two-way causality of these variables as a limitation that could influence the results (Busch & Hoffmann, 2011; Endrikat et al., 2014; Gallego-Álvarez et al., 2014; Testa & D'Amato, 2017). Waddock and Graves (1997) have already drawn a bidirectional and reverse connection between corporate social and financial performance, but further research must be conducted. Consequently, additional research is necessary to provide a more comprehensive understanding of this potentially reverse relationship, namely the effect of profitability on sustainability, and this thesis aims to address this research gap.

3.2.7. Research Gap, Relevance and Conclusion

The systematic literature review identified 69 promising studies on the relationship between financial performance and GHG emissions, published between 1997 and 2024. In recent years, a growing body of research has examined the relationship between GHG emissions and financial performance. The metrics predominantly used to approximate financial performance include the accounting-based Return on Assets (ROA) and the market-based Tobin's Q, which sometimes yield similar but contradictory results. Studies employ absolute and relative measures to assess carbon performance, which are defined primarily by the level of GHG emissions, which also seem to influence the outcomes. The primary findings suggest a negative correlation, indicating that firms with higher GHG emissions tend to exhibit poorer financial performance. Furthermore, this relationship appears to vary based on industry, time horizon, and firm size. However, a potential two-way or reverse causation is mentioned, and a lack of literature on the opposite directional relation is identified. A potential reason for this one-sided research could be the interest of companies to understand the factors increasing profitability and simultaneously improving their environmental footprints, making the win-win argument of Porter's Hypothesis more intuitive to analyse.

Building on the identified research gaps and the calls for further research on the relationship direction, namely the lack of research on how profitability impacts GHG emissions, this study aims to explore whether higher profitability is linked to lower GHG emissions. A potential relation based on the Slack Resources, Legitimacy and Stakeholder Theory, which will be further discussed in the next chapter. Understanding the factors influencing GHG reduction, particularly low GHG emission levels, is crucial for enhancing the implementation of effective carbon reduction strategies and regulations and achieving the reduction goals of the Paris Agreement. Furthermore, the lack of differentiation between the three emission Scopes is a significant shortcoming of current studies, as all three Scopes and the underlying business reason for their emissions differ substantially from each other (WRI & WBCSD, 2004). Lastly, a higher data quality and quantity is expected for the most recent years, due to the upcoming CSRD (European Union, 2022). Consequently, this thesis will analyse the separate relation of profitability between all three GHG Scopes and provide an overview of corporate GHG emission levels from 2017 to 2023.

This research is particularly significant due to new European Union regulations, such as the Corporate Sustainability Reporting Directive (CSRD), which mandate GHG emission disclosure for many firms under the criteria shown in Chapter 2.2.3. By offering current information on corporate GHG emissions in Europe and examining the relationship between firm profitability and emission levels, this study provides valuable insights for regulators, company management, and other stakeholders. Nonetheless, the potential limitations of this systematic literature review, such as publication or reporting biases and the search strategy, might lead to incomplete or misleading conclusions. For instance, publication bias could result in overrepresenting studies with significant findings, while underreporting of non-significant results might skew the overall understanding of the relationship between GHG emissions and financial performance. Additionally, the scope and search strategy employed in this review may have inadvertently excluded relevant studies, thereby limiting the comprehensiveness and generalisability of the findings. Such limitations necessitate caution in interpreting the results and highlight the importance of further research to validate and expand upon these initial insights.

These findings will be extended with quantitative research explored in the second part of this thesis. Based on this review and a further discussion on the theoretical background, the exact research hypotheses will be developed in the next chapters.

4. Theoretical Background

Individual theories do not seem to do justice to the complex topic of factors influencing companies' environmental performance outlined in the literature. Therefore, this chapter introduces the Slack Resources Theory, together with the Legitimacy and Stakeholder Theory, as theoretical frameworks to explain this thesis' research topic.

4.1. Slack Resources Theory

The first theoretical concept this research will be based on is the Slack Resources Theory, introduced by Cyert and March (1963). This theory posits that companies with slack resources have a greater capacity to adapt to change and invest in opportunities (Bourgeois, 1981). Slack resources are related to firm performance and, more specifically, profitability, fitting the argument of this thesis (Daniel et al., 2004; George, 2005). This theory makes a solid foundation for the research questions this thesis aims to answer, namely, the impact of profitability on companies' GHG emissions. Within the context of this study, the assumption is that firms with slack resources, therefore higher profitability, are likely to invest more in sustainable initiatives, which should result in lower GHG emissions. Previous research by Oestreich and Tsiakas (2023) has concluded that more profitable companies tend to emit fewer GHG emissions than less profitable companies. On the other hand, financial constraints are linked to enhanced carbon emissions (Rehman et al., 2024). However, it is crucial to differentiate between direct and indirect emissions, as the extent to which companies can influence these with their resources varies significantly.

4.2. Legitimacy and Stakeholder Theory

The Legitimacy Theory originates from Dowling and Pfeffer (1975) and posits that "*organizations seek to establish congruence between the social values associated with or implied by their activities and the norms of acceptable behavior in the larger social system of which they are a part*" (Dowling & Pfeffer, 1975, p. 122), and has been linked to explain CSR behaviour of companies in the past and also recent literature (Bachmann & Ingenhoff, 2016; J. C. Chen et al., 2008; Deegan, 2002; Palazzo & Scherer, 2006; Patten, 2020). Firms demonstrating higher profitability often achieve superior CSR scores (Coelho et al., 2023). According to the Legitimacy Theory, this phenomenon can be attributed to the ability and inclination of profitable companies to align with prevailing social values and norms. This is also in line with the Stakeholder Theory introduced by Freeman (1984). This framework highlights the evolution of corporate focus from purely economic concerns to a broader consideration of various stakeholder needs, including environmental and ethical concerns. Furthermore, with higher profitability comes greater responsibility, which can be explained by more significant stakeholder pressure and firms more willing to adhere to this pressure (Jakhar et al., 2019). Also, the visibility and resources of profitable firms make them more likely to be targeted by stakeholder demands (Gold et al., 2022). Therefore, this thesis argues, that under the frameworks of Legitimacy and Stakeholder Theory, firms with higher profitability face increased pressure from stakeholders to comply with social norms, resulting in lower GHG emission levels. However, it is essential to note that CSR can also improve financial performance reversely, and the relation between GHG emission and financial performance might go both ways, as discussed in the findings of the SLR.

Building on the Slack Resources, Legitimacy and Stakeholder Theory, a company's profitability is expected to negatively correlate with GHG emissions, meaning that more profitable companies are expected to emit less GHG emissions. This hypothesis will be formulated and expanded in the next chapter.

5. Hypothesis Development

This chapter elaborates on the analysis and regression hypotheses of this thesis, based on the findings of the literature review and the theoretical background.

Before the empirical analysis of the relationship between profitability and GHG emissions, there is a need to understand the distribution and trends of GHG emissions in Europe, including the differences between each Scope. Therefore, the first research question is as follows: "*What are the Scope 1, 2 and 3 GHG emissions levels for European companies from 2017-2023?*". This overview will be the foundation for the work on the second research question and help understand the dynamics of and between Scope 1, 2 and 3 GHG emissions. To accomplish this, the initial section will concentrate on the emission disclosures, statistics, and distribution, highlighting their different implications. Particularly noteworthy will be the examination of Scope 3 emissions, as their calculation remains challenging (Fouret et al., 2024). GHG trends over time and distribution among sample companies will further complete the analysis. Lastly, a comparative analysis across sectors and countries will be performed, as differences across sectors and geographical locations are expected (Ghose et al., 2023). These insights will help to identify the trends in sectors, companies and countries.

The SLR analysed studies focusing on the relationship between financial performance and GHG emissions and found that studies mostly focus on the relationship direction of whether it "pays to be green" but revealed a lack of studies examining the impact of profitability on GHG emissions. The lack of studies on this reverse relationship, together with mentions of potential reverse and two-way causation, was discussed in several papers (Busch & Hoffmann, 2011; Testa & D'Amato, 2017; Waddock & Graves, 1997), motivating the following analysis. This thesis argues that financial performance impacts GHG levels, adding to the research needs of scholars in this field. Based on the Slack Resources, Legitimacy, and Stakeholder theories, we hypothesise that profitability significantly negatively impacts GHG emissions, especially focusing on this directional relationship. Furthermore, most studies in the SLR do not distinguish between the three Scopes of GHG emissions. This lack of differentiation is problematic because the sources and implications of GHG emissions vary extensively across Scopes 1, 2 and 3, requiring different approaches and policies for effective reduction (WRI & WBCSD, 2004). This lack of distinction in current literature limits stakeholder interpretation and relevance. Motivated by this gap, we aim to differentiate and analyse the impact of profitability on individual Scopes of GHG emissions. Hence, the second research question is formulated: "*How does firm profitability impact total and Scope 1, 2 and 3 GHG emissions performance?*". To answer this question, we divided it into four sub-hypotheses to distinguish the effects on total, Scope 1, Scope 2, and Scope 3 GHG emissions. The first hypothesis we make is the following.

> *Hypothesis 1: Firm profitability is negatively associated with Total GHG emission levels.*

According to Slack Resources Theory, firms with higher profitability would have more financial resources to invest in GHG emissions reduction, therefore suggesting lower Total GHG emissions (Cyert & March, 1963). Hassan and Romilly (2018) did not find an impact of economic performance on emissions, but other studies suggest a negative or bidirectional relation (Meng et al., 2023; Testa & D'Amato, 2017; Waddock & Graves, 1997). This relation will be tested with GHG emissions numbers from LSEG Eikon and a proxy for financial performance. The metric choice was ROA, the most used accounting-based metric in the analysed studies. Further methodological choices will be outlined in the next chapter.

A significant research gap identified is the lack of differentiation between the three GHG emissions Scopes. Therefore, hypotheses 2, 3, and 4 focus on the effect of firm profitability on specific Scopes. The second hypothesis focuses on Scope 1 emissions and is formulated as follows.

> *Hypothesis H2: Firm profitability is negatively associated with Scope 1 GHG emission levels.*

Scope 1 GHG emissions refer to direct emissions from owned or controlled assets (WRI & WBCSD, 2004). As highlighted in our theoretical background using the Slack Resources Theory by Cyert and March (1963), higher profitability can allow firms to invest more in reducing these direct emissions. Additionally, firms face pressure from stakeholders to maintain legitimacy by reducing their GHG emissions (Dowling & Pfeffer, 1975; Freeman, 1984). In contrast to *Total GHG* emissions, Scope 1 emissions are, per definition, more related to the specific firm assets (WRI & WBCSD, 2004) and specific industries (Ghasemi et al., 2023), which might lead to different results for this regression. Scope 1 GHG emissions are emitted mainly by companies with energy-intensive processes, like in the energy, material, or manufacturing industry, and the reduction and potential decarbonisation strategies include the shift to low-carbon fuels, Carbon Capture and Storage (CCS), Process Optimisation, and Innovation (Cavaliere, 2019). All these solutions are considered resource-intensive (Cavaliere, 2019) and might require higher profitability. Out of five studies identified during the SLR approximating carbon performance through Scope 1 GHG emissions, three have a negative, one a mixed, and one has no significant correlation to financial performance, indicating mixed findings. However, innovation advantages reducing GHG emissions and improving operational efficiency, as suggested by Porter's win-win Hypothesis (Porter & van der Linde, 1995), could also lead to higher profitability, creating a potential two-way relationship and biasing results. Like H1, a negative correlation between firm profitability and Scope 1 GHG emissions is expected, but some caveats may influence this relation.

The third hypothesis focuses on Scope 2 GHG emissions and is the following.

> *Hypothesis H3: Firm profitability is negatively associated with Scope 2 GHG emission levels.*

Scope 2 GHG emissions refer to indirect emissions from the consumption of purchased energy (WRI & WBCSD, 2004). Reducing Scope 2 emissions often requires switching to renewable energy sources, which has become cheaper than fossil fuel electricity in the last few years (IRENA, 2022). Reducing Scope 2 GHG emissions by buying renewable energy can be both cost-saving and a demonstration of environmental responsibility, making this decision relatively straightforward for companies. Furthermore, reducing Scope 2 emissions is seemingly easier for a company to achieve than for Scope 1 emissions, which might lower the impact of profitability on this relationship (Bricheux et al., 2024). Another influencing factor of Scope 2 emission levels is the market- or location-based calculation methodology mentioned in Chapter 2.3.2, which highly influences the emission numbers (brightest, n.d.; WRI & WBCSD, 2004). This highlights the problem of energy purchase agreements, which might reduce the Scope 2 GHG emission with the market-based approach. However, a local production facility could be operated exclusively with local CO2-intensive energy. The location approach calculates Scope 2 emissions based on the local energy mix and provides a better picture of the emissions, but it is more resource-intensive to be influenced by companies (Roston et al., 2024). To conclude, a negative correlation between firm profitability and Scope 2 GHG emissions is anticipated, as firms with greater financial resources can readily reduce these emissions. However, the impact of profitability is expected to be less significant for Scope 2 emissions than other emission Scopes, as it is influenced by factors such as the company's sector, energy needs, and the local energy mix.

The fourth hypothesis focuses on Scope 3 GHG emissions and is the following:

> *Hypothesis H4: Firm profitability is negatively associated with Scope 3 GHG emission levels.*

Scope 3 GHG emissions are all other indirect emissions in a company's value chain (WBCSD, 2011). While firms have no direct influence on Scope 3 emissions, they can pressure and innovate along the entire supply chain to reduce their footprint (Patchell, 2018), which is expected to be more likely with more resources, thus higher firm profitability (Koh et al., 2023). However, business choices like outsourcing heavily affect these emissions, making the calculation potentially complex and small-scale (Mytton, 2020; Radonjič & Tompa, 2018). Furthermore, due to the complexity and comparability challenges associated with calculating and determining Scope 3 emissions (Fouret et al., 2024), a high number variance is expected, which could negatively influence the regression performance. Scope 3 emissions are also highly dependent on the industry and the company's products (Günther et al., 2015). Currently, to the best of the author's knowledge, no studies have examined the correlation between Scope 3 emissions and financial performance, making this a novel perspective. The impact of profitability on Scope 3 emissions is expected to be negative if the data situation allows for a significant regression model.

Accordingly, the null hypothesis for *H1*, *H2*, *H3* and *H4* is formulated as follows, and would indicate that profitability is not or positively associated with the respective GHG emission category:

> *Hypothesis H0: Firm profitability is not negatively associated with Total GHG, Scope 1, Scope 2 or Scope 3 emission levels.*

Having established the hypotheses to be tested in response to the second research question, the subsequent chapter will delineate the methodological framework adopted for this study.

6. Methodology

This chapter outlines the methodological approach for analysing GHG emissions from European companies, which are used to address both research questions. It begins with an overview of the sample companies and the data collection, followed by a discussion of the selection of dependent, independent, and control variables. The chapter then details the model specifications used for regression analysis, focusing specifically on answering the second research question: *How does firm profitability impact total and individual Scope 1, 2 and 3 GHG emissions?*

6.1. Sample and Data Collection

In order to provide an overview of Scope 1, 2 and 3 GHG emissions of European companies and analyse the impact of profitability on these GHG emissions, we opted for the companies in the STOXX Europe 600 index as our sample. The STOXX Europe 600 represents the 600 largest companies in 17 European and is well diversified by industries (STOXX, 2024). Furthermore, Europe is still seen as a pioneer in sustainability reporting and environmental responsibility (Barbu et al., 2022), which makes us expect solid and comparable data. The data will be collected from the LSEG Eikon database for the financial years 2017 to 2023, as 2017 marks the first year the NFRD regulation became mandatory in European Union member states and is a significant milestone for non-financial reporting (European Union, 2014). The GHG emission variation during the COVID-19 pandemic years may pose a challenge (A. Kumar et al., 2022). However, the time horizon of 7 years will help to get consistent results and is not far away from the average time horizon of 9.2 years identified in the SLR. All relevant data points for this research were initially accessed through the LSEG Eikon platform to minimise the use of multiple sources and rely on systematically sourced information. However, GHG emissions data for most firms for the year 2023 was not available in the LSEG Eikon database. Consequently, if available, the 2023 GHG emissions data was manually collected from annual or sustainability reports for all firms with missing values. The GHG emission figures were reviewed and updated during data collection to account for any retroactive changes in previous years. This step was necessary to ensure accuracy, as changes in companies' calculation methods sometimes resulted in significant deviations. The full dataset is available in Appendix 1.

6.2. Dependent Variables

Although there is extensive literature on the impact of GHG emissions on financial performance, vice versa, it is not the case. Accordingly, the dependent variables will be the GHG emissions across all Scopes of sample companies, collected for the financial years 2017 to 2023. In line with the four hypotheses, four dependent variables representing the GHG emissions are used. Scholars have used absolute and relative measures for GHG, as highlighted in the SLR. This research will be based on absolute emissions levels, as used by Mahapatra et al. (2021) or Porles-Ochoa and Guevara (2023), because, ultimately, only a reduction of absolute GHG emissions can reduce climate change. The four dependent variables for each hypothesis are, therefore, Total GHG emissions, representing the sum of Scope 1, 2 and 3 emissions (*Total GHG*) for Hypothesis 1, Scope 1 GHG emissions (*Scope 1*) for Hypothesis 2, Scope 2 GHG emissions (*Scope 2*) for Hypothesis 3 and Scope 3 GHG emissions (Scope 3) for Hypothesis 4.

6.3. Independent Variables

As an analysis of the impact of profitability on the dependent variables is the aim of this study, Profitability is the independent variable for the regressions. However, there is no clear consensus on how to approximate profitability in literature, but Return on Assets (*ROA*) is seen as a common metric for an accounting-based, short-term measure of financial performance (Benkraiem et al., 2023; Busch et al., 2022; Delmas et al., 2015; Feng et al., 2024), and Tobin's Q as a common metric for a market-based measure of long-term financial performance (Busch et al., 2022; Hassan & Romilly, 2018; Houqe et al., 2022; K. H. Lee et al., 2015). Considering profitability as the independent variable, an accounting-based measure of profitability is more appropriate than a future expectation-based market metric like Tobin's Q, which is based on expectations rather than actual profits. Therefore, profitability will be approximated by *ROA*, calculated as net income by total assets. Furthermore, to check the robustness of our regression, we will also test our hypotheses with *ROE*, calculated as net income by shareholders' equity and lastly, *ROS* as net income divided by sales.

6.4. Control Variables

Studies investigating the relationship between financial performance and carbon performance, as identified in the SLR, have used a common set of control variables (see Chapter 3.2.3) and the ones used for this regression analysis are outlined subsequently. This study implements four control variables to account for other effects next to probability, influencing GHG emissions. First, firm size (*SIZE*) has been linked to better socially responsible behaviour (Waddock & Graves, 1997), and we use the natural logarithm of the firm's revenues to define this metric (Alvarez, 2012; França et al., 2023). Second, as discussed by Velte (2023), board diversity (*BOARDDIV*) can drive GHG emissions performance. It will, therefore, be included as a control variable, calculated as a percentage number of women to total board members. Third, capital expenditures is found to be an indicator of GHG emissions (Xia & Cai, 2023), and the relative measure of capital intensity (*CAPINT*), calculated as CAPEX divided by total sales, is used in numerous studies (Busch et al., 2022; Desai et al., 2021; Meng et al., 2023). Additionally, sales growth (GROWTH) is the last control variable that accounts for the potential increased GHG emissions associated with output growth. Sales growth is calculated as percentual annual changes in sales, in line with similar studies (Desai et al., 2021; Gallego-Alvarez et al., 2015; Ghose et al., 2023; Lewandowski, 2017). Furthermore, the industry type of a company is undeniably a significant determinant of the amount of GHG emissions (Ritchie et al., 2020), which is why a classification in *Low* and *High-Emission-Sectors* is performed. The 11 sectors from the GICS sector classification are used and divided into both categories. Consumer discretionary, energy, industrials, materials, and utilities as *High-Emission-Sectors*, according to MSCI (2023) and the sector analysis performed later in Chapter 7.1.2. Financials, information technology, consumer staples, real estate, communication services and health care, as *Low-Emission-Sectors*. The method with which these sectors will be accounted for is discussed in the model specification chapter, as an invariant dummy variable is not suited for the planned fixed-effect regressions model (Wooldridge, 2012, p. 484-492).

6.5. Model Specifications

To test the hypotheses regarding the impact of firm profitability on GHG emissions, we will employ multiple linear regression models using an Ordinary Least Squares (OLS) approach (Greene, 2019). As Shahgholian (2019) notes in a literature review study, endogeneity between dependent and independent variables is a significant risk in the analysed relationship, 48 out of 80 studies of their literature review check for endogeneity. Because panel data from 2017 to 2023 is used, a fixed effects model is chosen to control for firm heterogeneity and endogeneity of variables that could bias the results, such as industry-specific effects or inherent company policies towards sustainability that do not change over time (Greene, 2019). Furthermore, the Hausman test will be performed to test the fixed-effects against random-effects and deduce the proper fit of the model (Hausman, 1978).

This approach allows to test for between-company variations over the study period and is used by several studies with similar panel data (Iwata & Okada, 2011; Lewandowski, 2017; J. Wang et al., 2021), providing a more accurate estimation of the relationship between profitability and GHG emissions. Literature and publications agree that industry is a significant determinant of GHG emissions. However, a fixed-effect model cannot process such an entity invariant variable separately in the model, only in combination with all the other potential fixed-effects. Since varying effects between *Low-* and *High-Emission-Sectors* are expected, the dataset will be split into two parts: entities from *Low-Emission-Sectors* and entities from *High-Emission-Sectors*, similar to the approach of Ghose et al. (2023). This separation accounts for the expected differences between these sectors, as noted in the literature (Ghasemi et al., 2023). Each regression model will be performed separately on the high-emission and low-emission datasets and compared against each other's.

The fixed-effects regression models for each dependent variable are specified as follows:

$$\begin{aligned}\text{H1: Total GHG}_{it} &= \beta_1 ROA_{it} + \beta_2 SIZE_{it} + \beta_3 BOARDDIV_{it}\\ &\quad + \beta_4 CAPINT_{it} + \beta_5 GROWTH_{it} + \mu_i + \varepsilon_{it}\\ \text{H2: Scope 1}_{it} &= \beta_1 ROA_{it} + \beta_2 SIZE_{it} + \beta_3 BOARDDIV_{it}\\ &\quad + \beta_4 CAPINT_{it} + \beta_5 GROWTH_{it} + \mu_i + \varepsilon_{it}\\ \text{H3: Scope 2}_{it} &= \beta_1 ROA_{it} + \beta_2 SIZE_{it} + \beta_3 BOARDDIV_{it}\\ &\quad + \beta_4 CAPINT_{it} + \beta_5 GROWTH_{it} + \mu_i + \varepsilon_{it}\\ \text{H4: Scope 3}_{it} &= \beta_1 ROA_{it} + \beta_2 SIZEE_{it} + \beta_3 BOARDDIV_{it}\\ &\quad + \beta_4 CAPINT_{it} + \beta_5 GROWTH_{it} + \mu_i + \varepsilon_{it}\end{aligned}$$

Where:

- β_1, β_2, β_3, β_4 and β_5 are the coefficients for the independent and control variables.
- μ_i represents the unobserved company-specific fixed effects.
- ε_{it} is the error term.

The regression results will be interpreted based on the coefficients' sign, magnitude, and statistical significance. The primary focus will be on the coefficient of ROA to understand its impact on GHG emissions across all Scopes. A negative coefficient would support the hypothesis that higher profitability is associated with lower GHG emissions. Control variables will also be interpreted to understand their influence on GHG emissions. The next chapter will now start with the analysis of the collected dataset.

7. Analysis and Results

This chapter is divided into two sections representing the two research questions relevant to this work and will focus on the quantitative and empirical analysis of the collected panel data. The first chapter will analyse the Scope 1, 2 and 3 GHG emissions of European companies, and the second chapter will focus on the impact of profitability on these GHG emission Scopes.

7.1. Quantitative Analysis: Scope 1, 2 and 3 GHG Emissions in Europe

Before delving into an in-depth analysis of Scope 1, 2 and 3 emissions, providing a brief introductory overview of the GHG emissions disclosures across the sample companies is essential. The following chapters will focus on the distinction between the individual Scope 1, 2 and 3 GHG emissions figures within the dataset. Before a detailed analysis of these figures, an overview of the data's quality and quantity will be provided. This overview will be the foundation for a more detailed examination of the emission numbers.

7.1.1. Overview of the GHG Emission Disclosures

As previously highlighted, the high quality and availability of data were anticipated due to stringent European regulations, Europe's dominant role in sustainability reporting, and corresponding research in Europe (Singhania & Chadha, 2023). An initial indicator of this data quality is the number of data points available for each company and each year, illustrated in Figure 7.

A clear trend of increasing data availability over the years is observable, as many companies disclose their GHG emissions across all Scopes. These results complement the findings of Barbu et al. (2022), which analysed the evolution of non-financial reporting and the impact of the NFRD on disclosures of European companies and found a positive influence over time. The slight decrease in data points for 2023 can likely be attributed to the manual collection of data from annual and sustainability reports rather than to an actual decline in data point numbers. With a maximum of 600 data points per Scope, constrained by the number of companies in the STOXX Europe 600 index, 98% reported their *Scope 1* and *Scope 2* GHG emissions in 2022, and 89% reported their *Scope 3* emissions. Notably, Scope 3 emissions were significantly less frequently published than *Scope 1* and *Scope 2* in the initial years, but this disparity has markedly narrowed recently. The same goes for the other Scopes, where a high disclosure increase has occurred. A positive trend was expected since all companies of our sample will be required to disclose their GHG emissions across the three Scopes when the CSRD comes into action for the FY2024 disclosures (European Union, 2022). Having addressed the availability of the data, we shall now examine the reported figures in depth.

7.1.2. Analysis of the Scope 1, 2 and 3 GHG Emissions

This chapter commences with a descriptive statistical analysis of the dataset to provide an overarching view of the data. Subsequently, a trend analysis is conducted to compare the dynamics of *Scope 1*, *Scope 2* and *Scope 3* emissions over time. Finally, a comparative analysis by sector and country is performed before the chapters end with a brief conclusion.

Descriptive Statistics

Before delving deeper into the data, it is essential to examine some basic statistics to better understand the dataset. Table 2 provides an overview of the most important numbers, which allows for the first insights.

Based on the mean emission numbers for each Scope over 2017–2023, it is possible to calculate the average share of total emissions of all Scopes. Figure 8 visualises the average total reported GHG emissions shares and shows the size differences between the three Scopes.

Scope 3 emissions have by far the most significant share of all three scopes, making up 88.5% of average *Total GHG* emissions, which is in line with expectations and findings in the literature (Matthews et al., 2008). *Scope 1* GHG Emissions, also called direct emissions are directly emitted by the companies and account for about 9,5%, and *Scope 2* represents

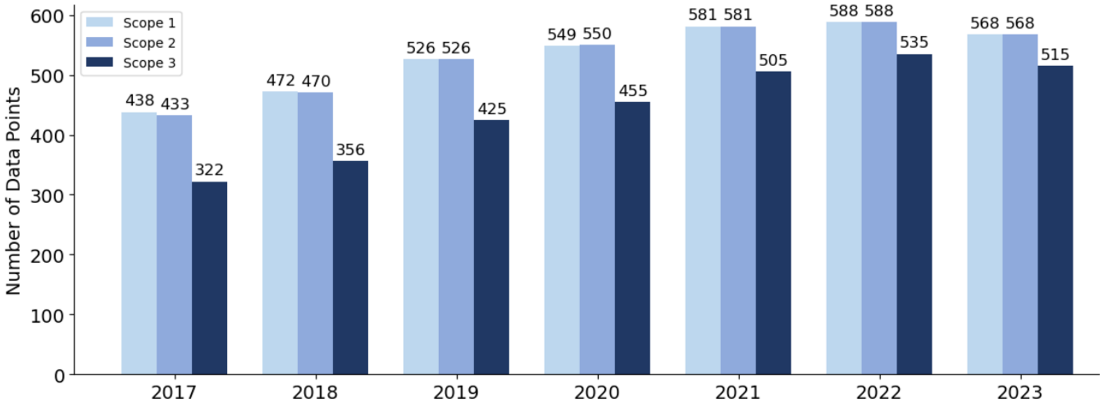


Figure 7: Number of Scope 1, 2 and 3 Data Points per Year

Table 2: Summary of Statistics for Scope 1, 2 and 3 GHG Emissions (in tons CO_2e)

Emission Types	N	Mean	SD	Min	25%	Median	75%	Max
Scope 1	3,722	2,611,755	12,406,433	0	4,859	33,833	269,580	179,700,000
Scope 2	3,720	431,302	1,425,410	0	8,513	46,078	225,205	22,057,000
Scope 3	3,113	24,466,782	126,861,683	0	31,500	539,638	6,300,000	2,823,000,000

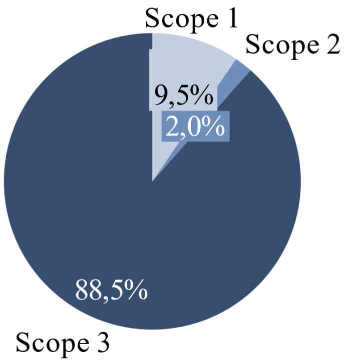


Figure 8: Average Share of the Three Scopes of Total Reported GHG Emissions

the purchased energy by companies, which represents the smallest portion of *Total GHG* emission in the sample, with about 2,0%. Furthermore, the wide range of values across all Scopes is notable, for example, with *Scope 1* emissions where values between the 25th and 75th percentiles differ by a factor of 55. This variability is expected, given the significant differences in the sizes of the companies within the STOXX Europe 600 index. However, this extensive range presents challenges for regression analysis performed in later chapters, as it can lead to heteroscedasticity (Gallego-Alvarez et al., 2015). To address this issue, we will apply natural logarithms to the emission values in our regressions, as discussed in Chapter 6.2. This transformation will help normalise the data and mitigate the impact of extreme values (Wooldridge, 2012). To ensure the comparability and quality of GHG emission data, the variance within entities is a crucial metric, as it helps to understand the deviation of these numbers from the mean. In this context, a high variance would suggest significant variability in GHG emissions across a specific Scope within an entity and over time, potentially undermining the reliability of the data or indicating significant changes in the calculation methodology. As mentioned in Chapter 5, we expect some challenges in the data quality of *Scope 3* emissions, as the calculation is complex and allows for a higher margin of discretion than *Scope 1* and *Scope 2* emissions. Figure 9 represents the standard deviation (SD) as a percentage of the mean within the GHG emission numbers of each entity in the data set from 2017 to 2023, and a clear difference between the Scopes can be seen.

The lowest standard deviation is observed for *Scope 1* emissions, with an average of 29.63% deviation from the mean, followed by *Scope 2* emissions at 37.41%. In contrast, *Scope 3* emissions exhibit a significantly higher standard deviation of 56.15%, indicating considerable variance among the *Scope 3* values reported within companies over time. During the manual collection of the latest 2023 values, this high deviation became apparent and is likely due to the lack of clear guidance, incomplete composition, and measurement diver-

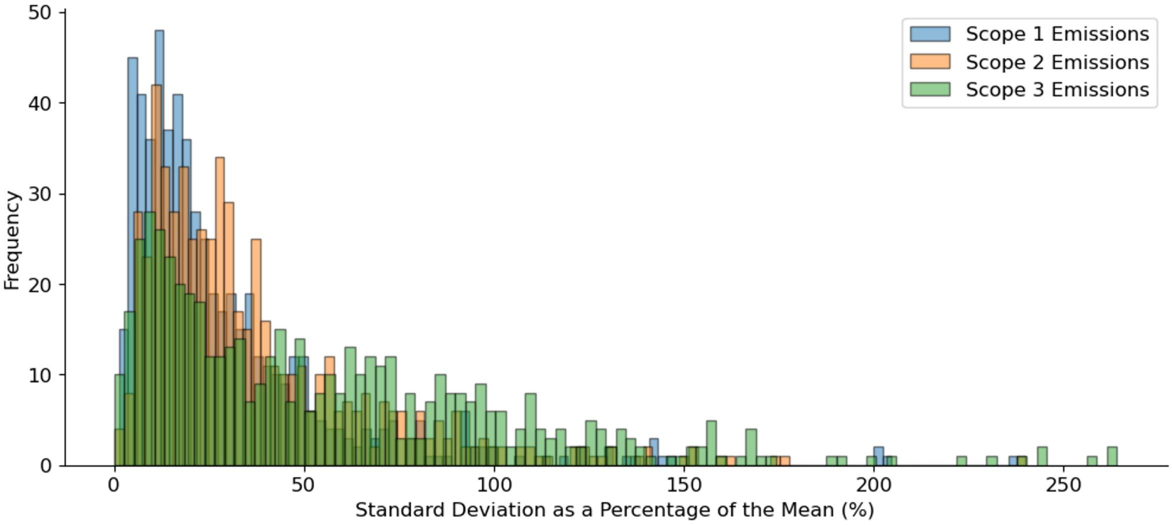


Figure 9: Comparison of the SD of Emissions as Percentage of the Mean by Entity

gence, three main problems cited by Nguyen et al. (2023), which studied the data quality of *Scope 3* emissions. The graph shows that the standard deviation distribution of *Scope 3* emission is less left skewed than for *Scope 1* and *Scope 2*, indicating more extreme values and underscoring the need for further standardisation of *Scope 3* emission reporting to enhance comparability in the future. Despite the high fluctuations in *Scope 3* emissions, the number of companies publishing all three scopes is promising at around 89% in 2022, which should allow us to conduct solid analyses afterward. In the next section, we will look at striking trends in the dataset.

Emission Trends by Scopes

In this analysis, shown in Figure 10, the trends in *Scope 1*, *Scope 2*, and *Scope 3* GHG emissions from European companies from 2017 to 2023 are examined. These findings allow us to gain insights into the carbon performance of these companies and will ideally identify a trend of negative GHG emissions growth. However, before delving into these trends, it is crucial to highlight certain peculiarities of the unbalanced panel data to avoid false interpretations. Specifically, the number of companies reporting their GHG emissions has increased over the years, making it impractical to analyse the trend in total emissions over the entire timeframe, as it would distort the results. Consequently, variables independent of the total number of reporting companies each year ensure a more meaningful analysis. To avoid the problem of outliers, the median growth rates of all three Scopes over the years are plotted in Figure 10, which offers interesting insights.

The initial observation is the pronounced decline in growth rates in 2020, marked by negative growth rates of -8.89% for *Scope 1*, -10.45% for *Scope 2*, and -10.28% for *Scope 3* GHG emissions. This decline was followed by a subsequent recovery beginning in 2021. In 2022 and 2023, the trends normalised, showing slightly negative growth rates for *Scope 1* and *Scope 2* GHG emissions, while *Scope 3* emissions exhibited a growth rate of approximately 1%. The exact growth rates can be found in Table 10 of Appendix 2. The sharp decline in GHG emission levels in 2020 is in line with expectations of the effects of the COVID-19 Pandemic. A. Kumar et al. (2022) analysed the impact of COVID-19 on GHG emissions and found similar results. These results were attributed, among other factors, to a decline in energy consumption, mobility, trade, and economic output (A. Kumar et al., 2022). In conclusion, the analysis reveals that *Scope 1* and *Scope 2* emissions have negative median growth rates between 2017 and 2023, registering at -2.02% and -5.44%, respectively. In contrast, *Scope 3* emissions exhibited a median annual growth rate of 1.37%, which can probably be attributed to the increasingly comprehensive methodologies employed in the calculation basis and the other challenges of *Scope 3* GHG emissions mentioned in Chapter 2.3.2 and 5.

Absolute Emission Levels by Companies

As mentioned, the absolute sum of GHG emissions per year is not a viable metric for unbalanced panel data. However, some absolute emission data comparison would provide valuable insights into the biggest GHG emitters of the STOXX Europe 600 index. To account for unbalanced numbers of entries per company, we opted for the average sum of *Scope 1* and *Scope 2* GHG emissions and the average revenues over the timeframe, plotted in Figure 11, helping to visualise significant outliers. *Scope 1* and *2* emissions are, per definition, the ones directly attributable to a firm. Therefore, the combination is an often-used metric to compare GHG emission levels across companies. These emissions are visualised against the average revenues of each company to provide a firm size metric as orientation.

Figure 11 illustrates the distribution of *Scope 1* and *Scope 2* GHG emissions among European companies. The data in-

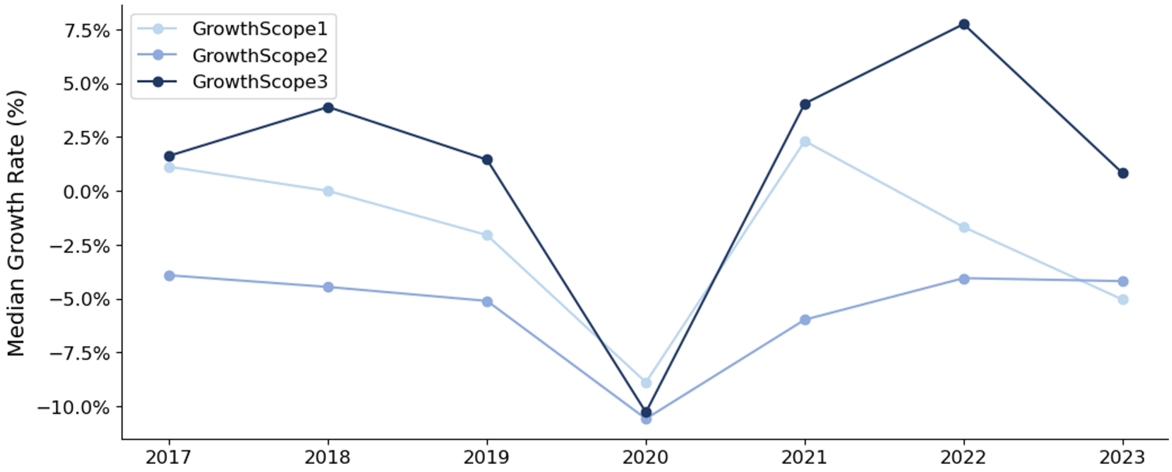


Figure 10: Median Growth Rates per Year, per Scope

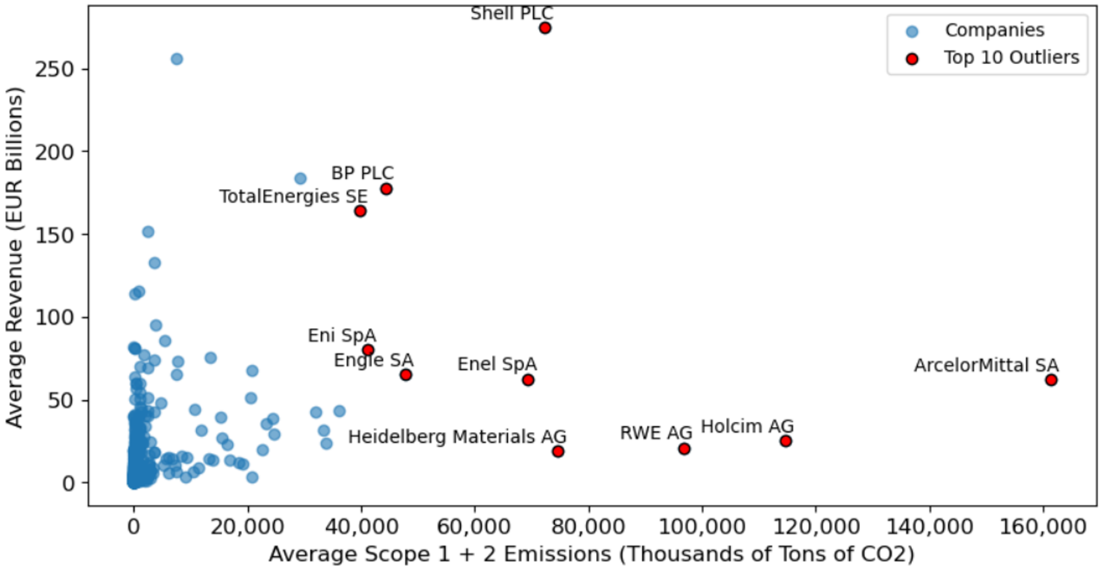


Figure 11: Average Scope 1 + 2 Emissions vs Revenues (2017-2023)

dicates that most companies generate low to medium GHG emissions. However, a significant portion of the emissions is concentrated among a small number of outliers. Specifically, the top ten companies with the highest emissions have been identified and labelled in the figure. Among these, seven companies belong to the energy sector, while the remaining three are in the materials sector, with two specialising in cement production and one in steel manufacturing. A report published by the CDP supports these findings, showing that 100 companies from the fossil fuel sector have been responsible for over 70% of industrial GHG emissions since 1988 (Griffin, 2017). Interestingly, when *Total GHG* instead of *Scope 1 + 2* emissions is plotted, the top 10 outliers change substantially. Figure 15 shows this plot and can be found in Appendix 2. After adding *Scope 3* GHG emissions, the energy sector is still dominant, but the materials sector not anymore. Firms with high fossil fuel consumption in their product life cycles predominate the list, including Airbus SE, Volkswagen AG, Siemens Energy, Siemens AG and Rolls-Royce Holdings PLC.

This distribution suggests that industry type and associated business models play a crucial role in determining a company's emission levels. It is also the reason for the decision to analyse the impact of profitability on GHG emissions on low- and high-emission companies separately. To further explore the relationship between industry sectors and emissions, the following chapter will examine the distribution of GHG emissions across various European sectors in greater detail.

Sector Analysis

Having observed that companies within the energy and materials sectors exhibit significantly higher GHG emissions than those being in other sectors, GHG emissions across different sectors in the sample are examined, because they are seemingly major determinants of absolute GHG emissions. Instead of using average figures as in the previous chapter, this industry analysis will rely on the most recent data from 2023. To facilitate meaningful comparisons without excessive granularity, we have adopted the Global Industry Classification Standard (GICS), categorising companies into 11 sectors. This classification is widely utilised by financial professionals, investors, and researchers due to its consistency and comprehensiveness (Bhojraj et al., 2003). The GICS 11-sector framework balances avoiding excessive fragmentation and capturing essential distinctions among different sectors. Figure 12 shows the Scope 1, 2 and 3 GHG emissions distribution across all 11 sectors for 2023.

For *Scope 1* emissions, displayed in Figure 12, the sectors materials, utilities, energy, and industrials are responsible for approximately 94.65% of the total *Scope 1* emissions, leaving the remaining seven industries to account for only 5.35%. This suggests that these four sectors significantly contribute to direct emissions through their business models, a fact also observed by other emissions reports (Polizu et al., 2023). In contrast, for *Scope 2* emissions, the materials sector is the most significant contributor, responsible for about 45.61% of the total *Scope 2* emissions. This indicates that companies in this sector purchase substantial amounts of energy for their business activities. The remaining emissions are more evenly distributed across the other sectors compared to *Scope 1* emissions. Utilities rank second with 13.76%, while all other industries contribute less than 10% of *Scope 2* emissions. *Scope 3* emissions, also referred to as value-chain emissions, have only gained attention in recent years, when the GHG Protocol published the Corporate Value Chain (Scope 3) standard in 2011 (WBCSD, 2011). However, their significance in the context of global GHG reduction is substantial (Matthews et al., 2008). As demonstrated in Chapter 7.1.2, *Scope 3* GHG emissions constitute approximately 90% of the *Total GHG* emissions for companies within the STOXX 600 Europe index. Consequently, their reduction is crucial to attain the goals of the Paris Agreement (United Nations, 2015a) and the responsibility lies with the companies and their respective business models. The sectors causing the highest amounts of *Scope 3* GHG emissions are industrials (32.79%), energy (27.72%), materials (13.68%), and consumer discretionary (13.22%). The consumer discretionary sector includes major automotive firms like Mercedes-Benz, Volvo, BMW, Stellantis, and Volkswagen, which report high *Scope 3* emissions due to the emissions in their value-chain and product life cycles (Wells & Nieuwenhuis, 2012). Similar to the distribution observed for *Scope 1* emissions, a few sectors are responsible for most GHG emissions.

In addition to examining the distribution of total emissions by sector, analysing GHG emission growth rates per sector serves as a valuable complement, as it shows the current trends. Figure 13 presents an overview of the median growth rates for all three Scopes across the 11 GICS sectors, with the number of data points per industry depicted in the *Scope 1* histogram. Between 133 and 805 data points represent each sector over the period from 2017 to 2023. This substantial dataset ensures the robustness of the median against outliers and provides meaningful insights into the trends of GHG reduction performance across different sectors.

The figures reveal that growth rates for *Scope 1* and *Scope 2* emissions are negative across all sectors, albeit with varying magnitudes. Notably, the financial sector exhibits the highest reduction rates for both types of emissions. For *Scope 1* emissions, the high-emission sectors, as shown in Figure 13, display relatively modest rates of decline between 0% and -1,5%, with the utilities sector as an outlier in the group, achieving the third best reduction rate at approximately 5%. Conversely, *Scope 2* emissions show significantly higher reduction rates across all sectors. This suggests that companies may find it easier to mitigate *Scope 2* emissions than *Scope 1* emissions, particularly in energy-intensive sectors such as materials, energy, and industrials. Improving *Scope 1* emissions often requires enhancing the energy efficiency of industrial processes, while *Scope 2* reductions can be more easily achieved through green energy purchase agreements, as highlighted by McKinsey in their report on the consumer goods industry (Bricheux et al., 2024), or a lower energy grid GHG footprint. Regarding *Scope 3* emissions, the growth rates for most sectors are positive, aligning with the trends discussed in Chapter 7.1.2. Unexpectedly, the information technology sector demonstrates the highest growth rates by a considerable margin. This anomaly may indicate that some companies have revised their assessment methodologies, leading to a higher attribution of *Scope 3* in this sector. The shift to cloud services by many IT companies could also be the reason for this trend, as such emissions are categorised as *Scope 3* under the GHG Protocol (WBCSD, 2011). This phenomenon has been previously examined in the literature (Mytton, 2020), and the results depicted in Figure 13 match these findings.

In conclusion, as observed in the previous chapter regarding company-specific emissions levels, the sector analysis shows that a limited number of sectors and companies are responsible for most GHG emissions. From both regulatory and research perspectives, focusing on these high-emission sectors is advisable, as they each require tailored solutions based on their specific business models and types of emissions. This development could also be observed in the SLR performed in the earlier chapters, where about one-quarter of the studies focused solely on companies in CO2-intensive sectors. Furthermore, regulations and policies have been increasingly targeted at these high-emission industries, with successful emission reductions (Pan et al., 2024; Yin et al., 2024), which also speaks in favour of our findings and the willingness to reduce global GHG emissions. In summary, the negative growth rates observed are encouraging in the context of combating climate change. However, this study does not assess whether these trends align with the climate targets outlined in the Paris Agreement. Additionally, the significant increase in *Scope 3* emissions within the IT sector highlights the potential for companies to shift their *Scope 1* or *Scope 2*

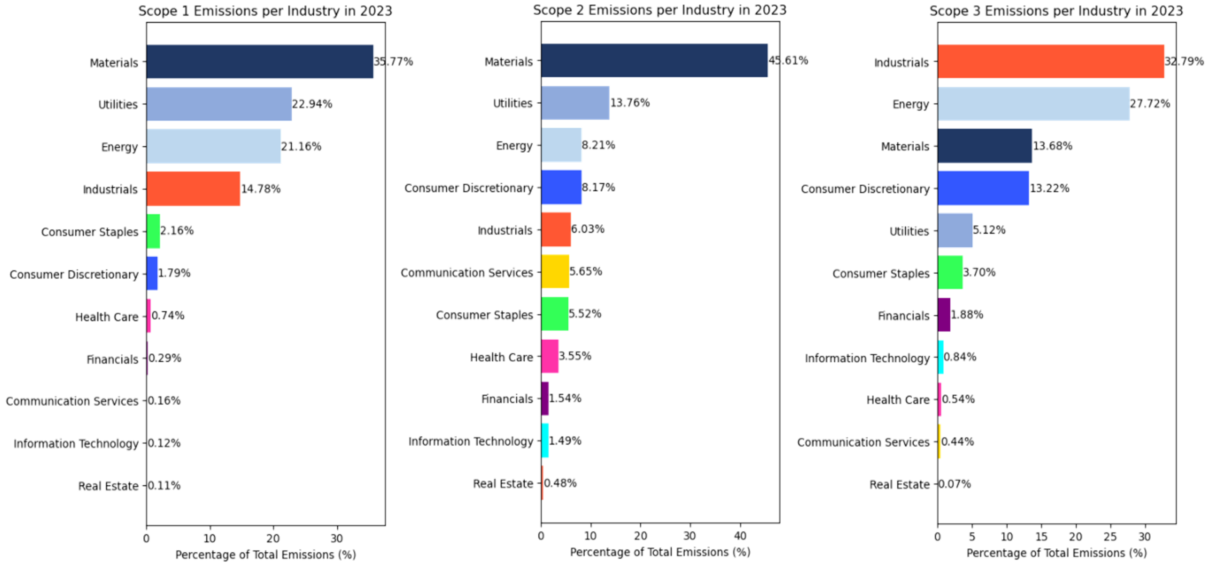


Figure 12: Distribution of Scope 1, 2 and 3 GHG Emissions across the GICS Sectors

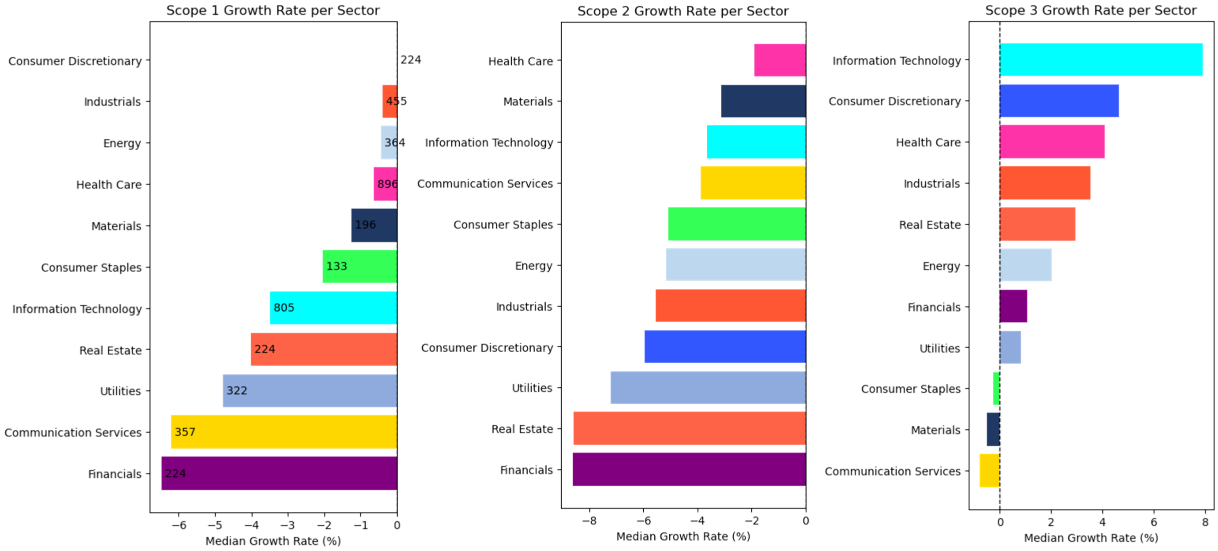


Figure 13: Median Growth Rates per Sector, per Scope

emissions through business practices. This underscores the importance of accurately accounting for *Scope 3* emissions to obtain a comprehensive picture of a company's environmental impact. Addressing these issues requires the collaboration of policymakers, regulators, and other stakeholders, who must respond swiftly to emerging trends across various sectors, making this sector analysis with the latest numbers from 2023 a valuable source of information.

In the concluding section of this analysis, we will investigate countries' Scope 1, 2 and 3 GHG emissions.

Country Analysis

In this chapter, we will analyse the growth rates of companies within the STOXX Europe 600 index by country. Consistent with the methodology employed in the previous chapter, the median growth rate has been used as the benchmark. The country-specific analysis in Figure 14 parallels the industry-specific breakdown across all three Scopes of emissions.

Notably, *Scope 1* emission reduction rates are generally lower than those for *Scope 2* emissions, with exceptions observed in countries such as Poland, Italy, and Belgium. Acknowledging that the STOXX Europe index encompasses only a limited selection of companies per country is essential. Thus, the provided chart offers insights specific to these companies rather than the national economy. In the case of *Scope 3* emissions, it is noteworthy that Germany exhibits a negative median growth rate despite having a significant number of companies represented. This suggests that German companies, or the industries prevalent in Germany,

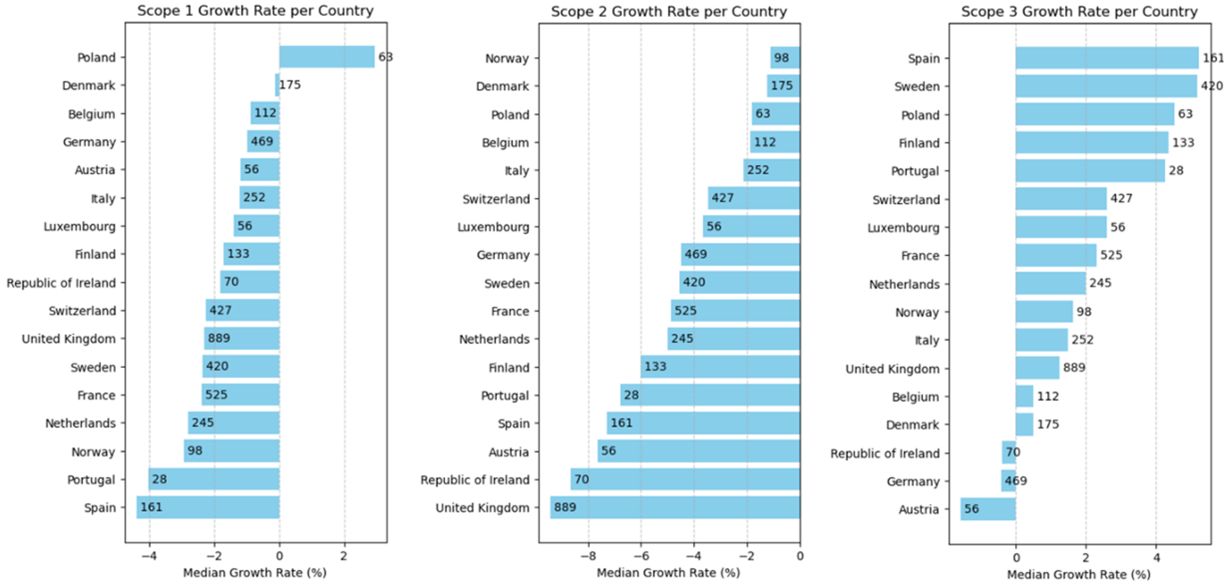


Figure 14: Median Growth Rates per Country, per Scope

may prioritise the reduction of Scope 3 emissions more than their counterparts in other countries included in the sample. While Ireland and Austria also show negative median growth rates, the limited number of data points for these countries makes the median growth rate less reliable.

Before proceeding to the second analytical section of this study, which will examine the impact of profitability on Scope 1, 2 and 3 GHG emissions, a summary of the key findings from the previous chapters is provided.

7.1.3. Key Findings and Discussion of Scope 1, 2 and 3 GHG Emissions in Europe

In this chapter, we quantitatively analysed the Scope 1, 2 and 3 GHG emissions of European companies within the STOXX Europe 600 index. Our investigation into the data quality and availability from 2017 to 2023 revealed that European companies demonstrate a high level of compliance with GHG emissions disclosures, a result of stringent European regulations (Barbu et al., 2022; European Union, 2014, 2022). The consistency in the number of companies reporting *Scope 1* and *Scope 2* emissions was notable, and there was a significant increase in the reporting of *Scope 3* emissions over the years. The descriptive statistical analysis underscored that *Scope 3* emissions constitute the largest share of *Total GHG* emissions, which aligns with expectations, given their broader definition and other publications on their respective share (Matthews et al., 2008). However, the data variability is substantial between and within companies, especially for *Scope 3* GHG emissions, as the standard deviation analysis revealed. This considerable variance in *Scope 3* values reported by companies underscores the need for further standardisation in *Scope 3* emission reporting to enhance comparability, as discussed by Nguyen et al. (2023).

The trend analysis illustrated a pronounced decline in growth rates in 2020, which can be attributed to the effects of the COVID-19 pandemic, as publications from other scholars also indicate (A. Kumar et al., 2022). This decline was followed by a recovery beginning in 2021. Overall, *Scope 1* and *Scope 2* emissions have negative median growth rates of -2.02% and -5.44%, respectively, from 2017 to 2023. Conversely, *Scope 3* emissions exhibited a median annual growth rate of 1.37%, which can be partially attributed to the challenges of *Scope 3* emission calculations (Nguyen et al., 2023). Examining absolute emission levels by companies revealed that a small number of outliers contribute significantly to the total emissions. Specifically, the energy and materials sectors were identified as the primary contributors. This finding was further substantiated by the sector analysis, which revealed that the sectors of materials, utilities, energy, and industrials are responsible for approximately 94.65% of total *Scope 1* emissions. Similarly, *Scope 2* emissions were predominantly from the materials sector, which accounted for about 45.61% of the total. A few sectors also dominated the distribution of *Scope 3* emissions, and the median growth rate for *Scope 3* emissions in the IT sectors was unexpectedly the highest by a high margin with about 8%. A potential reason could be the fast shift to cloud application in the whole sector, a trend that has already been the subject of other scientific work (Mytton, 2020). This analysis underscores the significant impact of business practices on the distribution of emissions across different Scopes. It highlights the critical importance of mitigating *Scope 3* emissions to meet the climate objectives outlined in the Paris Agreement. Therefore, it is not just important, but urgent for policymakers and other stakeholders to continuously monitor prevailing trends. This will enable them to respond effectively with appropriate measures when necessary, ensuring we stay on track to meet our climate goals.

Before concluding the chapter, it is essential to address the limitations of the data. The figures for the emission distribution across sectors are from a single year, making them potentially sensitive to outliers. *Scope 3* emissions can vary significantly within and between companies in the same industry. Furthermore, the median growth rate was selected due to extreme outliers, which can significantly distort the mean average and complicate its interpretation. Consequently, the evolution of *Total GHG* emissions by Scope may differ substantially from the patterns suggested by median growth rates. Given the unbalanced nature of the panel data set and the high variance observed between and within entities, the median growth rate remains the most suitable metric for this analysis. Additionally, the sample comprises companies represented in the STOXX Europe 600, which includes the 600 largest companies in Europe. This could result in an undervaluation of fragmented sectors with many small companies and an overvaluation of industries dominated by a few large firms.

In conclusion, this chapter provided a detailed examination of the Scope 1, 2 and 3 GHG emissions of European companies between 2017 and 2023, answering the first research question: *What are the Scope 1, 2 and 3 GHG emissions levels for European companies from 2017-2023*, highlighting significant trends and variances across sectors, countries, time and companies. The findings underscore the importance of continued efforts to standardise emission reporting and the need for targeted regulatory measures to address *High-Emission-Sectors*. To investigate one of the factors that might determine the GHG emissions levels, this thesis will analyse the impact of profitability on GHG emissions starting in the next chapter. This analysis aims to provide valuable insights for companies, managers, policymakers, and other stakeholders interested in understanding how a specific factor, such as profitability, influences GHG emissions.

7.2. Empirical Analysis: Impact of Profitability on GHG Emissions

This chapter presents the core analysis of this study, focusing on the relationship between profitability and *Total*, *Scope 1*, *Scope 2* and *Scope 3* GHG emissions of European companies, thereby addressing the second research question of this thesis. To begin with, the collected data undergoes descriptive analysis, including a summary of statistical measures and an examination of correlations among all variables. The subsequent section details the fixed-effects regression analysis and its results, with individual assessments of the four subhypotheses. The chapter concludes with the robustness tests conducted and the limitations inherent in the regressions, followed by a summary of the findings.

7.2.1. Descriptive Statistics

The section on descriptive statistics is divided into two segments. The first segment, a summary of statistics, involves the analysis of essential statistical characteristics of the data. The second segment addresses the analysis of correlations between the variables, wherein preliminary insights are derived.

Summary of the Statistics

To start the empirical analysis, the summary of statistics of both datasets is provided in Table 3, displaying basic metrics for the variables in the *Low-Emission-Sectors* and *High-Emission-Sectors*.

As stated earlier, the data originates almost entirely from the LSEG Eikon platform, and missing relevant data was added manually wherever possible. This included, in particular, the emission values for 2023, which were not yet available via Eikon for the most important companies. Although manual additions are prone to error, they enable the subsequent analysis to be conducted using the latest data. Further data transformations were conducted using Python, including removing rows with missing values and preparing variables for logarithmic transformation. As mentioned in the previous chapters, the natural logarithm of all emission variables is used to improve the fit for a regression, similar to many other studies using GHG emissions in regressions (Houqe et al., 2022; Mahapatra et al., 2021; Raval et al., 2021). A small constant with a value of 1e-2 was added to address zero values. This is why the minimum value of certain emission variables is below zero when the natural logarithm of zero plus the small constant is calculated. The complete Python code can be found in Appendix 1. All numbers presented are after the removal of missing values and thus represent the complete dataset used for the subsequent regression analyses. As shown in Table 3, the final dataset has 2,524 observations, representing 538 individual firms, or about 90% of the initial sample. The *Low-Emission-Sectors* dataset has 1,245 observations (N), corresponding to 267 individual firms, whereas the *High-Emission-Sectors* dataset has 1,279 observations, corresponding to 271 individual firms.

The difference between high-emission and low-emission firms also becomes apparent in the data when comparing mean, median, and max values, which are higher for all emission Scopes in *High-Emission-Sectors*. Comparing the *ROA* numbers, the mean and median numbers do not differ significantly, but the SD of 14.6 percentage points in *Low-Emission-Sectors* is much higher than 6.3 percentage points in *High-Emission-Sectors*. This indicates a higher range of values for *Low-Emission-Sectors* and a more constant number for *High-Emission-Sectors*. Variables for GHG emissions will not be analysed in depth again, as the natural logarithm makes the interpretation challenging, and Chapter 7.1.2 already discussed this matter. Along with this summary of statistics, the correlations between all variables help to understand the dataset and will be analysed using a Pearson correlation matrix in the following chapter.

Correlation Matrix

Pearson (1895) introduced the concept of linear correlation between two variables, which allows us to understand the relationship between two variables in both directions. A widely used tool in modern statistics is the Pearson Correlation Matrix, displayed in Table 4. This matrix shows the

Table 3: Summary of Statistics for the Regression Variables

Low-Emission-Sectors	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Total GHG	1,245	12.355	2.834	12.307	4.804	18.808
Scope 1	1,245	9.090	3.090	9.157	-4.605	15.126
Scope 2	1,245	9.890	2.668	10.087	-4.605	15.387
Scope 3	1,245	11.643	3.299	11.710	-4.605	18.807
ROA	1,245	0.066	0.146	0.044	-0.263	2.511
SIZE	1,245	22.171	1.567	22.143	17.809	25.691
BOARDDIV	1,245	0.208	0.149	0.200	0.000	1.000
CAPINT	1,245	0.075	0.138	0.034	-0.059	2.058
GROWTH	1,245	0.068	0.244	0.014	-1.689	5.041
High-Emission-Sectors	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Total GHG	1,279	15.212	2.587	15.279	7.271	21.789
Scope 1	1,279	12.231	2.958	11.865	-4.605	19.007
Scope 2	1,279	11.511	2.454	11.493	-4.605	16.909
Scope 3	1,279	14.618	3.024	14.927	2.231	21.761
ROA	1,279	0.070	0.063	0.060	-0.205	0.585
SIZE	1,279	22.816	1.259	22.709	19.260	26.599
BOARDDIV	1,279	0.173	0.138	0.167	0.000	0.714
CAPINT	1,279	0.075	0.142	0.041	0.000	2.208
GROWTH	1,279	0.115	0.364	0.064	-0.869	7.999

correlation between all variables of the data set and thus provides initial insights into their relationships.

Unsurprisingly, there is a high correlation at the 1% significance level between all four variables for GHG emissions. As discussed in Chapter 7.1.2, *Scope 3* emissions constitute the largest share of *Total GHG* emissions. Therefore, high correlations of 0.96 and 0.95 between *Total GHG* and *Scope 3* in both datasets are expected. Additionally, *ROA* appears to be negatively correlated with GHG variables in both datasets, although there are significant differences between high-emission and low-emission firms. The effect is approximately twice as large for *Scope 1* and *Scope 3* emissions in *High-Emission-Sectors* but does not differ significantly for *Scope 2* emissions. This correlation between financial performance, measured by *ROA*, and the Scopes of GHG emissions are consistently negative at the 1% significance level, indicating a strong relationship. Similarly, the *SIZE* of a company, measured by the natural logarithm of annual revenues, is strongly positively correlated with the amount of GHG emitted across all Scopes, consistent with prior research (Hassan & Romilly, 2018; Wells & Nieuwenhuis, 2012). Moreover, the high significance of the correlation between all variables for *Scope 1* and *Scope 2* in *Low-Emissions-Sectors*, is in contrast with the more nuanced picture for *Total GHG* and *Scope 3*.

In summary, no multicollinearity is detectable between the independent variables. Firm *SIZE* seems to be highly correlated with GHG emission levels across all Scopes, and *ROA* shows a moderate negative correlation. Most control variables have a significant correlation with *Scope 1* and *Scope 2* emissions. However, apart from firm *SIZE*, they do not significantly correlate with *Total GHG* and *Scope 3* emissions. The correlation matrix indicates differences between the *Low-* and *High-Emission-Sectors*, which supports the distinction between these two groups. Subsequently, to unilaterally examine the impact of profitability on Scope 1, 2 and 3 GHG emissions, the results of the linear regressions will be analysed and evaluated in the following chapters.

7.2.2. Multiple Linear Regressions Analysis

This chapter addresses the four sub-hypotheses of the second research question: *How does firm profitability impact total and individual Scope 1, 2 and 3 GHG emissions performance?* An overview of the regression results from both *Low-* and *High-Emission-Sectors* is provided, and the implications of the results will shortly be discussed. For each hypothesis, fixed-effect panel OLS regressions are performed, and the null hypothesis (*H0*) is rejected for p-values < 0.05, indicating a significance at the 5% level. Before the individual regression results, the basic assumptions for a fixed-effect Panel OLS regression should be met and tested. The following assumptions, as outlined by Wooldridge (2012) in his book Introductory Econometrics: A Modern Approach (5th Edition), are tested with specified tests or graphical analysis on all regression models performed in the subsequent chapters. Linear relationship between the dependent and independent variables, normality of residuals, no multicollinearity, exogeneity of independent variables, homoscedasticity of residuals, no autocorrelation, and specifically for fixed-effect models, the assumption that entity-fixed effects are constant over time.

Table 4: Pearson Correlation Matrix

Low-Emission-Sectors	Total GHG	Scope 1	Scope 2	Scope 3	ROA	SIZE	BOARDDIV	CAPINT
Total GHG								
Scope 1	0.749 ***							
Scope 2	0.742 ***	0.810 ***						
Scope 3	0.956 ***	0.637 ***	0.615 ***					
ROA	-0.103 ***	-0.080 ***	-0.135 ***	-0.074 **				
SIZE	0.640 ***	0.632 ***	0.612 ***	0.588 ***	-0.162 ***			
BOARDDIV	-0.008	-0.099 ***	-0.067 ***	0.009	0.182 ***	0.023		
CAPINT	0.036	0.076 **	0.129 **	-0.019	-0.077 ***	-0.126 ***	-0.033	
GROWTH	-0.044	-0.117 ***	-0.096 ***	-0.016	0.113 ***	-0.080 ***	0.019	-0.029
High-Emission-Sectors	Total GHG	Scope 1	Scope 2	Scope 3	ROA	SIZE	BOARDDIV	CAPINT
Total GHG								
Scope 1	0.729 ***							
Scope 2	0.600 ***	0.642 ***						
Scope 3	0.947 ***	0.585 ***	0.508 ***					
ROA	-0.251 ***	-0.298 ***	-0.179 ***	-0.185 ***				
SIZE	0.688 ***	0.623 ***	0.574 ***	0.610 ***	-0.198 ***			
BOARDDIV	0.056 **	-0.019	-0.052 *	0.076 ***	0.080 ***	-0.016		
CAPINT	0.029	0.108 ***	0.086 ***	-0.013	-0.128 ***	-0.158 ***	0.123 ***	
GROWTH	0.042	0.030	-0.061 **	0.050 *	0.082 ***	0.068 **	0.012	-0.022 ***

Linearity is assumed between the dependent and independent variables, and graphical analysis of the scatter plots between these variables tests the assumption. Most similar studies, identified in the previously performed SLR, assume the relationship between financial performance and GHG emissions as linear, with a few exceptions using nonlinear regression models and finding U-shaped and Inverted U-shaped relations. The plots in Appendix 3 indicate a strong linear relationship between GHG emission and firm *SIZE*, a weaker linear relation with *ROA,* and a more nuanced picture of the other variables.

High or perfect multicollinearity between independent variables can cause several problems in regression models (Wooldridge, 2012, pp.94–99). Multicollinearity can highly influence the estimation of the regression coefficient and is to be avoided (Wooldridge, 2012, pp. 94–99). No or low multicollinearity is a basic assumption of the fixed-effect regression model and can easily be tested with the Variance Inflation Factor (VIF) (Wooldridge, 2012, p. 98). The VIF tests the variance increase when variables are correlated and is usually interpreted in the following increments (Wooldridge, 2012, p. 98). VIF = 1, indicating no multicollinearity; 1 < VIF < 5, indicating moderate correlation not requiring specific measures; VIF> 5, indicating significant and potentially problematic correlation; and VIF >10, considered as a threshold for serious multicollinearity (Wooldridge, 2012, p. 98). Since all VIF values for the independent variables of the performed regressions are between 1.018 and 1.085, multicollinearity is not seen as a problem and is neglected in further analysis. The VIF values for each variable can be found in Appendix 3.

The following basic assumptions of linear regression models is the homoscedasticity of residuals, which refers to the constant variation of error terms across the whole range of the independent variables (Wooldridge, 2012, pp. 93–94) and non-autocorrelation of residuals, which means no correlation of residuals across time (Wooldridge, 2012, p. 353). To test for homoscedasticity, or the presence of the opposite, namely heteroscedasticity, the Breusch-pagan test by Breusch and Pagan (1980) can be used. Autocorrelation or serial correlation in panel data is tested with the commonly referred Wooldridge test (Wooldridge, 2010, pp. 176–178). After performing both tests, shown in Appendix 3, the results indicate that heteroscedasticity is present in most regressions, and there is evidence of serial correlation between the residuals. Violating the homoscedasticity assumption still allows for valid results but requires further adjustments in the model (Wooldridge, 2012, pp. 268–296). A commonly used way to account for heteroscedasticity and autocorrelation is using robust standard errors, which have no influence on the coefficients but the respective f-statistic and p-values (Wooldridge, 2012, pp. 268–296). Clustered standard errors are explicitly suited for panel data and will be used on all regressions to test the hypotheses (Petersen, 2009).

Another assumption is that the normality of residuals is expected when performing an OLS regression (Wooldridge, 2012, pp. 118–121). This assumption can be tested with the Shapiro-Wilk test (Shapiro & Wilk, 1965) and by a graphical analysis of the regression models' histograms and Q-Q plots of residuals. The Shapiro-Wilk test tests the hypothesis that a sample comes from a normally distributed population and was developed by Shapiro and Wilk (1965). The results of this test indicate a non-normal distribution of residuals for the regressions performed, which motivates the further graphical analysis of residuals. The plots can be found in Appendix 3 and show a normal distribution for the major parts of the data, but outliers seem to influence the regression in the tails. These outliers are to be expected with GHG emissions data (Griffin, 2017), and are consistent with the findings from the previous analysis of company Scope 1, 2 and 3 GHG emissions in Chapter 7.1.2, which indicates that the majority of GHG emissions come from a small number of companies, namely outliers, making these outliers an essential part of these regressions and hypotheses. Consequently, the assumption for normal distribution of residuals is not entirely met, but scholars indicate that it is not strictly needed for consistent results in fixed-effect regressions and emphasise the use of robust standard errors to address non-normality issues (Greene, 2019, pp. 987–988; Wooldridge, 2012, p. 490).

The two last assumptions discussed are especially relevant for the work with GHG emissions data and the specific fixed-effect regression models. Many studies mention endogeneity, referring to the missing significant influencing variables in the regression model, as a potential issue for the relationship between financial and sustainability performance (Busch et al., 2022; Delmas et al., 2015; Gallego-Alvarez et al., 2015; Shahgholian, 2019). The opposite, exogeneity is a major assumption of every regression, but can rarely be guaranteed in real-life situations, particularly for time series data (Wooldridge, 2012, p. 355). As mentioned earlier, endogeneity is tested with the Hausman test, which compares fixed-effect regression to random-effect regression (Hausman, 1978). The presence of endogeneity would make the fixed-effect regression model more significant than the random-effect, as all time-invariant fixed entity effects are accounted for by definition (Wooldridge, 2012, p. 496). The Hausman tests, shown in Appendix 3, indicate that there might be endogeneity issues in the regression models, and the use of a fixed-effect regression should help to diminish these issues. The last assumption is inherent to fixed-effect regressions, which assumes that fixed-entity effects are time-invariant (Wooldridge, 2012, pp. 484–492). This can hardly be tested but is assumed, since in the context of this work, company sector and business model seem to be significant factors influencing GHG emission levels and are both time-invariant in most cases (IPCC, 2023; WRI & WBCSD, 2004).

The following chapters show, analyse, and discuss the results of the fixed-effect regression models used to answer the four sub-hypotheses of the second research question.

Results for Hypothesis H1

The regression results for *Hypothesis H1* are displayed in Table 5. The regression analysed the link of Profitability, measured by *ROA*, with *Total GHG*, measured by the sum of Scope 1, 2 and 3 GHG emissions. To analyse the results correctly, it is essential to highlight that all dependent variables and *SIZE* were transformed as natural logarithms, and the other variables are in percentages, which requires caution for the interpretation of the coefficients. Furthermore, some studies calculate *Total GHG* emissions as the sum of *Scope 1* and *Scope 2* (Czerny & Letmathe, 2024; Ghose et al., 2023). Whereas this study sees *Scope 3* emissions as necessary for assessing a company's most realistic carbon footprint. An approach mentioned in the literature, due to the risk of potential carbon leakage, the shift of *Scope 1* or *Scope 2* emissions to *Scope 3* by business practices (Wei et al., 2020), or the misleading or distorting picture from only including *Scope 1* and *Scope 2* emissions (Radonjič & Tompa, 2018). However, this also means that the *Total GHG* emissions are significantly influenced by *Scope 3* emissions, as these account for the largest share, as previously analysed in Chapter 7.1.2. Next to the p-value, the significance of a variable is indicated by *, ** and ***, respectively standing for a 10%, 5% and 1% significance level.

The model demonstrates statistical significance in the *Low-Emission-Sectors*, as indicated by an F-statistic p-value of 0.000. However, only a few variables exhibit significant impacts on GHG emissions. Specifically, firm *SIZE* and *BOARDDIV* are statistically significant predictors, with p-values of 0.000 and 0.001, respectively. Firm *SIZE*, with a coefficient of 0.744, suggests that larger firms tend to have higher *Total GHG* emissions, likely due to a larger scope of operational activities and energy use, as suggested by J. Lee and Yu (2019). *BOARDDIV* shows the most substantial positive impact on emissions, with a coefficient of 1.890, indicating that increased *BOARDDIV* correlates with higher emissions. This suggests that diverse boards may face challenges in aligning sustainability goals with business objectives or are in companies where emissions are more complicated to manage. Although this relation seems unintuitive because board diversity is often associated with better CSR performance (Hossain et al., 2023), other studies mention challenges of more diverse boards, which could impact this relationship (R. B. Adams et al., 2015). *ROA* exhibits a negative coefficient of -0.528, suggesting that profitability potentially reduces emissions; however, this relationship is not statistically significant, with a p-value of 0.339, indicating that profitability does not substantially influence *Total GHG* emissions in *Low-Emission-Sectors*. Although this study did not explicitly analyse *Low-Emission-Sectors*, these findings corresponded to the research of Hassan and Romilly (2018), finding no significant impact of economic performance on GHG emissions. Other factors, such as capital intensity and growth, do not significantly affect emissions within these sectors.

In contrast, the regression model in the *High-Emission-Sectors* is robust, as indicated by a higher overall R^2 value of 0.562 compared to 0.339 in *Low-Emission-Sectors*. This suggests that the model explains a more significant portion of the variability in emissions in *High-Emission-Sectors*. Notably, four out of the five independent variables significantly affect GHG emissions at the 5% level. *ROA* has the most significant negative impact on emissions, with a coefficient of -4.046

Table 5: Regression Results for Hypothesis H1

Dependent variable				
Total GHG	**Low-Emission-Sectors**		**High-Emission-Sectors**	
Independent Variables	Coefficient	P-Value	Coefficient	P-Value
Intercept	-4.515	0.317	-19.203 ***	0.001
ROA	-0.528	0.339	-4.046 **	0.019
SIZE	0.744 ***	0.000	1.516 ***	0.000
BOARDDIV	1.890 ***	0.001	0.046	0.943
CAPINT	0.295	0.577	1.751 ***	0.008
GROWTH	-0.097	0.434	-0.200 **	0.028
No. Observations:	1,245		1,279	
No. Entities:	267		271	
F-statistic (robust):	5.916		9.917	
P-Value:	0.000		0.000	
R^2 (Between):	0.339		0.562	

and a p-value of 0.019. This result indicates that more profitable companies tend to have lower *Total GHG* emissions, possibly due to investments in cleaner technologies and more efficient processes or stakeholder pressure in *High-Emission-Sectors* to reduce emissions, supporting the theoretical framework outlined in this thesis. This finding also support the results of Meng et al. (2023), which found that financial performance enhances carbon performance, and Oestreich and Tsiakas (2023). Firm *SIZE* and *CAPINT* have significant positive impacts on emissions, with coefficients of 1.516 and 1.751 and p-values of 0.000 and 0.008, respectively. This suggests that larger and more capital-intensive companies in *High-Emission-Sectors* have higher emissions, likely due to the nature of their operations, which often involve energy-intensive processes (Ghasemi et al., 2023; Nishitani & Kokubu, 2012). The *GROWTH* variable also shows a small but significant negative effect on *Total GHG* emissions, with a coefficient of -0.200 and a p-value of 0.028, indicating that higher revenue *GROWTH* is negatively linked to GHG emissions, supporting the argument that growing business models are less based on GHG emissions. Environmental and technological innovation of companies have been linked to increasing environmental performance in several studies (Mo, 2022; Muthuswamy & Sharma, 2023; Wedari et al., 2023) and innovation is linked to revenue growth (Angus et al., 1996). This supports the argument that growing companies in *High-Emission-Sectors* might be more innovative or efficient and consequently have lower GHG emissions.

Based on these findings, we fail to reject *H0* for *Hypothesis H1* in *Low-Emission-Sectors*, as profitability does not significantly impact *Total GHG* emissions. However, for firms in *High-Emission-Sectors*, *H0* is rejected for *Hypothesis H1*, as increased profitability correlates with lower *Total GHG* emissions. This divergence underscores the need for sector-specific strategies to manage emissions, recognising that financial performance and its influence on sustainability initiatives differ markedly between low- and high-emission industries. Combining that argument with the findings in Chapter 7.1.2, that most emissions come from a few companies in *High-Emission-Sectors* and few from *Low-Emission-Sectors*. It could be that the GHG emissions from companies in *Low-Emission-Sectors* are not significant enough to establish a noteworthy relationship between some variables in the regression analysis. Furthermore, the results align with findings of a negative relationship between financial performance and GHG emissions from Meng et al. (2023) and support the Slack Resources Theory. Indicating that companies in *High-Emission-Sectors* with higher profitability might invest more money in sustainable business practices, leading to lower emission levels. This regression does not analyse the potential bidirectional relation of both variables mentioned by Busch and Hoffmann (2011) and Testa and D'Amato (2017). Still, the results of the SLR and this regression support the theory that the relationship between environmental performance and financial performance might go both ways. This demands caution when analysing only one direction of this relation in the future and is also a problem in this research.

Results for Hypothesis H2

Hypothesis H2 focuses on the impact of profitability on direct GHG emissions from owned assets, namely *Scope 1* GHG emissions. The regression results are shown in Table 6.

The regression model is statistically significant for *Low-Emission-Sectors*, with an F-statistic p-value of 0.029, but individual variables exhibit limited impact on *Scope 1* emissions. The firm *SIZE* is the only variable with a statistically significant (at 5% level) positive coefficient of 0.254 and a p-value of 0.038. This indicates that larger companies tend to have slightly higher *Scope 1* emissions, potentially due to increased operational activities directly emitting GHGs. *BOARDDIV* also approaches significance with a coefficient of -0.572 and a p-value of 0.054, suggesting a potential negative relationship where more diverse boards may help mitigate direct emissions through improved governance and strategic

Table 6: Regression Results for Hypothesis H2

Dependent variable

Scope 1	**Low-Emission-Sectors**		**High-Emission-Sectors**	
Independent Variables	Coefficient	P-Value	Coefficient	P-Value
Intercept	3.547	0.185	5.823 **	0.017
ROA	0.358	0.399	0.627 *	0.084
SIZE	0.254 **	0.038	0.281 ***	0.009
BOARDDIV	-0.572*	0.054	-0.435	0.177
CAPINT	0.186	0.537	0.460	0.128
GROWTH	-0.021	0.805	-0.046	0.574
No. Observations:	1,245		1,279	
No. Entities:	267		271	
F-statistic (robust):	2.502		3.830	
P-Value:	0.029		0.002	
R^2 (Between):	0.143		0.122	

decision-making. This adds to similar findings of Hossain et al. (2023) and Muktadir-Al-Mukit and Bhaiyat (2024), which found a negative relation between board diversity and GHG emission levels. The coefficient for *ROA* is positive but not significant, with a p-value of 0.399, indicating that profitability might not have a meaningful effect on *Scope 1* emissions in *Low-Emission-Sectors*, which could be due to the same reason mentioned in *Hypothesis H1*. Namely, the selected variables can only minimally explain the level of *Scope 1* GHG emissions, or the absolute level of these emissions needs to be larger to find further significant relation. These findings suggest that companies' profits in *Low-Emission-Sectors* are not or less linked to business practices causing GHG emissions.

The model is statistically significant in the *High-Emission-Sectors* with an F-statistic p-value of 0.002. However, similarly to the *Low-Emission-Sectors*, the model's explanatory power is limited, as indicated by a lower R^2 value of 0.122. Here, *ROA* shows a positive coefficient of 0.627, which is significant at the 10% level (p-value of 0.084). It is not significant at the targeted 5% significance but still contrasts with its negative impact on *Total GHG* emissions seen in *Hypothesis H1*. This suggests that high profitability may not translate into reduced direct emissions in *High-Emission-Sectors*. Possibly due to the underlying business models, with a continued reliance on carbon-intensive operations, challenging to decarbonise (Cavaliere, 2019). Firm *SIZE* remains significant, with a coefficient of 0.281 and a p-value of 0.009, indicating that larger firms have higher direct emissions. However, the effect size is smaller than its impact on *Total GHG* emissions. This reflects larger companies' inherent challenges in curbing emissions directly tied to their core operational activities. Other variables, such as *BOARDIV*, *CAPINT*, and *GROWTH*, do not significantly impact *Scope 1* emissions in *High-Emission-Sectors*, indicating that these factors might not directly influence operational-level emissions.

Overall, the regression results suggest that *Scope 1* emissions are less sensitive to the independent variables than *Total GHG* emissions in H1. The lack of a significant negative relationship between *ROA* and *Scope 1* emissions in *High-Emission-Sectors* indicates that profitability may not be linked to practices reducing direct emissions. Instead, the results may suggest that increased profit is not driving more sustainable activities but rather associated with more business activities that emit more GHG, in line with the research of L. Wang et al. (2014), which mentioned the strong mining industry in their sample as a potential reason for this relationship. These findings suggest that we cannot reject *H0* for *Hypothesis H2* in both *Low-* and *High-Emission-Sectors*, as *ROA* is not significantly negatively associated with *Scope 1* GHG emissions. These results highlight the mixed findings identified in the SLR and the importance of analysing the three Scopes individually, as differences in their relationships with financial performance were expected.

Results for Hypothesis H3

Scope 2 emissions are indirect emissions from purchased energy, and Table 7 presents the results for *Hypothesis H3*, which examines the link between profitability and these indirect emissions. If available, the LSEG Eikon Database uses location-based *Scope 2* emissions, allowing us to focus on their implications.

The regression analysis for *Low-Emission-Sectors* reveals that the model has limited explanatory power, as indicated by an overall R^2 of 0.121. Despite this, *ROA* is the only statistically significant variable at the 5% level, with a coefficient of 0.518 and a p-value of 0.016. This positive relationship suggests that more profitable firms in *Low-Emission-Sectors* have higher *Scope 2* emissions. This finding contradicts the Slack Resource Theory, which posits that more profitable firms have additional resources to invest in energy efficiency and emission reductions, but could also indicate that location-based *Scope 2* emissions are not easily reduced with higher financial resources due to the challenges of influencing the local grid energy-mix (Karlsson et al., 2009). Other variables such as firm *SIZE, BOARDDIV, CAPINT*, and *GROWTH* do not signifi-

Table 7: Regression Results for Hypothesis H3

Dependent Variable

Scope 2	**Low-Emission-Sectors**		**High-Emission-Sectors**	
Independent Variables	Coefficient	P-Value	Coefficient	P-Value
Intercept	5.895*	0.056	24.135**	0.027
ROA	0.518**	0.016	2.173	0.114
SIZE	0.186	0.172	-0.560	0.246
BOARDDIV	-0.793	0.123	-0.068	0.844
CAPINT	0.219	0.432	0.242	0.498
GROWTH	-0.129	0.143	-0.040	0.903
No. Observations:	1,245		1,279	
No. Entities:	267		271	
F-statistic (robust):	4.185		1.093	
P-Value:	0.001		0.362	
R^2 (Between):	0.121		-0.530	

cantly impact *Scope 2* emissions, suggesting that these factors may not directly influence energy consumption and associated emissions in *Low-Emission-Sectors*.

The regression model's explanatory power for *High-Emission-Sectors* is notably poor, with a p-value of 0.362, indicating a weak link between the dependent and the independent variables. None of the independent variables are statistically significant at the 5% level, and the model fails to effectively explain the variability in *Scope 2* emissions. However, *ROA* displays a positive coefficient of 2.173, which is marginally non-significant at the 10% level (p-value of 0.114). This suggests a potential trend where increased profitability is associated with higher *Scope 2* emissions, although the relationship lacks statistical significance. The absence of significant explanatory variables may imply that factors beyond the scope of the current model, such as energy-sourcing strategies or the local energy mix (Chuang et al., 2018; WRI & WBCSD, 2004), could play a more substantial role in influencing *Scope 2* emissions in *High-Emission-Sectors*. Interestingly, firm *SIZE* does not significantly impact *Scope 2* emissions in either sector, indicating that company *SIZE* alone may not determine energy consumption or indirect emission levels.

Overall, the results indicate that the examined variables less influence *Scope 2* emissions compared to *Total GHG* or *Scope 1* emissions, which was expected from the elaboration of *Hypothesis H3*. The positive relationship between profitability and *Scope 2* emissions in *Low-Emission-Sectors* suggests that more profitable companies rely more on purchased energy creating GHG emissions. Meanwhile, the lack of significant predictors in *High-Emission-Sectors* highlights the complexity of managing energy-related emissions and the differences between the Scopes and industries. Consequently, we fail to reject *H0* for *Hypothesis H3* in *Low-* and *High-Emission-Sectors*. In *Low-Emission-Sectors*, the positive impact of *ROA* contradicts the expectation. In *High-Emission-Sectors*, the model lacks significant explanatory variables, indicating a need for further research to uncover additional factors influencing *Scope 2* emissions.

Results for Hypothesis H4

The last hypothesis of this thesis evaluates the impact of profitability on *Scope 3* GHG emissions, and the regression results are shown in Table 8.

In *Low-Emission-Sectors*, the regression model is statistically significant overall, as indicated by an F-statistic p-value of 0.000, with an R^2 of 0.298. The *ROA* has a negative coefficient of -1.279 with a p-value of 0.034, the largest coefficient for *ROA* in the *Low-Emission-Sectors*, signifying that higher profitability is associated with lower *Scope 3* emissions. This negative relationship suggests that profitable firms might be investing in more sustainable supply chain practices, such as selecting environmentally conscious suppliers (Fagundes Alves et al., 2024), eco-friendly and durable product design (Asif et al., 2022; Booth et al., 2023) or optimising logistics and sourcing (Hertwich & Wood, 2018), reducing Value Chain emissions. Firm *SIZE* also significantly impacts *Scope 3* emissions, with a coefficient of 0.962 and a p-value of 0.002, indicating that larger firms tend to have higher *Scope 3* emissions. This could be due to larger firms having more extensive supply chains and greater product distribution requirements (Bode & Wagner, 2015). Interestingly, *BOARDDIV* has a significant positive impact on emissions, with a coefficient of 2.948 and a p-value of 0.000, suggesting that more diverse boards might face challenges in aligning sustainability objectives across complex value chains (R. B. Adams et al., 2015) or it might reflect diverse perspectives that increase reporting transparency without immediate reduction efforts (Liao et al., 2015; Tingbani et al., 2020).

The model also achieves statistical significance for *High-Emission-Sectors*, with an R^2 of 0.370. The effect of profitability on *Scope 3* emissions is more pronounced here, with an *ROA* coefficient of −6.234 and a p-value of 0.002. This stronger negative impact indicates that firms in *High-Emission-Sectors* might leverage profitability more effectively

Table 8: Regression Results for Hypothesis H4

Dependent variable				
Scope 3	Low-Emission-Sectors		High-Emission-Sectors	
Independent Variables	Coefficient	P-Value	Coefficient	P-Value
Intercept	-10.141	0.137	-33.194***	0.000
ROA	-1.279 **	0.034	-6.234 ***	0.002
SIZE	0.962 ***	0.002	2.106 ***	0.000
BOARDDIV	2.948 ***	0.000	0.463	0.592
CAPINT	-0.974	0.634	1.962 **	0.030
GROWTH	-0.095	0.568	-0.207	0.272
No. Observations:	1,245		1,279	
No. Entities:	267		271	
F-statistic (robust):	6.232		9.339	
P-Value:	0.000		0.000	
R^2 (Between):	0.298		0.370	

to engage in emissions-reduction activities across their value chains, as mentioned above. Firm *SIZE*, with a coefficient of 2.106 and a p-value of 0.000, is positively correlated with *Scope 3* emissions, further highlighting the complexity of the value chain and emissions in *High-Emission-Sectors*. *CAPINT* also exhibits a significant positive relationship with *Scope 3* emissions, with a coefficient of 1.962 and a p-value of 0.030, implying that capital-intensive firms are more likely to have higher other indirect emissions, possibly due to greater consumption of resources and energy throughout their supply chains, a finding supported by (Hertwich & Wood, 2018).

The findings support the rejection of *H0* for *Hypothesis H4*, as profitability is negatively associated with *Scope 3* GHG emissions for firms in both *Low-* and *High-Emission-Sectors*. The results suggest that financially successful companies could be potentially better positioned to implement sustainability initiatives that reduce emissions across their entire value chain. The large negative correlation between *ROA* and *Scope 3* emissions underscores the importance of integrating environmental sustainability into the broader strategic objectives of profitable firms. Overall, the regression analysis underscores the importance of addressing *Scope 3* emissions as part of a comprehensive climate strategy, given their significant contribution to a firm's overall carbon footprint (Matthews et al., 2008). The results also highlight the different relationships between each Scope by showing a significant negative relation with *ROA* for both *Low-* and *High-Emission-Sectors*, contrasting with the findings for *Scope 1* and *Scope 2*. By focusing on value chain emissions, companies can achieve meaningful reductions in their environmental impact, align with global sustainability goals, and enhance their reputation and competitive advantage in increasingly environmentally-conscious markets.

7.2.3. Robustness Tests

To ensure the reliability of the results, a series of robustness tests were conducted to examine the impact of profitability on GHG emissions. These checks focused on the direction and significance of the relationship between profitability and GHG emissions across various dimensions. The results of all robustness checks can be found in Appendix 4 and are only briefly discussed.

Different measures of profitability, such as *ROE* and *ROS*, were employed to test the differences and is a common way for robustness checks in this field (Busch et al., 2022; Hassan & Romilly, 2018). The analysis revealed that while the relationship direction remained consistent, the significance was less pronounced for *ROE* and non-existent for *ROS*. Similar directional results were observed using a GHG metric relative to firm revenues. However, the relationship between *ROA* and all GHG Scopes lacked statistical significance. Furthermore, incorporating the natural logarithm of total assets, instead of revenues, as a *SIZE* control variable did not alter the direction of the relationship between *ROA* and GHG Scopes but significantly diminished model performance and the significance of the findings. These mixed results are common in most studies analysing similar relationships (Busch & Lewandowski, 2018; Lewandowski, 2017), and indicate the high dependence of results on specific metrics, making it challenging to draw definitive conclusions.

Following a procedure similar as Hassan and Romilly (2018) to assess the influence of extreme outliers, the data was truncated at the 1st and 99th percentiles and the 5th and 95th percentiles. This results in outcomes comparable to the primary analysis without outlier removal, albeit with subtle differences in significance levels. Analysing the entire sample without distinguishing between *Low-* and *High-Emission-Sectors* produced results that aligned with both dataset's expectations. The relationship with *Scope 3* emissions was negative and significant, while the relationship with *Scope 2* emissions was positive and significant. In contrast, the relationship with *Scope 1* emissions was positive but not significant, and the relationship with *Total GHG* emissions was negative but not significant. Lastly, considering the significant disruption of business activities and GHG emissions during the COVID-19 pandemic in 2020, the exclusion of this year from the analysis did not alter the main findings. However, it did result in minor changes to the significance levels.

Overall, the robustness checks confirmed the general reliability of the analysis but indicated significant differences depending on the choice of variables. This finding is also consistent with current literature and shows the complexity and difficulties of understanding the relationship between financial performance and GHG emissions.

7.2.4. Limitations

In this chapter, we discuss the limitations encountered during this research, which includes model specification constraints, data limitations and broader contextual challenges. Recognising these limitations is essential for interpreting the findings accurately and understanding the scope of the study.

A significant theoretical limitation of this study is the scarcity of comparable research on the directional relationship between profitability and GHG emissions, as identified in the SLR, which served as the motivation for this thesis. This scarcity makes it difficult to benchmark the findings and highlights the need for further empirical research to validate the results of this thesis. Another challenge is the potential bidirectional relationship between profitability and GHG emissions, mentioned by several scholars (Busch & Hoffmann, 2011; Testa & D'Amato, 2017; Waddock & Graves, 1997). While this study focuses on the impact of profitability on GHG emissions, emissions may also affect profitability, as shown in the literature identified in the SLR. This introduces endogeneity concerns that could bias the results.

The fixed-effects model used in this study assumes that individual-specific effects are time-invariant and uncorrelated with explanatory variables, which may not always be accurate, potentially leading to biased estimates (Wooldridge, 2012, pp. 484–496). Furthermore, the model does not allow to estimate specific time-invariant variables like the company sectors, which are supposedly major determinants of GHG emissions (Ghasemi et al., 2023). The limitations of fixed-effect models or other OLS-based regressions are mentioned in several studies, which support the use of other models like quantile regressions or the Gaussian Mixture Model (Meng et al., 2023; Rodríguez-García et al., 2022). Furthermore, the model assumes linearity, which may not capture the complex, non-linear relationships or tipping points between profitability and GHG emissions identified in the literature (Misani & Pogutz, 2015; Ogunrinde et al., 2022). Since the regression models identify correlations rather than causations, we cannot make definitive causal claims without experimental or quasi-experimental designs. Exploring alternative methods, such as dynamic models or machine learning techniques, could better capture complex interactions and non-linearities, offering richer insights into these dynamics.

The study presents mixed results across different Scopes of GHG emissions and varying explanatory variables. These inconsistencies underscore the complexity of assessing the impact of profitability on GHG emissions and highlight the need for further investigation into potential moderating variables. A fundamental limitation of this study is the risk of omitted variables, particularly those influencing Scope 1, 2 and 3 emissions. Each Scope appears to have distinct determinants, as suggested by the significant variance in model performance across these Scopes. Additionally, classifying companies into low- and high-emissions sectors may be overly simplistic and fail to capture sectoral complexity, potentially leading to misinterpretations. Differentiating effects caused by specific business model characteristics is challenging without extensive detail. Future research should use more granular classifications or focus on individual high-emitting sectors. As discussed in the previous chapter, robustness tests show significant differences when using various measures and profitability metrics, like ROE and ROS, indicating that measurement choice can substantially influence findings. This requires cautious interpretation and more comprehensive robustness checks in future studies. Measurement errors can also result from inconsistencies in reporting standards and estimation methods of companies, which is a highlighted problem for *Scope 3* emissions (Fouret et al., 2024; Patchell, 2018), potentially affecting the study's findings.

The dataset used in this study is based on the STOXX Europe 600 index, which includes the largest 600 European companies. This focus on European companies limits the findings' generalisability to other regions with different regulatory environments, market dynamics, and environmental practices. Furthermore, small and medium-sized enterprises (SMEs) are not included in the sample, limiting the results' applicability to large corporations. SMEs may exhibit different dynamics in profitability and GHG emissions, so future research should include a broader range of companies. Although data availability has improved, the dataset is still unbalanced, with missing information that could introduce bias and affect reliability. As discussed in Chapter 2.2.3, the upcoming CSRD conforming reports are expected to improve data transparency, quality, and availability for both large and smaller firms, especially regarding GHG emissions disclosures across all three Scopes.

In summary, this study's limitations provide critical insights into the constraints and challenges faced during the research process. Acknowledging these limitations helps contextualise the findings and underscores the need for continued research. Future studies should address these limitations by incorporating more comprehensive datasets, exploring alternative model specifications, and cautiously examining the complex interactions between profitability and GHG emissions.

7.2.5. Key Findings and Summary of Results

This chapter aims to synthesise the results of the four previous hypotheses into key findings and a summary. The empirical analysis distinguishes between *Low-* and *High-Emission-Sectors*, uncovering a significant variance in the impact of profitability on GHG emissions depending on the scope and sector type. Table 9 shows the relationship di-

Table 9: Key Findings of Regressions H1-H4

	Correlation with ROA	
Emission Types	**Low-Emission-Sectors**	**High-Emission-Sectors**
H1: Total GHG	Negative, not significant	Negative, significant
H2: Scope 1	Positive, not significant	Positive, not significant
H3: Scope 2	Positive, significant	Positive, not significant
H4: Scope 3	Negative, significant	Negative, significant

rection from each regression and the significance level at 5%.

Profitability shows a significant negative correlation between *Total GHG* and *Scope 3* emissions in *High-Emission-Sectors* and *Scope 3* emissions in *Low-Emission-Sectors*. This aligns with the theoretical framework, indicating that more profitable companies may invest more in reducing GHG emissions to enhance legitimacy or stakeholder satisfaction. However, this could also indicate that more profitable companies inherently have more sustainable business models, which highlights the limitations of these regressions. In order to isolate the effect of profitability more from other factors, it would therefore be advisable to compare the performance of companies that differ as little as possible apart from profitability. This means preferably from the same industry, with the same business model, and the same geographical focus. Conversely, profitability is positively linked to *Scope 1* and *Scope 2* emissions, but only significant for *Scope 2* in *Low-Emission-Sectors*, suggesting that these models do not fully capture the determinants of emissions for these Scopes. Additionally, the models for *Scope 2* emissions are the least significant, implying that factors not included in the regressions, like local grid energy mix, may play a crucial role. The low explanatory power of the models for *Scope 2* emissions underscores the importance of external factors like local energy grids, suggesting that future research should incorporate explanatory variables specific to each Scope to achieve more conclusive results. Control variables present nuanced results: firm *SIZE* positively correlates with *Total GHG*, *Scope 1*, and *Scope 3* emissions across both *Low-* and *High-Emission-Sectors*, while *BOARDDIV* and *CAPINT* show significance only in specific contexts. Unexpectedly, *BOARDDIV* positively correlates with *Scope 3* emissions in *Low-Emission-Sectors*, which may reflect challenges in aligning diverse perspectives with sustainability goals, or increased transparency. Capital intensity is only significantly related to *Total GHG* and *Scope 3* emissions in *High-Emission-Sectors*, consistent with its association with GHG-intensive activities.

Overall, the findings show mixed results, significantly differing between *Low-* and *High-Emission-Sectors* as well as across the specific Scopes of emissions. The results don't allow a definitive conclusion on the impact of profitability on Total, Scope 1, 2 and 3 GHG emissions and indicate the need for further, sector and Scope specific research. The next and last chapter of this thesis is dedicated to the implications of this work and the final conclusion.

8. Implications and Conclusion

Climate change is increasingly causing severe challenges worldwide. One of the critical objectives in combating climate change, as outlined in the Paris Agreement, is the reduction of GHG emissions. Enhancing companies' sustainability reporting requirements is critical to achieving this goal. As sustainability reporting evolves rapidly, new regulations such as the Corporate Sustainability Reporting Directive are making the disclosure of sustainability-related information mandatory, including Scope 1, 2 and 3 GHG emissions. This aligns with the principle of "what gets measured gets managed" (Drucker, 2007), emphasising the importance of transparency and accountability in driving sustainable practices. One question that scholars have asked themselves frequently is whether "it pays to be green". Although findings indicate that it could pay to be green, studies also find mixed results and often describe the problem of potentially reverse causality and bidirectionality of this relation. In line with the Slack Resource Theory, a more profitable company could be spending more money on CSR and emission reduction initiatives. However, only scarce literature exists on the relation whether "profitability drives sustainability". This thesis aimed to close this research gap identified in the systematic literature review and help scholars, businesses, and politics better understand the effect of profitability on GHG emissions of companies. Two research questions were formulated to do the topic justice.

The first research question, "*What are the Scope 1, 2 and 3 GHG emissions levels for European companies from 2017-2023?",* aimed to provide an overview of GHG emissions in Europe across all Scopes, for the largest 600 companies in Europe, based on the STOXX Europe 600 index. The findings indicate a high level of disclosure for all three Scopes, with steady increases over the years. However, while significantly improved, Scope 3 emissions disclosures have not yet reached the level of Scope 1 and 2 disclosures. There remains considerable variability in Scope 3 emissions levels within and between companies, reflecting ongoing calculation, methodology, and comparability challenges. The data reveals that median growth trends for companies are negative for Scope 1 and 2 emissions, while Scope 3 emissions show a slight growth. This pattern is consistent across sectors, underscoring the importance of Scope 3 emissions in understanding the full picture of GHG emissions. The COVID-19 pandemic's effect is evident, with a notable drop in emissions during the major pandemic year and a subsequent recovery in 2021. Additionally, the sector analysis shows that a small number of companies from high-emitting sectors, such as energy, materials, and industrials, are responsible for most GHG emissions, highlighting the disproportionate impact of high-emitting industries and companies on global GHG emissions.

The second research question, "*How does firm profitability impact total and individual Scope 1, 2 and 3 GHG emissions?"*, builds on the emissions overview insights and analyses the profitability correlation with all GHG Scopes using fixed-effect regressions. The relationship between profitability and GHG emissions was examined by categorising companies into low-emission and high-emission sectors to capture the differences in emissions profiles accurately. The analysis revealed a negative correlation between profitability and Scope 3 emissions, which was the strongest across all regressions, highlighting the significant impact of profitability on this Scope. Due to the large contribution of Scope 3 emissions, the regression of Total GHG emissions yielded similar but less significant results. Interestingly, the direction of the relationship between profitability and Scope 1 and 2 emissions was unexpectedly positive for both high and low-emitting sectors, though these findings lacked statistical significance in most cases. Overall, the impact of profitability on all Scopes of GHG emissions was more pronounced in high-emitting sectors than low-emitting ones, underscoring the stronger connection of profits and GHG emission in these sectors. Conducted robustness tests generally confirm the reliability of the findings, although using relative measures of GHG emissions and alternative profitability metrics resulted in nuanced results. While these alternative approaches largely pointed in the same direction, they often showed less or no statistical significance. In conclusion to research question two, the relationship varies across each Scope, highlighting the need for further research.

The implications of the findings for scholars, businesses, and policymakers are multifaceted. Scholars must consider the potential bidirectional and reverse relationship between financial performance and GHG emissions. Because it may not only "pay to be green" but "profitability may drive sustainability", recognising this is important and should be accounted for in future research. Additionally, the mixed results between the specific Scopes indicate the need to account for Scope-specific determinants and focus on individual relationships rather than Total GHG emissions. Since business practices less influence GHG emissions in low-emission sectors, and a few sectors produce the most emissions, scholars should focus on the sectors where the most GHG reduction can be achieved first. Businesses must prioritise reducing Scope 3 emissions, as they constitute the majority of GHGs, and ensure accurate carbon accounting to manage emissions effectively. To achieve that, policymakers need to ensure that all material emissions are included, and that the comparability of Scope 3 emissions is improved, especially because Scope 1 and Scope 2 emissions can be shifted to Scope 3 by business practices like outsourcing. Similar to the focus of scholars, GHG reduction policies should focus on reducing emissions from high-emitting companies and sectors to achieve the most impact on the fight against climate change. The findings indicate that the emission of GHG is still part of many business models since profitability is positively, but for most, not significantly correlated with Scope 1 and 2 GHG emissions. This suggests that political measures should be reinforced to hold companies accountable for the environmental damages they cause, while also implementing stricter regulations to reduce or eliminate greenhouse gas emissions. Although carbon taxes and emissions trading are a good start, policies must go further to ensure companies fully internalize the environmental costs of their activities, thereby intensifying the urgency to achieve lower emissions.

However, the findings and implications of this study should be interpreted with caution, and the limitations must be acknowledged. A major theoretical constraint is the scarcity of comparable research on this topic, making it challenging to benchmark findings and emphasising the need for further empirical validation. Additionally, the potential bidirectional nature of the relationship introduces endogeneity concerns, as emissions can also affect profitability, complicating interpretation. Furthermore, the fixed effects model used in the analysis assumes the time-invariance of fixed effects and linearity of the relationship, which may not capture the complex, potentially non-linear relationship between profitability and GHG emissions. Besides, data limitations also impact the study's generalisability. The focus on the STOXX Europe 600 index, comprising the largest European companies, excludes small and medium-sized enterprises (SMEs) and limits applicability to other regions with different regulatory environments. Additionally, the dataset's unbalanced nature and missing information pose challenges to the reliability of the findings. Acknowledging these limitations underscores the importance of future research to validate and expand upon these findings, incorporating more comprehensive datasets and exploring alternative models to understand the intricate relationship between profitability and GHG emissions.

Based on the findings and the implications, a suggestion for future research would be to focus on single high-emission sectors to compare how the profitability of different firms in similar contexts influences Scope 1, 2 and 3 GHG emissions. The results have shown substantial differences between each Scope, and a clear distinction of these in future research is advisable. Furthermore, the mandatory disclosure of emissions across all three Scopes for firms falling under the CSRD is a chance to perform similar research with more and better data in the coming years. Hopefully allowing for extensive analyses over time.

Ultimately, this work highlights the complex relationship between profitability and GHG emissions, underscoring the challenge of drawing definitive conclusions while emphasising society's continued reliance on environmentally harmful business practices for economic gain. Only the efforts of businesses, policymakers, and society can mitigate the adverse effects of climate change and ensure a resilient and sustainable world for future generations. Proper carbon accounting and reporting are essential first steps, but are they enough?

References

Ababneh, A. (2019). The impact of carbon accounting on corporate financial performance: Evidence from the energy sector in Jordan. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 1157–1163. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079490808&partnerID=40&md5=b7e9f096bb30931418cd2b4f36da3a29

Adams, C. A., & Frost, G. R. (2008). Integrating sustainability reporting into management practices. *Accounting Forum*, *32*(4), 288–302. https://doi.org/10.1016/j.accfor.2008.05.002

Adams, R. B., de Haan, J., Terjesen, S., & van Ees, H. (2015). Board Diversity: Moving the Field Forward. *Corporate Governance: An International Review*, *23*(2), 77–82. https://doi.org/10.1111/corg.12106

Akhter, F., Hossain, M. R., Elrehail, H., Rehman, S. U., & Almansour, B. (2023). Environmental disclosures and corporate attributes, from the lens of legitimacy theory: a longitudinal analysis on a developing country. *European Journal of Management and Business Economics*, *32*(3), 342–369. https://doi.org/10.1108/EJMBE-01-2021-0008

Aluchna, M., Roszkowska-Menkes, M., & Kamiński, B. (2023). From talk to action: the effects of the non-financial reporting directive on ESG performance. *Meditari Accountancy Research*, *31*(7), 1–25. https://doi.org/10.1108/MEDAR-12-2021-1530

Alvarez, I. G. (2012). Impact of CO 2 Emission Variation on Firm Performance. *Business Strategy and the Environment*, *21*(7), 435–454. https://doi.org/10.1002/bse.1729

Ampofo, A. A., & Sellani, R. J. (2005). Examining the differences between United States Generally Accepted Accounting Principles (U.S. GAAP) and International Accounting Standards (IAS): implications for the harmonization of accounting standards. *Accounting Forum*, *29*(2), 219–231. https://doi.org/10.1016/j.accfor.2004.11.002

Angus, F. R., Goodman, A. L., Pfund, P., Wasson, R., Wyndrum, R., & Zoller, W. M. (1996). Reengineering for Revenue Growth. *Research-Technology Management*, *39*(2), 26–31. https://doi.org/10.1080/08956308.1996.11671047

Asif, M. S., Lau, H., Nakandala, D., Fan, Y., & Hurriyet, H. (2022). Case study research of green life cycle model for the evaluation and reduction of scope 3 emissions in food supply chains. *Corporate Social Responsibility and Environmental Management*, *29*(4), 1050–1066. https://doi.org/10.1002/csr.2253

Bachmann, P., & Ingenhoff, D. (2016). Legitimacy through CSR disclosures? The advantage outweighs the disadvantages. *Public Relations Review*, *42*(3), 386–394. https://doi.org/10.1016/j.pubrev.2016.02.008

Barbu, E. M., Ionescu-Feleagă, L., & Ferrat, Y. (2022). The Evolution of Environmental Reporting in Europe: The Role of Financial and Non-Financial Regulation. *The International Journal of Accounting*, *57*(02), 2250008. https://doi.org/10.1142/S1094406022500081

Barney, J. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99–120. https://doi.org/10.1177/014920639101700108

Baumüller, J., & Grbenic, S. (2021). Moving from non-financial to sustainability reporting: analyzing the EU Commission's proposal for a Corporate Sustainability Reporting Directive (CSRD). *Facta Universitatis Series Economics and Organization*, *18*, 369–381. https://doi.org/10.22190/FUEO210817026B

Baumüller, J., & Sopp, K. (2022). Double materiality and the shift from non-financial to European sustainability reporting: review, outlook and implications. *Journal of Applied Accounting Research*, *23*(1), 8–28. https://doi.org/10.1108/JAAR-04-2021-0114

Bellantuono, N., Pontrandolfo, P., & Scozzi, B. (2016). Capturing the Stakeholders' View in Sustainability Reporting: A Novel Approach. *Sustainability*, *8*(4), 379. https://doi.org/10.3390/su8040379

Benkraiem, R., Dubocage, E., Lelong, Y., & Shuwaikh, F. (2023). The effects of environmental performance and green innovation on corporate venture capital. *Ecological Economics*, *210*, 25, Article 107860. https://doi.org/10.1016/j.ecolecon.2023.107860

Bernerth, J. B., & Aguinis, H. (2016). A Critical Review and Best-Practice Recommendations for Control Variable Usage. *Personnel Psychology*, *69*(1), 229–283. https://doi.org/10.1111/peps.12103

Bhojraj, S., Lee, C. M. C., & Oler, D. K. (2003). What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research. *Journal of Accounting Research*, *41*(5), 745–774. https://doi.org/10.1046/j.1475-679X.2003.00122.x

Biermann, F., Kanie, N., & Kim, R. E. (2017). Global governance by goal-setting: the novel approach of the UN Sustainable Development Goals. *Current Opinion in Environmental Sustainability*, *26-27*, 26–31. https://doi.org/10.1016/j.cosust.2017.01.010

Bode, C., & Wagner, S. M. (2015). Structural drivers of upstream supply chain complexity and the frequency of supply chain disruptions. *Journal of Operations Management*, *36*, 215–228. https://doi.org/10.1016/j.jom.2014.12.004

Boiral, O. (2013). Sustainability reports as simulacra? A counter-account of A and A+ GRI reports. *Accounting, Auditing & Accountability Journal*, *26*(7), 1036–1071. https://doi.org/10.1108/AAAJ-04-2012-00998

Boiral, O., & Heras-Saizarbitoria, I. (2020). Sustainability reporting assurance: Creating stakeholder accountability through hyperreality? *Journal of Cleaner Production*, *243*, 118596. https://doi.org/10.1016/j.jclepro.2019.118596

Boiral, O., Heras-Saizarbitoria, I., & Brotherton, M.-C. (2019). Assessing and Improving the Quality of Sustainability Reports: The Auditors' Perspective. *Journal of Business Ethics*, *155*(3), 703–721. https://doi.org/10.1007/s10551-017-3516-4

Booth, A., Jager, A., Faulkner, S. D., Winchester, C. C., & Shaw, S. E. (2023). Pharmaceutical Company Targets and Strategies to Address Climate Change: Content Analysis of Public Reports from 20 Pharmaceutical Companies. *International Journal of Environmental Research and Public Health*, *20*(4), 3206. https://doi.org/10.3390/ijerph20043206

Bouaddi, M., Basuony, M. A. K., & Noureldin, N. (2023). The Heterogenous Effects of Carbon Emissions and Board Gender Diversity on a Firm's Performance. *Sustainability*, *15*(19), Article 14642. https://doi.org/10.3390/su151914642

Bourgeois, L. J. (1981). On the Measurement of Organizational Slack. *The Academy of Management Review*, *6*(1), 29–39. https://doi.org/10.2307/257138

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *The Review of Economic Studies*, *47*(1), 239–253. https://doi.org/10.2307/2297111

Bricheux, C., Gatzer, S., Lehr, J., & Ponbauer, L. (2024). Reducing the Scope 1 and 2 emissions of consumer goods companies. *McKinsey*. https://www.mckinsey.com/capabilities/sustainability/our-insights/sustainability-blog/reducing-the-scope-1-and-2-emissions-of-consumer-goods-companies

brightest. (n.d.). Location vs. Market Based Scope 2 Emissions - Which Should You Use? Retrieved August 10, 2024, from https://www.brightest.io/location-market-based-emissions-scope-2/#:~:text=Market%2Dbased%20Scope%202%20emissions%20are%20emissions%20calculated%20based%20on,contract%20or%20agreement%20for%20energy

Buallay, A. (2019). Between cost and value. *Journal of Applied Accounting Research*, *20*(4), 481–496. https://doi.org/10.1108/JAAR-12-2017-0137

Busch, T., Bassen, A., Lewandowski, S., & Sump, F. (2022). Corporate Carbon and Financial Performance Revisited. *Organization and Environment*, *35*(1), 154–171. https://doi.org/10.1177/1086026620935638

Busch, T., & Hoffmann, V. H. (2011). How hot is your bottom line? linking carbon and financial performance. *Business and Society*, *50*(2), 233–265. https://doi.org/10.1177/0007650311398780

Busch, T., & Lewandowski, S. (2018). Corporate Carbon and Financial Performance: A Meta-analysis. *Journal of Industrial Ecology*, *22*(4), 745–759. https://doi.org/10.1111/jiec.12591

Cardoni, A., Kiseleva, E., & Terzani, S. (2019). Evaluating the Intra-Industry Comparability of Sustainability Reports: The Case of the Oil and

Gas Industry. *Sustainability*, *11*(4), 1093. https://doi.org/10.3390/su11041093

Cavaliere, P. (2019). Clean Ironmaking and Steelmaking Processes: Efficient Technologies for Greenhouse Emissions Abatement. In P. Cavaliere (Ed.), *Clean Ironmaking and Steelmaking Processes: Efficient Technologies for Greenhouse Emissions Abatement* (pp. 1–37). Springer International Publishing. https://doi.org/10.1007/978-3-030-21209-4_1

CDP. (2023). CDP Climate Change 2023 Reporting Guidance. https://guidance.cdp.net/en/guidance?ctype=ExternalRef&idtype=RecordExternalRef&cid=C6.1&otype=Guidance&incchild=0%C2%B5site=0&gettags=0

Chen, H. B., & Manu, E. K. (2022). The impact of banks' financial performance on environmental performance in Africa. *Environmental Science and Pollution Research*, *29*(32), 49214–49233. https://doi.org/10.1007/s11356-022-19401-w

Chen, J. C., Patten, D. M., & Roberts, R. W. (2008). Corporate Charitable Contributions: A Corporate Social Performance or Legitimacy Strategy? *Journal of Business Ethics*, *82*(1), 131–144. https://doi.org/10.1007/s10551-007-9567-1

Chuang, J., Lien, H.-L., Den, W., Iskandar, L., & Liao, P.-H. (2018). The relationship between electricity emission factor and renewable energy certificate: The free rider and outsider effect. *Sustainable Environment Research*, *28*(6), 422–429. https://doi.org/10.1016/j.serj.2018.05.004

Coelho, R., Jayantilal, S., & Ferreira, J. J. (2023). The impact of social responsibility on corporate financial performance: A systematic literature review. *Corporate Social Responsibility and Environmental Management*, *30*(4), 1535–1560. https://doi.org/10.1002/csr.2446

Cote, C. (2021, April). Making the Business Case for Sustainability. https://online.hbs.edu/blog/post/business-case-for-sustainability

Cuomo, F., Gaia, S., Girardone, C., & Piserà, S. (2022). The effects of the EU non-financial reporting directive on corporate social responsibility. *The European Journal of Finance*, 1–27. https://doi.org/10.1080/1351847X.2022.2113812

Cyert, R. M., & March, J. G. (1963). *A Behavioral Theory of the Firm*. Prentice Hall/Pearson Education.

Czerny, A., & Letmathe, P. (2024). The productivity paradox in carbon-intensive companies: How eco-innovation affects corporate environmental and financial performance. *Business Strategy and the Environment*. https://doi.org/10.1002/bse.3776

Daniel, F., Lohrke, F. T., Fornaciari, C. J., & Turner, R. A. (2004). Slack resources and firm performance: a meta-analysis. *Journal of Business Research*, *57*(6), 565–574. https://doi.org/10.1016/S0148-2963(02)00439-3

de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. d. L. (2020). Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, *32*(1), 19. https://doi.org/10.1186/s12302-020-0300-3

Deegan, C. (2002). The legitimising effect of social and environmental disclosures - a theoretical foundation accounting. *Accounting, Auditing & Accountability Journal*, *15*(3), 282–311. https://doi.org/10.1108/09513570210435852

Delmas, M. A., Nairn-Birch, N., & Lim, J. H. (2015). Dynamics of Environmental and Financial Performance: The Case of Greenhouse Gas Emissions. *Organization & Environment*, *28*(4), 374–393. https://doi.org/10.1177/1086026615620238

Desai, R., Raval, A., Baser, N., & Desai, J. (2021). Impact of carbon emission on financial performance: empirical evidence from India. *South Asian Journal of Business Studies*. https://doi.org/10.1108/SAJBS-10-2020-0384

Di Pillo, F., Gastaldi, M., Levialdi, N., & Miliacca, M. (2017). Environmental performance versus economic-financial performance: Evidence from Italian firms. *International Journal of Energy Economics and Policy*, *7*(2), 98–108. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85017646997&partnerID=40&md5=bd561271b3bd1e5a9d199c832b88552e

Diaz-Sarachaga, J. M. (2021). Shortcomings in reporting contributions towards the sustainable development goals. *Corporate Social Responsibility and Environmental Management*, *28*(4), 1299–1312. https://doi.org/10.1002/csr.2129

Diouf, D., & Boiral, O. (2017). The quality of sustainability reports and impression management. *Accounting, Auditing & Accountability Journal*, *30*(3), 643–667. https://doi.org/10.1108/AAAJ-04-2015-2044

Dowling, J., & Pfeffer, J. (1975). Organizational Legitimacy: Social Values and Organizational Behavior. *The Pacific Sociological Review*, *18*(1), 122–136. https://doi.org/10.2307/1388226

Downie, J., & Stubbs, W. (2013). Evaluation of Australian companies' scope 3 greenhouse gas emissions assessments. *Journal of Cleaner Production*, *56*, 156–163. https://doi.org/10.1016/j.jclepro.2011.09.010

Drucker, P. (2007). *The effective executive* (1st ed.). Routledge. https://doi.org/10.4324/9780080549354

Ducoulombier, F. (2021). Understanding the Importance of Scope 3 Emissions and the Implications of Data Limitations. https://doi.org/10.3905/jesg.2021.1.018

Eccles, R. G., Lee, L.-E., & Stroehle, J. C. (2020). The Social Origins of ESG: An Analysis of Innovest and KLD. *Organization & Environment*, *33*(4), 575–596. https://doi.org/10.1177/1086026619888994

EFRAG. (2023a). EFRAG's Cover Letter on the Cost-benefit analysis. https://www.efrag.org/Assets/Download?assetUrl=%2Fsites%2Fwebpublishing%2FSiteAssets%2F05%2520EFRAGs%2520Cover%2520Letter%2520on%2520the%2520Cost-benefit%2520analysis.pdf&AspxAutoDetectCookieSupport=1

EFRAG. (2023b). ESRS E1 CLIMATE CHANGE. https://www.efrag.org/sites/default/files/sites/webpublishing/SiteAssets/ESRS%20E1%20Delegated-act-2023-5303-annex-1_en.pdf

Elalfy, A., Weber, O., & Geobey, S. (2021). The Sustainable Development Goals (SDGs): a rising tide lifts all boats? Global reporting implications in a post SDGs world. *Journal of Applied Accounting Research*, *22*(3), 557–575. https://doi.org/10.1108/JAAR-06-2020-0116

Endrikat, J., Guenther, E., & Hoppe, H. (2014). Making sense of conflicting empirical findings: A meta-analytic review of the relationship between corporate environmental and financial performance. *European Management Journal*, *32*(5), 735–751. https://doi.org/10.1016/j.emj.2013.12.004

envoria. (2022). What is double materiality in the CSRD? https://envoria.com/insights-news/what-is-double-materiality-in-the-csrd

Erkens, M., Paugam, L., & Stolowy, H. (2015). Non-financial information: State of the art and research perspectives based on a bibliometric study. *Comptabilité Contrôle Audit*, *21*(3), 15–92. https://doi.org/10.3917/cca.213.0015

ESMA. (n.d.). Electronic Reporting. https://www.esma.europa.eu/issuer-disclosure/electronic-reporting

European Broadcasting Union. (2023). Sustainability Rulebook: The Corporate Sustainability Reporting Directive. Retrieved July 1, 2024, from https://www.ebu.ch/case-studies/open/legal-policy/eu-sustainability-rulebook-the-corporate-sustainability-reporting-directive#1

European Commission. (n.d.). Corporate sustainability reporting. https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en

European Commission. (2023). The Commission adopts the European Sustainability Reporting Standards. https://finance.ec.europa.eu/news/commission-adopts-european-sustainability-reporting-standards-2023-07-31_en

European Parliament. (2022). Sustainable economy: Parliament adopts new reporting rules for multinationals. https://www.europarl.europa.eu/news/en/press-room/20221107IPR49611/sustainable-economy-parliament-adopts-new-reporting-rules-for-multinationals

European Union. (2014). Directive 2014/95/EU of the European Parliament and of the Council of 22 October 2014 amending Directive 2013/34/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups Text with EEA relevance. Retrieved July 1, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0095

European Union. (2019). Summary of: Directive 2014/95/EU on disclosure of non-financial and diversity information - Disclosure of non-financial and diversity information by large companies and groups. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=LEGISSUM%3A240601_2

European Union. (2022). Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Regulation (EU) No 537/2014, as regards corporate sustainability reporting. Retrieved June 1, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2464

Fagundes Alves, M. Y., Marques Vieira, L., & Beal Partyka, R. (2024). Suppliers' GHG mitigation strategies (Scope 3): the case of a steelmaking company. *Journal of Manufacturing Technology Management*, *35*(2), 383–402. https://doi.org/10.1108/JMTM-05-2023-0162

Feng, Z. Y., Wang, Y. C., & Wang, W. G. (2024). Corporate carbon reduction and tax avoidance: International evidence. *Journal of Contemporary Accounting & Economics*, *20*(2), 18, Article 100416. https://doi.org/10.1016/j.jcae.2024.100416

Fernandez-Feijoo, B., Romero, S., & Ruiz, S. (2014). Effect of Stakeholders' Pressure on Transparency of Sustainability Reports within the GRI Framework. *Journal of Business Ethics*, *122*(1), 53–63. https://doi.org/10.1007/s10551-013-1748-5

Fouret, F., Haalebos, R., Olesiewicz, M., Simmons, J., Jain, M., & Kooroshy, J. (2024). Scope for improvement - Solving the Scope 3 conundrum. Retrieved July 1, 2024, from https://www.lseg.com/content/dam/ftse-russell/en_us/documents/research/solving-scope-3-conundrum.pdf

França, A., López-Manuel, L., Sartal, A., & Vázquez, X. H. (2023). Adapting corporations to climate change: How decarbonization impacts the business strategy-performance nexus. *Business Strategy and the Environment*, *32*(8), 5615–5632. https://doi.org/10.1002/bse.3439

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Cambridge university press.

Fujii, H., Iwata, K., Kaneko, S., & Managi, S. (2013). Corporate Environmental and Economic Performance of Japanese Manufacturing Firms: Empirical Study for Sustainable Development. *Business Strategy and the Environment*, *22*(3), 187–201. https://doi.org/10.1002/bse.1747

Galama, J. T., & Scholtens, B. (2021). A meta-analysis of the relationship between companies' greenhouse gas emissions and financial performance. *Environmental Research Letters*, *16*(4), 24, Article 043006. https://doi.org/10.1088/1748-9326/abdf08

Gallego-Alvarez, I., Segura, L., & Martínez-Ferrero, J. (2015). Carbon emission reduction: the impact on the financial and operational performance of international companies. *Journal of Cleaner Production*, *103*, 149–159. https://doi.org/10.1016/j.jclepro.2014.08.047

Gallego-Álvarez, I., García-Sánchez, I. M., & da Silva Vieira, C. (2014). Climate Change and Financial Performance in Times of Crisis. *Business Strategy and the Environment*, *23*(6), 361–374. https://doi.org/10.1002/bse.1786

Ganda, F. (2022). Carbon performance, company financial performance, financial value, and transmission channel: an analysis of South African listed companies. *Environmental Science and Pollution Research*, *29*(19), 28166–28179. https://doi.org/10.1007/s11356-021-18467-2

Ganda, F., & Milondzo, K. S. (2018). The Impact of Carbon Emissions on Corporate Financial Performance: Evidence from the South African Firms. *Sustainability*, *10*(7), 22, Article 2398. https://doi.org/10.3390/su10072398

George, G. (2005). Slack resources and the performance of privately held firms. *Academy of Management Journal*, *48*(4), 661–676. https://doi.org/10.5465/amj.2005.17843944

Ghasemi, M., Rajabi, M., & Aghakhani, S. (2023). Towards sustainability: The effect of industries on CO2 emissions. *Journal of Future Sustainability*, *3*(2), 107–118. https://doi.org/10.5267/j.jfs.2022.12.002

Ghose, B., Makan, L. T., & Kabra, K. C. (2023). Impact of carbon productivity on firm performance: moderating role of industry type and firm size. *Managerial Finance*, *49*(5), 866–883. https://doi.org/10.1108/MF-07-2022-0319

Gold, N. O., Taib, F. M., & Ma, Y. (2022). Firm-Level Attributes, Industry-Specific Factors, Stakeholder Pressure, and Country-Level Attributes: Global Evidence of What Inspires Corporate Sustainability Practices and Performance. *Sustainability*, *14*(20), 13222. https://doi.org/10.3390/su142013222

Gordon, R. A. (1968). Issues in Multiple Regression. *American Journal of Sociology*, *73*(5), 592–616. https://doi.org/10.1086/224533

Green, J. F. (2010). Private Standards in the Climate Regime: The Greenhouse Gas Protocol. *Business and Politics*, *12*(3), 1–37. https://doi.org/10.2202/1469-3569.1318

Greene, W. (2019). *Econometric Analysis* (Eighth edition. Global edition ed.). Pearson.

GRI. (2022). Linking the SDGs and the GRI Standards. https://www.globalreporting.org/media/lbvnxb15/mapping-sdgs-gri-update-march.pdf

GRI. (2024). New resource on emissions reporting using GRI and ISSB standards. https://www.globalreporting.org/news/news-center/new-resource-on-emissions-reporting-using-gri-and-issb-standards/

Griffin, P. (2017). The Carbon Majors Database - CDP Carbon Majors Report 2017. https://cdn.cdp.net/cdp-production/cms/reports/documents/000/002/327/original/Carbon-Majors-Report-2017.pdf?1501833772

Günther, H. O., Kannegiesser, M., & Autenrieb, N. (2015). The role of electric vehicles for supply chain sustainability in the automotive industry. *Journal of Cleaner Production*, *90*, 220–233. https://doi.org/10.1016/j.jclepro.2014.11.058

Hahnkamper-Vandenbulcke, N. (2021). Non-financial Reporting Directive - Briefing. Retrieved July 1, 2024, from https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/654213/EPRS_BRI(2021)654213_EN.pdf

Hart, S. L. (1995). A Natural-Resource-Based View of the Firm. *Academy of Management Review*, *20*(4), 986–1014. https://doi.org/10.5465/amr.1995.9512280033

Hassan, O. A. G., & Romilly, P. (2018). Relations between corporate economic performance, environmental disclosure and greenhouse gas emissions: New insights. *Business Strategy and the Environment*, *27*(7), 893–909. https://doi.org/10.1002/bse.2040

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251–1271. https://doi.org/10.2307/1913827

Helbing. (2022). Robust sustainability reporting enables "impact" creation. Retrieved July 15, 2024, from https://helbling.ch/en/insights/robust-sustainability-reporting-enables-impact-creation

Herremans, I. M., Nazari, J. A., & Mahmoudian, F. (2016). Stakeholder Relationships, Engagement, and Sustainability Reporting. *Journal of Business Ethics*, *138*(3), 417–435. https://doi.org/10.1007/s10551-015-2634-0

Hertwich, E. G., & Wood, R. (2018). The growing importance of scope 3 greenhouse gas emissions from industry. *Environmental Research Letters*, *13*(10), 104013. https://doi.org/10.1088/1748-9326/aae19a

Hoang, T. H. V., Przychodzen, W., Przychodzen, J., & Segbotangni, E. A. (2020). Does it pay to be green? A disaggregated analysis of U.S. firms with green patents. *Business Strategy and the Environment*, *29*(3), 1331–1361. https://doi.org/10.1002/bse.2437

Homroy, S. (2023). GHG emissions and firm performance: The role of CEO gender socialization. *Journal of Banking and Finance*, *148*, Article 106721. https://doi.org/10.1016/j.jbankfin.2022.106721

Hossain, A. T., Hossain, A., Cooper, T., & Islam, M. (2023). Corporate sexual orientation equality and carbon emission. *Accounting and Finance*. https://doi.org/10.1111/acfi.13187

Houqe, M. N., Opare, S., Zahir-Ul-hassan, M. K., & Ahmed, K. (2022). The Effects of Carbon Emissions and Agency Costs on Firm Performance. *Journal of Risk and Financial Management*, *15*(4), Article 152. https://doi.org/10.3390/jrfm15040152

IFRS Foundation. (2018). Conceptual Framework for Financial Reporting. https://www.ifrs.org/content/dam/ifrs/publications/pdf-stand ards/english/2021/issued/part-a/conceptual-framework-for-financial-reporting.pdf

IFRS Foundation. (2024). International Sustainability Standards Board. https://www.ifrs.org/groups/international-sustainability-standards-board/

Ioannou, I., & Serafeim, G. (2017). The consequences of mandatory corporate sustainability reporting. https://doi.org/10.1093/oxfordhb/9780198802280.013.20

IPCC. (2023). Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. https://doi.org/10.59327/IPCC/AR6-9789291691647

IRENA. (2022). Renwable Power Generation - Costs in 2021. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2022/Jul/IRENA_Power_Generation_Costs_2021.pdf?rev=34c22a4b244d434da0accde7de7c73d8

Iwata, H., & Okada, K. (2011). How does environmental performance affect financial performance? Evidence from Japanese manufacturing firms. *Ecological Economics*, *70*(9), 1691–1700. https://doi.org/10.1016/j.ecolecon.2011.05.010

Jakhar, S. K., Mangla, S. K., Luthra, S., & Kusi-Sarpong, S. (2019). When stakeholder pressure drives the circular economy. *Management Decision*, *57*(4), 904–920. https://doi.org/10.1108/MD-09-2018-0990

James, M. L. (2015). The benefits of sustainability and integrated reporting: An investigation of accounting majors' perceptions. *J. Legal Ethical & Regul. Isses*, *18*(1), 1. https://www.researchgate.net/publication/282173799_The_benefits_of_sustainability_and_integrated_reporting_An_investigation_of_accounting_majors'_perceptions

Karlsson, M., Gebremedhin, A., Klugman, S., Henning, D., & Moshfegh, B. (2009). Regional energy system optimization – Potential for a regional heat market. *Applied Energy*, *86*(4), 441–451. https://doi.org/10.1016/j.apenergy.2008.09.012

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *J R Soc Med*, *96*(3), 118–121. https://doi.org/10.1177/014107680309600304

Kim, R. E. (2016). The Nexus between International Law and the Sustainable Development Goals. *Review of European, Comparative & International Environmental Law*, *25*(1), 15–26. https://doi.org/10.1111/reel.12148

Kiron, D., & Kruschwitz, N. (2015). Sustainability Reporting As a Tool for Better Risk Management. *MIT Sloan Management Review*, *56*(4). https://sloanreview.mit.edu/article/sustainability-reporting-as-a-tool-for-better-risk-management/

Koh, S. C. L., Jia, F., Gong, Y., Zheng, X., & Dolgui, A. (2023). Achieving carbon neutrality via supply chain management: position paper and editorial for IJPR special issue. *International Journal of Production Research*, *61*(18), 6081–6092. https://doi.org/10.1080/00207543.2023.2232652

KPMG. (2022). Big shifts, small steps - Survey of Sustainability Reporting 2022. https://assets.kpmg.com/content/dam/kpmg/se/pdf/komm/2022/Global-Survey-of-Sustainability-Reporting-2022.pdf

Kumar, A., Singh, P., Raizada, P., & Hussain, C. M. (2022). Impact of COVID-19 on greenhouse gases emissions: A critical review. *Science of the Total Environment*, *806*, 150349. https://doi.org/10.1016/j.scitotenv.2021.150349

Kumar, P., & Firoz, M. (2018). Corporate carbon intensity matter: Predicting firms' financial performance. *SCMS Journal of Indian Management*, *15*(4), 74–84. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067021355&partnerID=40&md5=fff6125e8253190165add54dcbebd8a1

Kuruppu, S. C., Milne, M. J., & Tilt, C. A. (2019). Gaining, maintaining and repairing organisational legitimacy. *Accounting, Auditing & Accountability Journal*, *32*(7), 2062–2087. https://doi.org/10.1108/AAAJ-03-2013-1282

Kuzey, C., & Uyar, A. (2017). Determinants of sustainability reporting and its impact on firm value: Evidence from the emerging market of Turkey. *Journal of Cleaner Production*, *143*, 27–39. https://doi.org/10.1016/j.jclepro.2016.12.153

Lee, J., & Yu, J. (2019). Heterogenous Energy Consumption Behavior by Firm Size: Evidence from Korean Environmental Regulations. *Sustainability*, *11*(11), 3226. https://doi.org/10.3390/su11113226

Lee, K. H., Min, B., & Yook, K. H. (2015). The impacts of carbon (CO2) emissions and environmental research and development (R&D) investment on firm performance. *International Journal of Production Economics*, *167*, 1–11. https://doi.org/10.1016/j.ijpe.2015.05.018

Lewandowski, S. (2017). Corporate Carbon and Financial Performance: The Role of Emission Reductions. *Business Strategy and the Environment*, *26*(8), 1196–1211. https://doi.org/10.1002/bse.1978

Liao, L., Luo, L., & Tang, Q. (2015). Gender diversity, board independence, environmental committee and greenhouse gas disclosure. *The British Accounting Review*, *47*(4), 409–424. https://doi.org/10.1016/j.bar.2014.01.002

Loh, L., Thomas, T., & Wang, Y. (2017). Sustainability Reporting and Firm Value: Evidence from Singapore-Listed Companies. *Sustainability*, *9*(11), 2112. https://doi.org/10.3390/su9112112

Mahapatra, S. K., Schoenherr, T., & Jayaram, J. (2021). An assessment of factors contributing to firms' carbon footprint reduction efforts. *International Journal of Production Economics*, *235*, 11, Article 108073. https://doi.org/10.1016/j.ijpe.2021.108073

Manabe, S. (2019). Role of greenhouse gas in climate change. *Tellus A: Dynamic Meteorology and Oceanography*, *71*(1), 1620078. https://doi.org/10.1080/16000870.2019.1620078

Manetti, G., & Toccafondi, S. (2012). The Role of Stakeholders in Sustainability Reporting Assurance. *Journal of Business Ethics*, *107*(3), 363–377. https://doi.org/10.1007/s10551-011-1044-1

Matthews, H. S., Hendrickson, C. T., & Weber, C. L. (2008). The Importance of Carbon Footprint Estimation Boundaries. *Environmental Science & Technology*, *42*(16), 5839–5842. https://doi.org/10.1021/es703112w

Meng, X., Gou, D., & Chen, L. (2023). The relationship between carbon performance and financial performance: evidence from China. *Environmental Science and Pollution Research*, *30*(13), 38269–38281. https://doi.org/10.1007/s11356-022-24974-7

Misani, N., & Pogutz, S. (2015). Unraveling the effects of environmental outcomes and processes on financial performance: A non-linear approach. *Ecological Economics*, *109*, 150–160. https://doi.org/10.1016/j.ecolecon.2014.11.010

Mo, J. Y. (2022). Technological innovation and its impact on carbon emissions: evidence from Korea manufacturing firms participating emission trading scheme. *Technology Analysis & Strategic Management*, *34*(1), 47–57. https://doi.org/10.1080/09537325.2021.1884675

MSCI. (2023). The MSCI Net-Zero Tracker. https://www.msci.com/documents/1296102/41874802/NetZero-Tracker-NOV-cbr-en.pdf/e03b09df-911a-143c-bdda-36fd572a8972?t=1699917087121

Muktadir-Al-Mukit, D., & Bhaiyat, F. H. (2024). Impact of corporate governance diversity on carbon emission under environmental policy via the mandatory nonfinancial reporting regulation. *Business Strategy and the Environment*, *33*(2), 1397–1417. https://doi.org/10.1002/bse.3555

Muthuswamy, V. V., & Sharma, A. (2023). Moderating Effects of Environmental Governance on Environmental Innovations and Carbon Dioxide Emissions. *AgBioForum*, *25*(1), 203–214. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175838745&partnerID=40&md5=6f783d9f796a86f6f5de1c4730843fef

Mytton, D. (2020). Hiding greenhouse gas emissions in the cloud. *Nature Climate Change*, *10*(8), 701–701. https://doi.org/10.1038/s41558-020-0837-6

Nguyen, Q., Diaz-Rainey, I., Kitto, A., McNeil, B. I., Pittman, N. A., & Zhang, R. (2023). Scope 3 emissions: Data quality and machine learning prediction accuracy. *PLOS Climate*, *2*(11), e0000208. https://doi.org/10.1371/journal.pclm.0000208

Nishitani, K., & Kokubu, K. (2012). Why Does the Reduction of Greenhouse Gas Emissions Enhance Firm Value? The Case of Japanese Manufacturing Firms. *Business Strategy and the Environment*, *21*(8), 517–529. https://doi.org/10.1002/bse.734

O'Dwyer, B., & Owen, D. L. (2005). Assurance statement practice in environmental, social and sustainability reporting: a critical evaluation. *The British Accounting Review*, *37*(2), 205–229. https://doi.org/10.1016/j.bar.2005.01.005

Oestreich, M., & Tsiakas, I. (2023). Carbon Emissions and Firm Profitability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4353433

Ogunrinde, O., Shittu, E., & Dhanda, K. K. (2022). Distilling the Interplay Between Corporate Environmental Management, Financial, and Emissions Performance: Evidence From U.S. Firms. *IEEE Transactions on Engineering Management*, *69*(6), 3407–3435. https://doi.org/10.1109/TEM.2020.3040158

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Palazzo, G., & Scherer, A. G. (2006). Corporate Legitimacy as Deliberation: A Communicative Framework. *Journal of Business Ethics*, *66*(1), 71–88. https://doi.org/10.1007/s10551-006-9044-2

Palmer, K., Oates, W. E., & Portney, P. R. (1995). Tightening Environmental Standards: The Benefit-Cost or the No-Cost Paradigm? *Journal of Economic Perspectives*, *9*(4), 119–132. https://doi.org/10.1257/jep.9.4.119

Pan, T., Zhang, J., Wang, Y., & Shang, Y. (2024). The Impact of Environmental Regulations on Carbon Emissions of Chinese Enterprises and Their Resource Heterogeneity. *Sustainability*, *16*(3), 1058. https://doi.org/10.3390/su16031058

Patchell, J. (2018). Can the implications of the GHG Protocol's scope 3 standard be realized? *Journal of Cleaner Production*, *185*, 941–958. https://doi.org/10.1016/j.jclepro.2018.03.003

Patten, D. M. (2020). Seeking legitimacy. *Sustainability Accounting, Management and Policy Journal*, *11*(6), 1009–1021. https://doi.org/10.1108/SAMPJ-12-2018-0332

Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, *58*, 240–242. http://www.jstor.org/stable/115794

Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *The Review of Financial Studies*, *22*(1), 435–480. https://doi.org/10.1093/rfs/hhn053

Polizu, C., Khan, A., Kernan, P., Ellis, T., & Georges, P. (2023). Sustainability Insights: Climate Transition Risk: Historical Greenhouse Gas Emissions Trends For Global Industries. https://www.spglobal.com/ratings/en/research/articles/231122-sustainability-insights-climate-transition-risk-historical-greenhouse-gas-emissions-trends-for-global-indus-12921229#:~:text=By%20industry%3A%20The%20industry%20groups,the%20total%20scope%201%20emissions

Porles-Ochoa, F., & Guevara, R. (2023). Moderation of Clean Energy Innovation in the Relationship between the Carbon Footprint and Profits in CO2e-Intensive Firms: A Quantitative Longitudinal Study. *Sustainability*, *15*(13), 19, Article 10326. https://doi.org/10.3390/su151310326

Porter, M. E. (1980). *Techniques for analyzing industries and competitors*. Competitive Strategy.

Porter, M. E., & van der Linde, C. (1995). Toward a New Conception of the Environment-Competitiveness Relationship. *Journal of Economic Perspectives*, *9*(4), 97–118. https://doi.org/10.1257/jep.9.4.97

Qian, W., & Xing, K. (2018). Linking Environmental and Financial Performance for Privately Owned Firms: Some Evidence from Australia. *Journal of Small Business Management*, *56*(2), 330–347. https://doi.org/10.1111/jsbm.12261

Radonjič, G., & Tompa, S. (2018). Carbon footprint calculation in telecommunications companies – The importance and relevance of scope 3 greenhouse gases emissions. *Renewable and Sustainable Energy Reviews*, *98*, 361–375. https://doi.org/10.1016/j.rser.2018.09.018

Raval, A., Desai, R., & Bhatt, K. (2021). Nexus between carbon emission and financial performance moderated by environmental sensitivity: evidence from emerging economy. *International Journal of Managerial and Financial Accounting*, *13*(3-4), 209–231. https://doi.org/10.1504/IJMFA.2021.120516

Rehman, I. U., Shahzad, F., Hanif, M. A., Arshad, A., & Sergi, B. S. (2024). Financial constraints and carbon emissions: an empirical investigation. *Social Responsibility Journal*, *20*(4), 761–782. https://doi.org/10.1108/SRJ-01-2023-0014

Ritchie, H., Rosado, P., & Roser, M. (2020). Breakdown of carbon dioxide, methane and nitrous oxide emissions by sector. https://ourworldindata.org/emissions-by-sector

Rodríguez-García, M. D. P., Galindo-Manrique, A. F., Cortez-Alejandro, K. A., & Méndez-Sáenz, A. B. (2022). Eco-efficiency and financial performance in Latin American countries: An environmental intensity approach. *Research in International Business and Finance*, *59*, Article 101547. https://doi.org/10.1016/j.ribaf.2021.101547

Rokhmawati, A., & Gunardi, A. (2017). Is going green good for profit? Empirical evidence from listed manufacturing firms in Indonesia. *International Journal of Energy Economics and Policy*, *7*(4), 181–192. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030531346&partnerID=40&md5=48d10b2d332e77f292032b1b03d4dff6

Rokhmawati, A., Gunardi, A., & Rossi, M. (2017). How powerful is your customers' reaction to carbon performance? Linking carbon and firm financial performance. *International Journal of Energy Economics and Policy*, *7*(6), 85–95. https://hdl.handle.net/11159/1403

Rokhmawati, A., Sathye, M., & Sathye, S. (2015). The Effect of GHG Emission, Environmental Performance, and Social Performance on Financial Performance of Listed Manufacturing Firms in Indonesia. *Procedia Social and Behavioral Sciences*, *211*, 461–470. https://doi.org/10.1016/j.sbspro.2015.11.061

Roston, M., Seiger, A., & Mathieson, A. (2024). What's Scope 2 Good For? https://doi.org/10.2139/ssrn.4638672

SBTi. (2024). SBTi Corporate Near-Term Criteria. https://sciencebasedtargets.org/resources/files/SBTi-criteria.pdf

Shahgholian, A. (2019). Unpacking the relationship between environmental profile and financial profile; literature review toward methodological best practice. *Journal of Cleaner Production*, *233*, 181–196. https://doi.org/10.1016/j.jclepro.2019.05.365

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples)†. *Biometrika*, *52*(3-4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, *68*(1), 45–54. https://doi.org/10.1093/biomet/68.1.45

Singhania, M., & Chadha, G. (2023). Thirty years of sustainability reporting research: a scientometric analysis. *Environmental Science and Pollution Research*, *30*(46), 102047–102082. https://doi.org/10.1007/s11356-023-29452-2

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Solomon, S., Plattner, G.-K., Knutti, R., & Friedlingstein, P. (2009). Irreversible climate change due to carbon dioxide emissions. *Proceedings of the National Academy of Sciences*, *106*(6), 1704–1709. https://doi.org/10.1073/pnas.0812721106

STOXX. (2024). STOXX Europe 600 and Derived Indices. https://stoxx.com/wp-content/uploads/2024/06/STOXX_Europe_600_202406.pdf

Subramaniam, N., Akbar, S., Situ, H., Ji, S., & Parikh, N. (2023). Sustainable development goal reporting: Contrasting effects of institutional and organisational factors. *Journal of Cleaner Production*, *411*, 137339. https://doi.org/10.1016/j.jclepro.2023.137339

Tarmizi, N. F. A., & Brahmana, R. K. (2023). Environmental performance, political connection, and financial performance: evidence from global oil and gas companies. *Environmental Science and Pollution Research*, *30*(4), 11081–11098. https://doi.org/10.1007/s11356-022-22881-5

Tarquinio, L., & Posadas, S. C. (2020). Exploring the term "non-financial information": an academics' view. *Meditari Accountancy Research*, *28*(5), 727–749. https://doi.org/10.1108/MEDAR-11-2019-0602

Tatsuo, K. (2010). An analysis of the eco-efficiency and economic performance of Japanese companies. *Asian Business and Management*, *9*(2), 209–222. https://doi.org/10.1057/abm.2010.3

Testa, M., & D'Amato, A. (2017). Corporate environmental responsibility and financial performance: does bidirectional causality work? Empirical evidence from the manufacturing industry. *Social Responsibility Journal*, *13*(2), 221–234. https://doi.org/10.1108/SRJ-02-2016-0031

Thyssenkrupp. (2024). thyssenkrupp AG - Climate Change 2023 - CDP Response. https://www.cdp.net/en/formatted_responses/responses?campaign_id=83630982&discloser_id=1030443&locale=en&organization_name=thyssenkrupp+AG&organization_number=19080&program=Investor&project_year=2023&redirect=https%3A%2F%2Fcdp.credit360.com%2Fsurveys%2F2023%2Fjwbhd7d6%2F291584&survey_id=82591262

Tingbani, I., Chithambo, L., Tauringana, V., & Papanikolaou, N. (2020). Board gender diversity, environmental committee and greenhouse gas voluntary disclosures. *Business Strategy and the Environment*, *29*(6), 2194–2210. https://doi.org/10.1002/bse.2495

Tobin, J. (1969). A General Equilibrium Approach To Monetary Theory. *Journal of Money, Credit and Banking*, *1*(1), 15–29. https://doi.org/10.2307/1991374

Tomar, S. (2022). Greenhouse gas disclosure and emissions benchmarking. *SMU Cox School of Business Research Paper*, (19-17). https://doi.org/10.1111/1475-679X.12473

Trucost. (2013). Natural Capital at Risk: The Top 100 Externalities of Business. https://capitalscoalition.org/wp-content/uploads/2016/07/Trucost-Nat-Cap-at-Risk-Final-Report-web.pdf

Tsalis, T. A., Malamateniou, K. E., Koulouriotis, D., & Nikolaou, I. E. (2020). New challenges for corporate sustainability reporting: United Nations' 2030 Agenda for sustainable development and the sustainable development goals. *Corporate Social Responsibility and Environmental Management*, *27*(4), 1617–1629. https://doi.org/10.1002/csr.1910

United Nations. (1998). Adoption of the Kyoto Protocol to the United Nations Framework Convention on Climate Change. https://unfccc.int/cop5/resource/docs/cop3/07a01.htm

United Nations. (2015a). The Paris Agreement. https://unfccc.int/process-and-meetings/the-paris-agreement

United Nations. (2015b). Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1). https://documents.un.org/doc/undoc/gen/n15/291/89/pdf/n1529189.pdf

Vaitiekuniene, R., Sutiene, K., Kovalov, B., & Krusinskas, R. (2024). Does the Financial and Innovation Performance of European and Asian–Oceanian Companies Coincide with the Targets of the Green Deal? *Sustainability (Switzerland)*, *16*(4), Article 1485. https://doi.org/10.3390/su16041485

van Emous, R., Krušinskas, R., & Westerman, W. (2021). Carbon emissions reduction and corporate financial performance: the influence of country-level characteristics. *Energies*, *14*(19), Article 6029. https://doi.org/10.3390/en14196029

van Vuuren, D. P., & Riahi, K. (2008). Do recent emission trends imply higher emissions forever? *Climatic Change*, *91*(3), 237–248. https://doi.org/10.1007/s10584-008-9485-y

Van Greuning, H., Scott, D., & Terblanche, S. (2011). *International financial reporting standards: a practical guide*. World Bank Publications.

Velte, P. (2023). Determinants and financial consequences of environmental performance and reporting: A literature review of European archival research. *Journal of Environmental Management*, *340*, Article 117916. https://doi.org/10.1016/j.jenvman.2023.117916

Waddock, S. A., & Graves, S. B. (1997). The Corporate Social Performance-Financial Performance Link. *Strategic Management Journal*, *18*(4), 303–319. http://www.jstor.org/stable/3088143

Wallage, P. (2000). Assurance on Sustainability Reporting: An Auditor's View. *AUDITING: A Journal of Practice & Theory*, *19*(s-1), 53–65. https://doi.org/10.2308/aud.2000.19.s-1.53

Wang, J., Li, J., & Zhang, Q. (2021). Does carbon efficiency improve financial performance? Evidence from Chinese firms. *Energy Economics*, *104*, Article 105658. https://doi.org/10.1016/j.eneco.2021.105658

Wang, L., Li, S., & Gao, S. (2014). Do Greenhouse Gas Emissions Affect Financial Performance? - An Empirical Examination of Australian Public Firms. *Business Strategy and the Environment*, *23*(8), 505–519. https://doi.org/10.1002/bse.1790

Wang, Q. (2023). Financial effects of carbon risk and carbon disclosure: A review. *Accounting and Finance*, *63*(4), 4175–4219. https://doi.org/10.1111/acfi.13090

WBCSD. (2011). Corporate Value Chain (Scope 3) Accounting and Reporting Standard. https://ghgprotocol.org/sites/default/files/standards/Corporate-Value-Chain-Accounting-Reporing-Standard_041613_2.pdf

Wedari, L. K., Moradi-Motlagh, A., & Jubb, C. (2023). The moderating effect of innovation on the relationship between environmental and financial performance: Evidence from high emitters in Australia. *Business Strategy and the Environment*, *32*(1), 654–672. https://doi.org/10.1002/bse.3167

Wei, W., Zhang, P., Yao, M., Xue, M., Miao, J., Liu, B., & Wang, F. (2020). Multi-scope electricity-related carbon emissions accounting: A case study of Shanghai. *Journal of Cleaner Production*, *252*, 119789. https://doi.org/10.1016/j.jclepro.2019.119789

Wells, P., & Nieuwenhuis, P. (2012). Transition failure: Understanding continuity in the automotive industry. *Technological Forecasting and Social Change*, *79*(9), 1681–1692. https://doi.org/10.1016/j.techfore.2012.06.008S

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal*, *5*(2), 171–180. https://doi.org/10.1002/smj.4250050207

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Wooldridge, J. M. (2012). *Introductory econometrics : a modern approach* (Fifth). South-Western Cengage Learning.

WRI & WBCSD. (2004). The GHG Protocol: A corporate reporting and accounting standard (revised edition). https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

Wu, S. R., Shao, C., & Chen, J. (2018). Approaches on the Screening Methods for Materiality in Sustainability Reporting. *Sustainability*, *10*(9), 3233. https://doi.org/10.3390/su10093233

Xia, M., & Cai, H. H. (2023). The driving factors of corporate carbon emissions: an application of the LASSO model with survey data. *Environmental Science and Pollution Research*, *30*(19), 56484–56512. https://doi.org/10.1007/s11356-023-26081-7

Yin, J., Ibrahim, S., Mohd, N. N. A., Zhong, C., & Mao, X. (2024). Can green finance and environmental regulations promote carbon emission reduction? Evidence from China. *Environmental Science and Pollution Research*, *31*(2), 2836–2850. https://doi.org/10.1007/s11356-023-31231-y

Young, E. (2023). The evolution of the ESG reporting landscape. *EY*. Retrieved June 1, 2025, from https://www.ey.com/en_us/insights/climate-change-sustainability-services/the-evolution-of-the-esg-reporting-landscape

Zsóka, Á., & Vajkai, É. (2018). Corporate sustainability reporting: Scrutinising the requirements of comparability, transparency and reflection of sustainability performance. *Society and Economy Soc Ec*, *40*(1), 19–44. https://doi.org/10.1556/204.2018.40.1.3

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Small but Powerful: The Impact of Shelf Talker Flags on Consumer Shopping Behavior

Günther Gamper

University of Innsbruck

Abstract

Unseen is unsold, which means that shoppers can only buy what they see in the store. Therefore, retailers use different in-store marketing techniques to increase visual exposure and stimulate purchases. In this paper, I investigate the effect of shelf talker flags on consumer shopping behavior. In doing so, I hypothesize that shelf talker flags increase the subjectively perceived search ease and purchases of marked products. A field experiment shows that shelf talker flags make products more visible and easier to find at the point of purchase, significantly increasing consumers' subjectively perceived search ease. Furthermore, the results suggest that shelf talker flags can influence consumer buying behavior and increase purchases of marked products. However, this result is only marginally significant.

Keywords: in-store marketing; search ease; shelf talker flags; unplanned purchases; visual attention

1. Introduction

In retail, the "unseen is unsold" paradigm often applies, meaning that goods that do not reach customers visually cannot be purchased (Wästlund et al., 2018). The basic function of the retailer is to confront the buyer with his offer so that he can satisfy his needs (Streicher et al., 2021). However, nowadays supermarkets often have several thousand stock-keeping units (SKUs) in their assortment (Schwartz, 2004), so this is not an easy task. That's why retailers use various techniques to increase product visibility and encourage purchases. This has benefits for both sides, for the retailers as well as for the shoppers. On the one hand, if shoppers see more products, they make more purchases, especially unplanned purchases. On the other hand, customers also benefit from an improved presentation of goods, for example by simplifying the search process in stores (Chandon et al., 2000). This is important because supermarkets in reality tend to have complex and large assortments, which otherwise often lead to decision-making difficulties for the shoppers (Iyengar & Lepper, 2000). So, what can retailers do? What they can do ranges from structural aspects in product presentation to shelf management strategies to promotional signals at the point of purchase. One specific promotional tool used to make products visually salient at the point of purchase is shelf talker flags. In America and England, they are already widespread. However, in Austria, they are not yet found in supermarkets. Additionally, while other visual sales promotions, such as in-store displays and their impact on consumer shopping behavior, are well studied (e.g., Chandon et al., 2009; Roggeveen et al., 2016), there is hardly any research that specifically addresses shelf talker flags. Therefore, this paper aims to fill this gap by investigating the effect of shelf talker flags on consumer shopping behavior.

Specifically, I hypothesize that shelf talker flags increase subjectively perceived search ease and purchases of marked products. A field experiment shows that shelf talker flags make products more visible and easier to find at the point of purchase, significantly increasing consumers' subjectively perceived search ease. Furthermore, the results suggest that shelf talker flags can influence consumer buying behavior and increase purchases of marked products. However, this result is only marginally significant.

I would like to express my sincerest gratitude to my Bachelor's thesis supervisor, Dr. Mathias Streicher. His extensive experience, invaluable guidance, and assistance with data collection and analysis have been crucial to the success of this thesis. I am deeply grateful for his support and the confidence he has placed in me.

DOI: https://doi.org/10.5282/jums/v10i2pp334-348

The remainder of the paper is divided into four sections. First, a review of the existing literature on unplanned purchases and visual attention is provided, and the hypotheses are derived. Second, the research design of the experiment is described. Third, the results are reported. Finally, the results and their implications as well as the limitations of this work and future research directions are discussed.

2. Literature Review

2.1. The relationship between unplanned purchases and visual attention

2.1.1. Unplanned Purchases

Consumer purchases in retail stores can be divided into planned purchases and unplanned purchases. With planned purchases, consumers plan what purchases they will make (e.g., using a shopping list) before they visit the store (Bucklin & Lattin, 1991). This contrasts with unplanned purchases, which are not planned a priori and are often triggered by in-store stimuli (Inman et al., 2009). As reported by the Point of Purchase Advertising Institute (POPAI), for the vast majority of purchases (74%), the decision to buy or not buy a product is made while shopping (POPAI, 1997). As a result, unplanned purchases account for a significant proportion of customers' overall shopping behavior. Approximately 62% of all consumer purchases at mass retailers (e.g., Target), are unplanned (POPAI, 2014). This is consistent with a study by Inman, Winer, and Ferraro (2009), in which the proportion of unplanned purchases was 60.9%. Unplanned purchases can therefore be seen as an important component of retailers' profits (Gilbride et al., 2015). Although unplanned purchases are often associated with negative effects for shoppers e.g., with a loss of self-control leading to excessive spending (Rook, 1987; Streicher et al., 2021), shoppers are not unaware of them. On the contrary, shoppers actually regulate their unplanned purchases by having an implicit budget for them (Stilley et al., 2010). As Stilley, Inman, and Wakefield (2010) report, consumers have a fixed budget in mind for their purchases, which consists of spending on planned purchases and a residual amount for in-store decisions and thus for unplanned purchases. They refer to the latter portion as "in-store slack." In addition, unplanned purchases can also have positive effects for buyers. For example, they can use an improved product presentation to find out about alternatives to their standard products, which may actually be better (Iyer, 1989).

Unplanned purchases often occur because in-store stimuli remind customers of forgotten needs or trigger new needs (Inman et al., 2009). When consumers perceive a stimulus in a store, they do not simply ignore it but evaluate it (Yeung & Wyer, 2004), and if it's useful for their purposes, this may trigger affective or cognitive responses (Inman et al., 2009). However, as Inman, Winer, and Ferraro (2009) show, not all stimuli are equally suitable for this. Coupons for instance tend to harm unplanned purchases (Inman et al., 2009) because consumers usually decide whether or not to use a coupon before entering a store (Kahn & Schmittlein, 1989). Therefore, coupons are more likely to play a role in planned purchases (Inman et al., 2009). In contrast, they find that displays have a positive impact on unplanned purchases, especially when it comes to frequently needed purchases. Interestingly, consumers seem to expect products in end-of-aisle displays to be discounted, which leads to an increase in sales of the products, even if they have a normal price (Inman et al., 1990). There are also differences between the product categories. The study by Inman, Winer, and Ferraro (2009) finds that less frequently visited and hedonistic categories have a higher probability of an unplanned purchase. As these examples show, in-store stimuli can lead to purchases, especially unplanned purchases. However, for these in-store stimuli to have an impact, they must receive the visual attention of consumers in order to be noticed at all.

2.1.2. Visual attention

Attention, specifically visual attention, has a strong influence on in-store decisions and therefore plays an important role in consumer decision-making (Orquin & Mueller Loose, 2013). According to Russo and Leclerc (1994), consumer decision-making in the store can be divided into three phases: Orientation, Evaluation, and Review. Visual attention refers to the process by which visual impressions are filtered and selected for subsequent processing and eventual incorporation into awareness (Paré & Dorris, 2012). Several studies show that the extent of visual attention is strongly related to eye movements (e.g., Deubel and Schneider, 1996; Hoffman and Subramaniam, 1995; Kowler et al., 1995). Henderson and Hollingworth (1999) report that people are able to subliminally process the essence of a visual impression through an initial fixation from the peripheral visual field. If Inow relate it to shopping situations, this means that people can get a first rough overview of the assortment in the store (Chandon et al., 2009). However, to visually process more detailed information, for example about a product, a fixation of the eyes is required (Burke & Leykin, 2014). Therefore, eye fixations are considered a good measure of visual attention in research (Chandon et al., 2009; Orquin & Mueller Loose, 2013).

As described by Elmo Lewis' AIDA model (cf. *Figure 1*), the customer's buying process begins with attention and ends with action (Heath & Feldwick, 2008). This means that a product must first attract the customer's attention so that the other phases (interest, desire, action) can take place at all. Fittingly, in retail, there is the well-known saying "unseen is unsold", which means that consumers can only buy those products that they also visually perceive (Wästlund et al., 2018). This can be well explained by the way visual processing works. If products do not attract visual attention, then they do not receive eye fixations from consumers and are thus not identified (Orquin & Mueller Loose, 2013). Thus, consumers do not even know that the products are available and therefore they remain unsold.

How much consumers see and, conversely, then buy depends, among other things, on the breadth of their atten-

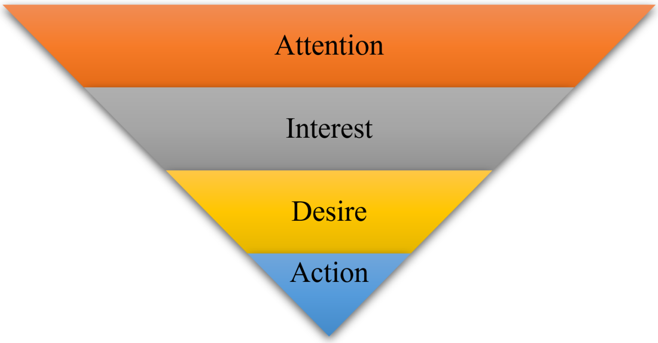


Figure 1: The AIDA model (Source: Adapted from Li and Yu (2013, p. 48))

tion (Streicher et al., 2021). Attentional breadth describes whether people focus their gaze on “a wider or a more limited visual area” (Friedman et al., 2003, p. 278). While people with narrow attention focus only on a fraction of all visual stimuli and ignore others (Wadlinger & Isaacowitz, 2006), people with broad attention are more susceptible to visual stimuli and tend to pay attention to a larger number of stimuli (Streicher et al., 2021). By manipulating consumers' attentional breadth in field and laboratory experiments, Streicher, Estes, and Büttner (2021) find that attentional breadth affects product choices and increases unplanned purchases. As they report, this is caused by broader attention activating an exploratory mindset, which leads consumers to explore the store more.

2.1.3. In-store exploration

In order for an SKU to have any chance at all of catching the attention of shoppers, they must first visit the area of the store where the product is positioned in the first place (Chen et al., 2021). For a long time, retailers assumed that customers go from aisle to aisle while shopping in a supermarket, walking through almost the entire supermarket (Hui et al., 2013). However, it seems that this doesn't quite correspond to reality. Recent studies that have used path tracking to examine consumer shopping behavior in retail stores have found that, on average, shoppers visit only about a third of the total store (Hui & Bradlow, 2012). Instead of walking through the entire store, the majority of shoppers stay mainly in the peripheral areas of the store and visit only those aisles that are relevant to their need fulfillment (Hui et al., 2009; Larson et al., 2005). As a result, most aisles are not even crossed by shoppers, leaving large parts of the assortment unseen (Hui et al., 2013). From a retailer's point of view, this is obviously a problem, because if shoppers don't see the majority of the product categories, much remains unsold. To counteract this, retailers often try to trigger exploratory shopping, a behavioral pattern of consumers that influences their shopping behavior (Streicher et al., 2021). Contrary to typical consumers, shoppers with an exploratory mindset, browse through the store and explore the merchandise by moving or looking around the store (Baumgartner & Steenkamp, 1996; Streicher et al., 2021). As Titus and Everett (1995) argue, shopper locomotion is significantly influenced by the nature of the search strategy. As consumers with an exploratory mindset browse more in-store, they see a greater number of product categories and stimuli in the store (Streicher et al., 2021), which in turn increases the likelihood of unplanned purchases (Granbois, 1968). Therefore, retailers try to encourage shoppers to explore the store. Exposure to in-store stimuli may even cause customers to experience some urge to buy, and make impulse purchases (Rook, 1987).

2.1.4. Impulse buying

Impulsive shoppers view shopping more positively, enjoy exploring the store and are more inclined to make in-store purchase decisions (Beatty & Ferrell, 1998). At this point, it's important to distinguish between impulse purchases and unplanned purchases, as they are similar but not the same thing. Although impulse purchases are also unplanned, the key differentiator is a certain urge to buy that consumers feel (Beatty & Ferrell, 1998). Impulsive shoppers are characterized, among other things, by the fact that they usually don't think long (Hoch & Loewenstein, 1991) and react strongly emotionally when shopping (Rook & Fisher, 1995; Verplanken & Herabadi, 2001). In addition, unplanned purchases by impulsive shoppers also occur because they browse around the store more than other shoppers (Beatty & Ferrell, 1998). Impulse buying is often associated with negative outcomes such as the buyer experiencing financial problems, feelings of guilt, or post-purchase disappointment (Rook, 1987). It can even go so far that consumers develop pathological buying behavior, where they are subject to a certain compulsion to buy (O'Guinn & Faber, 1989). Past research on impulse buying has suggested that for impulsive shoppers, the products purchased may be secondary, as it is primarily the shopping process that triggers positive feelings in them (O'Guinn & Faber, 1989; Verplanken & Herabadi, 2001). Additionally, the behavior of impulsive buyers is often characterized by hedonic motives such as variety and pleasure-seeking (Bayley & Nancarrow, 1998). So, you could say that it's not so much the products as the shopping itself that they enjoy.

Recently, a link between impulsive buying behavior and an attentional bias has also been found (Büttner et al., 2014).

Because shoppers with high buying impulsivity are more open to sudden purchases, they are more easily attracted to in-store stimuli (Hoch & Loewenstein, 1991). As a result, impulsive shoppers find it harder to control their visual attention, which leads them to be distracted more often by other products in the assortment (Büttner et al., 2014). By shifting their attention to more products, they notice more products, which in turn can then lead shoppers to make more unplanned purchases (Büttner et al., 2014). In addition, attention to enticing stimuli also makes self-control more difficult (Field & Eastwood, 2005) which further encourages impulse buying. Consistent with these results, Streicher, Estes, and Büttner (2021) show in their studies that the visual attention of impulsive shoppers has a causal effect on their shopping behavior.

However, in addition to visual attention, there are also situational factors and customer characteristics, that influence unplanned purchases.

2.1.5. Situational factors & shopper characteristics

Whether or not a customer makes unplanned purchases also depends in part on situational factors that occur during the shopping experience (Park et al., 1989). Consumer sentiment, for example, has a strong effect on the number of impulse purchases as various studies (e.g., Gardner and Rook, 1988; Rook, 1987) show. Consumers who view shopping positively and enjoy exploring the store have a higher likelihood of making purchase decisions in the store (Beatty & Ferrell, 1998). In addition, consumers are generally more likely to visit a store area if other shoppers are there (Hui et al., 2009). However, as soon as too many people are there, it becomes too crowded for shoppers, and the likelihood of them buying a product is reduced (Zhang et al., 2014). In addition, a study by Luo (2005) suggests that the presence of people influences perceived desire to buy, as participants reported varying degrees of urge to buy when they imagined that peers or family members were present. This may be due to the different social relationships between them and the buyer (Rook & Fisher, 1995). However, a study by Inman, Winer, and Ferraro (2009) shows that whether or not individuals are accompanied while shopping does not appear to influence unplanned purchases.

Furthermore, the time available for the purchase plays a role. According to Park, Iyer, and Smith (1989), shoppers who don't know the store well and are not in a hurry are those with the most unplanned purchases. Time pressure generally has a strong influence on consumers' in-store decision behavior (Iyer, 1989). When consumers are in a hurry due to time pressure, it has a negative impact on unplanned purchases (Iyer, 1989; Park et al., 1989). This is due to the fact that shoppers don't have much time to search for products or to explore the store (Beatty & Smith, 1987). Research on visual processing reports that when time pressure is present, consumers try to acquire information faster by fixating on individual objects for less time (Pieters & Warlop, 1999). Moreover, Hui, Bradlow, and Fader (2009) argue, that as shopping time progresses, shoppers begin to browse less and shop more efficiently when they realize that they have already spent too much time shopping. This in turn can lead to perceived time pressure that causes shoppers to move faster, stop less often, and thus come into less contact with products (Titus & Everett, 1995).

Another factor is the personal characteristics and inclinations of consumers. Shoppers vary in their susceptibility to in-store buying decisions (Streicher et al., 2021). Some shoppers find it very difficult to resist temptations and therefore repeatedly make unplanned purchases that they subsequently regret (Faber & Vohs, 2011). This consumer behavior is often associated with impulse buying (Rook & Fisher, 1995; Verplanken & Herabadi, 2001). Inman, Winer, and Ferraro (2009) also find a greater likelihood of unplanned purchases among female shoppers and larger households (e.g., consisting of five people). They argue that this is mainly the case because women tend to shop more often than men (Starrels, 1994) and larger households buy more products. As a result, both consumer groups come more into contact with stimuli in the store that can trigger certain needs (Inman et al., 2009). Furthermore, shopping regularity plays a role. When consumers shop regularly, they are likely to need only a few products and focus on buying them quickly and leaving the store (Inman et al., 2009), which negatively affects unplanned purchases. In contrast, when consumers shop infrequently and therefore make larger purchases at one time to satisfy their needs, unplanned purchases are more probable (Bell et al., 2011).

Consumers' purchase goals have a strong influence on their purchase intentions and their levels of in-store visual attention (Burke & Leykin, 2014). When consumers set their shopping goals only roughly in advance, unplanned purchases are more likely to occur (Bell et al., 2011). Some consumers see shopping as an experience and a fun activity, while others just want to quickly buy the groceries they need, making shopping a task that must be completed (Babin et al., 1994; Kaltcheva & Weitz, 2006). Buyers who are more task-oriented may use a shopping list to remind them of the needed products they want to buy (Block & Morwitz, 1999). Shopping lists are very helpful for planned purchases but have a negative effect on unplanned purchases (Inman et al., 2009). Even the payment method plays a role. Paying with card can make purchases more pleasant, as it is considered to be not as painful as buying with cash (Prelec & Loewenstein, 1998). Indeed, consumers who pay by card or check tend to make more unplanned purchases (Inman et al., 2009).

As this section has shown, unplanned purchases depend heavily on the visual attention that is attracted in the store. Since unplanned purchases make up a large part of retailers' revenues (Gilbride et al., 2015), they strive to create as much visual exposure as possible. Common methods such as travel distance, shelf management strategies, and promotional signals at the point of purchase are discussed in the next section.

2.2. In-store marketing techniques to create visual exposure

The first method retailers can use to enhance shoppers' exposure to in-store products is supporting in-store exploration by increasing in-store travel distance. Several studies (e.g., Kollat and Willett, 1967; Park et al., 1989) report that physical products in stores can serve as external memory aids for shoppers to activate new or forgotten needs. Thus, if more products come into the customer's field of vision as he or she moves further around the store, this can increase their awareness of potentially interesting products and trigger unplanned purchases (Granbois, 1968). In research, two primary metrics are used to determine the distance traveled in the store: the amount of aisles visited, and the time spent shopping in the store (Hui et al., 2013). Both have been shown to have a positive impact on unplanned purchases by increasing shoppers' exposure to in-store stimuli (Granbois, 1968; Hui et al., 2013; Inman et al., 2009). To extend in-store travel distance, retailers can use two techniques. On the one hand, retailers can relocate products in the store, and on the other hand, mobile promotions can be used (Hui et al., 2013).

The first strategy increases the travel distance by changing the structural aspects of the product presentation. This method is based on Granbois' (1968) study and is the classic method for increasing travel distance in the store (Hui et al., 2013). In this method, retailers place popular product categories (e.g., milk) scattered throughout the store so that customers have to travel a longer distance to make their planned purchases (Granbois, 1968; Iyer, 1989). Along the way, shoppers are then exposed to more in-store stimuli in hopes of increasing unplanned purchases (Hui et al., 2013). Fittingly, an old retail adage says that you should hide the milk in the back of the store (Hui et al., 2013). In this manner, retailers encourage shoppers to walk throughout the store and make unplanned purchases. A prominent example is IKEA. IKEA gives its customers a predefined path to follow, where it almost forces them to walk past all the products to get to the checkout (Streicher et al., 2021).

In the second strategy, retailers use mobile promotions to attract shoppers to low-traffic categories. This can be achieved through tailored promotions, e.g., coupon offers (Hui et al., 2013). Retailers can work with location-based grocery apps to get information about shoppers' locations and shopping lists. Then they can use this information to entice the shopper into unplanned categories through tailored coupon offers, to stimulate unplanned purchases (Hui et al., 2013).

In a field experiment, Hui, Inman, Huang, and Suher (2013) investigated the impact of both strategies on unplanned purchases by collecting path data from shoppers in stores with RFID tags. The results indicate that both strategies (product relocation vs. mobile promotions) can increase unplanned purchases. In the case of the product relocation strategy, they find that it may increase unplanned purchases by 7.2% versus 16.1% in the case of mobile promotions. As the researchers argue, the key benefit of the latter technique is the individualization option for each consumer. Since neither strategy precludes the other, retailers could use both simultaneously to increase unplanned purchases (Hui et al., 2013). However, there may be a trade-off between encouraging unplanned purchases by increasing travel distance and making it convenient for shoppers to purchase products. By lengthening the distance customers must travel to find their planned purchases, some customers may find the store unpleasant and avoid shopping there in the future (Hui et al., 2013). Therefore, retailers should carefully consider whether and to what extent they want to pursue this strategy.

In contrast to travel distance, which can be seen more as a store-wide strategy closely related to the store layout, there are also shelf-level methods retailers can use to increase visual exposure. Various eye-movement studies (e.g., Janiszewski, 1998; Lohse, 1997) report that attention can be influenced and even increased by advertising or catalog displays. In addition, according to a consumer study by Rook (1987), visual confrontation with products or stimuli in stores can trigger an increased feeling of compulsion to buy. Therefore, retailers use various methods to draw customers' attention to certain products, e.g., by increasing shelf space for certain product categories or positioning products several times in the store to increase their visibility and thus the chance of a purchase (Chandon et al., 2009; Streicher et al., 2021). Shelf space can be defined as the amount of space a product category occupies in the store (Campo & Gijsbrechts, 2005). By expanding the shelf space of product categories, sales of these products can be increased as Wilkinson, Mason, and Paksoy (1982) show. However, space constraints don't always allow retailers to increase shelf space. Alternatively, retailers can raise the number of facings of certain products in the store to draw more attention to them, while keeping the overall product category space constant (Chandon et al., 2009; Drèze et al., 1994). Indeed, as an eye-tracking study by Chandon, Hutchinson, Bradlow, and Young (2009) shows, an increase in the number of shelf facings can enhance product sales. This is because a higher number of shelf facings influences visual attention, which in turn acts as a mediator for brand evaluation (Chandon et al., 2009). The results also indicate that this is especially the case for low-market brands, young and educated shoppers, and regular customers of the brands. Further research indicates that this effect is partly influenced by shoppers' expectations. Shoppers may interpret a high number of facings as an indication of an important brand (Buchanan et al., 1999).

Last but not least, sales promotions and promotional signals are also often used at the point of purchase (POP) to advertise products (Chandon et al., 2000). Sales promotions can be "defined as temporary and tangible monetary or non-monetary incentives intended to have a direct impact on consumer behavior" (Chandon et al., 2000, p. 65). Monetary promotions such as discounts, coupons, and rebates (Chandon et al., 2009) are not the exception in the retail sector, but rather the rule. Price promotions can be used by retailers to attract visitors and increase traffic (Grewal et al., 1998) and a few studies also showed that they have an effect, even though short-lived, on brand performance (Dodson et al., 1978; Doob et al., 1969). In contrast to these positive effects, however, there are also some negative aspects. Price discounts don't affect sales in the long run (Dodson et al., 1978; Doob et al., 1969), are costly for retailers, and reduce profits (Jedidi et al., 1999). Moreover, price promotions can make customers more price-sensitive in the long run (Mela et al., 1997).

Because of these effects, retailers are increasingly also using nonmonetary promotions, which are designed to attract the attention of consumers at the point of sale (Chandon et al., 2009). In practice, such visual promotions include, for example, displays, shelf talker flags, in-store advertising, or flyers (Ailawadi et al., 2009). Gaining attention at the POP is essential because it strongly influences consumers' purchase decisions (Chandon et al., 2007). For example, Woodside and Waddle (1975) show that point-of-purchase signing can increase sales even when there is no price reduction. Displays also influence consumer behavior and can have a positive effect on retailers' sales by stimulating unplanned purchases as shown by various studies (e.g., Chandon et al., 2009; Roggeveen et al., 2016). Well-known brands such as Coca-Cola have long understood this and rely on creative store displays to boost product sales (Keh et al., 2021).

Regardless of whether the promotions are monetary or nonmonetary, the fundamental question is why consumers respond to promotions in the first place. As Chandon, Wansink, and Laurent (2000) argue, consumers respond to promotions, because they provide various benefits for them. The main reason why consumers respond to promotions is generally thought to be the associated cost savings, e.g., in the form of discounts or rebates (Blattberg & Neslin, 1993). In addition, promotions enable shoppers to switch to higher-quality products and facilitate the search process in the supermarket (Inman et al., 1990; Wansink et al., 1998). Finally, nonmonetary promotions in particular also offer shoppers the opportunity to satisfy hedonic needs such as entertainment, exploration, and value expression (Chandon et al., 2000).

Since this section is mainly about techniques retailers use to attract visual attention in their stores, the influence of promotional signals on it should be highlighted. Promotional signals visually highlight products at the point of purchase, greatly simplifying the search process and ultimately making the entire shopping experience more convenient for the consumer (Chandon et al., 2000). This function is essential because, as will be explained in the next section, too much visual exposure can also lead to decision-making difficulties.

2.3. Assortment – Less can be more

As shown in the previous section, retailers try to create as much visual exposure as possible through various techniques. However, too much exposure might negatively affect consumer decisions. In this context, of course, the assortment and especially the size of the assortment plays a role. Assortment can be defined as "the number of different items in a merchandise category" (Levy & Weitz, 1995, p. 30). Assortment size plays a key role in retailing and is an important factor for consumers, in the selection of the store (Iyengar & Lepper, 2000). Therefore, it's a frequently discussed topic, both in practice and in research.

By offering assortments that customers can use to meet their needs and wants, retailers increase the value of products to shoppers (Oppewal & Koelemeijer, 2005). As retailers generally try to present as many products as possible to consumers (Streicher et al., 2021), this often leads to large assortments. In modern society, generally dominates the assumption that more choices are better because it gives you some freedom of choice (Iyengar & Lepper, 2000). This attitude also applies to the retail world, where the myth prevails that a large assortment is always better. Overall, studies have shown that consumers like more choices and don't like to be limited in their decisions (Broniarczyk et al., 1998; Fitzsimons, 2000). Underlying this is the expectation of shoppers that they will be better able to satisfy their needs if they have more choices (Kahn & Lehmann, 1991). However, more recent research takes a more differentiated view on assortment size and suggests that a large assortment does not always have to be better; on the contrary, it can even have serious consequences. Apart from the fact that operating costs naturally increase with the number of SKUs in the store (Oppewal & Koelemeijer, 2005) too large assortments can influence purchase behavior negatively (Iyengar & Lepper, 2000).

For example, Diehl and Poynor (2010) find in their studies that purchases from a large assortment lead to lower satisfaction levels on the shoppers' side compared to purchases from a smaller assortment. Satisfaction in this context can be understood as the evaluation of a product after a decision has been made, while expectations express certain assumptions that often refer to the future (Oliver, 1996). The satisfaction with a purchase is strongly influenced by the expectations of shoppers, and when those expectations cannot be fulfilled by the assortment, shoppers are dissatisfied with their purchases (Diehl & Poynor, 2010). Additionally, if they make a decision, they are more likely to regret it after the fact (Iyengar & Lepper, 2000). Since a very large assortment means a large selection, shoppers have high expectations of finding a product that perfectly fits their needs. In reality, however, the perfect product often does not exist. This leads to consumers being disappointed in their expectations. With smaller assortments, on the other hand, customers' expectations are lower, which means they are more satisfied when they find a suitable product (Diehl & Poynor, 2010).

Furthermore, experimental studies by Iyengar and Lepper (2000) reveal that too large assortments can be overwhelming for shoppers because of the many choices available. The study shows that although shoppers generally find a wide range of choices attractive, they also have more difficulty making a decision. According to Iyengar and Lepper, this is partly due to the flood of options associated with the decision, but also due to an increased sense of responsibility associated with the decision. In some cases, this can even lead to people not making a decision at all (Diehl & Poynor, 2010; Iyengar & Lepper, 2000) and thus not buying anything.

A normal supermarket typically has more than 30.000 SKUs in its assortment (Schwartz, 2004). With so much choice, consumers are often uncertain (Dhar, 1997) and have difficulty choosing the right products (Diehl, 2005; Iyengar

& Lepper, 2000). Therefore, retailers need to simplify the search process.

2.3.1. Techniques to optimize the visual processing of the shelves

Humans cannot process all the stimuli they perceive in their environment at the same time (Chandon et al., 2009). This also applies to the visual processing of the assortment in retail stores. The part of the assortment that the shopper perceives is ultimately determined by his or her attention (Streicher et al., 2021). In light of this, and because supermarkets are becoming more and more complex nowadays, consumers need to be selective in how they use their visual attention to process information (Burke & Leykin, 2014). Indeed, shoppers often use different clues to navigate the store and to estimate the size of the variety of products offered in a store. These are, for example, the space occupied by the category, the presence of favorite products (Broniarczyk et al., 1998), or the arrangement and number of repetitions of products (Hoch et al., 1999). As a result, each consumer views the assortment through his or her own eyes and considers only that part of the assortment that is perceived in the decision-making process (Broniarczyk et al., 1998). Thus, the actual assortment is not as important as retailers often think, it's more about the perceived assortment. At the store level, already a familiar and well-organized layout can help consumers navigate the assortment more easily and find the products they need (Park et al., 1989). Furthermore, whether the products are well-organized or unorganized influences how the assortment is perceived by consumers and affects search ease (Hoch et al., 1999). When the customer finally stands in front of a shelf, all barriers to purchase must be minimized so as not to discourage him or her from buying (Burke, 2005). Retailers can use various techniques to facilitate the processing of the assortment.

A first, relatively simple way for retailers to counteract the negative consequences of a too large assortment and to facilitate the search process is to reduce the assortment by eliminating SKUs. Although the method seems simple, in reality, many retailers hesitate to do this because assortment is considered a critical factor in consumer store selection (Broniarczyk et al., 1998). However, research shows that supermarket shoppers make decisions with very low levels of engagement and are not particularly active in seeking alternatives (Dickson & Sawyer, 1990). This led researchers to investigate whether the assortment could be reduced without too serious consequences. Indeed, Broniarczyk, Hoyer, and McAllister (1998) show that retailers can reduce the number of products without negatively affecting the shopper's perception of the assortment. They argue in their paper that even with a reduction in assortment, profit can be increased under certain circumstances. However, for this strategy to work, two requirements must be met. First, consumers' favorite products must continue to be available, and second, the space of the product category must remain constant (Broniarczyk et al., 1998).

A second method to support the visual processing of the assortment is the choice of a suitable presentation method. It makes a difference whether the assortment is presented visually, e.g., with pictures, or described with text. This strategy is perhaps more suitable for retailers with online stores, but it's also fundamentally applicable to stationary retail. In general, consumers prefer a visual representation of the assortment in the form of images, because it allows them to scan the assortment faster (Townsend & Kahn, 2014). While humans have to process text step by step, images can be processed as a whole (Hart, 1997). Therefore, visual information can be grasped much faster (Townsend & Kahn, 2014). Because the visual presentation of assortments is easier for consumers to process, they also find it more enjoyable. This preference is what Townsend and Kahn (2014) call the "visual preference heuristic" in their paper. However, this doesn't mean that a visual presentation is always better. On the contrary, it depends on the size of the assortment. When the assortment is large, a visual presentation can be overwhelming and increase complexity for customers, and a presentation via text may be better (Townsend & Kahn, 2014).

Finally, as shoppers prefer to browse the assortment visually (Townsend & Kahn, 2014), it's important for retailers to pick up consumers on this level as well (Deng et al., 2016). In retailing, in-store displays are often used for this purpose. Not only can they stimulate sales (Roggeveen et al., 2016), but they can also simplify visual processing. As Deng and her colleagues (2016) have found, horizontal displays in particular seem to make it easier for shoppers to visually process the assortment. As they report, because of their horizontal field of view, humans can process horizontal displays faster and easier, and thus process the assortment more efficiently. The results of their study further show that this can even lead to a larger selection in the product category.

2.3.2. Promotion techniques to make products visually salient

The human brain can only deal with a limited number of visual impressions at the same time (Clement et al., 2013). In the same way, the visual attention of consumers in the supermarket is also limited. As supermarkets become more cluttered, marketers need to make sure their products are visible while shopping (Chandon et al., 2007). Visual saliency plays an important role here. The visual salience of a stimulus can be described as a feature that stands out and attracts attention (McArthur & Post, 1977). In a retail context, visual salience can be seen as "the likelihood that it will attract in-store attention" (Chandon et al., 2007, p. 228). When a brand attracts visual attention, it not only has benefits for that individual brand but also has a positive impact on other brands in the assortment, as it can trigger a memory-based consideration for other brands (Hutchinson et al., 1994).

According to prior research, the position of products affects consumers' attention and preferences (Valenzuela & Raghubir, 2009). This implies that not all positions receive the same amount of attention from consumers. Some positions receive more attention and others less. For the amount of attention, a product receives from consumers, especially the horizontal and vertical positioning on a shelf plays a major role. In general, it has been found that products positioned in the horizontal center of a shelf get more attention because shoppers are inclined to look there (Atalay et al., 2012). This behavior also influences consumers' purchase decisions. As Chandon, Hutchinson, Bradlow, and Young (2009) find in an eye-tracking study, brands positioned at the top or near the center of the shelf receive more attention and are evaluated better compared to products in other positions. As Valenzuela and Raghubir (2009) report this is also due to the consumers' belief that retailers place the most popular products in the center. Regarding the vertical position of a product, Chen, Burke, Hui, and Leykin (2021) find in an eye-tracking study that the optimal position for products is not at eye level as many retailers believe, rather it is about 14.7 inches (37.34 centimeters) lower. This means the optimal vertical position of a product is approximately at the height of the consumer's chest. Additionally, as further results of the study show shoppers' in-store attention is also influenced by a lateral bias. When shoppers cross an aisle, they are 21% more likely to notice products on the right (Chen et al., 2021).

Another method, that can be applied to make products visually more salient at the POP is imaginative displays. An imaginative display can be defined "as a product display constructed using multiple units of the same product in a novel yet aesthetically appealing form" (Keh et al., 2021, p. 111). Since these displays are significantly different from standard displays, they are more novel to consumers (Keh et al., 2021; Mugge & Schoormans, 2012). This makes the products stand out more from others and attract more attention from consumers (Raghubir & Greenleaf, 2006). Keh, Wang, and Yan (2021) find in their study, that imaginative displays can influence consumer buying behavior and increase sales when the shape of the display matches the benefit of the product to the consumer. However, they also report that retailers should use this method cautiously because if these two factors don't match, then it can negatively impact shopper behavior.

After all, not only the design of the product presentation but also the design of the products themselves can make them visually salient. As Clement, Kristensen, and Grønhaug (2013) report, product design features can capture consumer attention in two ways: either through physical features such as a unique shape and high contrast or by making the packaging as simple as possible to facilitate visual processing. In addition, Van der Lans, Pieters, and Wedel (2008) report that factors such as the brightness and color of a product package are important factors for brand search efficiency. Indeed, a study by Burke and Leykin (2014) shows that unique packaging can reduce product search times by up to 40%.

2.3.3. Shelf-Talker-Flags

A further promotional tool retailers can use to visually highlight a product at the point of purchase is shelf talker flags (STFs). Shelf talker flags, sometimes also called "Wobblers", belong to the category of nonmonetary promotions, and are small flags, which are attached to the shelf. By equipping products with shelf talker flags, they become visually more salient at the point of purchase. Moreover, promotional signals can facilitate consumers' in-store search process as products become visually more prominent and therefore easier to discover (Chandon et al., 2000). This effect has already been confirmed by prior research (Dickson & Sawyer, 1990; Inman et al., 1990). Thus, shelf talker flags should facilitate the search process by making products visually more prominent, which in turn should have a positive impact on subjectively perceived search ease. More formally:

> **H1**: Shelf talker flags, compared to a situation without shelf talker flags, increase subjectively perceived search ease.

By visually highlighting products, shelf talker flags could draw consumers' attention to more potentially interesting products. Since visual attention is an important factor in consumer decisions in stores (Chandon et al., 2007), the increased visual attention provided by shelf talker flags could lead to more purchases. Additionally, advertising signals can facilitate the purchase decision because they usually give consumers a reason why a product should be purchased (Chandon et al., 2000). For instance, the purchase quantity is often dictated by promotions, which further simplifies the purchase decision for the consumer (Wansink et al., 1998). An example of this would be product offers where a price reduction is only applied when two or more products are purchased. Generally, consumer decisions in grocery stores tend to be characterized by low engagement (Dickson & Sawyer, 1990). Moreover, Chaiken and Maheswaran (1994) report that when people are poorly motivated or unable to evaluate a product, they often use a "consensus heuristic" to shape their attitudes. Simply put, this means that shoppers rely on the opinions of others and assume that if many people like a product, it must be good (Valenzuela & Raghubir, 2009). If shelf talker flags now label a product as "Bestseller", consumers will become aware of popular products, which in turn could trigger the consensus heuristic, especially when consumers have little knowledge about a product category. Thus, this could also influence consumers' purchase behavior. This leads to the second hypothesis:

> **H2:** Shelf talker flags, compared to a situation without shelf talker flags, increase consumers' purchases of marked products.

Next, the study that was conducted to test the hypotheses and thus the effect of shelf talker flags on consumer shopping behavior is described.

3. Empirical part

3.1. Study Description

The study was conducted as a field experiment in cooperation with MPreis, an Austrian supermarket chain, in one of their stores in Innsbruck. Furthermore, it should be mentioned that the study was conducted as part of a university course together with four other fellow students. The study tests the effect of shelf talker flags on subjectively perceived search ease and purchases of marked products using a one-factor between-subjects design with two levels (shelf talker flags: with vs. without). It was conducted over a two-week period (Week 1 = without STFs, Week 2 = with STFs).

The research model (cf. *Figure 2*) describes the direct effect of the independent variable *shelf talker flags* on the two dependent variables: *subjectively perceived search ease* and *purchases of marked products*. First is tested whether manipulating the visual saliency of products by equipping them with shelf talker flags increases subjectively perceived search ease (H1). Then the effect of shelf talker flags on the purchases of marked products is tested (H2). For data collection, questionnaire-based interviews with shoppers were conducted in the store.

3.2. Field setting

Data collection took place in two weeks at the end of May respectively beginning of June 2022 with one week in between. The reason was that there were public holidays that might have otherwise skewed the study. The first week covered the period from 17.05 - 19.05 and the second week the period from 30.05 - 02.06. In both weeks, the survey took place, on Tuesday, Wednesday, and Thursday from 10 am to 5 pm. Random assignment of conditions was not possible due to the logistical effort associated with the field setting, so one condition (with vs. without STFs) was run per week. Therefore, the experiment can be considered a quasi-experiment as the conditions were not completely randomized. In the first week, the condition without shelf talker flags was carried out, and in the second week the condition with shelf talker flags. In order to make the shoppers familiar with the shelf talker flags a little bit, these were already attached after the first week of the experiment. After the end of the experiment period, all shelf talker flags were removed again. For data collection, two researchers were positioned after the checkout and near the exit of the store to intercept and interview shoppers after their purchases.

3.3. Sample

There were no predefined criteria for participants; every shopper who made a purchase in the store was approached to participate in the study. It is therefore a convenience sample. Across the two weeks of the experiment, 444 shoppers (65% female, $M_{age} = 51$, $SD_{age} = 21$) participated in the study. In the first week, 244 (54.95%) participated in the study, compared with 200 shoppers (45.05%) in the second week. The lower number of participants in the second week may be explained by the fact that some had already participated in the first week. Of the 444 participants, 26 (5.86%) were minors. These were excluded from the study due to the lack of representativeness of the overall sample, resulting in a final sample of 418 participants. All shoppers were interviewed only once and received no compensation, financial or otherwise, for their participation in the study.

3.4. Manipulation

The management of MPreis provided us with a list of 33 best-selling products from various different product categories. Then, shelf talker flags were designed to manipulate the visual salience of the selected products. The shelf talker flags were designed as 8cm × 5.5cm (3.15 × 2.17”) red rectangular signs and have "Bestseller" as the inscription. An example of the shelf talker flag design is shown in *Figure 3*. In the first week of the experiment, the condition without STFs, the status quo of the store was maintained, so nothing was changed in the store. In the second week, the condition with STFs, visual salience of the selected products was manipulated by placing shelf talker flags next to the products on the shelf. An example is shown in *Figure 4*. The shelf talker flags were not placed in any particular area of the store, rather they were spread throughout the supermarket. All 33 best-selling products were marked with shelf talker flags only once, with one exception: the 250ml can of Red Bull, which was marked twice in the store. So, in total, 34 shelf talker flags were placed in the store.

To ensure that all shelf talker flags were correctly positioned and still hanging next to the products, regular tours through the store were made. This happened at least twice a day, and if necessary, damaged, or lost shelf talker flags were replaced with new ones.

3.5. Main procedure

Data collection was conducted through a questionnaire-based interview with shoppers. Since the experiment was conducted in Austria, the interviews were conducted in German. Two researchers were located after the checkout and near the exit of the store. All shoppers who purchased something in-store were intercepted and asked to participate in the study. The questionnaire was constructed from previous literature and contained statements about *shelf attractiveness* and *search ease* as well as a control question regarding a shopping list. The statements were operationalized on a seven-point Likert-Scale ranging from -3 (strongly disagree) to +3 (strongly agree). In addition, there was a list of all 33 best-selling products on the back of the questionnaire. The entire questionnaire can be found in the *Appendix*. If consumers had purchased one of the best-selling products, the researchers noted this along with the number of products purchased. Other measures were expenses, age, and gender. *Table 1* shows the measures and reliability scores.

4. Results

4.1. Controls

There was no statistically significant difference between the two conditions for the two control variables gender ($p = .22$) and shelf attractiveness ($p = .99$). However, the variable

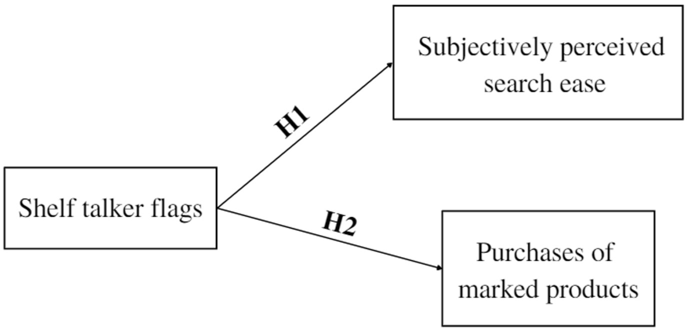

Subjectively perceived search ease
H1
Shelf talker flags
H2
Purchases of marked products

Figure 2: Research Design

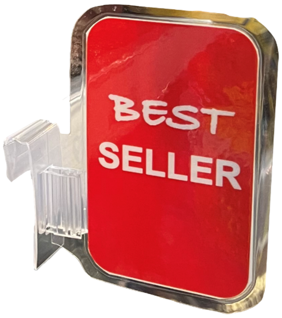


Figure 3: Shelf talker flag design

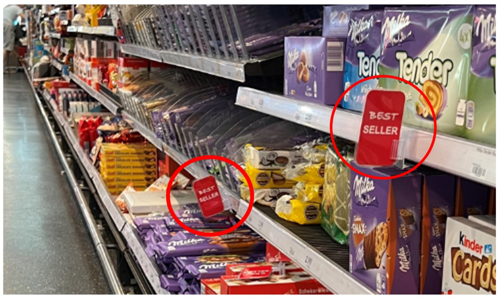


Figure 4: Exemplary positioning of shelf talker flags in a product category

Table 1: Measures and reliability scores of survey items

Variable	Items (-3 = strongly disagree; +3 = strongly agree)	Source
Shelf Attractiveness (α = .77)	Die Regale im Gang wirken gut organisiert!	Adapted from: Sevilla and Townsend (2016)
	Die Warenpräsentation in den Regalen ist optisch ansprechend!	
	Die Warenpräsentation ist visuell leicht zu verarbeiten!	
Search Ease (α = .79)	Es ist leicht, sich einen Überblick über das Produktangebot zu verschaffen!	
	Produkte, welche regelmäßig benötigt werden, sind leicht zu finden!	
	Die wichtigsten Produkte sind leicht zu entdecken!	
	Rabatt- oder Werbeschilder am Regal haben mich auf interessante Produkte aufmerksam gemacht!	Adapted from: Hilken et al. (2017)
	Rabatt- oder Werbeschilder am Regal waren nützlich für meinen Einkauf!	
	Rabatt- oder Werbeschilder am Regal haben mir bei der einen oder anderen Kaufentscheidung geholfen!	

expenses per shopper were significantly higher in the condition with STFs (M = €21.69) than in the condition without STFs (M = €17.89, $p < .05$). Also, the mean age of participants was significantly higher in the condition with STFs (M = 51 years) than in the condition without STFs (M = 47 years). Furthermore, age correlated positively with the expenses of the shoppers ($r = .17$). Thus, I included age as a covariate in the following analysis. However, the significance pattern did not differ when age was included as a covariate. Therefore, the covariate age is not discussed further.

4.2. Subjectively perceived search ease

An independent t-test with search ease as the dependent factor and the presence of shelf talker flags (i.e., yes, or no) as the independent factor was used to test the effect on subjectively perceived search ease. Results show that the shopper-reported search ease was significantly higher ($M = 1.35$) in the condition with STFs than in the condition without STFs [$M = 1.11$, $t(416) = 2.09$), $p < .05$]. Thus, hypothesis 1 is confirmed.

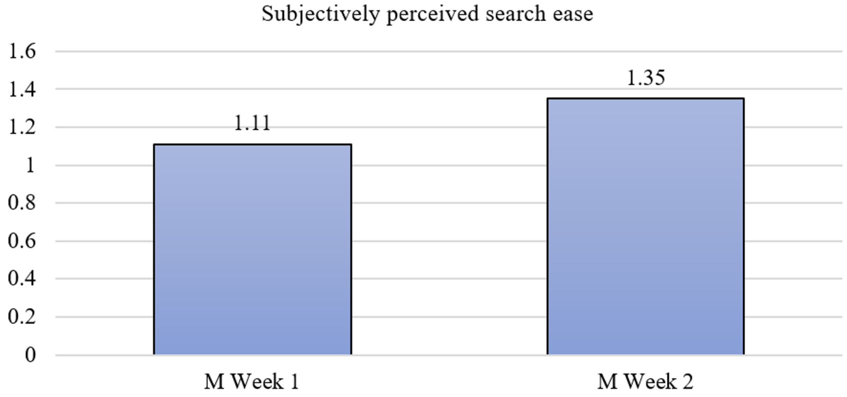


Figure 5: Subjectively perceived search ease

4.3. Purchases of marked products

An independent t-test with the purchases of marked products as the dependent factor and the presence of shelf talker flags (i.e., yes, or no) as the independent factor was used to test the effect on the purchases of marked products. Results show that shoppers on average purchased marginally more of the marked products in the condition with STFs ($M = .12$) compared to the condition without STFs [$M = .06$, $t(416) = 1.76$), $p = .08$]. This partially supports hypothesis 2.

5. Discussion

5.1. General discussion

This paper investigated the effect of shelf talker flags on consumer shopping behavior. First, I proposed that shelf talker flags positively affect the subjectively perceived search ease of consumers. Shelf talker flags make products visually more salient at the point of purchase. This simplifies the shopper's search process in the store as products can be discovered more easily in the assortment. Therefore, the subjectively perceived search ease of consumers increases. Second, it was hypothesized that shelf talker flags increase purchases of marked products. Shelf talker flags can increase customers' awareness sets of interesting products, by visually highlighting products. If consumers become aware of more products, this can increase sales (Deng et al., 2016) of marked products.

To investigate these two effects, I conducted a field experiment over a period of two weeks, where I manipulated the product presentation of 33 bestseller products by using shelf talker flags with the label "Bestseller". As assumed, in the condition with STFs, subjectively perceived search ease was evaluated significantly higher (H1). So, as our results show, shelf talker flags can significantly facilitate the search process of consumers in the store, which is becoming more and more important as retail stores offer large assortments and consumers are facing decision difficulties (Iyengar & Lepper, 2000). Furthermore, in the condition with STFs, I found an increase in the purchases of marked products (H2). However, this result is only marginally significant and should therefore be viewed with caution.

5.2. Theoretical contributions

The research on visual in-store marketing is constantly growing, as marketers and practitioners alike are highly interested in the factors that drive (unplanned) purchases. Prior literature has identified numerous techniques retailers can use to increase unplanned purchases, ranging from store-wide methods like travel distance (e.g., Granbois, 1968; Hui et al., 2009, 2013), over shelf management strategies (e.g., Campo and Gijsbrechts, 2005; Chandon et al., 2009; Chen et al., 2021; Drèze et al., 1994) to promotional signals at the point of purchase (e.g., Ailawadi et al., 2009; Chandon et al., 2000, 2007; Woodside and Waddle, 1975). Displays and their impact on consumer behavior have been studied frequently in research (Campo & Gijsbrechts, 2005). In contrast, shelf talker flags and their effects have hardly been studied so far. This paper adds to the existing literature, by investigating the effect of shelf talker flags on consumer shopping behavior.

As Chandon, Wansink, and Laurent (2000) argue, promotional signals can facilitate the search process of customers in the store and make shopping more convenient. While prior research has shown that this is indeed the case (e.g., Dickson and Sawyer, 1990; Inman et al., 1990), I extended the literature by showing that even a simple marker like a shelf talker flag can significantly facilitate the search process for shoppers in the store. In addition, I investigated whether or not shelf talker flags can increase purchases of marked products. By labeling the shelf talker flags with "Bestseller" I also incorporated the theory of the consensus heuristic (Chaiken & Maheswaran, 1994). However, I did not especially test for the consensus heuristic and our result on the purchases of marked products is only marginally significant.

5.3. Practical implications

Nowadays, retail stores are oftentimes complex and contain several thousand SKUs in their assortment (Schwartz, 2004). This makes it increasingly difficult for customers to find products that meet their needs and increases the difficulty of making decisions (Diehl & Poynor, 2010; Iyengar & Lepper, 2000). Therefore, retailers need to simplify

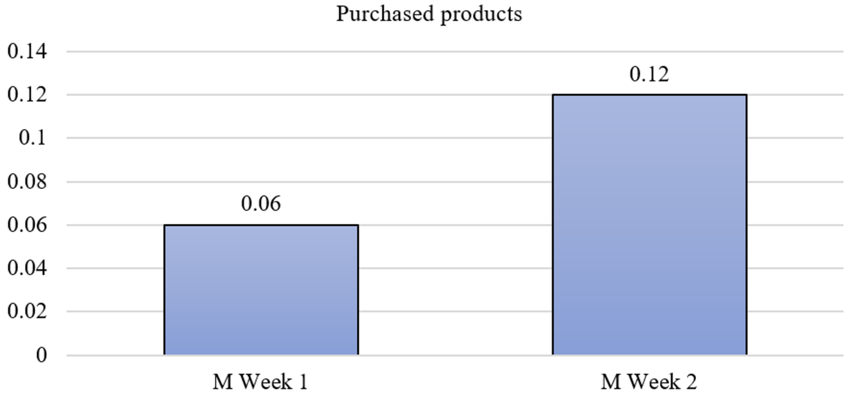


Figure 6: Purchases of marked products

the search process. Considering that a typical supermarket contains about 30.000 SKUs or more (Schwartz, 2004), I equipped only about 0.11% of them (33 products) with shelf talker flags. Yet I found a significant increase in the subjectively perceived search ease of consumers. Shelf talker flags are therefore a simple, but powerful method retailers can use to visually highlight products at the point of purchase to subsequently facilitate the search process of shoppers.

In addition, retailers often use monetary promotions (e.g., discounts or rebates) to attract shoppers to the store (Grewal et al., 1998) and to support purchases. However, these are costly and decrease profits for retailers (Jedidi et al., 1999). Shelf talker flags, in contrast, are cheap and easy to use for retailers and our results indicate that they can increase purchases of marked products. While I added shelf talker flags to best-selling products in our experiment, this may not be ideal for real-world use. In reality, other products probably make more sense since best-selling products sell well anyway. The additional effect of shelf talker flags on sales is therefore likely to be lower for best-selling products than for less popular products. Therefore, shelf talker flags could be used to show shoppers alternatives to popular products. For example, Tirola Kola could be highlighted as an alternative to Coca-Cola, which was one of the best-selling products in our experiment. However, I believe that the products should not be completely unknown and must be a good alternative. Therefore, the second best-selling product would be ideal for this purpose.

5.4. Limitations and Future Research

Our work has some important limitations and offers interesting possibilities for future research. First, our result on the effect of shelf talker flags on the purchases of marked products is only marginally significant. Therefore, a positive effect could not be completely proven. However, a possible explanation for the lack of significance could lie in the process of data collection. The data measuring the purchases of marked products is based solely on a questionnaire-based survey of in-store shoppers. Although the researchers conscientiously tried to collect all data accurately, not all shoppers agreed to let us look at their receipts or in their baskets. As a result, I had difficulty collecting the data accurately, which may have resulted in our data not being particularly precise. Nonetheless, our result suggests that there may be a positive relationship between shelf talker flags and purchases. Further research should measure this relationship more precisely, e.g., using a full sample of sales during the experimental period.

Second, the study was conducted as a field experiment. On the one hand, this is beneficial because the shelf talker flags could be tested in a natural, real-world environment and provide real shopper data. On the other hand, however, there are a lot of confounding variables that I could not exclude and that might have influenced our results. Situational factors such as time restrictions or the number of other customers in the store strongly affect in-store purchase decisions (Beatty & Smith, 1987; Hui et al., 2009; Park et al., 1989; Zhang et al., 2014) and could not be excluded. This may also explain the differences in the control variables in the two survey weeks. In addition, because I used real best-selling products in the experiment, I had the problem that some of the products I marked were almost always sold out. Also, it was not always possible for the store employees to restock the products immediately. Finally, the experimental conditions could not be completely randomized, making it a quasi-experiment. All of this could have potentially biased our study. To obtain more reliable data and to verify our results, a laboratory experiment would be useful.

Finally, with 34 shelf talker flags, I equipped only a small number of all available SKUs in the store. It would be interesting to investigate the effects of a larger number of STFs. More shelf talker flags could show an even greater effect on search ease. Conversely, it could also be that precisely because very few products were marked with shelf-talker flags, this caused them to stand out from the mass of products. However, Ibelieve there may be a tipping point where it could become a visual distraction for consumers. That, in turn, could negatively impact consumer shopping behavior. Thus, future research could investigate the impact of a larger number of shelf talker flags on consumer shopping behavior.

References

Ailawadi, K. L., Beauchamp, J., Donthu, N., Gauri, D. K., & Shankar, V. (2009). Communication and Promotion Decisions in Retailing: A Review and Directions for Future Research. *Journal of Retailing*, *85*(1), 42–55. https://doi.org/10.1016/j.jretai.2008.11.002

Atalay, A. S., Bodur, H. O., & Rasolofoarison, D. (2012). Shining in the Center: Central Gaze Cascade Effect on Product Choice. *Journal of Consumer Research*, *39*(4), 848–866. https://doi.org/10.1086/665984

Babin, B. J., Darden, W. R., & Griffin, M. (1994). Work and/or Fun: Measuring Hedonic and Utilitarian Shopping Value. *Journal of Consumer Research*, *20*(4), 644–656. https://doi.org/10.1086/209376

Baumgartner, H., & Steenkamp, J. B. E. (1996). Exploratory consumer buying behavior: Conceptualization and measurement. *International Journal of Research in Marketing*, *13*(2), 121–137. https://doi.org/10.1016/0167-8116(95)00037-2

Bayley, G., & Nancarrow, C. (1998). Impulse purchasing: a qualitative exploration of the phenomenon. *Qualitative Market Research: An International Journal*, *1*(2), 99–114. https://doi.org/10.1108/13522759810214271

Beatty, S. E., & Ferrell, M. E. (1998). Impulse buying: Modeling its precursors. *Journal of Retailing*, *74*(2), 169–191. https://doi.org/10.1016/s0022-4359(99)80092-x

Beatty, S. E., & Smith, S. M. (1987). External Search Effort: An Investigation Across Several Product Categories. *Journal of Consumer Research*, *14*(1), 83–95. https://doi.org/10.1086/209095

Bell, D. R., Corsten, D., & Knox, G. A. H. (2011). From point-of-purchase to path-to-purchase: How pre-shopping factors drive unplanned buying. *Journal of Marketing*, *75*(1), 31–45. https://doi.org/10.1509/jmkg.75.1.31

Blattberg, R. C., & Neslin, S. A. (1993). Chapter 12 Sales promotion models. *Handbooks in Operations Research and Management Science*, 553–609. https://doi.org/10.1016/s0927-0507(05)80035-0

Block, L. G., & Morwitz, V. G. (1999). Shopping Lists as an External Memory Aid for Grocery Shopping: Influences on List Writing and List Fulfillment. *Journal of Consumer Psychology*, *8*(4), 343–75.

Broniarczyk, S. M., Hoyer, W. D., & McAlister, L. (1998). Consumers' Perceptions of the Assortment Offered in a Grocery Category: The Impact of Item Reduction. *Journal of Marketing Research*, *35*(2), 166–176. https://doi.org/10.2307/3151845

Buchanan, L., Simmons, C. J., & Bickart, B. A. (1999). Brand equity dilution: Retailer display and context brand effects. *Journal of Marketing Research*, *36*(3), 345–355. https://doi.org/10.2307/3152081

Bucklin, R. E., & Lattin, J. M. (1991). A Two-State Model of Purchase Incidence and Brand Choice. *Marketing Science*, *10*(1), 24–39. https://doi.org/10.1287/mksc.10.1.24

Burke, R. R. (2005). Retail shoppability: A measure of the world's best stores. *Future retail now: 40 of the world's best stores*, 206–219.

Burke, R. R., & Leykin, A. (2014). Identifying the Drivers of Shopper Attention, Engagement, and Purchase. In *Review of Marketing Research* (pp. 147–187, Vol. 11). https://doi.org/10.1108/S1548-643520140000011006

Büttner, O. B., Florack, A., Leder, H., Paul, M. A., Serfas, B. G., & Schulz, A. M. (2014). Hard to Ignore: Impulsive Buyers Show an Attentional Bias in Shopping Situations. *Social Psychological and Personality Science*, *5*(3), 343–351. https://doi.org/10.1177/1948550613494024

Campo, K., & Gijsbrechts, E. (2005). Retail Assortment, Shelf and Stockout Management: Issues, Interplay and Future Challenges. *Applied Stochastic Models in Business and Industry*, *21*(4–5), 383–392. https://doi.org/10.1002/asmb.574

Chaiken, S., & Maheswaran, D. (1994). Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgment. *Journal of Personality and Social Psychology*, *66*(3), 460–473. https://doi.org/10.1037/0022-3514.66.3.460

Chandon, P., Hutchinson, J. W., Bradlow, E., & Young, S. H. (2007). Measuring the Value of Point-of-Purchase Marketing with Commercial Eye-Tracking Data. *INSEAD Working Papers Collection*, 225–258. https://doi.org/10.2139/ssrn.1032162

Chandon, P., Hutchinson, J. W., Bradlow, E. T., & Young, S. H. (2009). Does In-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase. *Journal of Marketing*, *73*(6), 1–17. https://doi.org/10.1509/jmkg.73.6.1

Chandon, P., Wansink, B., & Laurent, G. (2000). A Benefit Congruency Framework of Sales Promotion Effectiveness. *Journal of Marketing*, *64*(4), 65–81. https://doi.org/10.1509/jmkg.64.4.65.18071

Chen, M., Burke, R. R., Hui, S. K., & Leykin, A. (2021). Understanding Lateral and Vertical Biases in Consumer Attention: An In-Store Ambulatory Eye-Tracking Study. *Journal of Marketing Research*, *58*(6), 1120–1141. https://doi.org/10.1177/0022243721998375

Clement, J., Kristensen, T., & Grønhaug, K. (2013). Understanding consumers' in-store visual perception: The influence of package design features on visual attention. *Journal of Retailing and Consumer Services*, *20*(2), 234–239. https://doi.org/10.1016/j.jretconser.2013.01.003

Deng, X., Kahn, B. E., Unnava, H. R., & Lee, H. (2016). A "Wide" Variety: Effects of Horizontal versus Vertical Display on Assortment Processing, Perceived Variety, and Choice. *Journal of Marketing Research*, *53*(5), 682–698. https://doi.org/10.1509/jmr.13.0151

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

Dhar, R. (1997). Consumer Preference for a No-Choice Option. *Journal of Consumer Research*, *24*(2), 215–231. https://doi.org/10.1086/209506

Dickson, P. R., & Sawyer, A. G. (1990). The Price Knowledge and Search of Supermarket Shoppers. *Journal of Marketing*, *54*(3), 42–53. https://doi.org/10.1177/002224299005400304

Diehl, K. (2005). When Two Rights Make a Wrong: Searching Too Much in Ordered Environments. *Journal of Marketing Research*, *42*(3), 313–322. https://doi.org/10.1509/jmkr.2005.42.3.313

Diehl, K., & Poynor, C. (2010). Great Expectations?! Assortment Size, Expectations, and Satisfaction. *Journal of Marketing Research*, *47*(2), 312–322. https://doi.org/10.1509/jmkr.47.2.312

Dodson, J. A., Tybout, A. M., & Sternthal, B. (1978). Impact of Deals and Deal Retraction on Brand Switching. *Journal of Marketing Research*, *15*(1), 72–81. https://doi.org/10.1177/002224377801500109

Doob, A. N., Carlsmith, J. M., Freedman, J. L., Landauer, T. K., & Tom, J., S. (1969). Effect of initial selling price on subsequent sales. *Journal of Personality and Social Psychology*, *11*(4), 345–350. https://doi.org/10.1037/h0027415

Drèze, X., Hoch, S. J., & Purk, M. E. (1994). Shelf management and space elasticity. *Journal of Retailing*, *70*(4), 301–326. https://doi.org/10.1016/0022-4359(94)90002-7

Faber, R. J., & Vohs, K. D. (2011). Self-regulation and spending: Evidence from impulsive and compulsive buying. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of Self-regulation: research, theory, and applications, 2nd Edition* (pp. 537–551). Guilford Press.

Field, M., & Eastwood, B. (2005). Experimental manipulation of attentional bias increases the motivation to drink alcohol. *Psychopharmacology*, *183*(3), 350–357. https://doi.org/10.1007/s00213-005-0202-5

Fitzsimons, G. J. (2000). Consumer Response to Stockouts. *Journal of Consumer Research*, *27*(2), 249–266. https://doi.org/10.1086/314323

Friedman, R. S., Fishbach, A., Förster, J., & Werth, L. (2003). Attentional Priming Effects on Creativity. *Creativity Research Journal*, *15*(2–3), 277–286. https://doi.org/10.1080/10400419.2003.9651420

Gardner, M., & Rook, D. W. (1988). Effects of Impulse Purchases on Consumers' Affective States. *ACR North American Advances*.

Gilbride, T. J., Inman, J. J., & Stilley, K. M. (2015). The Role of Within-Trip Dynamics in Unplanned versus Planned Purchase Behavior. *Journal of Marketing*, *79*(3), 57–73. https://doi.org/10.1509/jm.13.0286

Granbois, D. H. (1968). Improving the Study of Customer In-store Behavior. *Journal of Marketing*, *32*(4), 28–33. https://doi.org/10.1177/002224296803200406

Grewal, D., Monroe, K. B., & Krishnan, R. (1998). The Effects of Price-Comparison Advertising on Buyers' Perceptions of Acquisition Value, Transaction Value, and Behavioral Intentions. *Journal of Marketing*, *62*(2), 46–59. https://doi.org/10.1177/002224299806200204

Hart, R. P. (1997). Analyzing Media. In R. P. Hart (Ed.), *Modern Rhetorical Criticism* (2nd, pp. 177–208). Allyn & Bacon.

Heath, R., & Feldwick, P. (2008). Fifty Years Using the Wrong Model of Advertising. *International Journal of Market Research*, *50*(1), 29–59. https://doi.org/10.1177/147078530805000105

Henderson, J. M., & Hollingworth, A. (1999). High-Level Scene Perception. *Annual Review of Psychology*, *50*(1), 243–271. https://doi.org/10.1146/annurev.psych.50.1.243

Hilken, T., de Ruyter, K., Chylinski, M., Mahr, D., & Keeling, D. I. (2017). Augmenting the eye of the beholder: exploring the strategic potential of augmented reality to enhance online service experiences. *Journal of the Academy of Marketing Science*, *45*(6), 884–905. https://doi.org/10.1007/s11747-017-0541-x

Hoch, S. J., Bradlow, E. T., & Wansink, B. (1999). The Variety of an Assortment. *Marketing Science*, *18*(4), 527–546. https://doi.org/10.1287/mksc.18.4.527

Hoch, S. J., & Loewenstein, G. F. (1991). Time-Inconsistent Preferences and Consumer Self-Control. *Journal of Consumer Research*, *17*(4), 492–507. https://doi.org/10.1086/208573

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787–795. https://doi.org/10.3758/bf03206794

Hui, S. K., & Bradlow, E. T. (2012). Bayesian multi-resolution spatial analysis with applications to marketing. *Quantitative Marketing and Economics*, *10*(4), 419–452. https://doi.org/10.1007/s11129-012-9122-y

Hui, S. K., Bradlow, E. T., & Fader, P. S. (2009). Testing Behavioral Hypotheses Using an Integrated Model of Grocery Store Shopping Path and Purchase Behavior. *Journal of Consumer Research*, *36*(3), 478–493. https://doi.org/10.1086/599046

Hui, S. K., Inman, J. J., Huang, Y., & Suher, J. (2013). The Effect of In-Store Travel Distance on Unplanned Spending: Applications to Mobile Promotion Strategies. *Journal of Marketing*, *77*(2), 1–16. https://doi.org/10.1509/jm.11.0436

Hutchinson, J. W., Raman, K., & Mantrala, M. K. (1994). Finding Choice Alternatives in Memory: Probability Models of Brand Name Recall. *Journal of Marketing Research*, *31*(4), 441–461. https://doi.org/10.2307/3151875

Inman, J. J., McAlister, L., & Hoyer, W. D. (1990). Promotion Signal: Proxy for a Price Cut? *Journal of Consumer Research*, *17*(1), 74–81. https://doi.org/10.1086/208538

Inman, J. J., Winer, R. S., & Ferraro, R. (2009). The Interplay among Category Characteristics, Customer Characteristics, and Customer Activities on in-Store Decision Making. *Journal of Marketing*, *73*(5), 19–29. https://doi.org/10.1509/jmkg.73.5.19

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

Iyer, E. S. (1989). Unplanned purchasing: Knowledge of shopping environment and time pressure. *Journal of Retailing*, *65*(1), 40–57.

Janiszewski, C. (1998). The Influence of Display Characteristics on Visual Exploratory Search Behavior. *Journal of Consumer Research*, *25*(3), 290–301. https://doi.org/10.1086/209540

Jedidi, K., Mela, C. F., & Gupta, S. (1999). Managing Advertising and Promotion for Long-Run Profitability. *Marketing Science*, *18*(1), 1–22. https://doi.org/10.1287/mksc.18.1.1

Kahn, B. E., & Lehmann, D. R. (1991). Modeling Choice Among Assortments. *Journal of Retailing*, *67*(3), 274–299. https://repository.upenn.edu/marketing_papers/241

Kahn, B. E., & Schmittlein, D. C. (1989). Shopping trip behavior: An empirical investigation. *Marketing Letters*, *1*(1), 55–69. https://doi.org/10.1007/bf00436149

Kaltcheva, V. D., & Weitz, B. A. (2006). When Should a Retailer Create an Exciting Store Environment? *Journal of Marketing*, *70*(1), 107–118. https://doi.org/10.1509/jmkg.70.1.107.qxd

Keh, H. T., Wang, D., & Yan, L. (2021). Gimmicky or Effective? The Effects of Imaginative Displays on Customers' Purchase Behavior. *Journal of Marketing*, *85*(5), 109–127. https://doi.org/10.1177/0022242921997359

Kollat, D. T., & Willett, R. P. (1967). Customer Impulse Purchasing Behavior. *Journal of Marketing Research*, *4*(1), 21–31. https://doi.org/10.1177/002224376700400102

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, *35*(13), 1897–1916. https://doi.org/10.1016/0042-6989(94)00279-u

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing*, *22*(4), 395–414. https://doi.org/10.1016/j.ijresmar.2005.09.005

Levy, M., & Weitz, B. A. (1995). *Retailing Management* (2nd). Richard D. Irwin.

Li, J., & Yu, H. (2013). An Innovative Marketing Model Based on AIDA: A Case from E-bank Campus-marketing by China Construction Bank. *iBusiness*, *05*(03), 47–51. https://doi.org/10.4236/ib.2013.53b010

Lohse, G. L. (1997). Consumer Eye Movement Patterns on Yellow Pages Advertising. *Journal of Advertising*, *26*(1), 61–73. https://doi.org/10.1080/00913367.1997.10673518

Luo, X. (2005). How Does Shopping With Others Influence Impulsive Purchasing? *Journal of Consumer Psychology*, *15*(4), 288–294. https://doi.org/10.1207/s15327663jcp1504_3

McArthur, L. Z., & Post, D. L. (1977). Figural emphasis and person perception. *Journal of Experimental Social Psychology*, *13*(6), 520–535. https://doi.org/10.1016/0022-1031(77)90051-8

Mela, C. F., Gupta, S., & Lehmann, D. R. (1997). The Long-Term Impact of Promotion and Advertising on Consumer Brand Choice. *Journal of Marketing Research*, *34*(2), 248–261. https://doi.org/10.1177/002224379703400205

Mugge, R., & Schoormans, J. P. (2012). Product design and apparent usability. The influence of novelty in product appearance. *Applied Ergonomics*, *43*(6), 1081–1088. https://doi.org/10.1016/j.apergo.2012.03.009

O'Guinn, T. C., & Faber, R. J. (1989). Compulsive Buying: A Phenomenological Exploration. *Journal of Consumer Research*, *16*(2), 147–157. https://doi.org/10.1086/209204

Oliver, R. L. (1996). *Satisfaction: A Behavioral Perspective on the Consumer*. McGraw-Hill Inc., US.

Oppewal, H., & Koelemeijer, K. (2005). More choice is better: Effects of assortment size and composition on assortment evaluation. *International Journal of Research in Marketing*, *22*(1), 45–60. https://doi.org/10.1016/j.ijresmar.2004.03.002

Orquin, J. L., & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*(1), 190–206. https://doi.org/10.1016/j.actpsy.2013.06.003

Paré, M., & Dorris, M. C. (2012). The role of posterior parietal cortex in the regulation of saccadic eye movements. In *Oxford Handbooks Online* (pp. 258–278). https://doi.org/10.1093/oxfordhb/9780199539789.013.0014

Park, C. W., Iyer, E. S., & Smith, D. C. (1989). The Effects of Situational Factors on In-Store Grocery Shopping Behavior: The Role of Store Environment and Time Available for Shopping. *Journal of Consumer Research*, *15*(4), 422–433. https://doi.org/10.1086/209182

Pieters, R., & Warlop, L. (1999). Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, *16*(1), 1–16. https://doi.org/10.1016/s0167-8116(98)00022-6

POPAI. (1997). Consumer buying habits study.

POPAI. (2014). The 2014 POPAI Mass Merchant Shopper Engagement Study.

Prelec, D., & Loewenstein, G. (1998). The Red and the Black: Mental Accounting of Savings and Debt. *Marketing Science*, *17*(1), 4–28. https://doi.org/10.1287/mksc.17.1.4

Raghubir, P., & Greenleaf, E. A. (2006). Ratios in Proportion: What Should the Shape of the Package Be? *Journal of Marketing*, *70*(2), 95–107. https://doi.org/10.1509/jmkg.70.2.095

Roggeveen, A. L., Nordfält, J., & Grewal, D. (2016). Do Digital Displays Enhance Sales? Role of Retail Format and Message Content. *Journal of Retailing*, *92*(1), 122–131. https://doi.org/10.1016/j.jretai.2015.08.001

Rook, D. W. (1987). The Buying Impulse. *Journal of Consumer Research*, *14*(2), 189. https://doi.org/10.1086/209105

Rook, D. W., & Fisher, R. J. (1995). Normative Influences on Impulsive Buying Behavior. *Journal of Consumer Research*, *22*(3), 305–313. https://doi.org/10.1086/209452

Russo, J. E., & Leclerc, F. (1994). An Eye-Fixation Analysis of Choice Processes for Consumer Nondurables. *Journal of Consumer Research*, *21*(2), 274–290. https://doi.org/10.1086/209397

Schwartz, B. (2004). *The Paradox of Choice: Why More Is Less*. Ecco.

Sevilla, J., & Townsend, C. (2016). The Space-to-Product Ratio Effect: How Interstitial Space Influences Product Aesthetic Appeal, Store Perceptions, and Product Preference. *Journal of Marketing Research*, *53*(5), 665–681. https://doi.org/10.1509/jmr.13.0601

Starrels, M. E. (1994). Husbands' involvement in female gender-typed household chores. *Sex roles*, *31*(7), 473–491.

Stilley, K. M., Inman, J. J., & Wakefield, K. L. (2010). Planning to Make Unplanned Purchases? The Role of In-Store Slack in Budget Deviation. *Journal of Consumer Research*, *37*(2), 264–278. https://doi.org/10.1086/651567

Streicher, M. C., Estes, Z., & Büttner, O. B. (2021). Exploratory Shopping: Attention Affects In-Store Exploration and Unplanned Purchasing. *Journal of Consumer Research*, *48*(1), 51–76. https://doi.org/10.1093/jcr/ucaa054

Titus, P. A., & Everett, P. B. (1995). The Consumer Retail Search Process: A Conceptual Model and Research Agenda. *Journal of the Academy of Marketing Science*, *23*(2), 106–119. https://doi.org/10.1177/0092070395232003

Townsend, C., & Kahn, B. E. (2014). The "Visual Preference Heuristic": The Influence of Visual versus Verbal Depiction on Assortment Processing, Perceived Variety, and Choice Overload. *Journal of Consumer Research*, *40*(5), 993–1015. https://doi.org/10.1086/673521

Valenzuela, A., & Raghubir, P. (2009). Position-based beliefs: The center-stage effect. *Journal of Consumer Psychology*, *19*(2), 185–196. https://doi.org/10.1016/j.jcps.2009.02.011

Van der Lans, R., Pieters, R., & Wedel, M. (2008). Competitive Brand Salience. *Marketing Science*, *27*(5), 922–931. https://doi.org/10.1287/mksc.1070.0327

Verplanken, B., & Herabadi, A. (2001). Individual differences in impulse buying tendency: feeling and no thinking. *European Journal of Personality*, *15*(1_suppl), 71–83. https://doi.org/10.1002/per.423

Wadlinger, H. A., & Isaacowitz, D. M. (2006). Positive mood broadens visual attention to positive stimuli. *Motivation and Emotion*, *30*(1), 87–99. https://doi.org/10.1007/s11031-006-9021-1

Wansink, B., Kent, R. J., & Hoch, S. J. (1998). An Anchoring and Adjustment Model of Purchase Quantity Decisions. *Journal of Marketing Research*, *35*(1), 71–81. https://doi.org/10.2307/3151931

Wästlund, E., Shams, P., & Otterbring, T. (2018). Unsold is unseen . . . or is it? Examining the role of peripheral vision in the consumer choice process using eye-tracking methodology. *Appetite*, *120*, 49–56. https://doi.org/10.1016/j.appet.2017.08.024

Wilkinson, J. B., Mason, J. B., & Paksoy, C. H. (1982). Assessing the Impact of Short-Term Supermarket Strategy Variables. *Journal of Marketing Research*, *19*(1), 72–86. https://doi.org/10.1177/002224378201900107

Woodside, A. G., & Waddle, G. L. (1975). Sales Effects of In-Store Advertising. *Journal of Advertising Research*, *15*(3), 29–33.

Yeung, C. W. M., & Wyer, R. S. (2004). Affect, Appraisal, and Consumer Judgment. *Journal of Consumer Research*, *31*(2), 412–424. https://doi.org/10.1086/422119

Zhang, X., Li, S., Burke, R. R., & Leykin, A. (2014). An Examination of Social Influence on Shopper Behavior Using Video Tracking Data. *Journal of Marketing*, *78*(5), 24–41. https://doi.org/10.1509/jm.12.0106

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Unravelling Collective Action Frames Through a Temporal Lens: A Case Study of an Environmental Movement in Germany

Sebastian Lüpnitz

Dresden University of Technology

Abstract

Organizing collective action in the face of climate change is one of the grand challenges of our time. Social movements and their approach to framing climate change are pivotal, as they are tasked with the role of challenging and redirecting dominant beliefs and narratives. Recent research suggests that time is at the core of framing and sustainability. However, there is scant research at the intersection of social movements and time. This study responds to this gap by examining how the framing of the environmental movement Letzte Generation in Germany constructs temporality. My findings reveal how the movement frames climate change as a catastrophe, representing itself as a fire alarm to create a shared sense of urgency and advocate for a crisis mode. Temporally, the framing constructs a clear chronology between a dominant past and an undesirable future and aims to redirect the focus to the present. As a result, the movement had to actively orchestrate a balance between disruptive strategies aimed at attention and polarization, and alignment strategies to foster resonance and support. By conceptualizing temporality in framing processes my study illustrates the pivotal role of time in research on social movements and framing. Moreover, it contributes to the discourse on time and sustainability by showing how actors emphasize a present-time perspective.

Keywords: climate crisis; polarization; social movements; strategic framing; time and temporality

1. Introduction

> *"Time is no longer on our side. [. . .] We have a choice: collective action or collective suicide."* *(Guterres, 2022)*

In the face of climate change, crafting convincing frames that foster collective action is crucial for stimulating change (Cornelissen & Werner, 2014; Nyberg et al., 2020). Although the need for sustainable development has been a topic of extensive discourse for over three decades (World Commission on Environment and Development, 1987), substantive societal transformation is still far from sufficient and the time window of action for securing a sustainable future is rapidly closing according to the latest IPCC report (Pörtner et al., 2022). In response, scholars within the sustainability discourse have recently shifted their attention towards questions of how to craft and enact desirable futures (Gümüsay & Reinecke, 2022). Yet, research in organization studies revealed how organizations translate sustainability into a business case which has led society to remain trapped in a "business-as-usual" paradigm, ultimately impeding rather than facilitating sustainable development (Wright & Nyberg, 2017). Hence, it is of particular relevance to study actors that aim to achieve societal change by challenging dominant framings, such as social movements (Wright et al., 2018). In fact, it might be the "plurality of future-making practices that contributes to extending the debates on climate change" (Wenzel et al., 2020, p. 1448) and the frames of movement actors that create alternative pathways to (re-)organize society (Munshi et al., 2022).

First, I want to thank Prof. Dr. Blagoy Blagoev for providing me with the freedom to explore my research interests while offering invaluable guidance and support whenever needed. I am also grateful to the whole Chair of Organization for their encouragement and inspiration throughout all phases of my work. Lastly, a special thanks goes to Jonas Angles for sharing the journey of writing our first empirical research paper. His willingness to listen, challenge my ideas, and meticulously proofread has been indispensable.

DOI: https://doi.org/10.5282/jums/v10i2pp349-368

Within the context of collective action, the literature at the intersection of framing and social movements conceptualizes frames as strategic devices that serve as the cornerstone for mobilizing others (Cornelissen & Werner, 2014). As a fundamental task, previous research emphasizes the importance of aligning frames with potential adherents to foster resonance and, consequently, ensure the effectiveness of framing (Benford & Snow, 2000; Snow & Benford, 1988). Studies provided empirical support for the resonance mechanisms among individuals´ preference for frames (Giorgi & Weber, 2015) and how resonating frames facilitate change at an institutional level (Zeng et al., 2019). In sum, resonance is understood as a key determinant for the success of framing activities.

However, this study examines a social movement that appears to deliberately employ non-resonating, polarizing frames. *Letzte Generation* (LG) is a recently formed environmental movement in Germany that is most known for its more disruptive forms of protest, such as road blockades. The movement rapidly caused a major public debate and provoked sharply contrasting reactions to the protests, resulting in a noticeable degree of polarization. Despite the prevailing rejection of LG within society and the predominantly severe criticism directed at its disruptive protests (Statista Research Department, 2023), the movement appears resolute in adhering to its strategy while remaining committed to the imperative of peacefulness. Therefore, the apparent absence of the strategic objective of alignment and resonance contradicts prior literature on framing in social movements.

To unravel framing activities of social movements, I suggest adopting a temporal lens. Framing inherently carries a temporal dimension, as it is rooted in an interpretation of the past, present, and future, intending to challenge and influence dominant temporal beliefs (Emirbayer & Mische, 1998; Nyberg et al., 2020). Particularly in the context of climate change, time has been argued to be the central element of sustainable development (Bansal & DesJardine, 2014). The pivotal temporal challenge of climate change deviates from the need for broader changes in the future whilst simultaneously requiring immediate action in the present (Slawinski & Bansal, 2015). Surprisingly, there is only little research that explicitly studies framing in social movements in relation to time. Moreover, I argue that a temporal perspective underscores the processual dynamics of framing and, therefore, counteracts the outcome-focused research due to the strong emphasis on the strategic aspects of framing in earlier literature (Cornelissen & Werner, 2014). Following this approach, and puzzled by the polarizing framing strategy of the case, I ask the following research questions:

> ***RQ1:*** *How is temporality constructed within the framing of an environmental movement?*
>
> ***RQ2:*** *How does the movement employ its framing strategically?*

To address these questions, I draw on interview, document, and observational data I collected over a period of nine months. Adhering to the phenomenological nature of my research, I employ an inductive approach to data analysis using the thematic analysis method (Braun & Clarke, 2021). My data show that the movement aims to shift the temporal focus to the present by claiming a crisis mode in the face of the climate catastrophe. Therefore, the movement employs two distinct strategic framing processes to achieve its objective: to convey urgency, the movement consistently engages in *pushing* to disrupt the present while simultaneously endeavouring to convince potential supporters by *translating* the climate catastrophe framing. Thereupon, I theorize that collective action frames, particularly in times of perceived or actual crisis, must strike a balance between disruptive strategies deliberately designed to trigger nonresonance and polarization, and alignment strategies aimed at fostering resonance and garnering support.

The contribution of my analysis is two-fold. First, I enhance the literature on framing and social movements by demonstrating the centrality of time in collective action frames and, thereby, illustrating the complexity of (strategic) framing activities. This study shows how the temporal dimension shapes the strategic processes used to employ a frame. Based on the temporal construction of a frame, my findings illustrate how resonance may not always be the primary objective of framing activities. Instead, polarizing frames can be a strategy to disrupt the present and enforce a temporal shift in the debate. Moreover, I show how alignment processes are influenced by the temporal construction of the frame by introducing the process of *translating*.

Second, this study contributes to the literature on time and sustainability by showing how actors value a present-time perspective. While previous literature predominantly argues for organizations to adopt a long-term, future-oriented time perspective to be sustainable, this case represents an example of a movement fighting for a sustainable future and simultaneously claiming a present-time perspective.

2. Theoretical Background

2.1. Unfolding Social Movements through a Framing Perspective

The extensive literature on framing and social movements offers a rich foundation to build on. In contrast to earlier approaches to studying social movements with a strong emphasis on structural aspects, such as resource mobilization theory (e.g., McCarthy and Zald, 1977), the framing perspective provides a theoretical lens to unravel how collective action is socially constructed (Johnston & Oliver, 2000). The very existence of social movements indicates that there are different interpretations – frames – of the same issue, which in this study´s context is climate change and climate action.

2.1.1. Defining Framing and Collective Action Frames

Frames are defined as “schemata of interpretation” (Goffman, 1974, p. 21). Thus, they serve as sensemaking devices that provide an interpretation of *what is going on* by compressing information from the environment. Therefore, frames can be understood as the "principles of organization" (Goffman, 1974, p. 11) or "set of rules" (Gamson, 1975, p. 604) that govern the assignment of meaning and the appropriate type of activity. Framing, in turn, constitutes the active process of defining *what is going on*, thus identifying what frames apply to a given event or situation (Goffman, 1974, p. 21). Consequently, framing signifies the process of constructing and attributing meaning, "an active, processual phenomenon that implies agency and contention at the level of reality construction" (Benford & Snow, 2000, p. 614).

Framing processes occur across all levels of analysis. In their review of the framing literature in management and organizational research, Cornelissen and Werner (2014) outline the various concepts at a micro, meso, and macro level. At a micro level, research investigates cognitive frames and how they shape sensemaking processes of individuals within the context of managerial decision-making in organizations. At a macro level, framing has been studied in institutional contexts to elucidate the processes by which meaning structures become institutionalized as "taken-for-granted realities" and, in turn, how these macro-level structures influence individuals´ interpretations and actions. At a meso level, the concept of framing has been used to examine how meaning is constructed and negotiated within organizations. To study social movements, the meso level is most appropriate as it focuses on how "strategic actors attempt to frame courses of actions and social identities in order to mobilize others to follow suit" (Cornelissen & Werner, 2014, p. 183). Hence, this study conceptualizes social movements as organized groups that aim to foster change by raising awareness and establishing a collective understanding of a problematic situation through framing activities.

With the aim of mobilizing individuals to take action, framing in social movements inherently encompasses a strategic dimension (Cornelissen & Werner, 2014). Snow and Benford (1988) identify three fundamental framing tasks that combined constitute the strategic facet: *diagnostic*, *prognostic*, and *motivational* framing. Diagnostic framing refers to the articulation of a problematic situation and the justification of why it is problematic in order to establish a consensus on the necessity of change. This task includes identifying the source of the problem by attributing blame and responsibility (Benford & Snow, 2000). Building upon the diagnosis of the situation, prognostic framing presents a proposed solution to address the problem. Lastly, motivational framing aims to provide individuals with a compelling reasoning for engaging in collective action. This encompasses providing a rationale through language that stirs motivation (Benford & Snow, 2000).

In summary, there are two key facets of framing within the context of collective action: framing as sensemaking, an ongoing interpretative process of meaning construction, and framing as a strategic tool for social movements to mobilize support and foster change. Consequently, frames in social movements have been referred to as *collective action frames* (Snow & Benford, 1988, p. 198) to emphasize the strong agentic nature of those frames in addition to their interpretative function, as they are "calling for action that problematizes and challenges existing authoritative views and framings of reality".

2.1.2. Understanding Framing as a Process

A large part of research in social movement studies focuses on the strategic processes of framing, thereby investigating how movements deliberately construct and deploy frames to mobilize support and legitimize collective action. Key to this understanding is the theory of frame alignment (Snow et al., 1986) which states that through frame alignment processes social movements "link their interests and interpretive frames with those of prospective constituents and actual or prospective resource providers" (Benford & Snow, 2000, p. 624). In their review of the framing literature, Benford and Snow (2000) outline four distinct alignment processes: (1) *bridging* – forming linkages between at least two ideologically congruent but yet unconnected frames concerning an issue or problem; (2) *amplification* – idealizing or invigorating specific existing cultural values or beliefs in the frames; (3) *extension* – enlarging the initial frames to incorporate issues and concerns that are seen as important to potential adherents; and (4) *transformation* – reframing old understandings and meanings and generating new values and frames.

Cornelissen and Werner (2014), however, critique that the strong emphasis on the strategic use of frames in empirical studies has overall led to an outcome-focused and static research neglecting the processual and dynamic nature of framing as meaning construction, originally proposed in the broader concept introduced by Goffman (1974). Consequently, frames are not just strategic messages that need to be deployed, but interpretations that allow actors to make sense of the world and make choices grounded in that understanding (Kaplan, 2008). In response to the "top-down" strategic approach to framing, recent studies propose a "bottom-up" interactional perspective on framing (Gray et al., 2015; Kaplan, 2008; Reinecke & Ansari, 2021). This shift aims to counter the outcome-focused and static research and align with the dynamic nature of framing processes. Focusing on the microprocesses of framing, this stream of research argues that collective action frames may be "subject to spontaneous emergence, reorientation, and shifts in new situations through dynamic meaning-making on the ground" (Reinecke & Ansari, 2021, p. 382). However, I argue that this approach to some extent overlooks the pronounced strategic dimension of framing in social movements. Therefore, for this study, I primarily focus on strategic processes, as it aligns most with my case. Nevertheless, it is important to recognize that framing is neither merely "top-down" nor "bottom-up", but rather encompasses elements of both. Strategic alignment processes are necessary for mobilizing support but rely also on interactional and situated accomplishments (Snow et al., 1986).

In my research, I follow the call from Cornelissen and Werner (2014) for more empirical research on framing as an "ongoing process of meaning construction" (p. 206). To explore the dynamics of framing and how meaning is constructed, I apply a temporal lens to my frame analysis (Nyberg et al., 2020). This enables me to extend the strategic perspective and acknowledge the complexity of framing processes.

2.1.3. Determine the Effectiveness of Framing: Frame Resonance

If collective action frames aim to foster change by alignment processes to mobilize people – whether as an explicit strategic objective or a more implicit interactional emergence – the question arises whether any characteristics determine the effectiveness of a frame. The literature argues that the degree of frame resonance (Snow & Benford, 1988) is at the core of explaining why certain frames are more successful in mobilizing people than others (Williams, 2004). In this context, resonance is understood as "an audience´s experienced personal connection with a frame" (p. 716), and thereby differs from neighbouring concepts like legitimacy or justification (Giorgi, 2017). Audience refers to whoever the framing of the movement targets, including individuals, media, or politics. Empirical evidence validates that whenever frames resonate, especially with key decision-makers and strategic partners like the media, the likelihood of policy change is enhanced (Zeng et al., 2019). Moreover, when exposed to multiple framings over time, audiences tend to prefer framings that resonate with their values, concerns, and needs (Giorgi & Weber, 2015).

Research following the "top-down" approach argues that frame alignment processes are directed towards achieving resonance up front (Snow et al., 1986). Resonance is, therefore, understood as a key ingredient of effective framing. Benford and Snow (2000) outline two sets of interacting factors that influence the degree of frame resonance: (1) *credibility* – the frame´s consistency in terms of coherence between beliefs, claims, and actions; the empirical credibility of the frame; and the perceived credibility of those articulating the frame; and (2) *salience* – the centrality of the movements´ frame to potential adherents; the experiential commensurability in terms of how the framing resonates with everyday experience; and the narrative fidelity and, therefore, cultural resonance of the framing. Research taking on a "bottom-up" lens argues that resonance is "contingent and situationally produced" (Reinecke & Ansari, 2021, p. 403). Therefore, the appeal of a frame to external audiences emerges iteratively by actors leveraging resonating frames and is validated through interactional processes with key actors in the field (Lee et al., 2018).

Overall, resonance is viewed as a key determinant of framing success and non-resonance is considered a problem, as those frames may "fall on deaf ears" with potential adherents (Snow & Corrigall-Brown, 2005) and are unlikely to promote change at an institutional level (Zeng et al., 2019). Regarding the very objective of social movements, which is to create a shared (collective action) frame, this holds particular relevance, especially in the context of environmental movements where the core issue of climate change affects everyone. In a case study on nonviolent resistance campaigns in Thailand, the choice for polarizing frames was found to further antagonize societal segments triggering countermobilization (Sombatpoonsiri, 2023). To my state of knowledge, there is no research examining how and why movements deliberately choose to employ non-resonating or polarizing frames.

2.2. Unravelling Framing through a Temporal Lens

To further enhance our understanding of framing in social movements, I propose a temporal lens. Studying the temporal construction of collective action frames enables us to unravel the dynamic processes that constitute and develop the frame. Moreover, I argue that insights into the temporal patterns within interpretative framing processes will aid in elucidating and interpreting the movements´ strategic choices to deploy their framing. To my knowledge, with few exceptions (Munshi et al., 2022; Nyberg et al., 2020; Vandevoordt & Fleischmann, 2021), there is very little research that explicitly studies framing and social movements in relation to time. Additionally, as time is argued to be the central element in sustainability (e.g., Bansal and DesJardine, 2014) the temporal lens is especially appropriate for studying framing in environmental movements.

2.2.1. The Centrality of Time in Collective Action Frames

Framing is a temporally embedded process. Building on the agentic dimension of social movements, as their capacity to construct and employ a framing based on their interpretation of an issue, framing processes are inherently "informed by the past (in its habitual aspect), but also oriented toward the future (as a capacity to imagine alternative possibilities) and toward the present (as a capacity to contextualize past habits and future projects within the contingencies of the moment)" (Emirbayer & Mische, 1998, p. 963). For example, prognostic framing suggests alternative ways into the future and diagnostic framing assigns responsibility for action (Snow & Benford, 1988), thus constituting the projective, future-oriented dimension of human agency (Emirbayer & Mische, 1998). Consequently, framing is a process of making sense of the past, present, and future. Literature on time in sensemaking processes conceptualizes "time as the very medium through which actors address and translate their realities" with the present as the "locus of defining pasts and futures" (Hernes & Schultz, 2020, p. 4). For this, past, present, and future are not understood as separate and linearly aligned (Reinecke & Ansari, 2015) or as stable temporal categories but constantly negotiated in an ongoing present (Schultz & Hernes, 2013). Hence, actors face multiple temporalities simultaneously at any given moment and shift their temporal orientations dynamically according to the context (Emirbayer & Mische, 1998). For example, Vandevoordt and Fleischmann (2021) investigate how social movements are confronted with somewhat conflicting temporalities while having to switch their temporal orientation between a focus on the present, i.e., in situations of emergency and crisis, and need to expand their temporal horizon towards the future, i.e., to incorporate and actively shape broader changes in the future.

Derived from the strategic nature of framing, collective action frames challenge dominant temporal beliefs and aim to change perceptions and interpretations of time. Literature on temporal work examines how actors strategically construct, navigate, and capitalize frames (Granqvist & Gustafsson, 2016) to align (Kaplan & Orlikowski, 2013), but also influence and redirect (Bansal et al., 2022) temporal assumptions or patterns. This stream of research argues that the more actors engage in temporal work, the more likely they enable concrete strategic action and choice that diverges from the prevailing status quo (Kaplan & Orlikowski, 2013). Consequently, framing in social movements can be understood as a form of temporal work with the objective of creating a shared belief of temporality to foster change (Granqvist & Gustafsson, 2016; Nyberg et al., 2020). In a study on framing contests in the UK debate on fracking, Nyberg et al. (2020) investigate how actors construct temporality in their framing to make them convincing. They introduce the theory of *temporal portability* (p. 189), stating that the construction of time within a framing makes it meaningful to act on. They argue that frames with a certain temporal linearity resulting from a clear chronology of connecting a dominant past with a recognized future are more convincing and, therefore, actionable. Thus, for environmental movements to be successful in challenging dominant frames, the counter-frames need to gain temporal portability through solidification processes of certainty, simplicity, and familiarity. While this study underlines the centrality of time in framing contests, it only provides limited insights into how temporality is actively constructed based on an actor´s interpretation of climate change and strategically employed.

2.2.2. Temporal Perspectives in Framing Climate Change

Research at the intersection of time and sustainability has highlighted how actors´ temporal perspectives matter in response to climate change (e.g., Lê, 2013; Slawinski and Bansal, 2012, 2015). In the literature, temporal perspective is characterized by its "degree of emphasis on the past, present, future" (p. 141) – the *temporal focus* – and "the distance looked into past and future" (p. 142) – the *temporal depth* (Bluedorn, 2002).

The long-term nature of climate change accounts for actors increasingly shifting their temporal focus to the future. In times of crisis surrounded by uncertainty, Wenzel et al. (2020) argue that "actors have begun to experience the future as a problematic, open-ended temporal category" (p. 1442). Actors struggle over different views on ecological futures and the complexity of the issue makes framing challenging. Climate change as a framing is abstract and lacks immediate actionable elements (Nyberg et al., 2020). Some research states that it is essential to partly decouple from the present and engage with distant futures as abstract, and potentially more radical, imaginations of "what might be", to collectively develop alternatives for a future in the face of climate change that, consequently, can become treated as as-if realities (Augustine et al., 2019). This is closely related to the notion of desirability, which allows actors to articulate desirable futures through acts of imagination and provide hope (Gümüsay & Reinecke, 2022). Hence, desirable futures are performative in that they actively shape the future in the present. However, research on future-oriented action presupposes that actors have sufficient time to develop, and craft shared collective futures. In moments of crisis that imply urgency and need for immediate action, those future imaginaries must be brought into the here and now to "disrupt present thinking" (De Cock et al., 2021, p. 470). While engaging with desirable futures invokes imagination for transformation, the construction of a shared sense of urgency by anticipating undesirable futures may be of equal importance to evoke collective action in the first place (Alimadadi et al., 2022).

Consequently, the question arises as to how organizations should adjust their time perspectives in response to climate change. Research argues that organizations face inter-temporal tensions resulting from the different temporal horizons of economic and environmental logics (Slawinski & Bansal, 2015). Thus, organizations must make trade-offs between benefits now, such as short-term financial profits, and benefits later, such as long-term sustainability targets. A present-time perspective favours pay-offs in the short-term at the expense of the long-term, therefore, contributes to short-termism (Laverty, 1996; Marginson & McAulay, 2008) which inherently prevents organizations from taking action towards sustainable development (Bansal & DesJardine, 2014; Slawinski et al., 2017). Thus, this stream of research emphasizes the need for organizations to adopt a long-term, future perspective that aligns with the temporality of the environment (Slawinski & Bansal, 2012, 2015). To be sustainable, organizations must be willing to make intertemporal trade-offs by balancing the different temporalities (Reinecke & Ansari, 2015; Slawinski & Bansal, 2015)).

Surprisingly, with one notable exception (Kim et al., 2019), the present-time perspective has received little attention in the literature on temporality and sustainability even though it is commonly agreed that climate change is an urgent issue that requires immediate action (Wenzel et al., 2020). Kim et al. (2019) challenge the assertion that a present-time perspective is not compatible with sustainable development by introducing the concept of a *long present*. By framing the present as an extended duration, rather than a distinct moment with no temporal depth, actors were able to see connections among processes that enabled rather than hindered sustainable development.

3. Methods

My study follows a phenomenon-driven case study approach (Yin, 1993) to understand the dynamics present within the environmental movement LG (Eisenhardt, 1989). Conducting a single-case study enabled me to engage deeply with the phenomenon and collect and analyze a rich set of data from multiple sources to ensure the robustness of my findings. Combining the lens of framing with a temporal perspective ensured a contemporary methodological approach to analyze my data and craft out insightful, new theory.

3.1. Data Collection

The data collection happened between December 2022 and September 2023 and included interviews with participants of LG, internal documents, and contextual observations. The triangulation allowed me to complement and contrast my different data sources, thereby providing robustness to my findings and theory (e.g., Eisenhardt, 1989). Appendix A provides an overview of the empirical material I collected.

3.1.1. Interviews

To gain deep insight into individual perspectives on climate change and LG, I conducted ten interviews with eight participants of the movement over a period of two months. They were all explicitly interviewed as individuals involved in LG, not as official spokespeople of the movement (Munshi et al., 2022). In the sampling process I selected people with different demographic characteristics (esp. in terms of age and gender), different levels of experience (esp. in terms of history in climate activism and hierarchical level at LG), and different functional areas within the movement to limit bias and ensure a diverse range of perspectives (Eisenhardt & Graebner, 2007). In the beginning, I approached individuals via Instagram and open chat groups of LG. From these initial contacts, I applied a snowballing principle and asked for further potential interview partners. In addition, one interviewee agreed to share my request in internal chat groups of LG. That way, I was able to reach out to individuals who acted more in the background or did not use social media.

The interviews followed a guideline that was adapted to the respective individual and to the thematic focus over time. Where possible, I collected information on my interviewees in social media postings or press releases before the interview. Initially, the guideline focused on understanding how the individuals frame climate change and climate action in general. To examine the role of temporality, I asked questions on their interpretation of the past and present in terms of political action and the role of environmental movements, and how they imagine the future in the face of climate change. Over time, the focus switched to the specific framing strategies of LG and the role of polarization. Afterwards, as the importance of networking activities emerged as a key finding, I conducted three interviews with a specific focus on understanding the mechanisms of networking.

Overall, a key challenge in the interviews was to create a safe space where the interviewees felt free to talk about their perspectives and experiences. The reason behind this was that because of the harsh criticism the movement faces, in interview situations the participants feel like they must continuously justify their actions. Early on I noticed that this would hinder me from getting individual and insightful responses. To address this challenge, I started the interviews with more open and personal questions, e.g., about the interviewees´ motivation to participate in the movement, how they feel about the current political situation, and their views on climate activism in general. By showing interest, I meant to create a pleasant atmosphere for the interview. Afterwards, I asked more specific and potentially critical questions, e.g., about the strategy and organization of LG. Furthermore, all meetings included informal talk before and after the interview. With two individuals, I conducted a second interview. One was motivated because of a different thematic focus, as the interviewee had two relevant roles at the time. The other one had the aim to understand how the attitude and understanding of a new participant of LG changed over time (seven weeks in between the interviews). Both interviews turned out to be very insightful as a certain feeling of trust was established.

In this thesis, I refer to the interviewees with the term "participants" (of the movement LG). This term emphasizes the active involvement of the individuals in the movement. I deliberately decided against the more common term "activists" as it may not accurately represent the self-identification of the interviewees. Some individuals explicitly criticized the term because it implies a kind of identity that suggests a social life of dropouts where activism is seen as an end in itself. Therefore, I argue that "participants" can be a more inclusive term that encompasses a broader range of individuals involved in the movement. In addition, it includes the diversity of different activities the interviewees engage in, ranging from "protesters" to "networkers". To protect respondent confidentiality I do not assign pseudonyms as potential conclusions on the gender may harm their anonymity (Kaiser, 2009).

The interviews lasted between 39 and 70 minutes with an average length of 56 minutes. All of them were conducted in German. Nine interviews took place online via Zoom and one interview took place in person. All interviews were audio-recorded and subsequently transcribed verbatim. In total, the approximately 9.5 hours of recorded interview material (9h 21min) resulted in 128 transcribed, and fully anonymized, pages. All interview excerpts used to present my findings were translated from German to English. Table 1 provides an overview of the interviews I conducted.

3.1.2. Documents

In addition to the interviews, I collected internal strategic plans as well as information material, guidelines, declarations, and organizational charts from the internal wiki of LG. The movement is transparent about its strategy, structure, and organization and I was able to access relevant data online via the website. In addition, the interviewees provided me access to transcripts from their presentations on the strategy as well as scientific papers the strategy is based upon. The documents supported me in creating an effective guideline, to ensure that I do not include superfluous questions, and in the interviews, to be able to talk about the topic and

Table 1: Interviews

Code	**Role Description***	**Date**	**Length** (minutes)
LG_01	Protestant	05.06.	44
LG_02	Protestant, Coordinator of regional protest team	14.06.	52
LG_03	Supporter of protest and mobilization	15.06.	60
LG_04	Protestant, Supporter of protest and networking	22.06.	52
LG_05	Protestant, Coordinator of regional networking, Communicator of strategy	23.06.	70
LG_06	Supporter of public relations and mobilization	05.07.	39
LG_07**	Protestant, Coordinator of regional networking, Communicator of strategy	19.07.	67
LG_08	Coordinator of a nationwide networking pillar	24.07.	56
LG_09	Coordinator of a nationwide networking pillar	26.07.	57
LG_10**	Supporter of protest and mobilization	02.08.	64

* Roles in critical positions (i.e., leading roles with decision-making power and personnel responsibility – called *Coordinators*) are clearly defined. Nevertheless, it is not unusual for participants of LG to have multiple roles at the same time. For interviewees not having an explicit role I used the term *Supporter* and mentioned the primary areas they supported.
** Interviews with recurring interviewees

ask subsequent questions. Moreover, I used the documents to underpin and enrich my data but also to contrast them with the statements my interviewees made. In total, I analyzed 161 written pages.

3.1.3. Contextual Observations

Moreover, I followed the public discourse and the development of LG in detail by observing meetings and collecting other publicly available material. In December 2022, I participated in a regular mobilization process of LG, which consisted of two online meetings, where a participant of LG presented the strategy and gave room for discussion. In the following weeks, I participated in the weekly update meetings of LG, to get a gist of how the movement works and organizes. Moreover, I listened to two podcasts, one directly produced by LG and one from journalists who investigated the movement over seven months. Over the whole time, I followed three open LG chat groups, including update and discussion rooms and relevant Instagram accounts, including the official account of LG and accounts from key strategic individuals. This helped me to contextualize my data and enrich as well as contrast emerging findings. Due to timely constraints, this data was not analyzed systematically.

3.2. Data Analysis

My analysis followed an inductive approach going from data to theory (Eisenhardt, 1989). Due to the phenomenon-driven nature of my research, I approached the data with a broadly scoped research question to ensure flexibility in the analysis process (Eisenhardt & Graebner, 2007). Puzzled by the societal rejection the movement faces and them holding on to their strategy, I initially entered the field with a general interest in how LG tries to foster change. Cycling between the literature and my data, framing emerged as a valuable lens to study social movements. The concept of framing and frame analysis (Gamson, 1975; Goffman, 1974), and the toolkit of framing tasks, framing processes, and framing features (e.g., Benford and Snow, 2000) provided me with a decent body of literature to comprehend my emerging findings. The role of time and temporality emerged as a core theme early in the analysis process. Consulting the literature, the temporal lens turned out to be a promising, yet mostly unstudied, perspective on framing, because it enabled me to unravel framing and framing strategies through a dynamic lens (Cornelissen & Werner, 2014; Nyberg et al., 2020). To craft out the themes and patterns of the temporal construction and the framing strategies, I broadly followed the methodological approach of reflexive thematic analysis suggested by Braun and Clarke (2021). Iterating between data collection and data analysis enabled me to adapt and specify my interview questions to emerging puzzles and findings. Following I describe the four main steps of my analysis process.

Step 1: Familiarize with the data. In the first step, I aimed to familiarize myself with the dataset and get a general overview of the various data sources. I skimmed through all the collected documents and thoroughly read and re-read through the interview transcripts to immerse and critically engage with the data (Braun & Clarke, 2021). I made textual and visual notes for every interview to get an understanding of the different perspectives on climate change and LG and afterwards compared the notes to look for recurring topics. The centrality of time as the constructing dimension of the framing of climate change and the resulting strategies from LG emerged in this step.

Step 2: Coding the data. Second, after uploading my data into MAXQDA software, I systemically coded the interview and document data employing an open coding approach. Following an inductive orientation, I shifted my attention to *asking questions* and *focusing on puzzles* (Grodal et al., 2021). For example, I noticed how the interviewees focused strongly on the present as a small time window for action, but highlighted the need for broad, systemic transformation in the future. I iteratively went back and forth between data sources, making sure to go through every transcript at least twice (Braun & Clarke, 2021; Locke et al., 2022). To enhance coding quality and ensure reflexivity within the process, I created memos after every analysis session (Birks et al., 2008) to reflect on my progress, challenges, uncertainties or emerging candidate themes. In this step, I noticed how the codes revolve around two central topics, namely the construction of urgency by a present focus and the creation of a shared sense of urgency by polarizing and aligning framing strategies.

Step 3: Develop themes from overarching patterns. Third, I organized the codes to overarching patterns across the data. Following an abductive approach, I iterated between data, literature, and emerging findings. To enhance my understanding of the findings on temporality, I consulted the literature on inter-temporal tensions (e.g., Slawinski and Bansal, 2015), desirability in temporal work (e.g., Alimadadi et al., 2022) and the small body of literature available on temporality in framing (e.g., Nyberg et al., 2020). Moreover, I contrasted my findings on polarizing as a framing strategy to the literature on frame resonance (e.g., Lee et al., 2018; Snow et al., 1986). To organize my thoughts, I moved away from textual memos and continuously developed and modified models to visualize my findings. In this step, I developed meaningful categories for the construction of temporality and the framing processes.

Step 4: Integrate and reflect on findings. In the last step, I used post-it notes and mind-mapping to integrate my findings into a conceptual model. I specifically looked at the interrelations between my two core findings, thus how the specific framing strategies resulted from the temporal construction of the initial frame on climate change. Finally, I reflected on my findings and contrasted them against my contextual observations and data sources.

4. Findings

4.1. Introduction to the Case: We are *The Last Generation!*

LG focuses on disruptive forms of protest. The start of the movement can be traced back to September 2021, before the federal election in Germany, where a few individuals initiated a hunger strike in pursuit of a public dialogue on the climate crisis with the candidates for chancellor. To date, some of these individuals form the core group of LG being responsible for the strategic orientation of the movement. The first protest under the name of *Last Generation* took place in January 2022 as a road blockade in Berlin. Since then, LG initiated various campaigns (e.g., turning off oil pipelines, soiling famous art paintings, spray-painting a private jet, and, more recently, larger unannounced protest marches), while the main and most frequent form of protest is still road blockades in all major cities in Germany. To achieve maximum disruption, the protesters usually glue themselves to the streets, prolonging their removal by authorities. The strategic foundation behind those protests builds on the idea of peaceful civil disobedience. The movement itself defines peaceful civil disobedience as "the strategic use of peaceful means by citizens who want to make a difference socially, politically, or economically" (Letzte Generation, 2023b). In practice, this includes deliberate acts of rule-breaking or violation of the law under the imperative of peacefulness to disturb the public and build up political pressure. Civil disobedience gained popularity in climate activism in recent years. An early, well-known example is the British movement *Extinction Rebellion* (XR). Interestingly, LG was predominantly initiated by individuals with prior involvement in the movement XR, who split off because of a perceived missing strategic and organizational clarity. Furthermore, together with several other environmental movements practicing civil disobedience, LG formed the international *A22 Network*.

Because of their disruptive protest, LG polarizes the public. Despite the young history and relatively small number of members (e.g., in comparison to *Fridays for Future* (FFF)), the movement has received great media attention and triggered a debate on climate activism in Germany. While the majority agrees on the importance of climate protection (Lehmphul, 2016), there are sharp divisions on the legitimacy of the protest of LG. The media uses fierce rhetoric, e.g., branding the participants of LG as *Klimakleber* (Climate-Gluers) or *Klimachaoten* (Climate-Anarchists). Also, leading politicians criticize the protest as inappropriate, e.g., the chancellor of Germany Olaf Scholz called the actions "completely crazy" (dpa, 2023), or even counter-productive, e.g., vice chancellor Robert Habeck argues that "this process prevents a majority in favour of climate projection" in (dpa & epd, 2023). Furthermore, the protests of LG are largely rejected in society (Statista Research Department, 2023). In addition, there is a lack of clarity on whether the protests of LG, particularly the road blockades, are legal. Many protesters face major legal repressions including temporary custody and monetary fines. In May 2023, the repressions culminated in a nationwide raid, where the homes of several key individuals of LG were searched, based on the suspicion that LG forms a criminal organization.[1] Despite the criticism and repressions, the movement to date continues with their strategy.

To ensure the ability to act in the face of criticism and repression, LG organizes itself in a centralized and hierarchical structure. A core team of three individuals, who are also referred to as the 'founding team´, owns the decision-making mandate and is responsible for the strategic direction of the movement. Together with three other individuals, they form the core group. The core group discusses strategic questions and provides strategic orientation for the regional groups. While they are not directly involved in operational tasks like campaign plannings, they do own the power to veto if decisions do not align with the broader strategy. Overarching operational decisions are delegated to a coordination

[1] As of the time of the writing process (October 2023), the suspicion has been temporarily dropped. However, the discussion persists, and the possibility of an indictment has not yet been ruled out.

group of ten. Significant tasks, including finance, press, IT, legal, organizational development, integration, networking, protest planning, and social support, are organized into independent nationwide teams that follow the strategic orientation of the core group. As the movement grew, additional regional teams were established. All roles require specific training, e.g., how to act and react in critical protest situations or how to speak to the media. The founding individuals of LG implemented the hierarchical structure in response to what they perceived as a lack of organizational clarity in prior movements, particularly from their experiences in XR. This approach is unconventional within the broader climate movement where many groups tend to adhere to a grassroots democratic approach, as seen in the case of FFF:

> "[. . .] that's something you don't have in any other movement following this grassroots democracy approach. [. . .] If someone is sick, if someone goes to jail, if someone has a burnout or doesn't feel like it anymore, or if there are massive conflicts. We have a structure to deal with that. And I had previously experienced at XR what it's like when suddenly everyone is gone. If suddenly people with key functions either fell out with someone or were no longer convinced of the cause or were sick and moved away or something. Of course, this can also happen at LG. But the structures prevent that." (LG_05, Pos. 11)

The movement claims transparency about its strategy and structure. Strategic plans and organizational charts including detailed role descriptions are provided publicly on the website.

The movement sees itself as a *fire alarm*. LG frames climate change as a catastrophe, that has not yet been understood as such by politics. The movement emphasizes the urgency and the need to act now. As one participant concluded, while all interviewees agreed on this, "right now, the urgency of the climate catastrophe is the most important thing" (LG_04, Pos. 33). LG directs all actions towards creating a shared sense of urgency. This aspect is also evident in the movement´s early agenda, featuring claims that may appear as too small and, therefore, insufficient (e.g., speed limit 100 on highways, 9-Euro public transportation ticket), but underscore the necessity of transitioning into a crisis mode. The framing of LG addresses politics and the current government directly, stating that "the federal government is leading us into climate hell and continues to press on the gas pedal" (Letzte Generation, 2023a).

Overall, LG provides a valuable case for studying framing and temporality for two reasons. First, as the movement centres its framing activities around creating a shared sense of urgency it emphasizes the temporal dimension of climate change. Studying how individuals involved in climate movements construct urgency offers the possibility to unravel the importance of time and temporality in crafting a convincing frame. Second, due to the major public discourse and the polarizing impact of the movement, it provides an interesting case for examining how a movement strategically provokes polarization while also managing it in practice to enhance resonance for their framing, and thus, promote rather than impede change. Below, I describe in detail (1) how urgency is constructed in the framing – *the temporal construction of the fire alarm* - and (2) how the movement employs its framing through protest and networking to create a shared sense of urgency – *the mechanisms that trigger the fire alarm*.

4.2. Constructing the *Fire Alarm* Temporally

My data show that the temporal construction of the climate catastrophe framing revolves around two core themes: *constructing a temporal chronology* between the past and the future, zooming out and providing a convincing framework for the need to change, and *focusing on the present*, zooming in and triggering a present-moment emergency call for immediate action. Table 2 provides an overview of the data structure including the first-order codes my analysis relies on.

4.2.1. Temporal Chronology

The participants of LG constructed a clear temporal chronology by declaring that society is continuing on a wrong path, thereby linking the past to the present, and projecting that the current path will lead to an ever-worsening catastrophe, thereby linking the present to the (near) future.

Linking Past to Present

LG highlights political inaction in the past and argues that effective climate protection should have started much earlier. While speaking about the motivation to become active in the movement, one interviewee argued that "from a political point of view, we have failed to take many decisive measures. This is addressed above all to the last Federal Government and also before (. . .)" (LG_02, Pos. 7). This is amplified by political inaction despite progressive political agreements in the past:

> "And I had to realize that nothing happened. And the final push or decision was the moment when I realized that they were not even going to stick to the Paris Agreement." (LG_04, Pos. 3)

Whereas the interviewees acknowledged the successes of earlier movements, especially FFF, in terms of raising collective awareness of the topic of climate change, they declared that these approaches have fallen short in creating a sense of urgency for change. One interviewee pointed to the need for more disruptive forms of protest because of this situation:

> "For me, this is the conclusion from FFF. They have activated millions and put them on the streets, but nothing really happened in politics and society. Of course, the climate issue has moved more into focus. But real change just hasn't happened." (LG_06, Pos. 17)

The framing of LG addresses the current government and blames it for not acting but ignoring the climate catastrophe in the present. In a key internal strategy document LG assessed Germanys' long-term goal of climate neutrality by 2045 as follows: "The fact is, the government won't save us. Their actions are objectively insufficient." (TheoryofChange, Pos. 15-16). One participant argued that the course taken by the government reveals the wilful ignorance of politics against the backdrop of scientific evidence:

> "And at the moment, the optimism that is being tried there is so far away from the realistic situation that in my opinion it is utopian. Things are being said that simply no longer match with the scientific state as I perceive it, or even with many scientists I talked to, and do not cover the reality at all." (LG_09, Pos. 35)

Linking Present to Future

When the interviewees talked about the future, they usually projected the disastrous consequences of climate change. The framing revolved around the narrative that "everything that makes life possible, or the life of future generations, is threatened by the climate crisis". (LG_02, Pos. 7). This involved framing climate change as a complex issue by connecting the climate crisis to other crises, e.g., social crisis, refugee crisis, democracy crisis, and food crisis. Climate change is emphasized as the central challenge upon which the exacerbation or mitigation of other crises depends. One participant described how this plays a fundamental role in the framing:

> "I think with LG or what I notice, the narrative changes insofar that more and more these social consequences are also taken into account. What does it mean when we have water scarcity and food scarcity? What does it mean when large parts of the world are no longer habitable for our society? Because somehow people didn't understand, okay that means war in many parts of the world. That means extreme refugee flows. This means extreme pressure on Europe. That also means civil war in our country if there is no more water and no more food." (LG_01, Pos. 21)

Nevertheless, the climate catastrophe is not depicted as a temporally and spatially distant future. To substantiate the framing, interviewees often refer to close and recent extreme weather events like the *Ahrtahl* flood disaster in 2021, wildfires in Europe, and droughts in Germany. One interviewee emphasized how climate catastrophe is a near future as "disasters are becoming increasingly visible here in Europe" (LG_02, Pos. 19). The participants highlighted the temporal proximity of this undesirable future and how it will affect everyone:

> "And that's what we're going to see. This closeness in time, but also as a picture, that it's not about some polar caps melting or the glaciers in our mountains, but really on our doorstep, in our supermarket, where we go shopping every day, there won't be enough food for everyone. I think that distinguishes the framing of LG because that future vision is temporally closer." (LG_01, Pos. 21)

LG points to the fast-closing time window for action to mitigate the catastrophe. Therefore, the movement argues that our action now determines our future. This becomes apparent in the name of the movement, which in the complete version is *Last Generation before the climate tipping points*. After those tipping points, the catastrophe the participants projected can no longer be mitigated.

The movement deliberately emphasizes the undesirable future states that will result from not acting appropriately in the face of climate change to construct urgency. One participant mentioned that "we actually have a world to win if we change the direction to a world that is so much better and more beautiful and more solidary and more fair", but right after argued that now "you have to stress the crisis, that´s clear" (LG_08, Pos. 25-27). Another interviewee explained the need to emphasize undesirable futures because "the scale of the crisis is simply being completely underestimated" (LG_09, Pos. 31).

4.2.2. Present Focus

The participants of LG shifted the temporal focus to the present by highlighting that climate change and time for action is now, thereby prioritizing the present, and arguing that now is not the time to pivot action around (desirable) futures, thereby de-prioritizing the future as a temporal category.

Prioritizing Present

Central to the framing of LG is that the climate catastrophe has already started. A participant described the situation as "an absolutely urgent emergency situation right now" (LG_09, Pos. 9). Most interviewees stressed the importance of a present focus by emphasizing how a lot of people, especially those living in the global south, suffer from the consequences of climate change in the present:

> "And given what's coming to billions of people and what's already a reality for millions of people, I think. . . for me it's a matter of conscience. How can I look myself in the eye if I'm not trying to do everything? When I see that human rights are already being trampled underfoot daily." (LG_08, Pos. 11)

The movement claims a crisis mode by shortening the time horizons for action. The participants pointed to the need for immediate action in the face of the impending catastrophe and argued that "it is really already too late" (LG_05, Pos. 9). One interviewee described the importance of the present focus as follows:

> "Because if we don't act now, we're losing a lot, a lot. And very few people are aware that this temporal component simply exists and because of that in ten years a lot is just too late and that's why I think it's such a major thing at LG. You just want to have done as much as you can in the decisive moments and have tried everything to foster this change." (LG_09, Pos. 31)

LG emphasizes the importance of taking the first steps for action by bringing smaller claims like a speed limit on highways or an affordable public transport ticket to the fore. While the participants are aware that "we need a lot more" (LG_05, Pos. 13), they highlight that those measures are "effective and quick" and show that "we understood the crisis as a crisis" (LG_02, Pos. 17). Moreover, according to one interviewee, the focus on the first steps addresses tangible responsibility amidst the extensive transformations required, thereby ensuring that action is initiated and maintained:

> "[T]hat is a completely different lever compared to always having this huge catalogue of demands. And some say that, and others say that. And the government does what it wants, and so does the economy. Everyone can continue on their course. And always say yes, yes, we do a part of it, we contribute, for example, to the *construction turn 2045*. It just doesn't get so concrete, like you could do it now, that would be right, but you decide not to." (LG_05, Pos. 13)

De-Prioritizing Future

In the framing of LG, climate change is "not a question of the future, but a question of the present or even the past" (LG_01, Pos. 25). Most participants highlighted the tension between the "need for a systemic change" (LG_03, Pos. 39), a long-term perspective that includes major transformations and a lot of time to craft visions, and the urgency of the climate catastrophe as a present issue where the time window for action is continuously shrinking. LG deals with this tension by focusing on the present, under the rationale that discussions about the future fail to address the pressing crisis at hand:

> "If we all fall over a cliff into the abyss, it doesn´t matter if we had a great vision before." (LG_04, Pos. 35)

Moreover, the movement criticizes a long-term perspective. One participant pointed out that the prevailing time horizon practiced in politics is far too long-term oriented, deliberately neglecting the urgency for immediate action:

> "And when Friedrich Merz argues that we still have ten years to spare, then even today there is very little opposition from the media - So ten years to set the course, then we can start with the change, which from a scientific perspective is total nonsense." (LG_09, Pos. 35)

Consequently, LG deliberately excludes "bigger" questions, like the system question, from their framing. Not because the movement assesses the discussion as unnecessary in general, but because is not the immediate priority at this juncture. On the contrary, I had the impression that most interviewees desire a system shift in the long term. However, it is also a strategic consideration as topics such as criticism of capitalism carry significant political implications:

> "So, about the System Change. It is indeed a strategic decision not to emphasize this. I think, simply in order not to have broad conservative masses against it immediately." (LG_08, Pos. 51)

The interviewed participants suggested a step-by-step approach for sequentially working towards a desirable future. For this, the first step is to create and agree on the urgency of the climate catastrophe and the need for immediate action by establishing a crisis mode. Taking the first steps implicitly sets the course for a path from which a desirable future can gradually emerge. For the participants, after agreeing on the urgency, explicitly crafting visions of a desirable future will gain importance in a second step:

> "I believe that we must first bring society to the point where the need for change is seen. And only when the necessity, when the necessity is discussed seriously, the visions of the future will become more relevant." (LG_03, Pos. 13)

As an initial step towards shaping a socially just path into the future, LG proposes the establishment of a *Gesellschaftsrat* (society council), tasked with developing a comprehensive plan for Germany to get out of fossil fuels by 2030. By doing so, LG effectively delegates all future inquiries and concerns to this foundational claim:

> "This means that citizens are selected, who are brought up to date by scientists, and then draw up a plan for how we manage to get out of fossil fuels by 2030. And 2030 is not what we have come up with either, but what is derived from the IPCC report of 2022 - and that is our vision. So, we want a plan to be there." (LG_06, Pos. 29)

Although the long-term goal is embedded in LG´s overall strategy, it received limited attention in the interviews and, based on my observations, was not prominently emphasized in the public framing, particularly during the early stages.

4.3. Triggering the *Fire Alarm* through Protest and Network

My data show that LG employs the climate catastrophe framing by engaging in two processes: *pushing*, deploying the framing through disruptive protest, and *translating*, aligning the framing through networking activities. Triggering the metaphorical fire alarm is directed towards creating a shared sense of urgency in society. The movement faces the

Table 2: Data structure temporality

First-order codes	Second-order themes	Aggregated dimensions
Highlighting political inaction in the past	Past → Present *We are continuing on the wrong path*	Past → Future *Constructing a Temporal chronology*
Declaring past approaches as failed to create urgency		
Blaming politics for inaction and ignorance in the present		
Projecting catastrophe by connecting climate crisis to other crises	Present → Future *The current path leads into a catastrophe*	
Pointing to the fast-closing time window for action to mitigate catastrophe		
Emphasizing undesirable, near futures		
Highlighting that the climate catastrophe has started	Prioritizing Present *Climate change and time for taking action is now*	Present Emergency Call *Focusing on the present*
Shortening time horizons for action by claiming a crisis mode		
Prioritizing first steps as crisis measures		
Excluding “bigger” questions	De-Prioritizing Future *Now is not the time to pivot our action around (desirable) futures*	
Suggesting that desirable futures will & can only emerge out of a crisis mode		

tension of having to focus on disruptive forms of protest to convey the urgency for change, which in turn has polarizing effects, and having to gain resonance for its framing to foster collective action and change. To succeed in both objectives, the movement engages in two processes that the participants described as “parallel” and “equally important” (LG_07, Pos. 11-15). These processes operate through distinct mechanisms but build upon each other to ensure the effectiveness of the fire alarm. Figure 1 illustrates the interactions of the framing processes that emerged from my analysis.

4.3.1. Pushing

The movement engages in disruptive forms of protest to display the urgency for change. The objective is to build up and uphold political pressure. While discussing the intensity of the protest, one participant described that consistency is essential:

> “So that would flatten out immediately I think if you didn’t keep it up. And I think you must put some pressure on the government. Above all, just to make clear that contrary to what is being said we´re not on the right track with the government’s current measures.” (LG_08, Pos. 33)

LG argues that because of the urgency of the climate catastrophe, it is necessary to interrupt everyday life. When talking about the street blockades in podcasts or press interviews participants often utter that they do not want to do or like the protests at all. However, as one participant told me, they see no other way to stop the crisis from being displayed:

> “That’s why we interfere with everyday life. That’s why we make our blockades. And we say watch out, the house is on fire, we must extinguish it. We must do something. And that is exactly what I see as our task. That we point out that we must do something about it.” (LG_06, Pos. 15)

The underlying mechanism of the disruptive protests lies in their confrontational nature. This approach enables the movement to capture (media) attention and stimulate a discourse. By the high frequency and unwavering consistency of the protest, LG compels the public to actively engage with and take a stance about the movement. One participant highlighted the importance of being unignorable to enforce a debate in the present:

> “No one’s ignoring us. That’s the ultimate target. We must be unignorable. That was also what immediately became clear to me. We must no longer allow ourselves to be ignored. We’re a fire alarm.” (LG_05, Pos. 17)

In doing so, LG deliberately triggers polarization. To ensure that the discourse receives the necessary critical attention, the “protest in general is just enormously important to create a certain social tension” (LG_09, Pos. 11). Moreover, one interviewee argued that due to the impending catastrophe polarization might be necessary to foster fast and comprehensive change:

> “But above all, I think that without this polarization we will not succeed in shifting the social discourse in time to one of the two poles, namely the crisis, crisis, crisis - pole.” (LG_10, Pos. 7)

The strategic objective of the disruptive protests is to deliberately elicit emotional responses among those affected by the actions. To achieve a "state of shock", the movement actively evokes emotions of "anger" and "rage" (LG_01, Pos. 29). Thereby, emotionalization is used to convey a heightened sense of urgency but also to amplify the discourse:

> "And to arouse these extremely strong emotions by simply blocking people on the street, and specifically as many as possible and as much as possible. And again, and again. That generates clicks. That garners attention." (LG_03, Pos. 17)

In all public actions, LG maintains a uniform and static framing, repeatedly emphasizing the urgency of the climate catastrophe and the imperative for immediate action. The attention and polarizing discourse from the protests provide the participants of LG with access to public stages to push their framing (e.g., press interviews, talk shows, and court hearings). The movement strategically leverages these moments to raise awareness about the climate catastrophe. One interviewee pointed out that "we just have to fight to stay present in the media and push our framing" (LG_03, Pos.17).

The disruptive protests turned the majority of society against the movement. When I asked participants how this makes sense strategically, the interviewees stated that mobilizing as many people as possible is in fact "not the main goal" (LG_04, Pos. 14-15). To gain support and solidarity, the movement relies on *Backfiring Moments*. The protest triggers over-reactions (e.g., home searches, major legal repression, exaggerated political rhetoric), that raise the question of whether those reactions are adequate to a movement fighting for climate action: "Why is she in jail now? Why isn't she sitting on the organ bench with us? Okay, she resisted. Why? What's going on? It's a climate catastrophe." (LG_05, Pos. 11). When society perceives the reactions as inappropriate, it amplifies the resonance for LG. Moreover, the movement strategically makes use of backfiring moments to confirm and reinforce the framing that politics did not understand the crisis:

> "By these overreactions the government then exposed itself and more and more people realize okay, it is not willing to end the injustice of the ongoing destruction of the world. And the government prefers to fight peaceful people who stand up for us all." (LG_02, Pos. 29)

By triggering backfiring moments, the movement indirectly fosters support and solidarity. Most participants described powerful examples to underline the success of the strategy:

> "At the moment of the house searches and the huge shock, so many people came to us at the same time. Unbelievable how many people showed solidarity." (LG_04, Pos. 5)

Nevertheless, participants acknowledge that "just disturbing is not enough for initiating a change process" (LG_09, Pos. 11).

4.3.2. Translating

To directly foster support and resonance, the movement aligns the framing by engaging in networking activities. LG practiced networking from the beginning, but it gained relevance over time as the attention from the protest acted as a door-opener: "Because we are so well known, it works extremely well that at the moment we are able to talk very easily, especially with very well-known and renowned scientists and institutions." (LG_09, Pos. 9). Compared to the protest, participants described networking as relatively invisible because it "primarily takes place behind the scenes" (LG_07, Pos. 31).

The objective of networking is to engage in discussions about climate change with all key actors in society, including religious institutions, political entities, law enforcement, and the scientific community. Besides explaining the strategy of LG, participants involved in networking aim to educate about and raise awareness for the climate catastrophe. The short-term goal is to gain support from institutions and individuals through public solidarity statements or resources that enable the protest, such as accommodation or legal assistance: "So please, you have to position yourselves in society right now and say that they are right, this is a fire alarm and not a false alarm." (LG_07, Pos. 7). The long-term goal is to stimulate transformational processes within these institutions, with the overarching goal of reorienting their actions to revolve around the climate catastrophe. One networking coordinator described the vision in networking with scientific institutions:

> "In the long term, together with Scientists Rebellion, we plan lectures at universities throughout Germany. Lectures where we try to call for more active science, where we argue that it is no longer enough to just keep on publishing while they are completely ignored by the general public anyway, not noticed at all. And politics do not refer to them either. And that more needs to be done to bring these scientific publications to the public. And that neutrality does not contradict - neutrality and passivity are not the same thing." (LG_09, Pos. 9)

In networking, participants translate the climate catastrophe framing of LG to the different pillars of society: "Networking means I translate, it's a fire alarm." (LG_07, Pos. 13). In the process, participants tailor the language used in networking discussions to align with the language common in the specific pillar. A senior participant leading a local networking team described this as follows:

> "This is not our normal LG lecture, but very specifically adapted for church contexts I´d say. Although - not adapted, it is rewritten. So, it is written differently now with a theological argumentation. Because basically within every area we address, there are always people who are

> well informed within the specific context and who have a clue." (LG_08, Pos. 5)

Translating means connecting the undesirable futures resulting from the climate catastrophe to the context of the specific pillar and making it tangible. Therefore, in networking talks, the participants "highlight what impact this can have for you and your area, in which you work." (LG_09, Pos. 13). For example, in networking with healthcare organizations LG frames "climate protection as health protection". In a strategy document that suggests various framing possibilities for networking with healthcare organizations, LG projects how the climate catastrophe will have direct, disastrous consequences on the pillar: "The health system is already at its limit – how is it supposed to survive the climate collapse?", but also how the pillar is specifically responsible to act: "Our job is to save lives!". While the portrayal of these undesirable futures is deliberately realistic and proximate, they are not framed as inevitable. Rather, participants underline the agency vested in the various actors to actively engage in mitigating the impending catastrophe, emphasizing the imperative of initiating immediate action. Appendix B provides a more extensive overview of strategically employed framings in the process of translating by presenting compelling examples sourced from internal documents.

Translating the catastrophe framing to the different pillars of society aims to create an emotional connection to climate change. Contrary to the protests, which primarily employ emotionalization to trigger anger and rage to intensify the discourse, emotionalization in networking focuses on "building connections" and displaying that "we actually want the same" (LG_07, Pos.21) to foster resonance for the framing of the climate catastrophe. According to the participants involved in networking, evoking emotions, especially emotions of fear and regret, is necessary to display urgency and foster immediate action:

> "[...] I think a partial emotionalization of the problem is necessary. Of course, it is always important to stick to the facts. But this is, I think, an important point because people in their everyday life are entangled in a displacement society where you simply get along with it and you don't necessarily want to deal with it because it's just a stupid subject. And I think that's why it's quite necessary to emotionalize it in parts." (LG_09, Pos. 13)

5. Discussion

This study investigated how time and temporality are constructed within and shape the framing of social movements. For this purpose, I examined the collective action frame of the environmental movement LG in Germany. I found that the movement frames climate change as a catastrophe and claims a crisis mode, thereby seeing itself as a fire alarm. To create a shared sense of urgency the movement seeks to shift the temporal focus to the present. Applying a temporal lens enabled me to unravel the initially puzzling framing strategy of polarization that contradicts earlier literature on frame resonance. Therefore, my findings enhance theories on frame alignment processes by proposing a temporal lens that underlines the complexity of framing. Below, I theorize my empirical findings and discuss their contributions. First, I integrate my findings in a theoretical model that links the temporal construction of the frame to the strategic framing processes and their primary objectives in terms of the degree of resonance in what I call *crisis frames*. Second, I discuss the contributions of my findings to the literature on framing in social movements and to the literature at the intersection of time and sustainability. Finally, I elaborate on the limitations of this study and propose potential avenues for future research.

5.1. A Theoretical Model of Temporality in Crisis Frames

Reflecting on the interrelations between the temporal construction and the framing processes, I integrate my two core findings into a conceptual model of temporality in framing processes, especially in crisis frames. Figure 2 illustrates how the framing processes reflect the construction of time in the initial framing of climate change and highlights how the interpretation of time determines the strategic framing processes.

In the face of the pressing climate crisis, the framing of LG is directed towards shifting the temporal focus to the present. Therefore, the movement challenges dominant temporal beliefs reflected in the past and present climate action of politics for being inadequate to the current situation. In doing so, the framing constructs a temporal chronology between a dominant past and a projected, undesirable future. The connections made between the dominant past and the undesirable future serve as diagnostic framing by addressing blame to politics in the past and present and assigning responsibility to the current political decision-makers for immediate action (Snow & Benford, 1988). By anticipating and strategically emphasizing undesirable futures the movement constructs urgency and emotionalizes the issue to increase actionability because the frame directly connects the need for action in the present to the possibility of undesirable states in the near future (Alimadadi et al., 2022). The constructed chronology has limited temporal depth. Thus, LG emphasizes the near future and the near past to make explicit connections to the present. Building on the constructed chronology, the framing temporally focuses on the present. The present-time perspective serves as prognostic framing by claiming a crisis mode as the proposed way to act in the face of the climate crisis (Snow & Benford, 1988). However, the diagnostic dimension predominates the framing, with the movement only offering a partial delineation of the desired crisis mode. The claimed first steps primarily function as symbolic claims that should display how present political action does not reflect a crisis mode. The present is framed as a limited, fast-closing time window for action that determines the future and as

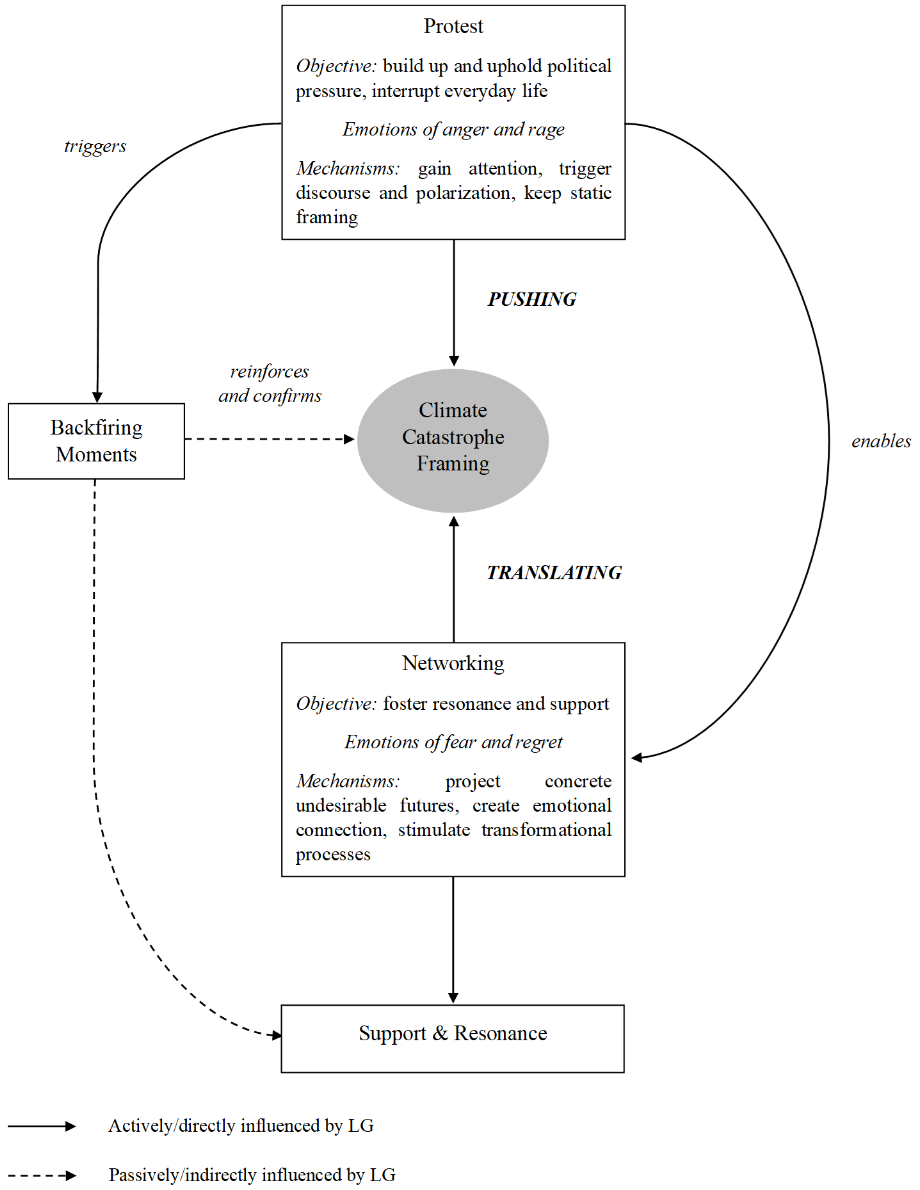

Protest
Objective: build up and uphold political pressure, interrupt everyday life
Emotions of anger and rage
Mechanisms: gain attention, trigger discourse and polarization, keep static framing
triggers
PUSHING
reinforces and confirms
Backfiring Moments
Climate Catastrophe Framing
enables
TRANSLATING
Networking
Objective: foster resonance and support
Emotions of fear and regret
Mechanisms: project concrete undesirable futures, create emotional connection, stimulate transformational processes
Support & Resonance
Actively/directly influenced by LG
Passively/indirectly influenced by LG

Figure 1: Strategic framing processes

the last possibility to prevent undesirable futures. Consequently, a better, desirable future can only emerge out of the crisis mode in the present. The movement intentionally deprioritizes the future as a temporal perspective to accentuate the sense of urgency and thereby strategically excludes questions for broader changes in the future. The way the movement constructs time forms the foundation of the framing and determines the processes through which the movement deploys its frame to foster collective action.

To convey the constructed urgency LG actively manages the balance between disrupting the present by keeping a static framing and aligning its framing to achieve a certain degree of resonance. This balancing act is reflected in the two interrelated but in practice decoupled processes of *pushing* and *translating*. While the movement utilizes synergies between the processes, such as leveraging the attention from the protest for access to networking partners, they build on fundamentally different mechanisms. Disruptive strategies, including all public activities like protests, press interviews, and court hearings, are employed to push the present focus. Therefore, the movement aims to disrupt the present and strategically trigger a certain degree of polarization. On the one hand, polarization is used to gain attention, be unignorable, and provoke a discourse in the here and now. This logically emanates from the framing of climate change as a catastrophe, emphasizing the limited time window for action. While specific campaigns, such as protests targeting affluent individuals or oil companies, garnered significantly greater resonance among the public audience, the movement observed a stark disparity in the level of attention received by

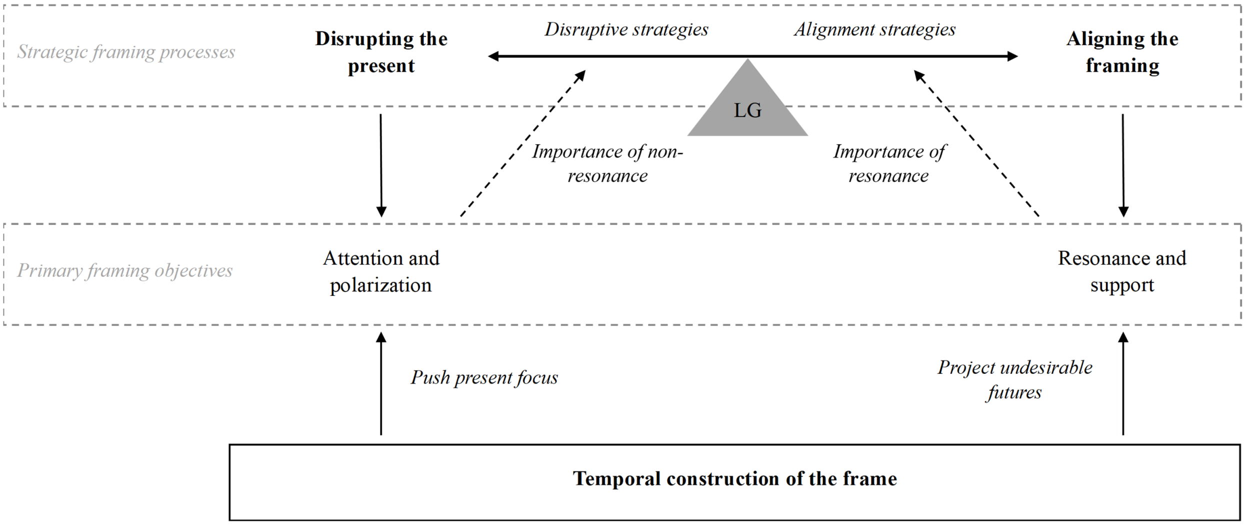

Strategic framing processes
Disrupting the present
Disruptive strategies
Alignment strategies
Aligning the framing
LG
Importance of non-resonance
Importance of resonance
Primary framing objectives
Attention and polarization
Resonance and support
Push present focus
Project undesirable futures
Temporal construction of the frame

Figure 2: A theoretical model of temporality in crisis frames: Balancing disruption and alignment

these actions compared to the road blockades. On the other hand, the framing inevitably generates polarization due to its present focus, implying direct and immediate influence on the actions of political and economic actors, rather than in the future. Thus, the temporal construction of the frame limits the margin in which the framing can be aligned with potential adherents. Consequently, within the context of a crisis, attaining a high degree of resonance may not necessarily be the primary objective of strategic framing processes, and triggering some degree of non-resonance might be crucial.

However, the movement endeavours to foster resonance by engaging in networking activities. LG aligns the climate catastrophe frame to persuade prospective supporters of the framing. By translating the climate catastrophe frame to the various pillars of society the movement projects concrete undesirable future states that will inevitably result from inaction in the present. To provide a tangible perspective, the complexity of crises intertwined with the climate catastrophe is deconstructed and contextualized within the framework of the institution or actor. Thus, networking also focuses on conveying the urgency and the need for immediate action in the here and now but partially aligns the framing to achieve that objective. By especially highlighting undesirable futures in the context of the respective networking partner, the movement triggers feelings of fear and regret and, therefore, aims to create an emotional connection to the climate catastrophe that in turn fosters resonance for the framing of LG and stimulates support. In the short term, the movement focuses on strategically powerful partners that can support the movement in public through solidarity statements and function as trustworthy carriers of the catastrophe framing in their specific context. In the long term, the movement engages in transformative processes within the institutions or actors by emphasizing that another future is possible.

In summary, the case illustrates how the strategic framing processes are linked to the temporal construction of the framing. For crisis frames that aim to create a shared sense of urgency, aligning the framing to achieve resonance may not constitute the primary objective of all strategic processes because of the inherent present focus. Instead, the movement deploys its framing in a dialectic process, actively orchestrating an intricate equilibrium between disruptive strategies, aimed at triggering non-resonance, and alignment strategies, aimed at fostering resonance. To shift the temporal focus to the present the case demonstrates that movements may strategically employ polarizing framing tactics. However, also for crisis frames resonance is necessary. The movement recognizes the importance of creating resonance within society to foster change. Nonetheless, for frames with a strong present focus, generating resonance and conveying urgency presents an enormous challenge. To partially align the framing, the case indicates how movements may initiate alignment processes independently, which subsequently take place decoupled from the main and visible public campaigns.

5.2. Contributions and Implications

The theoretical contribution of this study is twofold. First, it contributes to the research on framing in social movements by illustrating the fundamental role of time and temporality in framing processes and providing an in-depth temporal perspective on a unique, contemporary case. Second, it contributes to research at the intersection of time and sustainability by emphasizing the relevance of a present-time perspective in the face of climate change and, therefore, enhances theories on inter-temporal tensions and future-oriented action.

This study reveals how framing processes cannot be studied isolated from their temporal construction. Especially for collective action frames that serve an interpretative and a strategic function (Snow & Benford, 1988), time plays a fundamental role in how actors make sense of the past, present, and future but also aim to challenge and change dominant temporal beliefs. By unravelling how the movement strategically constructs a temporal chronology and tries to shift the temporal focus to the present, this study shows how framing can be understood as a form of temporal work (Bansal et al., 2022; Kaplan & Orlikowski, 2013; Nyberg et al., 2020). A temporal lens provides outcome-focused research on strategic framing processes and the somewhat static framing literature in general with a dynamic character to come up with the complexity of framing as an ongoing, interpretative process of meaning construction (Cornelissen & Werner, 2014).

However, this study extends beyond how actors construct time (Granqvist & Gustafsson, 2016) and mobilize temporality (Nyberg et al., 2020) by examining how the temporal construction and the strategic framing processes are inextricably intertwined. Building on the theory of temporal portability introduced by Nyberg et al. (2020) this case illustrates how a movement seeks to construct a temporal chronology to make the framing more convincing and actionable by for example "reducing the time frames for climate action" and "linking climate change to what are seen as legitimate and immediate concerns" (p. 192). The present-time perspective resulting from the movements´ interpretation of climate change as a catastrophe that requires immediate action affected the framing processes and objectives. The present focus posed a two-fold challenge, requiring the simultaneous cultivation of resonance and compelling conveyance of urgency, which in turn constrains frame alignment possibilities. Moreover, polarization is employed as a framing strategy in the process of pushing to display and enforce the present focus. While the framing literature conceptualizes resonance as a key mechanism for effectiveness (Benford & Snow, 2000; Cornelissen & Werner, 2014), this study illustrates how a movement had to balance strategies that aimed to create non-resonance and resonance to create a shared sense of urgency. This advances our understanding of frame resonance (Lee et al., 2018; Snow & Benford, 1988; Snow et al., 1986; Zeng et al., 2019) insofar as resonance might not always be the primary objective of all framing processes and employing polarizing frames can be a strategy to display the temporal perspective of the framing.

Furthermore, this study introduces the process of translating as an alternative form of frame alignment. Building on the theory of strategic alignment processes (Benford & Snow, 2000; Snow & Benford, 1988), translating contains elements of frame bridging because it involves making connections between two unconnected but ideologically related frames (e.g., networking with other climate justice movements or scientific institutions), and frame extension because it expands the scope and relevance of the frame by associating it with other issues or interests (e.g., networking with health organizations or church). However, it introduces a unique element in that it focuses specifically on the context of the different institutions or actors and tailors the framing to resonate with their specific concerns and values. The emphasis on creating an emotional connection to the issue in line with the temporal construction of the catastrophe framing distinguishes translating as a distinct process in the broader frame alignment theory. Translating, in this context, involves not only conveying but also persuading potential supporters to adopt the movement´s temporal beliefs, aiming to establish quick common ground during moments of crisis.

This study also contributes to the literature on intertemporal tensions in sustainable development (e.g., Bansal and DesJardine, 2014; Reinecke and Ansari, 2015; Slawinski and Bansal, 2015) by illustrating how a present-time perspective may be more valuable than previously assumed. The literature argues that organizations have to apply a long-term perspective to be sustainable and examine how actors navigate different temporalities and make trade-offs (Reinecke & Ansari, 2015; Slawinski & Bansal, 2015). Surprisingly, the case of LG represents an example of an actor who fights for a sustainable future and simultaneously claims a present-time perspective. Therefore, this study provides insights into how actors assess a long-term, future perspective as problematic in the face of climate change. First, because the undeniable urgency of the climate crisis requires immediate action to mitigate the ever-worsening situation and prevent the tipping points. By prioritizing the present, the movement underscores the idea that a socially just future can only be secured through immediate action in the present. Second, adopting a long-term perspective on climate change implies that there is enough time for change. As a result, actors might not perceive the urgency or "hide" behind long-term goals and, therefore, deliberately postpone climate action as a question for the distant future.

Thus, this study contributes to the discourse on how actors imagine collectively dealing with the tension between the need for immediate action because of the urgency and the need for broader changes in the future inherent in climate change (Slawinski & Bansal, 2015; Wenzel et al., 2020). The proposed crisis mode suggests that by prioritizing short-term goals, such as the implementation of first steps as crisis measures, concrete action and responsibility can be assigned. This enables a more effective evaluation compared to complex, long-term goals, determining whether society perceives the urgency and necessity for change. The short-term goals can further function as the starting point through which a catastrophe can be mitigated and a better, desirable future can emerge from. By perceiving the present as interconnected to the future rather than a distinct moment, the present-time perspective provides an alternative view of sustainable development that is not about trade-offs (Kim et al., 2019). Consequently, this study reveals how movements may not assess the missing long-term perspectives but the unwillingness to act in the present as the key issue that hinders sustainable development.

Moreover, this study contributes to research on future-oriented action by illustrating how actors make connections between the present and the future to construct urgency. While prior research focuses on how distant and desirable futures are imagined to invoke transformation (Augustine et al., 2019; Gümüsay & Reinecke, 2022), this study shows how actors deliberately project near and undesirable future states to emotionalize and create a shared sense of urgency (Alimadadi et al., 2022). To foster collective action the anticipation of undesirable futures directly connected to present states might be crucial to achieve a sense of urgency for change in the first place. Furthermore, this study advances research on future desirability (Alimadadi et al., 2022; Gümüsay & Reinecke, 2022) by showing how actors strategically exclude the amplification of desirable future possibilities. By not making connections between the present and a distant, desirable future the movement tries to avoid the discourse from derailing to political future imaginaries and thereby not tackling the actual urgency of the crisis in the present. The case depicts the challenge of connecting future social imaginaries to the present (De Cock et al., 2021; Nyberg et al., 2020) in urgent situations. In the face of a pressing crisis, the question arises as to how to achieve resonance for a framing that depicts a future vision that may be radically and ideologically different from the present, e.g., anti-capitalism. Therefore, the present-time perspective is used to deliberately distract from broader changes in the future to foster actionability in the present (Vandevoordt & Fleischmann, 2021).

On a practical note, movements trying to foster change need to craft collective action frames by constructing temporality in a convincing way (Nyberg et al., 2020). In times of crisis, movements may need to employ a strong focus on the present to construct urgency. However, focusing on the present and excluding broader future questions may result in a perceived imbalance between the movement´s claims and the action of the movement. While a simultaneous focus on the present and future might pose a risk for movements, potentially shrinking its power (Vandevoordt & Fleischmann, 2021), it is important to make strong connections between the present and the future and explain the temporal focus to achieve resonance, e.g., by engaging in networking activities.

5.3. Limitations and Future Research

An inherent limitation of this study stems from the single-case research design. LG provides a highly interesting and relevant case for studying social movements and the organization of time in framing because it is critical, in that the unusual framing strategies challenge existing theory, and is unique, in that the approach is new and triggered major discourse (Eisenhardt & Graebner, 2007; Yin, 1993). Therefore, conducting a single case study that allows immersion and a rich description of the phenomena can be a powerful example for extending existing theory and inspiring future research (Siggelkow, 2007). However, as a single case study only investigates one specific example the quality of the emergent theory is limited in terms of robustness, generalizability, and testability (Eisenhardt & Graebner, 2007). To enhance and strengthen my findings, future research could conduct multiple case studies (Yin, 1993). Specifically, I see great potential in comparative case studies that investigate polar types (Eisenhardt, 1989) of movements that seemingly have similar objectives but employ different framing strategies. Studies could examine how and why environmental movements construct time differently in their framing while agreeing on the urgency of climate change. FFF, for example, has a strong temporal focus on the future and explicitly claims systemic change. Research could then investigate the implications of diverse temporal constructions for collective action in the broader climate justice movement. Nonetheless, I also encourage future research to test the applicability of my findings in other contexts, such as social movements where there is less obvious temporal tension between urgency and broader chances in the future.

The short time horizon of the data collection process represents another limitation of this study. Derived from the processual and dynamic understanding of framing, it is certain that frames and framing strategies will change over time (Cornelissen & Werner, 2014). Although the strong and deliberately transparent strategic component of framing in the movement as well as observing the development over nine months enabled me to analyze and contextualize framing processes, the time horizon in which the interviews were conducted (two months), being the primary source of my findings, represents a rather short snapshot of time. If and how the framing of LG will change over time and how the movement will assess whether the claimed crisis mode is established requires further observation. An interesting question would be whether the temporal focus will shift back to the future once the movement perceives that a shared sense of urgency is achieved and how this affects framing strategies. While one plausible scenario could be that LG acts as a temporary organization that disbands upon completing its mission, an alternative scenario is that the movement significantly changes and takes on a new role in the climate change discourse. Furthermore, I see the need for studies that investigate the societal outcomes of polarizing framing strategies, e.g., in terms of policy change (Zeng et al., 2019) or the mobilization of counter-movements (Sombatpoonsiri, 2023). Future research could for example analyze framing contests to study the effectiveness of non-resonating frames (Nyberg et al., 2020). In times of crisis, the question arises as to how much resonance is needed to foster change and whether it is sufficient to solely address resonance on an emotional level, e.g., by activating emotions of fear and regret (Giorgi, 2017). In sum, I encourage future research to critically evaluate resonance as the key mechanism for frame effectiveness, particularly within ideological and political contexts (Giorgi & Weber, 2015).

Another limitation stems from the lack of observational data to capture the complexity of the phenomena. The observations I conducted solely had the objective to contextualize and contrast the emerging findings. I did not observe how the participants pushed their framing in actual situations like protests or court hearings and how participants translated the framing in networking talks to create an emotional connection. Future research could, therefore, benefit from collecting first-hand experiences by attending protests, internal plannings, or networking talks to further nuance the relation of temporality and framing processes. Moreover, contrasting emerging findings with perspectives of critics or fellow climate movement participants could provide valuable insights into the external evaluation of the strategy. Consequently, and building on the call from Cornelissen and Werner (2014), I encourage future research to conduct ethnographies over a longer period to come up with the complexity of framing processes and to enrich my findings.

Overall, this work should also be a call for management and organizational studies to incorporate more strongly atypical, other than business-related cases in their research. Especially in contexts like climate change that imply a need to change in the near and distant future, research should put a stronger emphasis on actors who challenge dominant beliefs and fight for change, like environmental movements (Gümüsay & Reinecke, 2022; Wenzel et al., 2020). Moreover, this case is a powerful example of how highly organized social movements can be, especially when they play a strong counterpart to dominant beliefs. Future research could investigate how movements in times of crisis reach this level of organization that goes beyond temporarily achieving 'organizationality' through identity claims (Dobusch & Schoeneborn, 2015) to establishing boundaries and constant actionability through structure that enables actors to build resilience to societal criticism and negative evaluations (Roulet, 2020).

6. Concluding Thoughts

Humanity is currently facing a self-induced yet existential crisis. Facing such a crisis requires a radical transformation of society, particularly our economic system. Scholars and practitioners are confronted with the pressing question of how societal change and political action can be organized in time to ensure a sustainable future. Although the necessity for change has been acknowledged for quite some time, there is still an underlying perception that progress remains insufficient. While the imperative of transcending the familiar paradigms of "business as usual" is apparent (Wright & Nyberg, 2017), especially within social actors like environmental movements, climate change as an actionable framing is "indeed foreign to our very sense of being" (Nyberg et al., 2020, p. 193).

This thesis sought out to unravel the framing activities of the environmental movement LG in Germany puzzled by the seemingly static and counterproductive strategy. By applying a temporal lens, this study illustrated the complexity of (strategic) framing processes. Framing climate change as a catastrophe involved a strong focus on the present and near undesirable future states among participants of the movement. My findings show how the temporal construction shaped the strategic framing processes. To convey the sense of urgency, framing activities were in parts deliberately targeted at triggering polarization by non-resonance. Therefore, the findings enhance research on framing in social movements by demonstrating the centrality of time and its consequences on strategic framing processes.

Moreover, the study holds valuable implications for the discourse on time and sustainability, highlighting the primary challenge of cultivating a collective sense of urgency to foster action in the face of the escalating climate crisis. Certain academic discourses within the field, such as those focusing on long-term perspectives or desirable futures, found limited applicability in this case. Instead, concepts of undesirable futures and emotionalization that resulted in a present-focused approach to climate action were far more prominent. On a final note, referring to the introductory quote by the Secretary-General of the United Nations António Guterres, "[t]ime is no longer on our side". This is further amplified by powerful and affluent actors who attempt to slow down sustainable development for their own gain, leading to a precarious situation where "delay is the new denial" (p. 2) in the discourse on climate action (Shue, 2023). Against the backdrop of scientific evidence, the urgency to redirect our discourse on sustainability, both in academia and practice, to the present and the means of immediate action has likely never been more pressing. I hope this study can inspire future research to embrace the present-time perspective and recognize the relevance of societal actors like social movements more strongly.

References

Alimadadi, S., Davies, A., & Tell, F. (2022). A palace fit for the future: Desirability in temporal work. *Strategic Organization*, *20*(1), 20–50.

Augustine, G., Soderstrom, S., Milner, D., & Weber, K. (2019). Constructing a Distant Future: Imaginaries in Geoengineering. *Academy of Management Journal*, *62*(6), 1930–1960.

Bansal, P., & DesJardine, M. R. (2014). Business sustainability: It is about time. *Strategic Organization*, *12*(1), 70–78.

Bansal, P., Reinecke, J., Suddaby, R., & Langley, A. (2022). Temporal Work: The Strategic Organization of Time. *Strategic Organization*, *20*(1), 6–19.

Benford, R. D., & Snow, D. A. (2000). Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, *26*, 611–639.

Birks, M., Chapman, Y., & Francis, K. (2008). Memoing in qualitative research: Probing data and processes. *Journal of Research in Nursing*, *13*(1), 68–75.

Bluedorn, A. C. (2002). *The Human Organization of Time: Temporal Realities and Experience*. Stanford University Press.

Braun, V., & Clarke, V. (2021). *Thematic Analysis: A Practical Guide to Understanding and Doing*. SAGE Publications. https://us.sagepub.com/en-us/nam/thematic-analysis/book248481

Cornelissen, J. P., & Werner, M. D. (2014). Putting Framing in Perspective: A Review of Framing and Frame Analysis across the Management and Organizational Literature. *Academy of Management Annals*, *8*(1), 181–235.

De Cock, C., Nyberg, D., & Wright, C. (2021). Disrupting climate change futures: Conceptual tools for lost histories. *Organization*, *28*(3), 468–482.

Dobusch, L., & Schoeneborn, D. (2015). Fluidity, Identity, and Organizationality: The Communicative Constitution of Anonymous. *Journal of Management Studies*, *52*(8), 1005–1035.

dpa. (2023). Scholz über Anklebe-Aktionen: „Finde das völlig bekloppt" [Sueddeutsche Zeitung]. https://www.sueddeutsche.de/politik/letzte-generation-kriminelle-vereinigung-justizsenatorin-berlin-pruefung-1.5874146

dpa & epd. (2023). Habeck: Gruppe "Letzte Generation" schadet Klimaschutz [ZDF]. https://www.zdf.de/uri/1b1cdd09-87ce-4ec3-8ead-9aecf85656fa

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, *14*(4), 532–550.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory Building From Cases: Opportunities And Challenges. *Academy of Management Journal*, *50*(1), 25–32.

Emirbayer, M., & Mische, A. (1998). What Is Agency? *American Journal of Sociology*, *103*(4), 962–1023.

Gamson, W. A. (1975). Review of Frame Analysis: An Essay on the Organization of Experience. *Contemporary Sociology*, *4*(6), 603–607.

Giorgi, S. (2017). The Mind and Heart of Resonance: The Role of Cognition and Emotions in Frame Effectiveness. *Journal of Management Studies*, *54*(5), 711–738.

Giorgi, S., & Weber, K. (2015). Marks of Distinction: Framing and Audience Appreciation in the Context of Investment Advice. *Administrative Science Quarterly*, *60*(2), 333–367.

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

Granqvist, N., & Gustafsson, R. (2016). Temporal Institutional Work. *Academy of Management Journal*, *59*(3), 1009–1035.

Gray, B., Purdy, J. M., & Ansari, S. ((2015). From Interactions to Institutions: Microprocesses of Framing and Mechanisms for the Structuring of Institutional Fields. *Academy of Management Review*, *40*(1), 115–143.

Grodal, S., Anteby, M., & Holm, A. L. (2021). Achieving Rigor in Qualitative Analysis: The Role of Active Categorization in Theory Building. *Academy of Management Review*, *46*(3), 591–612.

Gümüsay, A. A., & Reinecke, J. (2022). Researching for Desirable Futures: From Real Utopias to Imagining Alternatives. *Journal of Management Studies*, *59*(1), 236–242.

Guterres, A. (2022). Secretary-General's video message to the Petersberg Dialogue | United Nations Secretary-General. https://www.un.org/sg/en/content/sg/statement/2022-07-18/secretary-generals-video-message-the-petersberg-dialogue

Hernes, T., & Schultz, M. (2020). Translating the Distant into the Present: How actors address distant past and future events through situated activity. *Organization Theory*, *1*(1), 1–20.

Johnston, H., & Oliver, P. E. (2000). What a Good Idea! Frames and Ideologies in Social Movement Research. *Mobilization*, (5), 1–18.

Kaiser, K. (2009). Protecting Respondent Confidentiality in Qualitative Research. *Qualitative Health Research*, *19*(11), 1632–1641.

Kaplan, S. (2008). Framing Contests: Strategy Making Under Uncertainty. *Organization Science*, *19*(5), 729–752.

Kaplan, S., & Orlikowski, W. J. (2013). Temporal Work in Strategy Making. *Organization Science*, *24*(4), 965–995.

Kim, A., Bansal, P., & Haugh, H. (2019). No Time Like the Present: How a Present Time Perspective Can Foster Sustainable Development. *Academy of Management Journal*, *62*(2), 607–634.

Laverty, K. J. (1996). Economic "Short-Termism": The Debate, the Unresolved Issues, and the Implications for Management Practice and Research. *The Academy of Management Review*, *21*(3), 825–860.

Lê, J. K. (2013). How constructions of the future shape organizational responses: Climate change and the Canadian oil sands. *Organization*, *20*(5), 722–742.

Lee, M., Ramus, T., & Vaccaro, A. (2018). From Protest to Product: Strategic Frame Brokerage in a Commercial Social Movement Organization. *Academy of Management Journal*, *61*(6), 2130–2158.

Lehmphul, K. (2016, January). Umweltbewusstsein in Deutschland [Umweltbundesamt]. https://www.umweltbundesamt.de/themen/nachhaltigkeit-strategien-internationales/umweltbewusstsein-in-deutschland

Letzte Generation. (2023a). Letzte Generation vor den Kipppunkten. https://letztegeneration.org

Letzte Generation. (2023b). Ziviler Widerstand. https://letztegeneration.org/ziviler-widerstand

Locke, K., Feldman, M., & Golden-Biddle, K. (2022). Coding Practices and Iterativity: Beyond Templates for Analyzing Qualitative Data. *Organizational Research Methods*, *25*(2), 262–284.

Marginson, D., & McAulay, L. (2008). Exploring the debate on short-termism: A theoretical and empirical analysis. *Strategic Management Journal*, *29*(3), 273–292.

McCarthy, J. D., & Zald, M. N. (1977). Resource Mobilization and Social Movements: A Partial Theory. *American Journal of Sociology*. https://doi.org/10.1086/226464

Munshi, D., Cretney, R., Kurian, P., Morrison, S. L., & Edwards, A. (2022). Culture and politics in overlapping frames for the future: Multidimensional activist organizing and communicating on climate change in Aotearoa New Zealand. *Organization*. https://doi.org/13505084221131641

Nyberg, D., Wright, C., & Kirk, J. (2020). Fracking the Future: The Temporal Portability of Frames in Political Contests. *Organization Studies*, *41*(2), 175–196.

Pörtner, H.-O., Roberts, D. C., Tignor, M. M. B., Poloczanska, E. S., Mintenbeck, K., & et al. (2022). Summary for policymakers. Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.

Reinecke, J., & Ansari, S. (2015). When Times Collide: Temporal Brokerage at the Intersection of Markets and Developments. *Academy of Management Journal*, *58*(2), 618–648.

Reinecke, J., & Ansari, S. ((2021). Microfoundations of Framing: The Interactional Production of Collective Action Frames in the Occupy Movement. *Academy of Management Journal*, *64*(2), 378–408.

Roulet, T. J. (2020). *The Power of Being Divisive: Understanding Negative Social Evaluations*. Stanford University Press.

Schultz, M., & Hernes, T. (2013). A Temporal Perspective on Organizational Identity. *Organization Science*, *24*(1), 1–21.

Shue, H. (2023). Unseen urgency: Delay as the new denial. *WIREs Climate Change*, *14*(1), e809.

Siggelkow, N. (2007). Persuasion With Case Studies. *Academy of Management Journal*, *50*(1), 20–24.

Slawinski, N., & Bansal, P. (2012). A Matter of Time: The Temporal Perspectives of Organizational Responses to Climate Change. *Organization Studies*, *33*(11), 1537–1563.

Slawinski, N., & Bansal, P. (2015). Short on Time: Intertemporal Tensions in Business Sustainability. *Organization Science*, *26*(2), 531–549.

Slawinski, N., Pinkse, J., Busch, T., & Banerjee, S. B. (2017). The Role of Short-Termism and Uncertainty Avoidance in Organizational Inaction on Climate Change: A Multi-Level Framework. *Business & Society*, *56*(2), 253–282.

Snow, D. A., & Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization. *International Social Movement Research*, *1*(1), 197–217.

Snow, D. A., & Corrigall-Brown, C. (2005). Falling on Deaf Ears: Confronting the Prospect of Nonresonant Frames. In *Rhyming Hope and History: Activists, Academics, and Social Movement Scholarship* (pp. 222–238). University of Minnesota Press.

Snow, D. A., Rochford, E. B., Worden, S. K., & Benford, R. D. (1986). Frame Alignment Processes, Micromobilization, and Movement Participation. *American Sociological Review*, *51*(4), 464–481.

Sombatpoonsiri, J. (2023). 'A lot of people still love and worship the monarchy': How polarizing frames trigger countermobilization in Thailand. *Journal of Peace Research*, *60*(1), 88–106.

Statista Research Department. (2023). Bewertung der Klimaproteste von "Letzte Generation". https://de.statista.com/statistik/daten/studie/1345160/umfrage/bewertung-der-letzte-generation-klimaproteste

Vandevoordt, R., & Fleischmann, L. (2021). Impossible Futures? The Ambivalent Temporalities of Grassroots Humanitarian Action. *Critical Sociology*, *47*(2), 187–202.

Wenzel, M., Krämer, H., Koch, J., & Reckwitz, A. (2020). Future and Organization Studies: On the rediscovery of a problematic temporal category in organizations. *Organization Studies*, *41*(10), 1441–1455.

Williams, R. H. (2004). The Cultural Contexts of Collective Action: Constraints, Opportunities, and the Symbolic Life of Social Movements. In *The Blackwell Companion to Social Movements* (pp. 91–115).

World Commission on Environment and Development. (1987). *Our Common Future*. Oxford University Press.

Wright, C., & Nyberg, D. (2017). An Inconvenient Truth: How Organizations Translate Climate Change into Business as Usual. *Academy of Management Journal*, *60*(5), 1633–1661.

Wright, C., Nyberg, D., Rickards, L., & Freund, J. (2018). Organizing in the Anthropocene. *Organization*, *25*(4), 455–471.

Yin, R. K. (1993). *Case Study Research: Design and Methods* (Rev. ed., 13th print.). SAGE Publications.

Zeng, F., Dai, J., & Javed, J. (2019). Frame alignment and environmental advocacy: The influence of NGO strategies on policy outcomes in China. *Environmental Politics*, *28*(4), 747–770.

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Sustainability in the Corporate Sector: A News Textual Analysis Approach to Measuring ESG Performance

Mohammad Izzat Raihan Imron

Technical University of Munich

Abstract

Sustainability has become a crucial factor in the financial sector, making the assessment of a company's sustainability performance essential for informed decision-making. Recognizing the media's power to shape public perception of corporate sustainability issues, this study examines the use of news analysis to evaluate companies' performance against Environmental, Social, and Governance (ESG) criteria. Leveraging OpenAI's models, this research parses unstructured data within news articles and introduces a machine learning pipeline to score companies' ESG performance based on their media representation. The study uncovers several key findings: firstly, it demonstrates that a less costly, fine-tuned model can surpass the zero-shot capabilities of a more expensive model in classifying ESG content. Secondly, it identifies discrepancies in media coverage across industries, leading to unequal assessments of companies. Thirdly, it reveals a media tendency to underreport companies' environmental efforts. Finally, the study highlights areas where companies face media criticism, suggesting potential improvements in their ESG practices. These insights contribute to the understanding of how machine learning can assist in the critical evaluation of sustainability in the business domain.

Keywords: ESG; machine learning; natural language processing; news; NLP; sustainability

1. Introduction

Over the past decade, sustainability has become a prominent topic in the corporate finance world. The burgeoning interest in integrating values into the investment decision-making process underscores its significance. Although there is still a lack of clarity about the meaning, and it has been referred to by various names (Starks, 2023), there has been a remarkable shift in investors' focus from short-term profits to the consideration of long-term impacts and the non-financial risks associated with the ventures (Krappel et al., 2021).

Sustainability factors, including environmental stewardship, co-prosperity, ethical responsibility, and the preservation of a company's economic performance, are gaining increasing significance (J. Lee & Kim, 2023). As a result, environmental, social, and governance (ESG) practices have surfaced as a novel corporate management paradigm. The ESG concept was first introduced in 'Who Cares Wins,' published by the United Nations Global Compact in 2004, where it integrated the dimensions of ESG and the Principles for Responsible Investment (PRI) made the concepts become more popular in the following years (J. Lee & Kim, 2023; United Nations Global Compact, 2004).

According to PwC's Asset and Wealth Management Revolution 2022 Report, ESG-related assets under management (AuM) were at US$18.4 trillion in 2021 and are expected to reach US$33.9 trillion in 2026, making up 21.5% of assets under management (PwC, 2022). This substantial growth is more than a financial trend. It reflects a meaningful change in investor attitudes towards sustainable investing. Hartzmark and Sussman (2019) affirm this perspective, and they argue that sustainability is a valued attribute among investors, influencing their investment behaviors. Investors are navigating their decisions not solely based on financial returns but are also influenced by emotional and ethical factors that resonate with their moral values (Hartzmark & Sussman, 2019). Empirical evidence indicates that investors react more positively than usual to the latest positive news

DOI: https://doi.org/10.5282/jums/v10i2pp369-401

for a company with superior sustainability performance, subsequently elevating the company's stock price (Barka et al., 2023; Leite & Uysal, 2023). Consequently, a firm's commitment to sustainability strengthens its reputation while simultaneously enhancing investors' perceptions.

Assessing a company's sustainability performance has become an integral aspect of investment decision-making, intertwining financial prospects with value considerations. A collective of more than five thousand investors, managing a cumulative of over US$100 trillion in assets, have pledged to incorporate ESG data into their investment decision-making processes (Principles for Responsible Investment (PRI), 2023). Investors often resort to ESG scores or ratings provided by various institutions to measure a firm's sustainability quality. Organizations like MSCI, Sustainalytics, and Refinitiv offer ESG ratings that evaluate companies based on various sustainability metrics. However, it is critical to note that ESG ratings, despite their growing usage and influence, are not a perfect method for determining a company's commitment to or effectiveness in sustainability efforts.

A number of critiques emphasize that ESG ratings come with their own specific challenges and shortcomings. Ilango (2023), for example, argues that the system has the potential to disrupt financial market stability if it remains unregulated. A Financial Times article by Allen (2018) underscores that ESG ratings might be overly simplified and subject to subjective measurement. Chatterji et al. (2016) show a notable inconsistency in social ratings provided by six renowned raters, with disparities persisting even after adjusting for variances in the definition of Corporate Social Responsibility (CSR) among them, suggesting that the ratings might have low validity. Berg et al. (2022) describe further that the cause of the discrepancy is not just different definitions but a fundamental disagreement regarding the underlying data. The issue of inconsistency is prevalent not only between the raters but also within a single rater across different years. Results from Berg et al. (2020) reveal that there have been considerable adjustments to the historical evaluations by Refinitiv ESG (formerly ASSET4).

ESG rating providers, in an effort to be different, are likely to continue using their own methodologies and metrics, intentionally creating discrepancies in ratings (Brackley et al., 2022). Coupled with a lack of transparency in the disclosure of rating criteria and calculation methods, raises concerns about the reliable use of ESG ratings (J. Lee & Kim, 2023). It is important to directly gather ESG-related information from various data sources to effectively obtain objective information on a company's ESG efforts (J. Lee & Kim, 2023). However, collecting unstructured, scattered, and vast information can be arduous and financially demanding. To alleviate this, Natural Language Processing (NLP) can serve as a supportive tool to help streamline this process.

Several studies have explored the use of Natural Language Processing (NLP) to extract information from text, particularly in financial and environmental contexts. In the broad financial context, A. H. Huang et al. (2023) introduce FinBERT, a language model tailored for finance-related text. In a more niche context, Webersinke et al. (2021) develop ClimateBERT, a model specializing in climate-related texts. In the ESG realm, various recent studies such as those by H. Kang and Kim (2022), J. Lee and Kim (2023), Luccioni et al. (2020), Mehra et al. (2022), and Polignano et al. (2022) have utilized diverse NLP methods to analyze companies' sustainability reports, highlighting a growing interest in leveraging this tool to comprehend and evaluate ESG performances.

This study introduces a different approach to understanding companies' ESG performance by employing a third-person perspective, particularly through news sources. A modest amount of research has explored applying NLP to news text in the context of ESG. For instance, Nugent et al. (2020) introduce BERT_{RNA}, which is capable of performing multi-class ESG controversy classification tasks. Additionally, Fischbach et al. (2022) develop ESG-Miner, based on the BERT model, designed to classify the headlines of news articles from Twitter in the ESG context. A more detailed description of these studies is explained in Section 2.

According to the agenda-setting theory, the news media plays a crucial role in directing public attention to significant issues and shaping perceptions and knowledge about those topics, influencing how the public views and understands them (McCombs & Reynolds, 2002). It is also true in the ESG context that mass media, including news, has the ability to drive people's perceptions regarding the firms' environmental performance and other ESG issues (N. Brown & Deegan, 1998; Hammami & Hendijani Zadeh, 2019). The results of Serafeim and Yoon (2021) also indicate a link between positive ESG news and an increase in stock prices, and conversely, negative ESG news is connected with a drop in stock prices. In addition, Hammami and Hendijani Zadeh (2019) also find that public exposure from the news is one of the main drivers of ESG transparency that can help mitigate the information asymmetry between the companies and their stakeholders.

Considering the considerable influence of news, this study seeks to investigate the analysis of news as a means to evaluate the ESG performance of companies. However, due to the fact that not all news articles pertain to the subject of ESG, a careful selection must be made. In this context, I delve into the potential of Generative Pre-trained Transformer (GPT) models to perform multi-class classification tasks. In 2020, OpenAI released the largest language model at that time, GPT-3, with 175 billion trainable parameters (T. B. Brown et al., 2020). The model has been trained from various sources, including CommonCrawl, with a total of 499 billion tokens (T. B. Brown et al., 2020). The paper shows that the GPT-3 model performed better on the LAMBADA dataset test than the state-of-the-art of that time, Turing-NLG, with an accuracy of 72.5% (Zero-Shot) to 86.4% (Few-Shot).

In late 2022, OpenAI introduced ChatGPT, marking a significant stride in the progression of NLP and Artificial Intelligence (AI). Originating from the GPT-3.5 series, ChatGPT represents a refinement and advancement beyond its predecessor, GPT-3 (OpenAI, 2022). Not only does it generate responses that mimic human interaction, making it suitable for conversational applications, but it has also showcased remarkable capacities in technical fields. Notably, ChatGPT was reportedly capable of securing an entry-level software engineer position at Google and successfully navigating exams for law and business schools (Elias, 2023; Kelly, 2023). The substantial growth of ChatGPT is evidenced by its rapid user adoption, exceeding 100 million active users within two months post-launch and thus joining the ranks of the fastest-growing consumer applications in history, second only to Threads by Meta in the speed of user acquisition (Gordon, 2023).

Motivated by the surging interest in OpenAI's GPT models, I am drawn toward leveraging them to dissect and analyze the expansive and unstructured data within the news. This paper introduces a streamlined machine learning pipeline, employing a variety of models to ensure a smooth and intuitive measurement process. The initial phase involves collating the titles and leading paragraphs of news articles from the New York Times. Subsequently, named-entity recognition (NER) is implemented to isolate news that pertains to specific companies. Following this step, extract and categorize the news articles by using the text classifier. Lastly, sentiment analysis is applied to assign an ESG score to each company within each category. The main contributions of this thesis encompass:

1. A manually labeled, fine-grained ESG dataset comprising 4,500 news items.
2. An exploration into the capabilities of GPT models, evaluating their proficiency in performing classification tasks specifically within the ESG context.
3. A fine-tuned model based on the GPT model to enhance its functionalities in performing multi-class classification tasks in the ESG context.
4. A proposal of a systematic machine learning pipeline, which has been designed to assess companies' ESG performance utilizing information derived from news articles.
5. A thorough analysis of several companies' ESG performance, drawing from news articles to illuminate their activities and impacts in this domain.

The results of this study reveal that a fine-tuned model, which is less expensive and smaller in scale, can outperform the zero-shot prompting capabilities of a larger, more costly model in ESG classification tasks. Additionally, the machine learning pipeline used in this research demonstrates an ability to grasp the nuances and assess the sustainability performance of companies, as depicted in media reports. The analysis indicates that media attention varies across industries, resulting in a more thorough assessment of some companies over others. It is also worth mentioning that the environmental initiatives undertaken by companies are often underrepresented in media coverage, with a greater focus observed on governance and social aspects. Nevertheless, the findings from the machine learning pipeline also uncover areas where companies receive criticism from the media, pointing to opportunities for improvement in specific aspects.

The remainder of this paper is organized as follows. Section 2 provides an overview of the existing research on large language models and their applications in the ESG field. Section 3 outlines the research methods employed in this study. Section 4 portrays the findings of the study and the outcomes of the study, focusing on how companies are scored using the proposed approach. Section 5 goes into a detailed discussion of the findings and explains the limitations of the study. Finally, Section 6 concludes the paper with the conclusions.

2. Literature Review

2.1. Large Language Models

Over the past few years, there has been a significant shift in the field of NLP with the introduction of large language models (LLMs). These models are pre-trained foundational models that are self-supervised and are trained on extensive text datasets (Sejnowski, 2023; Sun et al., 2023). These LLMs, as described by A. H. Huang et al. (2023), use contextualized embeddings[1] that can represent a word with various vectors based on its context. This approach is a notable progression from the earlier word embedding models like word2vec and GloVe, where each word was tied to a single, fixed vector without taking into account the surrounding text (A. H. Huang et al., 2023). In contrast, LLMs adjust the vector representation of words based on their immediate textual context, making them especially effective in interpreting homographs[2] (A. H. Huang et al., 2023). As a result, LLMs offer more nuanced and accurate interpretations when analyzing texts compared to their predecessors.

Although the history of language models can be dated back more than 100 years, with early concepts being explored by Markov and Shannon based on probability theory (H. Li, 2022), a game-changing development came with the introduction of the transformer model by Vaswani et al. (2017). Originally designed for translation tasks, the transformer's defining idea is its self-attention mechanism, which allows the model to weigh the importance of various parts of an input sequence when producing an output (Douglas, 2023). Owing to its exceptional capabilities in language representation, transformer architecture has become foundational in almost all of today's pre-trained language models (H. Li, 2022).

The transformer architecture quickly gained popularity among researchers and split into three main categories: bidirectional, unidirectional, and sequence-to-sequence (Douglas, 2023; H. Li, 2022). First, there is a bidirectional model

[1] Embeddings are mathematical representations of objects or values, such as text, images, and audio, transformed into vectors based on their characteristics, attributes, and categorical associations.

[2] A homograph is a word that is spelled identically to another word yet differs in meaning.

using the encoder-only architecture like BERT (short for Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). BERT learns by hiding random words in a sentence and guessing them using the surrounding words for clues, which helps it get better results (Devlin et al., 2018; Douglas, 2023). Then, there is the unidirectional type, like GPT, that uses the decoder-only architecture (Radford et al., 2018). This architecture allows GPT to learn effectively from unlabeled data, which is helpful when there is no labeled data available (Radford et al., 2018). Finally, there is the sequence-to-sequence model, which uses both an encoder and a decoder. Some examples of this are BART, which stands for Bidirectional and Auto-Regressive Transformers (Lewis et al., 2019) and T5 (Raffel et al., 2019).

There are a lot of applications of LLMs across various industries and use cases. In healthcare, LLMs help answer medical questions. For example, Venigalla et al. (2022) create BioMedLM, a GPT model trained on biomedical data. Singhal et al. (2022, 2023) develop Med-PaLM and Med-PaLM 2, based on a pathways language model (PaLM), and Y. Li et al. (2023) make ChatDoctor, refining the large language model meta-AI (LLaMA) with a massive collection of dialogues between patients and doctors. In finance, Wu et al. (2023) introduce BloombergGPT, an LLM that is trained on a wide spectrum of financial data. In the legal world, Chalkidis et al. (2019) use LLMs to predict case outcomes from the European Court of Human Rights (ECHR), and Peric et al. (2020) modified a GPT-2 model to help write legal opinions.

Another fascinating area of research about LLMs is prompt engineering, which focuses on how to communicate with these models to get the best results effectively. While there are numerous plausible techniques for prompting, the field of prompt engineering still necessitates extensive experimentation to understand and perfect these methods (Kaddour et al., 2023). There is limited theoretical knowledge about why certain ways of wording a task work better, except for the fact that they produce superior practical outcomes (Kaddour et al., 2023). Various approaches have been investigated in prompt engineering, such as in-context learning, which enables LLMs to pick up new tasks without altering any parameters. There is also multi-turn prompting, which links a series of prompts and responses in a sequential manner, and chain-of-thought (CoT) prompting, which builds few-shot prompts through a sequence of intermediate reasoning steps leading to the final result, among other methods.

This study explores a prompt engineering technique by Sun et al. (2023). They introduce the ‘Clue And Reasoning Prompting’ (CARP) method, which adopts a step-by-step reasoning process designed for the intricate language patterns seen in text classification. It starts by nudging the language models to identify key clues. Next, it requests detailed reasoning, which ultimately helps in making the final classification. This technique impressively set a new state-of-the-art (SOTA) performance on four out of the five most common text classification benchmarks. Prompt engineering proves to be a valuable strategy when faced with constraints like limited datasets and computational resources for training or fine-tuning, managing to deliver results that can compete with more resource-intensive methods.

2.2. Large Language Models for ESG

Machine learning, in general, has been a key tool in ESG analysis and research. It has been used in a variety of ways, from shaping trading strategies to assessing risks. For instance, Chen and Liu (2020) use machine learning to understand a company’s ESG premium and capture the ESG alpha to build an automatic trading strategy. Impressively, their approach is proven to outperform the NASDAQ-100 and S&P 500 indexes over a ten-year test period. On another front, Nguyen et al. (2020) tap into machine learning to predict a company’s future emissions, which can help policymakers decide where to focus their efforts. Guo et al. (2020) look at ESG news to predict the volatility of stock prices.

Krappel et al. (2021) provide a heterogeneous ensemble to predict the ESG ratings of companies even if the company does not disclose its sustainability report. Furthermore, Polignano et al. (2022) and H. Kang and Kim (2022) utilize NLP to extract information from corporate sustainability reports and propose a machine learning pipeline to analyze the sustainability performance of companies. Incorporating news into the model, Borms et al. (2021) share a method to summarize ESG-related news using word patterns. They create a comparison using news-based ESG signals with the scores from an external data provider. Goutte et al. (2023) delve into how news sentiment, especially when it concerns ESG topics, can predict stock returns using data provided by the Global Database of Events, Languages, and Tone (GDELT) project.

The development of LLMs then further improves the nuance and comprehensiveness of textual data from different sources, such as reports or news articles. Different applications and improvements from the foundational models (e.g., BERT & GPT) have been explored by multiple researchers. Using a question-and-answering approach, Luccioni et al. (2020) develop ClimateQA using the BERT model, which assists in the analysis of financial reports regarding climate-relevant sections. Webersinke et al. (2021) build a language model, ClimateBERT, that was pre-trained on a dataset of over two million climate-related paragraphs, which improved the predictive performance in three climate-related tasks: text classification, risk and opportunity analysis, and fact-checking climate-related claims. This model was further applied to assess the disclosure of climate-related financial risks by Bingler et al. (2022).

Many researchers delve into topic modeling by leveraging LLM to extract ESG information. Mehra et al. (2022), for example, successfully fine-tune BERT using ESG texts from guides, case studies, blogs, reports, and other knowledge bases from the Accounting for Sustainability project. The fine-tuned model, called ESGBERT, is then applied to do classification tasks on the 10-Q filings of companies. Similarly, J. Lee and Kim (2023) make use of a BERT-based model to create a classification model based on Korean companies’ sustainability reports. Their model can sort sentences into four ESG categories, including those that are not relevant. They also tested their model on other types of data, such as news articles, to demonstrate its adaptability.

Aue et al. (2022) introduce a technique to predict ESG ratings by looking at patterns over time based on news articles. To identify news related to ESG and figure out the sentiment, they utilize BERT in their method. Fischbach et al. (2022) develop a model based on BERT to sort news headlines into three ESG categories and label irrelevant news, called ESG-Miner. While the model is good at identifying headlines related to the environment, its accuracy in determining ESG relevance could use some refinement. Nevertheless, specialized ESG classifier models are still developing, especially for fine-grained ESG topics. Tackling fine-grained ESG topics presents a significant challenge largely due to the absence of a universally accepted standard to define classes under the three ESG pillars. ESG rating institutions, such as MSCI, Refinitiv, and Sustainalytics, employ different methodologies to define the fine-grained classes.

The opportunity to further explore text classification within this domain remains substantial, considering there are only a handful of available models, and they present differing categorizations for the fine-grained classification. For instance, Nugent et al. (2020) pre-train a BERT model on financial news articles from the Reuters News Archive. They then used this trained model to sort texts into 20 different ESG topics and also to label them based on the United Nations Sustainable Development Goals (UN SDGs). H. Lee et al. (2023) use a different set of classes consisting of 35 ESG key issues by MSCI ESG rating guidelines. In their study, they utilize various models based on the BERT model to do the classification tasks.

Interestingly, the categories used by Nugent et al. (2020) and H. Lee et al. (2023) are divergent from those introduced by A. H. Huang et al. (2023). In a later study, A. H. Huang et al. (2023) create FinBERT, which is a version of BERT designed specifically for financial topics. They further provide several models for different purposes, such as FinBERT-tone for sentiment analysis of financial texts, FinBERT-ESG for ESG topic sorting, and FinBERT-ESG-9-categories for fine-grained ESG classification. In their detailed classification model, A. H. Huang et al. (2023) use 14,000 manually annotated sentences from ESG reports and annual reports to fine-tune FinBERT. This model can sort texts into nine distinct ESG categories: Climate Change, Natural Capital, Pollution and Waste, Human Capital, Product Liability, Community Relations, Corporate Governance, Business Ethics and Values, and Non-ESG.

Throughout this chapter, it is evident that a significant portion of research tends to favor the use of the BERT model for various applications, especially for fine-grained ESG classification. Despite this predominant focus on BERT, there exists a promising avenue of exploration in understanding and leveraging GPT models, particularly in the context of ESG topic identification from news articles. My aim is to contribute meaningful insights and perspectives on how GPT models can also be a good choice for this kind of task and provide a balanced view of the capabilities of different language models in handling the extraction of ESG-related information from textual data.

3. Methodology

In this section, I describe the proposed streamlined machine learning pipeline for measuring the sustainability performance of a company. It starts with the data collection from the news outlet to scoring the performance based on news sentiment analysis. I am using available machine learning models for NER tasks and sentiment analysis tasks, while I am using a fine-tuned GPT model for the fine-grained ESG classification tasks. The pipeline is shown in Figure 1. The pipeline is available in the Python language because there are models and libraries available to use for NLP tasks in this language.

3.1. Data Collection

This study extracts news articles from the New York Times, utilizing the Archive API. Recognized globally and based in New York City, the New York Times is a daily newspaper renowned for its wide-reaching influence and significant subscriber base, with over nine million subscribers worldwide (The New York Times Company, 2023). Besides its reputation and credibility to ensure the reliability and validity of the news data, the New York Times is chosen because it provides an API that allows for systematic data retrieval, unlike many other news outlets, which is pivotal for consistent and reproducible research practices. Furthermore, the availability of this API service free of charge aligns well with budgetary constraints and enables extensive data extraction without incurring additional costs. A collection of 1.8 million news articles, spanning from 2003 to 2022, forms the basis of data to be analyzed further. The chosen date range ensures that a substantial amount of data is available for analysis, providing a robust dataset to identify and analyze trends or patterns related to ESG performance over an extended period.

In alignment with the study's objectives, specific elements of the metadata—specifically headlines, lead paragraphs, web URLs, keywords, and published dates—were extracted for further analysis. The analytical focus of this study targets the combined text of headlines and lead paragraphs. Integrating both headline and lead paragraph ensures a coherent message is conveyed, as a headline or a lead paragraph alone often lacks the necessary context to fully understand the unfolding events. Keywords serve as a preliminary filter to isolate news pertaining to specific companies, streamlining the process of data analysis. Meanwhile, the publication dates provide a temporal framework to understand the sequence and timing of events.

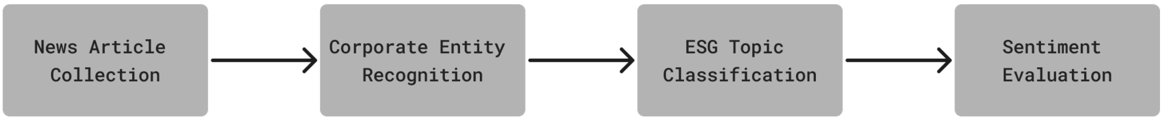


Figure 1: Machine Learning Pipeline for ESG Assessment

3.2. Company Selection and Named-Entity Recognition

In order to see the pipeline in action, this research selects eight companies to be examined. There are three criteria that are put in place to select the companies: market capitalization, sector representation, and news volume. All chosen companies are recognized as some of the largest entities within the S&P 500 based on their market capitalization to ensure it encompasses firms with significant impact and influence in the market. The companies were also selected to provide representation from diverse market sectors, giving breadth to the study by allowing for insights to be derived across different industries. Finally, companies with a more substantial presence in the news were prioritized, as a higher volume of available data enables a more in-depth and viable analysis.

The eight selected companies come from three different sectors: the technology/communication services sector (technology), consumer staples, and healthcare. Apple, Microsoft, Google, and Meta were chosen from the technology sector due to their gigantic market capitalization. Apple, Microsoft, and Google are the biggest companies within the S&P 500, with a total of over US$ 7 trillion in market capitalization (Johnston, 2022). Although Meta is not in the top four, it is included due to its significant presence in news articles and its operational similarities with the other three technology giants. It was challenging to select companies in the consumer staples and healthcare sectors because the amount of news for these sectors is not as many as in the technology sector. Therefore, I decided to select Coca-Cola and Pepsi for the consumer staples industry and Pfizer and Johnson & Johnson (J&J) for the healthcare industry because they tend to be more frequently featured in the news compared to other companies in their respective field.

It is worth mentioning that the list of keywords from the New York Times can sometimes include items that may not be directly related to the news article in question. Take the keyword "Apple Inc." as an example, and there could be instances where this is included even if the news piece primarily discusses only Microsoft's activities. To avoid this kind of impreciseness and to ensure the relevance of the company to the article, it is important to deploy a mechanism for accurate company name detection. In addressing this challenge, I utilize NER, which is notable for its capabilities to identify and categorize entities within the text into specified categories, including organizations, persons, and locations, among others. By applying NER, it is possible to extract the company of interest from the text, ensuring that the company is relevant to the article.

This research applies NER via the spaCy[3] library, specifically using the 'en_core_web_md' package, version 3.6.0, which is the medium-sized English model proficiently trained on web text, including news articles. This model is chosen for its capability to swiftly perform statistical entity recognition, such as identifying various types of named and numeric entities within text. The names of firms can be selectively extracted by targeting the label recognized as 'ORG'. Following the extraction, a string-matching method is applied to further refine the collection process based on the company of interest.

3.3. Text Classification

This study emphasizes the exploration of the capabilities of GPT models, particularly in executing ESG classification tasks. Therefore, unlike other machine learning tasks, including NER and sentiment analysis, this machine learning pipeline uses the results from this exploration. In this section, I outline the methods employed to investigate their potential to achieve fine-grained ESG classifications, such as the definitions of each class used in the model, the dataset for training and validating the model, and the procedure for fine-tuning the GPT models and validating zero-shot classification.

3.3.1. Class Definitions

The ESG topics used in this paper are inspired by MSCI ESG Key Issues (MSCI, 2023b) with minor adjustments. MSCI ESG Key Issues were selected as the basis for the class definitions in this study mostly due to its credibility, which makes it used by the majority of ESG exchange-traded funds (ETFs) (Hirai et al., 2021). The eight ESG topics are *Climate Change, Resource Stewardship, Environmental Opportunities, Human Capital, Product Liability, Social Opportunities, Corporate Governance,* and *Business Ethics.* One other class is *Non-ESG* to flags irrelevant news articles. The definition and example of each class are described as follows.

Climate Change

Pertains to news articles discussing topics such as carbon emissions reduction initiatives, the carbon footprint of products, climate change vulnerabilities, and financial initiatives or instruments designed to mitigate the impact of climate change. This can include policies, new technologies, or corporate strategies targeting climate change. The examples are as follows.

[3] From Honnibal and Montani (2017)

- *CEO of UK-based energy supplier Drax shares how the company, formerly 100% reliant on coal, reduced its carbon emissions by 85%. The company now has ambitions to not just be carbon neutral, but carbon negative.*
- *ExxonMobil Formally Joins The Net-Zero By 2050 Bandwagon. U.S. energy giant ExxonMobil announced Tuesday a formalized plan to cut its scope 1 and scope 2 carbon emissions to "net-zero" by the year 2050.*

Resource Stewardship

Involves articles highlighting how companies manage natural resources and waste. It encompasses a range of issues, including but not limited to water conservation, biodiversity, sustainable land use, responsible raw material sourcing, toxic emissions reduction, and effective waste management, including electronic waste. The examples are as follows.

- *Starbucks to Offer Reusable Cups in All EMEA Stores by 2025. Starbucks Corp will offer reusable cups in stores across Europe, the Middle East and Africa by 2025 in an effort to reduce the amount of single-use waste heading to landfill.*
- *The Coca-Cola Company And The Ocean Cleanup Join Forces To Clean Up 15 Of The World's Most Polluting Rivers Of Plastic Waste. The Coca Cola Company and The Ocean Clean-Up project have announced they will be collaborating on a ground-breaking partnership to clean up some of the world's worst polluting rivers - and collect plastic waste which can be recycled to make new bottles.*

Environmental Opportunities

Includes articles that focus on the potential opportunities arising from environmental conservation efforts. It covers green technology innovations, renewable energy initiatives, sustainable building projects, and financial investments targeting environmental sustainability. The examples are as follows.

- *China Clean Energy Giants Unveil World's Largest Wind Turbines. Ming Yang Smart Energy Group Ltd. unveiled the world's largest wind turbine, an offshore behemoth whose more than 140-meter-long blades will sweep across an area larger than nine soccer pitches.*
- *Honda Recommits To Fuel Cells As It Looks For New Markets. Honda is planning to offer up its new generation fuel cell systems for commercial vehicles, construction equipment and stationary power systems beginning in 2025.*

Human Capital

This class encompasses news items related to labor management, employee welfare, and workforce development. It can include articles about health & safety protocols, human capital development programs, and supply chain labor standards, including diversity, equity, and inclusion initiatives in the workplace. The examples are as follows.

- *Female employees file class-action discrimination suit against Black News Channel. Thirteen women who worked at Black News Channel say in lawsuit they were paid less than men and disciplined for being too aggressive in the workplace.*
- *Amazon warehouse workers suffer muscle and joint injuries at a rate 4 times higher than industry peers. Amazon workers are four times as likely to incur strains, sprains and other repetitive stress injuries as workers in non-Amazon warehouses.*

Product Liability

Relates to articles discussing aspects of product safety and quality, including chemical safety and consumer financial protection. It includes topics such as privacy and data security issues and socially responsible investment, which might impact product liability. The examples are as follows.

- *Google is facing a lawsuit after a privacy flaw in its contact tracing tech exposed Android users' data to third-party apps. The lawsuit alleges that Google exposed participants' private personal and medical information when they opted into using contact tracing apps.*
- *Nestle recalls more than 760,000 pounds of Hot Pockets because they might contain bits of plastic and glass. Four people contacted Nestlé when they found "extraneous material" in their Hot Pockets, the USDA said.*

Social Opportunities

This category covers articles focusing on the societal benefits generated through corporate initiatives. It includes news on community financing, enhancing healthcare access, nutrition and health opportunities, educational initiatives, and investments aimed at social development. The examples are as follows.

- *HBCUs Team Up With Wells Fargo To Improve Financial Wellness For College Students Of Color. Wells Fargo is investing $5.6 Million in a financial literacy and wellness program designed for college students of colorant their surrounding communities.*
- *DoorDash Establishes Grant Program For Women- And BIPOC-Owned Restaurants Disproportionately Affected*

By Covid. DoorDash's Main Street Strong Accelerator provides financial support and specialized educational resources specifically to women-, immigrant- and people of color-owned businesses that have been disproportionally impacted by the Covid-19 pandemic.

Corporate Governance

Deals with articles related to the structural and strategic management aspects of corporations. This includes topics such as ownership and control dynamics, board composition and performance, executive remuneration, accounting transparency, and significant executive team changes, including the hiring or resignation of C-level executives and directors. The examples are as follows.

- *Black employees are questioning Peloton about their pay, as the fitness giant's CEO pulls in a $17.8 million compensation package. Black employees are questioning Peloton execs about pay, while the CEO earned $17.8 million last fiscal year and the median employee earned $56,084.*
- *Discord Adds Ex-Netflix, Block Executives To Board Ahead Of Possible IPO. The two C-suiters' experience at public companies is the latest indication that Discord intends to go public soon.*

Business Ethics

Encompasses articles on the ethical considerations of business operations. It includes subjects such as tax transparency, anti-corruption measures, fraud prevention, and adherence to ethical business practices and regulations. The examples are as follows.

- *Former Netflix executive convicted of fraud after orchestrating more than $500,000 in bribes and kickbacks. As Netflix's IT chief, Michael Kail approved contracts with tech startups in exchange for kickbacks, even buying a house with the funds, a jury found.*
- *McDonald's to pay France $1.3 billion in tax fraud case. McDonald's France and related companies have agreed to pay $1.3 billion to the French state to settle a case in which the fast-food giant was accused of vast tax evasion*

Non-ESG

This class is for articles that do not fit into any of the above ESG categories. It can include a wide range of topics not directly related to environmental, social, and governance issues. The examples are as follows.

- *McDonald's temporarily removes the Chicken Big Mac from menus, saying that it's struggling to keep up with demand. The limited-edition burger was surprisingly popular, which led to its removal from UK restaurant menus while stocks are replenished.*
- *Ferrari Sparkling Wine Becomes The Official Toast Of Formula One. Italian sparkling wine will be at the podium.*

3.3.2. Dataset

For training and validating purposes, I used a different source of news articles than what has been described in Chapter 3.1 to minimize bias during the later application. In this step, I used the dataset from the GDELT project as the source of my news data. This project monitors news media from all over the world in over 100 languages (The GDELT Project, n.d.-a). It is considered one of the largest and most comprehensive open databases created by Kalev Leetaru in a research collaboration with a lot of institutions, including, among others, Google, JSTOR, and the Internet Archive (The GDELT Project, n.d.-b). It now consists of over a quarter-billion event records in over 300 categories covering events from 1979 to the present (The GDELT Project, n.d.-b).

GDELT offers different datasets for different purposes, such as GDELT Event Database, GDELT Global Knowledge Graph, and GDELT Article List. The most comprehensive database is the GDELT Event Database, which contains historical data from 1979 to date. However, for the purpose of this thesis, this database does not provide information on similar data as the lead paragraph from the New York Times API or a summary of the news articles. Only the GDELT Article List database provides this kind of information. Unfortunately, this database does not incorporate historical news, as it only started to collect the data in January 2020. Another challenge to using this database is the missing data from the news summary, which was only available in late 2020. This poses an obstacle where the usable data is only in a limited timeframe.

I have gathered news from several prominent media outlets to address the timeframe limitation and guarantee a sufficient volume of articles. These include the Wall Street Journal, Los Angeles Times, Bloomberg, Business Insider, the Economist, Forbes, the Washington Post, and ESG News. The collection spans from 2021 to April 2023, containing a total of 698,433 news pieces. These articles will then be selected and manually annotated to each ESG topic defined in the previous section. Nevertheless, the dataset has a high level of duplicate news articles, which needs to be removed. Moreover, there are a lot of irrelevant news articles because the dataset contains articles for all topics, including sports, arts, and others.

The duplicated news articles and unrelated ones are removed. Precisely, I aimed to include articles from the business, science, technology, and environment if it is identifiable through their URL structures. Given the vast amount of data, manually categorizing each article is daunting. To streamline this process, I initially employed the FinBERT ESG 9 Categories model by A. H. Huang et al. (2023). This model was particularly useful for basic categories like environmental, social, governance, and non-ESG. However, it was not foolproof. Some articles were not correctly sorted, and given our unique classification requirements, I still had to manually review and annotate many pieces to ensure their accurate categorization. This blend of automated and hands-on approaches allowed me to generate a more reliable dataset while saving time.

At first, I planned to add the stakeholder opposition topic and split the resource stewardship topic into natural capital and pollution & waste. However, finding news articles related to these topics in the dataset was challenging. Therefore, I excluded the stakeholder opposition and combined both the natural capital and pollution & waste topics into a broader resource stewardship category. Ultimately, I generated a manually annotated dataset with 4,500 articles in total, ensuring 500 articles for each topic. I then divided this dataset into training and validation sets using an 80:20 split.

3.3.3. Methods

This study investigates GPT models' capabilities in handling multi-class classification tasks, focusing on two distinct approaches: fine-tuning and zero-shot prompting. While fine-tuning refines a previously trained model for a particular task, zero-shot prompting relies on the model's inherent knowledge, eliminating the need for further training. GPT models, as large language models, possess a vast amount of knowledge, allowing them to make predictions, even for unfamiliar tasks. The fine-tuning approach is advantageous for models with a narrower knowledge base and is also cost-effective. Conversely, zero-shot prompting proves valuable when working with a sparse dataset.

This paper examines different models for fine-tuning and zero-shot prompting. For the fine-tuning method, I delve into the capabilities of the 'ada' model from the GPT-3 series. Being both cost-effective and fast, 'ada' emerges as a fitting choice for fine-tuning. Meanwhile, for zero-shot prompting experiments, I employ the 'gpt-3.5-turbo-0613' from the GPT-3.5 lineup. Its vast knowledge base makes it well-suited for such experiments. These model selections aim to optimize the balance between performance and efficiency. I hope the comparison of these two distinct approaches sheds light on the potential and limitations of each method.

For the fine-tuning method, OpenAI has furnished users with tools to streamline the process. While the GPT-3 model remains a closed-source offering, direct fine-tuning on personal hardware is not feasible. Consequently, I had to rely on OpenAI's dedicated services to carry out this task. There is a prerequisite to prepare the dataset in a specific manner. Particularly, it should comprise the 'prompt,' which serves as the input, and the 'completion,' which denotes the expected output. This dataset is to be formatted in the JSONL structure.

I conducted a series of experiments to achieve the optimal configuration for the fine-tuned model. There are two key techniques: the inclusion of a prompt in the input and the application of dummy label[4] in the output. This led to four distinct combinations to test:

(a) A version without any prompt and using the original label

(b) A version without any prompt but introducing the dummy label

(c) Incorporating the prompt while sticking to the original label

(d) Incorporating the prompt but introducing the dummy label

The pairing of the original label with the dummy label I used are detailed in Table 1. To provide a clearer understanding of these configurations, I illustrate the examples in Table 2.

In the zero-shot prompting method, I followed the approach of Sun et al. (2023) called CARP. However, CARP's initial tests were focused on sentiment analysis. In this study, I aim to evaluate how CARP performs when applied to multi-class classification tasks, particularly for ESG topics. Due to the absence of a standard definition for fine-grained ESG topics and to make sure the CARP output aligned with the objective of this paper, I included the definition of each class in every prompt. This ensures the model uses the exact definitions I used during labeling and only produces classes from the list. I split the prompt into two sections: system and user. In the system content, I prompted the model to act as an ESG news articles classifier and gave it the class definitions. In the user content, I structured the prompt following the CARP method.

3.3.4. Sentiment Analysis

The concluding step of this process involves analyzing the sentiment of the news articles and categorizing them as negative, neutral, or positive. There are many machine learning models designed for sentiment analysis tasks. To find the best fit, I chose four models tailored to financial news or texts, given their similarity to the dataset in this study. Precisely, I selected the top three models trained on the 'financial_phrasebank' dataset (Malo et al., 2013) based on accuracy rankings from the Papers With Code website (Papers With Code, n.d.), along with the sentiment analysis model from FinBERT by A. H. Huang et al. (2023). The three models trained on the 'financial_phrasebank' dataset are Sigma/financial-sentiment-analysis[5] (Sigma), Farshid/bert-large-uncased-financial-phrasebank-allagree2[6]

[4] Labeling data with random strings or placeholders, often referred to as 'metasyntactic variables,' is a technique used to mitigate bias in language models.

[5] https://huggingface.co/Sigma/financial-sentiment-analysis

[6] https://huggingface.co/Farshid/bert-large-uncased-financial-phrasebank-allagree2

Table 1: Pairing Between Original and Dummy Labels

Original Label	Dummy Label
Climate Change	baz
Resource Stewardship	qux
Environmental Opportunities	roc
Human Capital	tuv
Product Liability	dap
Social Opportunities	stu
Corporate Governance	klo
Business Ethics	xya
Non-ESG	nop

Table 2: Examples of Input and Output

Model	Input (Prompt)	Output (Completion)
(a)	CEO of UK-based energy supplier Drax shares how the company, formerly 100% reliant on coal, reduced its carbon emissions by 85%. The company now has ambitions to not just be carbon neutral, but carbon negative. ->	Climate Change
(b)	CEO of UK-based energy supplier Drax shares how the company, formerly 100% reliant on coal, reduced its carbon emissions by 85%. The company now has ambitions to not just be carbon neutral, but carbon negative. ->	baz
(c)	Classify the following text into one of the following classes: [' Climate Change', ' Resource Stewardship', ' Environmental Opportunities', ' Human Capital', ' Product Liability', ' Social Opportunities', ' Corporate Governance', ' Business Ethics', ' Non-ESG'] Text:\n"'CEO of UK-based energy supplier Drax shares how the company, formerly 100% reliant on coal, reduced its carbon emissions by 85%. The company now has ambitions to not just be carbon neutral, but carbon negative."' ->	Climate Change
(d)	Classify the following text into one of the following classes: [' baz', ' qux', ' roc', ' tuv', ' dap', ' stu', ' klo', ' xya', ' nop'] Text:\n"'CEO of UK-based energy supplier Drax shares how the company, formerly 100% reliant on coal, reduced its carbon emissions by 85%. The company now has ambitions to not just be carbon neutral, but carbon negative."' ->	baz

(Farshid), and mrm8488/distilRoberta-financial-sentiment[7] (mrm8488). All models are available on HuggingFace.

In determining the optimal model from the four contenders, I did a systematic and multi-step evaluation process. To begin, a preliminary assessment was conducted by evaluating the results from each model using the news articles from the New York Times to assess Apple's performance. I manually annotated 50 news articles to serve as ground truth for this evaluation. As a result, two of the four models showcased similar accuracy rates, which are FinBERT and the mrm8488's model, making them the frontrunners. Next, to differentiate the two top-performing models, a confusion matrix was examined, as shown in Table 3.

One notable observation from this matrix is that, in instances where sentiments are polar opposites, FinBERT frequently classified articles as negative, whereas the counterpart model leaned towards a positive classification. Given this intriguing divergence in sentiment classification, a focused annotation was undertaken. A subset of 25 articles, for which FinBERT predicted a negative sentiment and the alternative model predicted a positive, were manually annotated to understand the models' behavior better. After the detailed assessment, it was deduced that the FinBERT model provided more accurate and suitable classifications for the task in this study.

The sentiment conveyed in the news articles serves as the foundation to develop the company's performance score. For quantifying this sentiment score within each category, distinct weights are allocated to each sentiment type. Specifically, a weight of -1 is designated to negative sentiment, 0 to neutral, and 1 to positive sentiment. The cumulative score is derived by multiplying the total count of articles within each sentiment category by its respective weight. Given the potential for varied data distribution across categories and companies, this raw score is then normalized. This is achieved by dividing the cumulative score by the total sentiment count, ensuring a more equitable comparison of ratings across classes and companies. The formula for the normalized is described

[7] https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

Table 3: Confusion Matrix of the FinBERT and mrm8488 Model

		FinBERT		
		Negative	Neutral	Positive
mrm8488	Negative	157	93	9
	Neutral	127	845	71
	Positive	65	125	156

in the Equation 1.

$$\text{Normalized Score} = \frac{[(-1 \times \text{number of negative sentiments}) + (1 \times \text{number of positive sentiment})]}{\text{total number of sentiments}} \quad (1)$$

4. Empirical Results

4.1. Text Classification Results

In a series of experiments of fine-tuning the ‘ada’ model, one of the GPT-3 base models, the same settings were applied to the four scenarios. I adhered to the default configurations suggested by OpenAI for various parameters, including the number of epochs, batch size, and learning rate multiplier. In particular, the default number of epochs in training is four, the batch size is set to approximately 0.2% of the number of examples in the training set (with a maximum limit of 256), and the learning rate multiplier defaults to 0.05, 0.1, or 0.2 depending on final batch size (OpenAI, n.d.). As the number of examples was identical in every scenario, sticking to the default settings ensured uniformity across all experimental setups. This uniformity is crucial for a fair comparison between different scenarios, making sure that any variations in performance are due to the treatment itself.

Each fine-tuning session in our experiments lasted around 25 minutes on average, while the cost for fine-tuning the base model was $0.0004 for every 1,000 tokens. Notably, in the first setup (a), where I used the original labels without any prompts, it took about 25 minutes and 8 seconds, and the cost for training on 793,328 tokens was $0.32. In the second setup (b), where I applied dummy labels without prompts, the process was slightly quicker and cheaper. It took around 24 minutes and 47 seconds, costing $0.31 to train on 780,528 tokens. The third and fourth setups, (c) and (d), which involved the use of prompts, did not necessarily take more time but were more expensive. For setup (c), the fine-tuning lasted about 24 minutes and 51 seconds, incurring a cost of $0.67 to train on 1,687,176 tokens. In the final setup (d), the duration was almost identical at 24 minutes and 49 seconds, but it was slightly cheaper at $0.62 for training on 1,559,176 tokens. The summary of the duration, cost, and number of tokens trained is presented in Table 4.

In the experiment where I employed the zero-shot prompting method with the CARP approach, it was sufficient to assess the outcomes using the same validation set as I used for the fine-tuning. This approach focused solely on the validation phase, eliminating the need to track the duration since there was no training phase involved. In general, for the GPT-3.5 turbo model utilized in this particular experiment, the cost structure was as follows: $0.0015 per 1,000 input tokens and $0.002 per 1,000 output tokens. In the process of validating a set of 900 examples, the total incurred cost amounted to $1.21, covering the processing of 806,073 tokens. The comparison of the cost and number of tokens processed between fine-tuning and zero-shot prompting is summarized in Table 4.

During the fine-tuning process, we noticed distinct performance trends for each setup. In the first setup, the accuracy started at 0.81 at the 900th step and increased significantly to 0.85 after the 2700th step. By the end of the observed range, at 3600 steps, the accuracy only improved slightly to 0.86, suggesting that the model might be reaching its optimal performance. In contrast, the second setup showed a steady climb in accuracy from 0.78 to 0.86, indicating consistent learning without any major fluctuations. The third setup began with an accuracy of 0.82 at the 900th step, slightly outperforming the first and significantly surpassing the second at similar intervals. This improvement persisted until the 2700th step, reaching 0.875 accuracy, but then plateaued. The fourth setup began similarly to the second, with a 0.79 accuracy at the 900th step. Yet, it experienced a notable rise to 0.84 by the 1800th step and gradually increased to 0.86 and 0.87 at the 2700th and 3600th steps, respectively. The accuracy trends for these setups can be seen in Figure 2–5.

From the results above, we can observe that using the original label gives better results after the first round of training. This could be because the GPT-3 model already understands these labels, while I used dummy labels in the second and fourth cases. However, a better performance in the first round does not guarantee the best final results. For instance, the first scenario actually had the lowest accuracy among all, even though it started strong. In terms of optimal stopping point, it is worth mentioning that in the third model, additional training beyond the point where it plateaus might be unnecessary. In contrast, the other three setups could still improve from further training, as they continue to show better results with each round of training.

To gain a well-rounded understanding of the model’s performance, I leveraged training loss, accuracy, and weighted F1-score[8] to draw insights about the performance. In the first setup (a), the training loss was 0.021, and it achieved an accuracy of 0.862 with a weighted F1-score of 0.0862. The

[8] The weighted F1-score is a metric that assesses model accuracy by calculating and then averaging each class’s F1-score (the harmonic mean of precision and recall) according to its prevalence in the dataset.

Table 4: Comparison of Cost and Number of Tokens Processed

Scenario	Duration	Cost ($)	Tokens
(a)	25 minutes and 8 seconds	0.32	793,328
(b)	24 minutes and 47 seconds	0.31	780,528
(c)	24 minutes and 51 seconds	0.67	1,687,176
(d)	24 minutes and 49 seconds	0.62	1,559,176
CARP	n/a	1.21	806,073

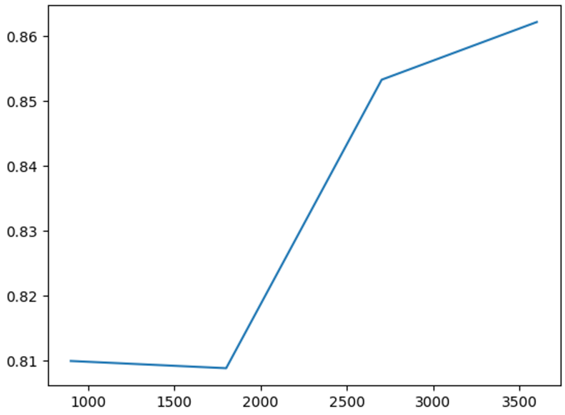


Figure 2: Accuracy and Steps Model (a)

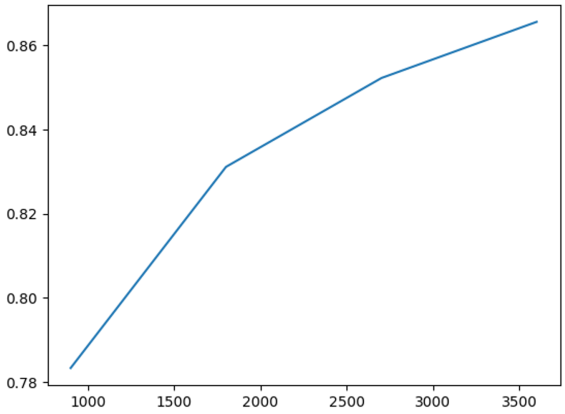


Figure 3: Accuracy and Steps Model (b)

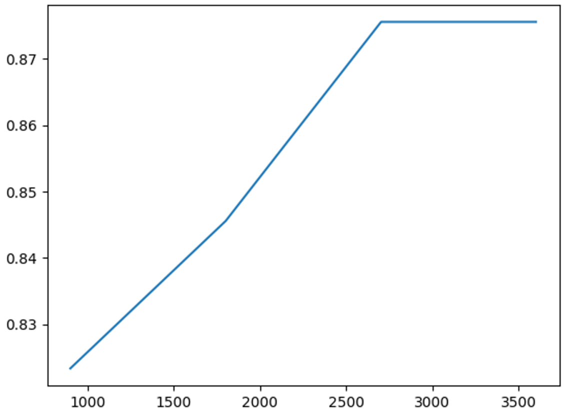


Figure 4: Accuracy and Steps Model (c)

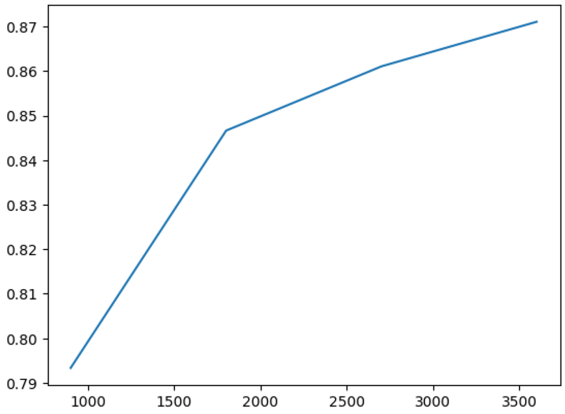


Figure 5: Accuracy and Steps Model (d)

second setup, setup (b), had a slightly higher training loss at 0.025 but managed to reach an accuracy and weighted F1-score of 0.865. Notably, the third scenario (c) demonstrated a superior efficacy with a lower training loss of 0.013 and an accompanying accuracy and weighted F1-score both at 0.875. Lastly, setup (d) had a training loss of 0.012, which is the lowest among the four, with slightly lower accuracy and weighted F1-score compared to (c), both at 0.871. From these findings, it becomes evident that setups (c) and (d) both exhibit low training losses, with setup (d) having the lowest. Low training loss is an indicator that the model has learned the underlying patterns from the data it was trained on. Furthermore, both (c) and (d) have high accuracy and weighted F1-score values, with (c) outperforming all. This suggests that model (c) is making correct predictions a high percentage of the time and is robust in terms of both false positives and false negatives.

From the zero-shot prompting experiment using CARP, it was observed that, from the 900-validation data, CARP achieved a moderate range of accuracy at 0.71 while the weighted F1-score was at 0.7. The similarity between these two values also indicates that it got balanced precision and recall. Compared to the results from fine-tuning, it is apparent that fine-tuning provides better performance. More consistency in both accuracy and weighted F1-score also indicates that the fine-tuned models are more discriminative, leading to fewer false positives and false negatives by the fine-tuned models. The superior performance of the fine-tuned models underscores the effectiveness of tailoring a model to a specific task, even though the pre-trained model has fewer parameters. Despite potentially lower performance than fine-tuned models, the results from the zero-shot prompting experiment showcase the utility of this approach. Being able to achieve moderate accuracy without any task-specific training is still commendable and can be particularly useful in scarce labeled data scenarios, or rapid deployment of a model is necessary. Table 5 summarizes the performance results of the fine-tuned models and CARP experiment.

Upon analyzing the experimental results from fine-tuning and zero-shot prompting methods, a number of valuable observations come to light. For the fine-grained ESG topic classification tasks, a fine-tuned smaller model tends to outperform a larger model employing zero-shot prompting. Furthermore, the financial implications of training the smaller model are considerably less than those associated with validating the efficacy of zero-shot prompting. While OpenAI's pricing is marginally higher for fine-tuned models at $0.0016 per 1,000 tokens, the token consumption is reduced due to the absence of the need for repeated class descriptions in every prompt. Therefore, in an economical and efficient manner, fine-tuned models present a more viable option for this task compared to the large GPT-3.5 model. Evaluating the fine-tuned models, both (c) and (d) demonstrated impressive performance, marked by low training losses. To achieve optimal results, I prioritized accuracy and F1-score. As a result, model (c) was selected for the classification step of the machine learning pipeline proposed in this research.

4.2. Corporate Sustainability Performance Analysis

As detailed in Chapter 3, the process for analyzing news articles begins with gathering content from the New York Times through its Archive API. I then narrow down these articles, selecting only those relevant to the eight chosen companies, utilizing both keywords provided by the New York Times API and NER for effective filtering. Subsequently, I employ the FinBERT-ESG-9-categories model[9] to further sift through the data, setting aside irrelevant articles. This step is essential as it helps in omitting unnecessary content from the analysis and saves on potential costs that might have been incurred if the fine-tuned model were applied to irrelevant articles. Instead, the fine-tuned model is only applied to categorize news pieces into the other classes specified in this research. In the final stage, I conduct sentiment analysis on each article, aiming to gauge a company's performance based on the news content.

The evaluation process encompasses eight diverse companies, including technology giants Apple, Microsoft, Google (under Alphabet), and Meta (formerly Facebook), alongside major players in the beverages industry such as The Coca-Cola Company and PepsiCo, as well as pharmaceutical firms Pfizer and Johnson & Johnson. This varied selection allows for a comprehensive analysis across different sectors. I aim to gain insights into how the assessments vary from one industry to another, considering the unique characteristics of each sector. The number of news articles related to each of these companies is laid out in Table 6. The detailed assessment for each company is described in the following sections.

4.2.1. Apple Inc.

Apple Inc. (Apple) is the largest company in the world in terms of market capitalization, having a staggering value of US$2.65 trillion as of 2022 (Johnston, 2022). Established in 1976 by Steve Jobs, Steve Wozniak, and Ronald Wayne in Los Altos, California, this American manufacturer has evolved to become a major player in the technology industry (Linzmayer, 2004). The company specializes in designing, manufacturing, and marketing a wide array of electronic devices and services, including smartphones, personal computers, tablets, wearables, peripherals, and various support services (Apple Inc., 2022). Some of its most famous products encompass the iPhone, Mac, and iPad. Apple's headquarters are located in California, and it employs approximately 164,000 full-time equivalent staff members (Apple Inc., 2022). Under the leadership of Chief Executive Officer (CEO) Tim Cook, who continued the reins from Steve Jobs in 2011 (Apple Inc., n.d.), Apple has continued to flourish and maintain its position at the forefront of technological advancement.

In our examination of 724 New York Times articles focused on Apple, the machine learning pipeline introduced in this study provided some insightful findings. Notably, Apple's societal impacts remain a central focus in media coverage. A dominant portion of the articles, amounting to 268, were classified under 'Product Liability,' reflecting a keen interest in the company's product offerings and related concerns. Additionally, governance, especially in terms of 'Business Ethics,' received significant attention with 237 articles. The 'Corporate Governance' category followed closely, comprising 107 articles, indicating a steady interest in Apple's corporate policies and practices. Environmental themes, on the other hand, were scarcely represented, with categories like 'Climate Change,' 'Resource Stewardship,' and 'Environmental Opportunities' garnering a combined total of just 12 articles. This limited environmental coverage suggests that the media might be underrepresenting Apple's efforts in this domain or that Apple's environmental initiatives were less newsworthy during the period of analysis.

Digging into the sentiment analysis of these articles reveals further insights into the media's perception of Apple's actions in each category. Overall, the sentiment expressed across the articles predominantly skews towards neutrality, with nearly 60% (428 out of 724) of the articles classified as such. This suggests a measured approach by the news outlet when reporting on Apple's multifaceted operations. In the 'Product Liability' theme, which held the largest share of articles, for example, a majority (153) maintained a neutral tone, while 86 leaned negative, and only 29 expressed a positive outlook. Moreover, in terms of 'Business Ethics,' the articles featured a neutral sentiment to a great degree (131), although negative perspectives were also evident in 81 news pieces. The remaining in this class were classified as positive. The 'Corporate Governance' subset showcased a largely neu-

[9] See A. Huang (2022)

Table 5: Comparative Evaluation Metrics of Different Models

Scenario	Training Loss	Accuracy	Weighted F1-Score
(a)	0.021	0.862	0.862
(b)	0.024	0.865	0.865
(c)	0.013	0.875	0.875
(d)	0.012	0.871	0.871
CARP	n/a	0.706	0.703

Table 6: Number of News Articles per Company Before and After Excluding Non-ESG Content

Company	Number of News Articles	Without Non-ESG
Apple	1,648	724
Microsoft	1,376	733
Alphabet (Google)	2,756	1506
Meta	2,113	1161
Coca-Cola	167	101
Pepsi	106	61
Pfizer	419	364
Johnson & Johnson	254	229

tral sentiment as well, accounting for 75 of its 107 articles. While environmental topics received limited attention, it is worth mentioning that articles focused on this aspect largely maintained a neutral to positive sentiment. The sentiment results for each category is described in Figure 6.

In terms of the scoring for Apple's performance using the sentiment, only the 'Environmental Opportunities' and the 'Social Opportunities' received a positive sentiment in general. Each of them got a total (normalized) score of 1 (0.2) and 4 (0.14), respectively. The 'Product Liability' and 'Business Ethics' aspects emerged as prominent areas of concern, with a total (normalized) score of -57 (-0.21) and -56 (-0.24). This indicates that despite a substantial volume of neutral coverage, the negative sentiment still slightly outweighs the positive. The critical stance is also echoed in the 'Human Capital' domain, with a total (normalized) score of -14 (-0.20). In regards to its performance in the 'Corporate Governance' realm, Apple hovers close to neutral with a modestly negative total (normalized) score of -2 (-0.02). Table 7 shows the comparison of scoring between the technology companies. The detailed scoring is available in the Appendix.

Diving deeper into the news articles, in regards to the 'Product Liability' aspect, there are various themes considered as negative. Concerns highlighted by media include privacy issues related to Apple's products and services, as well as the company's restrictive practices in managing its app store (e.g., Holpuch, 2022; Nicas, 2019). In terms of 'Business Ethics,' the discussion is often framed around antitrust concerns, with Apple's policies and fees for app developers being perceived as unfair (e.g., Nicas et al., 2020; Satariano, 2021). In the other facets, Apple was also scrutinized for its labor practices, where reports of stringent working conditions, particularly for workers in China, and friction with the labor unions (e.g., Gough and Chen, 2014; Zhuang, 2022). However, Apple has also gained positive attention for its initiatives aimed at enhancing healthcare, showcasing the company's concern for societal well-being (e.g., Singer, 2018). Despite this positive attention, it is noted that Apple has received less media attention compared to the more contentious topics.

According to MSCI's evaluation, Apple's approach to ESG matter is rated as moderate, placed in the middle range with a 'BBB' rating among 137 companies in the technology hardware, storage, and peripherals industry (MSCI, 2023a). When dissecting the components of the assessment, it is observed that Apple lags, particularly in areas concerning business ethics and labor practices within its supply chain (MSCI, 2023a). This viewpoint is echoed in the findings in the previous paragraph, where Apple is under significant scrutiny for ethical concerns and labor standards at its supplier locations. Interestingly, regarding issues of privacy and data security (part of product liability), MSCI categorizes Apple as average (MSCI, 2023a), contrasting with the more critical perspective of numerous negative sentiments identified in this research. At the same level, Apple is also considered average for its efforts in corporate governance and handling electronic waste (MSCI, 2023a). The present study parallels these findings, with the sentiment towards Apple's corporate governance being fairly mixed, albeit with a slight tilt toward the negative. Unfortunately, there were only a few articles talking about resource stewardship and none of them discussed Apple's waste management.

MSCI believes that Apple is a frontrunner compared to its peers in terms of human capital development and advancing clean technology. Despite Apple's efforts in these areas, they are still underreported by the news media. There are only a handful of articles talking about how well Apple develops its highly skilled workers, as well as the efforts made by Apple to develop technology for a better environment. This gap indicates a potential oversight in the media narrative, fail-

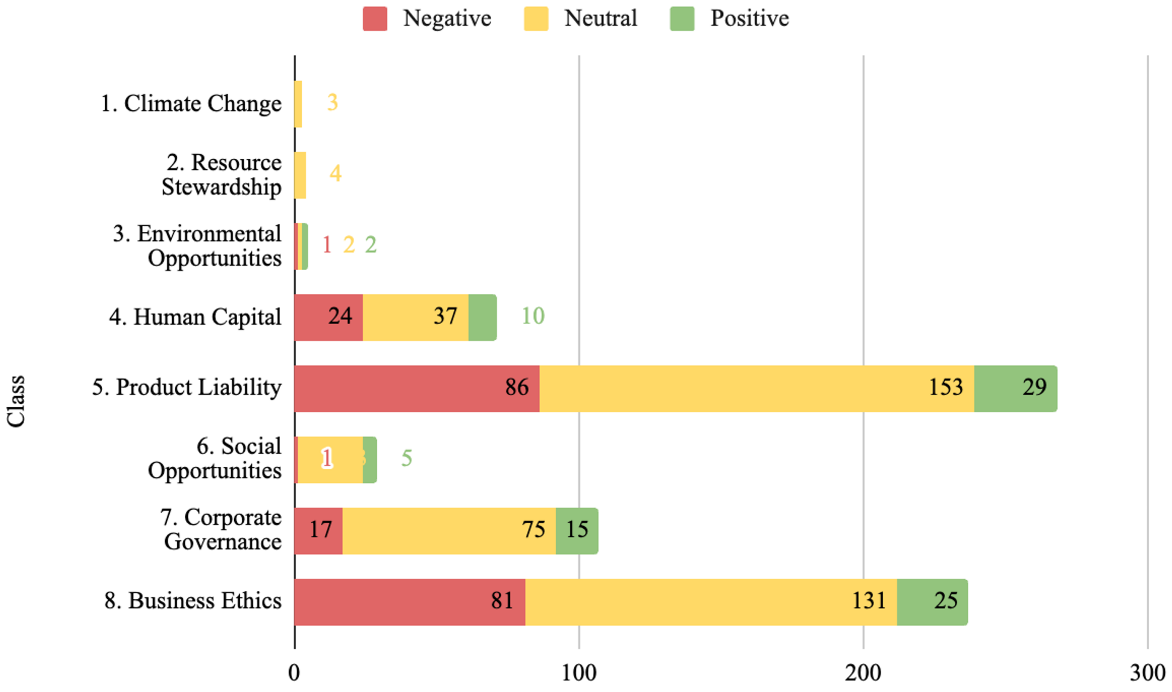


Figure 6: Apple's Sentiment Distribution Across Different ESG Categories

Table 7: Score Comparison Between Technology Companies

	Apple		**Microsoft**		**Alphabet**		**Meta**	
Categories	Score	Normalized Score	Score	Normalized Score	Score	Normalized Score	Score	Normalized Score
1. Climate Change	0	0.00	0	0.00	3	0.17	0	0.00
2. Resource Stewardship	0	0.00	0	0.00	1	0.17	1	0.17
3. Environmental Opportunities	1	0.20	1	1.00	2	0.09	2	0.67
4. Human Capital	-14	-0.20	-6	-0.11	-22	-0.15	-13	-0.21
5. Product Liability	-57	-0.21	-44	-0.26	-130	-0.28	-136	-0.29
6. Social Opportunities	4	0.14	14	0.24	15	0.13	-2	-0.04
7. Corporate Governance	-2	-0.02	-9	-0.05	-4	-0.02	-20	-0.13
8. Business Ethics	-56	-0.24	-97	-0.37	-186	-0.34	-128	-0.30

ing to capture the complete picture of Apple's initiatives and its impact on sustainable business practices and responsible resource management.

4.2.2. Microsoft

Microsoft stands as a global technological giant, with its main office situated in Washington, United States (Microsoft, n.d.-a). In terms of market capitalization, it is ranked as the third largest company worldwide, following Apple and Saudi Aramco, boasting a value of US$2.1 trillion as of 2022 (Johnston, 2022). Bill Gates and Paul Allen founded the company in 1975, initially focusing on the development and licensing of computer software, with their renowned operating system, Microsoft Windows, being a standout product (Zachary & Hall, 2023). Over the years, the company has expanded into several other product categories, including servers, productivity software, personal computers, consumer electronics, online advertising, and numerous other services (Zachary & Hall, 2023). The company has gained widespread recognition for its Windows operating system and Office Suite software. As of June 2023, Microsoft is a workplace for roughly 221,000 full-time employees around the globe (Microsoft, 2023). Currently, the company is under the leadership of Satya Nadella, who has been serving as the chairman and CEO since 2014 (Microsoft, n.d.-b).

There are marginally more articles from the New York Times discussing Microsoft's ESG initiatives compared to Apple, with 733 pieces. Of all the articles, the greatest focus was placed on 'Business Ethics,' which was the subject of 265 articles. The topic of 'Corporate Governance' also received considerable coverage, tallying up to 180 articles. This underscores the media's significant interest in the company's governance front, such as ethical considerations and leadership decisions. On the social facet, similar to Apple, 'Product Liability' topics became the most discussed area with 167 articles. 'Social Opportunities' and 'Human Capital' also featured prominently, with 58 and 55 articles, respectively. The coverage on the environmental aspect was still low, with a cumulative count of only eight articles, hinting at a potential underreporting of Microsoft's environmental endeavors similar to what was observed with Apple's coverage.

Analogous to the sentiment of Apple's news, it was evident that the neutrality sentiment prevails in the articles about Microsoft's ESG practices, with the majority of articles—433 out of 733—displaying a neutral stance. The media kept reporting the news without a strong prejudice toward the company. Despite the neutral tendency, it is worth noting that the discussion around 'Business Ethics' was significantly critical, with 113 articles classified as negative compared to 136 neutral and only 16 positive ones, signaling strong media scrutiny of Microsoft's ethical conduct. Nevertheless, for 'Corporate Governance' and 'Human Capital,' it could be observed that over half of the articles maintained this neutral tone, with 125 and 35 articles, respectively. Regarding the 'Product Liability' angle, the number of negative articles was moderately prominent, accounting for just above 35% with 59 articles, while 93 remained neutral out of 167 articles. The environmental categories—namely 'Climate Change,' 'Resource Stewardship,' and 'Environmental Opportunities'—show a mixed but limited sentiment profile, with a small number of articles suggesting these topics have not been as contentious in the media. The sentiment for each category is depicted in Figure 7.

Analyzing Microsoft's performance through sentiment scores unveils compelling insights. As discovered previously, 'Business Ethics' registered as the most concerning area, more prevalent than those of Apple, with a total (normalized) score of -97 (-0.37). The score indicates that Microsoft had worse ethical concerns than its peers. Another area full of criticisms was 'Product Liability,' where Microsoft got a total (normalized) score of -44 (-0.26), slightly worse than Apple. Comparable to Apple's 'Corporate Governance' score, Microsoft's score was also close to neutral, with a normalized (total) score of -0.05 (-9). Microsoft tended to be better in terms of 'Human Capital' than Apple, albeit still inclined to negative nuanced with a normalized (total) score of -0.11 (-6). Microsoft attained a much better score in terms of 'Social Opportunities' than Apple, with a normalized (total) score of 0.24 (14), indicating more societal actions from Microsoft got more attention from the media. Regrettably, on the environmental front, there were only a few news pieces available to be generalized. The detailed scoring is available in the Appendix.

Delving into the 'Business Ethics' news, Microsoft was often being criticized regarding antitrust and abuse of power matters (e.g., Lohr, 2020; Markoff, 2003; Weise and McCabe, 2022). A significant portion of this coverage stems from an earlier era, with 81 out of 113 articles concentrated between 2003 and 2009. In contrast, the subsequent 13 years accounted for just 32 articles, suggesting a potential shift toward more ethical operations by Microsoft, though criticisms persist. Concerning the 'Product Liability' subject, security vulnerabilities, particularly breaches in Microsoft's email systems, have been a focal point of media attention (e.g., Conger and Frenkel, 2021; Scott, 2015). Challenges in 'Human Capital' have also emerged, linked to immigration issues and job cuts (e.g., Frenkel, 2018; Wingfield, 2016a). Nonetheless, Microsoft has made notable efforts to contribute positively to society, including initiatives to support affordable housing and to enhance the accessibility of essential services for public servants (e.g., Vance, 2010; Weise, 2019).

MSCI places Microsoft within the software and services industry and grants it the highest rating (AAA) among 484 companies (MSCI, 2023a). These prestigious rankings originate from MSCI's assessment of Microsoft's exemplary performance in critical areas such as human capital development, safeguarding privacy and data, opportunities in clean tech, and mitigating carbon emissions (MSCI, 2023a). However, the results from analyzing the news articles show the investment in Microsoft's talented employees seems to be underrepresented. Interestingly, in terms of privacy & data security, MSCI evaluates Microsoft differently in regard to controversies from the media[10]. Microsoft is considered to be involved in severe-to-moderate level controversies—similar to what I found in the news—while MSCI still rates Microsoft as the leader in this aspect compared to its competitors (MSCI, 2023a). Additionally, the media's limited coverage of Microsoft's environmental attempts, with merely a single article on both its cleantech initiatives and carbon reduction strategies, raises questions about the adequacy of information to fully evaluate the company's environmental performance.

In a similar marking to Apple, MSCI assigns Microsoft a less favorable rating on the topic of business ethics and considers it to be average in corporate governance (MSCI, 2023a). Reciprocal with this rating, the sentiments found in the media reports suggest that Microsoft could benefit from a reassessment of its ethical procedures to address concerns about its dominant market position and abuse of power. This is also in line with the controversies indicator by MSCI, in which Microsoft is considered in moderate to severe controversies. As for corporate governance, the majority of media sentiments remain neutral but tilt slightly towards the negative, broadly matching MSCI's findings. This indicates that Microsoft's approach to governance—encompassing both its control structure and ethical conduct—requires further scrutiny and potential enhancement to uphold and improve its standing in these critical aspects.

4.2.3. Alphabet Inc. (Google)

Alphabet Inc. (Alphabet) is a conglomerate holding company of Google, a multinational technology company that provides a diverse range of products and services, including advertising, operating systems, hardware, web browsers,

[10] MSCI ESG Controversies serves as an additional indicator to assess companies' profiles, focusing on their actual or alleged involvement in activities that have adverse impacts, as reported by the media.

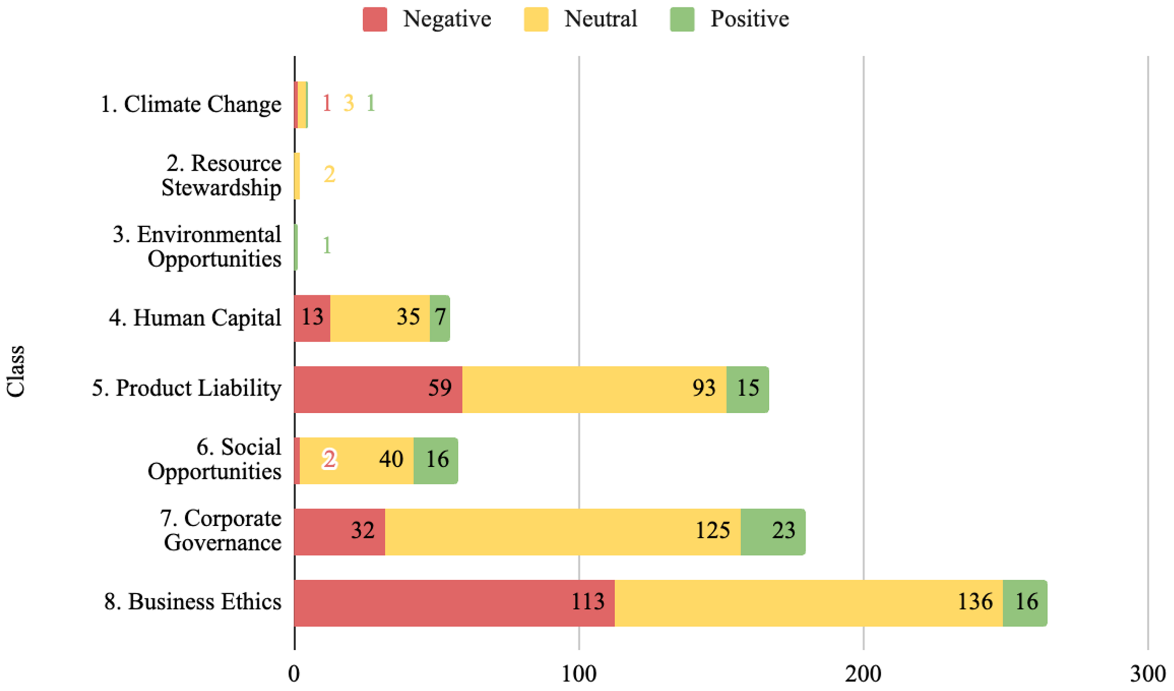


Figure 7: Microsoft's Sentiment Distribution Across Different ESG Categories

and cloud services (Alphabet Inc., 2022). Its market capitalization ranked fourth worldwide, with its worth recorded at US$1.54 trillion in 2022 (Johnston, 2022). Google, founded by Sergey Brin and Larry Page in 1998, initially started as a search engine and has since expanded to offer over 50 different internet services and products (Hall & Hosch, 2023). Some of its most notable offerings include Google Chrome, Gmail, Android, and Google Maps. In a significant move in 2015, Google underwent a reorganization, leading to the formation of Alphabet as a new public holding company (Alphabet Inc., 2015). The company's primary headquarters, known as the Googleplex, is situated in Mountain View, California, United States, and is the workplace for approximately 190,000 employees (Alphabet Inc., 2022). In 2019, Sundar Pichai took on the role of CEO for both Google and Alphabet, succeeding co-founders Larry Page and Sergey Brin (Alphabet Inc., 2019). Since Google is the biggest company under Alphabet, the discussion below mainly contains Google's news articles.

The New York Times has provided more extensive coverage of Alphabet and Google's ESG initiatives than it has for Apple and Microsoft. Out of 1,506 pieces spanning eight topics, 'Business Ethics' predominated, being the subject of 551 articles, which is over one-third of the total. The theme of 'Product Liability' also stood out with a significant count of 472 articles, over 30% of all articles. Meanwhile, 'Corporate Governance' and 'Human Capital' also received considerable attention, with 182 and 142 articles, respectively. The ranking of the most reported topics for Alphabet/Google closely mirrored that of Microsoft, with 'Business Ethics,' 'Product Liability,' and 'Corporate Governance' receiving the most spotlight. Social impact was also a notable subject, with 'Social Opportunities' discussed in 113 articles. The environmental category received more attention compared to their industry counterparts, with 46 articles in total, although it was still less compared to other subjects. In this spectrum, 'Environmental Opportunities' led with 22 articles, followed by 'Climate Change' with 18, and 'Resource Stewardship' with 6 articles.

The sentiment analysis of The New York Times articles on Google's ESG efforts reveals a predominately neutrality across the topics, with 933 out of 1506 articles maintaining an unbiased tone. However, in the domain of 'Business Ethics,' a critical perspective is noticeable, with 228 articles carrying a negative sentiment, almost outstripping the 281 neutral articles. This might suggest a keen media vigilance on Google's ethical practices. In contrast, 'Human Capital' and 'Corporate Governance' topics show a preference for neutrality, with 98 and 136 articles, respectively, indicating less contentious coverage. 'Product Liability' received quite controversial coverage, with a substantial 154 articles skewed towards a negative sentiment, highlighting concerns or issues in this area, although 294 articles remain neutral. 'Social Opportunities' presents a more positive outlook with 20 positive articles. Environmental issues showcase a diverse but generally limited sentiment range. 'Climate Change' has 13 neutral and 4 positive articles—only one article is considered negative. This could imply that while these topics are covered, they do not ignite as much as other topics in the media. The breakdown of sentiment across each category is illustrated in Figure 8.

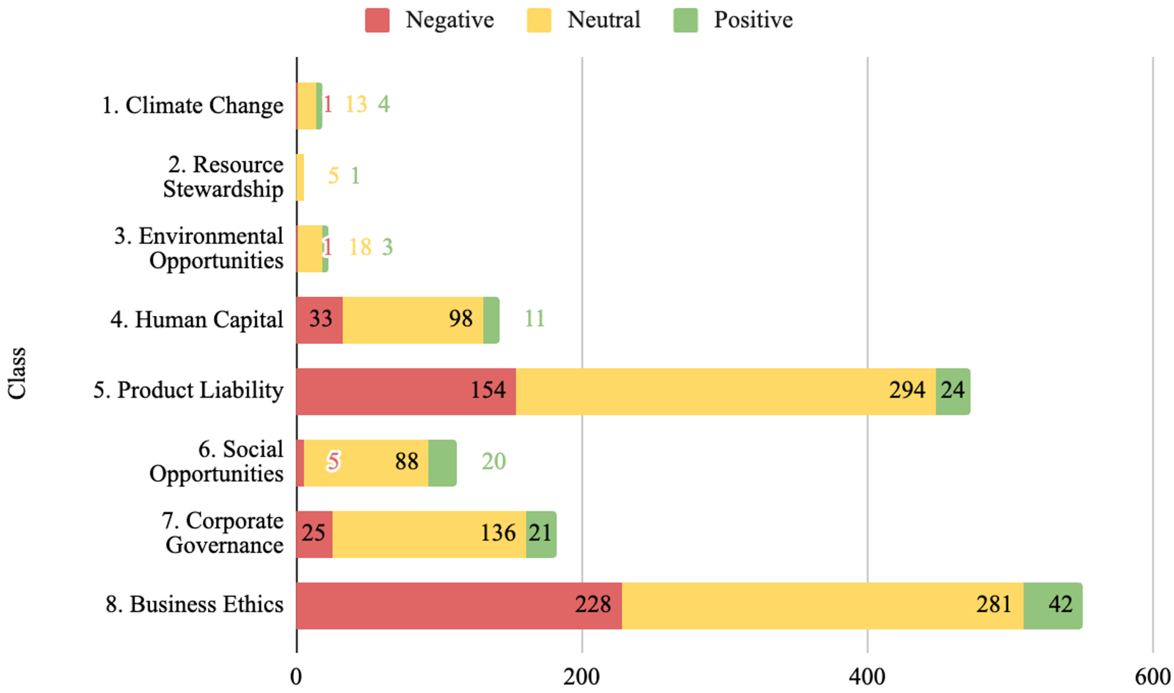


Figure 8: Alphabet's Sentiment Distribution Across Different ESG Categories

The sentiment assessment for Google's ESG performance presents a varied impression. In contrast to Apple and Microsoft, scoring the environmental features of Google's performance is more straightforward. This clarity comes from a broader range of sentiments represented across the categories. Specifically, Google demonstrates a positive tilt in this sector, with the areas of 'Climate Change,' 'Resource Stewardship,' and 'Environmental Opportunities' achieving normalized (total) scores of 0.17 (3), 0.17 (1), and 0.09 (2), respectively. These figures suggest a favorable perception of Google's efforts in the environmental matter. Conversely, 'Human Capital' reflects a notable drop with a normalized (total) score of -0.15 (-22), indicating areas of concern within this sphere. In the other social contexts, the 'Product Liability' category is markedly the most critical, with a significant negative normalized (total) score of -0.28 (-130), while 'Social Opportunities' conversely portrays a brighter spot with a positive normalized (total) score of 0.13 (15). The 'Product Liability' aspect is even more concerning than those of Microsoft and Apple, with a normalized score of -0.26 and -0.21, respectively. In terms of 'Social Opportunities,' Microsoft is also still leading with a normalized score of 0.24. The detailed scoring is available in the Appendix.

Moving to the governance domain, Google's 'Corporate Governance' presents nearly balanced sentiments with a thin negative normalized (total) score of -0.02 (-4), which aligns it closely with its industry counterparts. However, 'Business Ethics' emerges as the most concerning area, with a stark normalized (total) score of -0.34 (-186), signifying strong negative sentiment and possibly considerable criticism in this area. Although Google's normalized score in 'Business Ethics' is somewhat better than Microsoft's (-0.37), it is important to note that Google's total negative score is -186, compared to Microsoft's -97. This difference may point to Google facing more intensive media scrutiny over its ethical practices.

Reflecting on the media coverage of Google in the environmental landscape, there is a noticeable acknowledgment of the company's investments in renewable energy and other sustainable technologies. Several reports, although categorized as neutral, highlight Google's promising endeavors to address climate change (e.g., Austen, 2019; Hardy, 2016). In the social scene, particularly in 'Product Liability,' various issues have surfaced, such as problematic user data tracking, the misuse of products for extortion, and defects in product quality (e.g., Hill and McCabe, 2022; Morales, 2022; Woo, 2022). In the area of 'Human Capital,' Google has faced criticism regarding its treatment of employees, particularly temporary and contract workers (e.g., Scheiber, 2020; Wakabayashi, 2021). Nevertheless, the company has also gained recognition for its initiatives aimed at enhancing quality of life. An example of such efforts includes the development of healthcare tools designed to support doctors in making informed decisions (e.g., Grady, 2020). On the most concerning side, 'Business Ethics,' the media has intensely scrutinized Google's ethical practices, especially in relation to antitrust issues. Google's involvement in various lawsuits across different regions and for different reasons has drawn considerable attention from the news outlet (e.g., Satariano, 2022a; Tracy, 2021).

As stated by MSCI, Alphabet has been assigned a 'BBB' rating, positioning it as average within the interactive media & services sector, which comprises 65 companies (MSCI, 2023a). The assessment highlights several concerns, such as corporate conduct, governance practices, employee development, and the pursuit of clean technology innovations (MSCI, 2023a). Corresponding to the findings from the media, the company is heavily criticized for its ethical behaviors. In the sphere of corporate governance, Alphabet appears to maintain a neutral stance, comparable to its peers like Microsoft and Apple, though MSCI categorizes it as a laggard (MSCI, 2023a). Points of disagreement also emerge in the opportunities in clean tech, where MSCI evaluates the company as lagging behind (MSCI, 2023a), whereas news reports indicate the company has launched multiple initiatives in this field. Nonetheless, there is a consensus between this research and MSCI's findings that Alphabet has taken steps towards reducing its carbon footprint (MSCI, 2023a).

4.2.4. Meta Platforms, Inc.

Meta Platforms, Inc. (Meta), previously recognized as Facebook, is a major technology corporation based in Menlo Park, California (Meta Platforms Inc., 2022). The company owns and operates a variety of products and services, including Facebook, Instagram, Messenger, and WhatsApp (Meta Platforms Inc., 2022). Currently, the technology giant is ranked 10th in terms of market capitalization in the world, with a value of US$449 billion (Johnston, 2022). Since Mark Zuckerberg founded Facebook in 2004 along with Dustin Moskovits, Chris Hughes, and Eduardo Saverin, Meta has grown from a social media platform connecting friends and family to pioneering immersive experiences through augmented and virtual reality (Meta Platforms Inc., n.d., 2022). Zuckerberg continues to play a pivotal role, leading the company as both chairman and CEO (Meta Platforms Inc., n.d.). As of the end of December 2022, the company recorded a total of 86,482 employees, inclusive of the roughly 11,000 workers affected by the massive layoff announced in November of the same year (Meta Platforms Inc., 2022).

The New York Times' coverage of Meta's ESG efforts encompasses a total of 1,161 articles across the topics, more than Apple and Microsoft's coverage. Similar to Apple, the most prevalent subject in this collection is 'Product Liability,' which dominates the spotlight with 470 articles, making up a substantial portion of the total coverage. This is followed by 'Business Ethics,' which is the focus of 422 pieces. 'Corporate Governance' ranks third with 150 articles, similar order to its technology peers. On the social front, 'Human Capital' and 'Social Opportunities' are notable topics, covered in 61 and 47 articles, respectively. Environmental topics remain underrepresented, with a total of just 11 articles; only Alphabet receives more coverage in the technology sector, featuring in over 20 articles. In the environmental domain, 'Resource Stewardship' interestingly is the subject of a greater number of articles, with 6 pieces, compared to 'Climate Change' and 'Environmental Opportunities,' which are covered in 2 and 3 articles, respectively.

The sentiment analysis of Meta's performance shows a pattern consistent with the other three technology companies, where the impartial view dominates the results, accounting for over 60% of the articles. Yet, 'Product Liability' stands out with 160 negative mentions, marking it as the most criticized topic despite a generally neutral coverage. Similarly, the domain of 'Business Ethics' is also heavily critiqued, with 140 negative articles against 270 neutral ones. 'Business Ethics' is one of the most concerning topics for technology companies, not just Meta. The scrutiny suggests the media's close watch on the ethical practices of large tech firms. Coverage on 'Human Capital' and 'Corporate Governance' predominantly maintain a neutral stance, numbering 38 and 106, respectively, pointing to less controversial media reporting in these areas. 'Social Opportunities' and environmental topics display a mix of sentiments but are covered less extensively. Figure 9 depicts the sentiment distribution across different categories.

Meta's sentiment scores highlight areas of concern, particularly within the social and governance categories. While environmental aspects like 'Resource Stewardship,' 'Environmental Opportunities,' and 'Climate Change' exhibit neutral to positive scores, coverage in these areas is sparse. In contrast, 'Product Liability' and 'Business Ethics' face notable scrutiny, with significant negative normalized (total) scores of -0.29 (-136) and -0.30 (-128), respectively. These scores are almost identical to those of Alphabet, indicating that these two tech giants are closely watched by the media regarding their products and ethical practices. 'Human Capital' also registers a substantial negative normalized (total) score of -0.21 (-13), marking the lowest performance among the other technology companies. In the typically more positive 'Social Opportunities' category, Meta surprisingly scores slightly negative at -0.04 (-2) normalized (total) score. This negative sentiment extends to 'Corporate Governance,' where Meta's normalized (total) score of -0.13 (-20) is the least favorable when compared with its sector peers. The detailed scoring is available in the Appendix.

Upon examining the articles in detail, it becomes apparent that the 'Business Ethics' theme is largely characterized by improper business practices and antitrust litigation (e.g., C. Kang et al., 2019; Satariano, 2020). Moving to the other governance facet, Meta received criticisms for the decision-making and conduct of its leadership, as well as the departure of key executives (e.g., Isaac, 2018; Satariano and Frenkel, 2022). In the 'Product Liability' category, which garners the most negative attention for Meta, several issues are highlighted by the media, including privacy breaches, potential for dangerous behaviors, and unreliable services (e.g., Isaac, 2021; Mac, 2022; Satariano, 2022b). Common challenges like antitrust and privacy issues are also faced by Meta's industry peers. Nonetheless, given the significant role Meta's products play in social interactions, there is heightened media concern over the ways these products can harm individuals, and it is suggested that the company has not adequately addressed these risks.

MSCI grades Meta with the lowest rating, CCC, compared to the other companies within the interactive media & services industry, the same sector as Alphabet (MSCI, 2023a). Meta is particularly lagging in areas of ethical conduct and

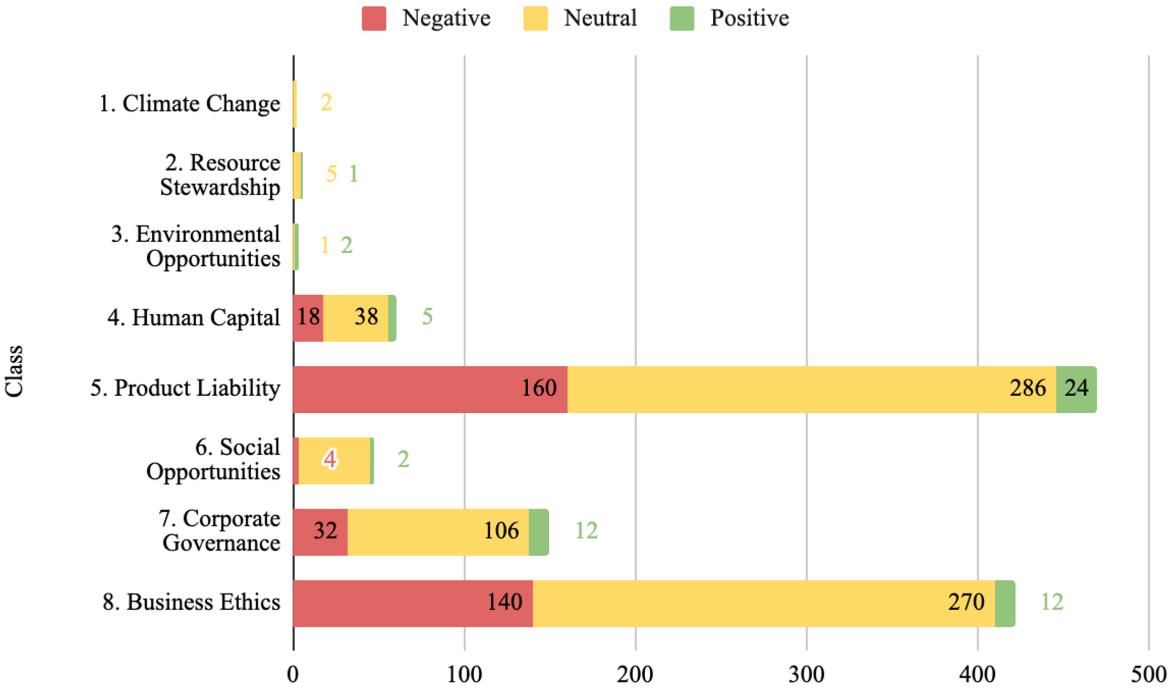


Figure 9: Meta Platforms' Sentiment Distribution Across Different ESG Categories

human capital development (MSCI, 2023a). These concerns echo the media's criticisms of Meta's business practices and allegations of anti-competitive actions. The media reports also disclose Meta's plans to pause hiring and reduce staff (e.g., Frenkel, 2022a, 2022b). MSCI's controversy indicator supports these points, suggesting Meta is entangled in moderate to severe labor rights disputes (MSCI, 2023a). However, when it comes to corporate governance and privacy, MSCI considers Meta's performance average, despite recognizing the company's involvement in moderate to severe governance disputes and serious privacy & data security issues (MSCI, 2023a). These MSCI assessments of controversy indicators are consistent with the results from this study, which indicate negative scores for Meta in these two categories. The sole area where MSCI acknowledges Meta's leadership is in its efforts to slash carbon emissions (MSCI, 2023a).

4.2.5. The Coca-Cola Company

Started in 1886 when Dr. John Pemberton sold his newly crafted syrup at Jacob's Pharmacy, laying the foundation for what would become a multinational total beverage corporation (The Coca-Cola Company, n.d.-b). Today, the company has grown far beyond its famous Coca-Cola drink, managing an array of brands like Sprite and Fanta and extending its reach to over 200 countries and territories (The Coca-Cola Company, n.d.-b). The company also offers a diverse product lineup that includes coffee, tea, juice, value-added dairy, plant-based beverages, and innovative new drinks (The Coca-Cola Company, 2022). At the helm is James Quincey, who has been with the company since 1996 and currently oversees its operations and 82,500 employees as the chairman and CEO (The Coca-Cola Company, n.d.-a, 2022). Under his leadership, this Atlanta-based company has continued to thrive, with a reported market value of US$242 billion (Britannica, 2023a).

The examination of Coca-Cola's ESG initiatives is covered in a total of 101 articles, a smaller number compared to the extensive coverage of major technology firms. Among these, 'Corporate Governance' emerges as the most discussed topic with 35 articles. 'Business Ethics' is another significant topic, featuring in 20 articles. The environment-related categories show varied coverage, with 'Resource Stewardship' leading with 13 articles. On social issues, the coverage for each topic is distributed relatively evenly. These articles are dominated by neutral sentiment, with a total of 74 out of 101 articles. 'Corporate Governance' sees an equal distribution of positive and negative sentiment, each with four articles. 'Business Ethics' faces the most scrutiny, reflected in six articles with a negative sentiment. Regarding environmental issues, 'Resource Stewardship' draws the most concern with five negative articles. All topics in the environmental sphere do not garner positive highlights from the media, indicating challenges in this area. Similarly, social topics also do not receive positive media attention. Instead, the coverage is primarily neutral, though some negative perspectives are present. Figure 10 provides a visual breakdown of the sentiments across these categories.

As reflected by sentiment scores, Coca-Cola's performance indicates several areas of concern across ESG topics. 'Resource Stewardship' incurs the most negative sentiment, with a normalized (total) score of -0.38 (-5). However, when taking a deeper look into the articles, there are some mis-

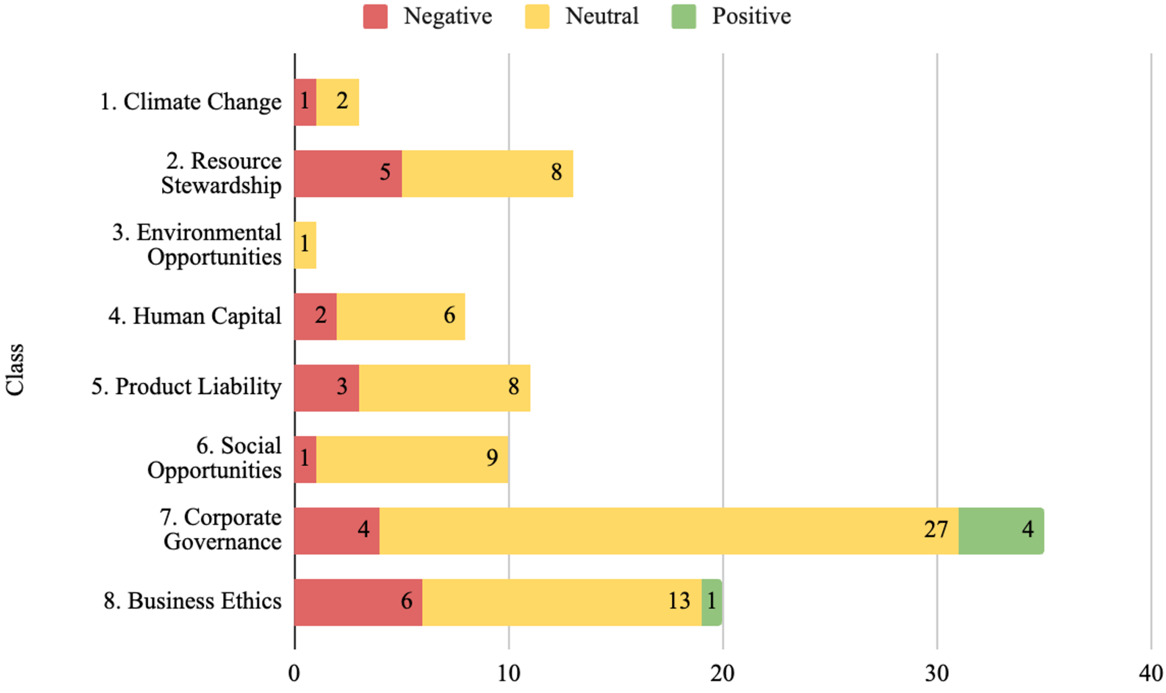


Figure 10: The Coca-Cola Company's Sentiment Distribution Across Different ESG Categories

classifications in terms of the sentiment, where the articles are considered negative and neutral, whereas they discuss positive efforts by Coca-Cola, such as recycling and water replenishment initiatives (e.g., Corkery, 2019; Schwartz, 2015). This means the scoring might not reflect the actual performance of the company in this domain. 'Business Ethics' also draws a notably negative view, sharing a normalized score of -0.25 with 'Human Capital.' Coca-Cola faces some critiques from the media for its ethical conduct, such as the alleged drug-smuggling operation and insider trading (e.g., Lattman, 2012; Mele, 2016). Several news pieces also covered the company's action to cut jobs (e.g., Strom, 2015). In another social theme, Coca-Cola scores -3 in total score and -0.27 in normalized score in 'Product Liability,' pointing to perceived issues in its products and logistics. Particularly, the media criticizes how Coca-Cola often sugarcoats the calories or nutrition facts, adds questionable ingredients to the products, and its supply chain management (e.g., Board, 2015; Southall, 2015; Strom, 2014a). 'Corporate Governance,' although it breaks even with a score of 0, the company still receives critiques about pay for its executives (e.g., Eavis, 2014). Table 8 exhibits the scoring comparison between the beverage and the pharmaceutical companies. The detailed scoring is available in the Appendix.

Rated as AAA, the highest rating in MSCI ESG Ratings, Coca-Cola stands out as the top performer among a hundred companies in the beverage sector (MSCI, 2023a). The company excels in several critical areas, such as management practices, developing healthful products, conserving water, ensuring the safety and quality of products, safeguarding employee health, and reducing the carbon emissions of its products (MSCI, 2023a). However, Coca-Cola's performance in ethical business conduct and waste management from packaging is seen as less impressive, with MSCI calling it mediocre (MSCI, 2023a). Sentiment analysis also casts a shadow on Coca-Cola's reputation, as ethical concerns arise from insider trading scandals. MSCI's data on controversies supports this view, indicating that Coca-Cola has been involved in a moderate amount of bribery and fraud cases (MSCI, 2023a).

Regarding 'Resource Stewardship,' MSCI views Coca-Cola's handling of packaging materials and waste as typical (MSCI, 2023a), but news outlets have noted the company's efforts to recycle. Nevertheless, both the media and MSCI acknowledge the company's efforts in its water management practices (MSCI, 2023a), such as the effort to replenish water that it uses around the globe (e.g., Schwartz, 2015). However, there's a difference of opinion on 'Product Liability.' MSCI ranks Coca-Cola highly in this category, suggesting leadership and responsibility in product nutrition (MSCI, 2023a). In contrast, sentiment analysis suggests Coca-Cola faces challenges here, with concerns over ingredients and nutritional information that may pose health risks to customers (e.g., Board, 2015).

4.2.6. PepsiCo

PepsiCo, Inc. was formed in 1965 through the merger of Pepsi-Cola Company and Frito-Lay, Inc (Britannica, 2023b). The original Pepsi-Cola, crafted by Caleb D. Bradham in 1898, gained formal corporate status in 1919 (Britannica, 2023b; PepsiCo Inc., 2022). Nowadays, PepsiCo is recognized as a prominent force in both the beverages and convenience foods sector, home to well-known brands such as Pepsi, Lay's, Cheetos, and Quaker (PepsiCo Inc., 2022).

Table 8: Score Comparison Between Beverage and Pharmaceutical Companies

	Coca-Cola		**PepsiCo**		**Pfizer**		**Johnson & Johnson**	
Categories	Score	Normalized Score	Score	Normalized Score	Score	Normalized Score	Score	Normalized Score
1. Climate Change	-1	-0.33	0	0.00	N/A	N/A	N/A	N/A
2. Resource Stewardship	-5	-0.38	-2	-0.67	0	0.00	-1	-1.00
3. Environmental Opportunities	0	0.00	N/A	N/A	0	0.00	1	1.00
4. Human Capital	-2	-0.25	0	0.00	-4	-0.27	-1	-0.20
5. Product Liability	-3	-0.27	-2	-0.29	-19	-0.19	-49	-0.40
6. Social Opportunities	-1	-0.10	4	0.40	43	0.30	11	0.23
7. Corporate Governance	0	0.00	0	0.00	-4	-0.09	-2	-0.13
8. Business Ethics	-5	-0.25	-5	-0.56	-14	-0.26	-24	-0.65

Similar to its industry peer, Coca-Cola, PepsiCo products are available in more than 200 countries and territories (PepsiCo Inc., 2022). The company operates out of its North Carolina base under the direction of Ramon L. Laguarta, who has been the CEO since 2018 and as Chairman since 2019 (PepsiCo Inc., n.d., 2022). With a global workforce of approximately 315,000 as of the end of 2022, PepsiCo continues to be a major player in the food and beverage industry (PepsiCo Inc., 2022).

Even less extensive than Coca-Cola's, PepsiCo's ESG efforts are only covered in 61 articles, with 'Corporate Governance' being the primary focus in 24 articles, indicating a significant media focus on the company's management practices and policies. 'Social Opportunities' is another area of interest, highlighted in 10 articles, showing the media's attention to Pepsi's efforts in social engagements and community impact. The topic of 'Business Ethics' is covered in 9 pieces, reflecting considerations of PepsiCo's ethical conduct. Environmental issues such as 'Climate Change' and 'Resource Stewardship' are less frequently discussed, with 5 and 3 articles, respectively, and there are no articles addressing 'Environmental Opportunities,' suggesting these areas may be less scrutinized or perhaps better managed. 'Human Capital' and 'Product Liability' categories are also observed, with 3 and 7 articles each. The distribution of the sentiment for PepsiCo is portrayed in Figure 11.

The sentiment analysis reflects a primarily neutral stance, with the majority of articles, 44 out of 61, not leaning towards either a positive or negative sentiment. 'Corporate Governance' is predominantly viewed in a neutral light, suggesting a balanced media perspective on the company's management practices. However, with over half of the publicity being negative, 'Business Ethics' becomes a topic of concern over PepsiCo's ethical conduct. 'Product Liability' is discussed neutrally in 5 articles, with 2 articles expressing a negative sentiment. 'Social Opportunities' stands out with a more favorable view, dominating the positive sentiments. Environmental categories receive limited but mostly neutral coverage, with 'Climate Change' not receiving any positive or negative sentiments, while 'Resource Stewardship' leans toward the negative.

PepsiCo's ESG performance sentiment scores highlight critiques in certain areas while also recognizing some positive actions. 'Resource Stewardship' and 'Business Ethics' have a negative sentiment with a normalized (total) score of -0.67 (-2) and -0.56 (-5), respectively. In contrast, 'Social Opportunities' stands out positively with a normalized (total) score of 0.4 (4), showcasing PepsiCo's commitment to social impact, such as increasing access to chickpeas in Ethiopia and awarding grants to businesses promoting social causes (e.g., Newman, 2011; Strom, 2011). However, PepsiCo's commitment to reducing sugary drink calories (e.g., Strom, 2014b), indicating proactive steps in 'Social Opportunities,' is classified as neutral instead of positive. Many articles under 'Corporate Governance' are regarded as neutral, such as those related to leadership changes and investor relations (e.g., Creswell, 2018; Picker, 2016). The deceptive marketing and infiltration of political agenda in its charity organization contributes to its 'Business Ethics' negative assessment (e.g., Strom, 2010; Vega, 2011). In the 'Product Liability' articles, the media highlights the risk of type 2 diabetes for countries with higher usage of high-fructose corn syrup, like in PepsiCo's products (e.g., Bittman, 2012). The detailed scoring is available in the Appendix.

In comparison with other companies in its sector, PepsiCo is ranked AA by MSCI, which is slightly lower than Coca-Cola, yet PepsiCo is still seen as a top company (MSCI, 2023a). The issue of how PepsiCo gets its raw materials is noted by MSCI as a concern, though it is not widely discussed in the news (MSCI, 2023a). MSCI considers PepsiCo's performance in areas like company management, business ethics, waste management, and worker health and safety to be average (MSCI, 2023a). This matches with sentiment analysis findings, which suggest that PepsiCo's approach to managing its business is balanced. Nonetheless, the sentiment analysis pointed out some problems with PepsiCo's moral practices, whilst MSCI rated these as typical for the industry. MSCI views PepsiCo as leading the way in promoting better nutrition and health, managing water use, and ensuring product safety and quality, similar to Coca-Cola (MSCI, 2023a). However, the sentiment analysis does not fully acknowledge PepsiCo's improvements in these areas; instead, it highlights potential health risks from its product ingredients.

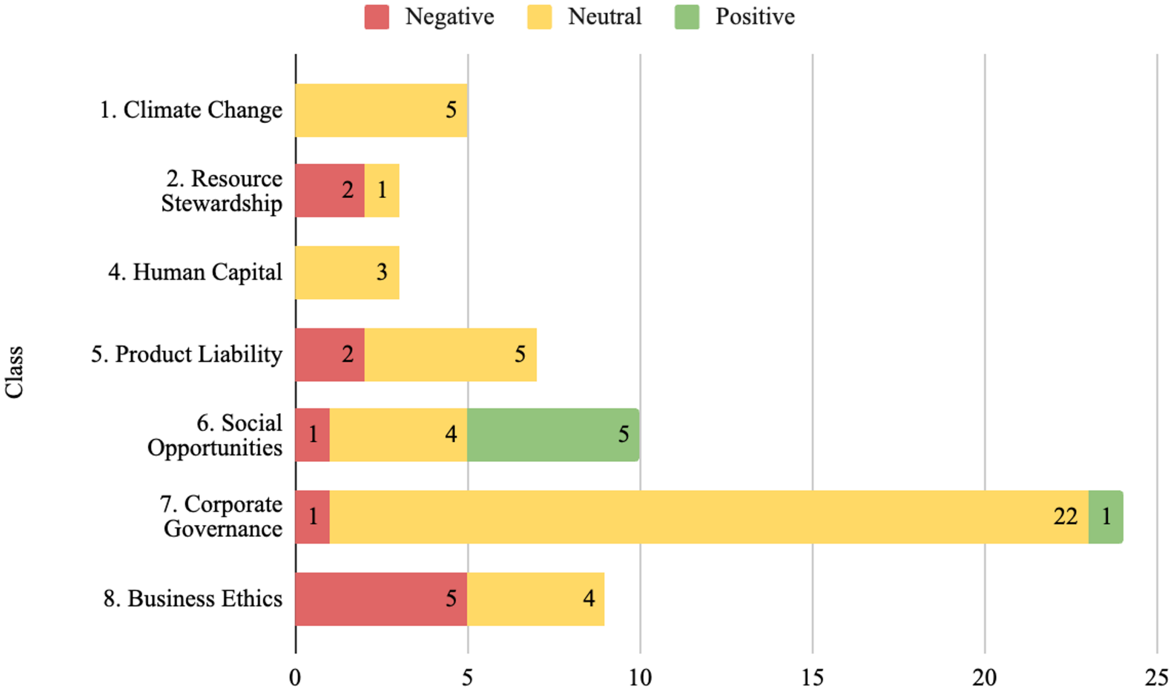


Figure 11: PepsiCo's Sentiment Distribution Across Different ESG Categories

4.2.7. Pfizer, Inc.

Pfizer, Inc. (Pfizer) is one of the biggest research-based pharmaceutical and biomedical corporations in the world, engaged in the creation, production, and global distribution of medical treatments (Nolen, 2023). Founded over a century ago in 1849 by the German chemist and entrepreneur Charles Pfizer and his cousin, Charles Erhart, the company has made significant strides in the field of medicine (Nolen, 2023). Notably, during the recent COVID-19 pandemic, Pfizer was at the forefront, developing the Comirnaty vaccine in collaboration with BioNTech, signifying an achievement in the global fight against the virus (Pfizer Inc., 2022). Pfizer is also recognized for pioneering medications such as Lipitor and Viagra, extending its reach to over 185 countries and territories (Pfizer Inc., 2022). The company is currently led by Dr. Albert Bourla as the chairman and CEO, guiding a dedicated workforce of 83,000 employees based out of their headquarters in New York (Pfizer Inc., n.d., 2022).

Pfizer receives extensive coverage of its ESG initiatives, spanning 364 articles. Among these, 'Social Opportunities' is the most discussed, with 143 articles highlighting the company's efforts in this area. 'Product Liability' also receives substantial attention, with 102 articles addressing the impacts of Pfizer's products. 'Business Ethics' is another focus, with 54 articles discussing the company's ethical practices. Following closely behind, 'Corporate Governance' is examined in 47 articles, reflecting scrutiny of the company's leadership and management decisions. 'Human Capital' emerges as well, with 15 articles potentially focusing on labor practices and workforce management. Environmental themes are less prominent; 'Resource Stewardship' and 'Environmental Opportunities' only account for 3 articles in total, pointing to a less pronounced focus on these issues. In addition, there is a noticeable absence of coverage of 'Climate Change,' indicating either a lack of effort or reporting in this critical area of ESG. Figure 12 visualizes the sentiment distribution for each category of Pfizer.

The sentiment analysis of Pfizer's ESG efforts skews towards a neutral viewpoint, with 204 articles maintaining an impartial tone. The 'Product Liability' discourse is mixed, with 43 articles reflecting negative sentiment and 24 articles expressing positivity, pointing to varied public reception and media reporting on the impact of Pfizer's products. 'Social Opportunities' reaps the brightest sentiment, with 49 articles acknowledging positive actions, significantly outweighing the 6 negative pieces, while environmental discussions under 'Resource Stewardship' and 'Environmental Opportunities' are discussed minimally but largely neutral. The majority of articles under 'Corporate Governance' depict a neutral stance, while 'Business Ethics' emerges as an area of concern, with 16 articles bearing a negative sentiment, possibly highlighting scrutiny over the company's ethical conduct.

It is evident that Pfizer's 'Social Opportunities' stands out with a positive total (normalized) score of 43 (0.30), suggesting a strong positive reception in this angle. However, areas such as 'Product Liability' and 'Business Ethics' are notable points of concern. 'Product Liability' has a particularly low total (normalized) score of -19 (-0.19), while 'Business Ethics' shows significant room for improvement with a total (normalized) score of -14 (-0.26). Concerns are also apparent in the 'Human Capital' category, where the negative sentiment is more pronounced at a normalized score of -0.27,

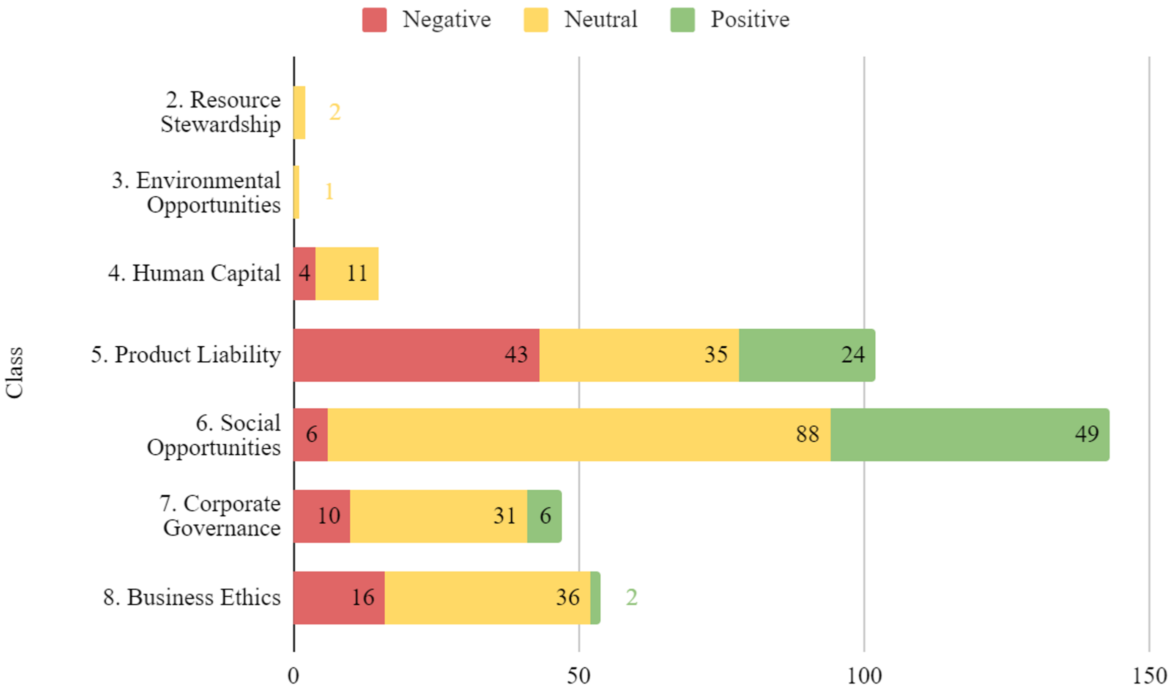


Figure 12: Pfizer's Sentiment Distribution Across Different ESG Categories

despite a total score of just -4. 'Corporate Governance' parallels this with a total score of -4 but with a less negative normalized score of -0.09, suggesting a predominantly neutral perception. In contrast, environmental categories have neither positive nor negative prominence, maintaining a neutral total score of 0. The detailed scoring is available in the Appendix.

Pfizer's efforts in tackling the global COVID-19 pandemic have been met with positive media coverage, especially regarding its role in providing the vaccines (e.g., LaFraniere, 2022; Zimmer, 2021a). Prior to the pandemic, Pfizer's contributions to healthcare innovation were also a frequent topic in the news (e.g., Pollack, 2014; Thomas, 2014). However, the company's products have attracted criticism for a variety of issues, including product defects and accusations of imitating rival technologies (e.g., Jewett, 2022; Robbins and Gross, 2022). Ethical issues have also surfaced, with Pfizer being implicated in acts of fraud such as tax avoidance and breaches of anti-racketeering laws (e.g., Barro, 2014; Bloomberg News, 2011). Discussions around corporate governance have been largely neutral, focusing on changes among senior leaders, but some negative attention has been drawn due to a failed merger attempt (e.g., Hughes, 2014).

MSCI assigns Pfizer a mid-range 'A' rating within the context of 260 pharmaceutical companies (MSCI, 2023a). This evaluation by MSCI acknowledges Pfizer's need for improvement in product safety and quality (MSCI, 2023a), which corresponds with sentiment analysis findings that point to negative news regarding product defects. Despite facing ethical issues, as reported in this study and indicated by MSCI's controversy measures, Pfizer is still rated as average, suggesting that its ethical challenges are less severe than those of many competitors. Notably, Pfizer is praised as a front-runner in providing healthcare access (MSCI, 2023a), highlighted by widespread news coverage of its COVID-19 vaccine development. Pfizer is also commended for its performance in corporate governance, investment in its workforce, and reducing harmful emissions and waste (MSCI, 2023a).

4.2.8. Johnson & Johnson

Established by the Johnson siblings—Robert, James, and Edward—in 1886, Johnson & Johnson has grown into a multinational powerhouse in both pharmaceuticals and medical device innovation (Johnson & Johnson, n.d.-a). Initially, their focus was on manufacturing sterile surgical products, such as sutures, absorbent cotton, and bandages (Johnson & Johnson, n.d.-a). In the present day, the company spans a wide range of health products, from everyday consumer goods like band-aids and baby powder to advanced pharmaceuticals for complex diseases and cutting-edge medical equipment for surgeries and orthopedics (Johnson & Johnson, 2023). During the COVID-19 pandemic, Johnson & Johnson expanded its health contributions by developing a vaccine to help curb the spread of the infectious disease. Under the leadership of Joaquin Duato as the CEO and Chairman, the company guides a dedicated manpower of over 150,000 people (Johnson & Johnson, n.d.-b, 2023). The global headquarters of Johnson & Johnson sit in New Brunswick, New Jersey, the same region where it began its journey (Johnson & Johnson, 2023).

Johnson & Johnson's ESG efforts are featured in 229 articles by The New York Times, less extensively covered than its competitor, Pfizer. Slightly different from Pfizer, 'Product Liability' emerges as the most scrutinized area, with 122 articles addressing concerns or developments related to the company's products and their impact on consumers. Following this, 'Social Opportunities' is covered in 48 articles, underscoring Johnson & Johnson's societal contributions. 'Business Ethics' is the subject of 37 articles, indicating a strong interest in the company's ethical standards. In contrast, 'Corporate Governance' features in only 15 articles, suggesting a more modest inquiry into the company's leadership. 'Human Capital' appears to be less of a focus, with only 5 articles. Environmental topics, namely 'Resource Stewardship' and 'Environmental Opportunities,' are the least represented in the coverage, with one article each. Neither Pfizer nor Johnson & Johnson has coverage related to 'Climate Change,' hinting at a possible underemphasis or lack of newsworthy events in this crucial ESG aspect for both companies. The distribution of the sentiments can be seen in Figure 13.

The analysis indicates a notable concentration of negative sentiment in 'Product Liability,' with 63 articles, highlighting scrutiny of the company's product safety and consumer impact. This is followed by 'Business Ethics,' where 25 articles carry a negative sentiment, suggesting a critical examination of the company's ethical conduct. Compared to Pfizer, Johnson & Johnson faces a more critical view in the category of 'Product Liability' and 'Business Ethics'. 'Social Opportunities' and 'Corporate Governance' exhibit more balanced coverage, with the majority of articles being marked as neutral. 'Human Capital' is also mostly discussed neutrally, while the environmental category, 'Resource Stewardship,' is viewed negatively in a single article, while 'Environmental Opportunities' gains a positive mention.

In assessing Johnson & Johnson's sentiment scores, it becomes clear that the company faces significant challenges in 'Product Liability,' where it has a notably negative total (normalized) score of -49 (-0.40). This is likely influenced by the company's role in the opioid epidemic, consumer litigation over product safety, and a number of product recalls (Hoffman, 2019; Hsu, 2021; Jiménez, 2021). 'Business Ethics' is another area with substantial negative sentiment, scoring -24 and normalized to -0.65, due to several lawsuits alleging improper marketing tactics, collusion, and bribery (Harris, 2011; Kanter & Thomas, 2013; Thomas, 2013). Conversely, 'Social Opportunities' shows a positive normalized (total) of 0.23, likely owing to Johnson & Johnson's efforts in developing and distributing COVID-19 vaccines and the development of new drugs (e.g., Pollack, 2012; Zimmer, 2021b). 'Corporate Governance' also carries a slight negative sentiment with a normalized (total) -0.13 (-2), slightly worse than its competitor, Pfizer. The detailed scoring is available in the Appendix.

It is important to note that the domains of 'Resource Stewardship,' 'Environmental Opportunities,' and 'Human Capital' are represented by a limited number of articles, resulting in total scores of -1, 1, and -1, with corresponding normalized scores -1.00, 1.00, and -0.20. The scant coverage in these areas introduces a degree of uncertainty to these scores, which may not fully capture the company's performance in these aspects. For instance, 'Environmental Opportunities' and 'Resource Stewardship' show very strong results, each with the highest possible normalized score. Yet, the trustworthiness of these scores is doubtful because they are based on a small amount of sentiment data.

MSCI rates Johnson & Johnson with an 'A' (Average), a ranking it shares with Pfizer within the pharmaceutical sector (MSCI, 2023a). Like Pfizer, Johnson & Johnson is recognized for its efforts to expand healthcare access, reduce harmful emissions and waste, and invest in employee development (MSCI, 2023a). The results from the sentiment score support this view, particularly highlighting Johnson & Johnson's role in distributing vaccines and developing new medications. However, the company's activities in employee development and waste management have not received much attention in the media. Regarding 'Corporate Governance,' Johnson & Johnson is considered average, while Pfizer is seen as performing better (MSCI, 2023a). Both MSCI's assessment and the sentiment scores align in identifying areas of concern for Johnson & Johnson, particularly in corporate behavior and product safety & quality, where MSCI marks it as underperforming (MSCI, 2023a). Ethical practices are a challenging area for Johnson & Johnson, with negative reports involving bribery and collusion. These ethical issues are confirmed by MSCI's controversy indicators, which note Johnson & Johnson's involvement in serious cases of bribery and fraud (MSCI, 2023a). In terms of product safety and quality, the company faces issues with product defects that cause health problems to the consumers, which are also underscored by MSCI's controversy indicators, showing the company's entanglement in significant controversies in this matter (MSCI, 2023a).

5. Discussions & Limitations

5.1. Discussions

Each industry discussed in Section 4 has its own characteristics. The technology sector, with its massive market capitalization, receives the biggest spotlight from the media. The number of articles featuring the gigantic technology companies' ESG activities is enormous compared to the coverage of other companies in other sectors. Following this, the pharmaceutical industry emerges as the second most covered sector, while the beverage sector receives the least media attention. In the technology sector, 'Business Ethics' and 'Product Liability' are the most discussed topics by the media. The ethical operations of these technology companies are mostly related to anticompetitive behavior. For instance, Apple had a case against Epic Games on how Apple abuses its power by charging developers unfairly high commissions to have a spot in Apple's App Store. Microsoft—the worst in terms of ethical conduct based on the findings, was under heavy scrutiny for its intent to acquire Activision since the deal might give an

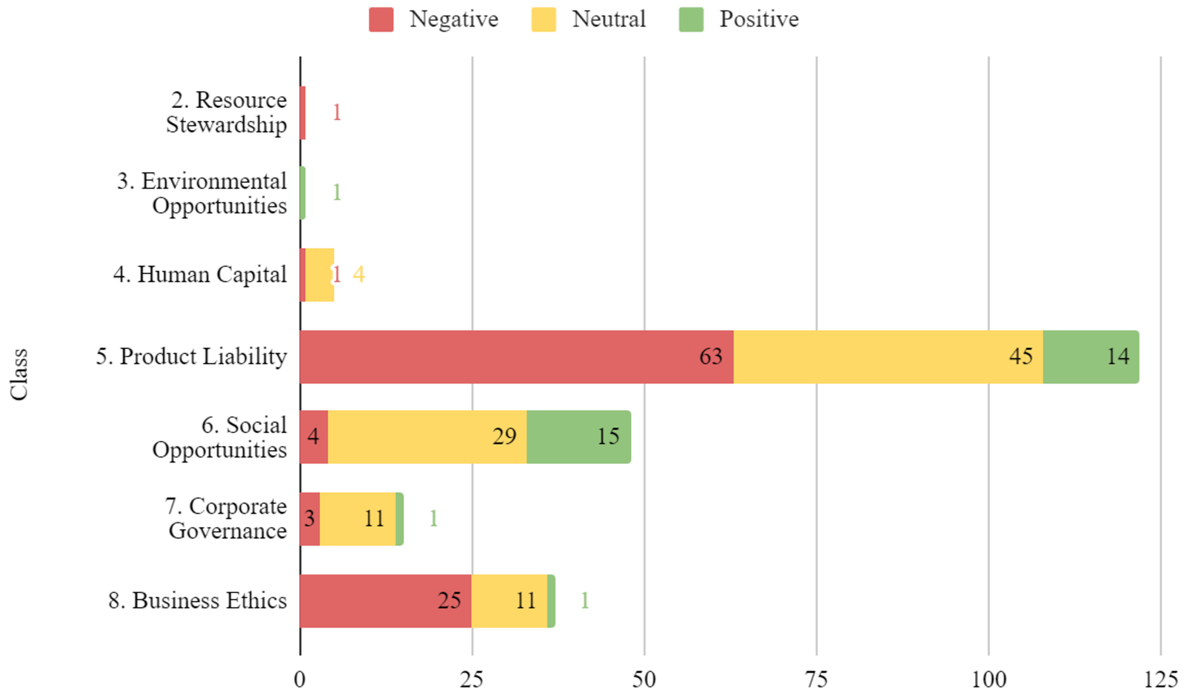


Figure 13: Johnson & Johnson's Sentiment Distribution Across Different ESG Categories

unfair competitive advantage to Microsoft (Weise & McCabe, 2022).

The ethical concerns for technology companies are distinct from those experienced in other industries. For other industries, a lot of issues are related to the company's unethical marketing and advertising tactics. In the beverage companies, for instance, the two companies have faced accusations of deceptive marketing strategies. The Coca-Cola Company encountered legal action over questionable claims regarding its juice products (Liptak, 2014), while PepsiCo and its subsidiary have been accused of engaging in misleading advertisements targeted at adolescents (Vega, 2011). The pharmaceutical sector also grapples with its share of ethical issues related to marketing. Pfizer was required to pay damages after being found in violation of federal racketeering laws through deceptive advertisements for one of its drugs (Bloomberg News, 2011). Johnson & Johnson faced penalties for unethical practices, including allegations of bribing doctors, which emerged during an investigation into its marketing tactics (Harris, 2011). These instances illustrate the varying ethical problems across different industries, shaped by the unique nature of the industry. Technology companies, often characterized by their large market share and dominant positions, face scrutiny over practices that could potentially harm competition, while other industries often confront ethical challenges centered around fraudulent marketing tactics.

The discourse surrounding 'Product Liability' in the technology sector also differs from that in other industries. This distinction primarily arises from the nature of the products and services offered by these industries. In the technology sector, the primary concerns highlighted in media discussions revolve around issues of data privacy and security, a direct consequence of the digital nature of the products. Taking Google as an example, it faced accusations of improperly collecting facial and voice recognition data without obtaining explicit consent from users (Hill & McCabe, 2022). Similarly, Meta was subjected to substantial fines for violating data privacy laws in Europe (Satariano, 2022b). In contrast, the pharmaceutical and beverage industries deal with more tangible products, leading to different kinds of product liability concerns. For example, in the pharmaceutical sector, a notable case is that of Johnson & Johnson, which had to recall sunscreen products due to contamination with harmful ingredients, raising serious questions about product safety and quality control (Jiménez, 2021). Meanwhile, beverage companies have been criticized for the potential health risks associated with their products (e.g., Bittman, 2012; Board, 2015).

While the technology industry has garnered media attention for ethical conduct and product liability issues, pharmaceutical companies have received prominent positive highlights in their contributions to social opportunities. Pfizer and Johnson & Johnson, not only because of their contributions in the recent pandemic, have been extensively covered by the media because of their breakthroughs in improving public health. A case in point was when Pfizer provided alternative contraceptives for developing countries beyond traditional daily birth control pills (Thomas, 2014). For Johnson & Johnson, the media acknowledged its effort in developing a new drug for prostate cancer (Pollack, 2012). In contrast, the media coverage of the beverage industry has been less favorable in terms of social contributions. Reports have highlighted a certain degree of hypocrisy, with companies like Coca-Cola and PepsiCo donating millions to health groups while simultaneously spending large sums to oppose public health legislation (O'Connor, 2016). In the technology sector, Microsoft is the leader in this area, with a normalized score of 0.24. However, the news articles do not necessarily report the social opportunities of its products. The articles can also include the altruistic actions of the company. For example, Microsoft pledged hundreds of millions of dollars to alleviate the housing crisis in the Seattle area (Weise, 2019). These varied narratives across industries reflect the diverse ways in which companies can impact society and how the media portrays these impacts.

Another interesting pattern that can be observed from the sentiment analysis is the limited coverage of the environmental efforts of the companies by the media. In the technology sector, which generally receives considerable media attention, a surprisingly small fraction, only about 1-3%, of the coverage is dedicated to the environmental theme. The situation appears even more pronounced in the pharmaceutical industry, where the share of environmental topics is less than 1%. Only the beverage sector receives a higher share of media coverage concerning environmental topics, by exceeding 10% of total coverage. That said, it is important to note that the news articles discussing the environmental aspect of PepsiCo predominantly feature outdated information, with seven out of eight articles being over 12 years old. In contrast, Coca-Cola has been the subject of a number of more recent news pieces. A notable example includes criticism from climate activists over Coca-Cola's sponsorship of a climate summit in Egypt in 2022, at a time when the company's production of plastics was reportedly on the rise (Engelbrecht, 2022). This pattern in media coverage might suggest that the nature of an industry significantly influences the extent of environmental reporting. The beverage industry's direct impact on resource utilization and waste production seemingly draws more media scrutiny compared to sectors like technology and pharmaceuticals.

It is crucial to underscore the importance of increased media coverage in the field of corporate environmental efforts. The current landscape, as indicated by the topic analysis results, shows a considerably low number of news articles covering companies' environmental actions. News outlets need to report more actions from the companies, both for the negative impact caused by the company and acknowledgment of positive milestones achieved by the companies. Negative coverage is essential not only for holding corporations accountable but also for informing and educating the public about the environmental impacts of various corporate actions. The media can report issues like carbon emissions, waste management, and resource utilization of the companies. Highlighting positive impacts is as crucial. This includes innovations in sustainable practices, successful implementation of environmentally friendly initiatives, significant reductions in carbon footprints, and investments in green and clean technology. Reporting on these areas can drive more informed consumer choices and potentially influence corporate policies.

Nevertheless, there is a prominent pattern in the 'Corporate Governance' theme that stands out as a resemblance across all industries. In this domain, companies from various sectors consistently exhibit scores that are either neutral or, at most, only slightly negative. Notably, Meta and Johnson & Johnson are at the lower spectrum with a score of -0.13, while Coca-Cola and PepsiCo maintain the highest scores with a neutral 0. This trend suggests a consistent but cautious media approach to corporate governance issues across various industries. A key aspect of the media coverage under this theme is the emphasis on significant changes in company leadership. For example, PepsiCo's CEO Indra Nooyi stepping down after 12 years (Creswell, 2018) and Sundar Pichai assuming the role of Alphabet's CEO are cases in point (Nicas & Wakabayashi, 2019). Such coverage is prevalent across various companies, lending a generally neutral tone to the discourse on corporate governance. Beyond leadership dynamics, the theme of corporate governance in media coverage extends to include topics like merger activities and executive pay. Instances such as Microsoft's acquisition of LinkedIn (Wingfield, 2016b) and Pfizer's proposed US$150 billion deal to buy Allergan (Merced, 2015) are illustrative of media focus on merger and acquisition cases. In another vein, executive compensation is a topic that caught media attention, particularly in the case of Coca-Cola, which faced shareholder criticism over its executive compensation (Eavis, 2014).

Compared to other research in this field, this study contributes to several key areas of improvement. J. Lee and Kim (2023) develop an ESG text classification model that allows researchers to extract ESG information from multiple sources, including reports and news articles. This study takes a leap forward by introducing a complete machine learning pipeline. It not only classifies ESG-related content but also extends to evaluating and scoring the companies' performance. In comparison to H. Kang and Kim (2022), who also utilized sentiment analysis but primarily focused on company sustainability reports, this research broadens the scope. It leverages machine learning to assess ESG performance from a third-party perspective, using news articles as the primary source. This approach provides a more balanced and external viewpoint on companies' ESG efforts compared to the self-reported nature of sustainability reports. Furthermore, this study surpasses the capabilities of the ESG-Miner tool presented by Fischbach et al. (2022). While ESG-Miner effectively identifies ESG relevance in news articles, it falls short in differentiating between the three main ESG categories. The model introduced in this study addresses this gap by providing a more detailed classification into nine distinct ESG categories, offering a finer and more insightful analysis of companies' ESG-related activities as portrayed in the media.

Furthermore, this study contributes a manually labeled ESG dataset comprising nine distinct classes, encompassing 4,500 news items. This dataset lays the groundwork for enhancing future ESG classification models. Additionally, the study illuminates the capabilities of GPT models, discovering that for classification tasks, fine-tuning a less costly model can yield better performance than using zero-shot prompting with more expensive models. The machine learning pipeline proposed in this research effectively demonstrates how companies are represented in media reports regarding their ESG performance, highlighting the value of machine learning as a tool for assessing corporate performance. The findings offer an in-depth analysis of companies' performance across eight different ESG categories, enabling the public to measure the achievement of these companies in these areas based on media reports. This approach not only provides a clearer picture of corporate sustainability performance but also underscores the potential of machine learning in extracting meaningful insights from complex datasets.

5.2. Limitations

As observed in Section 4, the prevailing sentiment expressed in news articles about companies' ESG performance is predominantly neutral. This trend could be attributed to a couple of reasons. Firstly, the media tends to maintain a neutral perspective when reporting news, which is reflected in the sentiment of the articles. Secondly, the sentiment analysis model employed in this study might have an inherent bias toward classifying sentiments as neutral. The news regarding PepsiCo's commitment to reducing sugary drink calories serves as a pertinent example (Strom, 2014b). This proactive step by PepsiCo is surprisingly deemed neutral by the model, whereas it could actually indicate a positive effort from the company. This bias could be due to the nature of the training data or the algorithmic design of the model. It is important to note that the model was primarily trained on financial communication texts rather than texts specifically focused on ESG topics. This difference in the nature of the training material could significantly influence the model's tendency to categorize ESG-related sentiments as neutral. This limitation suggests that the scores derived from the sentiment analysis may not comprehensively or accurately represent the actual ESG performance of the companies.

Another significant limitation in the current research methodology arises from the lack of integration between NER and sentiment analysis. While NER effectively identifies the presence of company names within the text, the sentiment analysis component captures the overall tone. However, this approach falls short of accurately discerning the nuanced sentiment directed specifically at a company's actions within the text. The complexity of this issue becomes evident when considering mixed messages. For instance, consider an article with the title and headline of 'Beverage Companies Embrace Recycling, Until It Costs Them. Recycling is struggling in much of the United States, and companies like Coca-Cola say they are committed to fixing it.' (Corkery, 2019). The model categorizes this article as 'Negative,' aligning with the overall tone that highlights the struggles in recycling efforts. However, this classification does not do justice to the specific sentiment towards Coca-Cola's roles. The company's commitment to addressing recycling issues, a potentially positive aspect, is overlooked in this analysis. This example demonstrates how the model's inability to parse and evaluate the sentiment related to specific actions or statements of a company can lead to an incomplete understanding of the sentiment in the text.

Employing entity-level sentiment analysis for evaluating a company's performance can enhance the accu-racy of a company's sentiment. Rather than deriving a generalized sentiment from the overall tone of the text, entity-level sentiment analysis focuses on the sentiment directly associated with a particular company mentioned in the text (Sinha et al., 2023). This method is proficient at unraveling the complex layers of sentiment related to multiple entities within a single piece of text. Future research could greatly benefit from adopting this method, as demonstrated in recent studies like those of Rønningstad et al. (2023), Sinha et al. (2023), and Tang et al. (2023). Sinha et al. (2023) and Tang et al. (2023) propose innovative frameworks for extracting sentiments specifically relevant to each entity in the financial domain. Implementing entity-level sentiment analysis would allow for a more targeted and accurate assessment of how companies are perceived in regard to their ESG efforts. Nevertheless, the limitations highlighted earlier in this study underscore the necessity of integrating human evaluation into the analysis to enhance its reliability. While entity-level sentiment analysis is advancing, especially in the context of ESG-related issues, it remains a developing field. The combination of automatic and human evaluation approaches would complement each other, potentially leading to a more accurate representation of companies' actual ESG performance.

The current ESG classification model is designed to categorize text into a single ESG class. This approach presents another limitation, where it fails to account for the complexity and multifaceted nature of many news articles. Multiple ESG-related topics might be intertwined within a single piece of text. Given this limitation, a promising avenue for future research would be to explore models capable of multi-class classification. This would allow for a more nuanced and detailed understanding of the ESG topics covered in the media. Further, augmenting this multi-class approach with an entity-level sentiment analysis could offer an ever richer and more accurate analysis. By not only identifying multiple ESG themes but also associating these themes with specific entities and their corresponding sentiments. This advancement could significantly enhance the depth of ESG performance assessments derived from media analysis.

As highlighted in the previous sections, one of the limitations identified in this study is the uneven coverage of certain topics and industries, leading to an incomplete assessment of companies' ESG performance. Environmental topics, for example, there are relatively few articles discussing the efforts of companies on this front. The limited media focus contrasts with evaluation from entities like MSCI, which provide comprehensive ratings of companies' performance in environmental initiatives. Many companies, including Microsoft, Alphabet, Coca-Cola, and PepsiCo, are recognized as leaders in each group in efforts to curb carbon emissions by MSCI. This leadership, however, is not fully reflected in the sentiment analysis due to the sparse coverage these initiatives receive in the media. Furthermore, the extent of media coverage varies significantly among companies. For instance, beverage companies like Coca-Cola and PepsiCo receive less media attention compared to larger technology sector companies. This imbalance presents a challenge in accurately gauging the ESG performance of less-covered companies, which may be actively making significant efforts in various ESG areas. Consequently, there exists a notable discrepancy between the ESG performance ratings provided by organizations like MSCI and the findings from sentiment analysis—especially for low-covered topics and companies, a disparity that can largely attributed to the limitations in media coverage. To address this challenge, future research could expand the coverage of news articles by incorporating additional news outlets as sources of input.

Nonetheless, this study demonstrates that the machine learning pipeline can be a valuable tool for analysts and the public to gauge companies' ESG performance as portrayed in the media. While the pipeline has shown potential, it requires further refinement for more accurate assessments, as outlined earlier. Despite its current limitations, this model serves as an excellent starting point for sorting through the vast quantity of news articles. This ability to filter and analyze large datasets is a significant advantage in understanding the often-complex narrative of corporate ESG performance from the media. It is also important to acknowledge the variability in ESG ratings across different organizations is a commonly observed phenomenon as described by Berg et al. (2022) and Chatterji et al. (2016). In this study, adding a news analysis can offer additional insights. By evaluating how companies are portrayed in the media, this study provides a different angle, potentially enriching the understanding of a company's ESG efforts.

6. Conclusion

Driven by the significant impact of media on public perception and the growing interest in OpenAI's GPT models, I embark on an analysis of news reports to assess companies' ESG performance. This assessment is conducted using a machine learning pipeline that incorporates various models to perform a range of tasks, including NER, classification, and sentiment analysis. While this paper utilizes available open-source models for performing NER and sentiment analysis tasks, this study also examines various GPT models for classification tasks. Interestingly, the findings reveal that a fine-tuned model, even one that is more cost-effective and smaller in scale (i.e., GPT-3, ada), can outperform the zero-shot prompting capabilities of a larger, more expensive model (i.e., GPT-3.5-turbo) in classifying nine distinct ESG topics.

In evaluating the sustainability performance of corporations, this paper focuses on articles published by the New York Times between 2003 and 2022. The analysis comprises eight public corporations spanning three distinct sectors. Despite covering a broad time span, the volume of media coverage varies significantly across these sectors. The technology sector receives the most attention, while the beverage industry garners the least number of reports in the media. Based on these media reports, the sentiment of the articles related to each ESG topic is assessed. These sentiment analysis results are then utilized to generate scores for the companies, providing a measure of their sustainability performance as portrayed in the media.

The results of this study exhibit unique characteristics in media reporting for each sector. The technology sector, in particular, faces considerable media scrutiny concerning its ethical operations and product liability. Key issues here include antitrust litigations and concerns related to privacy and data security. In contrast, the pharmaceutical industry attracts more media attention regarding its contributions to public health, especially due to the COVID-19 pandemic. Meanwhile, beverage companies receive a notably higher proportion of media coverage on environmental topics than other sectors. This interest primarily focuses on the natural resource usage and waste management practices of these companies. However, the media narrative appears more uniform across the companies studied when it comes to corporate governance, with a common emphasis placed on changes in company leadership. The sentiment in the corporate governance theme tends to be balanced and generally neutral across all eight companies.

This study highlights several limitations in the current research methodology, including limited media coverage for certain topics and sectors, inaccuracies in sentiment analysis classification, the lack of integration among machine learning models, and the inability to categorize complex texts into multiple classes. Addressing these limitations could greatly enhance further research in measuring sustainability performance using machine learning algorithms. For instance, incorporating a broader range of news sources could expand media coverage, potentially addressing the issue of limited coverage observed in this study. An improvement in the current machine learning pipeline could involve the integration of various models. Combining entity-level sentiment analysis with ESG topic analysis would lead to a more accurate assessment of companies' actions.

The potential impact of this study is a significant contribution to providing a comprehensive analysis of companies' ESG performance across various industries, using news articles as the primary data source. It provides an in-depth look at how various industries and companies are portrayed in the media with respect to their ESG initiatives. This approach is particularly insightful as it reveals the distinct nature of ESG reporting across different sectors. Moreover, this research sheds light on areas where the media could potentially enhance its role in shaping public perception of ESG efforts. It suggests that a more balanced and comprehensive coverage of ESG topics, including both achievements and areas of concern, could provide a more accurate picture of companies' ESG performance. Additionally, the study implies that increased coverage of underrepresented sectors and topics could contribute to a well-informed public discourse.

References

Allen, K. (2018). Lies, Damned Lies and ESG Rating Methodologies. *Financial Times*. https://www.ft.com/content/2e49171b-a018-3c3b-b66b-81fd7a170ab5

Alphabet Inc. (2015). Form 8-K 2015. https://www.sec.gov/Archives/edgar/data/1652044/000119312515336577/d82837d8k12b.htm

Alphabet Inc. (2019). Alphabet Management Change. https://abc.xyz/investor/news/2019/1203/

Alphabet Inc. (2022). Form 10-K 2022. https://www.sec.gov/Archives/edgar/data/1652044/000165204423000016/goog-20221231.htm

Apple Inc. (n.d.). Tim Cook. Retrieved October 31, 2023, from https://www.apple.com/leadership/tim-cook/

Apple Inc. (2022). Form 10-K 2022. https://s2.q4cdn.com/470004039/files/doc_financials/2022/q4/_10-K-2022-(As-Filed).pdf

Aue, T., Jatowt, A., & Färber, M. (2022). Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis. *arXiv preprint arXiv:2203.00000*.

Austen, I. (2019). Trash-Picking Robots? Park Bench Monitors? Toronto Debates Tech Giant's Waterfront Plans. *The New York Times*. https://www.nytimes.com/2019/06/24/world/canada/toronto-google-sidewalk-labs.html

Barka, Z., Hamza, T., & Mrad, S. (2023). Corporate ESG Scores and Equity Market Misvaluation: Toward Ethical Investor Behavior. *Economic Modelling*, *127*, 106467. https://doi.org/10.1016/j.econmod.2023.106467

Barro, J. (2014). Pfizer's Move Poses Challenge. Here's a Solution. *The New York Times*. https://www.nytimes.com/2014/04/30/upshot/radical-solution-to-challenge-of-corporate-taxes.html

Berg, F., Fabisik, K., & Sautner, Z. (2020). Rewriting History II: The (Un)Predictable Past of ESG Ratings. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3722087

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, *26*(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap Talk and Cherry-Picking: What ClimateBert has to Say on Corporate Climate Risk Disclosures. *Finance Research Letters*, *47*, 102776. https://doi.org/10.1016/j.frl.2022.102776

Bittman, M. (2012). Unseasonably Warm Winter Links. *The New York Times*. https://archive.nytimes.com/bittman.blogs.nytimes.com/2012/12/05/unseasonably-warm-winter-links/

Bloomberg News. (2011). Pfizer Told to Pay $142.1 Million Over Marketing of Epilepsy Drug. *The New York Times*. https://www.nytimes.com/2011/01/29/business/29pfizer.html

Board, T. E. (2015). Opinion: Coke Tries to Sugarcoat the Truth on Calories. *The New York Times*. https://www.nytimes.com/2015/08/14/opinion/coke-tries-to-sugarcoat-the-truth-on-calories.html

Borms, S., Boudt, K., Holle, F. V., & Willems, J. (2021). Semi-supervised Text Mining for Monitoring the News About the ESG Performance of Companies. *Data Science for Economics and Finance*, 217–239. https://doi.org/10.1007/978-3-030-66891-4_10

Brackley, A., Brock, E. K., & Nelson, J. (2022, October). Rating the Raters Yet Again: Six Challenges for ESG Ratings. *ERM Sustainability Institute*. https://www.sustainability.com/thinking/rating-the-raters-yet-again-six-challenges-for-esg-ratings/

Britannica. (2023a). PepsiCo, Inc. https://www.britannica.com/topic/PepsiCo-Inc

Britannica. (2023b). The Coca-Cola Company. https://www.britannica.com/topic/The-Coca-Cola-Company

Brown, N., & Deegan, C. (1998). The Public Disclosure of Environmental Performance Information–A Dual Test of Media Agenda Setting Theory and Legitimacy Theory. *Accounting and Business Research*, *29*(1), 21–41. https://doi.org/10.1080/00014788.1998.9729564

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language Models are Few-Shot Learners.

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural Legal Judgment Prediction in English.

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strategic Management Journal*, *37*(8), 1597–1614. https://doi.org/10.1002/smj.2407

Chen, Q., & Liu, X.-Y. (2020). Quantifying ESG Alpha Using Scholar Big Data. *Proceedings of the First ACM International Conference on AI in Finance*, 1–8. https://doi.org/10.1145/3383455.3422529

Conger, K., & Frenkel, S. (2021, March). Thousands of Microsoft Customers May Have Been Victims of Hack Tied to China.

Corkery, M. (2019). Beverage Companies Embrace Recycling, Until It Costs Them. *The New York Times*. https://www.nytimes.com/2019/07/04/business/plastic-recycling-bottle-bills.html

Creswell, J. (2018). Indra Nooyi, PepsiCo C.E.O. Who Pushed for Healthier Products, to Step Down. *The New York Times*. https://www.nytimes.com/2018/08/06/business/indra-nooyi-pepsi.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Douglas, M. R. (2023). Large Language Models.

Eavis, P. (2014). Coca-Cola, Yielding to Criticism, Revises Its Proposal for Executive Pay. *The New York Times*. https://www.nytimes.com/2014/10/02/business/coca-cola-yielding-to-criticism-revises-its-plan-for-executive-pay.html

Elias, J. (2023). Google is Asking Employees to Test Potential ChatGPT Competitors, Including a Chatbot Called 'Apprentice Bard'. *CNBC*. https://www.cnbc.com/2023/01/31/google-testing-chatgpt-like-chatbot-apprentice-bard-with-employees.html

Engelbrecht, C. (2022). Coke Is a Sponsor of the Climate Summit in Egypt. Some Activists Aren't Happy. *The New York Times*. https://www.nytimes.com/2022/11/07/climate/coca-cola-sponsor-cop27-climate-egypt.html

Fischbach, J., Adam, M., Dzhagatspanyan, V., Mendez, D., Frattini, J., Kosenkov, O., & Elahidoost, P. (2022). Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool.

Frenkel, S. (2018). Microsoft Employees Protest Work With ICE, as Tech Industry Mobilizes Over Immigration. *The New York Times*. https://www.nytimes.com/2018/06/19/technology/tech-companies-immigration-border.html

Frenkel, S. (2022a). Meta Will Freeze Most Hiring, Zuckerberg Tells Employees. *The New York Times*. https://www.nytimes.com/2022/09/29/technology/meta-hiring-freeze.html

Frenkel, S. (2022b). Meta Is Said to Plan Significant Job Cuts This Week. *The New York Times*. https://www.nytimes.com/2022/11/06/technology/meta-layoffs.html

Gordon, C. (2023). ChatGPT Is The Fastest Growing App In The History Of Web Applications. *Forbes*. https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/?sh=10b5be27678c

Gough, N., & Chen, B. X. (2014). Groups Accuse Apple Supplier in China of Labor Violations. *The New York Times*. https://www.nytimes.com/2014/09/05/business/Apple-Supplier-Is-Accused-of-Labor-Violations.html

Goutte, S., Le-Hoang, V. P., Liu, F., & von Mettenheim, H.-J. (2023). ESG Investing: A Sentiment Analysis Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4316108

Grady, D. (2020). A.I. Is Learning to Read Mammograms. *The New York Times*. https://www.nytimes.com/2020/01/01/health/breast-cancer-mammogram-artificial-intelligence.html

Guo, T., Jamet, N., Betrix, V., Piquet, L.-A., & Hauptmann, E. (2020). ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction.

Hall, M., & Hosch, W. L. (2023). Google. https://www.britannica.com/topic/Google-Inc

Hammami, A., & Hendijani Zadeh, M. (2019). Audit quality, media coverage, environmental, social, and governance disclosure and firm investment efficiency. *International Journal of Accounting & Information Management*, *28*(1), 45–72. https://doi.org/10.1108/IJAIM-03-2019-0041

Hardy, Q. (2016). Google Says It Will Run Entirely on Renewable Energy in 2017. *The New York Times*. https://www.nytimes.com/2016/12

/06/technology/google-says-it-will-run-entirely-on-renewable-energy-in-2017.html

Harris, G. (2011). Johnson & Johnson Settles Bribery Complaint for $70 Million in Fines. *The New York Times*. https://www.nytimes.com/2011/04/09/business/09drug.html

Hartzmark, S. M., & Sussman, A. B. (2019). Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows. *The Journal of Finance*, *74*(6), 2789–2837. https://doi.org/10.1111/jofi.12841

Hill, K., & McCabe, D. (2022). Texas Sues Google for Collecting Biometric Data Without Consent. *The New York Times*. https://www.nytimes.com/2022/10/20/technology/texas-google-privacy-lawsuit.html

Hirai, A., Brady, A., & Partners, S. (2021, July). Managing ESG Data and Rating Risk. https://corpgov.law.harvard.edu/2021/07/28/managing-esg-data-and-rating-risk/

Hoffman, J. (2019). Johnson & Johnson Ordered to Pay $572 Million in Landmark Opioid Trial. *The New York Times*. https://www.nytimes.com/2019/08/26/health/oklahoma-opioids-johnson-and-johnson.html

Holpuch, A. (2022). Two Women Sue Apple Over AirTag Stalking. *The New York Times*. https://www.nytimes.com/2022/12/06/business/apple-airtag-lawsuit.html

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hsu, T. (2021). Black Women's Group Sues Johnson & Johnson Over Talc Baby Powder. *The New York Times*. https://www.nytimes.com/2021/07/27/business/johnson-baby-powder-black-women.html

Huang, A. (2022). Description of 9-class Environmental, Social and Governance (ESG) Classification. https://www.allenhuang.org/uploads/2/6/5/5/26555246/esg_9-class_descriptions.pdf

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Hughes, C. (2014). The Lessons of a Drug Maker's Failed Deal. *The New York Times*. https://dealbook.nytimes.com/2014/05/27/the-lessons-of-a-failed-drug-deal/

Ilango, H. (2023, May). An Unregulated ESG Rating System Reveals Its Flaws. *Institute for Energy Economics & Financial Analysis*.

Isaac, M. (2018). Instagram's Co-Founders to Step Down From Company. *The New York Times*. https://www.nytimes.com/2018/09/24/technology/instagram-cofounders-resign.html

Isaac, M. (2021). Facebook and Its Apps Suffer Another Outage. *The New York Times*. https://www.nytimes.com/2021/10/08/technology/facebook-whatsapp-instagram-down.html

Jewett, C. (2022). Pfizer Recalls Some Blood Pressure Drugs, Citing Cancer Risk. *The New York Times*. https://www.nytimes.com/2022/03/23/health/pfizer-recall-blood-pressure-drug-cancer.html

Jiménez, J. (2021). Johnson & Johnson Recalls Sunscreen Because of Benzene Traces. *The New York Times*. https://www.nytimes.com/2021/07/14/us/johnson-johnson-sunscreen-recall-aveeno-neutrogena.html

Johnson & Johnson. (n.d.-a). Our Beginning. Retrieved October 31, 2023, from https://ourstory.jnj.com/our-beginning

Johnson & Johnson. (n.d.-b). Our Leadership Team. Retrieved October 31, 2023, from https://www.jnj.com/our-leadership-team

Johnson & Johnson. (2023). Form 10-K 2022. https://johnsonandjohnson.gcs-web.com/static-files/06bc3388-603b-4768-bf95-e6d43fda9fd3

Johnston, M. (2022). Biggest Companies in the World by Market Cap. *Investopedia*. https://www.investopedia.com/biggest-companies-in-the-world-by-market-cap-5212784

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and Applications of Large Language Models.

Kang, C., Isaac, M., & Popper, N. (2019). Facebook's Zuckerberg, Accused of Lying, Withstands a Washington 'Beating'. *The New York Times*. https://www.nytimes.com/2019/10/23/technology/facebook-zuckerberg-libra-congress.html

Kang, H., & Kim, J. (2022). Analyzing and Visualizing Text Information in Corporate Sustainability Reports Using Natural Language Processing Methods. *Applied Sciences*, *12*(11), 5614. https://doi.org/10.3390/app12115614

Kanter, J., & Thomas, K. (2013). Europe Says Drug Makers Paid to Delay a Generic. *The New York Times*. https://www.nytimes.com/2013/02/01/business/global/eu-says-drug-makers-paid-to-delay-generic-version.html

Kelly, S. M. (2023). ChatGPT Passes Exams from Law and Business Schools. *CNN Business*. https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html

Krappel, T., Bogun, A., & Borth, D. (2021). Heterogeneous Ensemble for ESG Ratings Prediction.

LaFraniere, S. (2022). New Booster Shot Targets Covid Variants More Effectively, Pfizer Says. *The New York Times*. https://www.nytimes.com/2022/11/04/us/politics/covid-booster-pfizer.html

Lattman, P. (2012). Former Coca-Cola Bottling Executive Charged With Insider Trading. *The New York Times*. https://dealbook.nytimes.com/2012/03/08/s-e-c-charges-former-coca-cola-bottling-executive-with-insider-trading/

Lee, H., Choi, J., Kwon, S., & Jung, S. (2023). EaSyGuide: ESG Issue Identification Framework Leveraging Abilities of Generative Large Language Models.

Lee, J., & Kim, M. (2023). ESG Information Extraction with Cross-Sectoral and Multi-Source Adaptation Based on Domain-Tuned Language Models. *Expert Systems with Applications*, *221*, 119726. https://doi.org/10.1016/j.eswa.2023.119726

Leite, B. J., & Uysal, V. B. (2023). Does ESG Matter to Investors? ESG Scores and the Stock Price Response to New Information. *Global Finance Journal*, *57*, 100851. https://doi.org/10.1016/j.gfj.2023.100851

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.

Li, H. (2022). Language Models: Past, Present, and Future. *Communications of the ACM*, *65*(7), 56–63. https://doi.org/10.1145/3490443

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge.

Linzmayer, O. (2004). *Apple Confidential 2.0: The Definitive History of the World's Most Colorful Company*. No Starch Press.

Liptak, A. (2014). Coke Can Be Sued by Rival Over Juice Claim, Court Says. *The New York Times*. https://www.nytimes.com/2014/06/13/business/supreme-court-says-coca-cola-can-be-sued-by-Pom-Wonderful.html

Lohr, S. (2020). Slack Accuses Microsoft of Illegally Crushing Competition. *The New York Times*. https://www.nytimes.com/2020/07/22/technology/slack-microsoft-antitrust.html

Luccioni, A., Baylor, E., & Duchene, N. (2020). Analyzing Sustainability Reports Using Natural Language Processing.

Mac, R. (2022). Lawsuit says Meta shares blame in the killing of a federal guard. *The New York Times*. https://www.nytimes.com/2022/01/06/technology/meta-facebook-lawsuit-security-guard.html

Malo, P., Sinha, A., Takala, P., Korhonen, P., & Wallenius, J. (2013). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts.

Markoff, J. (2003). TECHNOLOGY; RealNetworks Accuses Microsoft Of Restricting Competition. *The New York Times*. https://www.nytimes.com/2003/12/19/business/technology-realnetworks-accuses-microsoft-of-restricting-competition.html

McCombs, M., & Reynolds, A. (2002). *Media Effects* (J. Bryant, D. Zillmann, J. Bryant, & M. B. Oliver, Eds.). Routledge. https://doi.org/10.4324/9781410602428

Mehra, S., Louka, R., & Zhang, Y. (2022). ESGBERT: Language Model to Help with Classification Tasks Related to Companies' Environmental, Social, and Governance Practices. *Embedded Systems and Applications*, 183–190. https://doi.org/10.5121/csit.2022.120616

Mele, C. (2016). Bags of Cocaine Worth $56 Million Are Found at Coca-Cola Factory in France. *The New York Times*. https://www.nytimes.co m/2016/09/02/world/europe/bags-of-cocaine-worth-56-million-are-found-at-coca-cola-factory-in-france.html

Merced, M. J. d. l. (2015). Pfizer and Allergan Said to Be Near Merger for Up to $150 Billion. *The New York Times*. https://www.nytimes.com/2015/11/19/business/dealbook/pfizer-allergan-deal-for-up-to-150-billion-is-said-to-be-close-to-complete.html

Meta Platforms Inc. (n.d.). Mark Zuckerberg, Founder, Chairman and Chief Executive Officer. Retrieved October 31, 2023, from https://about.meta.com/media-gallery/executives/mark-zuckerberg/

Meta Platforms Inc. (2022). Form 10-K 2022. https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/e574646c-c642-42d9-9229-3892b13aabfb.pdf

Microsoft. (n.d.-a). Facts About Microsoft. Retrieved February 12, 2023, from https://news.microsoft.com/facts-about-microsoft/

Microsoft. (n.d.-b). Satya Nadella. Retrieved October 31, 2023, from https://news.microsoft.com/exec/satya-nadella/

Microsoft. (2023). Form 10-K 2022. https://microsoft.gcs-web.com/node/31736/html

Morales, C. (2022). Restaurants Face an Extortion Threat: A Bad Rating on Google. *The New York Times*. https://www.nytimes.com/2022/07/11/dining/google-one-star-review-scam-restaurants.html

MSCI. (2023a). ESG Ratings & Climate Search Tool. https://www.msci.com/our-solutions/esg-investing/esg-ratings-climate-search-tool

MSCI. (2023b). ESG Ratings Methodology.

Newman, A. A. (2011). Good/Corps Aims to Help Business Meet Social Goals. *The New York Times*. https://www.nytimes.com/2011/05/13/business/media/13adco.html

Nguyen, Q., Diaz-Rainey, I., & Kuruppuarachchi, D. (2020). Predicting Corporate Carbon Footprints for Climate Finance Risk Analyses: A Machine Learning Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3617175

Nicas, J. (2019). Apple Removes App That Helps Hong Kong Protesters Track the Police. *The New York Times*. https://www.nytimes.com/2019/10/09/technology/apple-hong-kong-app.html

Nicas, J., Browning, K., & Griffith, E. (2020). Fortnite Creator Sues Apple and Google After Ban From App Stores. *The New York Times*. https://www.nytimes.com/2020/08/13/technology/apple-fortnite-ban.html

Nicas, J., & Wakabayashi, D. (2019). Era Ends for Google as Founders Step Aside From a Pillar of Tech. *The New York Times*. https://www.nytimes.com/2019/12/03/technology/google-alphabet-ceo-larry-page-sundar-pichai.html

Nolen, J. L. (2023). Pfizer, Inc. https://www.britannica.com/topic/Pfizer-Inc

Nugent, T., Stelea, N., & Leidner, J. L. (2020). Detecting ESG topics using domain-specific language models and data augmentation approaches.

O'Connor, A. (2016). Coke and Pepsi Give Millions to Public Health, Then Lobby Against It. *The New York Times*. https://www.nytimes.com/2016/10/10/well/eat/coke-and-pepsi-give-millions-to-public-health-then-lobby-against-it.html

OpenAI. (n.d.). Fine-tuning (Legacy). Retrieved October 27, 2023, from https://platform.openai.com/docs/guides/legacy-fine-tuning/fine-tuning

OpenAI. (2022, October). Introducing ChatGPT. https://openai.com/blog/chatgpt

Papers With Code. (n.d.). Text Classification on financial_phrasebank. Retrieved October 18, 2023, from https://paperswithcode.com/sota/text-classification-on-financial-phrasebank

PepsiCo Inc. (n.d.). Our Leadership. Retrieved October 31, 2023, from https://www.pepsico.com/who-we-are/leadership

PepsiCo Inc. (2022). Form 10-K 2022. https://investors.pepsico.com/docs/default-source/investors/q4-2022/q4-2022-form-10k_hmielz4d40rd4s16.pdf

Peric, L., Mijic, S., Stammbach, D., & Ash, E. (2020). Legal Language Modeling with Transformers. *CEUR Workshop Proceedings*. https://doi.org/10.3929/ethz-b-000456079

Pfizer Inc. (n.d.). Executive Leadership. Retrieved October 31, 2023, from https://www.pfizer.com/about/people/executives

Pfizer Inc. (2022). Form 10-K 2022. https://www.sec.gov/ix?doc=/Archives/edgar/data/78003/000007800323000024/pfe-20221231.htm

Picker, L. (2016). Billionaire Investor Nelson Peltz Sells Stake in PepsiCo. *The New York Times*. https://www.nytimes.com/2016/05/14/business/dealbook/billionaire-investor-nelson-peltz-sells-stake-in-pepsico.html

Polignano, M., Bellantuono, N., Lagrasta, F. P., Caputo, S., Pontrandolfo, P., & Semeraro, G. (2022). An NLP Approach for the Analysis of Global Reporting Initiative Indexes from Corporate Sustainability Reports. *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, 1–8. https://aclanthology.org/2022.csrnlp-1.1

Pollack, A. (2012). Trial Shows Benefit in Using Prostate Cancer Drug Early. *The New York Times*. https://www.nytimes.com/2012/03/09/business/trial-shows-benefit-in-earlier-use-of-zytiga-for-prostate-cancer.html

Pollack, A. (2014). Guarded Optimism After Breast Cancer Drug Shows Promising Results. *The New York Times*. https://www.nytimes.com/2014/04/07/business/breast-cancer-drug-shows-groundbreaking-results.html

Principles for Responsible Investment (PRI). (2023). PRI Annual Report: Responsible investment ecosystems. https://www.unpri.org/annual-report-2023/responsible-investment-ecosystems#fn_link_1

PwC. (2022). Asset and wealth management revolution 2022: Exponential expectations for ESG. https://www.pwc.com/gx/en/financial-services/assets/pdf/pwc-awm-revolution-2022.pdf

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Robbins, R., & Gross, J. (2022). Moderna Sues Pfizer and BioNTech Over Covid Vaccine Technology. *The New York Times*. https://www.nytimes.com/2022/08/26/business/moderna-covid-vaccine-lawsuit.html

Rønningstad, E., Velldal, E., & Øvrelid, L. (2023). Entity-Level Sentiment Analysis (ELSA): An exploratory task survey. https://arxiv.org/pdf/2304.14241.pdf

Satariano, A. (2020). Facebook Loses Antitrust Decision in Germany Over Data Collection. *The New York Times*. https://www.nytimes.com/2020/06/23/technology/facebook-antitrust-germany.html

Satariano, A. (2021). Apple's App Store Draws E.U. Antitrust Charge. *The New York Times*. https://www.nytimes.com/2021/04/30/technology/apple-antitrust-eu-app-store.html

Satariano, A. (2022a). A secret ad deal between Google and Meta is under scrutiny in Europe. *The New York Times*. https://www.nytimes.com/2022/03/11/business/google-meta-eu-britain-inquiry.html

Satariano, A. (2022b). Meta Fined $275 Million for Breaking E.U. Data Privacy Law. *The New York Times*. https://www.nytimes.com/2022/11/28/business/meta-fine-eu-privacy.html

Satariano, A., & Frenkel, S. (2022). Oversight Board Criticizes Meta for Preferential Treatment. *The New York Times*. https://www.nytimes.com/2022/12/06/technology/meta-preferential-treatment.html

Scheiber, N. (2020). Labor Board Accuses Google Contractor of Violating Union Rights. *The New York Times*. https://www.nytimes.com/2020/10/08/business/google-nlrb-hcl-union.html

Schwartz, J. (2015). Coca-Cola Says It's Close to Water Replenishment Goal. *The New York Times*. https://www.nytimes.com/2015/08/26/business/coca-cola-expects-to-reach-its-water-replenishment-goal-5-years-early.html

Scott, M. (2015). Skype Service Problems for Some Users Worldwide. *The New York Times*. https://www.nytimes.com/2015/09/22/technology/skype-service-disrupted-for-some-users-worldwide.html

Sejnowski, T. J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, *35*(3), 309–342. https://doi.org/10.1162/neco_a_01563

Serafeim, G., & Yoon, A. (2021). Stock Price Reactions to ESG News: The Role of ESG Ratings and Disagreement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3765217

Singer, N. (2018). Apple, in Sign of Health Ambitions, Adds Medical Records Feature for iPhone. *The New York Times*. https://www.nytimes.com/2018/01/24/technology/apple-iphone-medical-records.html

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., Arcas, B. A. Y., Webster, D., & Natarajan, V. (2022). Large Language Models Encode Clinical Knowledge.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M. F., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A. Y., & Natarajan, V. (2023). Towards Expert-Level Medical Question Answering with Large Language Models.

Sinha, A., Kedas, S., Kumar, R., & Malo, P. (2023). SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News. https://doi.org/10.1002/asi.24634

Southall, A. (2015). One Dead After Truck Hits Apartment Building in the Bronx. *The New York Times*. https://www.nytimes.com/2015/10/13/nyregion/one-dead-after-truck-hits-building-scaffolding-in-the-bronx.html

Starks, L. T. (2023). Presidential Address: Sustainable Finance and ESG Issues—Value versus Values. *The Journal of Finance*, *78*(4), 1837–1872. https://doi.org/10.1111/jofi.13255

Strom, S. (2010). Pepsi Refresh Contestant Claims Rules Were Broken. *The New York Times*. https://www.nytimes.com/2010/10/01/business/01pepsi.html

Strom, S. (2011). PepsiCo to Foster Chickpeas in Ethiopia. *The New York Times*. https://www.nytimes.com/2011/09/21/business/pepsicos-chick-pea-plan-includes-taking-on-famine.html

Strom, S. (2014a). Coca-Cola to Remove an Ingredient Questioned by Consumers. *The New York Times*. https://www.nytimes.com/2014/05/06/business/coca-cola-to-remove-an-ingredient-questioned-by-consumers.html

Strom, S. (2014b). Soda Makers Coca-Cola, PepsiCo and Dr Pepper Join in Effort to Cut Americans' Drink Calories. *The New York Times*. https://www.nytimes.com/2014/09/24/business/big-soda-companies-agree-on-effort-to-cut-americans-drink-calories.html

Strom, S. (2015). Coca-Cola to Cut Up to 1,800 Jobs. *The New York Times*. https://www.nytimes.com/2015/01/09/business/coca-cola-announces-plan-to-cut-1600-to-1800-jobs.html

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text Classification via Large Language Models.

Tang, Y., Yang, Y., Huang, A. H., Tam, A., & Tang, J. Z. (2023). FinEntity: Entity-level Sentiment Classification for Financial Texts. https://arxiv.org/pdf/2310.12406.pdf

The Coca-Cola Company. (n.d.-a). James Quincey. Retrieved October 31, 2023, from https://www.coca-colacompany.com/about-us/leadership/james-quincey

The Coca-Cola Company. (n.d.-b). Our Company. Retrieved October 31, 2023, from https://www.coca-colacompany.com/about-us

The Coca-Cola Company. (2022). Form 10-K 2022. https://investors.coca-colacompany.com/filings-reports/all-sec-filings/content/0000021344-23-000011/ko-20221231.htm

The GDELT Project. (n.d.-a). The GDELT Project. Retrieved October 15, 2023, from https://www.gdeltproject.org/

The GDELT Project. (n.d.-b). The GDELT Story. Retrieved October 15, 2023, from https://www.gdeltproject.org/about.html

The New York Times Company. (2023). The New York Times Company Reports Second-Quarter 2023 Results.

Thomas, K. (2013). J.&J. to Pay $2.2 Billion in Risperdal Settlement. *The New York Times*. https://www.nytimes.com/2013/11/05/business/johnson-johnson-to-settle-risperdal-improper-marketing-case.html

Thomas, K. (2014). Pfizer and Aid Groups Team Up on Contraceptive for Developing World. *The New York Times*. https://www.nytimes.com/2014/11/14/business/pfizer-and-aid-groups-team-up-on-depo-provera-for-developing-world.html

Tracy, M. (2021). A West Virginia newspaper company is suing Google and Facebook over online ads. *The New York Times*. https://www.nytimes.com/live/2021/01/29/business/us-economy-coronavirus

United Nations Global Compact. (2004). Who Cares Wins: Connecting Financial Markets to a Changing World. https://www.unepfi.org/fileadmin/events/2004/stocks/who_cares_wins_global_compact_2004.pdf

Vance, A. (2010). Microsoft and New York in Software Deal. *The New York Times*. https://www.nytimes.com/2010/10/21/technology/21soft.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need.

Vega, T. (2011). Complaint Accuses Pepsi of Deceptive Marketing. *The New York Times*. https://archive.nytimes.com/mediadecoder.blogs.nytimes.com/2011/10/19/complaint-accuses-pepsi-of-deceptive-marketing/

Venigalla, A., Frankle, J., & Carbin, M. (2022, December). BioMedLM: a Domain-Specific Large Language Model for Biomedical Text. *Mosaic ML*. https://www.mosaicml.com/blog/introducing-pubmed-gpt

Wakabayashi, D. (2021). Google Temps Fought Loss of Pandemic Bonus. And Won. *The New York Times*. https://www.nytimes.com/2021/11/05/technology/google-workers.html

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBert: A Pretrained Language Model for Climate-Related Text.

Weise, K. (2019). Microsoft Pledges $500 Million for Affordable Housing in Seattle Area. *The New York Times*. https://www.nytimes.com/2019/01/16/technology/microsoft-affordable-housing-seattle.html

Weise, K., & McCabe, D. (2022). F.T.C. Sues to Block Microsoft's $69 Billion Acquisition of Activision. *The New York Times*. https://www.nytimes.com/2022/12/08/technology/ftc-microsoft-activision.html

Wingfield, N. (2016a). Microsoft Cutting 1,850 Jobs in Smartphone Unit. *The New York Times*.

Wingfield, N. (2016b). With LinkedIn, Microsoft Looks to Avoid Past Acquisition Busts. *The New York Times*. https://www.nytimes.com/2016/12/08/technology/with-linkedin-microsoft-looks-to-avoid-past-acquisition-busts.html

Woo, E. (2022). Fitbit recalls more than one million smart watches over a burn risk. *The New York Times*. https://www.nytimes.com/2022/03/02/business/fitbit-ionic-watch-recall.html

Wu, S., Irsoy, O., Lu, S.-w., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance.

Zachary, G. P., & Hall, M. (2023, February). Microsoft Corporation. https://www.britannica.com/topic/Microsoft-Corporation

Zhuang, Y. (2022). New Crack in Apple's Armor as Dozens Strike at Its Stores in Australia. *The New York Times*. https://www.nytimes.com/2022/10/17/business/apple-store-strike-australia.html

Zimmer, C. (2021a). Pfizer Says Its Vaccine Works Against Key Mutation in Contagious Variants. *The New York Times*. https://www.nytimes.com/2021/01/08/health/pfizer-covid-vaccine-variant-mutation.html

Zimmer, C. (2021b). J&J's Booster Shot Provides Strong Protection against Severe Disease from Omicron, a Study Says. *The New York Times*. https://www.nytimes.com/2021/12/30/health/johnson-vaccine-booster-omicron.html

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Understanding Emergent Leadership Across Cultural Levels: A Theoretical Framework

Elif Leman Bilgin

King's College London

Abstract

Emergent leadership literature emphasises identifying and nurturing leaders at all organisational levels to foster team harmony and align efforts toward shared goals. Since past studies focused largely on individual traits predicting leadership emergence, the interplay of different cultural levels, such as national culture, organisational culture and team culture in relation to individuals emerging as emergent leaders remains unexplored. This study extends beyond discussing the antecedents and outcomes of emergent leadership and provides an in-depth understanding of the phenomenon through different cultural levels. It introduces an overarching theoretical framework proposing that a) the unfolding of emergent leadership occurs at four levels, which are organic emergence, non-normative emergence, conditional emergence and non-emergence, based upon the type of interaction between cultural levels and potential emergent leaders, b) for emergent leadership to occur, potential emergent leaders must have or display some of the compatible antecedents, c) the approval of higher-level authority figures at the organisational or national level is a precondition for the occurrence of emergent leadership in stratified teams.

Keywords: emergent leadership; individual traits; national culture; organisational culture; team culture

1. Introduction

With the increasing need to respond to new challenges and opportunities and the continuously changing state of the world of work, organisations have been shifting towards informal and flat structures that promote the emergence of non-hierarchical and lateral leadership styles, such as emergent leadership (Kaplan et al., 2016; McClean et al., 2018). Early scholars broadly define this phenomenon as a type of horizontal leadership occurring in a flat team structure, where a team member gains an influence over other team members and is perceived as a leader by them, despite not having any formal authority or a role (Schneier & Goktepe, 1983). However, the literature presents a lack of unanimity in terms of what emergent leadership is and how this phenomenon unfolds (Wolfram Cox et al., 2022). The literature consists of various definitions of emergent leadership, with some researchers using the concepts of formal and informal leadership emergence synonymously (Judge et al., 2002, 2004). Conversely, some researchers strongly separated the two concepts and referred to emergent leadership as an individual or individuals having a leaderlike influence in the team without holding a formal position in the organisation (Barling & Weatherhead, 2016; McClean et al., 2018; Tabassum et al., 2023). Recently developed frameworks attempted to address the issues around construct clarity in emergent leadership research (e.g., Hanna et al., 2021; Wolfram Cox et al., 2022), laying the foundation for future research that is complemented by interdisciplinary views and complex social dynamics.

Since the first research on the topic was conducted by Murphy (1941), emergent leadership has drawn researchers' attention (Figure 1). The key role of emergent leaders in improving team performance and organisational outcomes makes this phenomenon particularly attractive to researchers as well as businesses (Spisak et al., 2015). For instance, global corporations including Google and General Electric recognise the critical role of emergent leaders for future success and growth while promoting emergent leadership through organisational leadership initiatives

DOI: https://doi.org/10.5282/jums/v10i2pp402-423

(Maloney, 2020). Likewise, considering the everchanging and competitive nature of the contemporary business world and the vast amount of resources organisations spend on leadership-related efforts and training programmes, it is crucial to further our understanding of the elements that impact the emergence of emergent leadership (Westfall, 2019).

Emergent leadership has been studied in connection to several research areas, including communication (J. Jiang et al., 2015; Rennie et al., 2023), team dispersion and colocation (Charlier et al., 2016), Big-Five personality traits (particularly extroversion) (Landis et al., 2022), team performance and self-managing teams (Doblinger, 2022), leadership effectiveness and virtual teams (Hoch & Dulebohn, 2017), as well as across different academic fields, such as education (Leeming, 2019), healthcare (Grimsley et al., 2021), management (Andersson & Tengblad, 2016; Hu et al., 2019), and psychology (Reichard et al., 2011; Schaumberg & Flynn, 2012). However, academic sources provide limited insight into the emergence of emergent leaders in cultural settings; particularly overlooking the influence of different cultural levels in which individuals are embedded in how such leadership unfolds (Hanna et al., 2021). Yet, it has been evident in cross-cultural studies that individuals' preferences towards leadership behaviours and perception of the ideal leader vary across different cultures (House et al., 2004; Javidan et al., 2006).

Likewise, the implicit leadership theories (ILTs; Lord and Maher, 1991; Lord et al., 1984) have been mentioned frequently in the emergent leadership literature; suggesting that culture and the differences in cultural norms influence how leadership is constructed (Javidan et al., 2006; Lord et al., 2020). Focusing on culture on a more individual level, Javidan et al.'s (2006) proposal of culturally endorsed ILTs argue that people's views about the features of successful or unsuccessful leadership are shaped by the culture in which they are embedded. This suggests that these culturally endorsed ILTs could play a pivotal role in how emergent leadership unfolds across different cultures.

Overall, the role of culture in how emergent leadership unfolds remained an unexplored area within the emergent leadership sphere. Since no previous studies have examined culture and emergent leadership jointly, the broad leadership literature determines the basis of the current understanding of how leadership may occur across cultures and borders. However, the majority of studies on leadership and national culture have been conducted in the United States (Figure 2). These studies are carried out in non-diverse settings; usually involving college students, and fail to provide sufficient insight into the genders or races of their participants. This is a critical issue, as in a world where most organisations are becoming increasingly diverse, the narrow perspective provided by previous meta-analyses may be inadequate (see Ensari et al., 2011). Extant leadership research also overlooks the multidimensional nature of culture, with no studies focusing on different cultural levels (i.e., national culture, organisational culture, team culture) collectively.

1.1. Research Purpose and Contribution

Understanding the role of different cultural levels in the way leadership emerges is crucial for most organisations, specifically for the ones that operate across different geographies and have culturally diverse work teams (Ely, 2004). Organisations with multicultural workforces often face challenges in terms of building team synergy within their diverse teams, which adversely affects the efficiency of day-to-day operations and business outcomes (Jehn & Bezrukova, 2004). Scholars suggest that business success lies within the identification and development of effective leaders across all organisational levels, as these leaders can positively influence the perceptions, attitudes and behaviours of their diverse team and motivate them to work towards a common organisational purpose (Butler et al., 2012; Osland et al., 2009). Correspondingly, emergent leaders can utilise their influential role to build harmony, thereby contributing to the growth of the company and even encouraging other leaders to emerge in the organisation (Zander et al., 2012). Likewise, employee behaviour is shaped by multiple internal and external factors, such as national culture, organisational culture, team culture as well as individual values and personality (Lok & Crawford, 2004; Smithikrai, 2008). Thus, providing an in-depth insight into how leadership emerges in relation to different cultural levels in which individuals are embedded has both academic and practical importance.

This dissertation, therefore, proposes an overarching theoretical framework to understand how the different cultural levels in which emergent leaders are embedded influence the way emergent leadership unfolds. So far, emergent leadership has not been explored across different cultural levels. To arrive at conclusions, the topic of emergent leadership must be thoroughly examined; not only its antecedents and outcomes but how it unfolds. Hence, the contributions of this dissertation are to:

a) review and present a comprehensive analysis of research findings on emergent leadership.

b) theorise about the influence of different cultures in which emergent leaders are embedded on how emergent leadership unfolds by examining national culture, organisational culture, team culture as well as the traits of potential emergent leaders.

c) generate implications for future research on emergent leadership across different cultural levels and provide practical suggestions for organisations.

The research question is as follows: How does the culture in which emergent leaders are embedded influence the manner in which emergent leadership unfolds?

1.2. Outline

To gain an in-depth understanding of emergent leadership, an integrative literature review is conducted, which is followed by a brief review of national culture in a leadership context with an emphasis on different cultural levels in which

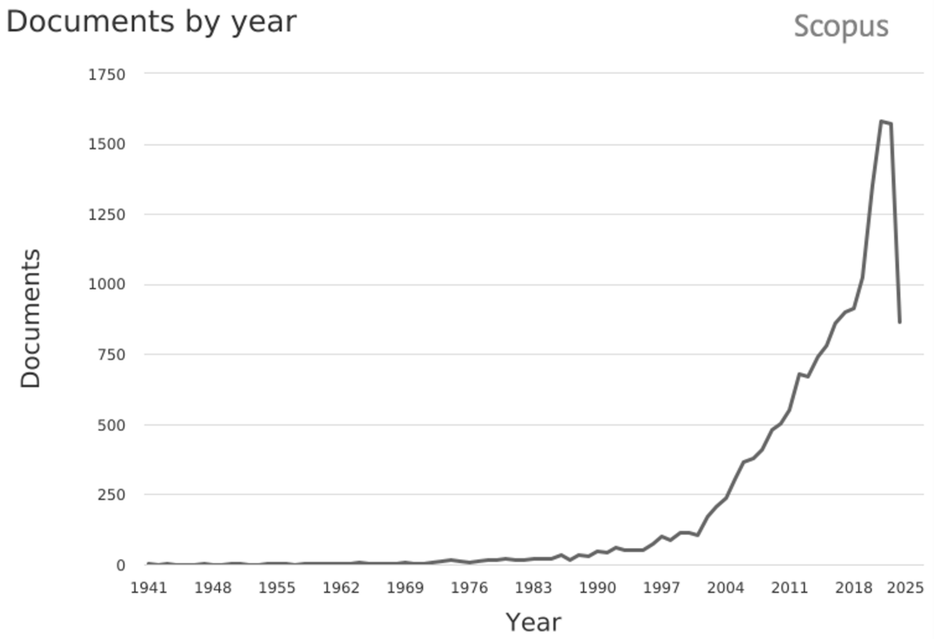


Figure 1: English language publications with article title, or abstract including “emergen*” AND “leader*”

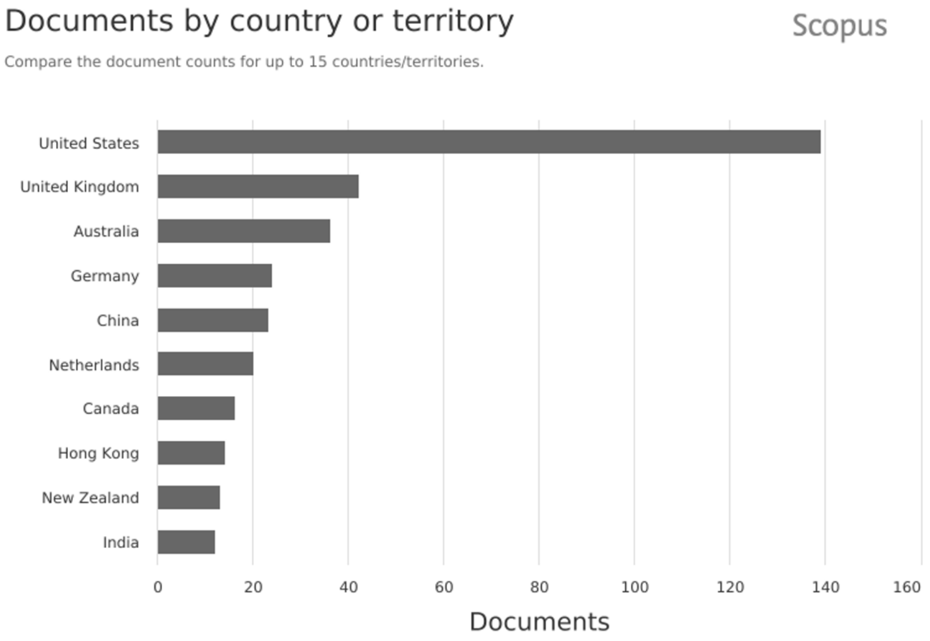


Figure 2: Publications with titles of "national culture" AND "leadership" published in the last 20 years in English language across countries.

individuals are embedded. To answer the research question, it is essential to explore the phenomenon of emergent leadership and how it occurs extensively. Since culture is a highly broad topic, it is explored briefly and the emphasis is placed on the areas that are relevant to answering the research question. Finally, the proposed framework is discussed in detail, followed by propositions and theoretical and practical implications.

2. Emergent Leadership

2.1. Operationalising Emergent Leadership

As outlined previously, the literature includes various definitions of emergent leadership since the term was first studied in 1941. Since then, the urgent need to operationalise emergent leadership has been pointed out by different scholars (Kickul & Neuman, 2000; Schneier & Goktepe, 1983) but the variations in definitions remained. While some studies used broad and vague references such as "a leader emerging in a group" or "champions" (Lanaj & Hollenbeck, 2015; Loignon & Kodydek, 2022; Taylor, 2009), others studied it as situations in which no formal leader exists (Taggar et al., 1999), referred to it as being perceived as a leader (Kent & Moss, 1990, 1994) or used informal leader and emergent leader interchangeably (Landis et al., 2022; Wu et al., 2021).

Most contemporary scholars agree that emerging as an emergent leader differs from emerging as a formal leader through the characteristics of the individuals who are involved in the process (Ensari et al., 2011; Gerpott et al., 2019; Schlamp et al., 2021), arguing that while formal leadership tend to rely on the views and judgements of senior members, for emergent leaders their peers or themselves determine how the process unfolds (DeRue & Ashford, 2010). Likewise, although both processes include different forms of social interactions such as granting and assuming leadership and creating influence over other members, emergent leaders do not have any formal positions in their organisations (Badura et al., 2022).

It can be argued that the lack of clarity and internal consistency around the definition of emergent leadership occurs because leadership emergence itself is seen as a mysterious magical process. For instance, Guastello and Bond Jr. (2007) state that leadership emergence studies often involve processes where "group participants might be measured on a number of traits that could be related to leadership behaviors. Members of the group then interact while carrying out a task. Then magic happens and a leader emerges from the group at the end of the discussion period" (p. 357).

So far, Hanna et al.'s (2021) review has been the most extensive attempt to conceptualise emergent leadership. Hence, to ensure conceptual clarity, for this review, Hanna et al.'s (2021) definition of emergent leadership, which is "the degree to which an individual with no formal status or authority is perceived by one or more team members as exhibiting leaderlike influence" (p. 82) is followed. With an aim to theorise about how different cultures in which emergent leaders are embedded influence how emergent leadership unfolds, this paper then builds onto Hanna et al.'s (2021) framework consisting of three key elements of emergent leadership, which are *lateral influence*, *unit of analysis* and *temporal duration*.

Lateral influence symbolises the emergent leader's ability to cause considerable influence over their team, including a vast series of behaviours (e.g., taking responsibility, planning and organising tasks) as well as roles (e.g., manager, motivator and mediator) (Hanna et al., 2021). From early scholars to contemporary studies, lateral influence has been perceived as a fundamental aspect of emergent leadership, illustrating the importance of being perceived as leaderlike by others in conceptualising this phenomenon (Lanaj & Hollenbeck, 2015). Many researchers corroborated this element; however, it is important to note that this influence may be both momentary and long-term. This aspect is explained further under temporal duration, which is outlined below.

Unit of analysis demonstrates the individuality of emergent leadership, as more than one individual in a team may be viewed as an emergent leader, thereby teams can contain multiple emergent leaders. However, it is essential to emphasise that even when there are several different emergent leaders, the influence these leaders generate does not arise collectively, but rather emerges from each individual in the team. Correspondingly, research indicates that not only there could be multiple emergent leaders in a team, but also having more than one emergent leader could benefit the team by increasing team effectiveness (Ziek & Smulowitz, 2014).

Temporal duration symbolises the fluid span of emergent leadership, stressing that informal leaders can emerge temporarily, and their influence can change or fade over time. This is due to the fact that various elements could affect who and how long a person is perceived as a leader in the team. For instance, Landis et al. (2022) argue that extroverted emergent leaders' influence over the group in which they operate may not always last long, as group members tend to stop perceiving them as leaders at some point and leave their leadership network over time. In their study with a sports team, Mertens et al. (2021) suggest that due to the competitiveness involved in these types of teams, leadership structures can change considerably throughout the season, thereby allowing players to engage in informal leadership roles at different times.

This review also identified another potential key element of emergent leadership: *knowledge dissemination*. Knowledge dissemination refers to the emergent leader's role in contributing to or facilitating knowledge-building practices in the team. Comfort and Okada (2013) suggested that particularly in times of uncertainty, emergent leaders act as an enabler of a wide exchange of knowledge amongst the group. Another study argued that knowledge sharing acts as a key function for emergent leaders to be recognised and deferred by other members, leading the knowledge shared by the leader to become a property of a team and build a collective cognition over time (Murase et al., 2013). However, defining knowledge dissemination as a definitive element of this phenomenon may be erroneous, as the influence of emergent leaders may not always be in the form of imparting knowledge to others.

2.2. Other Leadership Constructs and Related Concepts

Before reviewing the topic of emergent leadership further, it is essential to highlight other leadership constructs, as emergent leadership has also been studied alongside other leadership concepts. Examples include assigned leadership, shared leadership, participative leadership, team leadership and self-leadership (Huang et al., 2010; Landis et al., 2022; Wu et al., 2021). These concepts overlap to some extent, however, have all demonstrated to be separate constructs, as illustrated in both theoretical (Hoch & Dulebohn, 2017) and empirical studies (Wickham & Walther, 2009).

Assigned leadership is a contrasting concept to emergent leadership which incorporates a more traditional and top-down organisational hierarchy (Paunova, 2015). Assigned leaders are formally appointed by company management, receiving their leader status from top to bottom (Lucas, 2003). Summerfield (2014) suggests that the concept of emergent leadership encompasses the idea that leadership is not confined to only those holding titles like CEO, president, or chairperson, in other words, assigned leaders, thereby arguing that all individuals in the organisation can enact positive change.

Shared leadership on the other hand is similar to emergent leadership, as they are both informal leadership concepts, however, distinctively, shared leadership solely exists as a shared group level phenomenon (Pearce & Sims, 2000). D. Wang et al. (2014) define shared leadership as "an emergent team property of mutual influence and shared responsibility among team members, where they lead each other toward goal achievement" (p. 181). As per the definition, shared leadership differs from emergent leadership by emphasising the team generating a leaderlike influence collectively, as opposed to focusing on the individuals' journey of emerging as a leader (D'Innocenzo et al., 2016). Yet, many researchers studied both leadership constructs together due to their complementary nature (Carte et al., 2006; Hoch & Dulebohn, 2017; Van Zyl & Hofmeyr, 2021).

Participative leadership refers to a type of democratic leadership style in which all team members are intentionally involved in organisational decision-making processes (Sashkin, 1976). Similar to emergent leadership participative leadership may induce the feeling of empowerment among team members (Ahearne et al., 2005), however, it has always been conceptualised as a team-level construct whereas emergent leadership posits that leadership occurs at an individual level and originates from the individual (i.e., unit of analysis, Hanna et al., 2021).

Team leadership is an umbrella term that refers to all leadership activities taking place in a team. Due to its broad definition, team leadership is analogous to emergent leadership and involves all lateral influences among individuals in a team (Hackman & Wageman, 2004). What differentiates the two concepts from each other is that, unlike emergent leadership, team leadership includes formal leadership processes and influences (van Knippenberg, 2017).

Self-leadership posits the notion that although human behaviour is often influenced by external forces such as a leader, actions are ultimately governed by intrinsic rather than extrinsic factors (Manz, 1986). The concept of internal regulation has been studied alongside self-leadership, as self-leadership is more concerned with how individuals influence and control their behaviours rather than examining their influence on other team members (Kanfer et al., 2008). This point is the greatest distinction between emergent leadership and self-leadership; because even though both constructs can unfold at the individual level, emergent leadership also has a collective level dimension, lateral influence (Hanna et al., 2021; Stewart et al., 2011).

2.3. Theoretical Frameworks and Conceptualisations

As outlined above, despite not being a new idea, emergent leadership has limited theoretical development and coherence, which causes difficulties in studying this concept. This paper identified several key publications that proposed comprehensive theoretical frameworks on emergent leadership. Although these papers shed some light on processes, antecedents and outcomes of emergent leadership, they either overlooked the role of macro-level complex mechanisms, such as culture, in how emergent leadership occurs or only provided suggestions for future research. However, it is critical to mention the key theories of emergent leadership that influence the framework of this paper, such as functional leadership theory (McGrath, 1962) and the relational models theory of group-level leadership emergence (Wellman, 2017).

Functional leadership theory (McGrath, 1962; Morgeson et al., 2010) posits that leadership can originate internally or externally and occur in formal or informal ways, thereby identifying emergent leadership as internally originated and informal. Scholars who studied this theory reported critical findings that provided further insight into the role of team dynamics in emergent leadership. For instance, teamwork behaviour was found to be a strong predictor of the emergence of emergent leadership (Luria & Berson, 2013), with Wolff et al. (2002) reinforcing this argument by suggesting that emergent leaders improve team task coordination by showing empathy towards team members. These findings underline the importance of team level considerations in emergent leadership which suggests that assessing team culture and its role in the unfolding of emergent leadership is pivotal.

The relational models theory of group-level leadership emergence (Wellman, 2017) suggests that there are two ways for leaders to emerge: In the first instance "groups converge on an authority ranking relational model, in which leadership influence is afforded to a small number of members who are perceived to possess the greatest individual leadership capabilities" (p. 597). In the second instance, "groups converge on a communal sharing relational model, in which leadership is viewed as a shared group responsibility" (p. 597). This theory is found particularly helpful for the current review, as it provides insight into different forms of emergent leadership as well as how and why these differences occur.

In terms of other relevant frameworks, in their analysis, Acton et al. (2019) focused on a narrow aspect of emergent leadership and exclusively studied the cognitive perceptions of leaders and followers. This framework followed a complexity perspective; however, it was entirely process-oriented and did not inform readers of the numerous contexts in which leaders can emerge. Moreover, they only reviewed psychology and management literature, thereby potentially missing relevant theory and research on emergent leadership in different disciplines.

Hanna et al.'s (2021) review added conceptual clarity to the construct's disunited conceptualisations and developed a broad framework that helped elucidate the nomological

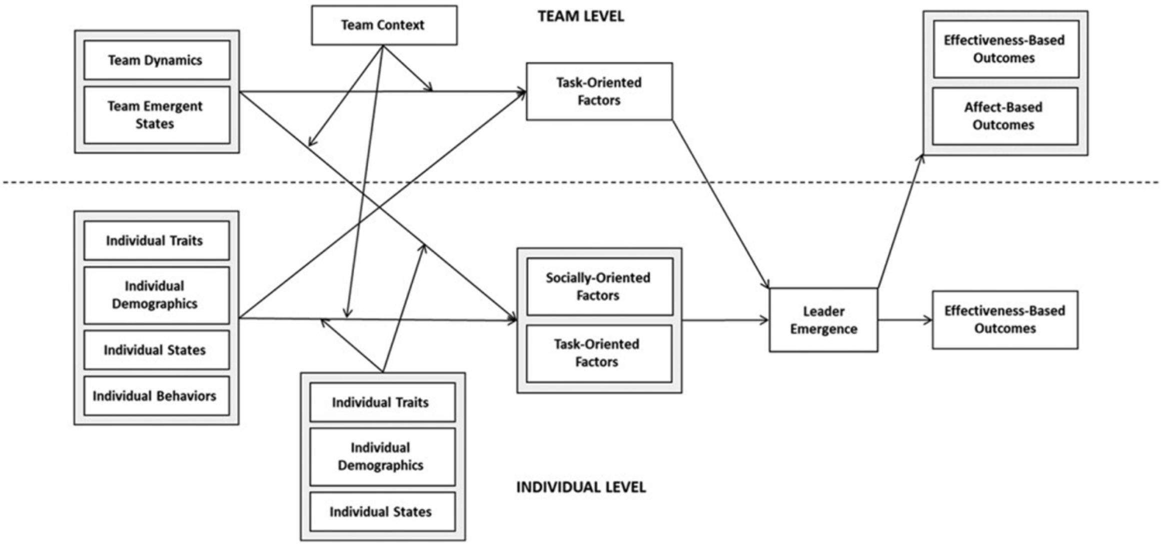


Figure 3: A High-Level Emergent Leadership Framework, Hanna et al. (2021).

network of this phenomenon by listing antecedents, consequences and mediators of emergent leadership (Figure 3). Differing from the previous papers, this review offered several diverse future research directions from methodological suggestions to theoretical questions, one of which has also provided the basis for this research. Another unique contribution of this review was that it pointed out the potential detrimental sides of emergent leadership, which is an idea that has not been mentioned in previous research. However, as suggested by the authors, Hanna et al.'s (2021) framework mainly plays an introductory role to further research and does not provide an in-depth insight into the occurrence of emergent leadership in relation to more complex social mechanisms, such as societies and cultures.

Addressing the previous calls, Tabassum et al. (2023) followed a social identity and implicit leadership perspective while offering a multi-level conceptualisation of emergent leadership (Figure 4). They adopted novel approaches by considering the dynamics beyond teams and looking into contextual attributes at organisational levels as well as examining the concepts of distributed leadership and empowering leadership. Nonetheless, their analysis of emergent leadership did not consider other macro-level actors and rather provided a detailed description of the relationship between emergent leadership and a range of factors including personal traits, skills, communication and self-perception.

As the latest review on the topic, Galvin et al. (2023) evaluated how leader emergence occurred in both formal and informal ways and studied individuals' potential as leaders and whether they would emerge as leaders (Figure 5). Their framework elaborated the notion of over-emergence which was introduced by Lanaj and Hollenbeck (2015); developed four types of leader emergence (over-emergence, congruent emergence, congruent non-emergence, and under-emergence); integrated leader emergence and effectiveness and provided some contextual considerations by briefly touching upon cultural and organisational factors. Yet, in terms of broad cultural factors, they solely considered how a leader might be perceived by others in different cultures and only provided comparisons between leadership emergence in collectivistic and individualistic cultures. Although beneficial, this study provided limited insights into the complex nature of human behaviour by making generalisations based solely on one type of culture.

2.4. The Unfolding of Emergent Leadership

Emergent leadership has been studied alongside several concepts such as team effectiveness, self-managing team, self-organisation, followership, trait affective presence, degree-centrality, vocal delivery and knowledge sharing (Carte et al., 2006; X. Jiang et al., 2021; Yoo et al., 2022; Ziek & Smulowitz, 2014). A substantial body of research has offered insight into the antecedents, mediators or outcomes of emergent leadership, emphasising who tends to emerge as leaders (Badura et al., 2022; Judge et al., 2002; Kaiser et al., 2008). Surprisingly, it is less common for researchers to study how this phenomenon unfolds, with most studies in this context being on other leadership concepts (i.e., leadership development, Hart, 2016), or assessing the relationship between emergent leadership and certain personality traits, intrinsic factors, team environment or organisational practices separately (Cogliser et al., 2012; Johnson & Bechler, 1998; Reichard et al., 2011).

Although variations exist, there has been a mutual theme amongst most emergent leadership scholars, which is that the context for emergent leadership has critical importance and there is merit in evaluating team, organisational and individual level factors together. In particular, Wellman (2017) argues that leadership emergence research overlooks "the potential for group level dynamics in the leadership emergence process" (p. 597). Likewise, Kozlowski and Klein (2000)

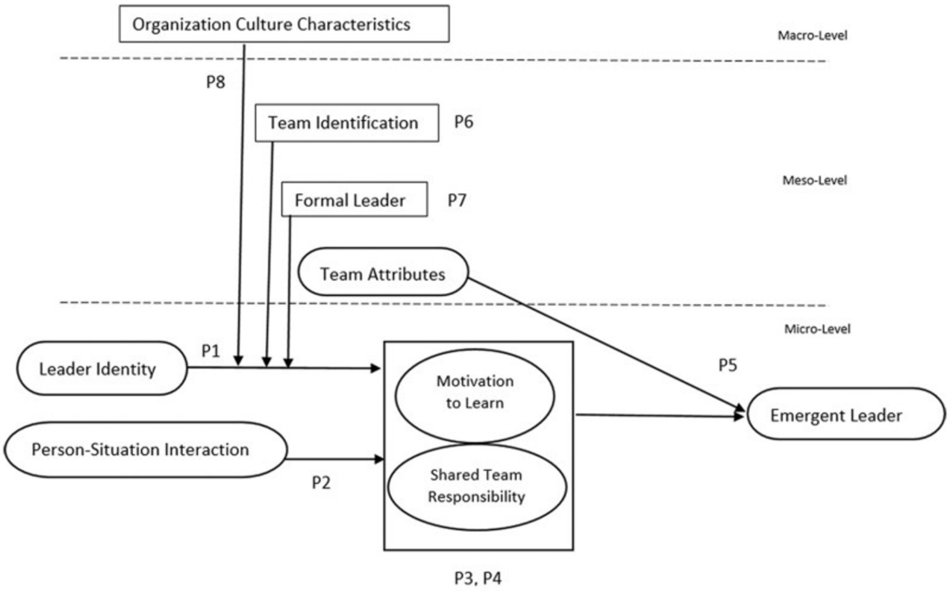


Figure 4: Emergent leadership: a multi-level overarching model, Tabassum et al. (2023).

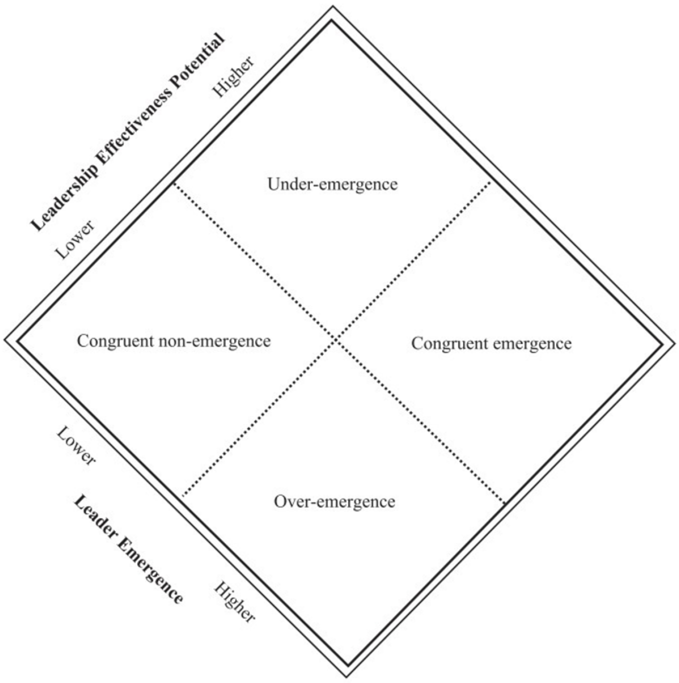


Figure 5: Typology of leader emergence, Galvin et al. (2023).

highlighted that "people in groups and subunits are exposed to common features, events, and processes. They interact, sharing interpretations, which over time may converge on consensual views of the group" (p. 10).

Hence to address the research question and theorise about how emergent leadership unfolds, in other words, 'the process', the relevant articles and studies identified in the literature as well as the findings of current frameworks have been incorporated, summarised and categorised into three main areas: organisational level, team level and individual level.

2.4.1. Organisational level

Organisational level considerations are critical, as it is acknowledged in the literature that the leadership emergence process is not only situated within a broad context of individual and team level dynamics but also in formal organisational structures and practices. As Kozlowski et al. (2013) emphasise "although it is not a core characteristic of emergence per se, contextual factors at the higher level shape and constrain the process dynamics of emergence" (p. 585).

In the organisational context, the concepts of *organisational sensemaking* and *organisational sense-giving* may help with understanding how emergent leadership unfolds (Weick et al., 2005). Organisational sensemaking refers to the process during which individuals perceive and eventually turn involved information into a clear narrative, with the narrative they form often mirroring their identity work (Ibarra & Barbulescu, 2010). Organisational sense-giving refers to the process by which a formed meaning is passed onto others (Hill & Levenhagen, 1995). This indicates that the message organisations convey through their vision, strategy, environment and culture can directly affect individuals' behaviour at work, including whether they emerge as leaders. Likewise, Tabassum et al. (2023) propose that "the characteristics of organization culture at a particular point in time moderates the relationship between leader identity and emergence of emergent leadership" (p. 11). They state that the overall positive impact of leader identity on the emergence of emergent leadership is higher when it is combined with an encouraging organisational culture; whereas the positive impact is weaker when organisations are not supportive of individuals to emerge as leaders (ibid). Hence, it can be argued that organisations' ability to align around a vision, strategy, and culture as well as to build a supportive environment determines the emergence of emergent leaders. Likewise, organisational complexity has also been studied in this context and defined in three facets: (a) the degree of knowledge that is required to comprehend the organisational environment, (b) the degree of unforeseen upcoming changes in the environment, and (c) the degree of available resources in the environment (Sharfman and Dean, 1991). Eisenbeiß and Giessner (2012) argued that the complexity of an organisational environment is negatively correlated with the emergence of leaders, particularly the ones who adopt ethical approaches to leadership, within an organisation.

Practical suggestions for organisations include informing formal leaders, from junior managers to senior ones, on the importance of being in support of emergent leaders in the organisation (Virtaharju & Liiri, 2019). Erkic (2022) stated that organisations should focus on building a culture that demonstrates support towards innovation, open communication and diverse perspectives while encouraging a culture of learning and knowledge sharing. Knowledge dissemination is also mentioned by Yoo et al. (2022), who suggested that implementing knowledge-building practices across the organisation can encourage emergent leadership and ultimately lead to an improvement in organisational effectiveness. Similarly, drawing on Schneider's (1975) seminal paper on organisational climate, Murase et al. (2013) argued that through wide-scale organisational practices and interactions between members, individuals' viewpoints and knowledge emerge to become a property of the organisation. They also suggested that the shared viewpoints and knowledge result in a positive influence on individuals' behaviours and allow members to function as a unified entity (ibid.). This is in line with Comfort & Okada's (2013) postulations, stating that developing knowledge commons in the event of uncertainty allows a broad exchange of knowledge and skills and results in the emergence of leaders who will aid members in demonstrating an appropriate collective action against the crisis.

2.4.2. Team level

Team-level research focused on team dynamics, in particular, team emergent states, which are defined as "the properties that develop during team interactions and describe members' attitudes and behaviour" (Fyhn et al., 2023, p. 1). Team-level exploration is particularly complex, due to the temporality of team phenomena, in other words, the fluid nature of these dynamics and properties. Nevertheless, many studies found team emergent states and team dynamics to affect leader emergence. For example, in their three-year ethnographic study, Smith et al. (2018) found that leadership emerged in line with the development of shared social identity and shared goals and evolved through interactions, processes and practices in a team. The impact of the team shared goals on leadership emergence is also reinforced by other scholars (Zhang et al., 2012). A supportive team environment in which members feel comfortable has also been reported to positively affect the possibility of the emergence of emergent leaders (DeRue et al., 2015; Mumford et al., 2008). Current evidence illustrates that not only do emergent leaders tend to emerge in teams with high levels of social support, but they can also promote a more supportive team environment and boost team performance (Crozier et al., 2017; Wellman et al., 2019; Zhang et al., 2012).

When unconventional team structures are concerned, such as virtual or geographically dispersed teams, it is found that traditional leadership approaches have been less effective in improving team performance (Van Zyl & Hofmeyr, 2021). Likewise, Ziek and Smulowitz (2014) suggested that in virtual teams, emergent leaders who had sound communication skills and promoted creativity across the team lead to team members being more successful at completing tasks and projects. Accordingly, unconventional teams are reported as more hospitable for emergent leadership, as they tend to propose a self-organised, congruous team environment that has a high level of team connectedness and quality leader-member exchanges (Przybilla et al., 2019; Zhang et al., 2012).

2.4.3. Individual level

The individual level or the self is consistently mentioned as a critical player in leadership emergence. In the context of individuals, Big Five personality traits are one of the most researched antecedents of emergent leadership. Extraversion, in particular, has been a highly frequently mentioned personality predictor of leadership (Serban et al., 2015; Wilson et al., 2021). Although there is extensive literature indicating that extraversion can assist individuals in emerging as leaders

(Lee & Farh, 2019; Moutafi et al., 2007), especially in leaderless groups (Ensari et al., 2011), it should also be considered that the majority of these studies are conducted in the US or the UK, where ideal leaders are generally portrayed as extroverted individuals (Hofstede, 1980). Likewise, Mitchell et al. (2022) argue that one of the primary theoretical assumptions linking extroversion and leadership may be due to extroverts having high levels of communication skills. Other researchers suggested that extroverts have been theorised to emerge as leaders because they are also dominant, assertive and communicative (e.g., Hu et al., 2019; Judge et al., 2002), which allows them to influence and lead others (e.g., Nahrgang et al., 2009).

Agreeableness has also been found to aid individuals in emerging as leaders in leaderless groups (Cogliser et al., 2012; Walter et al., 2012). However, studies that claim otherwise also exist (McClean et al., 2018). This may be a good indicator of the importance of studying one's personality in the context of other external social dynamics, such as the dynamics of a team, organisation or country.

Further, emotional stability and conscientiousness are found to have positive or neutral effects on the emergence of formal and informal leaders (e.g., Cogliser et al., 2012; Colbert et al., 2012; Emery, 2012; Wolff et al., 2002. Other noticeable individual traits affecting emergent leadership include creativity (Guastello, 1995), self-efficacy (Kwok et al., 2018), cognitive intelligence (Judge et al., 2004; Li et al., 2012; Rubin et al., 2002) and openness (Emery et al., 2013).

Beyond personality traits, research suggests that individuals' ability to understand a situation and manage it appropriately, in other words, social intelligence, allows them to cater to the needs of the team members and exhibit leadership (Byrne & Bradley, 2007; Zaccaro et al., 1991). In terms of emergent leadership, the impact of social intelligence is less straightforward, as there is evidence indicating that being socially intelligent can be advantageous (Gruber et al., 2018) but may have no effect on being perceived as a leader by team members (Emery et al., 2011). Yet, it can still be concluded that a low level of social intelligence does not impede members from emerging as a leader, but a high level of social intelligence can increase the likelihood of leadership emergence.

Among the individual behaviours that play a role in the emergence of emergent leadership, consideration and task-oriented behaviours have been listed as important parts of emergent leadership (Cogliser et al., 2012). For instance, Mitchell and Bommer (2018) looked into prosocial motivation and impression management concerning emergent leadership and found that prosocial motivation was positively linked to leadership emergence, irrespective of the amount of task coordination behaviour exhibited, whereas impression management motives only predicted leadership emergence when accompanied with high levels of task coordination behaviour. Finally, high levels of attention given to members of a team were found related to emergent leadership (Gerpott et al., 2018).

3. Culture in which Individuals are Embedded

Culture is a long-studied area subject that has been defined in various ways. Schein (1992) defined culture as "a pattern of shared basic assumptions that the group learned as it solved its problems that have worked well enough to be considered valid and is passed on to new members as the correct way to perceive, think, and feel in relation to those problems". This definition was focused more on culture in an organisational context and led further definitions to be formed, such as Adler's (2023) definition which approaches culture from an individual values perspective and describes it as cultural values translating into norms, perspectives and ethics, and being reflected in the rules and actions of the society. In this review, the definition provided by Project GLOBE, in which culture was defined as "shared motives, values, beliefs, identities, and interpretations or meanings of significant events that result from common experiences of members of collectivities" is utilised to conceptualise national, organisational and team culture (House et al., 1999, p. 13).

There are ample studies analysing cultural differences in organisations on both conceptual and empirical levels, with most scholars approaching the topic of culture in the context of national culture. Before Nardon and Steers's (2009) review of national culture, there was a divergence in organisational research that failed to promote parsimony and generated difficulties in making comparisons between studies and samples. Nardon and Steers (2009) referred to this complexity as the *culture theory jungle* and converged six key cultural models that focused on different aspects of societal beliefs, values and norms (Hall, 1993; Hofstede, 1980; House et al., 2004; Kluckhohn & Strodtbeck, 1961; Schwartz, 1994; Trompenaars & Hampden-Turner, 2004). Since all models have significant elements to contribute to the understanding of culture in organisational leadership contexts, as opposed to advocating one model over another, the convergent model of core cultural dimensions (CCDs) is preferred to be used in this paper. CCDs consist of five common themes that are hierarchy vs. equality, individualism vs. collectivism, mastery vs. harmony, monochronism vs. polychronism, and universalism vs. particularism.

The first theme of hierarchy-equality attempts to explain how individuals within a society structure their power relationships, more specifically, analyses whether the power in that particular society is allocated hierarchically or in a more egalitarian manner. Hofstede's (1980) *power distance* defines it as what individuals believe in terms of appropriateness of either significant or negligible variations in authority and power between the members of a group or society. In such countries, individuals of these cultures believe that it is acceptable or normal for some members of a group or society to employ substantial control over others (i.e., managers having high levels of authority over their subordinates).

On the other hand, countries with *low power distance* promote a more egalitarian and participative view of leadership. Such a view supports democratic approaches, encouraging subordinates to actively take part in decision-making and creating a more suitable environment for emergent leadership to unfold. Schwartz (1994) reinforces Hofstede's (1980) postulations and classifies China, Turkey and Thailand as hierarchical cultures, and Denmark, Sweden and Norway as egalitarian cultures. The GLOBE research (House et al., 2004) evaluates power distribution in societies through the lens of gender and points out the handicaps of gender egalitarianism. As a related yet different view, Trompenaars and Hampden-Turner (2004) studied how status and rewards are allocated in cultures, presenting "achievement" and "ascription" cultures. While achievement culture rewards individuals based on their attainment, ascription cultures emphasise seniority, class, age or gender in reward allocation. A recent cross-cultural study found that status was significantly effective in leadership emergence in both South Korean and American groups (Park, 2019). This indicates that how individuals receive their rewards, in this context, the reward being status, may differ, however status itself plays a critical role in emergent leadership in both ascription and achievement cultures.

In the context of emergent leadership, it can be assumed that in hierarchical countries with high power distance, there can be serious barriers to the emergence of emergent leadership in an organisation or work team. However, as Hanna et al. (2021) suggested, if individuals received approval from the authority figures in their workplace or team, there could be a chance for them to emerge as leaders. Accordingly, it is expected that the environment in egalitarian cultures would be more suitable for leadership emergence, with informal leadership attempts being acceptable and even favoured by society, organisations or team members. These assumptions are further analysed in the next section, where the overarching model proposed by the current review is introduced.

The second theme individualism-collectivism explores how different cultures carry out social organisation, with some being organised based on groups and others being organised based on individuals. Kluckhohn and Strodtbeck (1961) introduced three types of cultures in this context: *individualistic* countries that regard individual interests and aims as more important than collective goals and interests; *collateral* cultures where individuals see themselves as a part of a macro-level group and *lineal* cultures that places equal emphasis on groups and individuals.

Although all five frameworks studied concepts related to individualism and collectivism, Hofstede (1980) played a pioneering role in the introduction of these concepts. Hofstede (1980) states that individualistic cultures and collectivistic cultures differ in terms of their teachings around the sense of responsibility, with individualistic cultures stressing independence, being responsible for oneself and not relying on family or organisations, whereas collectivistic cultures value group interests, being responsible for the greater good, societal harmony and family relationships.

In terms of emergent leadership, considering that it occurs at the individual level and originates from the individual (unit of analysis) and involves influencing team members (lateral influence), it can be argued that individualistic cultures are likely to be hospitable towards emergent leaders by encouraging individuals to take initiative and lead others. As conformity is emphasised in collectivistic cultures, it is expected for such societies, organisations and teams to disapprove of informal and non-normative leadership approaches such as emergent leadership.

Mastery-harmony represents the differences across cultures in the degree to which individuals attempt to control their surroundings or choose to adapt to their environment. Cultures in high mastery believe that they should control, govern and change the environment around them, whereas cultures in high harmony believe that they should aspire to preserve harmony among the segments of the environment, including themselves (Dickson et al., 2012). Kluckhohn and Strodtbeck (1961) state that people in high-mastery cultures tend to be self-governed, assertive, competitive and achievement seekers. In contrast, people in high harmony cultures are likely to prioritise social ties and closeness over success and comfort to competitiveness. GLOBE (House et al., 2004) introduces the concepts of *humane orientation* (i.e., the extent to which society promotes being fair, caring and friendly towards others) and *assertiveness* (i.e., the degree to which people display assertive, dominant or aggressive behaviour in their social relationships in organisations or societies).

Assertiveness has been found linked to emergent leadership, as it requires taking initiative and being dominant at times to influence others (Hu et al., 2019; Judge et al., 2002). Hence, individuals who are in assertive cultures may effortlessly influence others through their dominant and assertive attitude, while also receiving support when they attempt to take on leadership in informal ways. Humane-oriented cultures, on the other hand, may reject emergent leadership, unless the leadership occurs in a way that is harmonious with the national or organisational culture (i.e., an emergent leader with a modest, altruistic attitude).

Monochronism vs. polychronism focuses on time orientation, with House et al. (2004) and Hofstede (2001) studying cultures' perceptions of time and the degree to which cultures are future-oriented and Trompenaars and Hampden-Turner (2004) assessing individuals' perceptions of the flow of time. Another important aspect of this theme is its focus on long term, future-oriented behaviours and rewards across cultures, where delayed gratification is encouraged or discouraged by society (House et al., 2004). Hall (1993) approached time orientation with a perspective of work activities, providing instances that monochronic cultures are likely to take on one task at a time, while polychronic cultures encourage focusing on multiple tasks simultaneously. This view is perhaps the most useful for leadership and management studies, as it can be argued that individuals' perceptions of the future are essentially an indication of their needs for achievement and perspectives on assuming control (Nardon & Steers, 2009).

Monochronistic cultures may expect individuals to be highly committed to their job as well as the organisation while displaying a single-minded approach towards their tasks and projects and focusing solely on individual tasks. Thus, monochronistic cultures may not tolerate emergent leadership given that it involves individuals taking on additional tasks and roles in their team in a non-normative manner. As emergent leadership requires individuals to adopt an interactive approach to tasks, projects and work-related planning, a polychronistic country will provide a more suitable environment for individuals to naturally assume leadership.

Universalism-particularism is the last theme of Nardon and Steers' (2009) framework which is concerned with the rules as a method of minimising uncertainty in societies. In principle, in universalistic cultures, there is a tendency to adhere to and respect formal and societal rules, and regulations as well as the law and bureaucracy. This tendency mainly stems from uncertainty avoidance, which can be described as the desire to control unanticipated actions or behaviours in society. In the context of business, universalistic culture takes the form of keeping constant and thorough records of dealings and organisational practices, while carrying out all processes "by the book". By contrast, particularistic cultures tend to use influential individuals, as opposed to notional and impersonal rules and regulations, as a method of controlling society. This culture of social control can be observed in families, organisations even in friendships in the form of influential individuals governing less influential ones. The longevity of this type of influence is secured by not rules but by mutual trust between parties. Individuals believe that trust should have precedence over formal rules and that some level of flexibility is required in bureaucracy.

Individuals assuming leadership in universalistic cultures could meet with resistance or be disapproved, due to the uncertain and ambiguous nature of emergent leadership. If a member were to attempt to take on leadership roles informally, this was expected to be done by following a formal procedure (i.e., issuing a contract, and providing a clear job description).

However, such formal agreement may also grant the individual "a formal status or authority" which contradicts the definition of emergent leadership (Hanna et al., 2021). Particularistic countries are expected to welcome emergent leaders and promote emergent leadership, considering that the culture relies on influential individuals and supports informal approaches and dynamics.

4. Methodology

The problem outlined above is addressed by conducting an integrative review of the literature on emergent leadership and culture. Integrative literature reviews involve reviewing, critically assessing, and synthesising respective literature on a certain topic in an integrative manner to create new frameworks and approaches on the matter (Torraco, 2005; Webster & Watson, 2002). Scholars underline the particular suitability of this topic when extant research has not been systematically analysed and integrated and when the problem area is relatively novel and unexplored (Snyder, 2019). Since this paper's area of focus fits the abovementioned description, in this paper, the literature on emergent leadership and culture is reviewed, assessed and synthesised into a framework that presents an overarching view of the topic.

This paper consists of one main body of literature, which is the literature on emergent leadership from an organisational, team-level and individual perspective. This is followed by a brief review of culture, in particular of national culture, which altogether provides a multifaceted understanding of emergent leaders who are embedded in the culture of their team, of their organisation and of the society in which they live. The decision to explore the topic of culture on both macro and micro level dimensions stemmed from the suggestions of scholars who indicated that leadership emergence should be studied as a socially constructed process involving complex social mechanisms and organisational actors (Virtaharju & Liiri, 2019; Wu et al., 2021). Likewise, ILTs suggest that people construct cognitive models of reality and employ these preexisting ideas to understand their environment and form their actions, which also extends to how they exhibit leadership behaviours (Den Hartog et al., 1999). Thus, as opposed to solely evaluating national culture and making macro-level generalisations, emergent leaders are analysed as individuals who are embedded in different cultural levels. All cultural levels are later evaluated across the main elements of emergent leadership that were outlined above to reach conclusions on the matter.

In terms of the process for selecting and reviewing the literature, Dwertmann & van Knippenberg's (2021) categorisation approach to integrative reviews was adopted. Subsequently, Elsbach and van Knippenberg's (2020) suggestions have guided the process of synthesising insights gained from the review to develop a new perspective on the literature. Since the ultimate aim of this dissertation is to theorise about the influence of different cultures in which individuals are embedded on how emergent leadership unfolds, emergent leadership literature has been reviewed thoroughly. Broad search terms ("emergen* AND "leader*") utilised on SCOPUS, including all papers published before July 2023. It is believed that the broad search terms would capture publications focusing on relevant concepts (leadership emergence, teams without assigned leaders) more effectively than narrow search terms.

Additionally, the references of prior reviews on emergent leadership have been manually checked to ensure utmost comprehensiveness (Badura et al., 2022; Badura et al., 2018; Hanna et al., 2021; Wolfram Cox et al., 2022)

The database search resulted in a total of 16,767 items. After deletions of duplicates by utilising Zotero software, 16,660 items remained for title screening. The titles of all 16,660 papers were scanned and publications that were explicitly disparate from emergent leadership were removed. Following that, a total of 519 papers remained and underwent abstract screening. The author further screened each paper to ensure that they met two criteria. Firstly, publications that were directly associated with emergent leadership or leadership emergence research were included. For instance, if a publication used the keywords 'emergence' and 'leadership' in its abstract, but not related to each other, the paper was omitted. Publications that focus on markedly different leadership concepts, such as assigned leadership, with no insight into emergent leadership were also removed. Secondly, any publications that were focused on nonhuman subjects were omitted (see Pugliese et al., 2015; Sueur, 2011; C. Wang et al., 2017).

The succeeding content analysis was conducted iteratively, and 134 records were coded using a coding criterion that has been developed to capture a comprehensive view across eight categories: the year of publication, country of publication, type of national culture dimensions, theories, research design, sample, measure(s) (if applicable) and conclusions (at individual, team and organisation level). While listing the country of publication and assessing the type of national culture dimensions to which publications belong, Nardon and Steers's (2009) core cultural dimensions (CCDs) of individualism-collectivism, hierarchy-equality, mastery-harmony, monochronism-polychronism and universalism-particularism is used as guidance.

5. Emergent Leadership across Cultural Levels: An Overarching Framework

Trompenaars and Hampden-Turner (2004) suggest that culture consists of layers “like an onion” and that analysing different layers is necessary to understand it. Likewise, based on the literature review, this paper proposes a theoretical framework that is shaped by the analysis of different levels of culture (i.e., individual, team, organisational and national culture) and theorises about the manner in which emergent leadership unfolds in such settings.

The Emergent Leadership across Cultural Levels (ELCL) framework (Figure 5) draws on Wellman's (2017) relational models leadership theory and proposes a novel approach by incorporating all cultural levels in which individuals are embedded while attempting to understand the emergent leadership phenomenon by evaluating the interaction between potential emergent leaders and each cultural level. It is a socially oriented framework that addresses earlier calls (Hanna et al., 2021; Tabassum et al., 2023), elaborates on recent findings (Galvin et al., 2023) and contributes to the extant theory while presenting a unique perspective that emphasises the role of different cultural levels in how emergent leadership unfolds. It is also believed to provide valuable insights and directions for future research due to the pioneering role it plays in intercultural emergent leadership research. In particular, this framework:

a) allows emergent leadership to be understood in complex social contexts involving different levels of cultural actors; the national culture in which potential emergent leaders live, the organisational culture in which potential emergent leaders operate, the team culture to which potential emergent leaders belong and the individual culture which consists of the attitudes and personality traits of potential emergent leaders.

b) classifies the unfolding of emergent leadership at four levels: organic emergence, nonnormative emergence, conditional emergence and non-emergence based upon the kind of interaction between cultural levels and potential emergent leaders.

c) presents practical implications for organisations regarding managing diverse workforces and ways of facilitating emergent leadership.

The ELCL framework categorises culture across three main cultural levels (i.e., national culture, organisational culture, team culture), discusses the role of individual features of potential emergent leaders (i.e., personality traits and attitudes) and proposes four different ways emergent leadership may unfold in relation to the interaction between cultures in which individuals are embedded. All components of this framework are explained, starting from an individual-level analysis of potential emergent leaders, followed by considerations of national, organisational and team culture and finally a detailed description of the four different ways that emergent leadership may unfold.

5.1. Potential Emergent Leaders

Following Trompenaars' (2004) onion analogy, the core of the onion is the individual culture which is an amalgam of the values, experiences, assumptions, knowledge, personality traits and many other factors that shape the individual as a person.

Wellman (2017) proposes that individuals emerge in a team based on either their leader prototypical qualities (i.e., intelligence, dedication, charisma, dominance) or group prototypical qualities (i.e., kindness, empathy, warmth, fairness). This first option generally occurs in more traditional and formal leadership arrangements, whereas the latter is more likely to occur in informal leadership styles, including emergent leadership. It is important to note that the literature has been more inclined to examine the positive aspects of emergent leaders, emphasising leaders who promote shared cognition, compassion, egalitarian values and participative decision-making (DeRue & Ashford, 2010; Morgeson et al., 2010). This paper recognises that emergent leadership may not always lead to positive outcomes, however, since the current literature provided little insight into the negative sides of this phenomenon, the ELCL model also focuses on the individual antecedents that are positively correlated with emergent leadership.

Considering that one of the key elements of emergent leadership, unit of analysis, underlines that this phenomenon originates from an individual and occurs at an individual level (Kickul & Neuman, 2000), the ELCL framework recognises the pivotal role of individuals in how emergent leadership unfolds. Drawing upon Wellman's (2017) postulations and extant literature on emergent leadership, this framework lists individual-level antecedents that facilitate or have a positive relationship with emergent leadership under the umbrella term of compatible individual antecedents (Table 1).

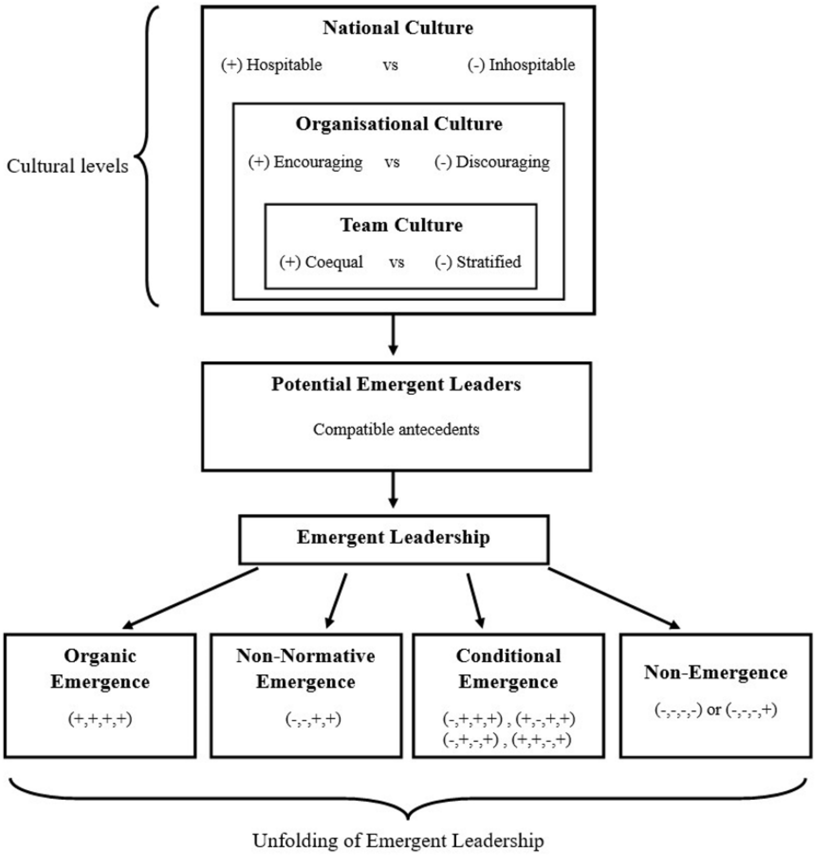


Figure 6: The Overarching Framework of Emergent Leadership across Cultural Levels.

Emergent leadership being an individual-level phenomenon also highlights that individuals assume a leadership role on their own account (DeRue & Ashford, 2010). Given that emergent leaders do not have any formal responsibility over the team or task outcomes, it can be theorised that individual-level antecedents that are positively linked to leadership emergence determine whether an individual chooses to assume leadership as well as whether emergence leadership occurs.

From a motivational point of view, individuals informally stepping up as leaders and taking on additional responsibilities are a form of effort that arises from individuals themselves (DeRue et al., 2015), thereby requiring prosocial motivation (Mitchell & Bommer, 2018). Moreover, it is critical to highlight the predictor role of certain personality traits (i.e., assertiveness, extraversion, agreeableness, emotional stability and conscientiousness) in the emergence of emergent leaders. Thus, the model proposes that:

> *Proposition 1.* For emergent leadership to occur, potential emergent leaders must have or display some of the compatible antecedents.

It is essential to highlight that this proposition does not undermine the key role of extrinsic factors, in particular, team culture, organisational culture and national culture in which individuals are embedded in how emergent leadership unfolds. It rather pinpoints that this phenomenon occurs at an individual level (Hanna et al., 2021), hence, unless individuals have some of the compatible antecedents, emergent leadership is unlikely to occur. Instances where individuals have compatible antecedents, however, external social barriers towards emergent leadership exist are discussed in the following sections.

5.2. National Culture

As the review on national culture and the latest, well-cited framework of core cultural dimensions (Nardon & Steers, 2009) illustrated, how societies structure power relationships, carry out social organisation, perceive time as well as the degree to which they attempt to control their surroundings and minimise uncertainty differ. While some countries have similar tendencies and preferences, some sit at opposite ends of the cultural spectrum. Likewise, Hofstede and Hofstede (2005) state that national culture is distinctive, as it is formed based on unique combinations of the language, religion, values, perceptions and behaviour of the people of that nation.

In the context of this paper, it is believed that there is merit in reevaluating Nardon and Steers' (2009) five dimensions through the lens of emergent leadership and categorising them as *hospitable* and *inhospitable* cultures for emergent leaders to emerge. Hospitable cultures refer to the cultures

Table 1: List of Compatible Individual Antecedents

Compatible Individual Antecedents	Key References
Extraversion	Judge et al., 2002
Agreeableness	Cogliser et al., 2012
Assertiveness	Hu et al., 2019; Judge et al., 2002
Conscientiousness	Cogliser et al., 2012; Lord et al., 1986
Creativity	Ensari et al., 2011
Social intelligence	Walter et al., 2012
Communication skills	Hu et al., 2019; Judge et al., 2002
Self-esteem	Andrews, 1984; Ellis et al., 1988
Emotional Stability	Judge et al., 2002
Openness	Emery et al., 2013
Self-efficacy	Kwok et al., 2018; Serban et al., 2015
Empathy	Wolff et al., 2002
Prosocial motivation	Mitchell and Bommer, 2018
Cognitive ability	Kickul and Neuman, 2000
Openness to experience	Kickul and Neuman, 2000
Self-monitoring	Dobbins et al., 1990
Positive body language	Sanchez-Cortes et al., 2010, 2012
Sense of achievement	Schlamp et al., 2021
Relation-oriented communication	Gerpott et al., 2019
Leadership Competency	Truninger et al., 2021
Warmth	DeRue et al., 2015
Leader-member exchange	Zhang et al., 2012

in which emergent leaders are likely to emerge and receive support. By contrast, inhospitable cultures refer to the cultures in which emergent leaders are unlikely to emerge and be supported. While hospitable cultures are located on the positive side of the spectrum, inhospitable cultures are located on the negative side of the spectrum. The full list of hospitable and inhospitable cultures is below (Table 2).

The ELCL framework acknowledges the dominant nature of national culture, proposing that national culture influences the culture of organisations that operate in that particular country (Lindholm, 2000). Research underpins that organisational cultures are usually a reflection of the values, beliefs and ideologies of the founders of the company, especially during the initial development stage (Robbins, 2003).

National culture not only affects the organisational culture but also the employees of those organisations (Buchanan & Huczynski, 2004). Given that individuals need to adapt to organisational culture to some extent, their behaviour is bound to be influenced accordingly (Thomas, 2008). When these individuals perform their jobs in a team, the team culture also gets affected by their behaviour (Jung & Hong, 2008), thereby linking national culture, organisational culture, team culture and the individual. It is based on this argument that the ELCL model theorises about the interaction between the cultural levels and individuals (in this case, potential emergent leaders).

> *Proposition 2.* The manner in which emergent leadership unfolds is determined by the type of interaction between cultural levels and potential emergent leaders who are embedded in those cultures.

Research corroborates that hospitable national culture may facilitate emergent leadership; however, it may not be sufficient to solely determine whether emergent leadership occurs or how it occurs (Steers et al., 2012). Although it is evident that national culture influences organisational and team cultures as well as individuals operating in them (Dickson et al., 2012), the social dynamics of the organisation and team, alongside individual characteristics should be entered into the equation when theorising about how emergent leadership unfolds.

5.3. Organisational Culture

Organisations convey messages through their vision, strategy, structure and reward systems, all of which can directly affect individuals' behaviour at work (Molina-Azorín et al., 2017; Turner et al., 2017), including whether they emerge as leaders (Meyer et al., 2005). These messages altogether form the organisational culture (Dickson et al., 2009), which can both encourage and discourage non-prototypical individuals from emerging as leaders. Correspondingly, some researchers advise organisations to recalibrate their structures and reward systems in a way that would facilitate emergent leadership in the organisation, in particular among the individuals who would not typically emerge as leaders (Wolfram Cox et al., 2022).

The ELCL model argues that a comprehensive approach involving an analysis of multiple cultural levels is warranted.

Table 2: List of hospitable and inhospitable cultures at a national level, Nardon and Steers (2009).

Hospitable	Inhospitable
Egalitarian	Hierarchical
Individualistic	Collectivistic
High in Mastery	Monochronistic
High in Harmony	Universalistic
Polychronistic	
Particularistic	

As the review on emergent leadership outlined, some scholars adopted a social constructionist perspective which states that leadership stems from contextual collective attempts in a sense-making process (DeRue & Ashford, 2010; Drath et al., 2008). During these processes, members tend to link ideal leadership with several actions, individuals or practices in organisational settings. Certain scholars have highlighted that the ideal leadership behaviour does not have a unique or exceptional character but is instead composed of typical organisational activities and procedures (Crevani et al., 2010; Larsson & Lundholm, 2010), highlighting the significant role of mundane, day-to-day organisational actions.

The constructionist view places individuals in the centre (Fairhurst & Grant, 2010), however, it provides limited insight into the social contexts and external actors, in other words, the "situational opportunities and constraints that affect the occurrence and meaning of organizational behavior" (Johns, 2006, p. 386). On the other hand, researchers who investigated the role of organisational culture in leadership emergence argued that organisational contexts act as the primary source for the emergence of leaders, downplaying the influence of individuals' own characteristics and other external social factors on leadership (Virtaharju & Liiri, 2019). The ELCL model recognises that merging previous perceptions and approaching the matter comprehensively is key while acknowledging that organisational contexts are beyond the situations and norms that solely impact leader effectiveness or that potential emergent leaders must conform (Luria et al., 2019; Tett & Guterman, 2000). Thus, based on prior research outcomes, the model introduces two categories of organisational culture which represent positive and negative organisational cultures that are either *encouraging* or *discouraging* towards emergent leadership respectively.

Encouraging organisational culture refers to the attitude and environment within an organisation that functions as a unified entity (Schneider, 1975), works towards building a shared vision and offers a supportive environment in which organisation-wide knowledge sharing, creativity and communication are promoted (Comfort & Okada, 2013; Erkic, 2022). Conversely, in organisations with discouraging organisational culture traditional structures and complexity are observed, where there is a lack of knowledge-sharing, innovation and available sources for individuals to informally assume leadership (Eisenbeiß & Giessner, 2012; Sharfman & Dean, 1991). In line with Wellman's (2017) arguments, organisations with encouraging culture send important cues to members emphasising that they are similar, whereas organisations with discouraging culture stress the differences among individuals who operate in that organisation.

5.4. Team Culture

Relational models leadership theory (Wellman, 2017) suggests that when members decide how leadership should be performed in the team, they either adopt the authority ranking model, which refers to a hierarchical approach where individuals assess all members and defer to the ones whom they perceive to have the most leadership qualities. In this model, the responsibility is solely given to the leaders and others are expected to support the leader, such as the military and police departments. Alternatively, members adopt the communal sharing model, which refers to all members of the team being perceived as equally valuable with relevant insights that can contribute to the team. This model emphasises consensus as opposed to deferring to the authority figure. Teams with informal and flat structures (i.e., holacracy at Zappos), can be a good example of this model (Perschel, 2010). This theory highlights that decision-making is directly influenced by the pressures within the wider environment while explaining how leadership unfolds within teams. In the teams that adopt the authority ranking model, leaders emerge based on their distinctive qualities. In contrast, in teams with communal sharing model, leaders emerge based on the similarities between them and other team members (Wellman, 2017).

The review on emergent leadership highlighted two main points on a team level. Firstly, when teams have a shared social identity and vision, this generates a suitable environment for individuals to emerge as leaders (Smith et al., 2018; Zhang et al., 2012). Secondly, a supportive team environment in which members feel comfortable facilitates the emergence of emergent leaders (DeRue et al., 2015; Mumford et al., 2008).

Building on Wellman's (2017) postulations and prior findings in the literature, the ELCL model presents two categories in which teams can be explored in the context of the focus of this paper. The first category of team culture is specified as *coequal* culture, where team members are seen as equals, shared goals are communicated effectively, decision-making occurs democratically, and creativity and initiative-taking are encouraged. The second category is *stratified* culture, referring to teams adopting hierarchical structures where members are expected to show respect to the leader and tend to behave and work according to the leader's orders with little room for individuals to "step up" or behave outside the norms. Coequal team culture is located on the positive side of the spectrum, whereas stratified team culture is located on the negative side of the spectrum.

> *Proposition 3.* For emergent leadership to occur in stratified teams, the approval of higher level authority figures at the organisational or national level are needed.

As described above, the culture in stratified teams is unlikely to organically provide ordinary team members with the opportunity to emerge as a leader. Moreover, it is expected such teams to show resistance towards a junior employee attempting to assume leadership roles, as this could be perceived as disrespecting the norms and senior members of the team and undermining the social dynamics of the team. However, the overarching perspective of the ELCL model recognises the possibility of certain cases where emergent leaders are approved by a high-level hierarchical figure and allowed to undertake informal leadership roles. This will be evaluated in the following sections when *conditional emergence* is discussed.

5.5. The Unfolding of Emergent Leadership

As noted previously, the literature on emergent leadership has predominantly identified antecedents that are associated with individuals emerging as leaders as well as team-level dynamics that affect emergent leadership (Galvin et al., 2023; Hanna et al., 2021). However, researchers have yet to reveal the influence of more macro-level elements, such as national culture and organisational culture, in the manner in which this phenomenon unfolds. Considering the evidence indicating that certain social dynamics may favour the emergence of specific types of leaders (Grint, 2005), it is expected to observe different leadership types emerge in different social contexts. Congruous with prior papers on emergent leadership (Badura et al., 2022) and the wider organisational literature (Johns, 2006), this framework suggests that based on the interactions between different cultural levels (i.e., national, organisational and team) and individual antecedents, how emergent leadership unfolds can be divided into four categories: *Organic emergence, non-normative emergence, conditional emergence* and *non-emergence*.

5.5.1. Organic Emergence

Organic emergence refers to the instances in which all cultural levels are on the positive side of the spectrum intertwining harmoniously, and the potential emergent leaders have compatible antecedents that will allow them to organically assume leadership responsibilities. To elaborate, when national culture is hospitable, organisational culture is encouraging, team culture is coequal and potential emergent leaders have compatible antecedents that are positively linked to emergent leadership, emergent leadership is expected to occur organically. In this instance, from the macro-level dynamics to individuals, the environment and conditions are highly suitable for potential emergent leaders to assume leadership.

> *Proposition 4.* When all cultural levels are on the positive side of the spectrum, and potential emergent leaders have compatible traits and consonant values, organic emergence occurs.

5.5.2. Non-normative Emergence

Non-normative emergence refers to situations where there are barriers to emergent leadership in national culture and organisational culture, yet the team culture is coequal, allowing potential emergent leaders with compatible antecedents to emerge as leaders. For instance, building on the example of Soluk and Kammerlander (2021), a socially intelligent, open, creative and extroverted individual with sound communication skills works in the technology team of a family-owned German firm located in Germany. The technology team has been formed recently as part of the company's digital transformation initiatives. Thus, the culture in this team is in contrast with the company's rule-oriented and rigid organisational culture (Nardon & Steers, 2009), as it has an agile and egalitarian culture that promotes initiative-taking. In this case, it is argued that the support of the team would enable the individual to assume informal leadership responsibilities, despite the barriers at the macro level. However, since the national culture and the company culture would strictly prefer individuals to carry out their work "by the book" (Hofstede, 1980), the individual can only emerge as an emergent leader with the support of their team and, more importantly, by going against the macro-level (national and organisational) norms.

> *Proposition 5.* Even when national culture and organisational culture are on the negative end of the spectrum, if team culture is on the positive side and potential emergent leaders have compatible antecedents, non-normative emergence occurs.

5.5.3. Conditional Emergence

Compared to non-normative emergence, in conditional emergence potential emergent leaders have a greater level of support, which comes either from an organisational or national level or team level and organisational and national level. Hence, when the individual emerges as an emergent leader, they do not go against the norms per se. However, since all cultural levels are not in the position of empowering or enabling emergent leadership, in this instance, the emergent leader remains limited in where and how they operate, and their leadership emerges based on conditions. These conditions include their leadership being blessed by high-level influential leaders of the organisation or nation. The Fuyao Glass and GM affair in 2014, where a Chinese glass manufacturer bought a former General Motors assembly plant in Ohio (Gawley & Dixon, 2020), can be a suitable example of how conditional emergence may occur. In this instance, adhering to rigid rules and formal procedures is a part of hierarchical and collectivistic Chinese culture, where all business operations must be approved and decided by the head of the organisation (Nardon & Steers, 2009). Conversely, American culture adopts egalitarian and individualistic approaches to business (Hofstede, 1980). When Fuyao bought the assembly plant, most of the workers were Americans who previously worked in an organisation that was influenced by American values. After this change, although workers were still located in the US, the organisational culture was bound to adapt to Fuyao's organisational culture (discouraging). In such settings, considering that the potential emergent leader's team has a coequal culture, the individual may only emerge as a leader on the condition of receiving approval from the high-level authority figures of the organisation.

> *Proposition 6.* Conditional emergence may occur in three ways a) when national culture is positive, organisational culture is negative, team culture is positive and potential emergent leaders have compatible antecedents b) when national culture is negative, but organisational culture is positive, team culture is positive and potential emergent leaders have compatible antecedents c) when national and organisational culture are on the positive side of the spectrum (i.e., hospitable, encouraging respectively), potential emergent leaders have compatible antecedents but team culture is negative.

5.5.4. Non-Emergence

Non-emergence refers to the situations in which emergent leadership does not occur or is unlikely to occur. The literature indicates that emergent leadership originates from the individual, but also is affected by external social contexts (DeRue & Ashford, 2010; Goktepe & Schneier, 1988). Given that one of the key elements of emergent leadership is being perceived as leaderlike by others (Hanna et al., 2021), without other members at the team level or organisational level viewing the individual as a leader, one of the main conditions of emergent leadership cannot be fulfilled. This proposition is also in line with Galvin et al.'s (2023) under-emergence, when well-equipped leaders who are not perceived as leaderlike fail to emerge as leaders. Hence, the model theorises that:

> *Proposition 7.* When all cultural levels are on the negative side of the spectrum, creating a combination of inhospitable, discouraging and stratified cultures, irrespective of whether potential emergent leaders have compatible antecedents, non-emergence occurs.

6. Conclusion

This paper extended the emergent leadership literature and equipped researchers, employees and business leaders with the knowledge that will allow them to better understand the dynamics involved in the emergence of emergent leadership across different cultural levels. In particular, the ELCL model proposed that for emergent leadership to occur, potential emergent leaders must have or display some of the compatible antecedents and how emergent leadership unfolds is determined by the type of interaction between cultural levels and potential emergent leaders who are embedded in those cultures.

Acknowledging the limitations concerning the assumptions of the ELCL framework is crucial. Primarily, it is accepted that leader emergence can be conceptualised in a variety of ways and can take place over a spectrum. Secondly, culture is fluid and norms attributed to societies may change over time (see Alkan et al., 2023). Further, cultures could be categorised in numerous ways along a spectrum, thereby identifying a culture as positive or negative in the context of leadership emergence may not always be straightforward. However, by using the most comprehensive cultural framework consisting of influential cross-cultural leadership studies which examined the relationship of cultural practices and values at the level of society and organisations (Nardon & Steers, 2009), and basing the model on key findings of the relevant literature, these limitations were attempted to be minimised.

Practical implications for organisations include utilising the knowledge of what is valued in a leader across different cultures in forming organisational learning and development practices, thereby enhancing cultural fluency across the organisation and increasing retention (Dorfman et al., 2004). Employers and managers can be trained on the characteristics of the national culture, organisational culture, and team culture in which they operate, while also learning about compatible individual antecedents of emergent leadership. This may be even more crucial in virtual teams, considering that the lack of physical interaction may generate an additional barrier for members to understand and interpret each other's messages.

Additionally, diverse perspectives and critiques of this paper will provide a foundation for future research in emergent leadership. Since the ELCL model is developed based on secondary research, its propositions and conclusions must be tested through both qualitative and quantitative empirical research.

References

Acton, B. P., Foti, R. J., Lord, R. G., & Gladfelter, J. A. (2019). Putting emergence back in leadership emergence: A dynamic, multilevel, process-oriented framework. *The Leadership Quarterly*, *30*(1), 145–164. https://doi.org/10.1016/j.leaqua.2018.07.002

Adler Jr, G. J. (2023). Culture, the civic, and religion: characteristics and contributions of cultural analysis through three exemplary books. *American Journal of Cultural Sociology*, *11*(3), 365–381.

Ahearne, M., Mathieu, J., & Rapp, A. (2005). To empower or not to empower your sales force? An empirical examination of the influence of leadership empowerment behavior on customer satisfaction and performance. *Journal of Applied psychology*, *90*(5), 945. https://doi.org/10.1057/s41290-022-00155-4

Alkan, D. P., Özbilgin, M. F., & Kamasak, R. (2023). The Leadership in Tackling the Unforeseen Consequences of the Covid-19 Pandemic: Who Is the Emergent Leader? In *Innovation, Leadership and Governance in Higher Education: Perspectives on the Covid-19 Recovery Strategies* (pp. 235–255). Springer Nature. https://doi.org/10.1007/978-981-19-7299-7_13

Andersson, T., & Tengblad, S. (2016). An experience based view on leader development: leadership as an emergent and complex accomplishment. *Development and Learning in Organizations: An International Journal*, *30*(6), 30–32.

Andrews, P. H. (1984). Performance-self-esteem and perceptions of leadership emergence: A comparative study of men and women. *Western Journal of Speech Communication*, *48*(1), 1–13. https://doi.org/10.1080/10570318409374137

Badura, K. L., Galvin, B. M., & Lee, M. Y. (2022). Leadership emergence: An integrative review. *Journal of Applied Psychology*, *107*(11), 2069–2100. https://doi.org/10.1037/apl0000997

Badura, K. L., Grijalva, E., Newman, D. A., Yan, T. T., & Jeon, G. (2018). Gender and leadership emergence: A meta-analysis and explanatory model. *Personnel Psychology*, *71*(3), 335–367. https://doi.org/10.1111/peps.12266

Barling, J., & Weatherhead, J. G. (2016). Persistent exposure to poverty during childhood limits later leader emergence. *Journal of Applied Psychology*, *101*, 1305–1318.

Buchanan, D. A., & Huczynski, A. (2004). *Organisational Behaviour: An Introductory Text* (5th). FT Prentice.

Butler, C. L., Zander, L., Mockaitis, A., & Sutton, C. (2012). The Global Leader as Boundary Spanner, Bridge Maker, and Blender. *Industrial and organizational psychology*, *5*(2), 240–243. https://doi.org/10.1111/j.1754-9434.2012.01439.x

Byrne, G. J., & Bradley, F. (2007). Culture's influence on leadership efficiency: How personal and national cultures affect leadership style. *Journal of Business Research*, *60*(2), 168–175. https://doi.org/10.1016/j.jbusres.2006.10.015

Carte, T. A., Chidambaram, L., & Becker, A. (2006). Emergent leadership in self-managed virtual teams: A longitudinal study of concentrated and shared leadership behaviors. *Group Decision and Negotiation*, *15*(4), 323–343. https://doi.org/10.1007/s10726-006-9045-7

Charlier, S. D., Stewart, G. L., Greco, L. M., & Reeves, C. J. (2016). Emergent leadership in virtual teams: A multilevel investigation of individual communication and team dispersion antecedents. *The Leadership Quarterly*, *27*(5), 745–764. https://doi.org/10.1016/j.leaqua.2016.05.002

Cogliser, C. C., Gardner, W. L., Gavin, M. B., & Broberg, J. C. (2012). Big five personality factors and leader emergence in virtual teams: Relationships with team trustworthiness, member performance contributions, and team performance. *Group & Organization Management*, *37*(6), 752–784. https://doi.org/10.1177/1059601112464266

Colbert, A. E., Judge, T. A., Choi, D., & Wang, G. (2012). Assessing the trait theory of leadership using self and observer ratings of personality: The mediating role of contributions to group success. *The leadership quarterly*, *23*(4), 670–685. https://doi.org/10.1016/j.leaqua.2012.03.004

Comfort, L. K., & Okada, A. (2013). Emergent leadership in extreme events: A knowledge commons for sustainable communities. *International Review of Public Administration*, *18*(1), 61–77. https://doi.org/10.1080/12294659.2013.10805240

Crevani, L., Lindgren, M., & Packendorff, J. (2010). Leadership, not leaders: On the study of leadership as practices and interactions. *Scandinavian journal of management*, *26*(1), 77–86. https://doi.org/10.1016/j.scaman.2009.12.003

Crozier, A. J., Loughead, T. M., & Munroe-Chandler, K. J. (2017). Top-down or shared leadership? Examining differences in athlete leadership behaviours based on leadership status in sport. *Physical Culture*, *71*(2), 86–98. https://doi.org/10.5937/fizkul1702086K

Den Hartog, D. N., House, R. J., Hanges, P. J., & Ruiz-Quintanilla, S. A. (1999). Culture specific and cross-culturally generalizable implicit leadership theories: Are attributes of charismatic/transformational leadership universally endorsed? *The Leadership Quarterly*, *10*(2), 219–256.

DeRue, D. S., & Ashford, S. J. (2010). Who will lead and who will follow? A social process of leadership identity construction in organizations. *Academy of management review*, *35*(4), 627–647.

DeRue, D. S., Nahrgang, J. D., & Ashford, S. J. (2015). Interpersonal perceptions and the emergence of leadership structures in groups: A network perspective. *Organization Science*, *26*(4), 1192–1209.

Dickson, M. W., Castaño, N., Magomaeva, A., & Den Hartog, D. N. (2012). Conceptualizing leadership across cultures. *Journal of World Business*, *47*(4), 483–492. https://doi.org/10.1016/j.jwb.2012.01.002

Dickson, M. W., Den Hartog, D., & Castaño, N. (2009). Understanding leadership across cultures. In R. Bhagat & R. Steers (Eds.), *Cambridge handbook of culture, organizations, and work* (pp. 219–243). Cambridge University Press.

D'Innocenzo, L., Mathieu, J. E., & Kukenberger, M. R. (2016). A Meta-Analysis of Different Forms of Shared Leadership–Team Performance Relations. *Journal of Management*, *42*(7), 1964–1991. https://doi.org/10.1177/0149206314525205

Dobbins, G. H., Long, W. S., Dedrick, E. J., & Clemons, T. C. (1990). The role of self-monitoring and gender on leader emergence: A laboratory and field study. *Journal of Management*, *16*(3), 609–618.

Doblinger, M. (2022). Individual competencies for self-managing team performance: A systematic literature review. *Small Group Research*, *53*(1), 128–180.

Dorfman, P. W., Hanges, P. J., & Brodbeck, F. C. (2004). Leadership and cultural variation: The identification of culturally endorsed leadership profiles. In R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, V. Gupta, & G. Associates (Eds.), *Culture, leadership, and organizations: The GLOBE study of 62 societies* (pp. 669–720). Sage.

Drath, W. H., McCauley, C. D., Palus, C. J., Van Velsor, E., O'Connor, P. M., & McGuire, J. B. (2008). Direction, alignment, commitment: Toward a more integrative ontology of leadership. *The leadership quarterly*, *19*(6), 635–653. https://doi.org/10.1016/j.leaqua.2008.09.003

Dwertmann, D. J., & van Knippenberg, D. (2021). Capturing the state of the science to change the state of the science: A categorization approach to integrative reviews. *Journal of Organizational Behavior*, *42*(2), 104–117. https://doi.org/10.1002/job.2474

Eisenbeiß, S. A., & Giessner, S. R. (2012). The emergence and maintenance of ethical leadership in organizations: A question of embeddedness? *Journal of Personnel Psychology*, *11*(1), 7–19. https://doi.org/10.1027/1866-5888/a000055

Ellis, R. J., Adamson, R. S., Deszca, G., & Cawsey, T. F. (1988). Self-monitoring and leadership emergence. *Small Group Research*, *19*(3), 312–324. https://doi.org/10.1177/104649648801900302

Elsbach, K. D., & van Knippenberg, D. (2020). Creating high-impact literature reviews: An argument for 'integrative reviews'. *Journal of management studies*, *57*(6), 1277–1289. https://doi.org/10.1111/joms.12581

Ely, R. J. (2004). A field study of group diversity, participation in diversity education programs, and performance. *Journal of Organizational Behavior*, *25*(6), 755–780.

Emery, C. (2012). Uncovering the role of emotional abilities in leadership emergence: A longitudinal analysis of leadership networks. *Social Networks*, *34*(4), 429–437. https://doi.org/10.1016/j.socnet.2012.02.001

Emery, C., Calvard, T. S., & Pierce, M. E. (2013). Leadership as an emergent group process: A social network study of personality and leadership. *Group Processes and Intergroup Relations*, *16*(1), 28–45. https://doi.org/10.1177/1368430212461835

Emery, C., Carnabuci, G., & Brinberg, D. (2011). Relational schemas to investigate the process of leadership emergence. *Acad. Manage. Annu. Meet. - West Meets East: Enlightening. Balancing. Transcending*. https://doi.org/10.5464/AMBPP.2011.143.a

Ensari, N., Riggio, R. E., Christian, J., & Carslaw, G. (2011). Who emerges as a leader? Meta-analyses of individual differences as predictors of leadership emergence. *Personality and Individual Differences*, *51*(4), 532–536. https://doi.org/10.1016/j.paid.2011.05.017

Erkic, A. (2022). Emergent leadership: why and how to let your team take the lead.

Fairhurst, G. T., & Grant, D. (2010). The social construction of leadership: A sailing guide. *Management communication quarterly*, *24*(2), 171–210. https://doi.org/10.1177/089331890935969

Fyhn, B., Schei, V., & Sverdrup, T. E. (2023). Taking the emergent in team emergent states seriously: A review and preview. *Human Resource Management Review*, *33*(1), 100928. https://doi.org/10.1016/j.hrmr.2022.100928

Galvin, B. M., Badura, K., LePine, J., & LePine, M. (2023). A theoretical integration of leader emergence and leadership effectiveness: Over, under, and congruent emergence. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2724

Gawley, T., & Dixon, S. (2020). Trading Health for Work: Recognizing Occupational Safety and Worker Health in the Film, American Factory. *NEW SOLUTIONS: A Journal of Environmental and Occupational Health Policy*, *31*, 104829112098072. https://doi.org/10.1177/1048291120980728

Gerpott, F. H., Lehmann-Willenbrock, N., Silvis, J. D., & Van Vugt, M. (2018). In the eye of the beholder? An eye-tracking experiment on emergent leadership in team interactions. *Leadership Quarterly*, *29*(4), 523–532. https://doi.org/10.1016/j.leaqua.2017.11.003

Gerpott, F. H., Lehmann-Willenbrock, N., Voelpel, S. C., & Van Vugt, M. (2019). It's Not Just What is Said, but When it's Said: A Temporal Account of Verbal Behaviors and Emergent Leadership in Self-Managed Teams. *Academy of Management Journal*, *62*(3), 717–738. https://doi.org/10.5465/amj.2017.0149

Goktepe, J. R., & Schneier, C. E. (1988). Sex and gender effects in evaluating emergent leaders in small groups. *Sex Roles*, *19*(1–2), 29–36. https://doi.org/10.1007/BF00292461

Grimsley, E. A., Cochrane, N. H., Keane, R. R., Sumner, B. D., Mullan, P. C., & O'Connell, K. J. (2021). A pulse check on leadership and teamwork: An evaluation of the first 5 minutes of emergency department resuscitation during pediatric cardiopulmonary arrests. *Pediatric Emergency Care*, *37*(12), E1122–E1127. https://doi.org/10.1097/PEC.0000000000001923

Grint, K. (2005). Problems, problems, problems: The social construction of 'leadership'. *Human relations*, *58*(11), 1467–1494. https://doi.org/10.1177/0018726705061314

Gruber, F. M., Veidt, C., & Ortner, T. M. (2018). Women who emerge as leaders in temporarily assigned work groups: Attractive and socially competent but not babyfaced or naïve? *Frontiers in Psychology*, *9*, 2553. https://doi.org/doi.org/10.3389/fpsyg.2018.02553

Guastello, S. J., & Bond Jr., R. W. (2007). A swallowtail catastrophe model for the emergence of leadership in coordination-intensive groups. *Nonlinear Dynamics, Psychology, and Life Sciences*, *11*(2), 235–251.

Guastello, S. J. (1995). Facilitative style, individual innovation, and emergent leadership in problem solving groups. *The Journal of Creative Behavior*, *29*(4), 225–239. https://doi.org/10.1002/j.2162-6057.1995.tb01397.x

Hackman, J. R., & Wageman, R. (2004). When and How Team Leaders Matter. *Research in Organizational Behavior*, *26*, 37–74. https://doi.org/10.1016/S0191-3085(04)26002-6

Hall, S. (1993). Culture, community, nation. *Cultural studies*, *7*(3), 349–363.

Hanna, A. A., Smith, T. A., Kirkman, B. L., & Griffin, R. W. (2021). The Emergence of Emergent Leadership: A Comprehensive Framework and Directions for Future Research. *Journal of Management*, *47*(1), 76–104. https://doi.org/10.1177/0149206320965683

Hart, R. K. (2016). Informal Virtual Mentoring for Team Leaders and Members: Emergence, Content, and Impact. *Advances in Developing Human Resources*, *18*(3), 352–368. https://doi.org/10.1177/1523422316645886

Hill, R. C., & Levenhagen, M. (1995). Metaphors and mental models: Sensemaking and sensegiving in innovative and entrepreneurial activities. *Journal of Management*, *21*(6), 1057–1074.

Hoch, J. E., & Dulebohn, J. H. (2017). Team personality composition, emergent leadership and shared leadership in virtual teams: A theoretical framework. *Human Resource Management Review*, *27*(4), 678–693. https://doi.org/10.1016/j.hrmr.2016.12.012

Hofstede, G. (1980). Motivation, leadership, and organization: Do American theories apply abroad? *Organizational Dynamics*, *9*, 42–63.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

Hofstede, G., & Hofstede, G. J. (2005). *Cultures and Organizations: Software of the Mind*. McGraw-Hill.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage Publications.

House, R. J., Hanges, P. J., Ruiz-Quintanilla, S. A., Dorfman, P. W., Javidan, M., Dickson, M., & Gupta, V. (1999). Cultural influences on leadership and organizations: Project GLOBE. *Advances in global leadership*, *1*(2), 171–233.

Hu, J., Zhang, Z., Jiang, K., & Chen, W. (2019). Getting ahead, getting along, and getting prosocial: Examining extraversion facets, peer reactions, and leadership emergence. *Journal of Applied Psychology*, *104*(11), 1369–1386. https://doi.org/10.1037/apl0000413

Huang, X., Iun, J., Liu, A., & Gong, Y. (2010). Does participative leadership enhance work performance by inducing empowerment or trust? The differential effects on managerial and non-managerial subordinates. *Journal of Organizational Behavior*, *31*(1), 122–143. https://doi.org/10.1002/job.636

Ibarra, H., & Barbulescu, R. (2010). Identity as narrative: Prevalence, effectiveness, and consequences of narrative identity work in macro work role transitions. *Academy of management review*, *35*(1), 135–154.

Javidan, M., Dorfman, P. W., De Luque, M. S., & House, R. J. (2006). In the eye of the beholder: Cross cultural lessons in leadership from project GLOBE. *Academy of management perspectives*, *20*(1), 67–90.

Jehn, K. A., & Bezrukova, K. (2004). A field study of group diversity, workgroup context, and performance. *Journal of Organizational Behavior*, *25*(6), 703–729.

Jiang, J., Chen, C., Dai, B., Shi, G., Ding, G., Liu, L., & Lu, C. (2015). Leader emergence through interpersonal neural synchronization. *Proceedings of the National Academy of Sciences*, *112*, 4274–4279.

Jiang, X., Snyder, K., Li, J., & Manz, C. C. (2021). How Followers Create Leaders: The Impact of Effective Followership on Leader Emergence in Self-Managing Teams. *Group Dynamics*, *25*(4), 303–318. https://doi.org/10.1037/gdn0000159

Johns, G. (2006). The Essential Impact of Context on Organizational Behavior. *The Academy of Management review*, *31*(2), 386–408.

Johnson, S. D., & Bechler, C. (1998). Examining the relationship between listening effectiveness and leadership emergence: Perceptions, behaviors, and recall. *Small group research*, *29*(4), 452–471. https://doi.org/10.1177/1046496498294003

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*, 765–780.

Judge, T. A., Piccolo, R. F., & Ilies, R. (2004). The forgotten ones? The validity of consideration and initiating structure in leadership research. *Journal of Applied Psychology*, *89*, 36–51.

Jung, J. Y., & Hong, S. (2008). Organizational citizenship behaviour (OCB), TQM and performance at the maquiladora. *International Journal of Quality & Reliability Management*, *25*(8), 793–808.

Kaiser, R. B., Hogan, R., & Craig, S. B. (2008). Leadership and the fate of organizations. *American psychologist*, *63*(2), 96. https://doi.org//10.1037/0003-066X.63.2.96

Kanfer, R., Chen, G., & Pritchard, R. D. (2008). The three C's of work motivation: Content, context, and change. In *Work motivation: Past, present, and future* (pp. 30–45). Routledge. https://doi.org/10.4324/9780203809501

Kaplan, M., Dollar, B., Melian, V., Van Durme, Y., & Wong, J. (2016). *Shape culture drive strategy*. https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2016/impact-of-cultureon-business-strategy.html

Kent, R. L., & Moss, S. E. (1990). Self-monitoring as a predictor of leader emergence. *Psychological Reports*, *66*(3), 875–881.

Kent, R. L., & Moss, S. E. (1994). Effects of sex and gender role on leader emergence. *Academy of management journal*, *37*(5), 1335–1346.

Kickul, J., & Neuman, G. (2000). Emergent leadership behaviors: The function of personality and cognitive ability in determining teamwork performance and KSAs. *Journal of Business and Psychology*, *15*, 27–51.

Kluckhohn, F. R., & Strodtbeck, F. L. (1961). *Variations in value orientations*. Row, Peterson.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions.*, 3–90.

Kozlowski, S. W., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational research methods*, *16*(4), 581–615.

Kwok, N., Hanig, S., Brown, D. J., & Shen, W. (2018). How leader role identity influences the process of leader emergence: A social network analysis. *The Leadership Quarterly*, *29*(6), 648–662.

Lanaj, K., & Hollenbeck, J. R. (2015). Leadership over-emergence in self-managing teams: The role of gender and countervailing biases. *Academy of Management Journal*, *58*(5), 1476–1494. https://doi.org/10.5465/amj.2013.0303

Landis, B., Jachimowicz, J. M., Wang, D. J., & Krause, R. W. (2022). Revisiting extraversion and leadership emergence: A social network churn perspective. *Journal of Personality and Social Psychology*, *123*(4), 811–829. https://doi.org/10.1037/pspp0000410

Larsson, M., & Lundholm, S. E. (2010). Leadership as work-embedded influence: A micro-discursive analysis of an everyday interaction in a bank. *Leadership*, *6*(2), 159–184. https://doi.org/10.1177/1742715010363208

Lee, S. M., & Farh, C. I. (2019). Dynamic leadership emergence: Differential impact of members' and peers' contributions in the idea generation and idea enactment phases of innovation project teams. *Journal of Applied Psychology*, *104*(3), 411–432. https://doi.org/10.1037/apl000038

Leeming, P. (2019). Emergent leadership and group interaction in the task-based language classroom. *Tesol Quarterly*, *53*(3), 768–793.

Li, Y., Chun, H., Ashkanasy, N. M., & Ahlstrom, D. (2012). A multi-level study of emergent group leadership: Effects of emotional stability and group conflict. *Asia Pacific Journal of Management*, *29*(2), 351–366. https://doi.org/10.1007/s10490-012-9298-4

Lindholm, N. (2000). National Culture and Performance Management in MNC Subsidiaries. *International Studies of Management & Organization*, *29*(4), 45–66.

Loignon, A.-C., & Kodydek, G. (2022). The Effects of Objective and Subjective Social Class on Leadership Emergence. *Journal of Management Studies*, *59*(5), 1162–1197. https://doi.org/10.1111/joms.12769

Lok, P., & Crawford, J. (2004). The effect of organisational culture and leadership style on job satisfaction and organisational commitment: A cross-national comparison. *Journal of Management Development*, *23*(4), 321–338. https://doi.org/10.1108/02621710410529785

Lord, R. G., De Vader, C. L., & Alliger, G. M. (1986). A Meta-Analysis of the Relation Between Personality Traits and Leadership Perceptions: An Application of Validity Generalization Procedures. *Journal of Applied Psychology*, *71*(3), 402–410. https://doi.org/10.1037/0021-9010.71.3.402

Lord, R. G., Epitropaki, O., Foti, R. J., & Hansbrough, T. K. (2020). Implicit Leadership Theories, Implicit Followership Theories, and Dynamic Processing of Leadership Information. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 49–74. https://doi.org/10.1146/annurev-orgpsych-012119-045434

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A Test of Leadership Categorization Theory: Internal Structure, Information Processing, and Leadership Perceptions. *Organizational Behavior and Human Performance*, *34*(3), 343–378. https://doi.org/10.1016/0030-5073(84)90043-6

Lord, R. G., & Maher, K. J. (1991). *Leadership and Information Processing: Linking Perceptions and Performance*. Unwin Hyman.

Lucas, J. W. (2003). Status Processes and the Institutionalization of Women as Leaders. *American Sociological Review*, *68*(3), 464–480. https://doi.org/10.2307/1519733

Luria, G., & Berson, Y. (2013). How do leadership motives affect informal and formal leadership emergence? *Journal of Organizational Behavior*, *34*(7), 995–1015. https://doi.org/10.1002/job.1836

Luria, G., Kahana, A., Goldenberg, J., & Noam, Y. (2019). Leadership Development: Leadership Emergence to Leadership Effectiveness. *Small Group Research*, *50*(5), 571–592. https://doi.org/10.1177/1046496419865326

Maloney, D. (2020). Emergent leadership: The art of letting your team take the lead. https://slack.com/blog/collaboration/emergent-leadership-team-trait

Manz, C. C. (1986). Self-Leadership: Toward an Expanded Theory of Self-Influence Processes in Organizations. *The Academy of Management Review*, *11*(3), 585–600. https://doi.org/10.2307/258312

McClean, E. J., Martin, S. R., Emich, K. J., & Woodruff, C. T. (2018). The social consequences of voice: An examination of voice type and gender on status and subsequent leader emergence. *Academy of Management Journal*, *61*, 1869–1891.

McGrath, J. E. (1962). *Leadership Behavior: Some Requirements for Leadership Training*. U.S. Civil Service Commission, Office of Career Development.

Mertens, N., Boen, F., Steffens, N. K., Haslam, S. A., & Fransen, K. (2021). Will the real leaders please stand up? The emergence of shared leadership in semi-professional soccer teams. *Journal of Science and Medicine in Sport*, *24*(3), 281–290. https://doi.org/10.1016/j.jsams.2020.09.007

Meyer, A. D., Gaba, V., & Colwell, K. A. (2005). Organizing Far from Equilibrium: Nonlinear Change in Organizational Fields. *Organization Science*, *16*(5), 456–473. https://doi.org/10.1287/orsc.1050.0135

Mitchell, T., & Bommer, W. H. (2018). The interactive effects of motives and task coordination on leadership emergence. *Group Dynamics*, *22*(4), 223–235. https://doi.org/10.1037/gdn0000092

Mitchell, T., Lemoine, G. J., & Lee, D. (2022). Inclined but less skilled? Disentangling extraversion, communication skill, and leadership emergence. *Journal of Applied Psychology*, *107*(9), 1524–1542. https://doi.org/10.1037/apl0000962

Molina-Azorín, J. F., Bergh, D. D., Corley, K. G., & Ketchen, D. J. (2017). Mixed Methods in the Organizational Sciences: Taking Stock and Moving Forward. *Organizational Research Methods*, *20*(2), 179–192. https://doi.org/10.1177/1094428116687026

Morgeson, F. P., DeRue, D. S., & Karam, E. P. (2010). Leadership in Teams: A Functional Approach to Understanding Leadership Structures and Processes. *Journal of Management*, *36*(1), 5–39. https://doi.org/10.1177/0149206309347376

Moutafi, J., Furnham, A., & Crump, J. (2007). Is Managerial Level Related to Personality? *British Journal of Management*, *18*(3), 272–280. https://doi.org/10.1111/j.1467-8551.2007.00511.x

Mumford, M. D., Antes, A. L., Caughron, J. J., & Friedrich, T. L. (2008). Charismatic, Ideological, and Pragmatic Leadership: Multi-Level Influences on Emergence and Performance. *Leadership Quarterly*, *19*(2), 144–160. https://doi.org/10.1016/j.leaqua.2008.01.002

Murase, T., Resick, C. J., Jiménez, M., Sanz, E., & DeChurch, L. A. (2013). Leadership and emergent collective cognition. In *Theories of Team Cognition: Cross-Disciplinary Perspectives* (pp. 117–144). Taylor and Francis. https://doi.org/10.4324/9780203813140-15

Murphy, A. J. (1941). A study of the leadership process. *American Sociological Review*, *6*, 674–687.

Nahrgang, J. D., Morgeson, F. P., & Ilies, R. (2009). The Development of Leader–Member Exchanges: Exploring How Personality and Performance Influence Leader and Member Relationships Over Time. *Organizational Behavior and Human Decision Processes*, *108*(2), 256–266. https://doi.org/10.1016/j.obhdp.2008.09.002

Nardon, L., & Steers, R. M. (2009). The culture theory jungle: Divergence and convergence in models of national culture. In R. Bhagat &

R. Steers (Eds.), *Cambridge Handbook of Culture, Organizations, and Work* (pp. 3–22). Cambridge University Press.

Osland, J., Taylor, S., & Mendenhall, M. (2009). Global leadership: Progress and challenges. In R. Bhagat & R. Steers (Eds.), *Cambridge Handbook of Culture, Organizations, and Work* (pp. 245–271). Cambridge University Press. https://doi.org/10.1017/CBO9780511581151.011

Park, J. (2019). Social networks and leadership emergence. *ECMLG 2019 15th European Conference on Management, Leadership and Governance*, 305–314. https://doi.org/10.34190/MLG.19.032

Paunova, M. (2015). The emergence of individual and collective leadership in task groups: A matter of achievement and ascription. *The Leadership Quarterly*, *26*(6), 935–957. https://doi.org/10.1016/j.leaqua.2015.10.002

Pearce, C. L., & Sims, H. P. (2000). Shared leadership: Toward a multi-level theory of leadership. In *Advances in Interdisciplinary Studies of Work Teams* (pp. 115–139). Emerald Group Publishing Limited (Advances in Interdisciplinary Studies of Work Teams). https://doi.org/10.1016/S1572-0977(00)07008-4

Perschel, A. (2010). Work-life flow: How individuals, Zappos, and other innovative companies achieve high engagement. *Global Business and Organizational Excellence*, *29*, 17–30. https://doi.org/10.1002/joe.20335

Przybilla, L., Wiesche, M., & Krcmar, H. (2019). Emergent leadership in agile teams – An initial exploration. *SIGMIS-CPR - Proceedings of the 2019 Computers and People Research Conference*, 176–179. https://doi.org/10.1145/3322385.3322423

Pugliese, F., Acerbi, A., & Marocco, D. (2015). Emergence of leadership in a group of autonomous robots. *PLoS ONE*, *10*(9). https://doi.org/10.1371/journal.pone.0137234

Reichard, R. J., Riggio, R. E., Guerin, D. W., Oliver, P. H., Gottfried, A. W., & Gottfried, A. E. (2011). A longitudinal analysis of relationships between adolescent personality and intelligence with adult leader emergence and transformational leadership. *Leadership Quarterly*, *22*(3), 471–481. https://doi.org/10.1016/j.leaqua.2011.04.005

Rennie, S. M., Prieur, L., & Platt, M. (2023). Communication style drives emergent leadership attribution in virtual teams. *Frontiers in Psychology*, *14*, 1095131. https://doi.org/10.3389/fpsyg.2023.1095131

Robbins, S. P. (2003). *Organizational Behavior*. Prentice-Hall.

Rubin, R. S., Bartels, L. K., & Bommer, W. H. (2002). Are leaders smarter or do they just seem that way? Exploring perceived intellectual competence and leadership emergence. *Social Behavior and Personality: an international journal*, *30*(2), 105–118. https://doi.org/10.2224/sbp.2002.30.2.105

Sanchez-Cortes, D., Aran, O., Schmid Mast, M., & Gatica-Perez, D. (2010). Identifying Emergent Leadership in Small Groups Using Nonverbal Communicative Cues. *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*, 1–6. https://doi.org/10.1145/1891903.1891953

Sanchez-Cortes, D., Motlicek, P., & Gatica-Perez, D. (2012). Assessing the Impact of Language Style on Emergent Leadership Perception from Ubiquitous Audio. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM)*, 1–4. https://doi.org/10.1145/2406367.2406408

Sashkin, M. (1976). Changing Toward Participative Management Approaches: A Model and Method. *Academy of Management Review*, *1*(3), 75–86. https://doi.org/10.5465/amr.1976.4400820

Schaumberg, R. L., & Flynn, F. J. (2012). Uneasy Lies the Head That Wears the Crown: The Link Between Guilt Proneness and Leadership. *Journal of Personality and Social Psychology*, *103*(2), 327–342. https://doi.org/10.1037/a0028127

Schein, E. H. (1992). *Organizational culture and leadership: A dynamic view*. Jossey-Bass.

Schlamp, S., Gerpott, F. H., & Voelpel, S. C. (2021). Same talk, different reaction? Communication, emergent leadership and gender. *Journal of Managerial Psychology*, *36*(1), 51–74. https://doi.org/10.1108/JMP-01-2019-0062

Schneider, B. (1975). Organizational Climate: Individual Preferences and Organizational Realities Revisited. *Journal of Applied Psychology*, *60*(4), 459–465. https://doi.org/10.1037/h0076919

Schneier, C. E., & Goktepe, J. R. (1983). Issues in emergent leadership: The contingency model of leadership, leader sex, leader behavior. In H. Blumberg, A. Hare, V. Kent, & M. Davies (Eds.), *Small groups and social interaction* (pp. 413–421). John Wiley & Sons.

Schwartz, S. H. (1994). Beyond individualism/collectivism: New cultural dimensions of values. In U. Kim, H. Triandis, C. Kagitcibasi, S. Choi, & G. Yoon (Eds.), *Individualism and collectivism: Theory, method, and applications* (pp. 85–119). Sage.

Serban, A., Yammarino, F. J., Dionne, S. D., Kahai, S. S., Hao, C., McHugh, K. A., Sotak, K. L., Mushore, A. B., Friedrich, T. L., & Peterson, D. R. (2015). Leadership emergence in face-to-face and virtual teams: A multi-level model with agent-based simulations, quasi-experimental and experimental tests. *Leadership Quarterly*, *26*(3), 402–418. https://doi.org/10.1016/j.leaqua.2015.02.006

Sharfman, M. P., & Dean, J. W. (1991). Conceptualizing and Measuring the Organizational Environment: A Multidimensional Approach. *Journal of Management*, *17*(4), 681–700. https://doi.org/10.1177/014920639101700403

Smith, P., Haslam, S. A., & Nielsen, J. F. (2018). In Search of Identity Leadership: An ethnographic study of emergent influence in an interorganizational R&D team. *Organization Studies*, *39*(10), 1425–1447. https://doi.org/10.1177/0170840617727781

Smithikrai, C. (2008). Moderating effect of situational strength on the relationship between personality traits and counterproductive work behaviour: Effect of situational strength. *Asian Journal of Social Psychology*, *11*(4), 253–263. https://doi.org/10.1111/j.1467-839X.2008.00265.x

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Soluk, J., & Kammerlander, N. (2021). Digital transformation in family-owned Mittelstand firms: A dynamic capabilities perspective. *European Journal of Information Systems*, *30*(6), 676–711. https://doi.org/10.1080/0960085X.2020.1857666

Spisak, B. R., O'Brien, M. J., Nicholson, N., & van Vugt, M. (2015). Niche construction and the evolution of leadership. *Academy of Management Review*, *40*, 291–306.

Steers, R. M., Sanchez-Runde, C., & Nardon, L. (2012). Leadership in a Global Context: New Directions in Research and Theory Development. *Journal of World Business*, *47*(4), 479–482. https://doi.org/10.1016/j.jwb.2012.01.001

Stewart, G. L., Courtright, S. H., & Manz, C. C. (2011). Self-Leadership: A Multilevel Review. *Journal of Management*, *37*(1), 185–222. https://doi.org/10.1177/0149206310383911

Sueur, C. (2011). Group Decision-Making in Chacma Baboons: Leadership, Order and Communication During Movement. *BMC Ecology*, *11*, 26. https://doi.org/10.1186/1472-6785-11-26

Summerfield, M. R. (2014). Leadership: A simple definition. *American Journal of Health-System Pharmacy*, *71*(3), 251–253. https://doi.org/10.2146/ajhp130435

Tabassum, M., Raziq, M. M., & Sarwar, N. (2023). Toward an overarching multi-level conceptualization of emergent leadership: Perspectives from social identity, and implicit leadership theories. *Human Resource Management Review*, *33*(2). https://doi.org/10.1016/j.hrmr.2022.100951

Taggar, S., Hackett, R., & Saha, S. (1999). Leadership emergence in autonomous work teams: Antecedents and outcomes. *Personnel Psychology*, *52*(4), 899–926. https://doi.org/10.1111/j.1744-6570.1999.tb00184.x

Taylor, A. C. (2009). Sustainable urban water management: understanding and fostering champions of change. *Water Science and Technology*, *59*(5), 883–891. https://doi.org/10.2166/wst.2009.033

Tett, R. P., & Guterman, H. A. (2000). Situation Trait Relevance, Trait Expression, and Cross-Situational Consistency: Testing a Principle of Trait Activation. *Journal of Research in Personality*, *34*(4), 397–423.

Thomas, D. C. (2008). *Cross-Cultural Management: Essential Concepts*. SAGE.

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. https://doi.org/10.1177/1534484305278283

Trompenaars, F., & Hampden-Turner, C. (2004). *Riding the Waves of Culture: Understanding Cultural Diversity in Business* (2nd). Nicholas Brealey.

Truninger, M., Ruderman, M. N., Clerkin, C., Fernandez, K. C., & Cancro, D. (2021). Sounds like a leader: An ascription–actuality approach to examining leader emergence and effectiveness. *Leadership Quarterly*, *32*(5), 101420. https://doi.org/10.1016/j.leaqua.2020.101420

Turner, S. F., Cardinal, L. B., & Burton, R. M. (2017). Research Design for Mixed Methods: A Triangulation-Based Framework and Roadmap. *Organizational Research Methods*, *20*(2), 243–267. https://doi.org/10.1177/1094428115610808

van Knippenberg, D. (2017). Team Leadership. In *The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes* (pp. 345–368). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118909997.ch15

Van Zyl, J., & Hofmeyr, K. (2021). Leadership behaviour that facilitates shared leadership emergence in internationally dispersed nonformal teams. *South African Journal of Business Management*, *52*(1). https://doi.org/10.4102/sajbm.v52i1.2695

Virtaharju, J. J., & Liiri, T. P. (2019). The supervisors who became leaders: Leadership emergence via changing organizational practices. *Leadership*, *15*(1), 103–122. https://doi.org/10.1177/1742715017736004

Walter, F., Cole, M. S., van der Vegt, G. S., Rubin, R. S., & Bommer, W. H. (2012). Emotion Recognition and Emergent Leadership: Unraveling Mediating Mechanisms and Boundary Conditions. *Leadership Quarterly*, *23*(5), 977–991. https://doi.org/10.1016/j.leaqua.2012.06.007

Wang, C., Chen, X., Xie, G., & Cao, M. (2017). Emergence of Leadership in a Robotic Fish Group Under Diverging Individual Personality Traits. *Royal Society Open Science*, *4*(5), 161015. https://doi.org/10.1098/rsos.161015

Wang, D., Waldman, D. A., & Zhang, Z. (2014). A Meta-Analysis of Shared Leadership and Team Effectiveness. *Journal of Applied Psychology*, *99*(2), 181–198. https://doi.org/10.1037/a0034531

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii. https://www.jstor.org/stable/4132319

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science*, *16*(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Wellman, N. (2017). Authority or community? A relational models theory of group-level leadership emergence. *Academy of Management Review*, *42*(4), 596–617. https://doi.org/10.5465/amr.2015.0375

Wellman, N., Newton, D. W., Wang, D., Wei, W., Waldman, D. A., & LePine, J. A. (2019). Meeting the Need or Falling in Line? The Effect of Laissez-Faire Formal Leaders on Informal Leadership. *Personnel Psychology*, *72*(3), 337–359. https://doi.org/10.1111/peps.12308

Westfall, C. (2019). Leadership development is a $366 billion industry: Here's why most programs don't work.

Wickham, K. R., & Walther, J. B. (2009). Perceived behaviors of emergent and assigned leaders in virtual groups. *Virtual Team Leadership and Collab. Eng. Advancements: Contemp. Issues and Implications*, 50–64. https://doi.org/10.4018/978-1-60566-110-0.ch005

Wilson, J. M., Fletcher, T. D., Pescosolido, T., & Major, D. A. (2021). Extraversion and Leadership Emergence: Differences in Virtual and Face-to-Face Teams. *Small Group Research*, *52*(5), 535–564. https://doi.org/10.1177/1046496420986620

Wolff, S. B., Pescosolido, A. T., & Druskat, V. U. (2002). Emotional intelligence as the basis of leadership emergence in self-managing teams. *Leadership Quarterly*, *13*(5), 505–522. https://doi.org/10.1016/S1048-9843(02)00141-8

Wolfram Cox, J., Madison, K., & Eva, N. (2022). Revisiting emergence in emergent leadership: An integrative, multi-perspective review. *Leadership Quarterly*, *33*(1). https://doi.org/10.1016/j.leaqua.2021.101579

Wu, C., Yao, H., Ning, X., & Wang, L. (2021). Emergence of Informal Safety Leadership: A Social–Cognitive Process for Accident Prevention. *Production and Operations Management*, *30*(11), 4288–4305. https://doi.org/10.1111/poms.13523

Yoo, S., Joo, B.-K., & Noh, J.-H. (2022). Team emergent states and team effectiveness: the roles of inclusive leadership and knowledge sharing. *Journal of Organizational Effectiveness*, *9*(3), 353–371. https://doi.org/10.1108/JOEPP-05-2021-0120

Zaccaro, S. J., Foti, R. J., & Kenny, D. A. (1991). Self-Monitoring and Trait-Based Variance in Leadership: An Investigation of Leader Flexibility Across Multiple Group Situations. *Journal of Applied Psychology*, *76*(2), 308–315. https://doi.org/10.1037/0021-9010.76.2.308

Zander, L., Mockaitis, A. I., & Butler, C. L. (2012). Leading global teams. *Journal of World Business*, *47*(4), 592–603.

Zhang, Z., Waldman, D. A., & Wang, Z. (2012). A multilevel investigation of leader–member exchange, informal leader emergence, and individual and team performance. *Personnel Psychology*, *65*(1), 49–78.

Ziek, P., & Smulowitz, S. (2014). The impact of emergent virtual leadership competencies on team effectiveness. *Leadership and Organization Development Journal*, *35*(2), 106–120. https://doi.org/10.1108/LODJ-03-2012-0043

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Analyzing the Retail Gasoline Market in Germany: Impact of Spatial Competition and Market Concentration on Prices

Nicolas Fiedler

Ludwig Maximilian University of Munich

Abstract

Given the changing landscape of fuel retailing, this study explores the impact of spatial competition and market concentration on diesel prices in Germany. The question of how population density and gas station density, i.e. the equilibrium pattern of locations of firms, are related is examined. In addition, the impact of gas station density, as a spatial measure of competition, and market concentration on diesel prices is investigated. Based on theory, population density should have a positive impact on gas station density. Gas station density should have a negative and market concentration a positive influence on the diesel price in a district. Using 2022 data on German gas stations and diesel prices, a positive effect of population density and on gas station density, a negative effect of gas station density on diesel price, and a positive effect of market concentration on diesel price were each found at the district level. The effects of gas station density and market concentration, however, were relatively small. The results show that fuel prices at gas stations are influenced by spatial competition and market concentration.

Keywords: diesel prices; gas station density; market concentration; retail gasoline market; spatial competition

1. Introduction

In 2022, fuel prices in Germany have come very much into focus. For the reason that it was the most expensive year ever for drivers in Germany to fill up their fuel tanks. A liter of diesel was priced at an average of 1.95 euros, and a liter of Super E10 at 1.86 euros over the year 2022. Gasoline was thus around 27 cents more expensive than in the previous record year of 2012, and the price per liter of diesel was even almost 47 cents higher than in 2012 (Prack, 2023). The mobility landscape is of great importance to society and economy. Mobility is not only a tool to fulfill individual needs, but also an important driver for economic activities and trade. In this context, fuel prices play a critical role, influencing to a large extent individual mobility costs and overall costs for the transportation sector. It is important for all population groups and economic sectors to purchase fuels efficiently and at reasonable prices. For this reason, it makes sense to examine in detail the factors that influence these prices and the underlying market structure.

Fuels can be considered a fairly homogeneous product due to their chemical composition. Nevertheless, gasoline prices at individual gas stations vary widely. Many papers explain the price differences by the intensity of local competition. This competition is determined in the retail gasoline market not only by the number of competitors, but also by the geographic distribution of gasoline stations in the area. Consumers tend to buy their fuel at gas stations close to their homes, as greater distance to gas stations results in higher transportation costs. It follows that competition in the retail gasoline market is highly localized (Bergantino et al., 2020).

The objective of this paper is to investigate the impact of spatial competition and market concentration on prices by looking at the retail gasoline market in Germany. The first question that arises is what affects the equilibrium pattern of company locations? And the second question is, what are the characteristics of equilibrium prices if there is spatial competition among firms? Clemenz and Gugler (2006) have also already studied these two questions in their work. It has already been observed that population density has a major impact on the density of gas stations (Clemenz & Gugler, 2006; Götz & Gugler, 2006). Based on this and the model of Salop

DOI: https://doi.org/10.5282/jums/v10i2pp424-440

(1979), the first hypothesis was developed that retail stores are more likely to be located more densely in areas with higher population density. Several researchers such as Barron et al. (2004), Cardoso et al. (2020), Clemenz and Gugler (2006), and Meerbeeck (2003) have looked at the relationship between spatial competition and prices at gas stations. Most of them found that higher spatial competition leads to lower prices. An important factor is the distance between gas stations. This leads to the second hypothesis that, under spatial competition, prices tend to be lower as the density of seller locations increases. The relationship between market concentration and prices has also been studied a few times, for example by Eckert and West (2005), Kihm et al. (2016), and Sen (2003). Their findings indicate that higher prices are related to higher market concentration. This also forms the third and last hypothesis of this paper, that prices tend to be higher, as market concentration increases.

To test these hypotheses, a dataset of German gas stations and their diesel prices from the year 2022 was used. The political districts in Germany were defined as local markets, following Clemenz and Gugler (2006). Based on this, average values such as diesel price, gas station density, one and four firm concentration ratio in 2022 were calculated for each district. All hypotheses were tested using multiple OLS regressions with different combinations of control variables. In the first step, a significant positive effect of population density on gas station density was found. In the second step, a significant negative effect of gas station density on diesel price was found. A robustness test was performed by removing 20% of the districts with the highest population density from the dataset, in order to exclude mainly metropolitan areas from the analysis. The effect was still negative but no longer significant. Lastly, a significant positive effect of market concentration on diesel price was found. This was also the case in the robustness test. But it should be noted that the effects of gas station density and market concentration are very small.

This paper has the following outline. Section 2 presents the theoretical background by means of relevant literature and description of important concepts and leads argumentatively to the hypotheses. Section 3 provides an overview of the German retail gasoline market. Section 4 describes the methodology used and section 5 presents the empirical results. Section 6 concludes the study with a discussion and comparison of the results with previous literature, limitations, theoretical and managerial implications.

2. Theory and Hypotheses

2.1. Gas Station Density

In this paragraph, following Clemenz and Gugler (2006), the aspects of the circular city model of Salop (1979) that are relevant to this research are described. An important feature of spatial competition is that consumers shop at the store where their total costs are lowest. These consist of the price (multiplied by the quantity) and any transport costs a customer must face. This results in each store having a "local monopoly", the geographical extent of which depends on the prices of the nearest competitors and the respective transport costs of each customer to the various stores. These transportation costs are generally substantially influenced by the distance between stores, and to some extent by road quality and the availability of public transportation, etc. The price a store can charge, increases with the distance to the nearest competitor as well as with the transport costs of a customer. Demand in such a local monopoly depends not only on the geographic size of the market, but also on the number of customers in the territory. In other words, the population density. A company wants to be at a location with many potential customers but few competitors, as the high demand is distributed among relatively few companies. The potentially high demand is distributed among relatively few companies, thus the expected profit is higher. In more densely populated areas, the gap between competitors can be smaller, as demand per square kilometer is usually much higher.

This relationship has already been observed in the retail gasoline market. Götz and Gugler (2006) looked at the relationship between population density and gas station density in Austria. Using an Ordinary Least Squares (OLS) regression, they found that if the population density in a market increases by one percentage point, then the density of gas stations increases on average by 0.835 percent. Clemenz and Gugler (2006) have defined the 121 political districts in Austria as local markets. Looking at the same relationship, they ran multiple regressions, each with different control variables, such as different variables for market concentration or number of motor vehicles per capita. Their results showed that in each of their models, the coefficient of population density is positive and significant at the 5% level or lower. Their coefficients range from 0.810 to 0.873, depending on the control variable. This means that if population station density increases by one percentage point, the gas station density increases by at least 0.8% on average. They also performed a robustness test by changing the market definition and defining each municipality as a local market. Their results have shown to be robust to the change in market definition.

This suggests that, on average, the higher the population density in a local market, the higher the gas station density. Based on this literature and the previous explanations, the following first hypothesis was developed:

> Hypothesis 1: Retail stores are more likely to be located more densely in areas with higher population density.

2.2. Diesel Price and Gas Station Density

Higher gas station density, in turn, leads to increased spatial competition. This raises the question of the extent to which this competition influences the prices at gas stations. From a theoretical standpoint, pricing differences can be explained by product differentiation.

In order to explain the concept of product differentiation, this paragraph refers to the book "The Theory of Industrial

Organization" by Tirole (1988). Two products are almost never perfect substitutes, in the sense that all customers are indifferent if the two goods have the same price. Products are almost always differentiated by a certain attribute. A good can be characterized as a bundle of different attributes: quality, availability, location, the information of the consumer about its availability and quality, etc. Each consumer has a ranking for the different characteristics. Researchers usually focus on a small part of these characteristics and a special description of the respective customer preferences. There are two cases that are used commonly. The first case is vertical differentiation. In a vertically differentiated product environment, all consumers agree on the preferred mix of features. In more general terms, they also agree on the order of preference. The most typical example of this is quality. The majority agree that a higher quality is to be favored. Given equal prices, there is a clear natural ordering over the characteristics space. For example, given the same price, most people prefer a powerful laptop over a less powerful laptop, or an Audi over a Toyota. The second case is horizontal differentiation. For some attributes, given equal prices, the optimal decision depends on the individual consumer. Preferences and tastes are very different. A common example is color, one person prefers blue, the other one red. Another important point in this context is the location. To exaggerate, it is likely that Munich residents prefer products that are available in Munich to physically identical products that are only available in New York. Broken down, customers prefer stores that are close to them. One model of horizontal differentiation is Salop's circle model, which was described at the beginning of the chapter.

In his work, Meerbeeck (2003) applied the concept of product differentiation to the gasoline market. He states that in the case of gas stations, both horizontal and vertical differentiation can be observed. Although gasoline can theoretically be considered a homogeneous commodity in terms of its physical and chemical characteristics, in practice product specifications vary due to the additives used by each brand. Although these factors could play a role, he found that location has the greatest impact on price differences between gas stations. Consumers generally prefer gas stations in their "neighborhood." Lee (2007) examined the nature of competition in San Diego's retail gasoline market using two years of panel data on weekly gas station prices. He was able to find that retail prices are strongly influenced by gas station characteristics such as brand name and amenities. But he also found that the relative geographic proximity of competing gas stations is an important factor in explaining price differences among gas stations. Gas stations compete most with gas stations that are less than 1 mile away. The intensity of competition continues to decrease with distance.

From these papers, it can be inferred that spatial competition is an important factor in the formation of prices in the gasoline market. Several researchers have already studied how this spatial competition affects equilibrium prices at gas stations. Often, local gas station density is defined using a fixed and arbitrary radius. Hosken et al. (2008) determined the number of gas stations in a 1.5-mile radius and the distance to the nearest gas station as the competitive measure. They looked at how this local competition affects prices at individual gas stations. However, no consistent relationship was found between local competition and the prices or margins of the gas stations. Barron et al. (2004) also analyzed the effect of the number of gas stations in a 1.5-mile radius on price. They observed that a 50% increase in the number of gas stations in this radius leads to a decrease in the price of about 0.5%.

Other studies have defined political areas as local markets. Meerbeeck (2003) has defined the gas stations density in a municipality as the local competition measure for the Belgian retail gasoline market. A higher number of gas stations leads to more intense competition. The results show that the number of local competitors has a negative effect on prices. This indicates that as the number of competing stations in a local market increases, diesel prices decrease. However, the magnitude of this effect is very small, but significant.

Clemenz and Gugler (2006) have defined the 121 political districts in Austria as local markets. They analyzed the influence of gas station density on the margin at gas stations. Several equations were analyzed with different control variables, such as market concentration or ALPS. ALPS describes the proportion of mountains and forest in each county. This is intended to serve as an additional proxy for the differences in transportation costs between the respective districts. Their results showed that in each of their models, the coefficient of gas station density is negative and significant at the 5% level or lower. Their coefficients range from -0.035 to -0.045. This means that if gas station density increases by one percentage point, the margin decreases by around 0.04% on average. They also performed a robustness test by changing the market definition and defining each municipality as a local market. Their results have shown to be robust to the change in market definition. This suggests that the closer competitors are to each other on average, the lower the margin.

Cardoso et al. (2020) have measured the effect on prices in the Brazilian gasoline market when a new company enters a local market. As a result of the increased spatial competition, prices decrease. The closer the location of the new competitor is to existing gas stations, the more prices are reduced.

In summary, most papers have shown that prices at gas stations are affected by spatial competition. It turns out that the distance between gas stations is a key factor. Gas station density reflects this average distance as an inverse proxy and thus can be used as a measure of spatial competition. Based on this and the model of Salop (1979), the second hypothesis was developed:

> Hypothesis 2 - Under spatial competition, prices tend to be lower as the density of seller locations increases.

As Clemenz and Gugler (2006) have already pointed out, hypothesis 2 has an obvious consequence for hypothesis 1: In the presence of spatial competition, the increase in store density has to be less than proportional to the increase in population density, since a higher station density lowers the equilibrium price.

2.3. Diesel Price and Market Concentration

So far, only gas stations have been considered individually and independently of the operating company. In reality, however, not every company operates just one gas station in each market, but in a number of cases several. For this reason, some companies could have more power than others in certain markets. This in turn could have an impact on prices.

In order to introduce the subject of the market power and oligopolies, this paragraph refers to the book "Industrial Organization: Markets and Strategies" by Belleflamme and Peitz (2015). Companies are expected to maximize profits, but the market conditions limit their ability to exploit consumers. If the paradigm of perfect competition is to be believed, firms will in the end sell at a price that equals their marginal cost. Firms do not have market power if compared to the industry they are small and price takers. But what happens when some companies have a large market share in the respective industry or local market? They cannot be described as price takers. But neither can they be described as pure price makers, compared to monopolists, as they still compete with other companies. Although these large companies unquestionably exercise market power, their smaller competitors do as well. Market power is the ability to set prices above the competitive level and thus generate profit. Industries in which a few companies compete with each other and thus market power is collectively shared are called oligopolies. Most industries are oligopolies, and so is the retail gasoline market. The characteristic feature of oligopolistic competition is that companies cannot be ignoring the behavior of their rivals. Companies' profits in the end depend on the combination of decisions taken by all companies in the market. In making decisions, firms must take into account the likely behavior of their competitors and respond to their own decisions. The analysis of this strategic interaction goes back as far as the nineteenth century. Augustin Cournot and Joseph Bertrand are considered the founding fathers of oligopoly theory. They each developed their own models for analyzing strategic interaction in such markets. It is beyond the scope of this paper to discuss them.

According to Cotterill (1986), there are several oligopoly models that predict that the price level in a market is positively associated with one or more measures of seller concentration, such as the four-firm ratio or one firm ratio.

Several researchers have already studied how market concentration affects equilibrium prices. Cotterill (1986) looked at the relationship between market concentration and prices using supermarkets in local markets in Vermont. He found that prices are significantly higher in more concentrated markets. Keeler et al. (1999) looked at this effect in for-profit and non-profit hospitals. Regarding market definition, they have assumed that the market for each hospital corresponds to the boundaries of the district in which the hospital is located. They were able to find that hospital prices were higher in markets with higher concentration. This was even the case for non-profit hospitals. Asplund and Sandin (1999) used 486 driving schools and their prices to analyze competition in 235 local markets in Sweden. The results showed that prices are increasing in firm concentration within a market. Newmark (2004) also presents several studies in his work that also support this relationship. He lists, for example, the papers of Connor (1990), Cotterill (1990), and Koller and Weiss (1989) and several more.

This relationship has also been studied by several researchers in the retail gasoline market. Sen (2003) investigated this effect for the Canadian retail gasoline market. To do this, he analyzed monthly averages in 10 major Canadian cities over a seven-year period. He found that in addition to higher average monthly wholesale prices, increasing local market concentration is also positively and significantly associated with higher retail prices. Eckert and West (2005) studied gas stations in Vancouver and also found that higher market concentration leads to higher prices. Kihm et al. (2016) found in the German gasoline market that higher market concentration, in the form of the Herfindahl-Hirschman index, leads to significantly higher prices within a 5 kilometer radius of the gas station. Clemenz and Gugler (2006) analyzed the impact of market concentration on prices at the district and municipality level in their paper. No significant relationship was found at the district level, but at the municipality level. Bergantino et al. (2018) have looked at the municipalities in Rome. They discovered that the higher the market concentration in a given municipality, the higher the price of both gasoline and diesel at gas stations. Hosken et al. (2008), on the other hand, did not identify a significant relationship. A large part of the literature, both in the retail gasoline market and in other industries, describes a positive relationship between market concentration and prices. Based on this, the third hypothesis was developed:

> Hypothesis 3 - Prices tend to be higher, as market concentration increases.

In general, it can be said that these mechanisms have been little studied in this form for the German retail gasoline market. In addition, many researchers have only looked at selected local markets in the respective countries, and very rarely have the entire country or all gas stations been included in the analysis. My work aims to fill these gaps in the literature.

3. The German Retail Gasoline Market

Oil is still the most important source of energy in Germany, accounting for 33% of total energy supply in 2018. German oil demand has fallen much more slowly than domestic oil production over the past decade, so Germany remains heavily dependent on oil imports. Germany has a high dependency on oil imports of about 97%. Oil import dependency is calculated as domestic oil production divided by total oil demand. Germany's top crude oil suppliers in 2018 were Russia (36%), Norway (12%), United Kingdom (8%), Kazakhstan (8%), Libya (8%), and Nigeria (6%). For oil products, Germany produces most of its demand in its domestic refineries. This covered 88% of total demand in 2018. Accounting for more than half of total oil usage, the transportation sector is the largest oil consumer. Diesel is the most consumed oil product, followed by gasoline (IEA, 2020). Germany has one of the largest refining capacities in Europe, with a total of 13 refineries spread across the entire country (Mineralölwirtschaftsverband, 2020). A map with the oil infrastructure of Germany is shown in Figure 1.

In this map are shown the locations of the 13 refineries, the oil pipelines and oil storage sites to get a rough idea of how the oil infrastructure looks like. In the retail fuel sector, there were 14460 active gas stations in Germany in 2022 (Statista, 2023a). Figure 2 shows a map of Germany with the individual gas stations to get an impression of how the gas stations are distributed in Germany.

In Germany, five companies together own 68% of the market share of total fuel sales in 2022. These five companies are Aral (21%), Shell (20%), Jet (10%), Total (9.5%) and Esso (7%). The market share of gas station operators by number of gas stations shows a quite similar situation. These five companies operate about half of all gas stations in Germany (Statista, 2023b).

The year 2022 was shaped by very high fuel prices in Germany. For drivers in Germany, it was the most expensive year ever to fill up their fuel tanks. A liter of diesel costs an average of 1.95 euros, and a liter of Super E10 1.86 euros. Gasoline was thus around 27 cents more expensive than in 2012, the previous record year, and the price per liter of diesel was even almost 47 cents higher than in 2012 (Prack, 2023). Figure 3 shows an overview of how the average daily fuel prices have developed in 2022.

The course of fuel prices this year has been shaped by a few striking factors. An extremely large increase in the prices of all types of fuel is seen at the end of February. On 23 February, the price of diesel was still at 1.67 euros per liter and has risen within about two and a half weeks by 38% percent to 2.31 euros per liter. This price shock was triggered by the invasion of Russian troops into Ukraine on February 24, 2022. As already described above, Russia is Germany's largest crude oil supplier.

The next sharp change in the trend took place on the first of June. Within one day, the E5 price has fallen by about 28 cents to 1.93 euro per liter. This was triggered by the so-called "Tankrabatt". In English, this means a government-funded fuel discount. Energy tax rates were reduced for a limited period of three months from June 1, 2022, for some types of fuel to the extent permitted under European law. For gasoline it was reduced by 29.55 cents per liter and for diesel by 14.04 cents per liter. As a consumption tax, the energy tax is intended to be passed on in full to the end consumer. The purpose of the "Tankrabatt" was to relieve people who depend on the car, such as commuters, families, as well as business people, especially in the crafts and logistics industries (Bundesregierung, 2022). On August 31, this benefit expired, which resulted in a huge increase in prices of all types of fuel on the first of September. The price of diesel has risen by about 10 cents, E5 by about 24 cents and E10 by about 23 cents per liter from one day to the next.

The question arises how the fuel prices for one liter are composed. Bft (2023) has illustrated this on the basis of one liter of diesel. A retail price of 1,762 per liter of diesel is used. Of this amount, 0.2813 euro is VAT, resulting in a net selling price of 1.4807 euro. 0.4704 euro is due to energy tax, 0.0950 euro is due to CO_2 pricing, 0.0030 euro is due to petroleum stockpiling levy and the value of goods excluding taxes (price of exploration, crude oil, processing and transportation) equals 0.9123 euro. This results in a total amount of statutory levies of 0.8497 euros per liter. This shows that the statutory levies form a large part of the fuel price.

4. Methodology

To test the three hypotheses, the following data and methodology is used. The methodology is strongly inspired by the paper "Locational choice and price competition: some empirical results for the austrian retail gasoline market" by Clemenz and Gugler (2006). I have a dataset of the German gas stations with the following information. For each gas station a separate ID, the company and information about the location, such as zip code, street, and coordinates. In addition, for each gas station for the most part hourly diesel prices for the complete year 2022. The data originates from the website Tankerkönig and was retrieved using an API. Based on this, first a daily and then an annual average diesel price was calculated for each gas station. Additional location information, such as the district key and federal state key, was added to each gas station based on the zip code using the OpenPLZ API. In Germany, each municipality can be precisely identified by means of the regional key. According to the Federal Statistical Office (2023c), the definition of the regional key is: "12-digit key for unambiguous identification of a municipality with the components: Federal state (2 digits), administrative district (1 digit), district (2 digits), association of municipalities (4 digits) and municipality (3 digits)." After data cleaning, i.e. removing incorrect information, duplicates and missing values, a dataset with 14097 gas stations remained. So, about 97.5% of the German gas stations were considered in the analysis. This is a very high percentage of a country's total gas stations included in the analysis, compared to many other studies. Some other studies use only a portion of a country's total gas stations such as for example Lee (2007) and Sen (2003) or Eckert and West (2005).

It is a relatively difficult thing to divide Germany into appropriate local gasoline markets. The market definition should not be too narrow and not too broad. Following Clemenz and Gugler (2006), each district was defined as a local market. If the respective markets were measured inaccurately, it is possible that the estimates underestimate the true relationship between the dependent and independent

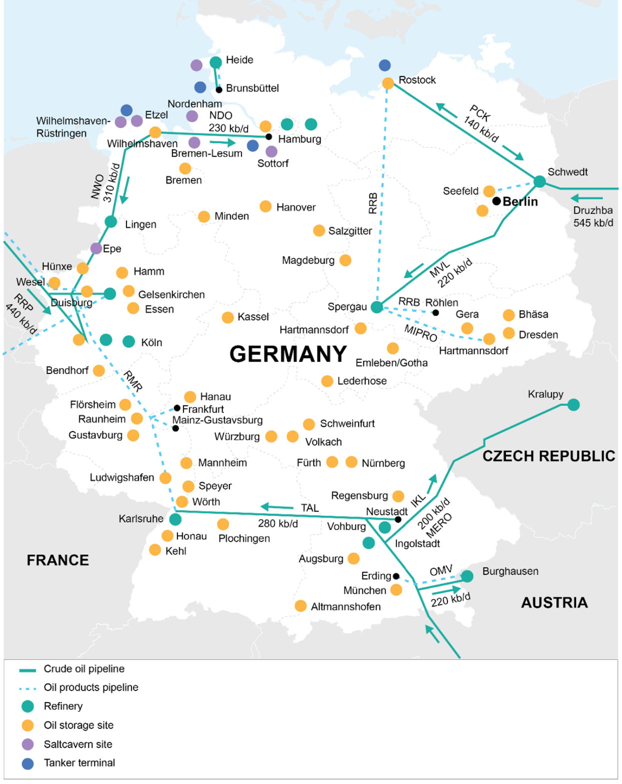


Figure 1: Map of Germany's oil infrastructure (Source: IEA (2020))

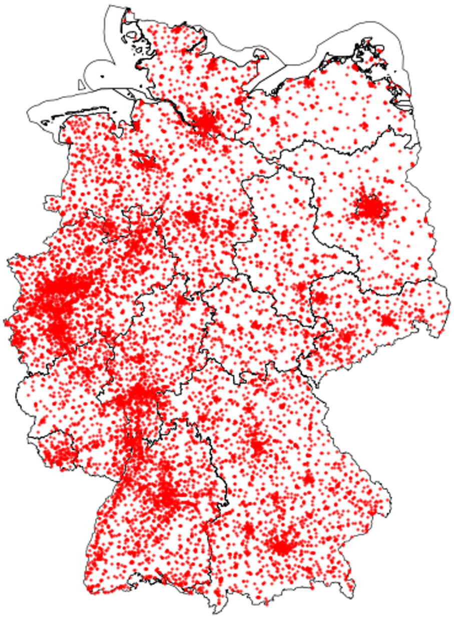


Figure 2: Map of the individual gas stations in Germany

variables. Unless these imprecisions are correlated with our relevant variables, increased white noise affecting statistical significance is the most likely consequence. In their paper, the two authors conducted a robustness test by defining municipalities in Austria as local markets. Their results are robust to the change in market definition. For this reason, it is assumed that for Germany, the districts also represent appropriate local gasoline markets. Based on this, the data was grouped by district and the corresponding values were calculated at the district level. The average diesel price ($diesel_k$) and the number of gas stations (N_k) for each district for the year 2022 were calculated. In addition, two measures of

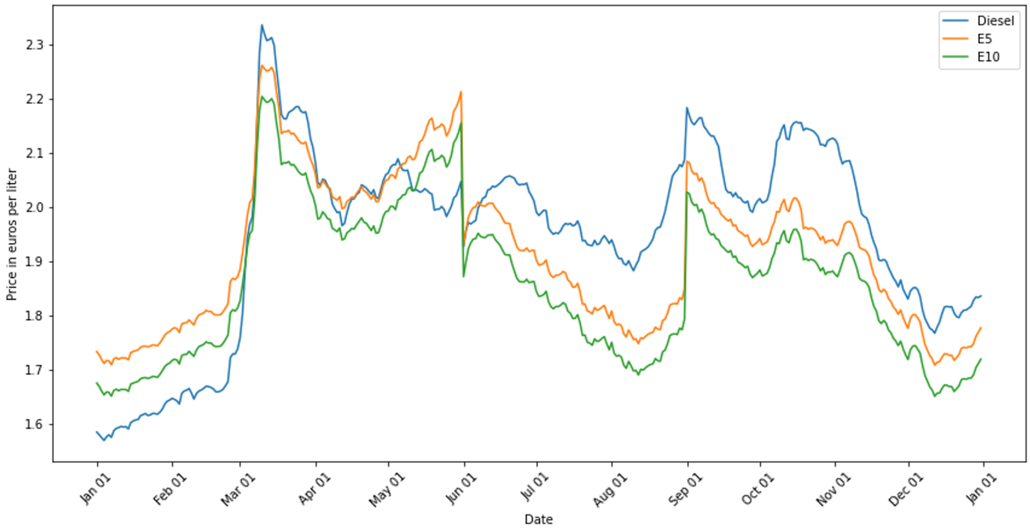


Figure 3: Daily average prices for diesel, E5 and E10 in Germany in 2022

market concentration were calculated for each district. First, the market share of the largest firm in the respective county $(C1_k)$. This was calculated as follows:

$$C1_k = \frac{N_{1,k}}{N_k}$$

$N_{1,k}$ is the number of gas stations of the largest company in district k. The largest company is called the company with the most gas stations in district k. N_k is the total number of gas stations in district k. In addition, the market share of the four largest companies combined $(C4_k)$ was calculated as follows:

$$C4_k = \frac{\sum_{n=1}^{4} N_{n,k}}{N_k}$$

$N_{n,k}$ is the number of gas stations operated by the n largest firms in district k. N_k is as above, the total number of gas stations in district k.

The area (A_k) (Bundesamt, 2023b), number of inhabitants (Pop_k) (Bundesamt, 2023a) and number of motor vehicles (VT_k) (Kraftfahrt-Bundesamt, 2023) were also added for each district. Based on this data, the gas station density (SD_k) was calculated as follows:

$$SD_k = \frac{N_k}{A_k}$$

The population density (PD_k) :

$$PD_k = \frac{Pop_k}{A_k}$$

The motor-vehicle density (VD_k) :

$$VD_k = \frac{VT_k}{A_k}$$

In addition, the number of motor vehicles per head (V_k). According to Clemenz and Gugler (2006), gas station density is a suitable (inverse) approximation of the average distance between gas stations. But only with the condition that gas stations do not cluster in certain areas. Their research suggests that a clustering of gas stations does not really occur often on average. Therefore, it can be assumed that gas station density is a suitable distance measure and can be used as a variable for spatial competition.

All the important variables are listed in the table 1 for sake of better overview.

4.1. Gas station density

To test the first hypothesis, several ordinary least squares regressions were performed. The superordinate equation, in analogy to (Clemenz & Gugler, 2006), looks as follows:

$$lnSD_k = \beta_0 + \beta_1 DEMAND_k + \beta_2 C_k + \varepsilon_k$$

k = 1, ..., 396 represents the districts in Germany, $lnSD_k$ is the natural logarithm of the density of gas stations per square kilometer of the respective district. $DEMAND_k = \{lnPD_k, lnVD_k\}$, with the natural logarithm respectively of population density, motor vehicle density, and motor vehicles per head. There is no sales data or the like available to reflect demand for diesel. But gasoline demand is not really price sensitive (Brons et al., 2008; Espey, 1998; Hanly et al., 2002). For this reason, the variable population density, motor vehicle density and motor vehicles per capita represent the different demand for diesel in the individual districts quite well.

Table 1: Variable Definition

Variable	Definition
Pop_k	Number of inhabitants in district k
A_k	Area of district k in square kilometers
N_k	Number of gasoline stations in district k
VT_k	Number of motor-vehicles in district k
avg_diesel_k	Retail price charged for diesel per liter averaged over the year 2022 averaged over all stations within district k in cent
$SD_k = N_k/A_k$	Density of gasoline stations in district k. (Inverse) proxy for distances between stations
$PD_k = Pop_k/A_k$	Population density in district k
$SP_k = Pop_k/N_k$	Inhabitants per gasoline station in district k
$C1_k$	Market share of the largest firm in district k defined as $C1_k = \frac{N_{1,k}}{N_k}$, where $N_{1,k}$ is the number of gasoline stations operated by the largest (most gas stations in district) firm in district k
$C4_k$	Sum of market shares of the largest four firms in district k defined as $C4_k = \frac{\sum_{n=1}^{4} N_{n,k}}{N_k}$, where $N_{n,k}$ is the number of gasoline stations operated by the n largest firm in district k
V_k	Degree of motorization defined as the number of motor-operated vehicles per head in district k
$VD_k = VT_k/A_k$	Motor-vehicle density in district k

The control variable in this model is market concentration. The natural logarithm of the market share of the largest ($lnC1_k$) or the largest four firms ($lnC4_k$) in district k. ε_k represents the error term. Here is an overview of the exact equations that were analyzed:

$$lnSD_k = \beta_0 + \beta_1 lnPD_k + \varepsilon_k \tag{1}$$

This equation was used to measure purely the effect of population density on gas station density, also it serves as a basis to compare the results to the equations with control variables.

$$lnSD_k = \beta_0 + \beta_1 lnVD_k + \varepsilon_k \tag{2}$$

In eq. 2, motor vehicle density serves as a proxy for demand. This variable could be a better proxy for demand than population density, since often in cities, i.e. very densely populated areas, there are relatively fewer people who own a motor vehicle. This would make the resulting demand less. For reasons of comparability with the results of Clemenz and Gugler (2006), this variable was not used in the other equations.

$$lnSD_k = \beta_0 + \beta_1 lnPD_k + \beta_2 lnC1_k + \varepsilon_k \tag{3}$$

$$lnSD_k = \beta_0 + \beta_1 lnPD_k + \beta_2 lnC4_k + \varepsilon_k \tag{4}$$

In eq. 3 and eq 4, market concentration was included as a control variable. In the retail gasoline market, it cannot be assumed that each company operates only one station. There is yet no clear answer as to how market concentration affects the number of gas stations in a local market. However, it is fairly safe to say that the number of stores a pure monopolist will set up in the absence of entry threat is the lower bound. On the other hand, the upper limit on the number of stores is given by the number of locations a monopolist will establish if the market is free to enter. Moreover, there is no clear indication of the relationship between market concentration and gas station density (Clemenz & Gugler, 2006). For this reason, the two market concentration measures were included in the analysis because they may alter the effect of population density.

$$lnSD_k = \beta_0 + \beta_1 lnPD_k + \beta_2 lnV_k + \varepsilon_k \tag{5}$$

In reality, the choice of a company's location is influenced by many other factors than the number of potential customers alone. The demand per individual potential customer is also a factor. For this reason, the number of cars per capita is included in equation 5 to get a better proxy for demand.

$$lnSD_k = \beta_0 + \beta_1 lnPD_k + \varepsilon_k \tag{6}$$

Equation 6 is, obviously, the same as Equation 1, but the data set that was analyzed is different. Population density varies considerably between districts, as urban areas are also considered. It could be that entry decisions of companies in cities are influenced by different factors than in rural areas, for example, availability of space, higher set up costs or higher rents, etc. For this reason, the robustness of the results is tested by removing the 20% of districts with the highest population density from the original dataset. This primarily removes cities from the analysis. Clemens & Gugler removed only the districts of Vienna from their analysis, leaving aside the fact that there are other large cities in Austria that could influence the effect. By removing the districts with the highest population density, the difference between urban and rural areas is much better controlled.

Several more equations were tested with different combinations of the control variables, but it was found that listed equations best represented the differences. OLS regression was performed for all equations.

4.2. Diesel price

To test the second and third hypothesis, several ordinary least squares regressions were performed. The superordinate equation, in analogy to Clemenz and Gugler (2006), looks as follows:

$$\begin{aligned} diesel_k = &\beta_0 + \beta_1 lnSD_k + \beta_2 lnC_k \\ &+ federalState\ fe + \ \varepsilon_k \end{aligned}$$

$k = 1, ..., 396$ again represents the districts in Germany, $diesel_k$ is the average diesel price over the year 2022 for district k. $lnSD_k$ is the natural logarithm of the density of gas stations per square kilometer of the respective district. This variable serves as an inverse proxy for the average distance between gas stations. A higher value of SD is indicative of higher local competition. Another independent variable in this model is market concentration. The logarithm of the market share of the largest ($lnC1_k$) or the largest four firms ($lnC4_k$) in district k. In this model, fixed effects were controlled for by federal S state. The spread of the diesel price is up to 5 cents per liter from state to state. These fixed effects explain a fair amount of the variation in diesel price between districts, which is why they were included in the analysis. ε_k represents the error term.

Here is an overview of the exact equations that were analyzed:

$$\begin{aligned} diesel_k = &\beta_0 + \beta_1 lnSD_k + \beta_2 lnC1_k \\ &+ federalState\ fe + \varepsilon_k \end{aligned} \tag{7}$$

$$\begin{aligned} diesel_k = &\beta_0 + \beta_1 lnSD_k + \beta_2 lnC4_k \\ &+ federalState\ fe + \varepsilon_k \end{aligned} \tag{8}$$

Equations 7 and 8 each measure the effect of gas station density and the two market concentration measures on the average price of diesel. They form the basis for testing hypotheses two and three.

$$\begin{aligned} diesel_k = &\beta_0 + \beta_1 \widehat{lnSD_k} + \beta_2 lnC4_k \\ &+ federalState\ fe + \varepsilon_k \end{aligned} \tag{9}$$

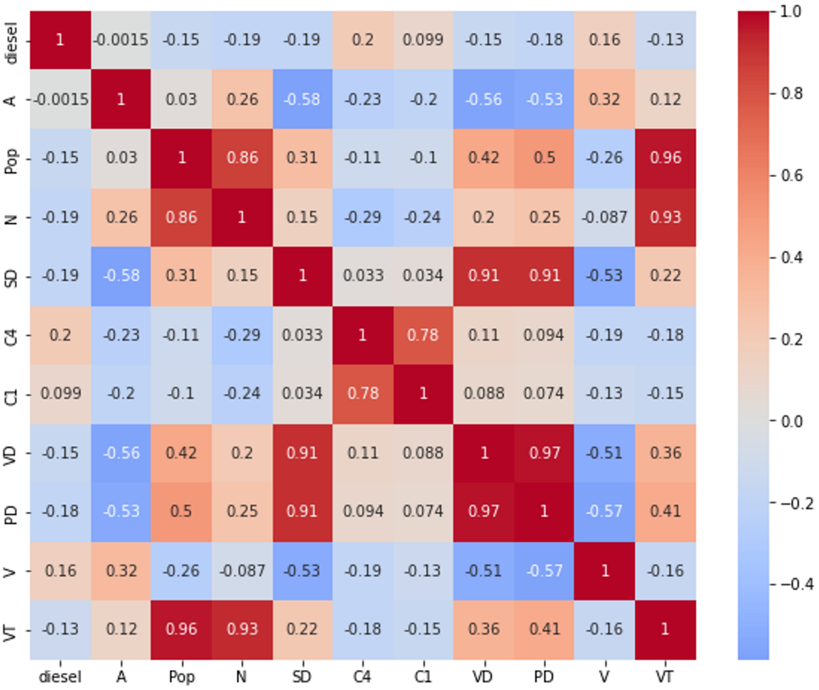
It might be a problem to estimate the diesel price by the gas station density and to ensure the direction of causality. It could be that the diesel price and the gas station density influence each other. If the diesel price is higher in certain regions due to other unidentifiable factors, it would make sense for companies to open additional gas stations in this region. This would increase the density of gas stations induced by the higher diesel price. This interaction could lead to biased results and might skew the true effect of gas station density on the price of diesel. To solve this potential endogeneity problem, a two-stage least squares (2SLS) approach is used. In the first stage, the density of gas stations is estimated by using the instrument, population density. This instrument variable is used to isolate the effect of population density on gas station density. In the second stage, the estimated lnSD is added to equation 9 to estimate the effect on the diesel price. In short, this approach instrumentalizes lnSD by lnPD. This seems like an ideal instrument since population density is exogenous to diesel prices and determines almost completely station density.

$$\begin{aligned} diesel_k = &\beta_0 + \beta_1 lnSD_k + \beta_2 lnC4_k \\ &+ federalState\ fe + \varepsilon_k \end{aligned} \tag{10}$$

Equation 10 is, obviously, the same as Equation 8, but the data set that was analyzed is different. The price of diesel in urban areas could still be influenced by additional factors, such as generally higher price levels or higher gas station operating costs due to higher rents, for example. For this reason, the same data set was used as in Equation 6, i.e., excluding the 20% of districts with the highest population density. Except for equation 9, as described above, an OLS regression was performed for all equations. Each of equations seven to ten was performed again with the logarithm of the diesel price as the dependent variable. This was done for the reason that one can later interpret the percentage change in diesel price. In addition, for a better comparability with other studies already listed in the theory part.

5. Results

To get a first feeling for the data it makes sense to look at the correlation matrix in Figure 4 with all the relevant variables. It is interesting to see that gas station density (SD_k) is very strongly positively correlated with population density (PD_k) and motor vehicle density (VD_k), the Pearson correlation coefficient being 0.91 in each case. On the other hand, however, gas station density is moderately negatively correlated with motor vehicles per capita (V_k) ($r = -0.53$). The average diesel price is only weakly correlated with the independent variables, positively as well as negatively. The highest correlation in terms of absolute value is with the combined market share of the four largest companies ($\mathrm{C4}_k$) with a coefficient of 0.2. Gas station density (SD_k) is weakly negatively correlated with average diesel price ($r = -0.19$). It is also still interesting to note that population density (PD_k) and motor vehicle density (VD_k) are almost completely positively correlated ($r = 0.97$).



	diesel	A	Pop	N	SD	C4	C1	VD	PD	V	VT
diesel	1	-0.0015	-0.15	-0.19	-0.19	0.2	0.099	-0.15	-0.18	0.16	-0.13
A	-0.0015	1	0.03	0.26	-0.58	-0.23	-0.2	-0.56	-0.53	0.32	0.12
Pop	-0.15	0.03	1	0.86	0.31	-0.11	-0.1	0.42	0.5	-0.26	0.96
N	-0.19	0.26	0.86	1	0.15	-0.29	-0.24	0.2	0.25	-0.087	0.93
SD	-0.19	-0.58	0.31	0.15	1	0.033	0.034	0.91	0.91	-0.53	0.22
C4	0.2	-0.23	-0.11	-0.29	0.033	1	0.78	0.11	0.094	-0.19	-0.18
C1	0.099	-0.2	-0.1	-0.24	0.034	0.78	1	0.088	0.074	-0.13	-0.15
VD	-0.15	-0.56	0.42	0.2	0.91	0.11	0.088	1	0.97	-0.51	0.36
PD	-0.18	-0.53	0.5	0.25	0.91	0.094	0.074	0.97	1	-0.57	0.41
V	0.16	0.32	-0.26	-0.087	-0.53	-0.19	-0.13	-0.51	-0.57	1	-0.16
VT	-0.13	0.12	0.96	0.93	0.22	-0.18	-0.15	0.36	0.41	-0.16	1

Figure 4: Correlation Matrix

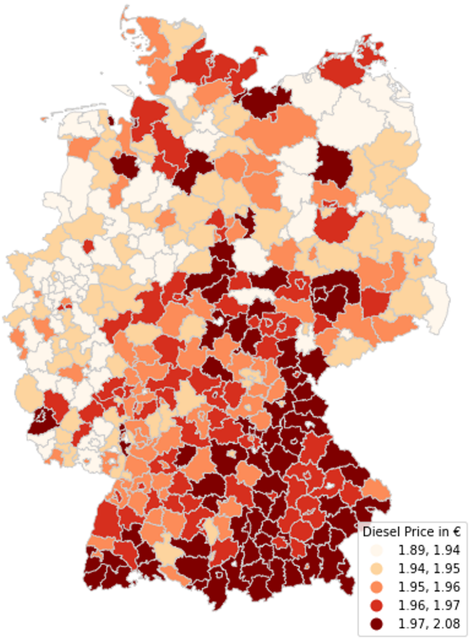


Figure 5: Average Diesel Prices on district level

To get a first impression of the average diesel prices it makes sense to look at the diesel prices by district on a map. In the annual average prices from 2022 there is a relatively large difference between the cheapest and most expensive district. In 2022, diesel was cheapest in the district of Mülheim an der Ruhr at 1.89 euros per liter, and most expensive in Trier-Saarburg at 2.08 euros per liter. That is a difference of 19 cents, or about 10%. On average across all districts, the diesel price in 2022 was 1.96 with a standard deviation of 0.024 euro per liter. The Bavarians have refueled, on average, for 1.98 euro per liter, thus approx. 5 cents more expensively than the citizens of Berlin, who had on average the most favorable prices. It can be seen very clearly that there are substantial price differences between the individual districts and even federal states.

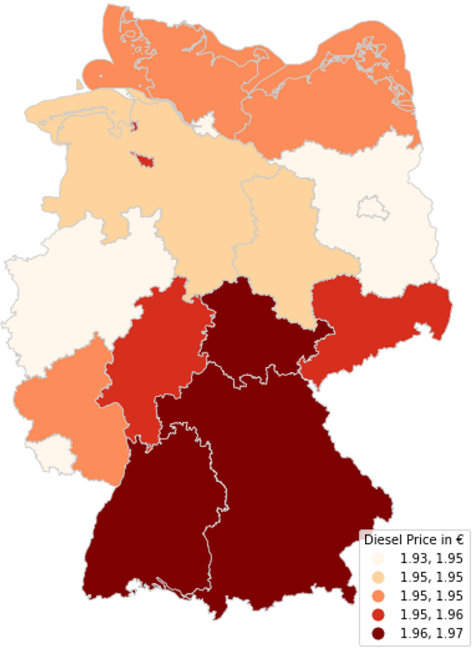


Figure 6: Average Diesel Prices on federal State

The average district in Germany has 209748.823 inhabitants with an area of 902.241 square kilometers. The largest district is Mecklenburgische Seenplatte with 5495.590 square kilometers, the smallest is Schweinfurt City with 35.7 square kilometers. The average population density is 533.264 inhabitants per square kilometer. The least densely populated district is Prignitz in Brandenburg with 35.339 inhabitants per square kilometer. The opposite is Munich City with 4788.246 inhabitants per square kilometer. In the average district, each inhabitant owns 0.758 motor vehicles. In Hohenlohekreis in Baden Württemberg, there are almost as many motor vehicles as inhabitants.

A German district has an average of 41.477 gas stations. On average, one gas station covers 25.005 square kilometers or 4991.13 inhabitants. The highest density of gas stations is in Schweinfurt City, the lowest in Uckermark in Brandenburg. The firm with the most gas stations operates an average of 21.9%, and the largest four firms together operate 58.8% of the gas stations in a district. Table 2 contains the descriptive statistics for all variables.

5.1. Hypothesis 1 – Population Density and Gas Station Density

In the scatterplot in Figure 7, there is a clear relationship between population density and gas station density. This trend shows that the higher the population density in a given district, the higher the density of gas stations. To test the hypothesis that retail stores are more likely to be located more densely in areas with higher population density, we look at the OLS regression results in table 3. The coefficients from the logarithm of population density are positive and significant at the 1% level for each equation in the model. The coefficient of 0.8212 ($p < 0.001$) shows that for every percentage point that the population density in a district increases, the gas station density increases about 0.82%. Models with control variables such as market concentration or number of motor vehicles per capita also support this trend, with coefficients for population density ranging from 0.8233 ($p < 0.001$) to 0.8289 ($p < 0.001$).

If the 20% of districts with the highest population density, i.e. primarily cities, are left out of the analysis, the effect becomes even slightly higher. The largest effect is seen for motor vehicle density, 0.9157 ($p < 0.001$). This means that if the density of motor vehicles increases by one percent, the density of gas stations increases by about 0.92%. The effect of the two market concentration measures is negative and significant. If the combined market share of the four largest firms increases by one percent, then gas station density decreases by about 0.42%.

These results support our hypothesis that retail stores are more likely to be located more densely in areas with higher population density.

5.2. Hypothesis 2 – Gas Station Density and Diesel Price

The scatterplot in Figure 8 indicates a relationship between the density of gas stations and the average price of diesel. This trend, that under spatial competition prices are lower the higher the seller density, is also supported by the results of the OLS regressions in table 4 and 5. For the equations with the absolute diesel price as dependent variable across all districts, the coefficients of ln SD range from -0.3268 ($p < 0.01$) to -0.3871 ($p < 0.01$). For the logarithm of the diesel price, the coefficients of ln SD range from -0.0020 ($p < 0.01$) to -0.0017 ($p < 0.01$). In other words, if the den-

Table 2: Descriptive Statistics

Variables	Count	Mean	Std. Deviation	Minimum	Maximum
Pop_k (inhabitants)	396	209748.823	246188.715	34091.000	3.677×10^6
A_k (in km^2)	396	902.241	722.673	35.700	5495.590
N_k (stations)	396	41.477	29.101	4.000	303.000
$diesel_k$ (in Cent)	396	195.554	2.376	188.877	207.774
$1/SD_k$ (km^2/station)	396	25.005	20.986	1.623	181.002
SP_k (inhabitants/station)	396	4991.13	1851.41	2298.2	18287.5
PD_k (inhabitants/km^2)	396	533.264	711.151	35.339	4788.246
VD_k (motor vehicles/km^2)	396	353.170	413.843	28.515	2857.350
$C1_k$ (in %)	396	0.219	0.064	0.105	0.500
$C4_k$ (in %)	396	0.588	0.109	0.344	1.000
V_k (motor vehicles/head)	396	0.758	0.139	0.001	0.985

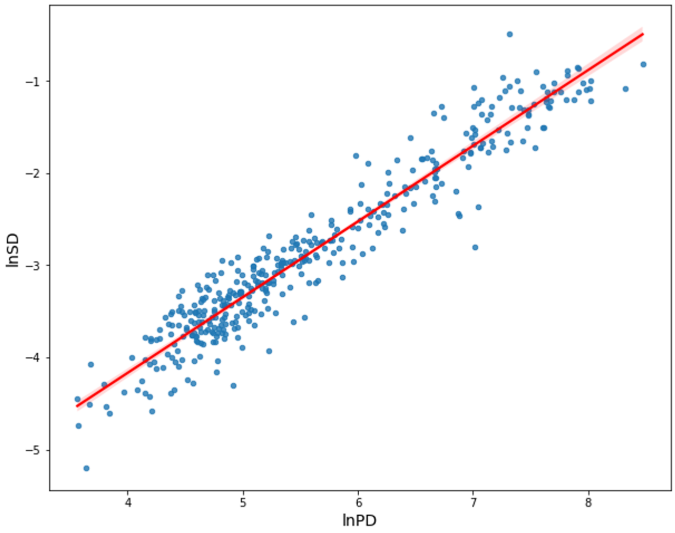


Figure 7: Scatterplot - gas station density and population density

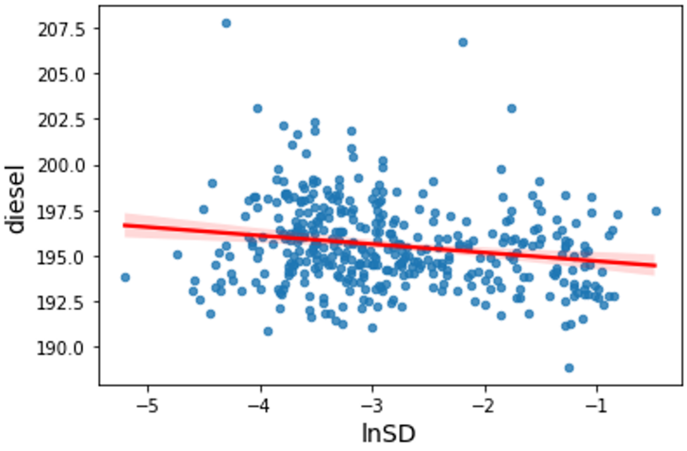


Figure 8: Scatterplot - average diesel prices and gas station density

sity of gas stations increases by one percent point, then the price of diesel falls by about 0.003 cents or 0.002%.

If the 20% of districts with the highest population density, i.e. mainly cities, are excluded from the analysis, the effect is still negative but no longer significant (p = 0.656 and p = 0.674).

Table 3: Regression Results gas station density

Dependent variable: ln SD						
Sample:	All districts				Without highest 20% PD	
Equation	1		2		3	
Independent variables	Coef	p-value	Coef	p-value	Coef	p-value
lnPD	0.8212	0.000			0.8254	0.000
lnVD			0.9157	0.000		
lnC1					-0.1698	0.002
lnC4						
lnV						
Intercept	-7.4542	0.000	-7.7333	0.000	-7.7420	0.000
Adjusted R2	0.921		0.922		0.923	
No Obs	396		396		396	
Equation	4		5		6	
Independent variables	Coef	p-value	Coef	p-value	Coef	p-value
lnPD	0.8289	0.000	0.8233	0.000	0.8325	0.000
lnVD						
lnC1						
lnC4	-0.4126	0.000				
lnV			0.0275	0.010		
Intercept	-7.7229	0.000	-7.4560	0.000	-7.5128	0.000
Adjusted R2	0.926		0.921		0.848	
No Obs	396		396		316	

Note: Standard Errors are heteroscedasticity robust (HC0)

Table 4: Regression Results absolute diesel price

Dependent variable: $diesel_k$								
Sample:	All districts						Without highest 20% PD	
Equation	7		8		9		10	
Method	OLS		OLS		2SLS		OLS	
Independent variables	Coef	p-value	Coef	p-value	Coef	p-value	Coef	p-value
$\ln SD_k$	-0.3406	0.009	-0.3871	0.003	-0.3268	0.006	-0.1139	0.656
$\ln C1_k$	1.0854	0.001						
$\ln C4_k$			2.8911	0.000	2.8563	0.000	3.0641	0.000
Intercept	196.0629	0.000	195.7961	0.000	195.9434	0.000	196.4656	0.000
Adjusted R^2	0.319		0.348		0.342		0.369	
federalState fixed effects	yes		yes		yes		yes	
No. Obs	396		396		396		316	

Note: Standard Errors are heteroscedasticity robust (HC0)

5.3. Hypothesis 3 – Market Concentration and Diesel Price

The scatterplot in Figure 9 indicates the tendency that the higher the combined market share of the four largest companies, the higher the price of diesel. This trend, that the higher the market concentration the higher the prices, is also supported by the OLS regressions. The coefficients of the two market concentration measurements are positive and significant across all equations, supporting the third hypothesis. If the market share of the largest company in a district increases by one percentage point, then the price of diesel increases by about 1 cent on average. Looking at the effect of the aggregate market share of the four largest firms in a district, the coefficients for absolute diesel prices are 2.8911 ($p < 0.001$) and 2.8563 ($p < 0.001$). This effect increases to 3.0641 ($p < 0.001$) when the 20% districts with the highest popula-

Table 5: Regression Results logarithm diesel price

Dependent variable: $\ln diesel_k$								
Sample:	All districts						Without highest 20% PD	
Equation	7		8		9		10	
Method	OLS		OLS		2SLS		OLS	
Independent variables	Coef	p-value	Coef	p-value	Coef	p-value	Coef	p-value
$\ln SD_k$	-0.0017	0.009	-0.0020	0.003	-0.0017	0.006	-0.0005	0.674
$\ln C1_k$	0.0055	0.001						
$\ln C4_k$			0.0146	0.000	0.0144	0.000	0.0155	0.000
Intercept	5.2783	0.000	5.2770	0.000	5.2777	0.000	5.2805	0.000
Adjusted R^2	0.322		0.350		0.344		0.371	
federalState fixed effects	yes		yes		yes		yes	
No. Obs	396		396		396		316	

Note: Standard Errors are heteroscedasticity robust (HC0)

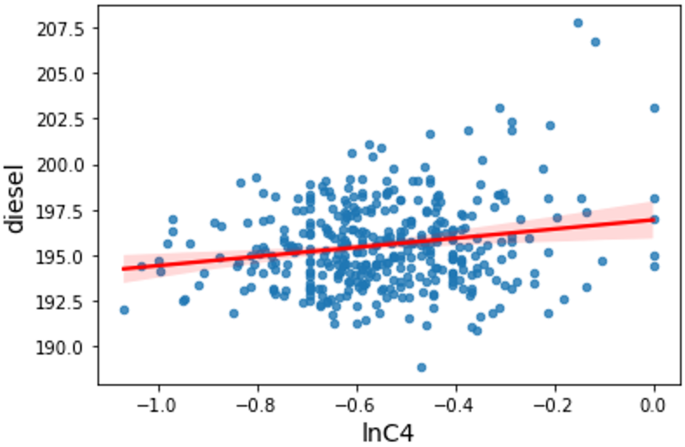


Figure 9: Scatterplot - average diesel prices and four firm ratio

tion density are excluded from the analysis. Thus, it could be concluded that the effect of market concentration is more pronounced for rural areas and has more influence on the diesel price there.

These results support our hypothesis that prices tend to be higher, as market concentration increases.

6. Discussion

The results shown above strongly support the first hypothesis that retail stores are more likely to be located more densely in areas with higher population density. This finding suggests that more densely populated areas also have more gas stations. One possible explanation could be that a larger number of potential customers in a market increases the profitability of gas stations. Areas with high population density may have higher demand for fuels. In these areas, consumers could benefit from higher availability of gas stations. Thus, potential customers can choose between a larger number of filling stations and potentially benefit from lower fuel prices. The relationship between population density and gas station density could also provide insights into the structure of the gas station industry and its ability to adapt. High population densities may be an indicator of economic activity and mobility. This, in turn, perhaps influences the development of gas stations.

As already shown in the results, the coefficient of lnD, in the equation without control variables is 0.8212 ($p < 0.000$). In the paper from Clemenz and Gugler (2006) this coefficient was 0.810. You can see that the results are pretty much the same. The coefficients of the other equations with control variables from the two scientists closely resemble my coefficients for population density. Götz and Gugler (2006) also found a very similar relationship. This is consistent with the predictions of several spatial competition models that the relationship between retail stores and population density is not perfectly proportional since a higher density of stores causes the equilibrium price to decrease.

The observation that the effect of motor vehicle density on gas station density is higher compared to the effect of population density shows the relevance of automobility. A higher number of motor vehicles per square kilometer could potentially lead to greater use of existing gas stations, thereby increasing the demand for new gas stations in that particular area. This is also shown by the coefficient of the logarithm of motor vehicles per capita in Equation 5. If motor vehicles per capita increase by ten percent in a district, then gas station density increases by approximately 0.275%. Including variables related to motor vehicles in the analysis makes sense to possibly mitigate the effect of cities to some extent. In densely populated urban areas, public transport is often better developed, which could lead to fewer people owning a car. In turn, this could have an effect on gas station density. When the 20% of districts with the highest population density, which are mostly cities, are removed from the analysis, the effect of the population density increases slightly.

Although market concentration was only a control variable in the test of the first hypothesis, it is interesting to look at this effect as well. Market concentration has a negative impact on gas station density, so the higher the market concentration, the lower the average gas station density. This suggests that markets with few dominant players tend to have fewer gas stations. One reason could be that the competitive barriers that new players have to overcome in a highly concentrated market mean that fewer new gas stations are opened. Finding out the exact causes, however, is beyond the scope of this paper and could be the subject of further research. In a concentrated market, potential customers could face more limited options to choose from. On the one hand, due to generally fewer gas stations, and on the other hand, the large companies operate several gas stations, which makes the number of gas stations operated by various other companies smaller. The results suggest that market concentration plays a role in shaping the structure of the gas station market. In terms of competition and regulatory policy, market concentration should be closely monitored further to ensure that it does not affect competitiveness and consumer choice.

Other studies have already found a somewhat similar effect. As already shown in the results section, the coefficient of the one firm ratio was -0.1698 ($p < 0.01$), while for Clemenz and Gugler (2006) it was -0.132. The coefficients of the four firm ratio are also within the same range. It can be predicted that if the market share of the largest company in the county increases by one percentage point, then the gas station density decreases by about 0.15% on average. As you can see, there is a rather small, but nevertheless significant effect. The results of Götz and Gugler (2006) also give support to this effect.

The results shown above support the second hypothesis that under spatial competition, prices tend to be lower as the density of seller locations increases. In concrete terms, it can be observed that a one percentage point increase in the density of gas stations is associated, on average, with a decrease of about 0.002% in the diesel price in a district. One cause may be the increased spatial competition that comes with an increasing number of gas stations in a geographically defined market. If the density of gas stations increases, then logically the average distance between the gas stations becomes smaller. As already explained in the theory section, a smaller distance between competing gas stations leads to more intense competition. This more intense competition may lead gas station operators to reduce prices in order to attract potential customers and thus secure or increase their market share.

Another possible cause could be the supply and demand relationship. A larger number of gas stations in a market could increase the supply of fuel. If the demand for fuel in a market is relatively constant, an increased supply could lead to a decrease in prices.

In their paper, Clemenz and Gugler (2006) looked at the impact of gas station density on the margin at gas stations at the district level. But there are drawbacks to calculating the margin in their study. They make some generalized assumptions for calculating the costs. For example, they make blanket assumptions for transportation costs to and within Austria and do not distinguish between gas stations or districts. They thus calculate a single numerical value for the costs with which they then calculate the margin for all gas stations. I chose the diesel price rather than the margin as the dependent variable because this way the results are not distorted by possible inaccurate assumptions. Even if it makes no sense to compare the absolute values, it is at least possible to compare the direction of the effect. The two researchers also found a significant negative effect of gas station density. They also performed a robustness test by excluding Vienna districts from the analysis. Their results are robust to the change in the analyzed data. I also performed a robustness test as described earlier by excluding the 20% of districts with the highest population density. The effect is still negative but no longer significant ($p = 0.656$). The reduced dataset has an average gas station density of 0.045 gas stations per square kilometer, or inversely 30.218 square kilometers per gas station. The excluded data, on the other hand, has an average gas station density of 0.267 gas stations per square kilometer or, inversely 4.410 square kilometers per gas station. You can see that there is a very big difference between the average density of gas stations. This substantial difference could be the reason why the results in the robustness test are no longer significant. The relatively high density of gas stations in the excluded districts shows a high degree of spatial competition. Lee (2007) was able to discover that gas stations compete most when they are less than a mile apart and the intensity of competition continues to decrease with distance. Cardoso et al. (2020) found that the closer the competitor's location is to existing gas stations, the more prices fall. In rural areas, the average distance between gas stations may be too large for competition to significantly affect prices. This could be the reason why the effect of gas station density on the diesel price is no longer significant in the reduced data set. It may be that the district is chosen too large as the local market in rural areas. Another reason could be that the relationship between gas station density and diesel price in urban areas may be influenced by other factors that are more pronounced in cities. These findings may open up new research perspectives to investigate in more detail the specific influence factors in urban and rural areas and to understand the relationship between gas station density and diesel price more extensively. Kaldor (1935) has found that firms compete not only in a local market but also with firms in neighboring markets. Competition is less intense, but it is still there. This is a factor that was not included in this paper, but could be considered in further research.

The results shown above support the third hypothesis that prices tend to be higher, as market concentration increases. In concrete terms, it can be observed that a one percent increase in the four firm ration is associated, on average, with an increase of about 0.0146% in the diesel price in a district. This effect could be due to various causes based on the market power of the companies. As market concentration increases, so does a company's market power, since there are fewer competitors who could force price cuts. In such a scenario, the different gas stations of the same company could have incentives to keep prices high. Gas stations of the same company will most likely not compete with each other. If one or a few companies have a large market share, then this could lead to less price competition overall. Higher market concentration or market share would not necessarily have to be due to market power, but could also be due to the firms' efficiency.

But Mueller (1986) made clear in his work that if high concentration and high market shares were due to higher efficiency, one would expect lower prices in concentrated markets, which he did not observe. This is consistent with the presented results.

High market concentration could also result in a high entry barrier for new market participants. It may be more difficult for new competitors to establish themselves if a few companies dominate the majority of the market. As already shown in the results, market concentration has a negative impact on gas station density. For this reason, there are, on average, proportionally fewer gas stations in a concentrated market that could be in competition with each other. This would also lead to less spatial competition.

Another reason for the higher prices could be tactical collusion. Pepall et al. (2014) listed several factors in their book that facilitate collusion, including high market concentration, significant entry barriers, and product homogeneity. These are all points that play a role in the German retail gasoline market. Cotterill (1990) has shown that tacit collusion does exist in concentrated markets. Tacit collusion is also often described as price followship. Balto (2001) has shown that the higher the market concentration, the greater the risk that a further increase will lead to higher prices through collusion.

As described in the theory section, some researchers have looked at the relationship between market concentration and prices in the retail gasoline market. Clemenz and Gugler (2006) did not find a significant relationship at the district level. In contrast, I was able to find a significant relationship. In the work of Bergantino et al. (2018), the coefficients of brand market share, three firm ratio, and HHI were also all positive and highly significant in the regressions for gasoline and diesel prices. The direction of the effect of market concentration is also consistent with the findings of many other researchers, such as Eckert and West (2004), Kihm et al. (2016), and Sen (2003).

Due to the lack of sales and revenue data, the market concentration was calculated on the basis of the number of gas stations operated by each company. It is important to note that this method may not capture all the subtle nuances of the market structure. Further research could consider alternative methods of measuring market concentration, if sales data from individual gas stations are available, to validate the results.

The effect of market concentration was robust to the reduction in the size of the data analyzed. This suggests that market concentration also has an impact on prices in less densely populated areas.

It should be noted that a potential endogeneity problem could be in the interaction between market concentration and diesel prices. Singh and Zhu (2008) clarify that the fundamental problem is that "market structures are not randomly assigned". They state that companies take into account demand and cost conditions as well as potential competitors when making market entry decisions. The market structure thus emerges as a consequence of these strategic decisions. They further point out the example that markets with unobserved high costs are likely to have higher prices. On the other hand, these markets are also likely to attract fewer market participants. This may in turn lead to a higher market concentration. The price would thus be influenced to a certain extent by the relatively higher costs in certain regions and not mainly by market concentration. A topic for further research would be to what extent the market structure or market concentration in the German retail gasoline market is endogenously determined by prices.

The findings of this paper not only contribute to the theoretical understanding of the spatial relationships between population density, gas station density, market concentration, and fuel prices, but also offer potential applications for both academic research and the corporate world. For further research, some approaches have already been touched upon. In summary, the findings obtained could serve as a basis to study similar spatial relationships in other industries or geographic regions. Applications in the corporate world are also emerging from the findings of this study. For gas station owners, the results offer valuable insights into location planning. The relationship between gas station density, market concentration and prices could lead companies to adjust their competitive strategies. In markets with high concentration, differentiation strategies or cooperations could be more relevant, while in highly competitive markets, price leadership could play a role. Companies could also make strategic decisions such as acquisitions of gas station locations based on the findings on the relationship between market concentration and prices. As a conclusion for companies, one could draw that on average it makes more sense to take over gas stations from other companies than to build new sites, especially in densely populated areas. This way, the density of gas stations does not increase and does not have a negative effect on the price, but the market concentration becomes higher and has a positive effect on the price.

References

Asplund, M., & Sandin, R. (1999). Competition in interrelated markets: An empirical study. *International Journal of Industrial Organization*, *17*(3), 353–369.

Balto, D. A. (2001). Supermarket merger enforcement. *Journal of Public Policy & Marketing*, *20*(1), 38–50.

Barron, J. M., Taylor, B. A., & Umbeck, J. R. (2004). Number of sellers, average prices, and price dispersion. *International Journal of Industrial Organization*, *22*(8), 1041–1066. https://doi.org/10.1016/j.ijindorg.2004.05.001

Belleflamme, P., & Peitz, M. (2015). *Industrial organization: markets and strategies*. Cambridge University Press.

Bergantino, A. S., Capozza, C., & Intini, M. (2018). Empirical investigation of retail gasoline prices.

Bergantino, A. S., Capozza, C., & Intini, M. (2020). Empirical investigation of retail fuel pricing: The impact of spatial interaction, competition and territorial factors. *Energy Economics*, *90*, 104876. https://doi.org/10.1016/j.eneco.2020.104876

Bft. (2023). Zusammensetzung des Benzin-/Dieselpreises. Retrieved August 21, 2023, from https://www.bft.de/daten-und-fakten/benzinpreis-zusammensetzung

Brons, M., Nijkamp, P., Pels, E., & Rietveld, P. (2008). A meta-analysis of the price elasticity of gasoline demand. A SUR approach. *Energy Economics*, *30*(5), 2105–2122. https://doi.org/10.1016/j.eneco.2007.08.004

Bundesamt. (2023a). Fortschreibung des Bevölkerungsstandes. Retrieved August 21, 2023, from https://www-genesis.destatis.de/genesis/online?operation=statistic%5C&levelindex=0%5C&levelid=1682174311638%5C&code=12411%5C#abreadcrumb

Bundesamt. (2023b). Gebietsfläche: Kreise, Stichtag. Retrieved August 21, 2023, from https://www-genesis.destatis.de/genesis/online%5C#astructure

Bundesamt. (2023c). Regionales - Regionalschlüssel (RS). Retrieved August 12, 2023, from https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Glossar/regionalschluessel.html

Bundesregierung. (2022). Fragen und Antworten zum „Tankrabatt". Retrieved August 21, 2023, from https://www.bundesregierung.de/breg-de/aktuelles/faq-energiesteuersenkung-2049702

Cardoso, L. C. B., Uchôa, F., Huamani, W., & Gomez, R. V. (2020). Price effects of spatial competition in Brazilian gas stations.

Clemenz, G., & Gugler, K. (2006). Locational choice and price competition: some empirical results for the austrian retail gasoline market. *Empirical Economics*, *31*(2), 291–312. https://doi.org/10.1007/s00181-005-0016-7

Connor, J. M. (1990). Empirical challenges in analyzing market performance in the US food system. *American Journal of Agricultural Economics*, *72*(5), 1219–1226.

Cotterill, R. W. (1986). Market power in the retail food industry: Evidence from Vermont. *The Review of Economics and Statistics*, 379–386.

Cotterill, R. W. (1990). Food mergers: Implications for performance and policy. *Review of Industrial Organization*, 189–202.

Eckert, A., & West, D. (2004). A Tale of Two Cities: Price Uniformity and Price Volatility in Gasoline Retailing. *The Annals of Regional Science*, *38*, 25–46. https://doi.org/10.1007/s00168-003-0144-y

Eckert, A., & West, D. S. (2005). Price uniformity and competition in a retail gasoline market. *Journal of Economic Behavior & Organization*, *56*(2), 219–237. https://doi.org/10.1016/j.jebo.2003.09.006

Espey, M. (1998). Gasoline demand revisited: An international meta-analysis of elasticities. *Energy Economics*, *20*(3), 273–295.

Götz, G., & Gugler, K. (2006). Market concentration and product variety under spatial competition: Evidence from retail gasoline. *Journal of Industry, Competition and Trade*, *6*, 225–234.

Hanly, M., Dargay, J., & Goodwin, P. (2002). Review of income and price elasticities in the demand for road traffic. *Department for Transport, London*.

Hosken, D. S., McMillan, R. S., & Taylor, C. T. (2008). Retail gasoline pricing: What do we know? *International Journal of Industrial Organization*, *26*(6), 1425–1436.

IEA. (2020). Germany 2020 - Energy Policy Review. https://www.iea.org/reports/germany-2020

Kaldor, N. (1935). Market imperfection and excess capacity. *Economica*, *2*(5), 33–50.

Keeler, E. B., Melnick, G., & Zwanziger, J. (1999). The changing effects of competition on non-profit and for-profit hospital pricing behavior. *Journal of Health Economics*, *18*(1), 69–86.

Kihm, A., Ritter, N., & Vance, C. (2016). Is the German retail gasoline market competitive? A spatial-temporal analysis using quantile regression. *Land Economics*, *92*(4), 718–736.

Koller, R. H., & Weiss, L. W. (1989). Price levels and seller concentration: The case of Portland Cement. *Concentration and Price*, 17–40.

Kraftfahrt-Bundesamt. (2023). Bestand nach Zulassungsbezirken (FZ 1). Retrieved August 12, 2023, from https://www.kba.de/DE/Statistik/Produktkatalog/produkte/Fahrzeuge/fz1%5C_b%5C_uebersicht.html

Lee, S.-y. (2007). Spatial Competition in the Retail Gasoline Market : An Equilibrium Approach Using SAR Models.

Meerbeeck, W. (2003). Competition and Local Market Conditions on the Belgian Retail Gasoline Market. *De Economist*, *151*, 369–388. https://doi.org/10.1023/B:ECOT.0000006590.66223.9a

Mineralölwirtschaftsverband. (2020). MWV-Jahresbericht 2020. https://en2x.de/service/publikationen/

Mueller, D. C. (1986). *Profits in the long run*. Cambridge University Press.

Newmark, C. M. (2004). Price-concentration studies: There you go again. *Antitrust Policy Issues*, 9–42.

Pepall, L., Richards, D., & Norman, G. (2014). *Industrial organization: Contemporary theory and empirical applications*. John Wiley & Sons.

Prack, N. (2023). Spritpreis-Entwicklung: Benzin- und Dieselpreise seit 1950. Retrieved August 21, 2023, from https://www.adac.de/verkehr/tanken-kraftstoff-antrieb/deutschland/kraftstoffpreisentwicklung/

Salop, S. (1979). Monopolistic Competition With Outside Goods. *Bell Journal of Economics*, *10*, 141–156. https://doi.org/10.2307/3003323

Sen, A. (2003). Higher prices at Canadian gas pumps: International crude oil prices or local market concentration? An empirical investigation. *Energy Economics*, *25*(3), 269–288.

Singh, V., & Zhu, T. (2008). Pricing and market concentration in oligopoly markets. *Marketing Science*, *27*(6), 1020–1035.

Statista. (2023a). Anzahl der Tankstellen in Deutschland nach Tankstellentyp von 1999 bis 2022. Retrieved August 12, 2023, from https://de.statista.com/statistik/daten/studie/72262/umfrage/anzahl-der-tankstellen-in-deutschland-nach-tankstellentyp-zeitreihe/%5C#:~:text=Im%5C%20Jahr%5C%202022%5C%20gab%5C%20es,im%5C%20Vergleich%5C%20zum%5C%20Vorjahr%5C%20nicht.

Statista. (2023b). Statistik-Report zum deutschen Tankstellenmarkt. https://de.statista.com/statistik/studie/id/26070/dokument/tankstellenmarkt-statista-dossier/

Tankerkönig. (2022). Spritpreise in Echtzeit. https://tankerkoenig.de/index.php

Tirole, J. (1988). *The theory of industrial organization* (Vol. 14). MIT Press.

Junior Management Science

www.jums.academy
ISSN: 2942-1861

The Role of Hierarchical Differentiation for the Effectiveness of Soccer Teams

Sebestyén András Huszár

Freie Universität Berlin

Abstract

The impact of hierarchical differentiation on team effectiveness is heavily discussed in scientific research with strong arguments lined up on both the pro and the contra sides. To contribute to this debate, I investigated the relationship between a specific facet of hierarchical differentiation, pay dispersion, and team effectiveness. I collected data from five seasons of Premier League and conducted a regression analysis to study the effect of pay dispersion on team performance, cooperation and aggressivity. The empirical results show that pay dispersion is positively and directly associated with aggressivity, whilst its relation with team performance and cooperation is moderated through the financial might of teams. The significant interaction effect for team performance means that pay dispersion has a significant negative effect for high financial might teams, and a weak positive effect for low financial might teams. For cooperation the interaction shows a significant positive effect for the low financial might teams and a weak negative effect for the high financial might teams. Thus, I conclude that pay dispersion indeed affects team effectiveness, however the economic power standing behind the teams needs to be considered.

Keywords: hierarchical differentiation; pay dispersion; Premier League; sports data; team performance

1. Introduction

After taking the reins of Manchester City in 2016 Pep Guardiola said “what we want is so simple: when the opponent has the ball, take it back as quick as possible. When we have the ball, try to move as quick as possible, to create as much chances as possible. That’s all. And good team spirit.” (Manchester City, 2022). In many ways he epitomized the quintessence of team success. Effectiveness and efficiency in reacting to the ever changing currents of the environment and creating chances of success, through cooperation and coordination, all the while retaining the internal harmony of the team. The simple, or leastwise simple to understand, ends however scarcely imply equally simple means. Neither in soccer, nor in management.

One hotly debated factor greatly impacting team success is the presence of hierarchical structures, and even more so of hierarchical differentiation. A predominant manifestation of hierarchical differentiation is the dispersion of pay, which, for better or worse, has the capacity to greatly disrupt existent hierarchical structures. In the 2021/22 season of Premier League, Cristiano Ronaldo joined his old club Manchester United with a salary that exceeded the salary of the second highest paid player (David De Gea) by 37%. That season Manchester United ranked the lowest it has in the last five seasons, 6th place, and it was the only season among the last five they failed to collect at least 60 points. Similarly, in the 2020/21 season Gareth Bale rejoined Tottenham with a salary exceeding that of the second highest paid players (Harry Kane and Tanguy Ndombele) by 200%, and the season ended with Tottenham being unable to crack the top 6 of Premier League, the only time they failed to do so in the last five seasons. The question arises, are these isolated cases of misfortune, or is there a deeper connection between pay dispersion and team success.

Literature indeed exists investigating the effects of pay dispersion on team effectiveness, yet there are competing theories and findings, with the debate reaching consens seeming quite improbable. Lazear and Rosen (1981) argued

I want to thank Prof. Dr. Stefan Razinskas for supporting me throughout the writing process. His feedback and insights were invaluable for the creation of this thesis.

DOI: https://doi.org/10.5282/jums/v10i2pp441-461

that increased pay dispersion will enhance group performance, through motivating increased competition and individual performance. On the contrary, other theories claim pay dispersion to have an adverse effect on group effectiveness, due to increased disconcert and disrupted cohesiveness (Akerlof & Yellen, 1990; Ramaswamy & Rowthorn, 1991). Examining sports teams, quite suitable to this stream of research due to the apparent ease of quantifying performance, Depken and Lureman (2018) found support for pay dispersion hampering team performance, whereas Torgler and Schmidt (2007) discovered both improving and hindering effects.

What is apparent, that both in the theoretical and in the empirical realm there is much controversy around the impact of hierarchical differentiation and pay dispersion on team effectiveness. And indeed, that is the aim of my present thesis, to contribute to this ongoing debate. My fundamental research question is, how does pay dispersion relate to team effectiveness. Strongest focus is laid upon pay dispersion's influence on the performance of the team, however I strive to shed light on pay dispersion's impact on cooperation and aggressivity as well. To that end I will be analyzing Premier League teams. Furthermore, as both in business organizations and in soccer teams, the financial might standing behind a team is a force of doubtless magnitude, I seek to uncover the interplay between financial might and pay dispersion, i. e. their possible interaction in shaping team effectiveness.

My thesis contributes to research by showing the significant moderating effect of financial might, which alters the influence of pay dispersion on team performance. I found evidence that pay dispersion has opposing effects on the performance of the high and low financial might teams. Pay dispersion negatively impacted the performance of high financial might teams, whereas it appeared to lightly increase it for the low financial might teams. Findings showed similarity for cooperation, yet this time the increasing effect for the low financial might teams being significantly positive, and the decreasing for high financial might teams appearing lightly negative. For aggressivity there was no interaction, yet pay dispersion itself did increase aggressive behavior. The explanation for the contrary effects caused by the interaction might lie in the different challenges teams face and the different perceptions players have of themselves and their teams. The discovery of this interaction effect furthers the understanding of how pay dispersion impacts team effectiveness.

The structure of the thesis is as follows. The next chapter offers a brief overview of relevant literature and offers insight in the usage of sports data in managerial research. The third chapter introduces the dataset and describes the regression analysis. In the fourth chapter the results of my hypothesis testing are presented, moreover the second half of the chapter contains a supplemental analysis, testing for interaction effects and the corresponding results. The last chapter reveals a discussion of the findings, detailing the theoretical implications of my thesis.

2. Theoretical background

As previously phrased in the research question, my research revolves in essence around investigating the impact of a certain type of hierarchical differentiation on the effectiveness of teams. Therefore, in this unfolding section I will delve deeper in the core concepts studied, and also offer a brief overview of the most relevant research papers assessing kin phenomena. Moreover, in this chapter I aim to derive my three hypotheses, that will serve as the foundations of my analysis. At last, I will argue for the noteworthy benefits of relying on sports data in researching business organizational and managerial phenomena.

Hierarchy is a concept as ancient as human history. Born out of what once had been a necessity for survival, it had shaped human societies for ages, and as its very construct is inherently and unswervingly human, it will continue to do so. Be it big-game hunting in the age of spear and bow, or the establishment of strategic alliances to reap the competitive benefits of interorganizational networking, for the ultimate success of human groups coordination and cooperation are of utmost importance (Halevy et al., 2011, p. 33). There ever was a need for structure that puts constraints on the adverse aspects of human nature, and at the same time enables virtues to strive and yield benefits (Halevy et al., 2011). Notwithstanding the doubtless existent possibilities (and examples) of failure, this indeed is a crucial facet of what hierarchy is meant to provide.

Magee and Galinsky define hierarchy as "an implicit or explicit rank order of individuals or groups with respect to a valued social dimension" (2008, p. 354). This definition highlights that the degree of awareness might vary greatly between individuals or groups. Some may not perceive that they are part of the hierarchy, but that does not change their embeddedness. Moreover, it emphasizes that there may exist a vast array of aspects by which hierarchy is determined ("valued social dimension"). There is no singular prime measure of hierarchy, although some dimensions do gain more importance over others in given situations. It is a process of social adoption, where certain dimensions might be selected to form the basis of formal hierarchies, or where certain dimensions might organically emerge as commonly valued and birth informal hierarchies. These processes resulting in the creation of hierarchical forms of social relationships constitute the phenomenon called *hierarchical differentiation* by Magee and Galinsky (2008, p. 354).

Groups and teams do form integral parts of any given organization. They are essential to functioning and they produce outcomes, therefore the desire to enhance their effectiveness is a natural implication of striving for the good of an organization. Moreover, for organizations embedded in any competitive environment, the analysis and improvement of the results produced by teams is essential for survival. The appeal of increasing team effectiveness is thus trivial, yet the means are not quite so. The definition of team effectiveness is neither straightforward, nor unitary. Dimas et al. (2023, p. 3) argue that to measure team effectiveness there cannot be one unanimously accepted criterion found, as team effectiveness inherently has different meanings and implications for different stakeholders (2023, p. 3). For instance, the manager in charge of a work group might have vastly different understanding of team effectiveness as opposed to the employee working in the group or the customer for whom the group is producing. Dimas et al. confirm this assumption by reviewing team effectiveness literature. Team effectiveness is a multidimensional construct of different facets (2023, p. 3).

An approach oftentimes employed in order to assess the economical dimension of team effectiveness is measuring team performance. A definition offered by Devine and Philips (2001, p. 512) explains team performance as "the degree to which a team accomplished its goal or mission". Although not always trivial, possibly a criteria should be chosen that best captures the overall achievements a team has made to accomplish its mission. In my thesis this facet of team effectiveness, team performance, stands in major focus and I aim to find an objective and quantifiable way to measure it.

The literature examining the relationship between hierarchical differentiation and team effectiveness dates back long ago. In 1968 Bridges et al. published a study examining the effects of hierarchical differentiation on the efficiency and productivity of teams, as well as on their risk-taking behavior. They found that hierarchical differentiation had adverse effects on the productivity and efficiency of groups, with the hierarchically undifferentiated groups triumphing in both regards. Moreover, it was showed that hierarchy also hampers risk-taking behavior. It needs to be mentioned however, that Bridges et al. examined groups of quite small size (merely four subjects pro group), and the research revolved mainly around problem solving and idea generation.

Quite contrarily Halevy et al. (2011) theorize a multilayered positive relationship between hierarchical differentiation and organizational success. They argue that hierarchy supports coordination and voluntary cooperation, whilst reducing conflicts. Moreover, that hierarchy incentivizes performance and thus increases motivation, and also constructs a psychologically rewarding environment. Albeit, concluding that the presence of hierarchical differentiation is overwhelmingly beneficial for the performance of an organization, they do identify certain moderating factors. Such as degree of task interdependence, which the higher, the more need it constitutes for hierarchy (for comparison, see the juxtaposition of basketball and baseball in Keidel (1987)), legitimacy of the hierarchical rank order, and the (mis)alignment of bases of hierarchy (as power, status, prestige, etc.)

Ronay et al. (2012) also investigated the effects of hierarchical differentiation on group productivity. The experiments they conducted showed that indeed, as Halevy et al. (2011) theorized, groups with considerable hierarchical differentiation outperformed undifferentiated groups when it came to tasks of high procedural interdependency. And at the same time they found no effect of hierarchical differentiation on tasks procedurally independent. Ronay et al. noted that their research focused mainly on hierarchical differentiation founded in differences of power and dominance, and other bases of hierarchy might be worthwhile to examine as well.

Kampkötter and Sliwka investigated the question in their paper (2018), whether supervisors should differentiate more between employees based on their performance, in the process of performance evaluation and bonus allocation. Their findings show that the willingness to differentiate between employees does have a positive effect on their performance, and consequently on the rise of future performance bonuses (individually and at large). However, notably this effect is stronger at the higher hierarchical echelons, whilst reversing to some degree at the lowest of levels. Therefore the authors suggest firms should employ stronger differentiation at the middle and higher levels, whereas being cautious of utilizing it at the bottom. Nevertheless, it seems, that much like with groups, hierarchical differentiation does have a complex and somewhat blurry effect on individuals, moderated by a variety of factors.

To tackle some of the highly difficult questions surrounding the inconsistent findings about the relationship of hierarchy and performance, Hays et al. (2022) offer a nuanced approach to study the impact of hierarchical differentiation on team performance, where they distinguish between two distinct types of differentiation, based on power and status. Hays et al. argue that the two hierarchies might interplay and co-effect team outcomes. They provide evidence that status differences beget a more competitive and less cooperative climate (p. 2098). Furthermore, they identify this as the determinant factor as in when power differences have detrimental effects on team performance (when both power and status differentiating is high).

In another recent paper To et al. (2022) propose a novel extension to models describing the relationship between hierarchy and performance. Their research understands team performance as not merely the result of hierarchy, but also as one of its future determinants. This is an idea quite similar to, although never explicitly stated in the paper to be derived from, the duality of structure, a core theorem of the highly influential structuration theory of Anthony Giddens (1984). As Giddens explains, structure enables agency, providing room for agents to act, and at the very same time structure is reproduced through agency and the action and interaction of agents. It is not a dualism where only one or the other exists, but a duality where one cannot exist without the other. What To et al. describe is kin to this perspective. They argue that hierarchical differentiation breeds performance success, and performance success reinforces or reshapes the hierarchical structure. Furthermore, they suggest the presence of an attribution process, as in a team greater influence is granted to members who are believed to be the causes of success.

As findings of research studies regarding the relationship of team effectiveness and hierarchical differentiation in particular, and hierarchy in general, were largely mixed, at best disaligned, at worst contradictory, Greer et al. (2018) commenced with an overarching meta-analysis review of this scientific landscape (evaluating 54 papers). They identified two competing perspectives among scholars investigating hierarchy, the "functionalist perspective" and the "conflict perspective". The perspective labeled as functionalist is a generally positive view on hierarchy, claiming hierarchy to enhance team effectiveness via improved coordination and cooperation processes. Whereas the conflict perspective offers a much harsher view on hierarchy, mainly highlighting its adverse effects on team effectiveness due to the amplification of a more conflict-laden environment. Greer et al. note that there has been a multitude of contingency factors identified in order to solve the discrepancy and explain the conflicting research results, such as task characteristics, team structure or form of hierarchy. Nevertheless, their results did not support the functionalist perspective on hierarchy (thus the existence of its significant improving effects on team effectiveness) in general. However, their research strongly supported the dysfunctional views on hierarchy, and showed that hierarchy negatively impacts team effectiveness, largely due to increasing "conflict-enabling states."

As mentioned above, the particular facet of team effectiveness that I strive to examine in my present thesis is team performance. And much like with the concept of team effectiveness, there are multiple aspects of the concepts hierarchy and hierarchical differentiation as well. In my research I am investigating soccer teams. In soccer teams formal hierarchy is not particularly prevalent or established, aside from the distinctive role of team captain, the chosen on-pitch leader of a team. Therefore, what I aim to examine is the informal hierarchy, and the impact of hierarchical differentiation resulting from this informal hierarchy.

Like previously explained, considering the definition of Magee and Galinsky (2008), there might be a variety of determinants of informal hierarchy present in any group of individuals. In my research sample, soccer teams, the seniority of a player in the team or the league, or the age of a player (which two, seniority and age, does not necessarily coincide), their current form in the season or their salaries, all may contribute to the shaping of the social rank order, known as hierarchy. The aspect of hierarchical differentiation I have chosen to understand and investigate within soccer teams was the differences in salaries, i. e. *pay dispersion*. This is an apt metric to symbolize hierarchical differences, as salaries are due to their very nature quantified, and not purely quantified, but in a form easy to grasp and perceive for all team members, or the average observers. Moreover, higher salaries are undoubtedly more desirable, and they do provide considerable incentives for soccer players. Higher salaries also show that a certain player is more valued, and more valuable to the club than other players are. Thus pay dispersion provides a within-team social rank order, one easy to discern, and one wherein every player of a team may be placed. To model pay dispersion I will use, in line with literature (Harrison & Klein, 2007), the metrics Gini index and the coefficient of variation (see more *Independent variables*).

Thus to redefine the research topic of my thesis (the relationship between hierarchical differentiation and team effectiveness) more accurately, I aim to shed light on the relationship between pay dispersion and team performance. Drawing on Greer et al.'s meta-analytic literature review (2018) and Hays et al.'s more recent study (2022) I surmise that pay dispersion causes detrimental fractures in a team's internal integrity, due to working as a causal agent for creating and escalating more conflicts, and therefore has an overall adverse impact on team performance. Consequently, I formulate my first hypothesis (H1) to test this effect as follows:

> **Hypothesis 1.** *Pay dispersion is negatively related to the performance of teams, as in higher pay dispersion resulting in lesser team success.*

When examining the functionalist perspective that claims hierarchy improves the cooperation and coordination processes of teams, Greer et al. (2018) found no significant support for this notion in their meta-analysis. Furthermore, Hays et al. (2022, p. 2098) found evidence that hierarchical differentiation does make a team's climate more competitive and less cooperative. Therefore I theorize that hierarchical differentiation diminishes cooperation, meaning that pay dispersion may adversely affect the cooperative behavior of teams. The duty of hypothesis two (H2) is to test this particular effect.

> **Hypothesis 2.** *Pay dispersion is negatively related to the cooperative behavior of teams, as in higher pay dispersion resulting in less cooperation within the team.*

The assumptions preceding the first two hypotheses suggest that hierarchical differentiation (in this thesis pay dispersion in particular) raises the level of competitiveness within a team (H2) and creates room for conflicts between individuals of the team (H1). Hierarchical differentiation thus is warranted to cause enhanced tensions within groups, and quite reasonably arises the question whether hierarchical differentiation can in fact result in explicit misbehavior. In a study examining interpersonal competitiveness Dumblekar (2010) found a close relation between competitiveness and aggressivity. Yet, research regarding this connection is not particularly exhaustive. Albeit, studies can be found in somewhat differing fields, as Schmierbach's paper from 2010, which examined the link between competitiveness and aggression in online gaming, or the paper of Krisnadewi and Soewarno (2020) wherein competitiveness was determined as a major factor in causing more aggressive organizational behavior. The line of argumentation is quite similar nevertheless in all the aforementioned cases. Competitition increases pressure, with pressure increasing frustration and evoking the need to perform better, which results in aggressive behavior. Now, given the hypothesized increased within-team competitiveness and the creation of a more conflict-laden environment I suspect the presence of a relationship between hierarchical differentiation and aggressivity. The difference in salaries doubtless creates tension within the soccer teams, as it makes it most easy to compare and understand the rank order, by making it quite evident which players are valued more by the club. Moreover, it is an apparent monetary benefit and thus a stark difference between teammates who, with a slight oversimplification, do have the same jobs. And this indeed I aim to test in my data analysis, therefore the third hypothesis of my research reads as follows.

> **Hypothesis 3.** *Pay dispersion is positively related to aggressivity, with higher pay dispersion enhancing aggressive behavior of the team.*

As stated above, to examine my research question and to test the three derived hypotheses I will utilize sports data. Now, the questions may arise, why use sports data to study business organizational phenomena, what benefits does it bring and what constraints does it mean for the research.

The usage of data collected from professional sports in management research is not a new-found approach. In 1984 Robert Keidel had already published an article (and the beginnings of using sports data in management research date even further back, see e. g. Gamson and Scotch (1964)), where he argued for the applicability of sports data in organizational setting, and determined the structure and management of the three major professional sports (in the US, baseball, basketball and football, providing models for vastly different team structures) as useful for understanding and shaping business organizations, or as he phrased for "determining their best game plans" (1984, p. 5). What Keidel showed in his article were the striking similarities between sports teams performing on the pitch and organizational groups, and between sports teams and organizations as a whole. Among the many parallels, and reasons why sports teams are apt models for businesses, Keidel listed the "need to compete externally" alongside with the "need to cooperate internally", furthermore the necessity of strategic human resource management and the fact that sports teams do resemble generic structures (1984, p. 12). These structures aid managers in understanding how their organizations work.

In his 1987 work Keidel expanded his triadic sports-model framework integrating it with several core constructs of organizational literature. The three major US professional sports serve as metaphors for autonomy of organizational parts and independence (baseball), hierarchical control and dependence (football) and voluntary cooperation and interdependence (basketball, interchangeable with soccer according to Keidel) (1987, p. 592 and 596). Moreover, Keidel argues that sports data has further benefits, such as the easy accessibility of high quantities of high quality objective data and unified measures that unambiguously quantify performance and success (1987, p. 608).

In their article from 2012, Day et al. reviewed studies that combined sports science and the field of organizational behavior, to assess the core themes and contributions of such endeavors. Their findings reinforce that professional sports can be used excellently to model the fundamental issues between competition ("getting ahead") and cooperation ("getting along"), moreover also to study succession, performance and motivation and dealing with pressure. They argue that sports are ever so suited to be analyzed as they offer a "living laboratory" where "life simplified" may be observed. The rules are explicit and known to all players and agents, moreover there are clear boundaries constraining the action, and winners and losers may be unanimously identified (2012, p. 399). Furthermore, in professional sports large stakes are dependent upon individual and group performance, and as much as entertainment, it is the constant generating of high revenues that stands in cardinal focus. The need to attain high, or leastwise sufficient, financial performance is another factor showing considerable parallel between sports teams and business organizations.

As of writing this thesis, the most recently published review of studies using sports data in management context is the work of Fonti et al. (2023), a literature review of great magnitude, where they identified and assessed 249 papers from the last five decades. Fonti et al. list a multitude of research areas in the field of management, where sports data was utilized to great success in the past years. Such areas include literature around the resource based view, status and reputation, risk-taking behavior, leadership, motivation and many more. The genuine impact of sports data on management research is thus quite apparent in this retrospective view. Furthermore, Fonti et al. highlighted how sports data may advance management research, as in aiding theory building and theory testing, radical theorizing (i. e. moving away from the traditional settings, and thus lessening the binding influence of "taken-for-granted" theoretical perspectives and creating room for novel views (p. 336)), and exploring emerging phenomena (due to high visibility of actors). The authors argue that beyond that sports data could even help alleviating certain concerns regarding management research, as in increasing validity through offering methods for triangulation or providing opportunities to replicate findings of management studies in different and data rich contexts.

However, as Fonti et al. pointed out there do exist certain drawbacks of relying on sports data (2023, pp. 346-348). Accessibility of sports data might diminish to some extent in the future, as organizations who compile and collect high quality datasets are growing increasingly aware of the value of such datasets and will protect them. A deeper understanding of the sporting context is a necessary precondition for analysing such data, which may not be the case for all researchers. For testing certain theories some sport settings might fit only to a limited extent, or simply not at all. Moreover, the researchers need to be aware of the possibilities of path dependencies, which might greatly influence the investigated phenomena, but which are lost once the analysis is transitioned into the realm of sports. And maybe most importantly the question of generalizability. Researchers have to be mindful how, and to what extent, the findings may be extended to business organizational settings, and reflect on the boundaries of sports data when understanding managerial phenomena.

However, the benefits duly outweigh the drawbacks of using sports data, (and Fonti et al. even offer remedies and mitigation approaches). As the last part of my theoretical research I examined the literature that relied on sports data, whilst studying pay dispersion and its effects.

Halevy et al. (2012) investigated the effects of hierarchical differentiation on the sporting success of NBA teams (basketball). The authors defined hierarchical differentiation as the dispersion of pay and of participation in games. They found that hierarchical differentiation increased team performance, due to increased cooperation and coordination. These findings stand in stark contrast to papers predicting hierarchical differentiation to have malign effects on team success, such as Greer et al. (2018) or Hays et al. (2022), and it is contradictory to my predicted H1. Given that according to Keidel (1987) basketball and soccer teams are interchangeable for examining managerial phenomena, this is an interesting premise.

To contribute to research surrounding human resource values and pay allocation Hill et al. (2017) examined teams from the MLB (baseball). They theorized that there is a congruence between the dispersion of human resource values and pay dispersion, moreover that this congruence is positively related to team performance. They found support for their theory, and in addition showed, that this congruence between value and pay dispersion positively moderates the relation between overall resource value and team performance. Whilst these results are doubtless interesting, it has to be noted that, as referred above by the drawbacks listed in Fonti et al. (2023) the characteristics of the scrutinized sport have to be taken into account. As heavily emphasized in Keidel's works (1984, 1987) baseball is a somewhat peculiar sport, where task independency is exceptionally high, and a sport where within team cooperation is not as essential as in other sports (e. g. football or basketball). Therefore to which organizational settings these results may be transferred to have to be carefully chosen.

Mondello and Maxcy (2009) evaluated the impact of pay dispersion and pay incentives on team performance in NFL (football) teams. Contrary to the findings of Halevy et al. (2012) and Hill et al. (2017) and in line with the dysfunctional view of hierarchical differentiation, the authors found a strong negative relationship between pay dispersion and on-field team performance, as in when pay dispersion increases, team performance significantly decreases. An interesting finding of the study was that pay dispersion however did positively correlate with team revenue earned. This is a somewhat paradoxical conclusion, as winning (negatively related to pay dispersion) and team revenue (positively related to pay dispersion) also correlate positively among themselves.

Franck and Nüesch (2011) analyzed the impact of pay dispersion on team performance in professional soccer teams (Bundesliga) and hypothesized the presence of a nonlinear effect. The authors found evidence for a U-formed relationship between pay dispersion and team performance, with moderate pay dispersion being the most detrimental for sporting success and very low levels of pay dispersion and high levels of pay dispersion enhancing team performance. Thus according to Franck and Nüesch teams should either follow an approach revolving around strong individualism or foster a "culture of cooperation" (2011, p. 3047). They also showed that the structure of salaries does effect a team's playing style, with greater pay dispersion increasing offensive and individualistic initiatives, which may be viewed in essence as increased risk-taking behavior.

Also in the realm of soccer did Bucciol et al. (2014) conduct their research, where they investigated the relationship between the performance of Serie A teams and their pay dispersion. A novel contribution of their study was the approach to analyze this relationship whilst employing different definitions of what constitutes a team. Bucciol et al. found, that if taken the narrowest definition of team (i. e. the players who play in a given game), then pay dispersion has a significant negative effect on team performance. In their research this effect of pay dispersion disappeared once they used the wider definition of team, taking into account not only those who directly contributed to the outcome. An interesting and somewhat unexpected finding of their research was moreover, that the detrimental impact caused by high pay dispersion could be attributed to worsening individual performances, but not to decreasing cooperation within the team. This stands partially against the traditional dysfunctional view on hierarchical differentiation, as that assigns the burdened cooperation as one of its pivotal arguments against the benefits of hierarchy.

In a more recent study Di Domizio et al. (2022) examined pay dispersion and team performance, also studying teams of Serie A, the top Italian professional soccer league. They employed weighted wages, meaning that they adjusted salaries with the ratio between the average salary of a given team and the league average. Di Domizio et al. confirmed the significant positive effect of relative wages on team performance, supporting the widespread notion that financially mightier teams do perform better. Moreover their results show a significant negative impact of pay dispersion on team performance. These results contradict to some extent the results of the study of Bucciol et al. (2014), which did not show a significant negative relation between pay dispersion and team performance once looking at the entire roster (only for players involved in the outcome). Albeit both papers studied the Serie A, Bucciol et al. took the results of single matches to measure team performance, whereas Di Domizio et al. focused on the ultimate outcome of the entire season. The discrepancy may lie in the approach Di Domizio et al. took to measure weighted wages, where they attempted to control for the financial power of teams, but at the same time Bucciol et al. also used average pay in their models as a control variable for a very similar end.

Thus there is evidence for both positive and negative effects of pay dispersion on team effectiveness once investigating sports data. Yet, arguments and findings on the side showing a negative impact appear to duly outweigh those standing for a positive. In this thesis I aim to replicate these findings conducting an analysis using data collected from professional soccer, whilst relying on a different setting. In my analysis I will be examining teams in the Premier League, the most high profile soccer league around the world. To my best knowledge, and as of writing this thesis, no study was published that did so. Furthermore, in order to control for the financial power standing behind the teams I take a notably different approach in this thesis (see *Control Variables*, financial_might), compared to what the previously referred studies did.

3. Methodology

3.1. Data collection and model

In the following section of the thesis the research design is introduced, detailing the empirical approach for data collection and analysis. As I have argued above, for its various and numerous advantages, I employed data collected from a professional sport. The chosen sport was soccer and I collected the dataset from the Premier League.

The Premier League is the upperest echelon of the English soccer league system. It was chosen to be the center of my analysis as it is historically understood as one of the most competitive professional soccer leagues (Ramchandani, 2012), and as of writing this thesis it is by large the most valuable soccer league. According to the widely accepted, and in the realm of scientific sports research oftentimes relied on website *www.transfermarkt.de* (Franck & Nüesch, 2011), the entire Premier League constitutes an overall market value exceeding 10.42 billion euros. In comparison, the second most valuable league, LaLiga, the highest Spanish soccer league, is estimated to have an overall market value of merely 4.78 billion euros, less than half of what Premier League is valued (Transfermarkt, 2023). In every season there are 20 teams competing in the Premier League and every team plays against every other competitor twice, thus a team plays altogether 38 games in any given season. A team may win, draw or lose a game, earning three, one or zero points, respectively. At the end of the season three teams are relegated to the Championship, the second highest English soccer league, and in return three teams are promoted to the Premier League. The aim of all teams ever is to collect as many points as possible, avoid relegation and strive for the top of the table to attain the prestigious and very lucrative qualification for the UEFA Champions League.

The final dataset contains seasonal level data of teams, therefore the unit of analysis is a given team's given season. I collected the data from the last five seasons of the Premier League that were completed at the time of writing this thesis, beginning with the 2017/18 season and ending with the 2021/22 season. Consequently, there are in total 100 team-season observations. A variety of metrics (e. g. number of points attained, number of passes completed, number of fouls committed or history, measured in seasons played in the Premier League) was collected using the official website of Premier League, *www.premierleague.com*. Individual-level statistics of players from every season were collected as well, utilizing the aforementioned *www.transfermarkt.de* (e. g. Premier League experience or age, and even tenure of coaches). For payroll data I have used the website *www.spotrac.com*. The individual-level data contributed to computing team-level metrics, such as the average of age, the average of Premier League experience or the variables measuring pay dispersion (see below, *Independent Variables*). The financial data of clubs is made publicly available through *www.gov.uk*, an official public sector information website of the United Kingdom, where an extensive record of all companies' filing history can be found (see more, *Control Variables*, financial_might). The validity of data collected from *www.gov.uk* and from the official website of Premier League is naturally exceptionally high. As I have mentioned above, *www.transfermarkt.de* is indeed a website frequently employed over a lengthy period of time in scientific research, and a website considered reliable (Franck & Nüesch, 2008; Frick, 2011; Lepschy et al., 2020; Torgler & Schmidt, 2007). The website *Spotrac* is also utilized in scientific research for player compensation and contract information (Mills & Winfree, 2018; Soebbing et al., 2022). Therefore concerns regarding the validity and reliability of the data are deemed to be unfounded.

In order to test the hypotheses and investigate the impact of pay dispersion on team performance, cooperation and aggressivity, I estimated several linear regression models. The simple structure of these models was the following (see Equation 1). Explanation for the individual variables follows in detail. I discuss the necessary assumptions for OLS (ordinary least squares) regression at the end of this section of my thesis.

The equation of linear regression models used:

$$Y = \beta_0 + \beta_1 * \text{pay dispersion} + \beta_i * \text{control variables} + \varepsilon \quad (1)$$

where,

- Y = dependent variable (team performance, cooperation or aggressivity)
- β_0 = the intercept
- β_1 = coefficient associated with the measure of pay dispersion, here the Gini index or the coefficient of variation
- pay dispersion = independent variable, Gini index or coefficient of variation
- β_i = coefficients of the control variables
- control variables = history, financial might, average age, average experience, average pay, roster size, coach tenure and past performance
- ε = error term

3.2. Variables

3.2.1. Dependent variables

To measure a team's seasonal performance I used the variable team_performance, which is in essence the number of points collected throughout the given season. This is an apt metric for the purpose, as collecting as many points as possible is what ultimately decides the final league placement of the teams (and higher league placement is an unequivocally positive and desired outcome). The more games a team wins (or does not lose, as draws do bring points), the more points it may collect. 20 teams competed in all the five seasons I have observed (and as a matter of fact ever since the 1994/95 season it always has been 20), and each team in each season did indeed play all of its 38 games. Thus I used the absolute number of points achieved for the variable without any transformation (e. g. using the percentage of maximum points achieved). The maximum number of points theoretically achievable was 114 in all five seasons.

In soccer the most frequent element of play that signifies cooperation and interaction between players of a team is passing the ball. Therefore, as in Franck and Nüesch (2011, p. 3046), to judge the display of cooperativeness within teams I applied the total number of passes a team had completed in a given season as proxy. This second dependent variable I named cooperation.

Several options presented themselves to account for the variable measuring the overall aggressivity of a team, such as the number of yellow cards, red cards or fouls. Albeit red cards undoubtedly signify the most severe and impactful violations of the game and its rules, the idea of relying on them was quickly discarded, as they are far too rare and far too impactful (a red card inevitably means a team losing a player, which in nigh all cases is a most grave setback). A multitude of fouls is awarded with yellow cards, yet the correlation between the number of yellow cards and the number of fouls in the dataset proved to be quite low, 0.317. There may be multiple reasons behind this low correlation. From one side a large portion of fouls is punished with a free kick for the opponent. These may very well be aggressive or hostile actions, yet not necessarily the like of which that results in the player being officially cautioned by the referee (receiving a yellow card). Whereas from the other side there are indeed cases of misconduct punished with yellow cards, that do not coincide with openly aggressive behavior, as delaying the restart of the game or removing the shirt in celebration. Therefore, I chose the number of fouls committed throughout the season by a given team to be the proxy for aggressivity, as it represents the aggressive tone of a team more suitably. This third dependent variable I called aggressivity_f.

3.2.2. Independent variables

The independent variable of this research is the one measuring pay dispersion within a team. This type of within-unit diversity may be categorized as disparity according to Harrison and Klein (2007, p. 1200), and in line with their recommendations I computed the Gini index and the coefficient of variation in order to model pay dispersion (p. 1210). Every single test I performed and every single model I estimated, I repeated with both coefficients in order to check for robustness of my findings. For clarity, only the numbers retained from the calculations with the Gini index are reported in my thesis. The calculated variables for the Gini index and the coefficient of variation show a remarkably high correlation between themselves, with a value slightly exceeding 0.960.

Normally, the Gini index is calculated from dividing the mean of the differences between all possible pairings of units with the size of the mean. However, in case all units are of the same size and they are ordered according to size, then the following Equation 2 yields the Gini index (Damgaard & Weiner, 2000):

$$G = \frac{1}{n}\left(n+1-2\left(\frac{\sum_{i=1}^{n}(n+1-i)\,x_i}{\sum_{i=1}^{n}x_i}\right)\right) \tag{2}$$

where,

- n = roster size, the number of players within a team
- i = index of the population, goes from 1 to n, the order is nondecreasing, with $x_i \leq x_{i+1}$
- x_i = the salary of i player

Still, Biemann and Kearney (2010, p. 591) have shown that the Gini index, as most estimators of diversity, does entail a certain bias. According to the authors, this systematic bias of the Gini index is most burdensome for smaller groups, where the level of disparity is underestimated (p. 591). They suggest using a corrected formula of the Gini index, see Equation 3. As the size of population in my present research corresponds the size of soccer teams, this bias-corrected version of the Gini index is indeed employed to avoid distortion caused by small population sizes.

Bias-corrected formula of Gini index:

$$G_n = G * \frac{n}{n-1} \tag{3}$$

where,

- G = Gini index calculated with the formula of Equation 2
- n = the size of population, here the roster size, the number of players within a team

This bias-corrected version of the Gini index calculated with Equation 2 and 3 is the independent variable, called pay_dispersion_gini.

The index I used to compute the second pay dispersion variable, for robustness tests, was the coefficient of variation (CoV). By definition the coefficient of variation is calculated by dividing the standard deviation with the mean (Bedeian & Mossholder, 2000, p. 286). The formula is present in Equation 4.

Coefficient of variation:

$$CoV = \frac{SD}{\overline{x}} = \frac{\sqrt{\frac{1}{n} * \sum_{i=1}^{n}(x_i - \overline{x})^2}}{\overline{x}} \tag{4}$$

where,

- SD = the standard deviation of player salaries
- $\overline{x}$ = the mean of player salaries
- n = roster size, the number of players within a team
- x_i = the salary of i player

Once again however, Biemann and Kearney (2010, p. 592) advises caution, as they argue the coefficient of variation is inherently biased as well, and once again, more significantly for smaller groups. Given that the standard deviation, which is used to calculate the coefficient of variation, is a biased measure itself, this is unsurprising. The authors recommend using the bias-corrected formula of standard deviation as basis for calculating the coefficient of variation, and show in their paper that this correction indeed yields significant improvement. In order to attain an unbiased estimation of the standard deviation the following formula shall be calculated (p. 589):

$$SD_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{q}} \tag{5}$$

This estimation noticeably uses a q value instead of the n, the group size. This q is the denominator of the unbiased estimate of the standard deviation, and is derived from Cureton's value table (1968) to any given group size. Consequently, the bias-corrected version of the coefficient of variation is the ratio of SD_n and the mean. This is the second version of the independent variable, and I named it pay_dispersion_cov.

3.2.3. Control variables

Following the praxis of scientific sports research I included a number of control variables in my model. These control variables I have drawn from relevant research papers and logic too, and they are expected to influence the variation of the dependent variables investigated.

A most reasonable assumption, shared by researchers and habitual soccer fans alike, is expecting greater money to beget greater results. Indeed, most research papers attempt to control for the differing economic might of sport clubs, which, especially in European soccer, and especially in the Premier League does vary to a great degree. Di Domizio et al. (2022) use the size of the hometown of a given team to account for a potential market size and thus source of revenue. Halevy et al. (2012) employed the average pay of players to control for the economic forces behind a team. Bucciol et al. (2014) took the same approach and Hill et al. (2017) included both aforementioned variables as controls. Franck and Nüesch (2011) on the other hand use the absolute wage expenditures. Whilst these are all judicious applications, in my thesis I follow a notably different route. The usage of the population of the home town of a team is quite problematic. For instance the population of Berlin (3.6 million) far exceeds the population of Munich (1.5 million), and even greater is the difference between the population of London (8.9 million) and the population of Manchester (0.5 million), however claiming that the financial might behind the clubs of Berlin or London far outweighs that of the clubs of Munich or Manchester, respectively, is dubious at best, baseless at worst. Average pay or total pay are both better proxies for financial might, yet a team can make a lot of expenditures not directly towards its players that contribute to its success. Better training facilities, equipment, academy, medical teams and analysts are all potentially helpful investments to increase sporting success. Therefore, in this thesis I use the total revenue of teams from their annual financial statements to form a better understanding of the financial might that stands behind the teams. This data was available through the *www.gov.uk*, where all UK based companies have to report their accounts. The control variable is called financial_might. Nevertheless, I also included the most frequently used control variable for economic power, average_pay in the model.

Sometimes success does breed success, thus to control for the effects of this potential performance trajectory I used the league placement of the last season. Previous performance may strongly influence performance in the next season (Hill et al., 2017, p. 1941). I transformed this variable for it to intuitively function and the higher numbers thus indicate better positions (20 equals first place, 19 for second place,...). I named it past_performance. The three teams that were promoted in each season were treated had they reached the last three places of Premier League in the previous season, with the team winning the Championship being assigned to place 18, the runner-up of the Championship to place 19, and the last team to qualify to 20 (consequently receiving scores of 3, 2 and 1, respectively).

In the research model I controlled for the experience the coach has with the team, coach_tenure (Bucciol et al., 2014) and also for the average age of the roster (Di Domizio et al., 2022) and the size of the roster (average_age and roster_size, respectively). Moreover, I included a control variable for league seniority, in order to account for the teams' Premier League experience (Di Betta & Amenta, 2010), counted as the number of seasons a team has competed in the highest of English football leagues (history). Lastly, I utilized the variable average_experience to control for the average amount of years the players of a club have played in the Premier League.

A concise overview of all the dependent, independent and control variables is shown in Table 1, containing the names, definitions and sources of each of them.

3.2.4. Summary statistics

Summary statistics of all variables are reported in Table 2. The dataset contains a 100 observations throughout the five seasons. Altogether data stems from 28 different teams and 14 teams have competed in all five of the seasons investigated.

Table 1: Variable overview

Variable	Definition	Source
DEPENDENT		
team_performance	absolute number of points collected by team *i* at the end of season *t*	www.premierleague.com
cooperation	total number of passes completed by the players of team *i* throughout season *t*	www.premierleague.com
aggressivity_f	total number of fouls committed by the players of team *i* throughout season *t*	www.premierleague.com
INDEPENDENT		
pay_dispersion_gini	Gini index describing the dispersion of players' salaries in team *i* in season *t*	
pay_dispersion_cov	coefficient of variation describing the dispersion of players' salaries in team *i* in season *t*	
CONTROL		
financial_might	total revenue of team *i* reported in the financial statement of the fiscal year *t*	www.gov.uk
average_pay	the mean of players' salaries in team *i* for season *t*	www.spotrac.com
past_performance	the number indicating the league placement of team *i* in season *t-1*, corresponding in reverse order, for the larger numbers to indicate higher placements	www.premierleague.com
coach_tenure	the number of seasons the coach in charge of team *i* has spent with team *i* till the end of season *t*	www.transfermarkt.com
average_age	the mean of players' ages in team *i* in season *t*	www.transfermarkt.com
roster_size	the number of players team *i* has in season *t*	www.spotrac.com
history	the number of seasons team *i* has played in the Premier League by the end of season *t*	www.premierleague.com
average_experience	the mean of the number of seasons each player of team *i* has played in the Premier League till the end of season *t* (not necessarily as part of team *i*)	www.transfermarkt.com

On average, the teams succeeded in collecting 52 points in a season, whilst completing more than 17,400 passes and committing more than 220 fouls. The mean of the variable history is quite high, signifying that most teams have taken part in the highest league more than 17 times. Interesting is, that the median of the variable coach_tenure is but 2 seasons, which speaks for the rarity of coach longevity in the Premier League. The mean is considerably higher, yet this is due to outlier extreme values (e. g. Arsene Wenger has spent 22 years coaching Arsenal).

Notably, there is a vast difference between the minimum and maximum values of the variable financial_might. 26.4 million and 627.1 million stand on the two ends of the scale (more than 0.6 billion difference), which further showcases the variety in the extent of the economic power standing behind a team. The average of the Gini index is 0.325, and it ranges from 0.139 to 0.518 in the dataset. Normally, the Gini index has a range between 0 and 1-(1/n), but the bias-corrected version does range between 0 and 1, with the higher values being associated with greater disparity.

All in all, the descriptive statistics tell that the variables do show sufficient variation, which is a necessity for performing regression analysis. The further imperative assumptions of linear regression will be discussed in the next sub-section.

3.3. Assumptions of linear regression

To be able to use the OLS (ordinary least squares) method for estimating the models of the three dependent variables (team performance, cooperation, aggressivity) certain assumptions need to be investigated first. To test these assumptions, and to calculate the linear regression models later on, I used two different softwares designed for statistical data analysis, JMP and STATA.

Firstly, regression analysis requires a complete selection of relevant exogenous variables. As most real life phenomena is a result of processes of great complexity, naturally utter completeness of a model is impossible, even on a theoretical level. Nevertheless, I selected the variables after careful examination of scientific practice in research papers akin to my thesis, and the presence of these variables is supported by logic too. Thus the models should be deemed as sufficiently complete. Due to the nature of data collection, systematic measurement errors are quite unlikely. Furthermore, as mentioned above the variables show significant variation (see Table 2).

I calculated the correlation between all independent and control variables in order to detect multicollinearity. This was a necessity as high correlation between independent and control variables might render the regression model

Table 2: Descriptive statistics

Variable	Mean	Median	Std. dev.	Min	Max
DEPENDENT					
team_performance	52.670	49.500	18.511	16	100
cooperation	17,447.510	16,541	3,884.648	10,226	28,241
aggressivity_f	222.520	223	27.499	161	307
INDEPENDENT					
pay_dispersion_gini	0.325	0.329	0.076	0.139	0.518
pay_dispersion_cov	0.602	0.601	0.157	0.250	1.161
CONTROL					
financial_might	246,152,601	173,462,500	157,825,173	26,400,000	627,122,000
average_pay	2,965,710.700	2,561,033.300	1,606,817.600	729,955.560	7,660,178.600
past_performance	10.500	10.500	5.795	1	20
coach_tenure	3.110	2	2.964	1	22
average_age	25.773	25.661	1.173	23.409	29.158
roster_size	25.020	25	3.012	17	32
history	17.490	21	9.869	1	30
average_experience	4.673	4.960	1.394	1.222	8.095

strongly biased. The correlation matrix can be found in Table 3. Four control variables showed strong correlation between themselves, history, financial_might, average_pay and past_performance. Between these four all pairwise correlations exceeded 0.6. As financial_might was expected to have the strongest influence on sporting success, I removed the other three from all the models. Now, none of the independent or control variables showed a correlation exceeding 0.6 between each other (with the highest value being 0.531 between average_experience and average_age, and all other correlations being under 0.5). Moreover the Variance Inflation Factors (VIF) were all under or barely above 2 (Mean VIF = 1.49). Therefore going forward the issue of multicollinearity is solved and it does not hinder the models any further.

To test the linearity of the parameters I employed Ramsey's Regression Specification Error test (RESET). The RESET test tests whether the inclusion of nonlinear combinations of fitted values improves explaining the variation of the dependent variable (Volkova & Pankina, 2013, p. 265). In my dataset the RESET test was not significant for team performance and cooperation, with p-values of 0.725 and 0.783, respectively. Unfortunately, for aggressivity the RESET test was significant with a p-value of 0.017. For this model the functional values were likely not defined correctly. In situations like this a possible solution could be transforming the variables, and this I indeed attempted. As the RESET test still remained significant after x^2-, x^3-, and log-transforming the independent variable, I will not handle this issue further in this thesis. The results of the models for aggressivity however, need to be viewed with this knowledge in mind.

Furthermore, for linear regression the error terms need to be normally distributed. This tends to be an issue mostly with smaller sample sizes, nevertheless I performed skewness and kurtosis tests in STATA (called *sktest*). The skewness and kurtosis tests were not significant for any of the models, with retaining p-values of 0.336 (for team performance), 0.733 (for cooperation) and 0.191 (for aggressivity). Thus the non-normality of residuals is not a concern for the regression models.

Next, the assumption of homoscedasticity had to be tested. Homoscedasticity means the absence of heteroscedasticity, thus assumes a constant variance of the residuals. To test for heteroscedasticity (the variance of the residuals being not constant) I performed the Breusch-Pagan test for the models. Neither for team performance (p-value = 0.110), nor for aggressivity (p-value = 0.130) was the Breusch-Pagan test significant, however it was significant for cooperation with a p-value of 0.001. Fortunately, STATA offers a command (called *vce(robust)*) to use White's heteroscedasticity-corrected standard errors, also known as robust standard errors. These robust standard errors will be used onwards to estimate the regression models for cooperation.

Testing for autocorrelation proved non-trivial, as the dataset was not a normal time-series dataset, but panel data, containing a multitude of short time-series independent from one another. Therefore neither the commonly employed Durbin-Watson test, nor the Breusch-Godfrey test was applicable. In order to test for serial correlation in panel data I used the Wooldridge test recommended by Drukker (2003), after Wooldridge (2002). The Wooldridge test was not significant for team performance (p-value = 0.098), yet it was significant for cooperation (p-value = 0.019) and also for aggressivity (p-value = 0.012). Given the characteristics of my dataset, tackling this issue would be most burdensome, as the number of observations within the panel time series is far too few (in some cases only one or two). Still in a more exhaustive research approach one could attempt to nest the data on a seasonal or club level and proceed with a different regression estimation method. As the presence

Table 3: Correlation matrix of the independent and control variables

Variable	1	2	3	4	5	6	7	8	9
1. pay_dispersion_gini	1.000								
2. history	0.204	1.000							
3. financial_might	0.382	0.667	1.000						
4. average_age	-0.428	-0.296	-0.293	1.000					
5. average_experience	-0.141	0.337	0.223	0.531	1.000				
6. average_pay	0.319	0.729	0.896	-0.205	0.373	1.000			
7. roster_size	0.270	0.072	0.144	-0.035	0.031	0.091	1.000		
8. coach_tenure	-0.129	-0.180	0.045	0.159	0.165	-0.061	-0.009	1.000	
9. past_performance	0.347	0.621	0.812	-0.243	0.403	0.805	0.121	0.089	1.000

of autocorrelation does not automatically render the whole model unusable (though it does make the coefficients not efficient and the confidence intervals biased) I will proceed with the OLS method for all dependent variables.

After examining the necessary assumptions of linear regression and OLS I estimated the regression models to test for the hypotheses derived in the *Theoretical background* chapter of this thesis. The results are reported in the next section.

4. Results

4.1. Hypothesis testing

In my statistical analysis I commenced with examining the models for team_performance, the dependent variable measuring sporting success. First, I ran the model with the five control variables (three control variables were removed from the models, see *Assumptions of linear regression*). The regression model was significant (p-value < 0.001), with an F-value of 48.639 and an R^2 of 0.721 (adjusted R^2 = 0.706). Financial_might had a strong significant effect on team_performance, with a p-value below 0.001 and a coefficient of $9.178 * 10^{-8}$, moreover average_age had a significant negative effect at a ten percent significance level, with a p-value of 0.075 and a coefficient of -2,120. The other control variables did not have significant effects on the dependent variable.

Next, I added the independent variable, pay_dispersion_gini, to the model. The expanded model was overall significant as well (p-value < 0.001), with an F-value of 40.700 and an R^2 of 0.724 (adjusted R^2 being 0.706). Financial_might had once again a strong significant effect with a p-value below 0.001 and a coefficient of $9.385 * 10^{-8}$, and the effect of average_age was now significant at a five percent level, with a p-value equal to 0.048 and a coefficient of -2.444. The independent variable, pay dispersion did not have a significant effect on team performance, thus Hypothesis 1 was not supported. The other three control variables remained non-significant.

To examine Hypothesis 2, I estimated the model for the dependent variable cooperation. Once again as before, first only the five control variables were included. For this model I used White's heteroscedasticity-corrected standard errors in STATA (see *Assumptions of linear regression*). The model was significant (p-value < 0.001), retaining an F-value of 49.980 and an R^2 of 0.748 (with the adjusted R^2 being 0.735). The effect of financial_might was strongly significant, its p-value being below 0.001, and the coefficient taking the value of $2.003 * 10^{-5}$. Average_age had a significant negative effect with a p-value equal to 0.027 and a coefficient of -517.016. Coach_tenure had a significant positive effect, with a p-value equal to 0.024 and a coefficient of 109.697. The model showed no significant effect for the other two control variables.

In the next step, the independent variable, pay dispersion was included in the model. Once again the model was computed using robust standard errors. The expanded model remained significant (p-value < 0.001), having an F-value of 40.700. The R^2 had the value of 0.750 (adjusted R^2 being 0.733). The influence of the variable financial_might was still strongly significant, with a p-value smaller than 0.001 and a coefficient of $1.972 * 10^{-5}$. The control variable average_age showed a significant negative effect, having a p-value of 0.044, with a coefficient of -468.690. Coach_tenure remained significant as well with a p-value of 0.018 and a coefficient of 114.828. The other two control variables had still no significant effect on the examined construct. Given that in the model pay dispersion the independent variable, did not have a significant effect, Hypothesis 2 remained unsupported.

At last I tested for Hypothesis 3, estimating a regression model explaining aggressivity_f as the dependent variable. The model containing the five control variables was significant (having a p-value of 0.047) and had an F-value of 2.347. The value of R^2 took 0.111 (with the adjusted R^2 being 0.064). The variable coach_tenure had a significant negative effect with a p-value of 0.011 and a coefficient of -2.372. None of the other control variables showed a significant effect.

Including the independent variable pay dispersion in the model proved to be a remarkable improvement. The expanded model was significant (with a p-value of 0.010) and had an F-value of 3.040. The R^2 was considerably higher than previously, now retaining a value of 0.164 (adjusted R^2 = 0.110). Coach_tenure still had a significant negative effect with a p-value of 0.019 and a coefficient of -2.156. Furthermore, now financial_might also had a significant negative effect with a p-value equal to 0.039 and a coefficient of -4.292 * 10^{-8}. The other control variables remained non-significant. However, pay dispersion did have a significant positive effect on the dependent variable, with a p-value of 0.017 and a coefficient of 100.632. Consequently, Hypothesis 3 was supported by the model.

The results of the regression analysis are summarized in Table 4. Model 1, 2 and 3 correspond to the dependent variables team performance, cooperation and aggressivity, respectively. I reported the values only for the models including the pay dispersion, for the table to serve as overview for hypothesis testing. The coefficients are unstandardized and the standard errors are reported in the brackets.

To offer a brief summary of my endeavors for direct effect testing, I have to state, that the models provided mixed results. Hypotheses 1 and 2 were not supported by the regression analysis, therefore no significant effect could have been determined for pay dispersion to influence the performance of teams or their displayed degree of cooperation. Consequently, according to my regression analysis, sporting success and cooperative behavior does not significantly vary between soccer teams, who take vastly different approach in allocating the salaries among their players. However, Hypothesis 3 was supported by the analysis, and pay dispersion had a significant positive effect on the examined dependent variable, aggressivity. This result means, that indeed, higher pay dispersion tends to lead to an overall more aggressive tone within a given team, resulting in a greater number of fouls committed on the field. Among the control variables one that remarkably stood out was financial might, for it had a strong significant effect throughout all three models, undoubtebly testifying for the pivotal role economical power plays in the sport of soccer.

As the results of the models were less than satisfactory, particularly for Hypotheses 1 and 2, and as one control variable showed far greater dominance than any other variable, I decided to conduct further analysis. Therefore, an examination of the possible relationship between the independent variable and the control variable financial might follows in the second half of this chapter.

4.2. Supplemental analysis

The relationship between two variables in management research in particular, and quantitative research in general, is oftentimes dependent upon a third variable. These third variables are called moderating variables and they might exert great influence on the nature or strength of a relationship of two (or more) other variables, or on both (Dawson, 2014, p. 1). As after the theoretical research, and the eventually supported Hypothesis 3, I still firmly believed that the independent variable pay dispersion does need to have an influence on the dependent variables of the Hypotheses 1 and 2 (team performance and cooperation), I decided to conduct tests for moderation effect. My conjecture was that the most dominant control variable of my regression models, financial might, could potentially affect the relationship between pay dispersion and the dependent variables. Therefore, I decided to test for these interaction effects in Model 1 and 2.

In order to operationalize this endeavor, I relied on Jeremy Dawson's (2014) most helpful article, which extensively covers moderation in management research. As I strived to examine the interaction between one independent (pay dispersion) and one moderator variable (financial might) I tested for simple two-way interaction. In order to test for two-way interaction an interaction term needs to be included in the simple OLS regression model. For the interaction term I computed a new variable in JMP. In line with Jeremy Dawson's recommendation I mean-centered the two variables first, the two components of interaction. Mean-centering a variable means subtracting the mean from its values, thus creating a new variable with a mean value equal to zero. The interaction term is calculated then as the product of the mean-centered independent variable and the mean-centered control variable. I called this new interaction variable gini_x_financial_might. The interaction term itself shall not be mean-centered, and the dependent variables remain in their raw forms as well (p. 2). Using the method of mean-centering over others (e. g. z-standardization) is beneficial, as thus regression coefficients may be interpreted directly for the original variables. Jeremy Dawson strongly advises to mean-center all other control variables included in the model as well (p. 12). Consequently, in the coming models all variables, barring the dependent and the interaction, appear in their mean-centered forms. In order to interpret the results it is fundamental to include the main effects of the independent and control variable in the model, besides the interaction term (p. 2).

Thus the expanded regression model is as follows:

$$\begin{aligned} Y = \beta_0 + \beta_1 * \mathit{pay\ dispersion\ c} + \beta_2 * \mathit{financial\ might\ c} \\ + \beta_3 * \mathit{interaction\ term} + \beta_i * \mathit{control\ variables\ c} + \varepsilon \end{aligned} \tag{6}$$

where,

- Y = dependent variable (team performance or cooperation)
- β_0 = the intercept
- β_1 = coefficient associated with the measure of pay dispersion, here the Gini index or the coefficient of variation
- pay dispersion c = independent variable, the Gini index or the coefficient of variation, mean-centered
- β_2 = the coefficient associated with financial might
- financial might c = moderator variable, mean-centered
- β_3 = coefficient associated with the interaction term
- interaction term = product of the mean-centered independent variable (pay dispersion) and the mean-centered moderator variable (financial might), not mean-centered itself

Table 4: Estimation results of linear regression models

	Model 1	**Model 2**	**Model 3**
Constant	93.309*** (31.583)	26,775.88*** (6,305.958)	282.040*** (81.690)
pay_dispersion_gini	-16.038 (16.027)	2,385.367 (2,672.962)	100.632** (41.458)
financial_might	9.385×10^{-8}*** (7.903×10^{-9})	1.970×10^{-5}*** (1.980×10^{-6})	-4.292×10^{-8}** (2.044×10^{-8})
average_age	-2.444** (1.219)	-468.690** (229.433)	-2.775 (3.154)
average_experience	1.347 (0.966)	-294.697 (196.669)	1.915 (2.499)
roster_size	-0.098 (0.350)	-74.329 (73.405)	-0.497 (0.905)
coach_tenure	0.201 (0.348)	114.828** (47.807)	-2.156 (0.901)
R^2	0.724	0.750	0.164

Notes: There were a 100 observations included in all the models.
Robust standard errors were used for Model 2.
Statistical significance levels: * = 10%, ** = 5%, *** = 1%

- β_i = coefficients of the control variables
- control variables c = control variables average age, average experience, roster size and coach tenure, all mean-centered
- ε = error term

Thus I estimated this improved regression model of Equation 6, and tested first for team performance. The model was significant (p-value < 0.001) and had an F-value of 40.807. It had an R^2 of 0.756 (adjusted R^2 being 0.738). This seems to be an improvement compared to the previous model for team performance that did not have the interaction term included (R^2 was 0.724, and the adjusted R^2 was 0.706). In this model the control variables did not have a significant effect on team performance. However, the interaction between pay dispersion and financial might proved to be significant, with the interaction term having a p-value of 0.001, and a coefficient of -3.738 * 10^{-7}. My conjecture regarding the presence of a moderation effect therefore found support. The significant presence of the interaction term means, that the relationship between team performance and pay dispersion varies given the level of financial might (see Dawson, 2014, p. 3). A summary of the results of this model can be found in Table 5, along with those of the expanded Model 2.

In order to gain a better understanding of this significant interaction, I used the Excel template offered by Jeremy Dawson at his webpage, *http://www.jeremydawson.com/slopes.htm*. Entering the required data, unstandardized regression coefficients, means and standard deviations, the template plots the effect and visualizes it in a graph. This graph for the expanded model of team performance can be seen in Figure 1.

At the very first glimpse it is striking that the lines associated with low and high financial might have opposing steepness. The figure shows that the relationship between team performance and pay dispersion is positive when the financial might of a team is low. On the contrary, for teams of high financial might, the relationship between team performance and pay dispersion appears to be negative.

As the significance of the interaction term in the model only tells about the existence of a divide between low and high values of the moderator, and not about the significance of the relationship between the independent variable and the dependent variable within these groups, I performed simple slope tests to gain further insight (Dawson, 2014, p. 3). The simple slope tests can be computed with the Excel template of Dawson, used for the visualization in Figure 1. Adding the variance of the coefficient of the independent variable and the moderator, and their covariance of coefficients, I calculated the simple slope tests for the slopes plotted on Figure 1, which are one standard deviation above and below the mean of the moderator variable. The simple slope test for the group of low financial might showed that the slope, as observable, has a positive steepness of 28.889. The test was not significant at a five percent significance-level, with its p-value being 0.150. The t-value of the slope was 1.453. On the other hand, the simple slope test for the group of high financial might was significant, with a p-value of 0.001. The slope had a t-value of -3.446 and a negative steepness of -89.101.

This means, that for teams of high financial might pay dispersion does have a significant negative effect on team performance. The plotted slopes (Figure 1) hint at a weak positive relationship between pay dispersion and team performance amidst teams of low financial might, however this effect is not significant.

For the next step, I tested the interaction model presented in Equation 6 for the second dependent variable of my re-

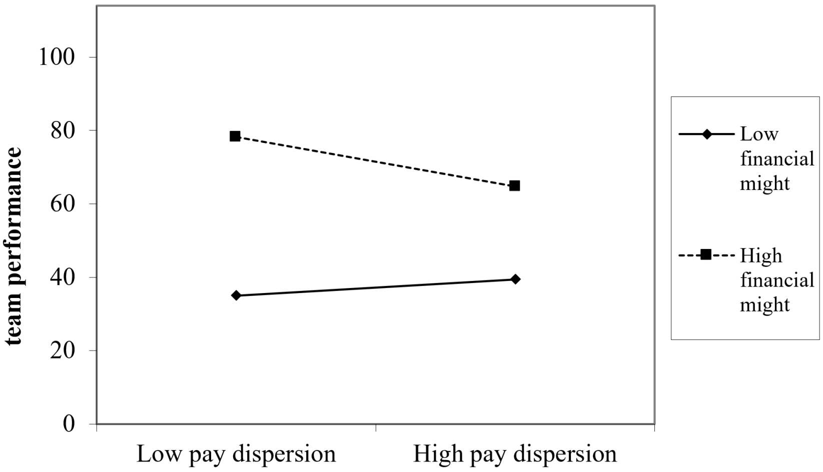


Figure 1: Moderating effect of financial might on the relationship between pay dispersion and team performance

search, cooperation. As before, to control for heteroscedasticity I computed the model using White's heteroscedasticity-corrected standard errors in STATA. This new model was significant as well (with a p-value below 0.001), and had an F-value of 36.680. The R^2 was 0.763 (with the adjusted R^2 taking the value of 0.745). Compared to the previous model for cooperation, where the interaction was not included, the R^2 appears to increase. It had the previous value of 0.750 (and adjusted R^2 was 0.733). Out of the control variables coach_tenure had a significant positive effect, with a p-value of 0.012 and a coefficient of 115.647. Moreover, average_age did have a, now weak, significant effect (only at a ten percent significance level), with a p-value equal to 0.087, and a coefficient of -383.985. Much like earlier, the other two control variables showed no significant effects. Significant was however the interaction term with the p-value of 0.011, and the coefficient of -0.001. This proves that, as with team performance in the expanded Model 1, the moderating variable, financial might, does influence the independent variable's relationship with cooperation. Given the values of financial might, there is significant difference between the relation of pay dispersion and cooperation within a team. For an overview of the results of this model see Table 5.

To attain an apt presentation of the interaction effect, I again relied on Jeremy Dawson's Excel template, to visualize the two-way interaction between pay dispersion and financial might, this time for cooperation. This visualization may be found in Figure 2.

Figure 2 shows much similarity to Figure 1 at first observation. The lines depicting the groups of low and high financial might have opposing steepness. It appears that for teams of lower financial status pay dispersion increases the number of passes and thus cooperation. On the other hand, pay dispersion within teams of higher financial status seems to have an adverse effect on cooperation.

As before, the significance of the interaction term within my model, and the plotting of the slopes in Figure 2, only testifies for the existence of a significant difference between the two groups, determined as one standard deviation above and below the mean of the moderator variable (high and low financial might, respectively). Therefore I performed simple slope tests to test for the significance of the individual relationships.

The results of the simple slope tests were the following. For the group of teams characterized by high financial might the slope had a steepness of -7515.308. The simple slope test was not significant for this group, having a p-value of 0.132. Its t-value was -1.520. The slope of the low financial might group had a positive steepness of 8474.292. Contrary to the test for the high financial might group, this simple slope test was strongly significant with a p-value of 0.007. The test's t-value equaled 2.745.

Testing for the interaction effect in the models of the second dependent variable showed that there is a significant difference between how pay dispersion affects cooperation given the level of financial power. The simple slope tests allowed a deeper understanding of this finding. Their results suggested, that within teams of high financial status pay dispersion does not significantly influence the level of cooperation, albeit the visual plotting of the slope implies a weak negative relationship (Figure 2). Yet, for teams of low financial status pay dispersion does indeed have a significant positive impact on the level of cooperation displayed.

Albeit the original hypothesis did find support in Model 3, I also tested for the presence of interaction for the third dependent variable, aggressivity. Although the expanded Model 3 was significant (having a p-value of 0.015), the interaction term itself was not significant, with a p-value of 0.522. Therefore, I did not investigate in this direction any further.

For robustness check I have calculated every test and model using the coefficient of variation as the independent variable pay dispersion, instead of the Gini index. The calculations were robust, with the tests (assumptions of linear regression) being significant precisely where they were with the Gini index and being non-significant in synchron

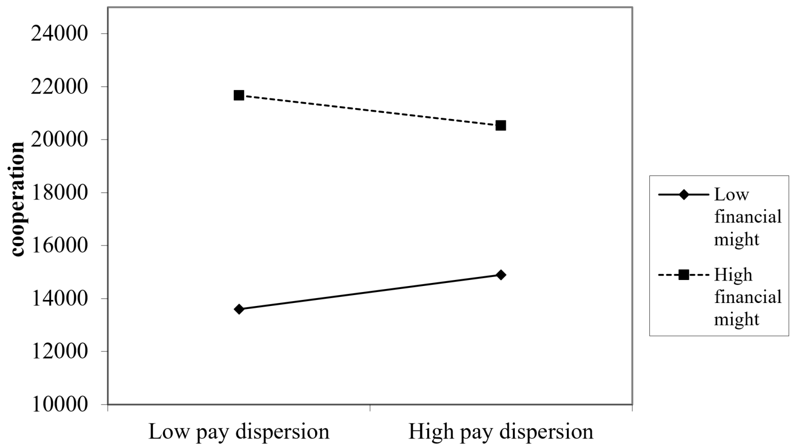


Figure 2: Moderating effect of financial might on the relationship between pay dispersion and cooperation

Table 5: Estimation results of the expanded linear regression models

	Model 1'	Model 2'
Constant	54.361*** (1.065)	17676.580*** (212.782)
interaction_term	-3.740 × 10^{-7}*** (1.070 × 10^{-7})	-5.060 × 10^{-5}** (1.960 × 10^{-5})
average_age	-1.819 (1.166)	-383.985 (221.825)
average_experience	1.281 (0.913)	-303.709 (194.530)
roster_size	-0.110 (0.331)	-76.003 (73.549)
coach_tenure	0.207 (0.329)	115.647** (45.297)
R^2	0.756	0.763

Notes: There were a 100 observations included in all the models.
The coefficients are unstandardized and the standard errors are reported in the brackets.
Robust standard errors were used for Model 2'.
Statistical significance levels: * = 10%, ** = 5%, *** = 1%

with the Gini index. Estimating the models retained largely the same results considering when variables were significant and at what significance level, with three small discrepancies. Firstly, in the regression model testing for pay dispersion's direct effect on team performance (Model 1), the control variable average_age was only weakly significant for the coefficient of variation model, with a p-value of 0.056. This small divergence may be deemed inconsequential for my thesis. Far more interesting is however, that using the coefficient of variation not only replicated, but even strengthened the findings surrounding the interaction effects, particularly the simple slope tests. As with the coefficient of variation, for both interaction models, both slopes were significant. As in, for team performance pay dispersion had a weakly significant positive effect on the low financial might teams, with the p-value of 0.051 (t-value 1.977, steepness 19.846). For cooperation pay dispersion had a weakly significant negative effect on the high financial might teams, having a p-value of 0.095 (t-value -1.687, steepness -3606.326). Therefore my findings can be considered robust, and employing the coefficient of variation even confirms effects (at a 10% significance level) that were only surmised with models of the Gini index.

In the next chapter follows a discussion of the results of my analysis.

5. Discussion

5.1. Theoretical implications

The findings of my thesis offer insight into how pay dispersion influences team effectiveness in soccer teams. My results showed, that at large the effectiveness and outcomes of teams are related to pay dispersion, however this relationship might be moderated by another variable in certain aspects. The presence and conspicuous nature of this moderating effect is what could be considered the chief contribution of my thesis to scientific literature studying hierarchical differentiation and team effectiveness. In sport setting in particular and in management environment in general.

Albeit the influence of hierarchy on the effectiveness of organizations and groups is highly debated, lately the consens has somewhat shifted in the direction of hierarchy and hierarchical differentiation being a detriment to positive outcomes (Greer et al., 2018). When testing for the direct effect of pay dispersion on team performance, my regression analysis showed no significant effect, and consequently Hypothesis 1 was not confirmed. This, to a certain degree contradicts both the studies that argued for an improving influence of hierarchy (such as Halevy et al. (2011), Kampkötter and Sliwka (2018), Ronay et al. (2012), and To et al. (2022)), and also the ones proclaiming adverse effects of hierarchy (such as Hays et al. (2022) or Greer et al. (2018)). This unexpected, and at first glimpse somewhat unsatisfying, outcome can be explained by the presence of a strong control variable (financial might) and its imposed interaction on pay dispersion. The results of the interaction analysis showed, that pay dispersion has significantly different effects on team performance given the degree of financial might, i. e. economical power behind the team. For the economically more powerful teams pay dispersion had adverse effects on sporting success, thus confirming the unified findings of the meta-analysis of Greer et al. (2018). For the economically less mighty teams the effect, albeit appearing slightly positive, was not significant in nature (for the Gini index, though weakly significant for the coefficient of variation), therefore the results of my thesis do not primarily support the row of studies arguing for hierarchy breeding success.

What stands behind this separation of effects is a compelling question and cannot be answered with utter surety based on my conducted research alone. Yet, one possible explanation I would propose here. The economically less powerful teams are usually the teams on the lower end of the Premier League rankings, thus the main concern for such teams is to avoid relegation and carve themselves a deserved and stable place in the highest of English soccer leagues. For teams like that having ‘star players’, differentiated from the other players with a disproportionately higher salary, might very well be regarded as means for attaining a much needed legitimacy. The presence of one or two such top players, separated in status, signals the belonging of a club to the Premier League (which is a primary concern of smaller clubs, as the competition is hard with a high profile second league below them, the Championship), and therefore likely be viewed more positively by other team members, as it contributes to a common and strongly desired end. And for that end it is likely recognized as a necessity, and less likely causes internal tension and conflict, which would considerably hamper team performance. On the other hand, for the economically powerful teams, as the ‘Big Six’[1], relegation and attaining sporting legitimacy is of no relevant concern, and their endeavors are largely directed towards winning trophies and qualifying for the Champions League. Now, such teams possessing great financial might are likely constituted from top-class players, who all were or could have been regarded as ‘star players’ in their respective carreers, yet even among them salary differences do and will occur. I believe a major psychological contributor to the negative impact of pay dispersion on the performance of economically powerful teams could be the greatly increased tension. As in this case the positive aspect of increasing a team’s status and legitimacy is virtually absent, as is the fear of relegation, and at the very same time ‘star players’ are selected (due to great differences in absolute salaries) among players who all, and maybe even rightfully, consider themselves ‘star players’. Differentiating between players who mostly find themselves quite worthy of eminence may easily create tension, and without the alleviating circumstance of, for lack of a better word, fighting for survival in the Premier League, this internal tension can cause detrimental consequences on sporting success. This is but a theory, and it may be interesting to investigate the possibility of this psychological connection. For that end, conducting deep interviews with players might be a necessity.

Hays et al. (2022) argued in their study for how hierarchical differentiation decreases cooperation within teams, through increased competition, and Greer et al. (2018) found as well that cooperation decreases as a result of increased differentiation through the birthing of more and severer conflicts. Testing for direct effect on cooperation, I found no significant effect in my regression analysis, thus neither Hypothesis 2 was confirmed. This appeared to contradict the theoretical background of my research, as null-effect was scarcely, if ever, theorized, yet once again testing for interaction effects helped to shed light on the multifaceted nature of this relationship. Financial might indeed influenced the relationship between pay dispersion and cooperation, and there was a significant difference between the high and low financial might clubs. For the economically more powerful clubs pay dispersion appeared to decrease cooperation, yet the effect was not significant (for the Gini index, but weakly significant for the coefficient of variation). For the economically less powerful clubs however pay dispersion improved cooperation. This finding, for the low financial might clubs, supports the findings of Halevy et al. (2011), where they argued for an increased voluntary cooperation and contradicts the studies of Greer et al. (2018) and Hays et al. (2022).

Explanations for this dual nature of the relationship between pay dispersion and cooperation could follow a similar line of argumentation as for team performance (see above). It even supports the previous argumentation, as it shows a significant improvement in cooperation for the highly differentiated, yet economically less powerful teams. In other words, it appears that elevating the status of players, or ‘creating superstars’, might be rewarded with a positive echo from teammates if they recognize it as a pivotal step for achieving the club’s goals. And similarly as before, it is notable that this positive impact is absent for high financial might teams. The parallelisms in findings and possible explanation schematas for team performance and cooperation are unsurprising as these two markers of team effectiveness themselves are highly correlated with each other. In my dataset the correlation of the variables team effectiveness and cooperation was 0.822.

Though it is not an overmuch studied relationship I theorized that pay dispersion does affect the aggressivity of a team. Testing for this direct effect my regression analysis showed a significant positive effect of pay dispersion on aggressivity, thus Hypothesis 3 was confirmed. The way it was modeled this finding means that the higher the dispersion in salaries the more explicit aggressive acts will be committed

[1] The term ’Big Six’ refers to the six most dominant clubs of the Premier League during and after the 2010s. In alphabetical order they are: Arsenal F. C., Chelsea F. C., Liverpool F. C., Manchester City F. C., Manchester United F. C. and Tottenham Hotspur F. C.

by the team members. My research investigated solely aggressive acts directed outwards from the team and towards the opponent (as fouls were used as proxy), therefore this positive association shows that within-team pay dispersion increase may result in enhanced between-teams aggression. The connection between aggression and pay dispersion might be explained by the increased within-team competition. Hays et al. (2022) showed that pay dispersion increases competitive behavior and there are studies linking competitiveness and aggressive behavior together, such as Dumblekar (2010), Krisnadewi and Soewarno (2020), and Schmierbach (2010), all showing positive effect. Albeit the Hypothesis 3 was confirmed I tested for the presence of an interaction effect with financial might as before, yet no such significant effect was found. Financial might itself however did have a significant negative effect on aggressivity, which may not be as surprising as most fouls are committed whilst defending, therefore teams who spend more time attacking and making plays will naturally commit less fouls.

Considering studies that examined pay dispersion within sport settings, my research contradicts Halevy et al. (2012) (basketball) and Hill et al. (2017) (baseball) as both showed positive relationship between pay dispersion and team performance. As Hill et al. investigated baseball teams, this may be attributed to the stark differences between a soccer and a baseball game, with baseball teams relying much less on cooperation, and baseball being overall a sport where outstanding individual performances might amount to great sporting success, whereas in soccer not so much or far more rarely. The divergence from Halevy et al.'s study is not as straightforward to explain, yet it might be due to the different team sizes (5 for basketball vs 11 for soccer), and overall smaller pitch, which results in tighter interplay between team members, where a hierarchical figure might exert greater influence upon the team, and coordination benefits are easier to reap. Yet again, my knowledge of the game of basketball is not sufficiently deep to wholly explain this discrepancy. On the contrary my research did confirm, leastwise for the high financial might teams, the results of Mondello and Maxcy (2009) (football), as they argued for pay dispersion diminishing team performance.

For studies analyzing soccer teams, Franck and Nüesch (2011) showed a U-formed relationship between pay dispersion and team performance. Although the corresponding tests were attempted, I did not find proof of such a U-formed or other nonlinear effect in my data. The results of the study of Bucciol et al. (2014) were to some extent confirmed however, as they predicted a negative relationship between pay dispersion and team performance, although only for the part of the team active on the pitch. Bucciol et al. (2014) moreover argued that the decreased performance is not due to decreased cooperation, which my analysis, again partially, supports, as for the low financial might teams pay dispersion did indeed increase cooperation, and for the high financial might teams it did not decrease it significantly. The results of my thesis also support the findings of Di Domizio et al. (2022), as they showed that the financially mightier teams perform better in terms of seasonal success. They derived this from using relative wages to control for financial power, whereas I used the total revenue, which two measures do correlate, but under no circumstance can be understood as the same. Furthermore, Di Domizio et al. found a significant negative association between pay dispersion and team performance, which my thesis also determined, albeit only for the financially mightier teams.

A cardinal difference between my thesis and most papers investigating the effects of pay dispersion using sports data, was the chosen control variable financial might and how it was defined. Money is a force of great magnitude in professional sports, therefore to aptly control for it is paramount. As argued in the *Methodology* section I find the usage of market size, as in the population of a given team's hometown (see Di Domizio et al. (2022) or Hill et al. (2017)), suboptimal to say the least, as it vastly downplays the economic power of some of the biggest clubs (e. g. Manchester City or Liverpool). Employing average or total pay is a far more accurate approach (see Bucciol et al. (2014) or Franck and Nüesch (2011)), yet I believe neither that would be the optimum, as several expenditures may contribute to a team's success aside from those directed towards the players, such as medical teams, analysts or training facilities. Consequently, I employed total revenue or group turnover. This variable did indeed prove to be strongly significant throughout all the models, moreover it helped to shed light on the interaction effect between financial might and pay dispersion. The interaction effect, which may be accredited as the main contribution of my present thesis, and which according to my best knowledge, and as of writing this thesis, was not thus far examined.

At last considering the other control variables, their impact could be determined as mixed at best. Whilst average age and coach tenure were significant in some of the models, notably both of them in models estimating cooperation (original and expanded too), the other control variables were far less impactful. Average experience and roster size were not significant in any of the models, neither in those that tested for direct effect, nor in the ones that estimated interaction. This means that the experience the players have collected in the Premier League is not a significant determinant in how the team will fare in competition or how strongly the players will cooperate, moreover that the size of the roster is largely inconsequential for success and cooperation. Less surprising is, that neither of these factors affect aggressivity. Nonetheless this result propounds the question whether there is any merit in their inclusion as controls in further studies investigating pay dispersion in professional sports.

5.2. Practical implications

My research offers some important practical implications for managers. Based on my findings, an approach to salary distribution should most definitely take into consideration the overall economic or financial power standing behind a team, as the effects of pay dispersion do impact differently the team's performance given the magnitude of financial might. In organizations or teams of high financial power, pay dispersion adversely affects team performance, thus it should be opted for a more egalitarian pay structure. Whereas in a lower financial power setting, pay dispersion weakly, and not significantly, but improves performance. In other words, for instance within the very same organization, it might be prudent to differentiate in pay levels on the lower echelons of the organizational ladder, e. g. teams of low-level operational tasks, but at the same time prioritize even distribution of salaries within the upper management teams and executives. It may be beneficial to that wise communicate the possibility of hierarchical ascension at the lower levels of the organization, yet emphasize the notion among the highly ranked employees that they are valued nigh equally within the organization. Creating 'superstars' or star employees may as well function as a driving force in teams of lower hierarchical standing, but it will cause tension and conflict in teams or groups where everyone, or most, consider themselves as an illustrious member of the organization and a major contributor to its successful functioning and past achievements.

Similarly, pay dispersion exerts opposing influence on the cooperation of teams, once again given the surrounding financial might. Managers shall be mindful that among highly paid employees (i. e. teams of high financial might), dispersion of pay will not advance cooperation, it might even lightly hinder this type of endeavor. For teams of lesser wage however, differentiated pay might increase voluntary cooperation and cause overall greater individual effort to contribute. A further notable implication is, that pay dispersion and the resulting hierarchical differentiation may increase aggressivity and competitiveness, thus if organizations have to deal with issues caused by aggressive behavior of employees, the differences in pay might be significant contributor to it. Albeit pay dispersion does significantly increase aggressive behavior, it has to be noted, that it is by no means the only, or the primary determinant (given the somewhat low R^2 of the model, i. e. the portion of variance explained), and many other, and especially circumstantial, factors play a role.

5.3. Limitations and future research

At long last certain limitations of my research need to be addressed. First and foremost it has to be understood that the sport setting itself is a very unique environment, and the findings of my thesis shall not be heedlessly adapted. Generalizability is ever the question for research relying on sports data, and not all sports teams are apt models for all organizations or teams. A profound understanding of the sport and the structure of its competing teams is needed in order to truly make use of the implications, for the better of the organization. Here, I recommend reading the paper of Keidel (1987), where he provides guidance for applying knowledge collected from sports in organizational setting. Soccer may be substituted for basketball in Keidel's framework, due to the similarities of the nature of the two games (Keidel, 1987, p. 592).

Using sports data in managerial research is akin to a large scale field experiment, where there are clear rules, boundaries and controls. It delivers compelling and rich data, but it does not perfectly model life. Examining soccer teams shows some peculiarities, which set this environment apart from most business organizations. For instance the salaries are absurdly high compared to normal organizations. Moreover, the players are under near complete public exposure, which rarely happens in the world of business. In my research, the teams I investigated showed no diversity in terms of sex. From one side, conducting a research to test for the robustness of the findings whilst collecting data from female teams should not be overly difficult. Quite difficult is however to find mixed teams, which do not exist in high level professional sports, yet are quite prevalent in organizations.

Similarly, the diversity of age was much smaller in my dataset compared to normal organizations, and individuals far too young to be commonly employed did have a lasting impact on the outcomes of certain teams (e. g. 19 year old Bukayo Saka was a major contributor to Arsenal's 2021/22 season, playing in all 38 games (scoring 11 goals), and so was 21 year old Phil Foden for Manchester City, playing 28 games in that season (scoring 9 goals)). Given the smaller degree of age diversity and the overall low average age of soccer teams, the implications of my research may be even more applicable to start-ups, than to mature organizations. Testing my findings in the setting of start-up organizations might be a promising direction for future research.

Some shortcomings of the collected dataset were revealed whilst inspecting the assumptions of OLS regression. The RESET test was significant for aggressivity, which means the functional values for that model were not perfectly defined. Likely there is a nonlinear relationship between pay dispersion and aggressivity. Models estimating this relationship should try for transforming the independent variable, pay dispersion. Notably however, the squared, cubical and log-transformed versions of the variable did not bring sufficient results, and the RESET test remained significant. Therefore other mathematical transformations may be needed to lift this burden.

Another weakness of the dataset was the presence of autocorrelation. The Wooldridge test was significant for the models of cooperation and aggressivity. To tackle this issue, which is quite difficult due to the panel data being consisted of several short time series (spanning from one to five observations), nesting the data on a seasonal level might offer help. Alternatively using the generalized least squares (GLS) method can prove to be fruitful, as GLS is less prone to be biased by autocorrelation than OLS.

It would be highly interesting to unearth the psychological factors standing behind the interaction effects discovered in my thesis. I theorize that in smaller clubs the players view much more favorably 'superstars', as they recognize the need for their presence, as opposed to bigger clubs where elevating the status of some players begets a negative echo among others. This, however is most difficult to prove by merely relying on quantitative data. For future research, I recommend the collection of qualitative data by conducting deep interviews with players, where they are asked to share their perceptions, regarding the presence of 'superstars' and what it means for their clubs. Incorporating a mixed methods approach in our analysis could advance the understanding of these complex processes.

Regarding my models for aggressivity I do believe that the number of fouls is the best accessible proxy, however it should be viewed somewhat askance. Umphress et al. (2010) coined the term unethical pro-organizational misconduct, where they refer to a specific type of misconduct, one committed intentionally and in order to benefit the organization. Unethical pro-organizational behavior is highly prevalent in professional soccer and its most common form is tactical, or professional, fouls. This is where the line becomes a trifle blurry. Most soccer fans vividly remember the finals of Euro 2020, where, in minute 90+6 Giorgio Chiellini (ITA) grabbed Bukayo Saka's (ENG) shirt lugging him to the ground, thus halting a dangerous outbreak of the striker and sending the game into extra time (after which Italy won the game and the trophy in penalty shootout). This foul hardly seemed an act of uncontrolled aggression resulting from the within-team tension and much more cold calculation, where the benefits apparently outweighed the risks for the perpetrator. However on the contrary, even tactical fouls may indeed result from increased tension and competitiveness within the team, as players sense a higher pressure to perform and advance their teams. To differentiate between acts of aggression and unethical pro-organizational behavior without context is highly difficult, thus the one solution I could recommend, would be closely observing soccer games and evaluating the actions as they unfold.

Further research could examine different facets of hierarchical differentiation, alone or in combination with pay dispersion, on team performance. Seniority on the team, and even position on the field may be a crucial determinant of the social rank order. Moreover, it could be interesting to attempt to replicate my findings, particularly regarding the interaction of financial might and pay dispersion, in other sports with teams of different structure, such as (American) football or baseball.

In conclusion, my findings support the notion that hierarchical differentiation, and pay dispersion in particular, play a pivotal role in the effectiveness of teams. However, my analysis also proved that for sporting success the effect of pay dispersion is contingent upon the financial power standing behind the teams. Similarly, the interplay of financial might and pay dispersion is a strong determinant of within-team cooperation, yet, the overall aggressivity of a team they influence independently and contrarily. I hope for my thesis to inspire future works utilizing soccer data in managerial research.

References

Akerlof, G. A., & Yellen, J. L. (1990). The Fair Wage-Effort Hypothesis and Unemployment. *The Quarterly Journal of Economics*, *105*(2), 255–283. https://doi.org/10.2307/2937787

Bedeian, A. G., & Mossholder, K. W. (2000). On the Use of the Coefficient of Variation as a Measure of Diversity. *Organizational Research Methods*, *3*(3), 285–297. https://doi.org/10.1177/109442810033005

Biemann, T., & Kearney, E. (2010). Size Does Matter: How Varying Group Sizes in a Sample Affect the Most Common Measures of Group Diversity. *Organizational Research Methods*, *13*(3), 582–599. https://doi.org/10.1177/1094428109338875

Bridges, E. M., Doyle, W. J., & Mahan, D. J. (1968). Effects of Hierarchical Differentiation on Group Productivity, Efficiency, and Risk Taking. *Administrative Science Quarterly*, *13*(2), 305–319. https://doi.org/10.2307/2391457

Bucciol, A., Foss, N. J., & Piovesan, M. (2014). Pay Dispersion and Performance in Teams. *PLOS ONE*, *9*(11), e112631. https://doi.org/10.1371/journal.pone.0112631

Cureton, E. E. (1968). The Teacher's Corner: Unbiased Estimation of the Standard Deviation. *The American Statistician*, *22*(1), 22–22. https://doi.org/10.1080/00031305.1968.10480435

Damgaard, C., & Weiner, J. (2000). Describing Inequality in Plant Size or Fecundity. *Ecology*, *81*(4), 1139–1142. https://doi.org/10.1890/0012-9658(2000)081[1139:DIIPSO]2.0.CO;2

Dawson, J. F. (n.d.). Interpreting Interaction Effects. Retrieved July 20, 2023, from http://www.jeremydawson.com/slopes.htm

Dawson, J. F. (2014). Moderation in Management Research: What, Why, When, and How. *Journal of Business and Psychology*, *29*(1), 1–19. https://doi.org/10.1007/s10869-013-9308-7

Day, D. V., Gordon, S., & Fink, C. (2012). The Sporting Life: Exploring Organizations through the Lens of Sport. *Academy of Management Annals*, *6*(1), 397–433. https://doi.org/10.5465/19416520.2012.678697

Depken, C. A., & Lureman, J. (2018). Wage Disparity, Team Performance, and the 2005 NHL Collective Bargaining Agreement. *Contemporary Economic Policy*, *36*(1), 192–199. https://doi.org/10.1111/coep.12220

Devine, D. J., & Philips, J. L. (2001). Do Smarter Teams Do Better: A Meta-Analysis of Cognitive Ability and Team Performance. *Small Group Research*, *32*(5), 507–532. https://doi.org/10.1177/104649640103200501

Di Betta, P., & Amenta, C. (2010). A Die-Hard Aristocracy: Competitive Balance in Italian Soccer 1929-2009. *Rivista di Diritto ed Economia dello Sport*, *6*, 13–40.

Di Domizio, M., Bellavite Pellegrini, C., & Caruso, R. (2022). Payroll Dispersion and Performance in Soccer: A Seasonal Perspective Analysis for Italian Serie A (2007–2021). *Contemporary Economic Policy*, *40*(3), 513–525. https://doi.org/10.1111/coep.12566

Dimas, I. D., Torres, P., Rebelo, T., & Lourenço, P. R. (2023). Paths to Team Success: A Configurational Analysis of Team Effectiveness. *Human Performance*, 1–25. https://doi.org/10.1080/08959285.2023.2222272

Drukker, D. M. (2003). Testing for Serial Correlation in Linear Panel-data Models. *The Stata Journal*, *3*(2), 168–177. https://doi.org/10.1177/1536867X0300300206

Dumblekar, D. V. (2010). Interpersonal Competitiveness - A Study Of Simulation Game Participants' Behaviour. *Paradigm*, *14*(2), 13–19. https://doi.org/10.1177/0971890720100203

Fonti, F., Ross, J.-M., & Aversa, P. (2023). Using Sports Data to Advance Management Research: A Review and a Guide for Future Studies. *Journal of Management*, *49*(1), 325–362. https://doi.org/10.1177/01492063221117525

Franck, E., & Nüesch, S. (2008). Mechanisms of Superstar Formation in German Soccer: Empirical Evidence. *European Sport Management Quarterly*, *8*(2), 145–164. https://doi.org/10.1080/16184740802024450

Franck, E., & Nüesch, S. (2011). The Effect of Wage Dispersion on Team Outcome and the Way Team Outcome is Produced. *Applied Economics*, *43*(23), 3037–3049. https://doi.org/10.1080/00036840903427224

Frick, B. (2011). Performance, Salaries and Contract Length: Empirical Evidence from German Soccer. *International Journal of Sport Finance*, *6*(2), 87–118. https://EconPapers.repec.org/RePEc:jsf:intjsf:v:6:y:2011:i:2:p:87-118

Gamson, W. A., & Scotch, N. A. (1964). Scapegoating in Baseball. *American Journal of Sociology*, *70*(1), 69–72. https://doi.org/10.1086/223739

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration*. Polity.

gov.uk. (n.d.). Retrieved July 20, 2023, from https://www.gov.uk/

Greer, L. L., de Jong, B. A., Schouten, M. E., & Dannals, J. E. (2018). Why and When Hierarchy Impacts Team Effectiveness: A Meta-Analytic Integration. *Journal of Applied Psychology*, *103*(6), 591–613. https://doi.org/10.1037/apl0000291

Halevy, N., Chou, E. Y., & Galinsky, A. D. (2011). A Functional Model of Hierarchy: Why, How, and When Vertical Differentiation Enhances Group Performance. *Organizational Psychology Review*, *1*(1), 32–52. https://doi.org/10.1177/2041386610380991

Halevy, N., Chou, E. Y., Galinsky, A. D., & Murnighan, J. K. (2012). When Hierarchy Wins: Evidence From the National Basketball Association. *Social Psychological and Personality Science*, *3*(4), 398–406. https://doi.org/10.1177/1948550611424225

Harrison, D. A., & Klein, K. J. (2007). What's the Difference? Diversity Constructs as Separation, Variety, or Disparity in Organizations. *The Academy of Management Review*, *32*(4), 1199–1228. https://doi.org/10.2307/20159363

Hays, N. A., Li, H., Yang, X., Oh, J. K., Yu, A., Chen, Y.-R., Hollenbeck, J. R., & Jamieson, B. B. (2022). A Tale of Two Hierarchies: Interactive Effects of Power Differentiation and Status Differentiation on Team Performance. *Organization Science*, *33*(6), 2085–2105. https://doi.org/10.1287/orsc.2021.1540

Hill, A. D., Aime, F., & Ridge, J. W. (2017). The Performance Implications of Resource and Pay Dispersion: The Case of Major League Baseball. *Strategic Management Journal*, *38*(9), 1935–1947. https://doi.org/10.1002/smj.2616

Kampkötter, P., & Sliwka, D. (2018). More Dispersion, Higher Bonuses? On Differentiation in Subjective Performance Evaluations. *Journal of Labor Economics*, *36*(2), 511–549. https://doi.org/10.1086/694588

Keidel, R. W. (1984). Baseball, Football, and Basketball: Models for Business. *Organizational Dynamics*, *12*(3), 5–18. https://doi.org/10.1016/0090-2616(84)90021-4

Keidel, R. W. (1987). Team Sports Models as a Generic Organizational Framework. *Human Relations*, *40*(9), 591–612. https://doi.org/10.1177/001872678704000904

Krisnadewi, K. A., & Soewarno, N. (2020). Competitiveness and Cost Behaviour: Evidence from the Retail Industry. *Journal of Applied Accounting Research*, *21*(1), 125–141. https://doi.org/10.1108/JAAR-08-2018-0120

Lazear, E. P., & Rosen, S. (1981). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy*, *89*(5), 841–864. http://www.jstor.org/stable/1830810

Lepschy, H., Wäsche, H., & Woll, A. (2020). Success Factors in Football: An Analysis of the German Bundesliga. *International Journal of Performance Analysis in Sport*, *20*(2), 150–164. https://doi.org/10.1080/24748668.2020.1726157

Magee, J. C., & Galinsky, A. D. (2008). Social Hierarchy: The Self-Reinforcing Nature of Power and Status. *The Academy of Management Annals*, *2*(1), 351–398. https://doi.org/10.1080/19416520802211628

Manchester City. (2022). Pep Guardiola: Five-year anniversary - 13 memorable Pep quotes... Retrieved July 20, 2023, from https://www.mancity.com/features/ipep-quotes/

Mills, B., & Winfree, J. (2018). Athlete Pay and Competitive Balance in College Athletics. *Review of Industrial Organization*, *52*(2), 211–229. https://doi.org/10.1007/s11151-017-9606-8

Mondello, M., & Maxcy, J. (2009). The Impact of Salary Dispersion and Performance Bonuses in NFL Organizations. *Management Decision*, *47*(1), 110–123. https://doi.org/10.1108/00251740910929731

Premier League. (n.d.). Retrieved July 20, 2023, from https://www.premierleague.com/

Ramaswamy, R., & Rowthorn, R. E. (1991). Efficiency Wages and Wage Dispersion. *Economica*, *58*(232), 501–514. https://doi.org/10.2307/2554695

Ramchandani, G. (2012). Competitiveness of the English Premier League (1992-2010) and Ten European Football Leagues (2010). *International Journal of Performance Analysis in Sport*, *12*(2), 346–360. https://doi.org/10.1080/24748668.2012.11868603

Ronay, R., Greenaway, K., Anicich, E. M., & Galinsky, A. D. (2012). The Path to Glory Is Paved With Hierarchy: When Hierarchical Differentiation Increases Group Effectiveness. *Psychological Science*, *23*(6), 669–677. http://www.jstor.org/stable/41489754

Schmierbach, M. (2010). Killing Spree: Exploring the Connection Between Competitive Game Play and Aggressive Cognition. *Communication Research*, *37*(2), 256–274. https://doi.org/10.1177/0093650209356394

Soebbing, B. P., Wicker, P., & Watanabe, N. M. (2022). NFL Player Career Earnings and Off-Field Behavior. *The Review of Black Political Economy*, *50*(1), 81–96. https://doi.org/10.1177/00346446221076868

Spotrac. (n.d.). Retrieved July 20, 2023, from https://www.spotrac.com/

To, C., Yan, T., & Sherf, E. N. (2022). Victorious and Hierarchical: Past Performance as a Determinant of Team Hierarchical Differentiation. *Organization Science*, *33*(6), 2346–2363. https://doi.org/10.1287/orsc.2021.1528

Torgler, B., & Schmidt, S. L. (2007). What Shapes Player Performance in Soccer? Empirical Findings from a Panel Analysis. *Applied Economics*, *39*(18), 2355–2369. https://doi.org/10.1080/00036840600660739

Transfermarkt. (2023). European Leagues & Cups. Retrieved July 20, 2023, from https://www.transfermarkt.com/wettbewerbe/europa

Umphress, E. E., Bingham, J. B., & Mitchell, M. S. (2010). Unethical Behavior in the Name of the Company: The Moderating Effect of Organizational Identification and Positive Reciprocity Beliefs on Unethical Pro-Organizational Behavior. *Journal of Applied Psychology*, *95*(4), 769–780. https://doi.org/10.1037/a0019214

Volkova, V. M., & Pankina, V. L. (2013). The Research of Distribution of the Ramsey RESET-Test Statistic. *Applied Methods of Statistical Analysis*, 265–267. https://www.amsa.conf.nstu.ru/amsa2013/AMSA2013_proceedings.pdf

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. http://www.jstor.org/stable/j.ctt5hhcfr

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Waiting Time Estimation for Ride-Hailing Fleets Using Graph Neural Networks

Hashmatullah Sadid

Technical University of Munich

Abstract

Ride-hailing services are part of intermodal transport systems, allowing passengers to use various transport modes for their trip. The optimal choice for a request in the intermodal system depends on the passenger's waiting time for the ride-hailing service. Estimating this waiting time is crucial for efficient system operation. The prediction of waiting time depends on the spatial dependency of the transport network and traffic flow elements. Graph neural network (GNN) approaches have gained attention for capturing spatial dependencies in various applications, though less attention has been given to ride-hailing waiting time prediction. The aim of this master thesis is to implement a GNN-based method to predict waiting time for ride-hailing requests in the network. Simulation-based waiting time data is used for model training and validation. MATSim is chosen for generating waiting time data under different demand and supply scenarios. Graph Convolutional Network (GCN) and Gated Attention Network (GAT) are used as prediction models. Regression and MLP methods are used as baselines to compare model performance. Results show GCN outperforms regression by 15%, while GAT performs 14% better than regression.

Keywords: graph convolutional network; ride-hailing service; waiting time estimation

1. Introduction

1.1. Background and motivation

The wide implementation of smartphone-based on-demand mobility services has already started changing the transport pattern of many cities. Ride-hailing platforms such as Uber, Lyft, Cabify, or Didi provide location-based and door-to-door taxi-like services by connecting riders and drivers in a centralized automated manner. These services provide the opportunity for passengers to order a customized ride using a smartphone application (Anderson, 2014; Henao & Marshall, 2019; Xu et al., 2020; C. Yan et al., 2020; Zha et al., 2016). Ride-hailing technologies create more business opportunities even for individuals with a car capable to offer taxi-like services (de Souza Silva et al., 2018; Lee et al., 2018). On the other hand, the complex dispatching algorithms used in these systems enable efficient matching of the customers and drivers considering both their spatial and temporal distributions. This leads to a significant reduction in search frictions in on-demand mobility services, as well as boosting the ordering and payment processes. As a result, lower costs are incurred both for riders and drivers (Anderson, 2014; Lee et al., 2018; Zha et al., 2016).

The successful implementation of a ride-hailing service depends on both customers' expectations and satisfaction as well as the suppliers' advantage. From a customer perspective, the price of the service, waiting time for the vehicle, reliability of the system, trip comfort, travel time among others are the key important perceived values (Gilibert & Ribas, 2019). Whereas for the suppliers, the accurate estimation of the demand and the recognition of potential service areas where competitors do not provide high-quality services (e.g. long passengers' waiting time) are the crucial decision elements in their business models. Among other influential factors, waiting time plays a vital role in generating ride-hailing demand and thus makes it an important factor for suppliers to find their optimal service areas.

Meanwhile, ride-hailing services are part of intermodal-transport systems, allowing passengers to use various transport modes for their entire trip. The optimal choice for a request in the intermodal-transport system, therefore, depends on the waiting time of a passenger until served by the ride-hailing service. Hence, estimating this waiting time is a crucial element for the efficient operation of the system. The waiting time of a passenger is the exact time that a passen-

DOI: https://doi.org/10.5282/jums/v10i2pp462-490

ger needs to wait in the mobility service platform until picked up.

Prediction and estimation of the waiting time of a customer for a ride-hailing service depend both on the spatial dependency of the transport network and traffic flow elements of the network. The spatial information includes the road network, demand and supply of the system, request location, and more, where the traffic flow elements demonstrate the variation of traffic flow variables and their relation to the waiting time in different time intervals. Thus, to predict and estimate the waiting time considering these influential factors, it is important to find an correlation among them. This is typically done by traditional statistical approaches or model-based methods. Statistical methods require more powerful feature engineering and assumptions (often leads to inaccurate estimations), where model-based (simulation tools) approaches require detailed modelling efforts and need high computational resources. Hence, machine learning-based approaches especially deep learning techniques have been widely used to improve prediction accuracy and have been utilized in many fields including traffic forecasting (C. Chen et al., 2019; K. Chen et al., 2021; Fang et al., 2020; Jin, Yan, et al., 2021; Q. Wang et al., 2021; Zhang et al., 2021).

Deep learning methods learn multiple layers of features by extracting more complex non-linear relationships. A convolutional neural network (CNN) for instance has been proved to reflect the spatial features of the network by modelling the whole city as a grid (Jiang & Zhang, 2019). However, due to non-Euclidean structure of the road network, this method is not optimal (Bronstein et al., 2017; Jiang & Luo, 2021; Z. Wu et al., 2021). Therefore, Graph neural networks (GNNs) have attracted attention for traffic forecasting problems, as they could capture spatial dependencies of a road network as a graph (i.e., intersections as nodes of the graph and road connections as edges of the graph) (C. Chen et al., 2019; Fang et al., 2020; Jiang & Luo, 2021; Q. Wang et al., 2021). For instance, Jin et al., 2022 and Jin, Yan, et al., 2021 used spatio-temporal GNN to estimate the travel time using real-world datasets, where X. Wang et al., 2020 employed spatio-temporal GNN for traffic flow prediction. Despite many studies in traffic forecasting problems (Fang et al., 2020; Q. Wang et al., 2021; Zhang et al., 2021), less attentions have been made to estimate the ride-hailing waiting time in a service area using deep learning approaches. Thus, it is imperative to design a deep learning method to predict ride-hailing waiting time in a service area considering spatial and operational features of network and traffic flow elements.

Meanwhile, deep learning algorithms require a large amount of data for training, testing, and validation. However, ride-hailing-related data, especially the waiting time information are not publicly available or cost-deficient for academic purposes. Thus it is advantageous to use simulation-based approaches to extract waiting time data in a service area.

1.2. Research Objective

The main goal of this master thesis is to implement a GNN-based method to predict the waiting time of a ride-hailing request in the transport network. This study employs a traffic simulation tool (i.e., PTV Vissim, SUMO, MATSim, etc.) to model traffic network features and ride-hailing service scenarios aiming to extract waiting time data for GNN implementation.

The following sub-aims are also included as a licentiate part of this work:

1. A literature review on the basics of neural networks, the theory of GNNs, different types of GNNs, and their applications, and
2. Developing a simulation-based platform for extracting ride-hailing waiting time data.

1.3. Structure of the thesis

This thesis is divided into six chapters. Chapter 1 introduces the background, motivation, and objectives of the study. Chapter 2 gives an overview of topics related to this thesis. First, an overview of ride-hailing simulation methods is presented. Subsequently, the basic theory of neural networks, and GNNs together with different types of GNNs are outlined.

In chapter 3, the methodology of this master thesis which contains the waiting time extraction approach and the proposed GNN framework which is the main contribution of this master thesis are introduced.

In chapter 4, an experimental setup is designed to generate waiting time data for the Cottbus city network, and implement it in the proposed models. Chapter 5 presents the main findings of the experiment together with a sensitivity analysis.

Chapter 6 gives a summary of the main contributions of this thesis, followed by a conclusion of the study and providing outlook for future researches.

2. Literature Review

In this chapter, we review related topics to this master thesis. First, an overview of on-demand mobility simulation including taxi modelling in MATSim is presented. Second, we introduce the basics of neural networks, GNN, different types of GNN, as well as their applications.

2.1. Ride-hailing Simulation

Waiting time is an important indicator for the efficient implementation of on-demand mobility concepts both from the demand and supply perspectives. Customers prefer choosing an on-demand mobility service (e.g., a taxi) if the waiting time is reasonable. On the supply side, the service providers attempt to place the taxis and dispatch them in the area where existing suppliers do not meet the customers' expectations (long waiting times). Since waiting time data is a core advantage of a service provider, this data is not shared openly. Meanwhile, accessing such data for academic purposes is cost-deficient, thus, an efficient approach is to generate waiting time data using traffic simulation models.

There are a variety of traffic models (macro-, meso-, and microscopic) to simulate and evaluate the implementation of a mobility concept (H. U. Ahmed et al., 2021; Grau & Romeu, 2015; Jing et al., 2020). However, selecting an appropriate traffic simulation tool depends on whether it is free to use, and include multi-mode simulation. In addition, the tool should be implemented in large-scale network, and be computationally efficient. Hence, in this study, we select an agent-based simulation tool (MATSim) with functionality to model different on-demand mobility concepts (Horni et al., 2016).

2.1.1. Agent-based Modelling

In this study, we use the Multi-Agent Transport Simulation (MATSim) tool to model the on-demand mobility service and extract waiting times. MATSim is an agent-based modelling framework with an iterative, co-evolutionary learning approach, where each agent attempts to maximize their daily utility by selecting from the set of planned activities. Agents receive rewards (positive scores) for performing scheduled activities and penalties (negative rewards) for long traveling or late arriving to an activity. After each iteration, agents score their executed plans with a resulting score. By learning from the experience, agents try to either adjust their plans (e.g., choosing a new route, selecting another mode) or choose among the best plans based on their scores (Horni et al., 2016).

In MATSim, a simulation run requires three input files namely: the configuration file, the population file, and the network file. The configuration file is the core input that assigns the simulation settings under which the simulation is carried out (HÖrl, 2017). The population file describes the daily plans of each agent by providing information about the socio-demographic attributes as well as their daily trips. The network file contains the links and nodes of the study area including specific information about the road network (type, capacity, mode, etc.).

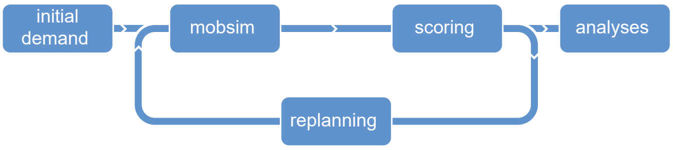


Figure 1: MATSim execution loop, source: (Horni et al., 2016)

The typical execution loop of MATSim is shown in Figure 1. The initial demand (population file) is inserted into the loop, where the simulation, scoring, and replanning of agents' plans are executed. In the first stage, all agents are moved along a physical network using Mobility Simulation (MobSim) unit. MobSim utilizes a spatial queue-based approach for traffic simulation without considering the microscopic driving behavior (car following and lane changing behaviors). In the scoring module, each agent's plan receives a reward using a utility function (e.g., performing an activity means a positive score, traveling gets negative rewards, etc.). Finally, the replanning module enables some agents to either adjust their existing plans or choose among the best plans (Horni et al., 2016). An example of the scoring scheme is shown in Figure 2.

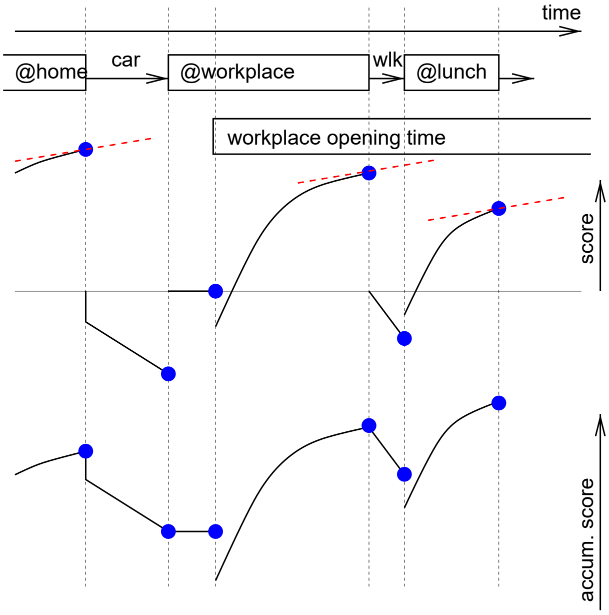


Figure 2: Scoring function representation, source: (Horni et al., 2016)

2.1.2. Taxi Modelling in MATSim

MATSim provides several extensions to model on-demand mobility services (Bischoff et al., 2017; Ruch et al., 2018, 2021). For taxi modelling, we utilize the dynamic vehicle routing problem (DVRP) extension integrated into MATSim by Maciejewski and Nagel, 2012. When a request with a coordinate and time is sent into the dispatching algorithm, the algorithm searches for a vehicle that can serve the request within a defined maximum waiting time (for more details on how the dispatching algorithm functions, readers are referred to (Maciejewski & Nagel, 2012)). To run a simple taxi simulation in MATSim, despite three main files (config, population, and network), we require the taxi location file to be added into the simulation framework. Worth mentioning that the taxi demand is defined within the population file, by modifying the mode choice of agents to taxis. The number of taxis in the network to serve a specific mode share of taxi demand is selected based on the minimum number of empty taxis during the peak hour demand.

When a request is submitted into the dispatching system, the dispatching algorithm assigns a taxi to this request. Once the taxi reaches a passenger and picks up the passenger, the taxi trip begins. The difference between the time a request is submitted and picked up is defined as the passenger waiting time. In this thesis, we try to extract the taxi waiting time for all agents under different demand scenarios.

2.2. Artificial Neural Networks (ANNs)

2.2.1. The basics of neural networks

Neural networks are a set of algorithms inspired by the human brain to mimic the relations and patterns from data (Maind & Wankar, 2014). Artificial neurons are the basic units of the neural network and are based on the Perceptron (Rosenblatt, 1957). The structure of an artificial neuron is displayed in Figure 3. Similar to biological neurons, a simple processing unit receives the input information and generates an output depending on the inputs. Overall, the information processing is conducted in five steps: First, the processing unit obtains the information as inputs $h_1, h_2, h_3, ..., h_n$. Second, each input is weighted by its corresponding weight denoted as $w_1, w_2, w_3, ..., w_n$. Third, a bias term b is added to the sum of all weighted inputs. Fourth, a non-linear activation function is applied and finally, the output y is generated. Mathematically a neuron output is expressed as:

$$y_j = \sigma\left(\sum_{i=1}^{n} h_i.w_i + b_j\right) \tag{1}$$

where y_j is the output of neuron, σ is the activation function, h_i is the input information, w_i is the corresponding learnable weight of neuron i, b_j is the bias term, and n is the number of inputs.

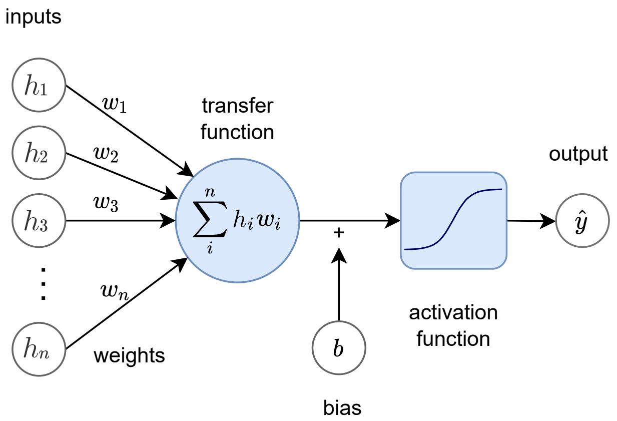


Figure 3: Illustration of an artificial neuron, perceptron. The input features are multiplied with the corresponding weights and the sum is added with a bias b, which then passes to an activation function. The outcome is the output value of the neuron.

The overall structure of neural networks consists of a number of interconnected neurons, which are arranged in at least three different layers namely: input layer, hidden layer(s), and output layer as shown in Figure 4. The input layer acts as an interface between the data and the network, and no calculation is done in this stage. The hidden layer(s) contains a predefined number of connected neurons, which transmit the information from the input layer to the neighboring hidden layers and then to the output layer. Finally, the output layer presents the results of the information process and depending on the type of prediction and the activation function may have different dimensions.

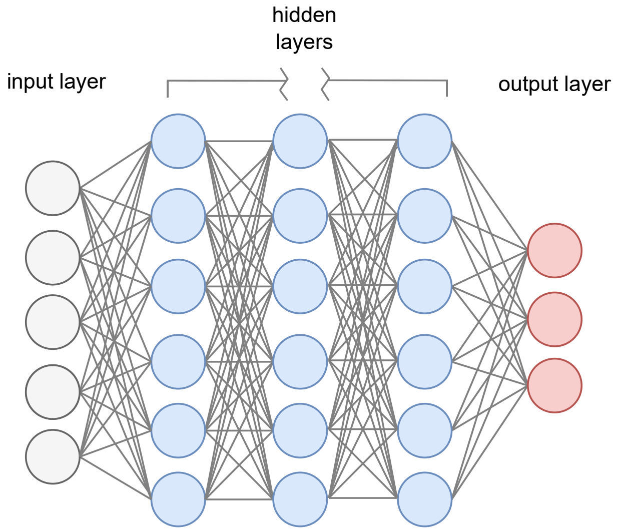


Figure 4: The structure of a fully connected 3-layer neural network. Source: (Saracoglu & Altural, 2010)

The connection between neurons in the hidden layer categorizes the neural networks into feed-forward and feedback networks. In a feed-forward neural network, information is fed in a forward direction from the input layer to hidden layers and finally to the output layer. The output of each neuron in the hidden layer is transmitted to the next neuron in the next layer. This continues until the information process reaches the output layer. On the other hand, in a feedback network, the information can be transmitted in both directions, from hidden layer l to hidden layers $l-1$ and $l+1$. In a feedback network, it is possible to create loops, where information can propagate continuously until an equilibrium condition is reached. Worth-mentioning that neural networks are widely used in many learning tasks including pattern recognition, classification, regression, signal processing, and more. In this section, we present multi-layer perceptron (MLP) in detail. MLP will be further used in this master thesis as a comparison method.

2.2.2. Multi-layer Perceptron (MLP)

MLP is a feed-forward artificial neural network that is comprised of connected layers namely: an input layer, one or more hidden layers, and an output layer (Bishop, 1995). In a fully-connected MLP, every neuron is linked to the consecutive neurons in the next layer. Depending on the prediction task, the output layer predicts and/or classifies the samples. In case, the MLP is used for classification purposes, the Softmax and its variants are utilized as the activation function in the output layer. For numerical prediction, however, Rectified Linear Unit (ReLU) is the most used activation function. The details of some activation functions are described at the end of this section. Formally, in a $(K+1)$ layers perceptron as depicted in Figure 5, where one input layer, (K) hidden layers, and one output layer are structured, the information processing of neuron (i) in layer (k) is expressed as:

$$y_i^k = \sigma\left(\sum_{j=1}^{m_{(k-1)}} y_j^{k-1}.w_{i,j}^k + b_{i,0}^k\right) \tag{2}$$

where (y_i^k) is the output of neuron i in layer k, σ is the activation function, m_l denotes the number of neurons stacked in hidden layer k, $w_{i,j}^k$ is the weight from the neuron j in layer $(k-1)$ to the neuron i in the layer (k), and $b_{i,0}^k$ is the bias for the neuron i in layer k.

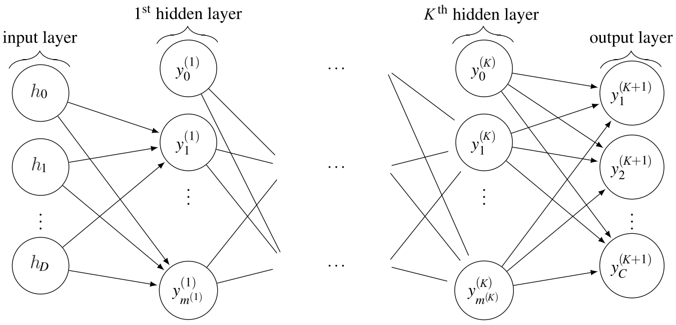


Figure 5: The structure of a fully connected MLP with K+1 layers, D inputs and C output neurons. Adopted from (Stutz, 2014)

The MLP model requires a training procedure to predict a realistic output. The training process takes advantage of the true values and/or labels of samples, and tries to minimize the difference between the predicted and true values using a loss function. Let's assume $\hat{y}$ as the MLP output and y as the true value (target), the backpropagation algorithm attempts to change the learnable weights w_{ij}, as well as the bias b in such a way to minimize the loss function L. Backpropagation algorithms are based on gradient descent and aim to optimize the model by converging the loss function to zero.

2.2.3. Activation Functions

Activation functions are the most important features of deep learning algorithms. In ANNs, the activation function of a neuron determines the output of that neuron with the given set of inputs, by simply mapping weighted inputs into an output. There are many activation functions in the literature, however, in this section, we present the widely used activation functions.

Sigmoid Function

The sigmoid function is a special form of logistic function which converts the model outputs into a probability score (between 0 and 1). As shown in Figure 6 (a), the sigmoid function generates a S-shaped curve and has a non-zero gradient throughout its domain. Hence, this function is a good candidate when applying the backpropagation algorithm.

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

Rectified Linear Unit (ReLU)

ReLU is a non-linear activation function that returns 0 if it receives any negative input, and returns the input value for any positive values (see Figure 6 (b)).

$$f(x) = \begin{cases} 0, & if \quad x < 0 \\ x, & else \end{cases} \tag{4}$$

Hyperbolic Tangent Function

The tangent hyperbolic function denoted as tanh maps the input values between -1 and +1. Similar to the Sigmoid function, it also creates a S-shape curve as shown in Figure 6 (c).

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5}$$

Heaviside Step Activation Function

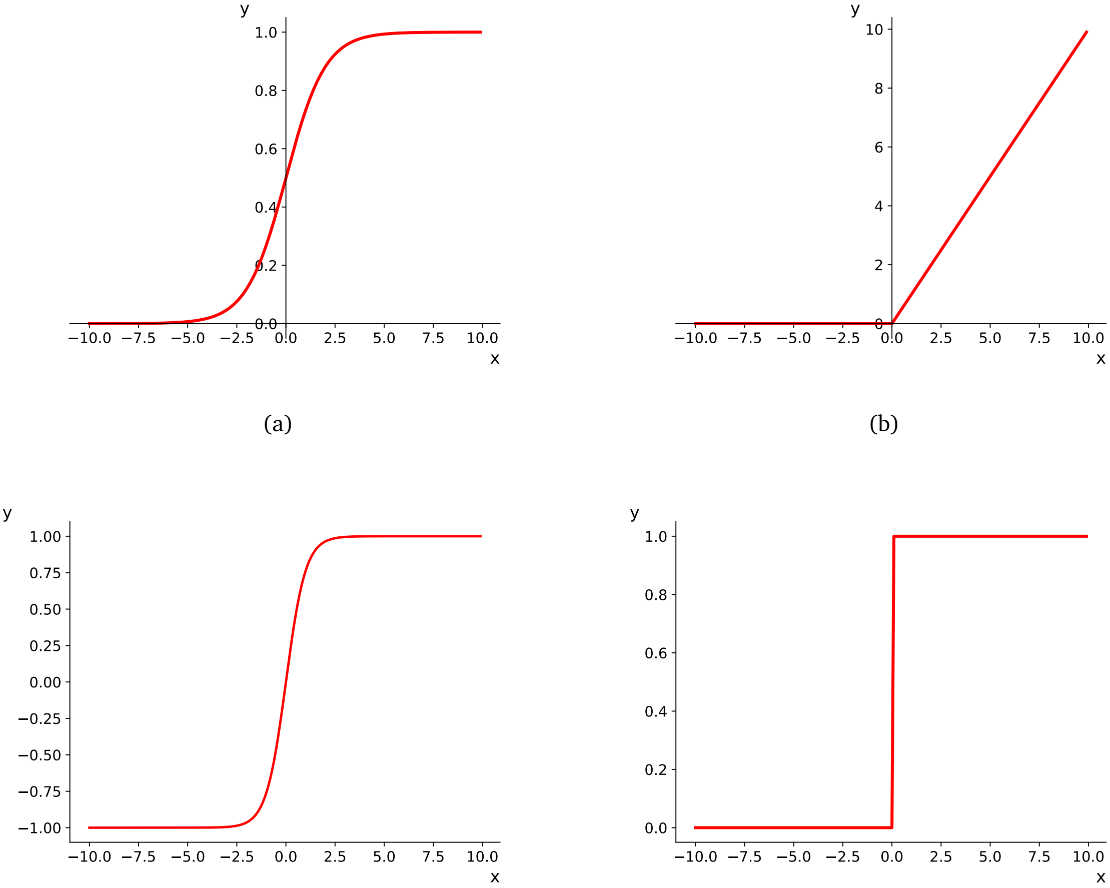
The Heaviside step activation function is a type of threshold function which generates the output 1 for positive inputs and 0 otherwise as displayed in Figure 6 (d).

$$f(x) = \begin{cases} 0, & if \quad x < 0 \\ 1, & if \quad x \geq 0 \end{cases} \tag{6}$$

2.3. Graph Neural Networks (GNNs)

Over the past decade, deep learning paradigms such as convolutional neural networks (CNNs) (Lecun & Bengio, 1995), recurrent neural networks (RNNs) (Hochreiter & Schmidhuber, 1997), and autoencoders (Vincent et al., 2010) have become trending topics in artificial intelligence and machine learning. The superior performance of deep learning in many domains such as object detection, machine translation, speech recognition, etc., is partially related to the recent advancement in computational resources, the availability of big data, and the power of deep learning algorithms in extracting latent representation from Euclidean data. Although deep learning algorithms could efficiently replicate the hidden patterns of Euclidean data, their utilization for applications where data are generated from non-Euclidean domains in form of graphs is challenging (Bronstein et al., 2017).

Graphs such as social networks (Y. Wu et al., 2020), biological and chemical networks (Fout et al., 2017), transport networks (K. Chen et al., 2021; Jin, Yan, et al., 2021; Q. Wang et al., 2021), and other applications (Dai et al., 2018) are complex data structures that have created important challenges on existing deep learning methods (Zhou et al., 2020). These challenges include the irregularity of graph structure,



(a)

(b)

(c)

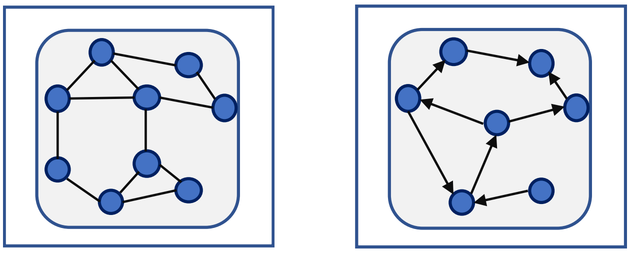
(d)

Figure 6: Plots of (a) Sigmoid, (b) ReLU, (c) tanh, and (d) Heaviside step activation functions

complex relationships, and interdependencies among the objects of a graph, as well as large-scale graphs (Jiang & Luo, 2021; Liu & Zhou, 2020; Z. Wu et al., 2021; Zhou et al., 2020). Hence, there is an increasing interest in extending the existing deep learning algorithms to mimic the graph data. Graph Neural Networks (GNNs) are widely used and the most successful in learning graphs data in various applications. In this section, we explore the basics of graph theory, GNNs, and different types of GNNs and their applications. This literature review aims to develop the methodology of this master thesis, which is described in the next chapter.

2.3.1. Graphs

Before introducing the GNNs, we briefly discuss the structure of a graph. A graph is a data structure that consists of two components namely vertices (nodes) and edges. A graph G can be defined as $G = (V, E)$, where V is the set of nodes, and E are the edges between them. The type of dependency between nodes categorizes the graphs into directed and undirected graphs. The information regarding the connection of nodes is often represented by adjacency matrix A. In a directed graph, all edges are directed from one node to another, whereas in an undirected graph, connected nodes are directed by a pair of edges with inverse directions (Z. Wu et al., 2021).



(a)

(b)

Figure 7: An illustration of (a) undirected, and (b) directed graphs

2.3.2. Graph Embeddings

Graph embedding is a method which is used to transform graph elements and their features into a low-dimensional space. The main goal of graph embedding is to map the original graph into a more computationally efficient format while keeping the original graph characteristics including the geometric relations and features of the graph (Gharaee et al., 2021; W. L. Hamilton, 2020; Hoff et al., 2002; Khoshraftar & An, 2022). Hence, the similarity in the embedding space approximates similarity in the graph or network.

The graph embedding could also be described using the encoder-decoder framework. An encoder model maps each graph element (i.e, nodes) into a low-dimensional vector or embedding. Whereas a decoder model takes the low-dimensional graph embeddings in the latent space and uses them to rebuild information about each node's neighborhood in the original graph. There are different embedding methods proposed in the literature. Khoshraftar and An, 2022 categorized graph embedding into traditional [e.g., Node2vec (Grover & Leskovec, 2016), Deepwalk (Perozzi et al., 2014), Graph factorization (A. Ahmed et al., 2013), Line (Tang et al., 2015), etc.] and GNN-based [e.g., RecGNN (Scarselli et al., 2009), GCN (Kipf & Welling, 2017), GraphSAGE (W. Hamilton et al., 2017), GCRN (Seo et al., 2016), DGCN (Zhuang & Ma, 2018), GAT (Veličković et al., 2018), and more] approaches. For more detailed information, readers are referred to (W. L. Hamilton, 2020; W. L. Hamilton et al., 2017; Khoshraftar & An, 2022).

2.3.3. The Concept of GNN

GNNs are deep learning methods operated for graph structure data. The basic idea of GNN is to iteratively update the representation of a node by aggregating its own representation and the representation of its neighboring nodes. GNNs use the representation of graph data including the node features and the connection between nodes. The output of a GNN model is the new representation of each node called embedding which contains the structural and feature information of other nodes. More specifically, each node knows about other nodes, the connection of nodes, and its context to the graph. The embeddings are further used for prediction.

The GNNs learn the representation vector of a node (h_v) by combining its own features and the neighboring node features (so-called message passing) with two important functions:

- **AGGREGATE**: The aggregate uses the states of all direct neighbors (u) of a node (v) and aggregates them in a specific method.
- **UPDATE**: The "update" operation uses the current state in time step (k) and combines it with aggregated neighbor states.

Within each message passing iteration in a GNN, a hidden embedding ($h_v^{(k)}$) related to each node $v \epsilon V$ is updated based on the information aggregated from the v node's neighbors ($N(v)$). The general framework of the GNN can be defined mathematically as follows:

$$\begin{aligned} {h_v}^{(k)} &= UPDATE({h_v}^{(k-1)}, \\ &\quad AGGREGATE\left(\left\{{h_u}^{(k-1)} : u \in N(v)\right\}\right)) \\ &= UPDATE^{(k-1)}\left({h_v}^{(k-1)}, {m_{N(v)}}^{(k)}\right) \end{aligned} \tag{7}$$

where $m_{N(v)}^{(k)}$ is the message that is aggregated from the v's neighborhood ($N(u)$).

More specifically, at each iteration k, the **AGGREGATE** function takes the information of each node (v) from its neighborhood $N(v)$ and generates a message $m(k)$. The embeddings of node (v) in iteration $k-1$ is combined with the message m(k) by the **UPDATE** function. The output of this process is the updated embeddings of node v ($h_v^{(k)}$). Worth-mentioning that at iteration $k = 0$, the initial embeddings of every node is basically the input features of all nodes ($h^{(0)} = x_u$). The final output after K iteration describes each node's embeddings in the embedding space.

Depending on the **AGGREGATE** and **UPDATE** operations, there are many variants of GNN models proposed in the literature. According to Z. Wu et al., 2021, GNNs are categorized into four groups namely: Recurrent GNN (RecGNN), Convolutional GNN (GCN), graph autoencoders (GAEs), and Spatio-temporal graph neural networks (STGNNs). In this thesis, we present the most relevant types of GNN including convolutional GNN, recurrent GNN, and graph attention networks. However, before presenting different types of GNNs, we briefly explain the basic **AGGREGATE** and **UPDATE** functions of GNN.

According to Merkwirth and Lengauer, 2005 and Scarselli et al., 2009 in the basic GNN, the **AGGREGATE** function contains trainable parameters which is defined as follows:

$${h_v}^{(k)} = \sigma(W_{self}^{(k)} {h_v}^{(k-1)} + W_{neigh}^{(k)} \sum_{v \in N(v)} {h_v}^{(k-1)} + b^{(k)}) \tag{8}$$

where $W_{self}^{(k)}, W_{neigh}^{(k)} \in \mathbb{R}^{d^{(k)} \times d^{(k-1)}}$ are parameters matrices and σ shows an element-wise non-linearity function such as tanh, ReLU, and $b^{(k)}$ is the bias term which is often eliminated for simplicity of the model.

In the basic GNN framework, the message passing is conducted similarly to a multi-layer perceptron (MLP), where linear operations are sent to a single element-wise non-linearity operation. More specifically, a linear combination of the sum of information received from the neighboring nodes and the previous embedding of a node itself is followed by a non-linear function.

Moreover, to omit the explicit update step and perform the message passing by only the aggregation function, the self-loops approach is proposed. In this approach the update function is defined through the aggregation method and the message passing is expressed as:

$${h_v}^{(k)} = AGGREGATE\left(\left\{{h_u}^{(k-1)}, \forall u \in N(v) \cup \{v\}\right\}\right) \tag{9}$$

Besides these two general methods, there are other generalized methods proposed for the AGGREGATE operator(i.e, neighborhood normalization). In the following subsections, we introduce different types of GNN mentioned in the above-section together with their aggregation methods.

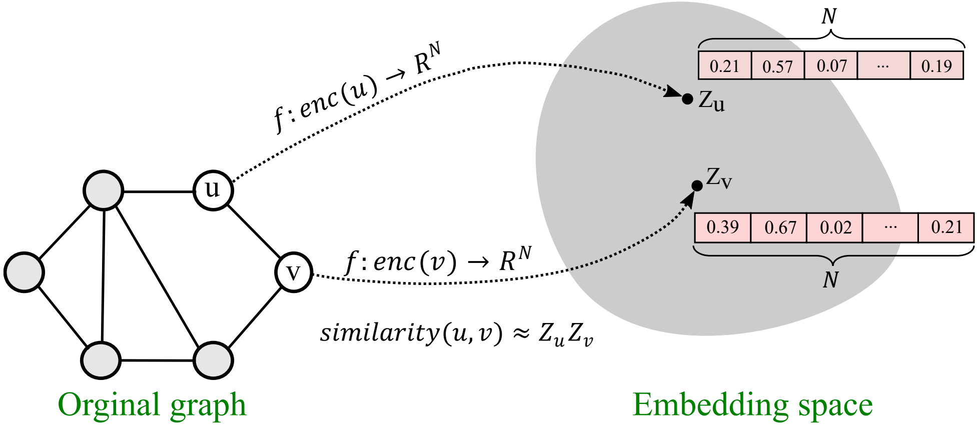

$f: enc(u) \rightarrow R^N$
$f: enc(v) \rightarrow R^N$
N
0.21 0.57 0.07 ··· 0.19
Zu
Zv
0.39 0.67 0.02 ··· 0.21
N
u
v
$similarity(u, v) \approx Z_u Z_v$
Orginal graph
Embedding space

Figure 8: Illustration of the node embedding of a graph with 6 nodes, 7 edges, and each node has N features.

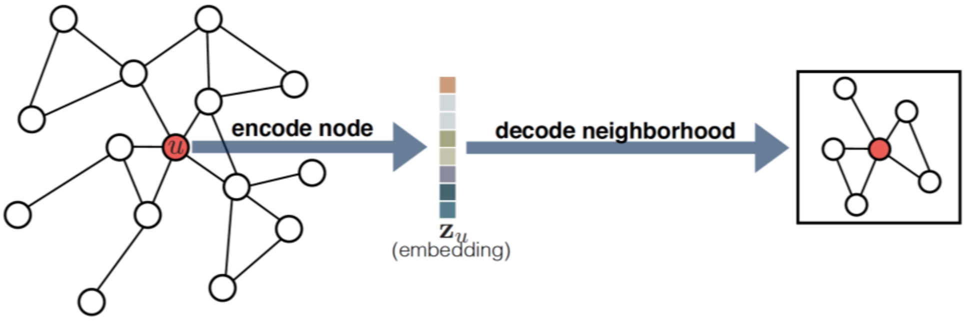


Figure 9: Overview of encoder-decoder framework. source: (W. L. Hamilton, 2020)

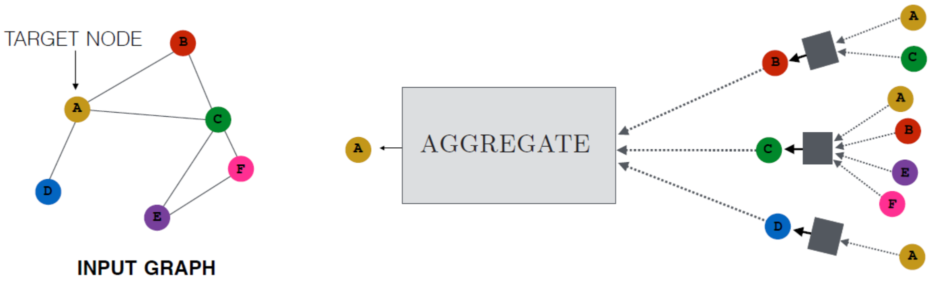


Figure 10: The information aggregation of a single node from its local neighborhood. source: (W. L. Hamilton, 2020)

Graph convolutional networks (GCNs)

GCN is the most popular GNN type which is extracted from the idea of the normal convolutional network. GCN can handle the cyclic mutual dependencies architecturally using a pre-defined number of layers with different weights in each layer. There are two types of GCN proposed in the literature namely spectral-based and spatial-based GCN. The first spectral-based method was proposed by Bruna et al., 2013. In this approach, graph convolution is defined by introducing filters from the view of graph signal processing. The information propagation in spectral GCN could be similar to signal propagation along the nodes. In spectral GCN, the convolution operation is defined in the Fourier domain by calculating the Eigen-decomposition of graph Laplacian matrix (for in-depth details, the reader is referred to (Z. Wu et al., 2021)).

On the other hand, spatial-based approaches consider the information propagation by operating on spatially close neighbors to define graph convolution. In the method proposed by Kipf and Welling, 2017, a symmetric-normalized aggregation with self-loop update operation is employed. The message-passing function of the GCN model is expressed as follows:

$$h_v{}^{(k)} = \sigma\left(W^{(k)} \sum_{u \in N(v) \cup \{v\}} \frac{h_u}{\sqrt{|N(v)|\,|N(u)|}}\right) \tag{10}$$

Worth-mentioning that the simplified spectral-method proposed by Kipf and Welling, 2017 could also be considered as spatial-based GCN. The details of this GCN method is presented in the methodology section of this master thesis.

Recurrent Graph Neural Networks (RGNN)

RGNN is a particular class of recurrent neural networks (RNNs) that is applied to sequential data. RGNN uses the same set of parameters as in GNN recurrently over nodes in a graph to extract high-level node representations. The conventional prediction methods of RNN encounter computational challenges, especially for long-term information propagation. Thus, to address this issue and reduce the exploding and gradient vanishing problems, Gated recurrent unit (GRU) and Long-short term memory (LSTM) are introduced. GRNN employed with a GRU or LSTM is called gated GNN - GGNN.

GGNN reduces the recurrence to a fixed number of steps and also it does not limit the parameters for convergence. The general framework of GNN in case it is employed with a GRU unit is defined as:

$$h_v{}^{(t)} = GRU\left(h_v{}^{(t-1)}, \sum_{u \in N(v)} W h_u^{(t-1)}\right) \tag{11}$$

Since there is a strong interdependency between the elements of traffic in the transport network, a gated method is effective to be used for capturing the sequential relationship of the data.

Graph Attention Neural Networks (GAT)

In the aggregation process, the basic GNN framework puts equal weight on all neighbor nodes. However, not all neighbors are equally important. The aim of GAT is to apply attention to neighbor nodes to indicate the importance of each node during the aggregation step. For instance, in a road network, several links connecting to one intersection might have different traffic load, and thus should be captured with different scores. Veličković et al., 2018 proposed attention weights which define a weighted sum of the neighbors into the propagation steps. In the approach, the aggregated message passing is expressed as:

$$a_{ij} = \frac{exp(\sigma(a^\top [Wh_i] || [Wh_j]))}{\sum_{k\epsilon N_i} exp(\sigma(a^\top [Wh_i] || [Wh_k]))} \tag{12}$$

where a_{ij} is the attention coefficient of node j to i, N_i indicates the neighborhoods of node i in the graph. In addition, σ is the non-linear activation function (LeakyReLU), $h\epsilon R^{N\times F}$ is the input node features (N: number of nodes, F: dimension of the features), W is the weight matrix, and a is the learnable attention vector. Thus, the final output feature of each node is predicted as follows:

$$h_i = \sigma \cdot \Big(\sum_{j\epsilon N_i} \alpha_{ij} W h_j \Big) \tag{13}$$

In addition, the multi-head attention mechanism similar to (Vaswani et al., 2017) stabilizes the learning process. Thus, in each layer, the K-independent attention mechanism (each with different parameters) is applied to compute the feature-wise aggregated output (normally by average). The equation 13 is transformed, and their features are concatenated as follows:

$$h_i = \coprod_{k=1}^{K} \sigma \cdot \Big(\sum_{j\epsilon N_i} \alpha_{ij}^k W^k h_j \Big) \tag{14}$$

where $\coprod$ indicates concatenation, and K is the number of attention mechanism.

Since multi-head attention is performed on the final prediction, therefore the average of the resulting attentions are considered. The final representation after averaging takes the following form:

$$h_i = \sigma \cdot \Big(\frac{1}{K} \sum_{k=1}^{K} \sum_{j\epsilon N_i} \alpha_{ij}^k W^k h_j \Big) \tag{15}$$

An illustration of the aggregation process of multi-head attention layer is shown in Figure 11.

Spatio-temporal Graph Neural Network (STGNN)

In real-world applications, graphs could have dynamic characteristics both in terms of the graph structure and features (Z. Wu et al., 2021). For instance, in a transport network, the link features (e.g., speed, travel time, traffic flow) change during the day and thus it is required to capture both the spatial dependency, and the temporal variation of the graph. Spatio-temporal graph neural networks (STGNN) have attracted attentions in mimicking the spatial and temporal properties of graphs simultaneously. In many STGNN related studies, GCN are integrated with a temporal block (e.g., GRU, LSTM) to capture the spatial and temporal dependencies respectively.

STGNNs have been implemented in many applications including transportation (Li et al., 2018; X. Wang et al., 2020; Yu et al., 2018), driving behavior prediction (Jain et al., 2016), environment monitoring (Jin, Sha, et al., 2021; S. Wang et al., 2020), human action recognition (S. Yan et al., 2018) and more. Figure 12 shows an illustration of STGNN which is comprised of GCN for capturing spatial dependency, and a temporal block- GRU for representing the temporal features.

To summarize, GNNs are powerful techniques in capturing graph data. For mimicking only the spatial dependencies of the graph data, GCN and GAT could be applied, where for capturing the temporal variation of a graph data, STGNNs are widely utilized in literature. Since our extracted simulation-based data will only consider a request point with its associated waiting time without considering the variation of the

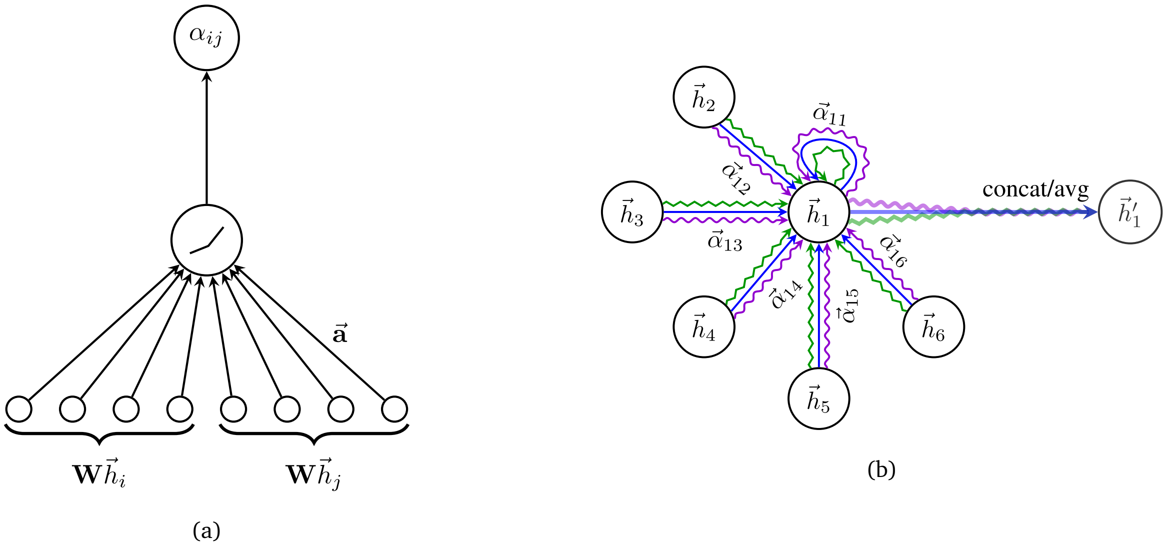


Figure 11: A display of GAT model (a) the attention mechanism of the model, and (b) the multi-head attention differentiated with different colors by node 1 and it's neighbors. Source: (Veličković et al., 2018)

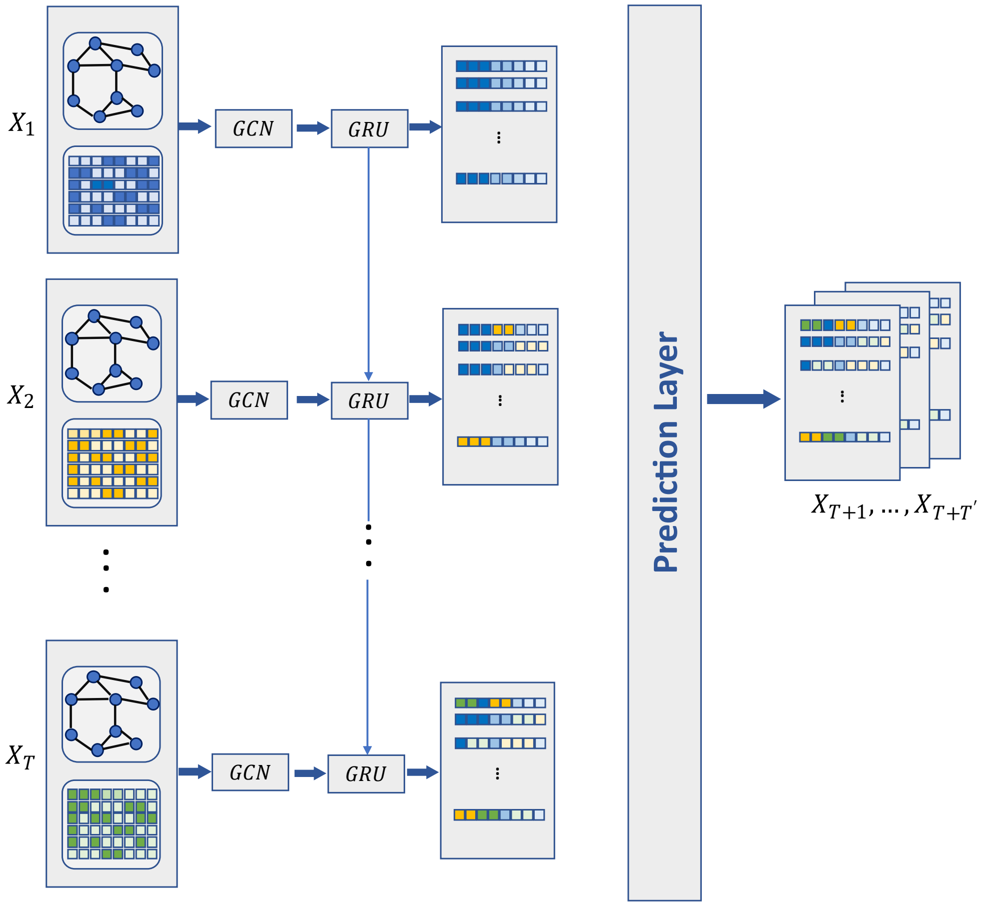


Figure 12: A sample structure of STGNN, where GCN captures the spatial dependenciy , and the temporal variation is represented by GRU, (own illustration)

waiting time in different time periods during the day, we utilize a model to predict the spatio-operational dependencies of the graph data in regards to the waiting time values. Hence, GCN and GAT model are the best candidates and are utilized in this master thesis.

3. Methodology

The methodology of this master thesis is comprised of four major sections namely: (i) the data generation, (ii) the proposed framework, (iii) the model evaluation, and (iv) the model transferability. In the data generation section, we introduce the process of ride-hailing simulation and extracting the waiting time data, followed by the map matching in QGIS and finally the transformation of the data for the final implementation in the proposed model. The proposed framework is the main contribution of this master thesis, which describes the models that are used in the experimental setup in the next chapter. Third, the model evaluation section presents the design of the loss function as well as evaluation matrices for the assessment of the proposed models' performance. Finally, in the model transferability section, the analysis of generalization capability of the trained models on a different dataset is described.

3.1. Data Generation

In this study, we utilize MATSim tool to model the ride-hailing and extract requests' waiting time. The simulation process contains running different scenarios including demand and supply variations as well as matching the outputs of the simulation runs with the road network geometry.

3.1.1. Ride-hailing simulation in MATSim

To conduct a simple simulation run in MATSim, config, population, and network files are required as describe in section 2.1.1. To include taxi modelling, we further need the definition of taxi demand within the population file and the taxi distribution data. The definition of the demand and optimal supply is a crucial step in extracting the waiting time data. Foremost, we extract the waiting time information under different penetration rates of ride-hailing services, with three different supply policies, namely: (i) Optimal supply, (ii) 20% above optimal supply, and (iii) 20% less than the optimal supply. This is done due to the fact, that we have excess and shortages of supply in reality. A schematic of the simulation scenarios including different demand and supply is shown in Figure 13.

Theoretically, we need the waiting time of all agents in the network, as for the proposed GNN, the waiting time information is needed for all nodes. However, a 100% penetration rate of ride-hailing in the network generates a huge number of empty vehicles driving in the network and thus create a huge congestion. This of course results in unrealistic waiting times. Hence, to avoid such an issue and meanwhile generate waiting time information for all agents, we propose a method which extract waiting time for all agents, however, not in one simulation setup, but under several simulation runs. To clarify, let's assume a service area where **x**% of all agents use ride-hailing for their daily trip activities. However, it is not known which agents exactly use ride-hailing and consequently any agent could be a potential candidate. Thus, first we randomly select **x**% of all agents to use ride-hailing and run the first simulation and store the resulting output. Second, another **x**% of all agents are chosen where they were not selected in step 1 and change their mode to ride-hailing and conduct the second simulation run and same here we store the output. Based on the value of **x**, a total of **n** different plan files are generate and sent to the MATSim simulator. Using this approach, it is intended to have ride-hailing requests for all agents but under **x**% penetration rate, and meanwhile we achieve more realistic data. The final extracted waiting times includes the waiting of request for all agents, however, under **x**% penetration rate of ride-hailing demand. Figure 14 shows the schematic overview of the agents plan creation and simulation.

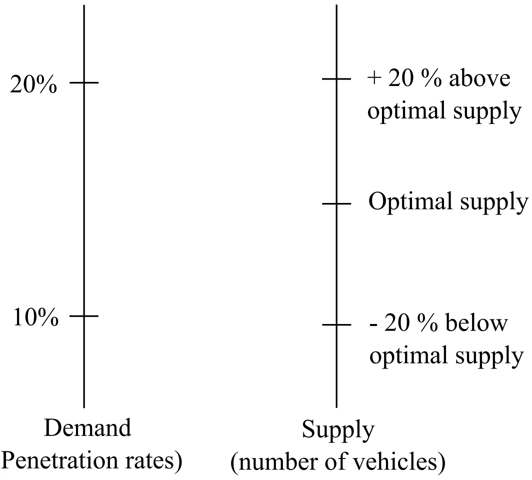

20%
10%
+ 20 % above optimal supply
Optimal supply
- 20 % below optimal supply
Demand (Penetration rates)
Supply (number of vehicles)

Figure 13: Different scenarios for demand scales in terms of ride-hailing penetration rates,and supply in terms of number of vehicles.

On the supply side, determination of the optimal number of ride-hailing vehicles follows a trial and error approach. Depending on the network and demand size, several simulation runs have been conducted to find an optimal number of vehicles. First, we run the model with bigger fleet sizes and check the number of vehicles being idle for most of the day. Similarly, simulation runs with smaller fleet sizes are also done. Hence, simulation runs continue from bigger and smaller fleet sizes until an optimal number of vehicles is reached.

3.1.2. Map Matching

The outputs of various simulation scenarios are the information of requests points with their locations (home, work, leisure, etc.) and simulated waiting times. To allocate these requests points to their nearby links and nodes, QGIS tool is used for the purpose of map matching. The distribution of the request points could have different relations to the nearby links. First, several request points could be located nearby a single link, which requires averaging the waiting times and then allocating a single waiting time to the link. Second, some request points, are far away from the nearby links and thus, it is difficult to programmatically allocate them to those links. To address this issue and better allocate request points to corresponding links, we propose creating grid cells.

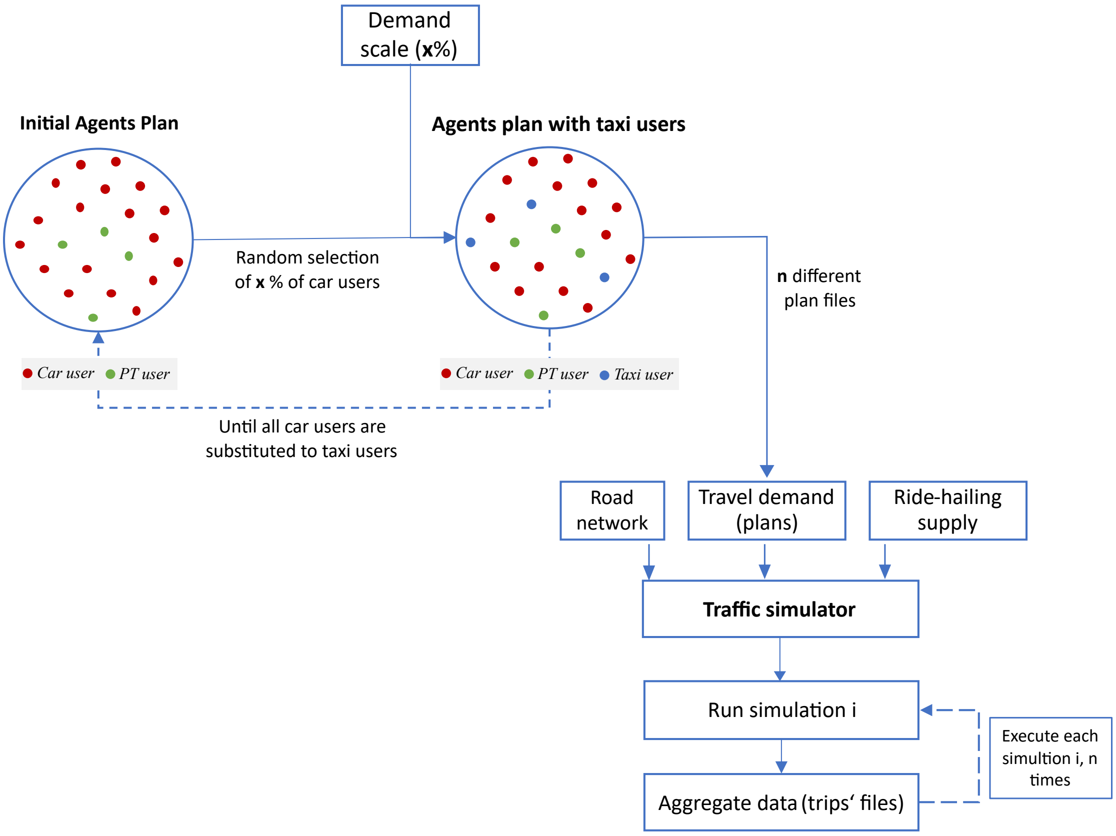

Demand scale (x%)
Initial Agents Plan
Agents plan with taxi users
Random selection of x % of car users
n different plan files
● Car user ● PT user
● Car user ● PT user ● Taxi user
Until all car users are substituted to taxi users
Road network
Travel demand (plans)
Ride-hailing supply
Traffic simulator
Run simulation i
Aggregate data (trips' files)
Execute each simultion i, n times

Figure 14: Schematic overview of the ride-hailing demand creation and MATSim simulation

However, since we conduct node prediction in our final model, and this requires obtaining the waiting time information from the links emerging from a single node, it saves one step to directly acquire the waiting information for each node from the grid cells rather than the links. Hence, the node-level waiting time information will be directly extracted from the grid cells.

We created grid cells of 250*250 meters on our network, these grid cells contain both nodes, and request points as shown in Figure 15. First, the average waiting time of each cell is determined by counting the number of request points in each cell and then the attributes of request points are averaged. Of course, it is only needed to know what is the average waiting time for ride-hailing in each cell. Second, after generating the waiting time of each cell, we assign them to the containing nodes in a cell. For instance, let' assume that cell (a) contains three different nodes, thus the waiting time of those nodes could be potentially the same as the grid cell itself. Worth-mentioning that we deal with missing cell waiting time by assigning the nearby cell waiting time to it.

Meanwhile, the census data is used to extract the population density in each grid cell. This information is important, since ride-hailing request closely relate to the population density of the area. Similar to waiting time extraction for each node, we allocate the population density around a node by assigning the corresponding cell population density to it.

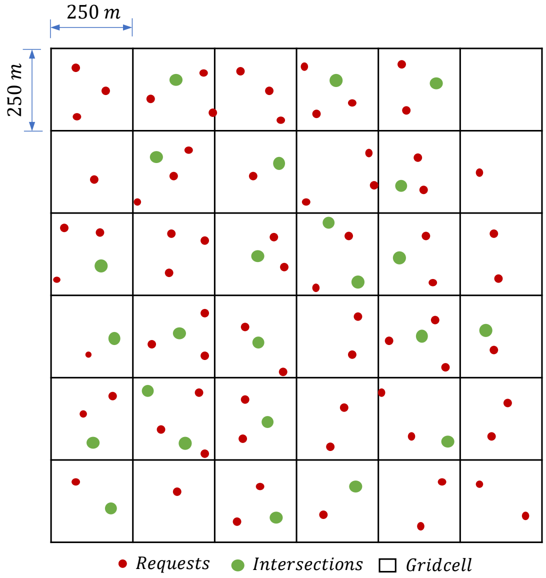


Figure 15: Illustration of a sample grid cells, request points, and road network

In addition, to include the ride-hailing vehicles distribution in our features, we map the vehicles (assigned randomly in the network during the simulation) in our network. Since, vehicles are assigned to links and their exact locations are not known, the link ID information is used to relate availability of a vehicle to a node using a dummy variable. We allocate the information from a link (0 if there is no vehicle to the link, and 1 otherwise) to the associate node and consequently create the vehicle distribution feature for the nodes.

To summarize, in this step the information for each node including population density, vehicle distribution, and the waiting time are created. The first two information will be utilized as a constant features of the nodes, where waiting time is the targets of our data for model training.

3.1.3. Feature Generation and Data Transformation

For the implementation of a GNN pipeline, we need a set of nodes with their features and the links connecting them. In a road network, most of the features are associated with links in the graph, whereas for the intersections (nodes), only a few pieces of information such as the location, type, etc are available. Therefore, for the GNN model, we should perform link prediction by using a dual graph to change the nodes to links and the links to nodes. However, the application of a dual graph changes the real typology of a road network. Thus, we propose a method so-called flow-out approach which aggregates the information of the links to the corresponding node.

The flow-out approach is straightforward. Each link is indeed emerging from a node and sinks to another node. Suppose, four links are emerging from a single node (e.g., node u in Figure 16), thus, the average of the attributes of these links could be assigned as the aggregated features of the corresponding node. As depicted in Figure 16, the aggregated attributes of the node u and v are calculated as follows:

$$\begin{aligned} h^{(u)} &= mean(h^{(e_{u,v})}, h^{(e_{u,1})}, h^{(e_{u,2})}, h^{(e_{u,3})}, h^{(e_{u,4})}) \\ h^{(v)} &= mean(h^{(e_{v,u})}, h^{(e_{v,5})}, h^{(e_{v,6})}, h^{(e_{v,7})}) \end{aligned} \tag{16}$$

where $h^{(u)}$ and $h^{(v)}$ are the aggregated features of node u and v respectively, and $h^{(e_{u,v})}$, $h^{(e_{u,1})}$,...,$h^{(e_{u,4})}$ are the features of emerging edges from node v.

Using this approach,the graph data are successfully created, which comprises nodes with their features and a set of links that shows the connection between the nodes. Worth-mentioning, that except for population density, and vehicle distribution which come from the grid cells and node degree attribute which belongs to each node, all other features are aggregated from the corresponding links. A sample presentation of the feature data as well as the edge information are depicted in Table 1.

Finally, the data is further processed and prepared for the final implementation. The detailed description of the data transformation is presented in section 4.5.

3.2. Proposed Framework

The proposed framework in this master thesis consists of two components, the embedding module, and the spatial dependency learning module. The embedding module aims to map the constant and variable properties of each node to a low-dimensional space and to further insert them into the learning module. On the other hand, the spatial dependency learning module requires the nodes embedding and edges information for node representation learning. In the following sections, each module is described in details.

3.2.1. Embedding Module

Since waiting time for ride-hailing requests is affected by many spatial factors such as the location of the request, the population density of the request area, connectivity of the nearby roads as well as the operational factors including capacity of links, speed limit, and ride-hailing vehicles distribution in the network. We initialize the node embedding by concatenating these features and mapping them into a low-dimensional latent space (one-dimensional tensor).

The embedding of nodes can be formulated as follows:

$$h_v = \sigma(W_v \cdot [L \parallel C \parallel S \parallel P \parallel V \parallel T \parallel F\ ] + b_v) \tag{17}$$

where W_v and b_v are the learnable weight and bias respectively, || is the concatenation operator, and L, C, S, P, V, T, F are the node features (please see Table 3 for description of each feature).

3.2.2. Spatial Dependency Learning Module

This module is the core contribution of this master thesis. The main function of this module is to predict the waiting time based on given spatio-operational attributes of the road network by implementing both GCN and GAT.

a) Graph Convolutional Network (GCN)

In a traffic network, nearby roads are more likely to share common attributes such as capacity, speed limit, and more. Thus, closely located nodes and links share both spatial and operational features. On the other hand, a GCN model with its $l-$layers is capable to aggregate and average the hidden representation of each node with its neighbors. We can apply the GCN approach, to predict the feature of a single node by aggregating its own features and the features of the neighboring nodes. Let's consider a graph $G = (V, E)$ with the following descriptions:

- A feature matrix $H^{(0)} = X_{in} \epsilon R^{N \times D}$, with N : number of nodes, and D : number of input features.
- An adjacency matrix A which describes the graph structure and their relations.

The output of the model $H^{(l)} \epsilon R^{N \times F}$ can be generated as follows:

$$H^{(l)} = f(H^{(l-1)}, A) \tag{18}$$

with $H^{(0)} = X_{in}$ the initial nodes' representations, $H^{(L)} = X_{out}$ is the final nodes' representations, and L is the number of convolutional layer.

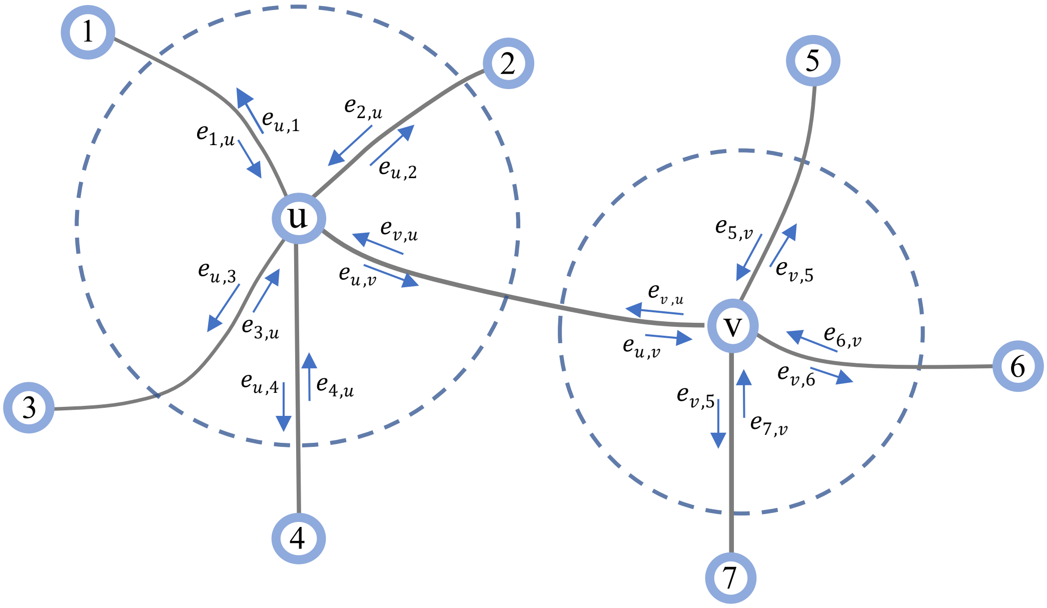


Figure 16: Illustration of the flow-out approach and allocation of links attributes to corresponding nodes

Table 1: A sample presentation of the graph data (node features and edge information)

Node	Capacity	speed limit	...	Vehicle
1	500	50		0
2	300	30		2
3	620	70		1

(a)

Link	fromNode	toNode
1	60	81
2	123	5
2	6	51

(b)

In this master thesis, we utilize the propagation rule introduced in (Kipf & Welling, 2017) as follows:

$$H^{(l+1)} = f(H^{(l)}, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{19}$$

with $\hat{A} = A + I$, where I is the identity matrix, and $\hat{D}$ is diagonal node degree matrix of $\hat{A}$.

The implementation of this GCN model could be simply described in three steps namely: (i) Feature propagation, (ii) Feature transformation, and (iii) Activation layer.

Feature propagation: In each layer, we take the average of the feature vectors of the nodes' neighbors.

$$H^{(l)} = \tilde{A}H^{(l-1)}$$

where $\tilde{A} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix including the self loops.

Feature Transformation: Each layer contains learnable weight matrix, which linearly transform the smoothed hidden feature representation to the next layer.

$$\tilde{H} = H^{(l)}W^{(l)}$$

where $\tilde{H}$ is the hidden layer representation.

Activation Layer: To ensure non-linearity, a non-linear component is added to the propagation and thus the final hidden feature representation is expressed as:

$$H^{(l)} = \sigma\tilde{H}$$

where σ is the non-linear activation function such as ReLU, sigmoid, tanh and more.

Based on the structure of our data, we consider a two-layer convolution operation similar to (Kipf & Welling, 2017). The schematic diagram (Figure 17) shows the two-layer GCN utilized in this master thesis.

The overall forward model takes the following forms:

$$H^{(1)} = f(H^{(0)}, A) = \sigma_1\left(\tilde{A}H^{(0)}W^{(0)}\right),$$
$$H^{(2)} = f(H^{(1)}, A) = \sigma_2\left(\tilde{A}H^{(1)}W^{(1)}\right)$$

combining the above two equations, the compact form of the representation is as follows:

$$H^{(2)} = \sigma_2(\tilde{A}\sigma_1(\tilde{A}H^{(0)}W^{(0)})W^{(1)}) \tag{20}$$

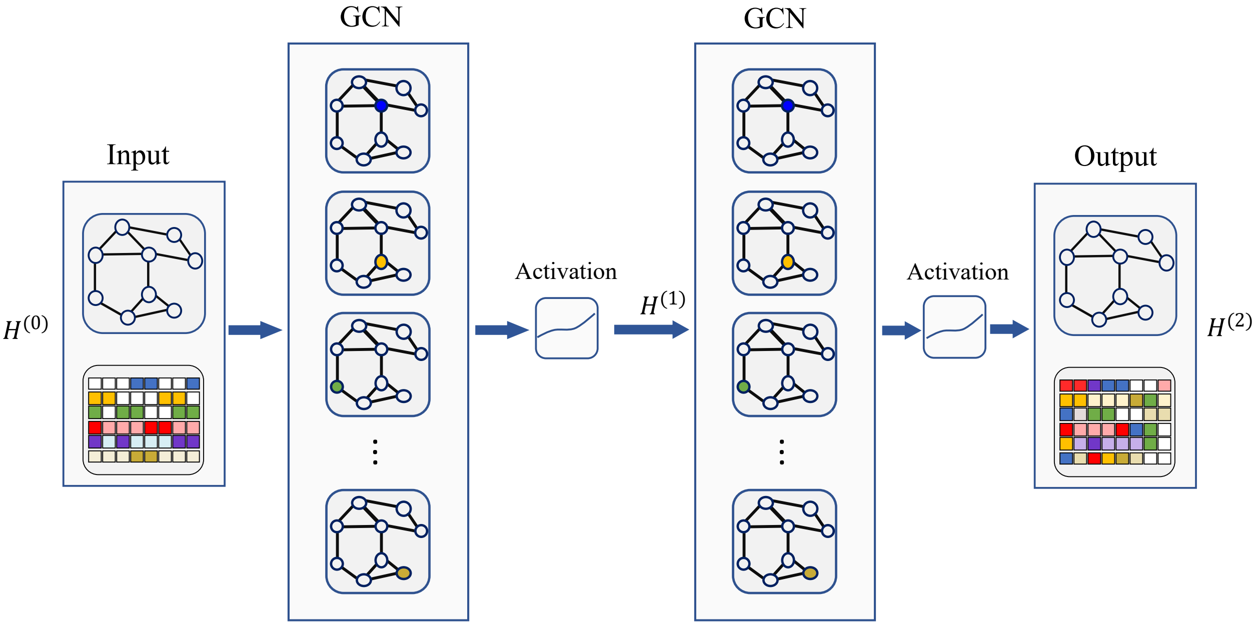


Figure 17: Overview of how a single node aggregates messages from its local

Tanh and Relu are selected as the activation functions (σ_1 and σ_2) respectively. The description of these activation functions are described in section 2.2.3.

b) Graph Attention Network (GAT)

In this model, we consider the attention mechanism in the feature propagation step. The attention mechanism proposed by Veličković et al., 2018 calculates the graph attention coefficient and adds it to the GCN operation. For details, the reader is referred to section 2.3.3. Similar to GCN model implementation, the GAT application could also be simply described in three steps, however, with adding a relation coefficient matrix.

Feature propagation: In each layer, we take the weighted average of the feature vectors of the nodes' neighbors.

$$H^{(l)} = \tilde{R}H^{(l-1)}$$

where $\tilde{R}$ is the relation matrix.

Feature transformation: The learnable weigh matrix is added, which transforms the hidden feature embeddings to the next layer.

$$\tilde{H} = H^{(l)}W^{(l)}$$

where $\tilde{H}$ is the hidden layer representation.

Activation layer: A non-linearity component is added to the propagation and thus the final hidden representation is expressed as:

$$H^{(l)} = \sigma\tilde{H}$$

where σ is the non-linear activation function.

Similar to section 3.2.2, we utilized a two-lyer GAT model, and thus GCN operation can take the following form:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{R}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{21}$$

where σ is the activation functions, $W^{(0)}$ and $W^{(1)}$ are the two learnable transformation matrices, and $\tilde{R} = mask(R) + I_N$ is the relation matrix. The mask is used to sparsify the relation matrix as follows:

$$mask(R) = \begin{cases} R_{ij} & ,if\ \tilde{A}_{ij} > 0 \\ 0 & ,otherwise \end{cases} \tag{22}$$

3.3. Evaluation module

During the training process for both proposed models (GCN and GAT), we select Mean Absolute Percentage Error (MAPE) as a loss function.

MAPE measures the percentage of average absolute error in comparison to the predicted value (how large is the difference between the predicted and actual values in comparison to the predicted value, see equation 25).

In addition, in this master thesis, we select three evaluation matrices namely: (i) Mean Absolute Error (MAE), (ii) Root Mean Square Error (RMSE), and (iii) Mean Absolute Percentage Error (MAPE) to evaluate the proposed models. MAE and RMSE are used to estimate the absolute error between the actual and predicted values, where RMSE is more sensitive in capturing large errors. On the other hand, MAPE is used to measure the estimation accuracy based on the percentage error. For these three matrices, the lower values indicate the better performance of a model. The definition of each matrix is described as follows:

$$\boxed{RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{23}$$

$$\boxed{MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \bar{Y}\right|} \tag{24}$$

$$\boxed{MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \bar{Y}_i}{Y_i}\right|} \tag{25}$$

where Y_i and $\bar{Y}_i$ are the predicted and true values of the i-th sample respectively, and N total number of predictions.

3.4. Model Transferability

We train both the proposed and baseline models (described in 4.7) with different datasets and evaluate the performance of each model. However, to check the generalization capability of the models, the best-trained model on the source data is selected and applied to a different dataset (we call it prediction dataset). The performance of each model is then assessed with the new dataset. This method is called transfer learning. Both the training datasets and the prediction dataset have the same number of features, however, the distribution of features varies among datasets.

4. Experimental Setup

In this section, we conduct an experiment to test the proposed models (see section 3.2.2) using simulated waiting time data. The following sections describe the study area, the extracted simulation-based data, the experiment settings, and methods for comparison which will be followed by results in the next chapter.

4.1. Study Area and Simulation Settings

In this master thesis, we utilize a synthetic MATSim model of the city of Cottbus to model ride-hailing demand. Cottbus is a city in the federal state of Brandenburg with around 100 000 inhabitants. The city is located roughly 110 km south of Berlin. The Cottbus transport network as shown in Figure 18, is comprised of 4470 nodes and 10729 links used for MATSim simulation. The population file contains 70000 agents and their work-related trips. To achieve more realistic results, the ride-hailing simulation runs are conducted together with the public transport lines which are operated within the city and its close surroundings.

Furthermore, this thesis uses the DRT extension of MATSim, which performs a centralized, on-the-fly assignment of vehicles to passengers as soon as the passengers' request to use the services. The ride-hailing is assumed to be door-to-door service, where the maximum waiting time of a passenger is set to 30 min. If a request is not served by any vehicles within the pre-defined constraint, the request is rejected and not considered in the simulation. In addition, vehicles are assumed to operate the whole day (24 h), without considering the consumed time for fueling or charging, operational issues or maintenance.

The data used for simulation runs as well as feature extraction in QGIS are collected from different sources. Cottbus network, agents plan file, and other MATSim-related files are gathered from Institut für Land- und Seeverkehr (ILS) - TU Berlin, and demographic information are used from the 2011 census data. In addition, OpenStreetMap is used as a base map in all maps of this master thesis.

4.2. Simulation Runs

As discussed in section 3.1.1, various simulation runs for different scenarios are conducted. On the demand side, 10% and 20% of all private trips are set to ride-hailing, whereas on the supply side, we have determined the optimum number of vehicles for both demand cases (10% and 20%) to be 1000 and 2000 vehicles respectively. Moreover, we run an additional simulation run with 15% demand and 1500 vehicles and extract another dataset (prediction dataset) to evaluate the transferability of the models.

First, for a 10% demand scenario, we select randomly 10% of agents from the population file and change their transport mode from private vehicle to taxi. To achieve the waiting time of all agents, we create 8 different sets of plans' files each with 10% taxi users, which contains 8*10% = 80% of all agents. The remaining 20% of agents use public transport for their daily activities, and therefore we do not change their mode of transport. Each of these 8 plan files together with one supply scenario (e.g., 750 taxis in the network) are simulated in MATSim. In total, 8*3 =24 different simulation scenarios have been conducted, and the sum of 24 trip files has been generated. Each trip file corresponds to 10% demand, and a certain supply in terms of the number of vehicles. A trip file contains a set of information (e.g., person ID, location, the start of trip, ...) for all generated trips during the simulation. Since we are interested in only taxi trips, we filter our data to include only the necessary information about taxis (e.g., request ID, location, waiting time,...). For each supply scenario, we concatenate 8 taxi trips data and as a result, we achieve the final taxi requests data for 80% of agents.

Similarly, for a 20% demand scenario, we select randomly 20% of agents from the population file and thus a total of 4 different sets of plans' files are created, and the same procedure is repeated. To exclude outliers from our data, we simply remove, waiting times which are less than a minute and more than 30 minutes. Finally, we create 6 different trips files (3 for 10% demand, and 3 for 20% demand), where each

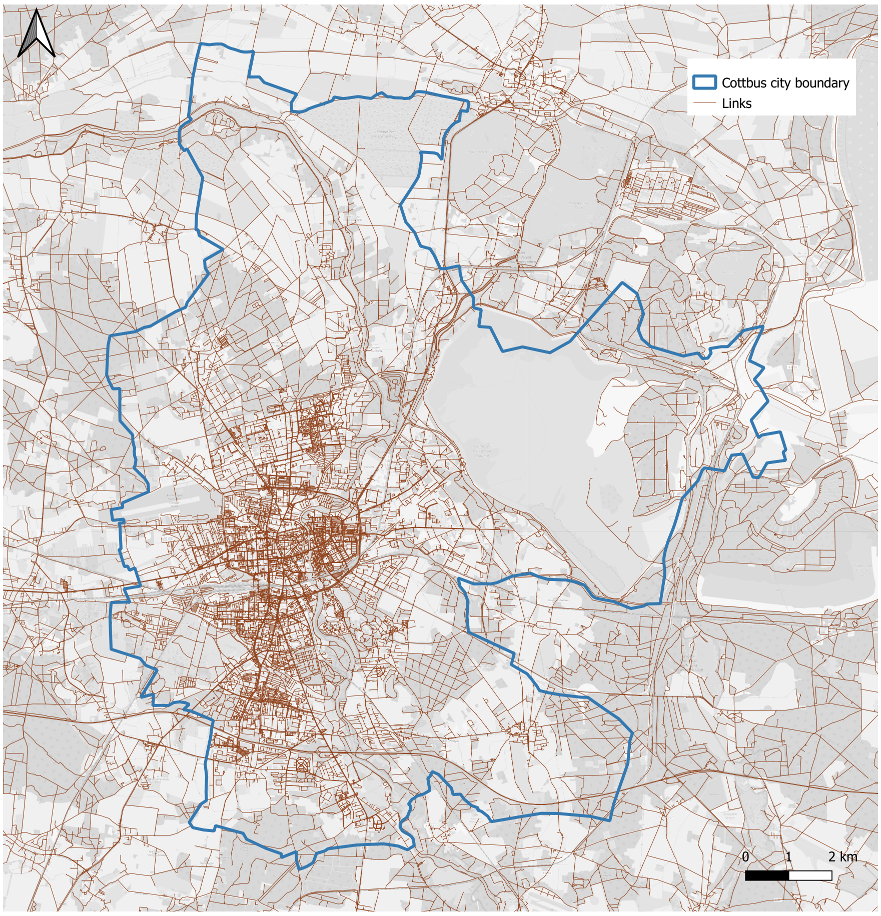


Figure 18: The Cottbus city road network.

is differentiated with the number of taxis in the network and these files are exported to QGIS for map matching.

4.3. Map Matching output

In the map matching process, we simply import 6 trip files and add them as trip points on the Cottbus network. Each point corresponds to a taxi trip request and the simulated waiting time. We create the grid cells (250*250) over the Cottbus network, and using the processing toolbox of QGIS (join attribute by location), we assign each trip point to a grid cell. The distribution of trip requests in grid cells is as shown in Figure 19. For visualization purposes, we only show the inner city map. Hence, for each trip file, a grid cell file is generated.

Similarly, we plot the nodes over the grid cells and assign each node to a grid. Consequently, the grid cell file includes information about the nodes and trip points located in each cell. Since it is possible that more than one trip point is located in a grid cell, we take the average of trip points in a cell and get the cell mean waiting time. Finally, each cell's waiting time corresponds to the node(s) waiting time located in that cell. The average waiting time distribution throughout the network using grid cells for two demand scenarios and associated optimal supply cases after map matching in QGIS is depicted in Figure 20.

On the hand, the taxi distribution file contains information about the links to which the taxis are located. Since the coordinates of taxis are not known, assigning a taxi to a grid cell is not possible. Therefore, we simply merge the taxi distribution file using the QGIS processing toolbox (join attribute by field value) with the links file, and thus the new link file will also have information on whether the link has a taxi or not. As a result, for node feature extraction, we use both the nodes' features extracted from the grid cells, as well as the information from the links file.

4.4. Features Generation

In the map-matching process, we successfully extract each node's waiting time information. As discussed in section 3.1.3, to allocate link features to a node, we assign the attributes of emerging links from a node using the QGIS processing toolbox (join attribute by field value). As a result, each node might contain features of different links including length, capacity, speed limit, travel time, average traffic flow, and vehicle availability. To have a unique features list for each node, we take the mean of features allocated to each node. Finally, the node file contains the feature (waiting time) from grid cells, and link attributes from the link file.

Furthermore, features such as population density and nodes' location-related information are independent of the

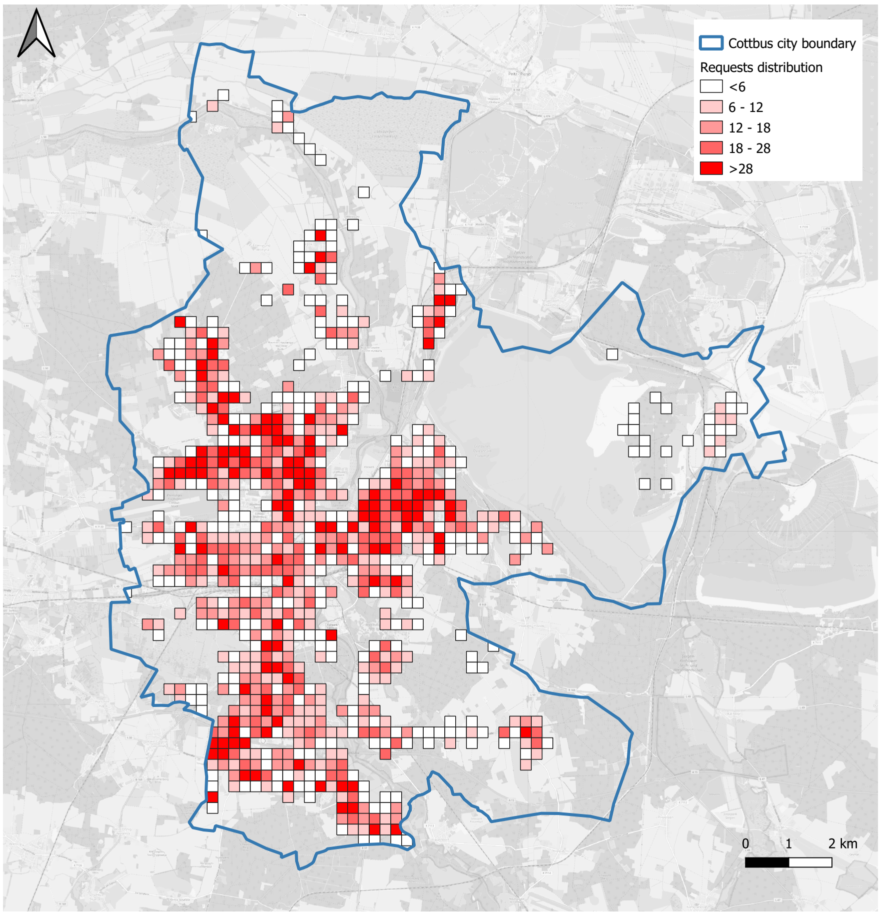


Figure 19: Illustration of the requests distribution under 20% demand and 2500 vehicles within Cottbus city.

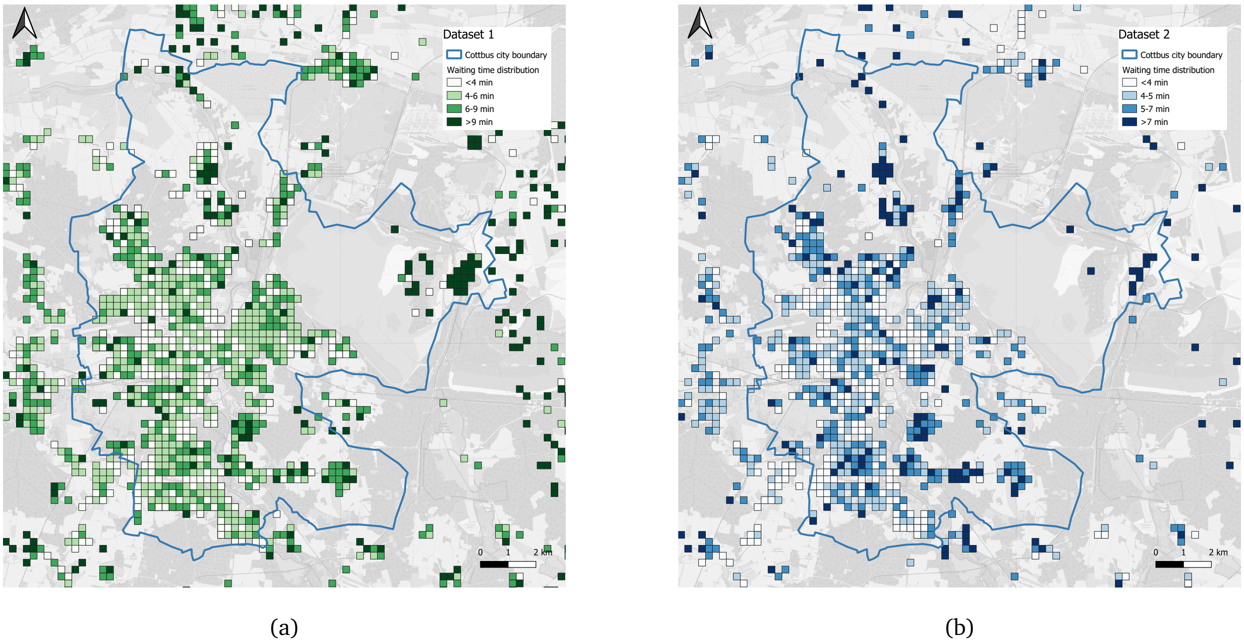


Figure 20: The average waiting time in each grid cell under: (a) 10% demand and 1000 vehicle, and (b) 20% demand and 2000 vehicles.

simulation output. We utilize census data which contains the population density points corresponding to 100-meter grid cells for Cottbus city. To estimate the population density in our proposed grid cells (250*250), we import the census data and overlay them in our grid cells and simply sum the values of population points in each grid cell. Meanwhile, we extract the node-related information such as betweenness, closeness, degree, and more by using the processing toolbox of QGIS (network centrality) as displayed in Table 2. Worth-mentioning that only degree and closeness are considered in the final features set.

Finally, combined features of nodes could be categorized into constant and variable features as presented in Table 3. The constant features include node degree, closeness, the average length, capacity and speed limit of the links emerging from a node, and average population density around a node, where variable features depends on each simulation scenario. The variable features contain ride-hailing vehicles distribution, average travel time and traffic flow of the emerging links from a node.

We train and test the proposed models, with a total of 6 datasets each containing the ride-hailing demand and the number of ride-hailing vehicles. Table 4 shows part of data for 10% demand and 1000 ride-hailing vehicles.

Meanwhile, we investigate the correlation among the features as well as between features and the waiting time for 10% and 20% demand datasets. As shown in Figure 21, there is not a strong correlation between the features and the waiting time in both datasets. However, length and travel time as well as capacity and speed limit are strongly correlated with each other.

In addition, to have a clear understanding of the waiting time distribution in each demand and supply scenario, we plot the histogram of each in Figure 22.

4.5. Data Transformation

After successful map matching of the ride-hailing request in QGIS and extraction of the nodes' features, the data is ready for the GCN and GAT implementation. However, the data will be further processed to easily implement them in the deep learning pipeline. First, we load each dataset as our node data, and network links as our edge data in python. The node data includes node features and target values (waiting times). We convert node features and target values to Numpy arrays. Numpy is a Python library that allows easy and efficient manipulation of arrays (Harris et al., 2020). The Numpy arrays are required, since in most machine learning techniques, they are the input of model. In addition, we use row-normalization technique to normalize the features data. A sample of transformed data is depicted in Table 5.

4.6. GCN and GAT settings

For the GCN model, We train a three-layer GCN as presented in section 3.2.2, and evaluate the performance of the model with our datasets. We select randomly 40% of the dataset for training, 30% for validation, and the rest for testing. In addition, we integrate an optimization module (differential evolution) to find the best set of hyperparameters.

Differential evolution (DE) is a stochastic population-based optimization method which is used for global optimization problems. DE does not require gradient information and hence could be efficiently used for nonlinear optimization problems (Georgioudakis & Plevris, 2020). The algorithm iteratively searches the design space to improve a candidate solution with regard to pre-defined targets. The candidate solution moves around the design space to check whether an improvement for the objective function is achieved. In case, a new candidate solution outperforms its parent, it replaces the parent, otherwise, it's simply discarded. In this thesis, the value of the objective function is the loss value of the validation, where the input variables are the hyperparameters. DE tries to change the hyperparameters within their boundary conditions aiming to find the minimum loss value for validation of the model.

The outcome of the optimization module for a 100 number of epochs depicts 53 hidden units, 0.0096 learning rate, 0.5 dropout rate for all layers, and $4e-5$ for L_2 regularization factor for the first layer. However, for simplicity, we use 64 as the number of hidden units, 0.01 learning rate, and 0.5 dropout rate.

Similarly, for the GAT model, we train a two-layer GAT model. The split of the dataset for training, validation, and testing remains the same as in the GCN model. Moreover, we use the optimized hyperparameters obtained for the GCN model to the GAT model, with additional hyperparameters for GAT namely as $K = 3$ number of attention heads, and $\alpha = 0.2$ for leakyReLU activation function.

Both models take features and adjacency matrix as inputs and predict numerical values as output. These numerical values are compared with the target values (true waiting times) until the loss function is minimized. For both models, we utilize Adam optimizer for minimization of the loss function. Regarding parameters analysis, we perform a sensitivity analysis of different number of layers of GCN and number of hidden units in regards to the GCN model performance as well as the number of attention heads and number of hidden units for GAT model.

4.7. Methods for comparison

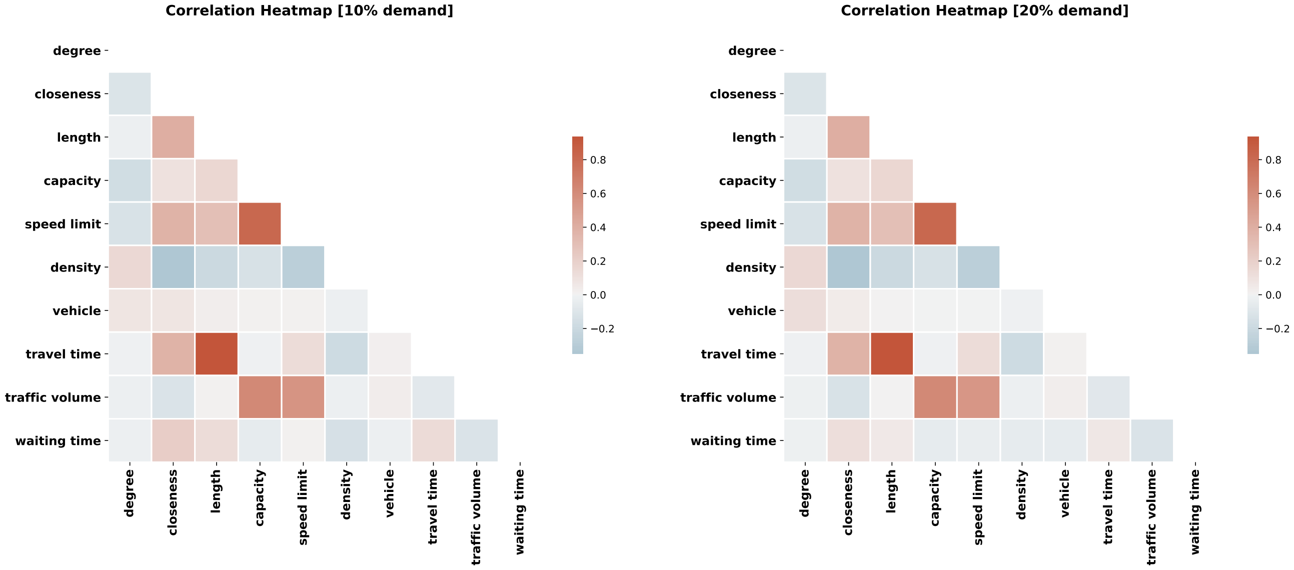
To verify the performance of the proposed models, we compare them with two baselines namely: (i) regression, and (ii) MLP models. The simple regression model takes the features as the independent variables and the waiting time as the dependent variable and creates a relation between them. On the other hand, Multiple layer perceptron (MLP) is a simple fully connected neural network model which takes the features as input and predicts the waiting time. In our experiments, we utilize three number of layers and ReLU as the activation function. The number of hidden units is same as in GCN and GAT models.

4.8. Software and Tools

In this master thesis, we use Python programming language to process the data and implement the proposed models. Python is a powerful programming language that has an extensive collection of libraries for data processing, computation, and visualization. The machine learning frameworks

Table 2: Description of the centrality features of the network nodes.

Node	degree	closeness	betweenness	eigenvector	xcoord	ycoord
60	0.000447	22629.9376	0	0	452370.2496	5747706.58
61	0.000447	7881.35775	0	0	447532.3292	5733172.71
62	0.000893	12573.66711	-5.56192E+12	0	452999.6389	5734564.663
63	0.00134	13370.23422	-5.74345E+12	0	455877.7776	5731834.515
64	0.00067	11899.66277	-4.2045E+12	0	454049.0625	5733308.502
65	0.00134	13336.2458	-3.51704E+12	0	455864.8711	5731921.305
66	0.00134	13520.31327	-5.92705E+12	0	456027.2457	5731850.125



(a)

(b)

Figure 21: Correlation matrix among features in (a) 10% demand and 1000 vehicles, and (b) 20% demand and 2000 vehicles datasets.

Table 3: Description of the nodes' features

Constant features	Variable features
D: degree Cl: closeness L: average length C: average capacity S: average speed limit P: population density	V: vehicle distribution T: average travel time F: average traffic flow

have been widely implemented in Python among the scientific community. The list of the prominent python libraries and tools used in this master thesis is as follows:

1. Computation: Numpy
2. Data handling and manipulation: Pandas
3. Visualization and plotting: Matplotlib, Seaborn
4. Machine learning: PyTorch

The hardware used for the study is a 2020 Lenovo Flex 5 with i7 processor and 16 GB RAM as well as a Desktop PC with almost the same specifications.

5. Experiment Results

In this chapter, the main findings of this master thesis are presented. First, we investigate the learning process of the implemented models by analyzing the convergence of loss functions for each model and in each dataset. Second, we evaluate the performance of the proposed models and compare them with regression and MLP models. Third, the findings of the analysis of the parameters under different hyperparameters settings for GCN and GAT models are presented. Finally, we analyze the transferability of each model, which are trained with the 10% and 20% demands and optimal supplies.

For simplicity, in the first and third sections, we only depict the plots for two datasets namely: (i) dataset 1, which include the 10% ride-hailing demand and the optimal number of vehicles (1000 vehicles) scenario, and (ii) dataset 2, which is the data under 20% ride-hailing demand and 2000 vehicles.

Table 4: Node features and waiting time values for a 10% demand scenario and 1000 ride-hailing vehicles

Node	D	Cl	L	C	S	P	V	T	F	W*
60	0.00045	22629.94	461.88	600.00	12.50	1.00	0.00	36.95	0.00	457.19
61	0.00045	7881.36	345.20	60.00	14.00	23.00	0.00	24.66	0.00	457.19
62	0.00089	12573.67	336.13	60.00	14.00	190.00	0.00	24.01	0.00	218.50
63	0.00134	13370.23	118.88	600.00	8.33	118.00	0.00	14.33	9.37	532.05
64	0.00067	11899.66	160.14	1800.00	15.28	80.00	1.00	11.27	1011.45	106.50
65	0.00134	13336.25	174.37	600.00	8.33	32.00	0.00	21.04	10.45	568.00
66	0.00134	13520.31	61.03	600.00	8.33	166.00	0.00	7.32	2.46	453.50
67	0.00134	12053.81	226.48	600.00	9.72	10.00	0.00	25.47	56.97	207.80
68	0.00134	12457.21	152.56	600.00	8.33	453.00	0.00	18.39	3.41	150.80
69	0.00089	11985.13	64.71	1800.00	17.36	23.00	1.00	4.00	1023.85	239.90
70	0.00134	11981.92	35.75	1400.00	15.74	23.00	1.00	2.67	717.53	239.90

W^*: average waiting time

Table 5: Normalized node-features for a 10% demand scenario and 1000 ride-hailing vehicles

Node	D	Cl	L	C	S	P	V	T	F
60	1.88E-08	0.9531	0.0195	0.0253	0.00053	4.21E-05	0.00000	0.001556	0.00000
61	5.35E-08	0.9441	0.0414	0.0072	0.00168	2.76E-03	0.00000	0.002954	0.00000
62	6.77E-08	0.9527	0.0255	0.0045	0.00106	1.44E-02	0.00000	0.001819	0.00000
63	9.41E-08	0.9390	0.0083	0.0421	0.00059	8.29E-03	0.00000	0.001007	0.00066
64	4.47E-08	0.7944	0.0107	0.1202	0.00102	5.34E-03	0.00007	0.000752	0.06753
65	9.45E-08	0.9403	0.0123	0.0423	0.00059	2.26E-03	0.00000	0.001483	0.00074
66	9.33E-08	0.9412	0.0042	0.0418	0.00058	1.16E-02	0.00000	0.000510	0.00017
67	1.03E-07	0.9285	0.0174	0.0462	0.00075	7.70E-04	0.00000	0.001962	0.00439
68	9.79E-08	0.9098	0.0111	0.0438	0.00061	3.31E-02	0.00000	0.001343	0.00025
69	5.99E-08	0.8033	0.0043	0.1207	0.00116	1.54E-03	0.00007	0.000268	0.06863
70	9.45E-08	0.8451	0.0025	0.0987	0.00111	1.62E-03	0.00007	0.000188	0.05061

5.1. Convergence Analysis

The initial results of this study reveal that in each model, the loss function (MAPE) successfully converges and learns the learnable weights. However, the speed of converges and the value of the loss function after 300 epochs varies depending on the model. For the GCN model, the loss function in training phase reduces from 1 to 0.35 after 300 epochs under dataset 1, and from 1 to 0.40 in dataset 2. For both scenarios, the loss function begins converging after 70 epochs in training phase as shown in Figure 23 (a) left , where in validation phase, the loss function converges already after 60 epochs (see Figure 23 (a) right). In comparison to training phase, where the fluctuation in loss function still exists after 70 epochs, in validation phase, loss function does not fluctuate after 60 epochs. This determines how fast the model learns the parameters.

Similarly, for GAT model, the loss function decreases from 4 to 0.37 during training phase in dataset 1, and from 5 to 0.39 when implementing dataset 2. In comparison to GCN model, GAT model learns to converge later and after around 150 epochs (see Figure 23 (b) left). The same could be seen for validation, the loss function converges after 150 epochs as depicted in Figure 23 (b) right. In addition, the loss function fluctuates higher than in GCN model. Although, GCN model is simple to implement, but able to learn faster and achieve higher performance.

On the other hand, for the regression model, the loss function starts from 18 and converges to 0.44 in dataset 1, and from 24 to 0.45 in dataset 2. For visualization reason, the loss values after epoch 25 is displayed in Figure 24. The model converges after 130 epochs in the training phase, and 125 epochs in the validation phase as depicted in Figure 24 (a).

Furthermore, the loss function reduces from 1 to 0.38 in dataset 1, and from 1 to 0.40 in dataset 2 when implementing the MLP model as displayed in Figure 24 (b). The loss function converges after 100 epochs in the training and validation phases. In comparison, to the regression model, the loss function fluctuates higher in MLP model.

5.2. Model Evaluation

Comparing the performance of GCN and GAT models considering regression and MLP models as baselines, we can find that regression model has a weaker performance than deep learning models. The reason might be the due to the linearity of regression model as well as the limited capabilities of regression model in capturing the complex structure of graph data. On the other hand, GCN and GAT model show better

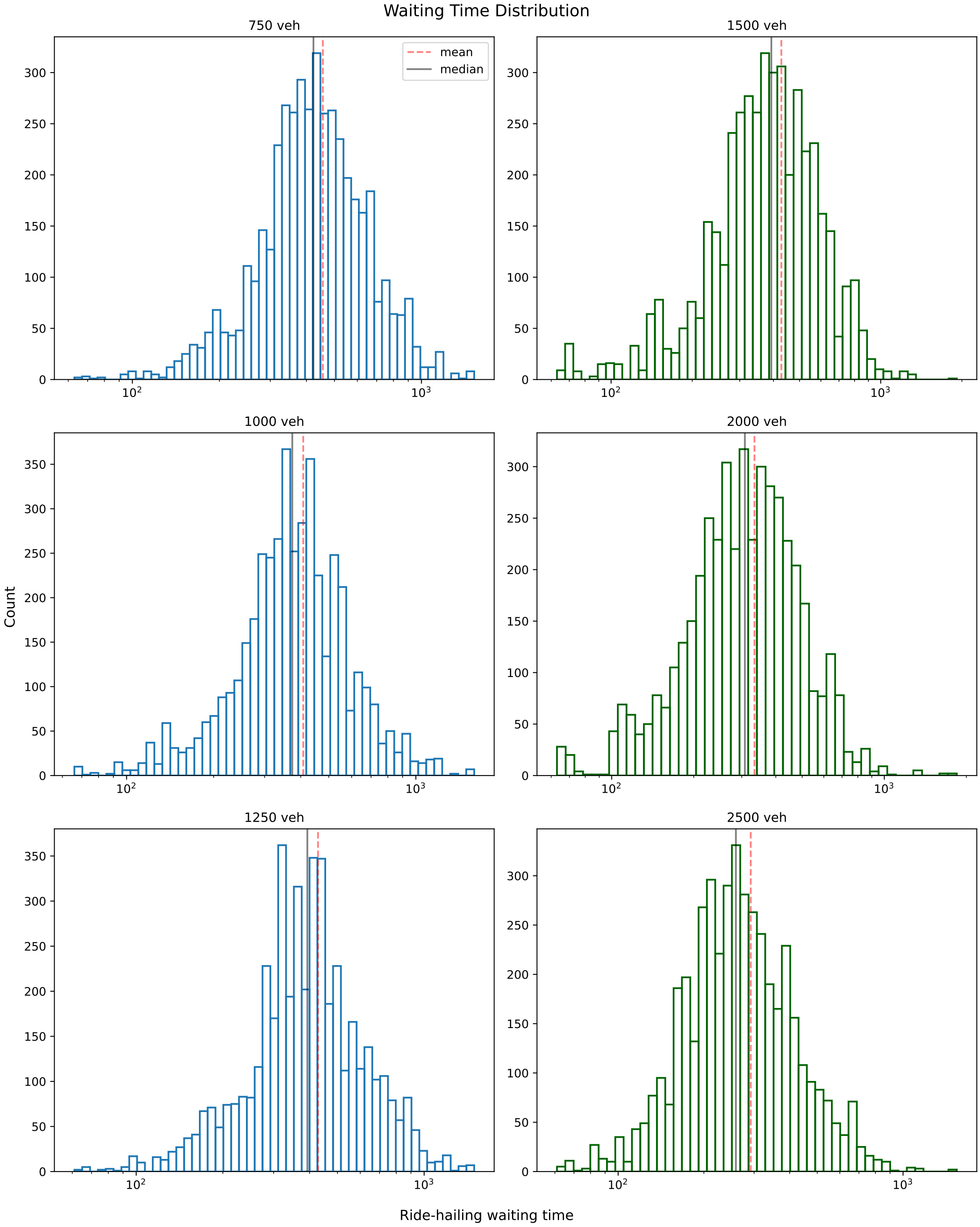
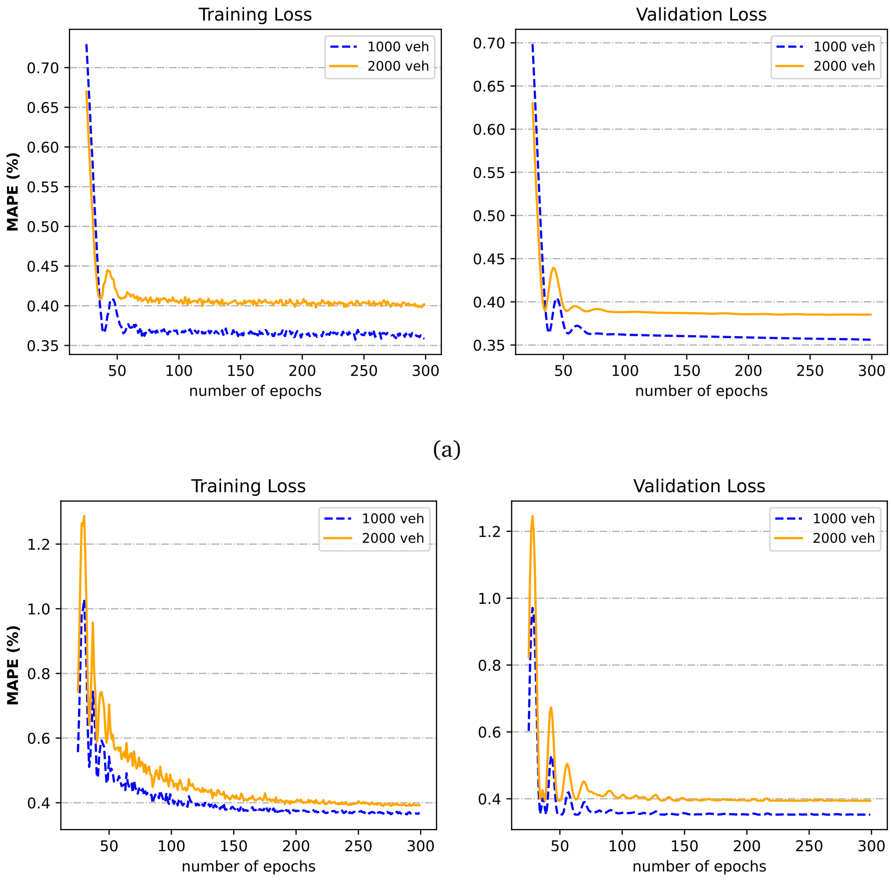


Figure 22: Waiting time data variation with 10 % demand on the left , and 20 % demand on the right under three different supply scenarios.

performance in all datasets. First, the findings of the models performance for 10% demand and three different supply scenarios depict that GCN model outperforms regression model as an average of 21.5%, 17.5% and 11.4% in MAPE, MAE, and RMSE respectively as depicted in Table 6. The comparison of GCN and MLP models reveal that GCN has better performance approximately 3% in MAE and 2% in RMSE, however, MAPE does not change.

Regarding the GAT model, it outperforms regression model about 21.5%, 17.5% and 11.5% in three evaluation matrices respectively. Similar to GCN, GAT shows around 2.7% and 2.3% better performance than MLP in MAE and



(a)

(b)

Figure 23: Training and validation losses for (a) GCN model, and (b) GAT model under 10% and 20% demand scenarios

RMSE, where no improvement has achieved for MAPE. In addition, the comparison of GCN and GAT models performance, it is found that both have almost the same performance (see Table 6).

Furthermore, the findings for the 20% demand scale and three supply scenarios reveal that the GCN model outperforms the regression model by an average of 15.4%, 16.4%, and 10.2% in MAPE, MAE, and RMSE respectively. However, when compared with MLP, GCN show around 1% in MAPE, 5.5% in MAE, and 5.3% in RMSE better performance. Meanwhile, the GAT model outperforms the regression model by an average of 13.8%, 13.4%, and 10.2% in MAPE, MAE, and RMSE respectively. However, when comparing with MLP, GAT shows 2.3% in MAE, and 2.5% in RMSE higher performance, whereas MAPE depicts a slightly higher loss value as shown in Table 7.

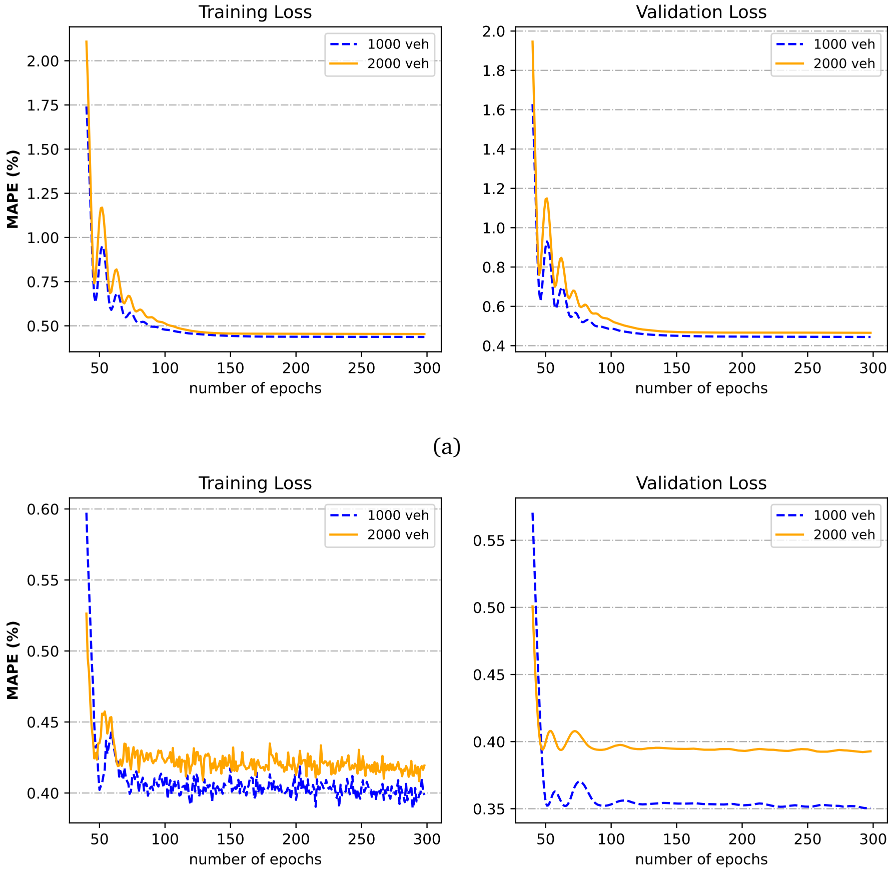
Furthermore, the analysis of the evaluation matrices reveals that by increasing the supply in terms of the number of vehicles, the overall models' performance improves. This also applies considering both demand scales. As depicted in Tables 6, and 7, all models have the best performance in the dataset with 20% demand and 2500 vehicles supply. On the hand, increasing the number of vehicles in the network results in decreasing the average waiting time of the network as well as the spread of the data in terms of standard deviation as displayed in Table 8. Hence, a less dispersion in the dataset (e.g., waiting time values) might best match with real-world data, and could be better linked with spatio-operational features of the graph.

5.3. Sensitivity Analysis

To investigate the effectiveness of different parameters in the performance of a model (e.g., GCN, GAT models), we conducted several experiments under various parameters settings. For GCN model, we choose number of GCN-layer, and the number of hidden units. On the other hand, number of attention heads and number of hidden units are selected for sensitivity analysis of GAT model performance.

First, fixing the number of hidden units $(n = 64)$, we run GCN model by changing the number of layers from 1 to 8. The findings reveal that the change in number of layers in GCN model does not have huge impact in performance of the model in all evaluation matrices. Still, with three-layers, GCN shows better performance when considering the datasets, and matrices as depicted in Figure 25.

Second, we fix the number of GCN-layer $(K = 3)$, and change the number of hidden units to [4,8,16,32,64,128].



(a)

(b)

Figure 24: Training and validation losses for (a) regression model, and (b) MLP model under 10% and 20% demand scenarios

Table 6: Performance of the proposed models in comparison to baseline models for estimation of waiting time under 10% ride-hailing demand and supply scenarios.

Scenario	- 20% optimal supply			optimal supply			+ 20 optimal supply		
Method	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
Regression	0.42	182.77	239.07	0.44	170.11	223.27	0.44	183.32	244.80
MLP	0.32	150.94	209.88	0.35	148.08	208.09	0.35	155.75	221.40
GCN	0.32	146.95	204.10	0.35	144.56	206.43	0.35	150.83	215.04
GAT	0.32	146.60	204.54	0.35	144.40	203.35	0.35	151.60	216.66
Improvement (%)									
*GCN**	**23.8**	**19.6**	**14.6**	**20.5**	**15.0**	**7.5**	**20.5**	**17.7**	**12.2**
*GCN***	**0.0**	**2.6**	**2.8**	**0.0**	**2.4**	**0.8**	**0.0**	**3.2**	**2.9**
*GAT**	**23.8**	**19.8**	**14.4**	**20.5**	**15.1**	**8.9**	**20.5**	**17.3**	**11.5**
*GAT***	**0.0**	**2.9**	**2.5**	**0.0**	**2.5**	**2.3**	**0.0**	**2.7**	**2.1**

(*, **) Comparison with regression and MLP models respectively.

As shown in Figure 26, by increasing the number of hidden units, the model performance improves in all scenarios. However, the slope of the change in GCN model performance is different with regards to the change in the number of hidden units. For instance, the change in number of hidden units from 8 to 16 has higher impact on the model performance than the change from 32 to 64 units. Meanwhile, the training time increase with higher number of hidden units. Hence, we found out that a 64 number of hidden units is a an optimal choice for GCN model evaluation.

As displayed in figures 25, and 26, the change in the number of units has more impact on the model performance in all matrices in comparison to the change in the number of GCN layers.

Table 7: Performance of the proposed models in comparison to baseline model for estimation of waiting time under 20% ride-hailing demand and supply scenarios.

Scenario	- 20% optimal supply			optimal supply			+ 20 optimal supply		
Method	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
Regression	0.42	184.30	238.80	0.47	155.28	206.70	0.40	119.11	160.17
MLP	0.37	154.17	213.83	0.39	138.58	194.10	0.34	110.78	162.04
GCN	0.39	153.40	211.41	0.38	125.79	177.76	0.32	103.13	151.88
GAT	0.38	151.13	208.20	0.39	135.33	189.14	0.34	108.00	157.91
Improvement (%)									
*GCN**	**7.1**	**16.8**	**11.5**	**19.1**	**19.0**	**14.0**	**20.0**	**13.4**	**5.2**
*GCN***	**-5.4**	**0.5**	**1.1**	**2.6**	**9.2**	**8.4**	**5.9**	**6.9**	**6.3**
*GAT**	**9.5**	**18.0**	**12.8**	**17.0**	**12.8**	**8.4**	**15.0**	**19.3**	**1.4**
*GAT***	**-2.7**	**2.0**	**2.6**	**0.0**	**2.3**	**2.4**	**0.0**	**2.5**	**2.5**

(*, **) Comparison with regression and MLP models respectively.

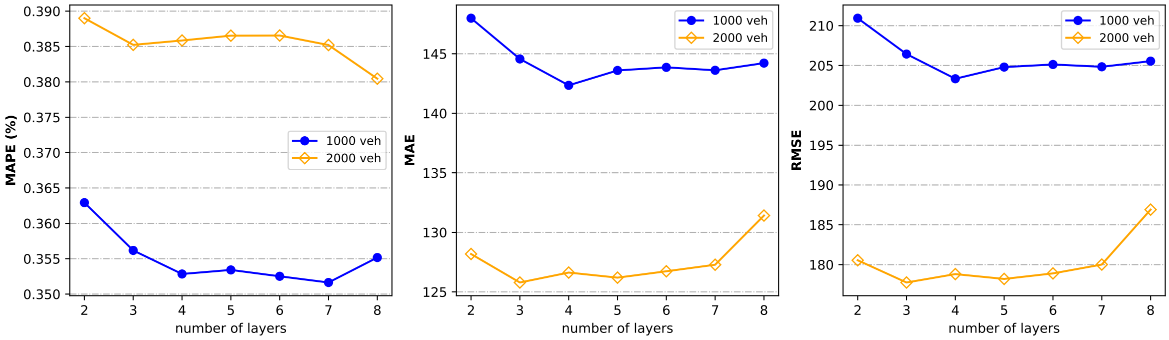


Figure 25: Variation of the GCN model performance with different number of layers under the optimal supply scenarios (1000 and 2000 vehicles) for both demand scenarios (10% and 20%) respectively.

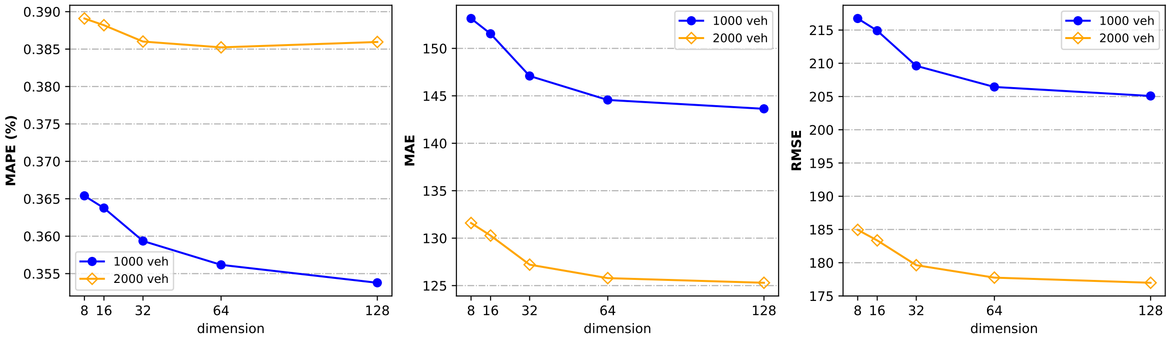


Figure 26: Variation of the GCN model performance with different number of hidden units under the optimal supply scenarios (1000 and 2000 vehicles) for both demand scenarios (10% and 20%) respectively.

Regarding the GAT model, first, we fix the number of hidden units to (n=64) and change the number of attention heads from 1 to 6. The findings show that does not change significantly, however, MAE and RMSE reduce considerably when the number of attention heads is set to 2 in all scenarios. Meanwhile, increasing the number of attention heads results in higher MAE and RMSE as depicted in Figure 27.

Meanwhile, the change in the number of hidden units [4,8,16,32,64,128] while keeping the number attention heads to (K =2), depicts that a higher number of hidden units results in poor performance in evaluation matrices and in both datasets. Similarly, a higher number of hidden units increases the training time, and thus selection of hidden units to 64 might be an optimal choice considering all evaluation matrices.

5.4. Model Transferability Analysis

The model performances on prediction data are shown in Table 9 and 10. Depending on the data used for training the models, different findings are extracted. For instance,

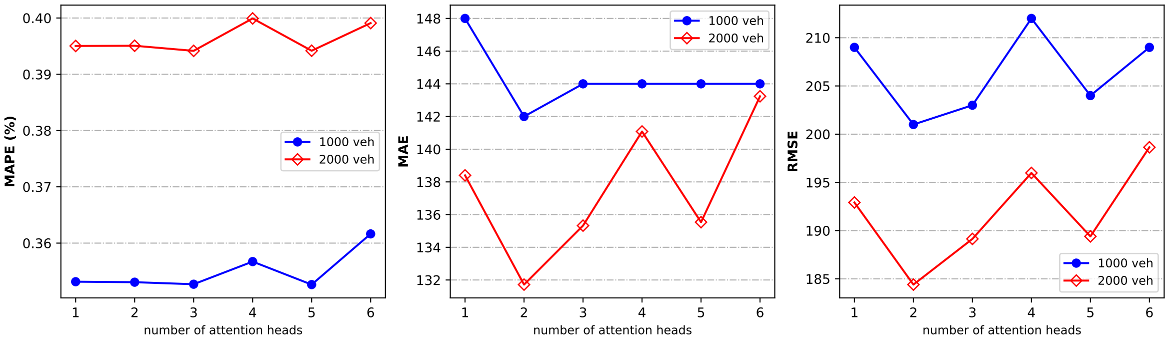


Figure 27: Variation of the GAT model performance with different number of attention heads under the optimal supply scenarios (1000 and 2000 vehicles) for both demand scenarios (10% and 20%) respectively.

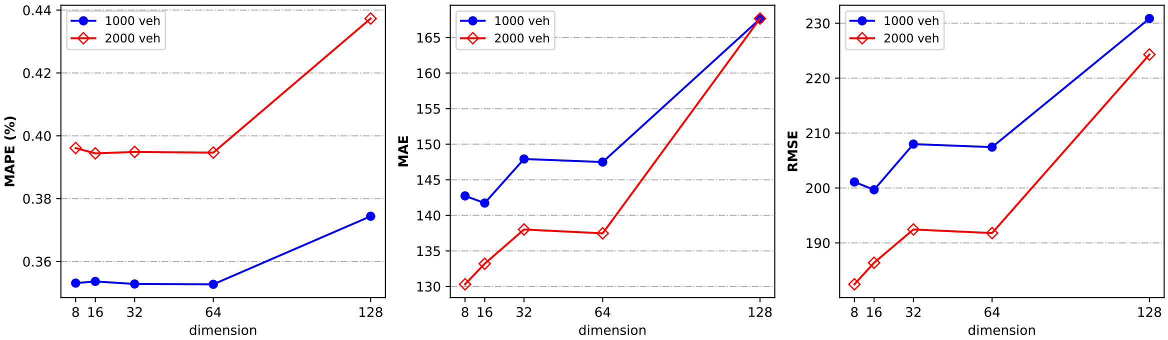


Figure 28: Variation of the GAT model performance with different number of hidden units under the optimal supply scenarios (1000 and 2000 vehicles) for both demand scenarios (10% and 20%) respectively.

Table 8: The description of mean and standard deviation of waiting time in each dataset.

Dataset	Mean	St.dev
750	456	194
1000	408	187
1250	428	198
1500	407	181
2000	334	160
2500	292	139

when the models are trained with the dataset 1, their performances on prediction data show improvements in most of evaluation matrices. As shown in Table 9, all models depict better performance in MAE and RMSE, where in MAPE, they show higher values. On the other hand, when the models are trained with the dataset 2, their performances in the prediction dataset deteriorate in all evaluation matrices as displayed in Table 10.

6. Conclusion

This chapter presents a conclusion of the contents of the whole thesis with the main findings as well as the future outlooks derived from this study.

In this thesis, we aim to estimate the ride-hailing requests' waiting time using the graph neural network (GNN) and considering the spatio-operational features of the transport network. The study begins with a comprehensive literature review, where we review the agent-based simulation models as well as the basics of neural networks and GNNs. The objectives are: (i) to choose a suitable simulation tool and settings for modeling the ride-hailing model and extract waiting time data, and (ii) to find which GNN-based approach(es) could predict the ride-hailing waiting time considering the features of the transport network data.

The methodology of this master thesis contains four main parts namely: (i) data generation, (ii) the proposed framework, (iii) the evaluation scheme, and (iv) model transferability. Multi-source data including the transport network, and agents plan data for running simulations, and OSM data, population density, and more are utilized for features' extraction in this master thesis. The final nodes data including the impacting features and the links information are prepared for the GNN implementation. In light of the application of different GNN-based approaches, GCN and GAT models are utilized to predict the waiting time in a service area. Both models take nodes' information together with each node's features as well as the relation between the nodes (links data) as inputs and predict the ride-hailing waiting time.

An experimental setup is developed to run MATSim-based

Table 9: Transferability analysis of the trained models with dataset 1 [10% demand and optimal supply]

	MAPE			MAE			RMSE		
Model	Eval	Pred	Change	Eval	Pred	Change	Eval	Pred	Change
Reg	0.44	0.52	(0.08)	170.11	146.14	23.97	223.27	199.37	23.90
MLP	0.35	0.47	(0.12)	148.08	126.05	22.03	208.09	181.75	26.34
GCN	0.35	0.50	(0.15)	144.56	125.63	18.93	206.43	175.32	31.11
GAT	0.35	0.49	(0.14)	144.40	126.89	17.51	203.35	180.27	23.08

Table 10: Transferability analysis of the trained models with dataset 2 [20% demand and optimal supply]

	MAPE			MAE			RMSE		
Model	Eval	Pred	Change	Eval	Pred	Change	Eval	Pred	Change
Reg	0.47	0.46	0.01	155.28	155.33	(0.05)	206.70	214.86	(8.16)
MLP	0.39	0.42	(0.03)	138.58	140.61	(2.03)	194.10	204.94	(10.84)
GCN	0.38	0.42	(0.04)	125.79	134.34	(8.55)	177.76	194.10	(16.34)
GAT	0.39	0.42	(0.03)	135.33	134.60	0.73	189.14	197.69	(8.55)

simulation runs under various demand and supply scenarios for the Cottbus city network. The extracted trips information together with other related features are concatenated and a total of 6 datasets are generated for the final implementation. In addition, regression and MLP models are selected as baselines for comparing the performance of the utilized models. We select MAPE, MAE, and RMSE matrices for the evaluation of each model.

The findings of the experiment test reveal that deep learning-based approaches have better performance than the regression model. For instance, GCN outperforms the regression model as an average of 15% in all datasets and evaluation matrices, whereas in comparison to MLP, GCN shows 3% better performance. Similarly, the GAT model depicts 14%, and 1.5% better performance than regression and MLP models respectively. In addition, it is found that all models have their best performance in the dataset with 20% demand scale and 2500 vehicles in terms of supply. Meanwhile, to test the effectiveness of the models' hyperparameters in regard to the performance of each model, a sensitivity analysis is carried out. The results depict that in the GCN model, the change in the number of GCN layers does not have a huge effect on the performance of the model, however changing the number of hidden units from low to high results in better performance of the model. On the other hand, the change in the number of attention heads in the GAT model does have a significant impact, however, by increasing the number of hidden units, the model performance deteriorates. Finally, the model transferability analysis shows that the models trained with the dataset 1 depict better performance in prediction dataset in comparison to models trained with the dataset 2.

The contribution of this master thesis successfully achieved the main goal of this study. We implemented several deep learning methods including regression model, MLP, GCN and GAT to estimate waiting time. Although GNN-based models are powerful in extracting graph data, their performance could be well-differentiated from other models (regression and MLP) with more complex data including several features. Meanwhile, there are some limitations in this master thesis and therefore it raises several new lines of work that could be pursued as valuable research topics in the future.

First, in this master thesis we used the waiting time data extracted from a simulation platform, however, utilization of real-world data for extraction of the graph features and further implementation of such data in the GCN and GAT models might have different outcomes. Second, in our data generation process, apart from waiting time information, and the traffic-related features, we also utilized population density as an extra feature that might have a direct impact on the ride-hailing waiting time. However, additional features such as land-use type, build-environmental characteristics (e.g., points of interest), public transport stops, and more impacting factors on waiting time could be considered. Thus, a study with more rich features might bring new insights into the waiting time prediction models. Third, our data is limited to spatio-operational features of the network. Each ride-hailing trip is requested in a specific location within a day. However, to include the impacts of traffic flow and congestion level, the temporal variation of the requests could be considered. For instance, a request waiting time in different time intervals should be added to the graph features. Using this data, which includes both the spatial and temporal variation of the request points, the STGNN model could be implemented to estimate the waiting time. Hence, a study could be conducted to extract such data from a microscopic simulation platform (e.g., Vissim, SUMO) and implement it in STGNN. Fourth, we used the features of the links to allocate node features. However, it is also possible to utilize a dual graph approach and directly conduct the prediction on links. Therefore, a study to transform a graph to its dual, and further do the prediction might have valuable outputs.

References

Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., & Smola, A. J. (2013). Distributed large-scale natural graph factorization. *Proceedings of the 22nd International Conference on World Wide*

Web - WWW '13, 37–48. https://doi.org/10.1145/2488388.2488393

Ahmed, H. U., Huang, Y., & Lu, P. (2021). A Review of Car-Following Models and Modeling Tools for Human and Autonomous-Ready Driving Behaviors in Micro-Simulation. *Smart Cities*, *4*(1), 314–335. https://doi.org/10.3390/smartcities4010019

Anderson, D. N. (2014). "Not just a taxi"? For-profit ridesharing, driver strategies, and VMT. *Transportation*, *41*(5).

Bischoff, J., Maciejewski, M., & Nagel, K. (2017). City-wide shared taxis: A simulation study in Berlin. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 275–280. https://doi.org/10.1109/ITSC.2017.8317926

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. *Oxford University Press, USA*, 498.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42. https://doi.org/10.1109/MSP.2017.2693418

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2013). Spectral Networks and Locally Connected Networks on Graphs.

Chen, C., Li, K., Teo, S. G., Zou, X., Wang, K., Wang, J., & Zeng, Z. (2019). Gated Residual Recurrent Graph Neural Networks for Traffic Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 485–492. https://doi.org/10.1609/aaai.v33i01.3301485

Chen, K., Deng, M., & Shi, Y. (2021). A Temporal Directed Graph Convolution Network for Traffic Forecasting Using Taxi Trajectory Data. *ISPRS International Journal of Geo-Information*, *10*(9), 624. https://doi.org/10.3390/ijgi10090624

Dai, H., Khalil, E. B., Zhang, Y., Dilkina, B., & Song, L. (2018, February). Learning Combinatorial Optimization Algorithms over Graphs [Comment: NIPS 2017]. https://doi.org/10.48550/arXiv.1704.01665

de Souza Silva, L. A., de Andrade, M. O., & Alves Maia, M. L. (2018). How does the ride-hailing systems demand affect individual transport regulation? *Research in Transportation Economics*, *69*, 600–606. https://doi.org/10.1016/j.retrec.2018.06.010

Fang, X., Huang, J., Wang, F., Zeng, L., Liang, H., & Wang, H. (2020). ConSTGAT: Contextual Spatial-Temporal Graph Attention Network for Travel Time Estimation at Baidu Maps. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2697–2705. https://doi.org/10.1145/3394486.3403320

Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein Interface Prediction using Graph Convolutional Networks. *Advances in Neural Information Processing Systems*, *30*.

Georgioudakis, M., & Plevris, V. (2020). A Comparative Study of Differential Evolution Variants in Constrained Structural Optimization. *Frontiers in Built Environment*, *6*, 102. https://doi.org/10.3389/fbuil.2020.00102

Gharaee, Z., Kowshik, S., Stromann, O., & Felsberg, M. (2021). Graph representation learning for road type classification. *Pattern Recognition*, *120*, 108174. https://doi.org/10.1016/j.patcog.2021.108174

Gilibert, M., & Ribas, I. (2019). Main design factors for shared ride-hailing services from a user perspective. *International Journal of Transport Development and Integration*, *3*(3), 195–206. https://doi.org/10.2495/TDI-V3-N3-195-206

Grau, J. M. S., & Romeu, M. A. E. (2015). Agent Based Modelling for Simulating Taxi Services. *Procedia Computer Science*, *52*, 902–907. https://doi.org/10.1016/j.procs.2015.05.162

Grover, A., & Leskovec, J. (2016, July). Node2vec: Scalable Feature Learning for Networks [Comment: In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016].

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems*, *30*.

Hamilton, W. L. (2020). *Graph Representation Learning* (tech. rep.).

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation Learning on Graphs: Methods and Applications [Comment: Published in the IEEE Data Engineering Bulletin, September 2017; version with minor corrections]. https://doi.org/10.48550/arXiv.1709.05584

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Henao, A., & Marshall, W. E. (2019). The impact of ride-hailing on vehicle miles traveled. *Transportation*, *46*(6), 2173–2194. https://doi.org/10.1007/s11116-018-9923-2

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

HÖrl, S. (2017). Agent-based simulation of autonomous taxi services with dynamic demand responses. *Procedia Computer Science*, *109*, 899–904. https://doi.org/10.1016/j.procs.2017.05.418

Horni, A., Nagel, K., & W. Axhausen, K. (Eds.). (2016). *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press. https://doi.org/10.5334/baw

Jain, A., Zamir, A. R., Savarese, S., & Saxena, A. (2016, April). Structural-RNN: Deep Learning on Spatio-Temporal Graphs [Comment: CVPR 2016 (Oral)].

Jiang, W., & Luo, J. (2021). Graph Neural Network for Traffic Forecasting: A Survey. *arXiv:2101.11174 [cs]*.

Jiang, W., & Zhang, L. (2019). Geospatial data to images: A deep-learning framework for traffic forecasting. *Tsinghua Science and Technology*, *24*(1), 52–64. https://doi.org/10.26599/TST.2018.9010033

Jin, G., Sha, H., Feng, Y., Cheng, Q., & Huang, J. (2021). GSEN: An ensemble deep learning benchmark model for urban hotspots spatiotemporal prediction. *Neurocomputing*, *455*, 353–367. https://doi.org/10.1016/j.neucom.2021.05.008

Jin, G., Wang, M., Zhang, J., Sha, H., & Huang, J. (2022). STGNN-TTE: Travel time estimation via spatial–temporal graph neural network. *Future Generation Computer Systems*, *126*, 70–81. https://doi.org/10.1016/j.future.2021.07.012

Jin, G., Yan, H., Li, F., Huang, J., & Li, Y. (2021). Spatio-Temporal Dual Graph Neural Networks for Travel Time Estimation. *arXiv:2105.13591 [cs]*.

Jing, P., Hu, H., Zhan, F., Chen, Y., & Shi, Y. (2020). Agent-Based Simulation of Autonomous Vehicles: A Systematic Literature Review. *IEEE Access*, *8*, 79089–79103. https://doi.org/10.1109/ACCESS.2020.2990295

Khoshraftar, S., & An, A. (2022, June). A Survey on Graph Representation Learning Methods.

Kipf, T. N., & Welling, M. (2017, February). Semi-Supervised Classification with Graph Convolutional Networks [Comment: Published as a conference paper at ICLR 2017].

Lecun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In M. Arbib (Ed.), *The handbook of brain theory and neural networks*. MIT Press.

Lee, K., Jin, Q., Animesh, A., & Ramaprasad, J. (2018). Are Ride-Hailing Platforms Sustainable? Impact of Uber on Public Transportation and Traffic Congestion. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3244207

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018, February). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting [Comment: Published as a conference paper at ICLR 2018].

Liu, Z., & Zhou, J. (2020). Introduction to Graph Neural Networks. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *14*(2), 1–127. https://doi.org/10.2200/S00980ED1V01Y202001AIM045

Maciejewski, M., & Nagel, K. (2012). Towards Multi-Agent Simulation of the Dynamic Vehicle Routing Problem in MATSim. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, R. Wyrzykowski, J. Dongarra, K. Karczewski, & J. Waśniewski (Eds.), *Parallel Process-*

ing and Applied Mathematics (pp. 551–560, Vol. 7204). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31500-8_57

Maind, M. S. B., & Wankar, M. P. (2014). Research Paper on Basic of Artificial Neural Network. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(1), 96–100. https://doi.org/10.17762/ijritcc.v2i1.2920

Merkwirth, C., & Lengauer, T. (2005). Automatic Generation of Complementary Descriptors with Molecular Graph Networks. *Journal of Chemical Information and Modeling*, *45*(5), 1159–1168. https://doi.org/10.1021/ci049613b

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online Learning of Social Representations [Comment: 10 pages, 5 figures, 4 tables]. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. https://doi.org/10.1145/2623330.2623732

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton : (Project Para)* (tech. rep.). Buffalo, NY.

Ruch, C., Hörl, S., & Frazzoli, E. (2018). AMoDeus, a Simulation-Based Testbed for Autonomous Mobility-on-Demand Systems. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3639–3644. https://doi.org/10.1109/ITSC.2018.8569961

Ruch, C., Lu, C., Sieber, L., & Frazzoli, E. (2021). Quantifying the Efficiency of Ride Sharing. *IEEE Transactions on Intelligent Transportation Systems*, *22*(9), 5811–5816. https://doi.org/10.1109/TITS.2020.2990202

Saracoglu, Ö. G., & Altural, H. (2010). Color Regeneration from Reflective Color Sensor Using an Artificial Intelligent Technique. *Sensors*, *10*(9), 8363–8374. https://doi.org/10.3390/s100908363

Scarselli, F., Gori, M., Ah Chung Tsoi, Hagenbuchner, M., & Monfardini, G. (2009). Computational Capabilities of Graph Neural Networks. *IEEE Transactions on Neural Networks*, *20*(1), 81–102. https://doi.org/10.1109/TNN.2008.2005141

Seo, Y., Defferrard, M., Vandergheynst, P., & Bresson, X. (2016, December). Structured Sequence Modeling with Graph Convolutional Recurrent Networks.

Stutz, D. (2014). *Understanding Convolutional Neural Networks* (tech. rep.).

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale Information Network Embedding [Comment: WWW 2015]. *Proceedings of the 24th International Conference on World Wide Web*, 1067–1077. https://doi.org/10.1145/2736277.2741093

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, December). Attention Is All You Need [Comment: 15 pages, 5 figures].

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018, February). Graph Attention Networks [Comment: To appear at ICLR 2018. 12 pages, 2 figures].

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion, 38.

Wang, Q., Xu, C., Zhang, W., & Li, J. (2021). GraphTTE: Travel Time Estimation Based on Attention-Spatiotemporal Graphs. *IEEE Signal Processing Letters*, *28*, 239–243. https://doi.org/10.1109/LSP.2020.3048849

Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., & Gao, F. (2020). PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2.5 Forecasting. *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 163–166. https://doi.org/10.1145/3397536.3422208

Wang, X., Ma, Y., Wang, Y., Jin, W., Wang, X., Tang, J., Jia, C., & Yu, J. (2020). Traffic Flow Prediction via Spatial Temporal Graph Neural Network. *Proceedings of The Web Conference 2020*, 1082–1092. https://doi.org/10.1145/3366423.3380186

Wu, Y., Lian, D., Xu, Y., Wu, L., & Chen, E. (2020). Graph Convolutional Networks with Markov Random Field Reasoning for Social Spammer Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 1054–1061. https://doi.org/10.1609/aaai.v34i01.5455

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks [Comment: Minor revision (updated tables and references)]. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Xu, Z., Yin, Y., & Ye, J. (2020). On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*, *132*, 29–43. https://doi.org/10.1016/j.trb.2019.02.011

Yan, C., Zhu, H., Korolko, N., & Woodard, D. (2020). Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, *67*(8), 705–724. https://doi.org/10.1002/nav.21872

Yan, S., Xiong, Y., & Lin, D. (2018, January). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition [Comment: Accepted by AAAI 2018].

Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting [Comment: Proceedings of the 27th International Joint Conference on Artificial Intelligence]. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Zha, L., Yin, Y., & Yang, H. (2016). Economic analysis of ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, *71*, 249–266. https://doi.org/10.1016/j.trc.2016.07.010

Zhang, K., Zhao, X., Li, X., You, X., & Zhu, Y. (2021). Network Traffic Prediction via Deep Graph-Sequence Spatiotemporal Modeling Based on Mobile Virtual Reality Technology (B. Nagaraj, Ed.). *Wireless Communications and Mobile Computing*, *2021*, 1–12. https://doi.org/10.1155/2021/2353875

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

Zhuang, C., & Ma, Q. (2018). Dual Graph Convolutional Networks for Graph-Based Semi-Supervised Classification. *Proceedings of the 2018 World Wide Web Conference*, 499–508. https://doi.org/10.1145/3178876.3186116

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Weggegangen, Platz vergangen - Non-territoriale Arbeitsplatzformen im Rahmen des hybriden Arbeitens aus Sicht der Arbeitnehmenden

Gone, Space Gone - Non-Territorial Workplace Models in the Context of Hybrid Working From the Employees' Perspective

Licia Reckersdrees

Paderborn University

Abstract

This study examines the effect of a non-territorial & hybrid workplace model on employees' preferences and marginal willingness to pay compared to territorial & non-hybrid, territorial & hybrid and non-territorial & non-hybrid work. A vignette study is used to evaluate individual preferences concerning various job offers which differ in terms of workplace model and salary. The correlation is analyzed using a chi-square test and a rank ordered logit regression. Possible underlying mechanisms are investigated by asking about the influence of workplace models on the working conditions of territoriality, privacy, collaboration and autonomy. The data is analyzed using the Kruskal-Wallis and Dunn's test. The results show a significant influence of workplace models on preference and willingness to pay. Employees prefer territorial & hybrid working and would forego an average of €2,175 compared to territorial & non-hybrid working. All four working conditions are influenced by the workplace models and can act as mechanisms. Overall, the conditions are rated best for territorial & hybrid working.

Zusammenfassung

Diese Studie untersucht die Wirkung einer non-territorialen & hybriden Arbeitsplatzform auf die Präferenz und marginale Zahlungsbereitschaft von Arbeitnehmenden im Vergleich zu territorialer & nicht-hybrider, territorialer & hybrider und non-territorialer & nicht-hybrider Arbeit. Mittels Vignetten-Studie werden Jobangebote entsprechend der Präferenz bewertet, die sich hinsichtlich der Arbeitsplatzform und des Gehalts unterscheiden. Der Zusammenhang wird durch einen Chi-Quadrat-Test und eine Rank Ordered Logit-Regression analysiert. Per Umfrage werden mögliche zugrundeliegenden Mechanismen untersucht, indem der Einfluss der Arbeitsplatzform auf die Arbeitsbedingungen Territorialität, Privatsphäre, Zusammenarbeit und Autonomie abgefragt und durch den Kruskal-Wallis und Dunn's Test ausgewertet wird. Die Ergebnisse zeigen einen signifikanten Einfluss der Arbeitsplatzformen auf die Präferenz und Zahlungsbereitschaft. Arbeitnehmende präferieren territoriales & hybrides Arbeiten und würden dafür auf durchschnittlich 2.175 € im Vergleich zum territorialen & nicht-hybriden Arbeiten verzichten. Alle vier Arbeitsbedingungen werden von den Arbeitsplatzformen beeinflusst und können als Mechanismen wirken, am besten bewertet werden sie insgesamt bei territorialem & hybridem Arbeiten.

Keywords: hybrid working; marginal willingness to pay; non-territorial working; vignette study; working conditions

1. Einleitung

Die physische Arbeitsumgebung hat einen wesentlichen Einfluss sowohl auf individueller als auch auf organisationaler Ebene, indem etwa die Arbeitszufriedenheit der Arbeit-

Gerne möchte ich mich bei Prof. Dr. Martin Schneider für die gute Betreuung bedanken, die genug Freiraum für eigene Entscheidungen geschaffen hat und gleichzeitig als enge Beratung sehr wertvoll war.

DOI: https://doi.org/10.5282/jums/v10i2pp491-521

nehmenden[1] und damit die Leistung von Unternehmen beeinflusst wird (Ashkanasy et al., 2014, S. 1174; Shobe, 2018, S. 4). Der Einfluss der Arbeitsumgebung wurde zwar bereits in vielfältiger Weise untersucht (Khazanchi et al., 2018, S. 590), die Digitalisierung und damit einhergehende Einführung von Informations- und Kommunikationstechnologien (IKT) und angepasste Gestaltung von Arbeitsprozessen haben allerdings erneut zu einem maßgeblichen Wandel der Arbeit geführt (De Croon et al., 2005, S. 120). Im Jahr 2023 kann gesteigerte Flexibilität für Arbeitnehmende durch hybrides Arbeiten als ein besonders wichtiger Trend angesehen werden (Castrillon, 2022), welcher zu einer veränderten Arbeitsumgebung führt. Hybrides Arbeiten meint dabei nach Halford (2005, S. 20) durch IKT sowohl von zu Hause aus, als auch im Unternehmen arbeiten zu können, sodass arbeitsbezogene und private Belastungen besser vereint werden können. Laut einer Studie mit 1.089 Befragten in Deutschland gaben 65% an, dass sie in Zukunft am liebsten in hybriden Arbeitsplätzen arbeiten möchten, 21% möchten nur von zu Hause aus arbeiten und lediglich 14% wollen ausschließlich im Büro arbeiten (Statista Research Department, 2022). Durch flexible Arbeitsgestaltung und die Einführung des hybriden Arbeitens verringert sich die Anwesenheitsquote der Arbeitnehmenden, sodass weitere innovative Arbeitsplatzformen besonders praktikabel werden, welche platz- und damit kostensparend sind (De Croon et al., 2005, S. 120). Darunter fallen non-territoriale Arbeitsplatzformen, welche nach Turner und Myerson (1998, S. 33) gemeinsam genutzte Arbeitsbereiche umfassen, die von Arbeitnehmenden täglich reserviert und nach Ablauf der Reservierungszeit vollständig geräumt werden müssen und durch Begriffe wie „Hot Desking“ oder „Hotelling“ bezeichnet werden. Bei dieser Arbeitsplatzform steht in der Regel weniger als ein Arbeitsplatz pro Person zur Verfügung (Gatt & Jiang, 2021, S. 955–956).

Obwohl hybride und non-territoriale Arbeitsplatzformen separat bereits vielfach untersucht wurden, gibt es gemischte Ergebnisse hinsichtlich der Auswirkungen der Arbeitsplatzformen einzeln betrachtet (Gatt und Jiang, 2021, S. 955; Morrow et al., 2012, S. 100). Außerdem liegt nur eine geringe Anzahl an Studien vor, welche die Auswirkungen von Arbeitsplatzformen prospektiv oder mittels Labordesign erheben. Dadurch können kausale Aussagen über Zusammenhänge nur in begrenztem Maß getroffen werden, obwohl hybrides und non-territoriales Arbeiten weit verbreitet sind (De Croon et al., 2005, S. 130).

Darüber hinaus besteht hinsichtlich der Auswirkungen der Kombination non-territorialer und hybrider Arbeitsplatzformen eine breite Forschungslücke. Obwohl diese in Kombination besonders praktikabel sind, ist nicht bekannt, welche Auswirkungen diese gemeinsam haben. Ebenso sind dem Zusammenhang zugrundeliegende Mechanismen bisher kaum erforscht (Wohlers & Hertel, 2017, S. 480).

Obwohl der Arbeitsort dem zweithöchsten Kostenfaktor entspricht (Gordon Brown, 2008, S. 5), sind neue Arbeitsplatzformen nicht nur aufgrund des Kostenfaktors relevant. Für Unternehmen ist es auch wichtig die Anforderungen der Arbeitnehmenden zu beachten und flexible Arbeitsformen anzubieten, um als attraktiverer Arbeitgeber wahrgenommen zu werden und im Wettbewerb um Arbeitnehmende mithalten zu können (Thompson et al., 2015, S. 727), da die Arbeitsmarktsituation aktuell zugunsten der Arbeitnehmenden ausfällt, sodass diese Forderungen in Bezug auf Arbeitsangebote stellen können (Castrillon, 2022). Entsprechend sollte untersucht werden, wie die Kombination einer non-territorialen und hybriden Arbeitsplatzform von potenziellen Arbeitnehmenden bewertet wird, da bisher nicht bekannt ist, wie sich diese auf die Präferenz auswirkt. Dem Einfluss der Arbeitsplatzform auf die Präferenz von potenziellen Arbeitnehmenden zugrundeliegende Mechanismen sind ebenfalls nicht erforscht, obwohl die Untersuchung dieser helfen kann, die Beziehung zwischen Arbeitsplatzform und Präferenz der Arbeitnehmenden besser zu verstehen.

Die vorliegende Arbeit hat entsprechend als Ziel den Zusammenhang zwischen den Auswirkungen einer non-territorialen und hybriden Arbeitsplatzform und deren Merkmalen auf die Präferenz von potenziellen Arbeitnehmenden empirisch zu untersuchen und zugrundeliegende Mechanismen herauszustellen. Dazu werden zwei Forschungsfragen anhand mittels einer Umfrage erhobener Primärdaten untersucht. Im Rahmen der ersten Forschungsfrage wird betrachtet, welche Auswirkungen die Arbeitsplatzform auf Arbeitnehmende hat, indem auf deren Präferenz und Bereitschaft, auf Gehalt zu verzichten, eingegangen wird. Da die Kombination aus non-territorial und hybrid beachtet werden soll, werden insgesamt vier sich aus non-territorial vs. territorial und hybrid vs. nicht-hybrid ergebende Arbeitsplatzformen im Vergleich untersucht: Arbeitsplatzform 1 (territorial & nicht-hybrid), Arbeitsplatzform 2 (territorial & hybrid), Arbeitsplatzform 3 (non-territorial & nicht-hybrid) und Arbeitsplatzform 4 (non-territorial & hybrid). Daraus werden in Kombination mit unterschiedlichen Gehältern Vignetten erstellt, sodass eine Conjoint-Analyse mittels Regression durchgeführt werden kann, um zu erfahren, welche Arbeitsplatzform präferiert wird und ob Arbeitnehmende bereit sind, dafür auf einen Teil des Gehaltes zu verzichten. Anschließend werden im Rahmen der zweiten Forschungsfrage dem Zusammenhang zugrundeliegende Mechanismen untersucht, indem die Wirkung der vier Arbeitsplatzformen auf die anhand der Literatur abgeleiteten Arbeitsbedingungen Territorialität, Privatsphäre, Zusammenarbeit und Autonomie analysiert werden. Dazu werden ein Kruskal-Wallis-Test und Dunn's-Test durchgeführt.

In der nachfolgenden Arbeit wird dazu zunächst in Kapitel 2 die bisherige Literatur über Auswirkungen hybriden und non-territorialen Arbeitens getrennt voneinander betrachtet zusammengefasst. Das *Job-Demands-Resources*-Modell (JDR-Modell) nach Bakker und Demerouti (2007) wird als theoretische Grundlage hinzugezogen, da es den Zusammenhang zwischen Arbeitsplatzformen und arbeitsbezogenen Konsequenzen im Allgemeinen abbildet. Anhand der in der Litera-

[1] In der vorliegenden Arbeit werden, wenn möglich, genderneutrale Formulierungen gewählt. Die Personenbezeichnungen beziehen sich, wenn nicht anders angegeben, auf alle Geschlechter.

tur identifizierten relevanten Arbeitsbedingungen und vorherigen Studien werden Hypothesen über die dem Zusammenhang zugrundeliegenden Mechanismen aufgestellt. In Kapitel 3 wird die Methode vorgestellt, indem auf das Studiendesign in Form der experimentellen Vignetten-Methode als Conjoint-Analyse eingegangen wird. Außerdem wird die Datenerhebung und die Datenanalyse beschrieben. Die Ergebnisse der Datenanalyse werden in Kapitel 4 unterteilt nach den beiden Forschungsfragen präsentiert. In Kapitel 5 werden die Ergebnisse diskutiert und deren Relevanz herausgestellt, bevor abschließend Limitationen und Implikationen für die Forschung und Praxis aufgezeigt werden.

2. Theoretische Grundlagen

In der Literatur herrscht selbst zwischen Forschern verschiedener Disziplinen ein weitgehender Konsens darüber, dass die physische Arbeitsumgebung vielfältige Auswirkungen auf Arbeitnehmende und Unternehmen hat (Morrow et al., 2012, S. 100; Shobe, 2018, S. 4). Insgesamt liegt allerdings kein klares Verständnis darüber vor, welche Effekte die Arbeitsumgebung auf das Verhalten von Arbeitnehmenden und weitere Ergebnisse hat (Ashkanasy et al., 2014, S. 1169). Dabei wirkt erschwerend, dass eine Vielzahl an möglichen Arbeitsumgebungen untersucht werden können, wobei Studien von Mikro-Themen wie der Platzierung der Tische, bis zu Makro-Themen reichen, bei denen das gesamte Unternehmen als physische Arbeitsumgebung betrachtet wird (Morrow et al., 2012, S. 100). Wenn betrachtet wird, welche Auswirkungen eine spezifische Arbeitsumgebung hat, liegen oftmals nur veraltete und gemischte Ergebnisse vor, die sich teilweise sogar widersprechen (Morrow et al., 2012, S. 100). Außerdem wurden bisher nicht ausreichend viele Studien mit unterschiedlichen methodischen Ansätzen zu den einzelnen Arbeitsumgebungen durchgeführt (Ashkanasy et al., 2014, S. 1174). Des Weiteren stellt die Operationalisierung und Messung der physischen Elemente oftmals eine Herausforderung dar (Morrow et al., 2012, S. 101). Den Auswirkungen des physischen Arbeitsumfeldes liegen komplexe Mechanismen zugrunde, welche neben den physischen Folgen auch Folgen für soziale Interaktionen in Organisationen haben, welche bisher nicht vollständig aufgedeckt werden konnten (Gonsalves, 2023, S. 3). Entsprechend sollten insbesondere auch psychologische Aspekte beachtet werden, da durch die Veränderung der Arbeitsumgebung Bedürfnisse von Arbeitnehmenden in einem anderen Maß erfüllt werden können (Frankó et al., 2022, S. 241).

Merkmale neuer Arbeitsplatzkonzepte wie dem hybriden und non-territorialen Arbeiten beeinflussen die Arbeitsumgebung maßgeblich, wurden bisher allerdings unzureichend erforscht (De Croon et al., 2005, S. 130). Bei dem hybriden Arbeiten verrichten Arbeitnehmende ihre Arbeit nicht mehr nur am Arbeitsplatz im Unternehmen vor Ort, sondern zumindest einen Teil dezentral an anderen Orten wie von zu Hause aus, indem vermehrt auf IKT zurückgegriffen wird. Durch den Ortswechsel kommt es zu einer physischen Entfernung, wodurch es infolge auch zu einer psychischen Entfernung und zu weniger Interaktionen kommen kann (Gajendran & Harrison, 2007, S. 1525). Non-territoriale Arbeitsplätze zeichnen sich dadurch aus, dass der Großteil der Arbeitsplätze nicht einzelnen Personen zugewiesen wird, sondern geteilt wird (Gatt & Jiang, 2021, S. 955–956). Außerdem handelt es sich dabei häufig um offene Räume (Inamizuu, 2013, S. 118), sodass die Arbeitsplätze effizienter genutzt werden können und dadurch die Interaktionen mit anderen Personen beeinflusst werden (Volker & van der Voordt, 2005, S. 241). Da die Merkmale der beiden Arbeitsplatzformen die Arbeitsumgebung beeinflussen, sollten auch deren Auswirkungen und die zugrundeliegenden Mechanismen erforscht werden. Da die Kombination non-territorial und hybrid in der Realität besonders häufig zur Anwendung kommt, sollte diese Forschungslücke insbesondere adressiert werden.

Um Aufschluss darüber zu bekommen, welche Auswirkungen die Kombination der non-territorialen & hybriden Arbeitsplatzform hat, werden in der vorliegenden Studie die sich aus non-territorialem vs. territorialem und hybridem vs. nicht-hybridem ergebenden vier möglichen Kombinationen untersucht. Dabei sollen die individuellen Konsequenzen auf Ebene der Arbeitnehmenden untersucht werden, da daraus anschließend auch Konsequenzen auf organisationaler Ebene abzuleiten sind. Entsprechend lautet die übergeordnete Forschungsfrage:

> ***F1:*** *„Welchen Einfluss haben die vier hier betrachteten Arbeitsplatzformen auf Arbeitnehmende?*

Die Auswirkungen unterschiedlicher Job-Charakteristiken auf das Wohlbefinden von Arbeitnehmenden wurden bereits vor fast einem halben Jahrhundert untersucht und es wurde versucht diese mithilfe von Modellen, wie dem *Demand-Control*-Modell von Karasek (1979) und dem *Effort-Reward-Imbalance*-Modell von Siegrist (1996) zu erklären. Das *Demand-Control*-Modell besagt, dass mentaler Stress das Ergebnis zu hoher Anforderungen und einem zu geringem Entscheidungsspielraum bei der Arbeit ist und eine Vergrößerung dieses Entscheidungsspielraums bei gleichbleibenden Anforderungen das Stresslevel senken kann, während die Produktivität aufrechterhalten wird (Karasek, 1979, S. 285). Das *Effort-Reward-Imbalance*-Modell besagt hingegen, dass Stress entsteht, wenn die Belohnung im Verhältnis zur wahrgenommenen Anstrengung als zu gering angesehen wird (Siegrist, 1996, S. 27). Da beide Modelle eine begrenzte Anzahl an Faktoren zur Erklärung der Entstehung von Stress hinzuziehen, welche nicht auf alle Arbeitsplätze zutreffen, und sich auf die negativen Folgen wie Stress fokussieren (Bakker & Demerouti, 2007, S. 309–310) wurde ein weiteres Modell entwickelt, welches im Folgenden näher beschrieben wird.

2.1. Job-Demands-Resources-Modell

Eine Weiterentwicklung des *Demand-Control*-Modells von Karasek (1979) und des *Effort-Reward-Imbalance*-Modells von Siegrist (1996) stellt das *Job-Demands-Resources*-Modell von Demerouti et al. (2001) dar. Es handelt sich dabei um das bekannteste Modell, dass die Beziehung zwischen Job-Charakteristiken und dem Wohlbefinden der Arbeitnehmenden abbildet (Lesener et al., 2019, S. 76). Das Modell wurde auf Grundlage empirischer Daten entwickelt, welche belegen, dass Burn-out in unterschiedlichen Arbeitskontexten auftreten kann und von bestimmten Kombinationen von Arbeitsbedingungen hervorgerufen wird (Demerouti et al., 2001, S. 508).

Die Besonderheit des Modells ist, dass sowohl negative, als auch positive Arbeitsbedingungen betrachtet werden, die sich auf das Wohlbefinden der Arbeitnehmenden auswirken können (Bakker & Demerouti, 2007, S. 310). Die Risikofaktoren für Burn-out sind zwar individuell, können aber laut Modell in zwei Kategorien unterteilt werden, welche als *Job Demands* und *Job Resources* bezeichnet werden (Bakker & Demerouti, 2007, S. 312), sodass das Modell auf eine Vielzahl von Beschäftigungen angewendet werden kann (Bakker & Demerouti, 2007, S. 309). Unter *Job Demands* werden Anforderungen verstanden, welche mit physischen oder psychologischen Anstrengungen einhergehen (z.B. eine ungünstige Arbeitsumgebung), und entsprechend mit Kosten verbunden sind, *Job Resources* bezeichnen hingegen Ressourcen, welche Anforderungen reduzieren, Weiterentwicklung fördern oder bei der Zielerreichung helfen (Bakker & Demerouti, 2007, S. 312). Entsprechend können Ressourcen auch alleine das Ergebnis positiv beeinflussen und nicht nur über eine Reduzierung der Anforderungen wirken (Wohlers & Hertel, 2017, S. 474). So können soziale Unterstützung, Feedback und Autonomie etwa motivierend wirken und zu einem gesteigerten Engagement führen (Müller et al., 2022, S. 4). Die Betrachtung von Anforderungen und Ressourcen macht das Modell besonders flexibel anwendbar, da somit Stärken und Schwächen von Jobs als auch positive und negative Indikatoren des Wohlbefindens der Arbeitnehmenden identifiziert werden können (Bakker & Demerouti, 2007, S. 309).

Weiterhin bildet das Modell zwei psychologische Prozesse ab, indem die Entstehung von *Job Strain* (Arbeitsbelastung) und *Motivation* betrachtet wird. Es wird davon ausgegangen, dass sich Anforderungen negativ auswirken können, da Individuen versuchen diese durch größere Anstrengungen auszugleichen und so die Arbeitsbelastung steigt. Andererseits können Ressourcen die Motivation von Arbeitnehmenden fördern, indem sie intrinsisch wirken, etwa wenn persönliches Wachstum gefördert wird, oder extrinsisch, wenn sie dabei helfen Arbeitsziele zu erreichen (Bakker & Demerouti, 2007, S. 313).Auch hier spielt die Interaktion der Anforderungen und Ressourcen eine Rolle, indem Ressourcen die Wirkung der Anforderungen auf die Arbeitsbelastung abschwächen können und Anforderungen den positiven Effekt der Ressourcen auf die Motivation verringern können (Bakker & Demerouti, 2007, S. 314). Laut dem Modell sind Ressourcen bei hohen Arbeitsanforderungen besonders wichtig, um die Arbeitsmotivation aufrecht zu erhalten (Bakker und Demerouti, 2007, S. 315;Bakker und Demerouti, 2017, S. 282).

Mithilfe einer Metaanalyse konnten Lesener et al. (2019, S. 93) die Annahmen des Modells belegen, wobei hochqualitative Studien mit Längsschnittdaten untersucht wurden, welche die durch das Modell vorhergesagten kausalen Beziehungen belegen Lesener et al. (2019, S. 92). Das *Job-Demands-Resources*-Modell bildet entsprechend belegt den Einfluss von Job-Charakteristiken auf das Wohlbefinden von Mitarbeitenden ab, indem deren Motivation und Gesundheit durch Anforderungen und Ressourcen verändert wird, sodass deutlich wird, dass diese beachtet werden sollten und Ressourcen gefördert werden sollten (Lesener et al., 2019, S. 95). Im Gegensatz zu vorherigen Modellen werden in dem weiterentwickelten *Job-Demands-Resources*-Modell allerdings nicht nur die Auswirkungen durch Motivation und Arbeitsbelastung auf negative abhängige Variablen wie Burn-out und Stress oder das Wohlbefinden der Arbeitnehmenden betrachtet. Neben der individuellen Ebene werden auch organisationale Ergebnisse betrachtet, worunter diverse Variablen fallen, wie die Arbeitsperformance und Abwesenheitsquote (Bakker & Demerouti, 2007, S. 310).

Das *Job-Demands-Resources*-Modell kann entsprechend hinzugezogen werden, um die Auswirkungen der Arbeitsplatzgestaltung zu untersuchen, da es Raum für eine Vielzahl an Arbeitsbedingungen lässt (Bakker und Demerouti, 2007, S. 310; Gatt und Jiang, 2021, S. 954) und diese je nach physischer Arbeitsumgebung unterschiedlich ausfallen können. Diesen Zusammenhang identifizieren auch De Croon et al. (2005) mithilfe einer systematischen Literaturanalyse. Diese stellen heraus, dass Bürokonzepte Arbeitsanforderungen, Arbeitsressourcen und kurzfristige Reaktionen der Mitarbeitenden beeinflussen, da diese mit unterschiedlichen Arbeitsbedingungen einhergehen (De Croon et al., 2005, S. 129). Des Weiteren kann die Arbeitsumgebung die Arbeitseinstellung sowohl positiv, als auch negativ beeinflussen (Gatt & Jiang, 2021, S. 954) und diese Dualität wird vom Modell berücksichtigt (Bakker & Demerouti, 2007, S. 310). Durch die Flexibilität des Modells, hinsichtlich der abhängigen Variable, kann mithilfe des Modells auf individueller Ebene geschaut werden, welche Form der Arbeitsplatzgestaltung Arbeitnehmende präferieren, um die Auswirkungen der Arbeitsplatzgestaltung zu untersuchen. Die Präferenz von Arbeitnehmenden bezüglich diverser Arbeitsplatzformen ist wiederum relevant, da sie unter anderem die Arbeitgeberattraktivität auf übergeordneter organisationaler Ebene beeinflusst (Maier et al., 2022). Unter Betrachtung bisheriger Forschungsergebnisse und des *Job-Demands-Resources*-Modells lässt sich darauf in Bezug auf die vier betrachteten Arbeitsplatzformen (territorial & nicht-hybrid, territorial & hybrid, non-territorial & nicht-hybrid und non-territorial & hybrid) folgende Annahme treffen:

> ***A1:*** *Die vier* ***Arbeitsplatzformen beeinflussen die Präferenz*** *der Arbeitnehmenden für ein Jobangebot.*

In verwandter Literatur, in der Auswirkungen anderer Arbeitsplatzformen wie beispielsweise flexiblen Arbeitszeiten und Orten untersucht werden, werden den Befragten verschiedene Jobangebote unterbreitet um mehr über deren Präferenzen zu erfahren (Bustelo et al., 2020; He et al., 2021; Mas & Pallais, 2017; Pouliakas & Theodossiou, 2010; Thompson et al., 2015). Die unterbreiteten Jobangebote unterscheiden sich teilweise neben der Arbeitsplatzform auch hinsichtlich des Gehaltes. Dadurch kann neben der Wirkung der Arbeitsplatzform auf die Präferenz der Befragten auch der mit dem Jobangebot einhergehende Nutzen bestimmt werden und dieser durch die Bereitschaft auf Gehalt zu verzichten quantifizierbar gemacht werden (Bustelo et al., 2020, S. 12). Da diese Experimente sowohl in Form von Feldexperimenten als auch in Form von Laborexperimenten Auswirkungen unterschiedlichster Arbeitsplatzformen auf die Bereitschaft auf Gehalt zu verzichten nachweisen, wird angenommen, dass sich auch die vier hier untersuchten Arbeitsplatzformen nicht nur auf die Präferenz der Arbeitnehmenden auswirken. Da die Arbeitsplatzformen vermutlich einen unterschiedlich großen Nutzen bringen, kann dies auch die Bereitschaft beeinflussen, für ein Jobangebot mit einer der Präferenz entsprechenden Arbeitsplatzform auf einen Teil des Gehaltes zu verzichten, im Vergleich zu weniger präferierten Jobangeboten. Entsprechend wird folgende zweite Annahme getroffen:

> ***A2:*** *Die vier* ***Arbeitsplatzformen beeinflussen*** *die* ***Bereitschaft*** *der Arbeitnehmenden* ***auf Gehalt zu verzichten****.*

2.2. Mechanismen

Um die Auswirkungen verschiedener Arbeitsplatzformen genauer vorhersagen zu können, sollte nicht nur betrachtet werden, ob die Präferenz und Bereitschaft der Arbeitnehmenden auf Gehalt zu verzichten dadurch beeinflusst wird, sondern auch welche Mechanismen dem zu Grunde liegen. Daraus ergibt sich die folgende zweite Forschungsfrage:

> ***F2:*** *Über welche* ***Mechanismen*** *beeinflussen die vier Arbeitsplatzformen die Präferenz und Bereitschaft der Arbeitnehmenden auf Gehalt zu verzichten?*

Um zu verstehen, wie eine bestimmte Arbeitsplatzform oder verschiedene Arbeitsplatzformen im Vergleich die Präferenz der Arbeitnehmenden beeinflussen, sollten die Eigenschaften betrachtet werden, die diese Arbeitsplatzform charakterisieren, da sie die Arbeitsbedingungen maßgeblich beeinflussen (Wohlers & Hertel, 2017, S. 470). Die Auswirkungen hybrider und non-territorialer Arbeitsplätze werden entsprechend von den Eigenschaften der zwei Merkmale *hybrid* und *non-territorial* gemeinsam beeinflusst. Nach Wohlers und Hertel (2017, S. 470) haben die Merkmale der Arbeitsplatzform Auswirkungen auf die Arbeitnehmenden auf individueller Ebene und damit auch auf organisationaler Ebene. Es kommt insgesamt zu arbeitsbezogenen Konsequenzen auf zwei Ebenen, indem die Arbeitsplatzform und damit einhergehende Merkmale in einem Zwischenschritt die Arbeitsbedingungen beeinflussen, wie in Abbildung 1 zu sehen ist.

Da die Auswirkungen des hybriden und non-territorialen Arbeitens in Kombination bisher unzureichend untersucht wurden (De Croon et al., 2005, S. 130), können die zu Grunde liegenden Mechanismen nicht aus spezifischer Literatur entnommen werden, welche sich mit den Auswirkungen dieser Arbeitsplatzform beschäftigt. Angelehnt an Wohlers und Hertel (2017, S. 470) kann sich allerdings an vorhandener Literatur orientiert werden, welche verwandte Arbeitsplatzkonzepte untersucht. Dadurch können relevante Mechanismen identifiziert werden, welche Vorhersagen über die Auswirkungen der Arbeitsplatzkonzepte auf die Arbeitsbedingungen und damit die individuellen und organisationalen Konsequenzen ermöglichen. In Bezug auf hybride und non-territoriale Arbeitsplatzformen konnten vier Mechanismen identifiziert werden. Diese lassen einen Zusammenhang zwischen der Arbeitsplatzform und den Arbeitsbedingungen *Territorialität*, *Privatsphäre*, *Zusammenarbeit* und *Autonomie* annehmen (Ashkanasy et al., 2014, S. 1170; Bencivenga und Camocini, 2022, S. 103; Brunia und Hartjes-Gosselink, 2009, S. 172; De Croon et al., 2005, S. 121; Elsbach und Pratt, 2007, S. 184; Wohlers und Hertel, 2017, S. 470). Darüber hinaus können mithilfe der Mechanismen mögliche Konsequenzen vorhergesagt werden, welche aus den veränderten Arbeitsbedingungen folgen, die durch verschiedene Arbeitsplatzformen beeinflusst werden.

Die vier betrachteten Arbeitsbedingungen werden im Folgenden separat definiert und der theoretische Hintergrund und die Relevanz der Arbeitsbedingung wird erläutert. Außerdem wird eine Tendenz aus der Literatur abgeleitet, wie hybride und non-territoriale Arbeitsplätze getrennt voneinander die Arbeitsbedingung beeinflussen. Unter der Annahme, dass die jeweilige Arbeitsbedingung durch die Arbeitsplatzform beeinflusst wird, werden anhand der Tendenzen jeweils zwei Hypothesen aufgestellt. Diese beziehen sich darauf, welche der vier betrachteten Kombinationen der Arbeitsplatzform mit der größten beziehungsweise geringsten Ausprägung der jeweiligen Arbeitsbedingung einhergeht.

2.2.1. Territorialität

Der erste Mechanismus beschreibt einen möglichen Effekt der Arbeitsplatzform auf arbeitsbezogene Konsequenzen, indem zunächst die Arbeitsbedingung *Territorialität* beeinflusst wird (Brunia & Hartjes-Gosselink, 2009; Wells, 2000; Wohlers & Hertel, 2017). Territorialität lässt sich „als räumliche Strategie zur Beeinträchtigung, Beeinflussung oder Kontrolle von Ressourcen und Menschen, durch die Kontrolle eines Gebiets“ (Sack, 1986, S. 1) verstehen. In Bezug auf den Arbeitsplatz meint Territorialität nach Wohlers und Hertel (2017, S. 471) ein Gefühl von Eigentum und einen eigenen Bereich zu haben. Dieser Bereich kann durch eine angepasste Anordnung und persönliche oder arbeitsbezogene Gegenstände personalisiert werden (E. D. Sundstrom & Sundstrom, 1986, S. 218). Die Personalisierung stellt nach E. D. Sundstrom und Sundstrom (1986, S. 225) einen wesentlichen Aspekt der Territorialität am Arbeitsplatz dar. Deshalb wird diese in der vorliegenden Arbeit nicht als einzelne Arbeitsbedingung betrachtet, sondern zur Territorialität attribuiert. Territorialität ist als Arbeitsbedingung eine wichtige Ressource, da die Personalisierung des Arbeitsplatzes positiv mit dem

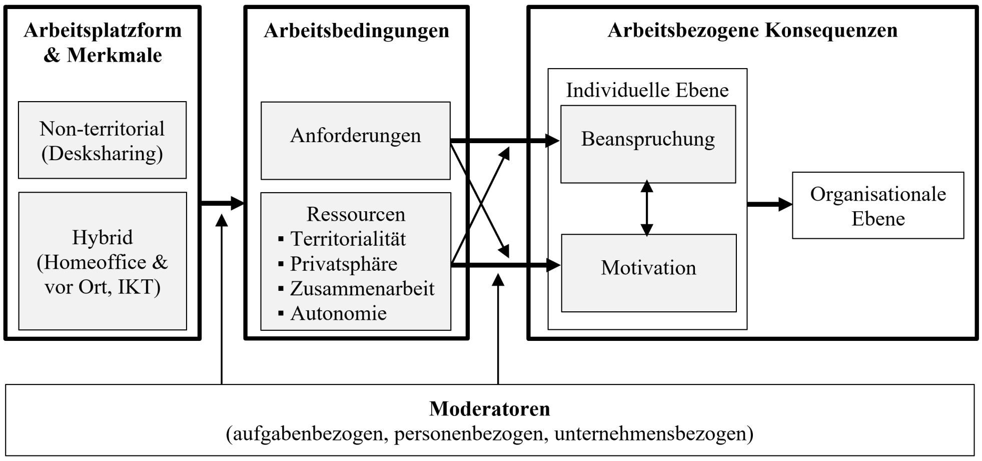

Arbeitsplatzform & Merkmale
Non-territorial (Desksharing)
Hybrid (Homeoffice & vor Ort, IKT)
Arbeitsbedingungen
Anforderungen
Ressourcen
▪ Territorialität
▪ Privatsphäre
▪ Zusammenarbeit
▪ Autonomie
Arbeitsbezogene Konsequenzen
Individuelle Ebene
Beanspruchung
Motivation
Organisationale Ebene
Moderatoren
(aufgabenbezogen, personenbezogen, unternehmensbezogen)

Abbildung 1: JDR-Modell angewandt auf hybrides & non-territoriales Arbeiten
(Quelle: eigene Darstellung in Anlehnung an Wohlers und Hertel (2017, S. 470) und Bakker und Demerouti (2007, S. 313))

Wohlbefinden des Arbeitnehmenden assoziiert wird (Wells, 2000, S. 249; Wohlers und Hertel, 2017, S. 471). Außerdem konnte Wells (2000) durch eine quantitative Studie und ergänzende Interviews in 20 Firmen belegen, dass Personalisierung signifikant mit der Zufriedenheit in Bezug auf die Arbeitsumgebung assoziiert ist (Wells, 2000, S. 247). Durch Personalisierung wird demonstriert, dass der Bereich von einer Person kontrolliert wird (E. D. Sundstrom & Sundstrom, 1986, S. 225). Durch die Kontrolle eines Bereiches kann auf individueller Ebene das Bedürfnis nach Wirksamkeit befriedigt werden (G. Brown, 2009, S. 47). Die Bedürfniserfüllung kann laut *Self-determination* Theorie (Deci et al., 2017), auf die später genauer eingegangen wird, zu arbeitsbezogenen Ergebnissen führen (Wohlers & Hertel, 2017, S. 471), sodass auf organisationaler Ebene der Einsatz gesteigert werden kann (G. Brown et al., 2005, S. 588). Des Weiteren lässt sich der eigene Bereich durch Personalisierung von anderen unterscheiden, sodass die eigene Identität reflektiert werden kann (Wells, 2000, S. 239). Dies kann nach Ashforth und Mael (1989) wiederum zur Aufdeckung von Ähnlichkeiten mit anderen Individuen beitragen, welche bei der Identifikation der Zugehörigkeit zu einer Gruppe behilflich sind. Laut *Social Identity* Theorie definieren sich Individuen partiell über Gruppenzugehörigkeiten (Ashforth & Mael, 1989, S. 34). Die Identifikation mit einer Gruppe wird mit gesteigerter Arbeitszufriedenheit und Motivation assoziiert (Wohlers & Hertel, 2017, S. 471–472).

Da *hybrides Arbeiten* soziale, physische und psychische Merkmale des Arbeitsumfeldes beeinflusst, ist davon auszugehen, dass die Identität der Arbeitnehmenden davon beeinflusst wird (Thatcher & Zhu, 2006, S. 1086) und damit auch die wahrgenommene Territorialität. Das Arbeiten im Homeoffice kann zu einer gesteigerten Identifikation mit der Organisation führen, wenn das Ausmaß des Arbeitens im Homeoffice als angemessen empfunden wird, bei Verpflichtung allerdings gegenteilige Konsequenzen haben (Thatcher & Zhu, 2006, S. 1081). In Bezug auf die Personalisierung bietet hybrides Arbeiten jedoch einen größeren Freiraum als die ausschließliche Arbeit in dem Unternehmen vor Ort. Arbeitnehmende können im Homeoffice selbst entscheiden, wie sie ihren Arbeitsplatz gestalten wollen und diesen mit persönlichen Gegenständen dekorieren. Dies wird mit Zufriedenheit in Bezug auf die Arbeitsumgebung assoziiert (Wells, 2000, S. 247). Außerdem sind sie mit dem eigenen Zuhause vertrauter (Brunia & Hartjes-Gosselink, 2009, S. 170) und haben mehr Kontrolle über den Bereich als über den Arbeitsplatz im Unternehmen (Brunia & Hartjes-Gosselink, 2009, S. 171). Tendenziell steigert hybrides Arbeiten somit laut Literatur die Territorialität der Arbeitnehmenden, da sie im Homeoffice einen eigenen Bereich haben, den sie personalisieren können, solange das Ausmaß als angemessen erachtet wird.

Non-territoriale Arbeitsplätze können hingegen zu einem geringeren Zugehörigkeitsgefühl und weniger Arbeitszufriedenheit führen (G. Brown, 2009, S. 24; Gatt und Jiang, 2021, S. 956), da die Möglichkeiten zur Personalisierung der Arbeitsumgebung limitiert sind (Wohlers & Hertel, 2017, S. 471). Dies wird auch in einer empirischen, qualitativen Studie von Elsbach (2003) belegt, welche ergeben hat, dass sich Arbeitnehmende durch non-territoriale Arbeitsplätze in ihrer Identität gefährdet fühlen können, da sie sich weniger durch Personalisierung des Arbeitsplatzes von anderen abgrenzen können. Um dem entgegen zu wirken werden teilweise neue Taktiken entwickelt, um die Arbeitsumgebung zu personalisieren und die Identität zu verdeutlichen (Brunia und Hartjes-Gosselink, 2009, S. 169; Elsbach, 2003, S. 634). Die Tendenz in der Literatur besagt entsprechend, dass non-territoriale Arbeitsplätze, bei denen sich jeden Tag ein freier Sitzplatz gesucht werden muss, im Vergleich zu territorialen Arbeitsplätzen, die Arbeitsbedingung Territorialität negativ beeinflussen.

Aufgrund der vorliegenden Ergebnisse der Wirkung hybrider und non-territorialer Arbeitsplätze auf die Arbeitsbedingung Territorialität wird folgender Mechanismus angenommen und werden folgende Hypothesen über deren Kombinationen aufgestellt:

A3:** Die Arbeitsplatzform beeinflusst die **Territorialität.

***H3a:** Ein **territorialer & hybrider** Arbeitsplatz geht mit der **größten** Territorialität einher.*

***H3b:** Ein **non-territorialer & nicht-hybrider** Arbeitsplatz geht mit der **geringsten** Territorialität einher.*

2.2.2. Privatsphäre

Neben der Territorialität wird in verwandter Literatur die Arbeitsbedingung *Privatsphäre* betrachtet (Ashkanasy et al., 2014; De Croon et al., 2005; Elsbach & Pratt, 2007; Wohlers & Hertel, 2017). Der zweite angenommene Mechanismus beschreibt demnach einen Effekt der Arbeitsplatzform auf die Arbeitsbedingung *Privatsphäre*, welche wiederum die Präferenz von Arbeitnehmenden beeinflussen kann. Dabei kann nach E. Sundstrom et al. (1980) zwischen zwei Arten von Privatsphäre unterschieden werden, der physischen und psychologischen Privatsphäre. Die physische Privatsphäre, auch architektonische Privatsphäre genannt, wird durch akustische und visuelle Isolation von der Umgebung hervorgerufen (E. Sundstrom et al., 1980, S. 102). Diese kann die psychologische Privatsphäre positiv beeinflussen (Wohlers & Hertel, 2017, S. 473), indem durch physische Privatsphäre das Gefühl der Kontrolle über den Zugang zu einem selbst oder der eigenen Gruppe hervorgerufen wird (E. Sundstrom et al., 1980, S. 101). Beide Arten werden in der Literatur in Bezug auf die Arbeitsumgebung untersucht, da die architektonische Privatsphäre auch in diesem Kontext die wahrgenommene Privatsphäre beeinflusst.

Bei der Optimierung der Privatsphäre handelt es sich laut der *Privacy Regulation*-Theorie nach Altman (1975)) um einen dynamischen Prozess, da Arbeitnehmende je nach Kontext unterschiedliche Level an Privatsphäre bevorzugen können (De Been & Beijer, 2014, S. 144). Obwohl es auch möglich ist zu viel Privatsphäre zu haben, liegt der Fokus der Literatur auf einem Mangel an Privatsphäre (Gove, 1978, S. 638). E. Sundstrom et al. (1980) haben damit einhergehend in drei Studien herausgefunden, dass Arbeitnehmende Privatsphäre im Gegensatz zur Erreichbarkeit und einem Mangel an Privatsphäre präferieren, da diese mit weniger Ablenkungen einhergeht (E. Sundstrom et al., 1980, S. 113). Dies kann durch das *Concept of overload* nach S. Cohen (1978) erklärt werden. Das Konzept besagt, dass eine größere wahrgenommene Kontrolle über äußere Einflüsse zu einer besseren Entspannung bei der Überwachung unvorhersehbarer Einflüsse führt, sodass mehr Aufmerksamkeit für weitere Anforderungen zur Verfügung steht. Eine Unzufriedenheit bezüglich der Privatsphäre kann somit bei zu vielen äußeren Einflüssen die Fähigkeit verschlechtern, die eigentlichen Arbeitsaufgaben zu erfüllen (De Been & Beijer, 2014, S. 144). Des Weiteren wird es als essentiell betrachtet, dem Bedürfnis der Arbeitnehmenden nach ausreichender Privatsphäre nachzugehen, um eine positive Wahrnehmung der Arbeitsumgebung zu erzielen (Haapakangas et al., 2018, S. 74) und die Arbeitsmotivation aufrecht zu erhalten (Inamizuu, 2013, S. 115).

In Bezug auf das *hybride Arbeiten* konnte mittels eines Experimentes von Appel-Meulenbroek et al. (2022) bestätigt werden, dass Arbeitnehmende nur bereit sind vermehrt im Unternehmen zu arbeiten, wenn der Arbeitsplatz genug private Bereiche bietet, um konzentriert arbeiten zu können. Daraus lässt sich schlussfolgern, dass hybrides Arbeiten durch den teilweisen Verbleib im Homeoffice mehr Privatsphäre bietet. Eine weitere qualitative Studie von Sewell und Taskin (2015) hat ergeben, dass das Arbeiten im Homeoffice genutzt werden kann, um Ablenkungen durch Lärm am Arbeitsplatz und Unterbrechungen durch Teammitgliedern zu entkommen. Die Gefahr der zu starken Isolierung durch zu große Privatsphäre beim hybriden Arbeiten wird in der Literatur nur in geringem Maß belegt, da Arbeitnehmende alternative Ressourcen wie IKT nutzen, um erreichbar zu bleiben (Halford, 2005, S. 28–29). In der Literatur wird entsprechend überwiegend eine Verbesserung der Privatsphäre durch das hybride Arbeiten belegt, da das teilweise Arbeiten im Homeoffice die Kontrolle über den Zugang zum Arbeitsplatz erhöht und die Sichtbarkeit reduziert.

Eine Fallstudie von Volker und van der Voordt (2005), in der Auswirkungen des *non-territorialen Arbeitens* untersucht werden, zeigt hingegen, dass zwar nicht non-territoriales Arbeiten selbst zu einer geringeren Zufriedenheit der Arbeitnehmenden führt, aber mit einem Mangel an visueller und akustischer Privatsphäre einhergeht. Zu diesem Ergebnis kommen auch Kim et al. (2016), die herausstellen, dass mit dem non-territorialen Arbeiten einhergehende Merkmale und räumliche Faktoren zu geringerer Privatsphäre führen können (Kim et al., 2016, S. 203). Dies lässt sich darauf zurückführen, dass non-territoriale Arbeitskonzepte häufig in Großraumbüros angewandt werden, welche die Privatsphäre verringern können (De Croon et al., 2005, S. 127). Darin kann es zu mehr Unterbrechungen kommen, welche Arbeitnehmende überfordern können (Volker & van der Voordt, 2005, S. 242), da weniger Kontrolle über die Interaktionen ausgeübt werden kann (De Croon et al., 2005, S. 129). Durch die reduzierte Privatsphäre kommt es bei fast 40% der Befragten in der Fallstudie von Volker und van der Voordt (2005) zu Unzufriedenheit. Insbesondere Arbeitnehmende, die sich regelmäßig konzentrieren müssen, wünschen sich häufiger territoriale Arbeitsplätze (Volker & van der Voordt, 2005, S. 247). Neben den negativen Auswirkungen auf die Privatsphäre durch die mit non-territorialem Arbeiten oftmals einhergehenden Großraumbüros belegt ein Experiment von Allen und Gerstberger (1973) jedoch, dass auch der Aspekt der freien Sitzplatzwahl beachtet werden muss. Diese wurde zuvor wenig betrachtet, kann aber dazu beitragen, mehr Kontrolle über Unterbrechungen und Interaktionen zu erlangen (Inamizuu, 2013, S. 111) und wird in aktuelleren Studien berücksichtigt. Folglich lässt sich daraus schließen, dass beide Aspekte nicht getrennt voneinander betrachtet werden sollten, insgesamt sich aber die Tendenz anhand der Literatur ableiten lässt, dass non-territoriales Arbeiten die Privatsphäre reduziert.

In Kombination mit hybridem Arbeiten wird folgender zweiter Mechanismus angenommen:

> ***A4:*** *Die Arbeitsplatzform beeinflusst die* ***Privatsphäre****.*
>
> ***H4a:*** *Ein* ***territorialer & hybrider*** *Arbeitsplatz geht mit der* ***größten*** *Privatsphäre einher.*
>
> ***H4b:*** *Ein* ***non-territorialer & nicht-hybrider*** *Arbeitsplatz geht mit der* ***geringsten*** *Privatsphäre einher.*

2.2.3. Zusammenarbeit

Als weitere relevante Merkmale der Arbeitsplatzform werden in der Literatur, die verwandte Arbeitsplatzkonzepte untersucht, die Dichte an Arbeitnehmenden in einer Umgebung, die physische Nähe zwischen den Arbeitnehmenden und die Sichtbarkeit herausgestellt, welche neben der Territorialität und Privatsphäre auch die Interaktionen und die Kommunikation zwischen den Arbeitnehmenden beeinflussen (Allen und Gerstberger, 1973, S. 487; Ashkanasy et al., 2014, S. 1172; Wohlers und Hertel, 2017, S. 475). Diese Aspekte können unter der Arbeitsbedingung *Zusammenarbeit* zusammengefasst werden. Zusammenarbeit meint dabei das Zusammenkommen verschiedener Menschen und Interessen, mit der Absicht ein gemeinsames Ziel zu erreichen, durch Interaktionen, den Austausch von Informationen und die Koordinierung der Aktivitäten (Jassawalla & Sashittal, 1998, S. 239). Der dritte angenommene Mechanismus beschreibt demnach einen Effekt der Arbeitsplatzform auf die Zusammenarbeit, wodurch die Präferenz von Arbeitnehmenden beeinflusst werden kann.

Die Beeinflussung der Zusammenarbeit durch die zuvor genannten Merkmale der Dichte und physischen Nähe kann mittels der *Social Interference*-Theorie begründet werden, welche besagt, dass die Möglichkeiten der Kommunikation und Interaktion mit anderen Arbeitnehmenden in Umgebungen mit größerer Nähe zwischen den Individuen erhöht sind (Davis et al., 2020, S. 950). Dies betrifft insbesondere ungeplante Interaktionen, die beispielsweise durch eine größere Sichtbarkeit entstehen können. Geplante Meetings und absichtliche Treffen werden von den physischen Gegebenheiten in geringerem Maß beeinflusst (Appel-Meulenbroek et al., 2022, S. 3; Gordon Brown, 2008, S. 12; Wohlers und Hertel, 2017, S. 475). Obwohl ungeplante Interaktionen einerseits die zuvor beschriebene Ressource Privatsphäre reduzieren können, können sie andererseits die Kommunikation zwischen den Arbeitnehmenden erleichtern und damit auch zu erhöhter Zusammenarbeit führen (Ashkanasy et al., 2014, S. 1172; De Been und Beijer, 2014, S. 145). Die Zusammenarbeit ist sowohl für die Arbeitsleistung (Allen & Gerstberger, 1973, S. 487) als auch für die Entstehung engerer Beziehungen zwischen Arbeitnehmenden relevant (Müller et al., 2022, S. 4). Dass vermehrte Zusammenarbeit als positiv wahrgenommen werden kann, hat eine empirische Studie von Morrow et al. (2012) ergeben. Sie konnten belegen, dass Zusammenarbeit als Mediator zwischen Veränderungen der Arbeitsumgebung und dem Zugehörigkeitsgefühl der Arbeitnehmenden agiert und erhöhte Zusammenarbeit mit einem höheren Zugehörigkeitsgefühl einhergeht (Morrow et al., 2012, S. 103).

Durch *hybrides Arbeiten* kann die Dichte und physische Nähe zwischen Arbeitnehmenden verringert werden, da dieses mit einer zumindest teilweisen örtlichen Trennung einhergeht. Infolge können Arbeitnehmende die Isolation fürchten (Müller et al., 2022, S. 4; Wohlers und Hertel, 2017, S. 476). Um einer verringerten Anzahl an ungeplanten Interaktionen und weniger Kommunikation entgegen zu wirken, können Arbeitnehmende den Wunsch verspüren vermehrt vor Ort im Unternehmen zu arbeiten, um an geplanter Kommunikation teilzunehmen oder ungeplante Interaktionen wahrscheinlicher zu machen (Appel-Meulenbroek et al., 2022, S. 3). Alternativ können Meetings geplant werden, um spontane Interaktionen zu ersetzen, diese Interaktionen werden aber in der Studie von Sewell und Taskin (2015) als unflexibler und starrer wahrgenommen. Besonders für Arbeitnehmende, deren Arbeit viel Kommunikation erfordert, war in der bereits beschriebenen Studie von Appel-Meulenbroek et al. (2022) eine Rückkehr aus dem Homeoffice attraktiver. Wenige Studien belegen jedoch, dass hybrides Arbeiten keinen negativen Effekt auf die Beziehungen der Arbeitnehmenden haben muss, da beispielsweise geplante Interaktionen qualitativ hochwertiger sein können und somit die Beziehung zwischen Arbeitnehmenden verbessern können (Gajendran und Harrison, 2007, S. 1538; Halford, 2005, S. 28). Insgesamt lässt sich dennoch feststellen, dass hybrides Arbeiten die Interaktion und Kommunikation zwischen Arbeitnehmenden aufgrund der teilweisen räumlichen Trennung verringert (Wohlers & Hertel, 2017, S. 475) und damit die Zusammenarbeit nicht gefördert wird.

Non-territoriales Arbeiten kann hingegen die Kommunikation zwischen Arbeitnehmenden steigern, da durch das Desksharing und damit einhergehende Platzwechsel vermehrte Optionen zur Interaktion mit unterschiedlichen Arbeitnehmenden zur Verfügung stehen (De Croon et al., 2005, S. 131; Kim et al., 2016, S. 206). Dies betrifft besonders Arbeitnehmende, die nicht im eigenen Team sind. Teammitglieder können durch variable Entfernungen nur in verringertem Umfang interagieren und kommunizieren, da sie keinen festen gemeinsam genutzten Bereich mehr haben (Allen und Gerstberger, 1973, S. 495; Wohlers und Hertel, 2017, S. 475). Wenn genug freie Plätze zur Auswahl stehen, können non-territoriale Arbeitsplätze den Arbeitnehmenden jedoch erlauben zu entscheiden, mit wem sie interagieren wollen und damit die Kommunikation und Zusammenarbeit erleichtern (Gatt und Jiang, 2021, S. 954; Gonsalves, 2023, S. 10; Volker und van der Voordt, 2005, S. 247). Entsprechend kann die Zusammenarbeit durch non-territoriales Arbeiten besonders zwischen Arbeitnehmenden unterschiedlicher Teams gefördert werden, die Zusammenarbeit mit Teammitgliedern jedoch nur wenn Sitzplätze in der Nähe voneinander gewählt werden.

Es lässt sich folgender dritter Mechanismus und die Wirkung der Kombination hybriden und non-territorialen Arbeitens hypothetisieren:

> ***A5:*** *Die Arbeitsplatzform beeinflusst die* ***Zusammenarbeit****.*
>
> ***H5a:*** *Ein* ***non-territorialer & nicht-hybrider*** *Arbeitsplatz geht mit der* ***größten*** *Zusammenarbeit einher.*
>
> ***H5b:*** *Ein* ***territorialer & hybrider*** *Arbeitsplatz geht mit der* ***geringsten*** *Zusammenarbeit einher.*

2.2.4. Autonomie

Der vierte und damit letzte angenommene Mechanismus in Bezug auf hybrides und non-territoriales Arbeiten beschreibt einen Einfluss der Arbeitsplatzform auf die Arbeitsbedingung *Autonomie* und daraus folgende arbeitsbezogene Konsequenzen (Gatt & Jiang, 2021; Wohlers & Hertel, 2017). Autonomie meint nach Deci und Ryan (1987) den eigenen Handlungen selbst zuzustimmen und das Gefühl zu haben, dass diese Handlungen von einem selbst ausgehen und man für sie verantwortlich ist (Deci & Ryan, 1987, S. 1025). Dabei wird eine Wahlmöglichkeit erlebt, sodass Handlungen mit Willenskraft und Unabhängigkeit ausgeführt werden (Deci & Ryan, 2000, S. 231). Das Umfeld kann Autonomie fördern, wenn Individuen durch Umweltbedingungen dabei unterstützt werden, über eigene Handlungen zu entscheiden (Gatt & Jiang, 2021, S. 960). In Bezug auf die Arbeitswelt definieren Hackman und Oldham (1975) Autonomie als den Grad der Freiheit, Unabhängigkeit und Diskretion, den der Job bietet, indem Arbeitnehmende über die Zeiteinteilung und verwendete Methoden selbst bestimmen können (Hackman & Oldham, 1975, S. 162). Neue Arbeitsformen weisen neben einer flexibleren zeitlichen Einteilung außerdem diverse Optionen in Bezug auf den Arbeitsort auf, wodurch die Autonomie gesteigert werden kann (Demerouti et al., 2014, S. 124). Im Kontext der neuen Arbeitsplatzformen wird folglich insbesondere die Autonomie der Arbeitnehmenden bezüglich der Arbeitszeit und des Arbeitsortes betrachtet (Wohlers & Hertel, 2017, S. 472).

Autonomie kann laut der *Self-determination*-Theorie nach Deci und Ryan (2000) neben den zwei weiteren psychologischen Grundbedürfnissen, Kompetenz und soziale Eingebundenheit, als Mediator zwischen diversen Einflüssen der Umwelt und der Motivation der Arbeitnehmenden agieren (Deci et al., 2017, S. 22). Die Arbeitsplatzform kann Einflüsse der Umwelt verändern, welche wiederum die Autonomie als Arbeitsbedingung beeinflussen. Wenn sich Arbeitnehmende durch eine Arbeitsplatzform in ihrer Autonomie gestärkt fühlen, kann die Arbeitszufriedenheit gesteigert werden, da eine Unterstützung der Autonomie mit einer höheren intrinsischen Motivation assoziiert wird (Deci & Ryan, 1987, S. 1024). Die Theorie besagt weiterhin, dass die Befriedigung der psychologischen Grundbedürfnisse neben einer gesteigerten autonomen Motivation zu einem höheren Wohlbefinden und effektiveren Leistungen führt (Deci et al., 2017, S. 39, S. 20). Arbeitsplätze, die einen geringeren Grad an Autonomie aufweisen, können hingegen eine Umgebung darstellen, die nicht zur Befriedigung der Grundbedürfnisse beiträgt (Gatt & Jiang, 2021, S. 959). Dies kann sich negativ auf die Motivation der Individuen auswirken, sodass externe Motivatoren vermehrt notwendig werden können (Deci et al., 2017, S. 22).

In Bezug auf das *hybride Arbeiten* liegt die Annahme nah, dass Arbeitnehmende dabei mehr Autonomie besitzen, da die Flexibilität bezüglich des Arbeitsortes größer ist. Diese erhöht wiederum auch die Freiheit in der Ausführung, was die zeitliche Einteilung wie Pausen betrifft, oder weitere Faktoren wie die Gestaltung der Arbeitsumgebung (Gajendran & Harrison, 2007, S. 1526). Wie ein systematischer Literaturüberblick von De Croon et al. (2005) zeigt, konnte dem Arbeitsort allerdings kein eindeutiger Effekt auf die Autonomie zugeschrieben werden. Während manche Studien eine erhöhte Autonomie beim hybriden Arbeiten zeigen, konnte bei anderen kein Effekt belegt werden (De Croon et al., 2005, S. 124). Vereinzelte Studien zeigen, dass hybrides Arbeiten zwar zu einem höheren Grad an Autonomie führen kann, Manager dafür aber andere Bereiche stärker kontrollieren können (Sewell & Taskin, 2015, S. 1525). Gajendran und Harrison (2007) haben mithilfe einer Meta-Analyse die Wirkung von hybridem Arbeiten auf die Autonomie als psychologischen Mediator anhand von 46 Studien untersucht. Sie konnten zwar nur einen kleinen, aber positiven Effekt auf die wahrgenommene Autonomie und eine gesteigerte Arbeitszufriedenheit nachweisen (Gajendran & Harrison, 2007, S. 1538). Damit lässt sich insgesamt die Tendenz ableiten, dass hybrides Arbeiten die Autonomie leicht steigert, auch wenn nicht alle Studien einen Effekt belegen.

Demerouti et al. (2014) stellen heraus, dass die charakteristischen Merkmale neuer Arbeitsplatzkonzepte durch neue Technologien eine größere Autonomie bei der Gestaltung des Arbeitstages in Bezug auf den Ort, die Zeit und die Kommunikation bieten können (Demerouti et al., 2014, S. 124, S. 130). Neben dem hybriden Arbeiten gilt vorheriges auch für das *non-territoriale Arbeiten*, da Desksharing und IKT den Arbeitnehmenden eine hohe Flexibilität bieten. Dies betrifft insbesondere den Arbeitsort, die Arbeitszeit (Wohlers & Hertel, 2017, S. 472) und die Personen, mit denen man zusammensitzt (Gatt & Jiang, 2021, S. 971). Non-territoriales Arbeiten beeinflusst die wahrgenommene Autonomie der Arbeitnehmenden, da es die Möglichkeit schafft, den Arbeitsort auszuwählen (Kim et al., 2016, S. 204). Eine Studie von Gatt und Jiang (2021) kann empirisch belegen, dass die Befriedigung des Bedürfnisses nach Autonomie positive Auswirkungen vorhersagen kann, da Autonomie über den Arbeitsort als Mediator mit einer höheren Arbeitszufriedenheit der Arbeitnehmenden einhergeht (Gatt & Jiang, 2021, S. 954). Dies kann mit der Self-determination Theorie erklärt werden, da eine größere Selbstbestimmung möglich ist (Gatt & Jiang, 2021, S. 960). Es liegt somit eine klare Tendenz vor, dass non-territoriales Arbeiten mit einer größeren Autonomie einhergeht.

In Bezug auf die Arbeitsplatzform und die Kombination hybriden und non-territorialen Arbeitens lässt sich folgender Mechanismus hypothetisieren:

> ***A6:*** *Die Arbeitsplatzform beeinflusst die* ***Autonomie****.*
>
> ***H6a:*** *Ein* ***non-territorialer & hybrider*** *Arbeitsplatz geht mit der* ***größten*** *Autonomie einher.*
>
> ***H6b:*** *Ein* ***territorialer & nicht-hybrider*** *Arbeitsplatz geht mit der* ***geringsten*** *Autonomie einher.*

2.2.5. Präferenz und Zahlungsbereitschaft

Die zuvor im Rahmen der Betrachtung der zugrunde liegenden Mechanismen getroffenen Annahmen und Hypothesen auf Basis der Theorie und Literatur werden folgend genutzt, um Hypothesen darüber aufzustellen, welche der vier Arbeitsplatzformen insgesamt mit der größten beziehungsweise geringsten Präferenz und Bereitschaft auf Gehalt zu verzichten einhergehen.

In Bezug auf hybrides Arbeiten weisen die betrachteten Mechanismen darauf hin, dass dieses gegenüber dem nicht-hybriden Arbeiten bevorzugt wird, da Territorialität, Privatsphäre und Autonomie laut Literatur tendenziell gesteigert werden und nur die Zusammenarbeit bei nicht-hybridem Arbeiten gesteigert wird. Dies bestätigt auch die Meta-Analyse nach Gajendran und Harrison (2007, S. 1538), welche mit über 46 Studien und mehr als 12.000 Arbeitnehmenden belegt, dass die positiven Auswirkungen durch hybrides Arbeiten die negativen leicht überwiegen, sodass dieses gegenüber nicht-hybridem Arbeiten tendenziell bevorzugt wird.

In Bezug auf non-territoriales und territoriales Arbeiten weisen die vermuteten Mechanismen weniger eindeutige Ergebnisse auf, da territoriale Arbeitsformen tendenziell mit größerer Territorialität und Privatsphäre einhergehen, non-territoriale Arbeitsplatzformen aber die Zusammenarbeit und Autonomie positiv zu beeinflussen scheinen. Morrison und Macky (2017, S. 112) führen jedoch eine Studie auf Basis des JDR-Modells durch, welche belegt, dass non-territoriale Arbeitsplatzformen mit erhöhten Anforderungen einhergehen. Insgesamt kann somit trotz kontroverser Ergebnisse hinsichtlich der Arbeitsbedingungen als Ressourcen davon ausgegangen werden, dass territoriales Arbeiten gegenüber dem non-territorialen Arbeiten leicht bevorzugt wird.

Daraus abgeleitet werden folgende Hypothesen darüber aufgestellt, welche Arbeitsplatzformen mit der größten und welche Arbeitsplatzform mit der geringsten Präferenz beziehungsweise Bereitschaft auf Gehalt zu verzichten einhergehen:

> ***H1:*** *Insgesamt weist die* ***territoriale & hybride*** *Arbeitsplatzform die* ***größte Präferenz*** *auf und die* ***non-territoriale & nicht-hybride*** *Arbeitsplatzform die* ***geringste.***
>
> ***H2:*** *Insgesamt sind Arbeitnehmende für die* ***territoriale & hybride*** *Arbeitsplatzform bereit auf* ***am meisten Gehalt zu verzichten****, für die* ***non-territoriale & nicht-hybride*** *Arbeitsplatzform sind sie* ***am wenigsten*** *bereit auf* ***Gehalt zu verzichten.***

3. Methode

Um die zuvor anhand bestehender Theorien aufgestellten Annahmen und Hypothesen deduktiv empirisch zu testen, werden Primärdaten mittels einer Online-Umfrage erhoben. Es werden keine Sekundärdaten genutzt, da keine inhaltlich geeigneten identifiziert werden konnten (Kaya, 2009, S. 50), welche die Auswirkungen der Kombination hybrider und non-territorialer Arbeitsplatzformen ausreichend erfassen. In der Umfrage wird eine experimentelle Vignettenstudie in Form einer Conjoint-Analyse durchgeführt, um quantitative Daten bezüglich der Präferenz der Arbeitsplatzform zu sammeln. Die Wirkung der vier unterschiedlichen Kombinationen der Arbeitsplatzformen auf die Arbeitsbedingungen wird anhand eines standardisierten Fragebogens mithilfe verschiedener Items ebenfalls quantitativ abgefragt. Im Folgenden werden das verwendete Studiendesign inklusive Operationalisierung der Variablen, die Datenerhebung inklusive Stichprobenauswahl und Aufbau der finalen Umfrage, und die einzelnen Schritte der durchgeführten Datenanalyse genauer vorgestellt.

3.1. Studiendesign

Da Experimente helfen können systematisch hypothetische Kausalbeziehungen zu überprüfen (Aguinis und Bradley, 2014, S. 352; Rack und Christophersen, 2009, S. 31; Thompson et al., 2015, S. 727), werden diese auch im Kontext der Jobsuche genutzt. Dabei kann etwa untersucht werden, wie Job-Charakteristiken die Präferenz von Arbeitnehmenden bei der Jobauswahl beeinflussen und welcher Wert diesen zugeschrieben wird (Bartz und Schwand, 2017, S. 4; Dalal und Singh, 1986, S. 555; R. Singh, 1975, S. 623). Dies kann mithilfe der *Experimental-Vignette*-Methode (EVM) nach Aguinis und Bradley (2014) erfolgen, welche in dieser Arbeit als Framework verwendet und durch eigene Aspekte ergänzt wird.

Die EVM besteht in der Regel aus zwei Teilen, dem Vignetten-Experiment als Kernelement und einer traditionellen Befragung zur Erfassung weiterer Merkmale und Charakteristiken der Befragten (Atzmüller & Steiner, 2010, S. 128). Das Vignetten-Experiment wird hier genutzt, um die Präferenz und die Bereitschaft auf Gehalt zu verzichten, in Bezug auf Jobangebote mit verschiedenen Arbeitsplatzkonzepten zu untersuchen. Damit soll die erste Forschungsfrage beantwortet werden, indem Annahme 1 und 2 und Hypothese 1 und 2 getestet werden. Die traditionelle Befragung zielt auf Forschungsfrage 2 und das Testen der Hypothesen 3-6 ab, welche mögliche Mechanismen vorhersagen.

Bei der EVM werden Befragten alternative Vignetten präsentiert, die kurze konstruierte Beschreibungen einer Person, eines Gegenstandes oder eines Szenariums beinhalten, welche systematisch verschiedene Kombinationen der Charakteristiken aufweisen (Atzmüller & Steiner, 2010, S. 128). Dadurch können abhängige Variablen, wie die Einstellung der Befragten, erfasst werden (Aguinis & Bradley, 2014, S. 352). Es kann auch die Relevanz der in den Vignetten enthaltenen Charakteristiken ermittelt werden, welche die realitätsnahe Entscheidung kausal beeinflussen (Atzmüller & Steiner, 2010, S. 129). Entsprechend ist die EVM für diese Arbeit eine geeignete Methode, wie *Entscheidungspunkt 1* fordert (Aguinis & Bradley, 2014, S. 357), da realitätsnahe Szenarien erstellt werden können, unabhängige Variablen aber gleichzeitig kontrolliert und manipuliert werden können, sodass die interne und externe Validität gesteigert werden kann (Aguinis & Bradley, 2014, S. 352). Damit kann die Verwendung der EVM zu wertvollen Einblicken führen, da Wissen über kausale Beziehungen gewonnen werden kann (Aguinis & Bradley, 2014, S. 353). Insbesondere die Kombination des Vignetten-Experiments mit der traditionellen Befragung liefert einen Mehrwert, da die Schwachstellen beider Methoden ausgeglichen werden können. Während Befragungen durch Repräsentativität tendenziell eine hohe externe Validität aufweisen, kann die interne Validität durch passive Erfassung ohne Intervention schwieriger sichergestellt werden, wohingegen Experimente durch manipulierte Designs mit interner Validität überzeugen, während die externe Validität durch die Vereinfachung der Situation leiden kann (Atzmüller & Steiner, 2010, S. 128).

Die EVM kann in Form einer *Conjoint-Analyse* angewendet werden, bei der Befragte Entscheidungen zwischen verschiedenen Vignetten mit manipulierten Variablen im Vergleich treffen müssen, um das implizite Urteil zu erfassen, etwa indem die Präferenz bezüglich der Vignetten im Vergleich angegeben werden muss (Aguinis & Bradley, 2014, S. 354). Im Kontext der Arbeitsplatzform können dazu Jobangebote mit wechselnden Charakteristiken unterbreitet werden, welche anschließend entsprechend der Präferenz in eine Rangfolge gebracht werden müssen (Bartz & Schwand, 2017, S. 4). Die Conjoint-Analyse kann dabei für diese Studie von Vorteil sein, wie *Entscheidungspunkt 2* fordert, da Jobangebote mit verschiedenen Charakteristiken näher an realen Situationen sind als separate Bewertungen einzelner Jobcharakteristiken, da bei den Jobangeboten zwischen unterschiedlichen Charakteristiken abgewogen werden muss (Radermacher et al., 2017, S. 78).

In *Entscheidungspunkt 3* wird die Art des Designs gewählt (Aguinis & Bradley, 2014, S. 360). Es wird ein *Within-Subject-Design* gewählt, bei dem jede Person dieselben Vignetten sieht, sodass Vergleiche zwischen den Vignetten gezogen werden können und der Bewertungsprozess aufgedeckt werden kann (Aguinis und Bradley, 2014, S. 361; Thompson et al. (2015, S. 735)).

In *Entscheidungspunkt 4* wird das Level der Immersion bestimmt (Aguinis & Bradley, 2014, S. 361). Da diese kostspielig ist, wird auf eine realistischere Darstellung des Szenariums durch Bilder oder ähnliches allerdings verzichtet. Das Szenario wird lediglich beschrieben, indem Befragte sich vorstellen sollen, dass sie eine Einstiegsposition als Controller*in suchen und in naher Zukunft ihr Masterstudium der Betriebswirtschaftslehre absolvieren. Anschließend werden in *Entscheidungspunkt 5* die Anzahl und die Ausprägungen der Charakteristiken anhand der Theorie festgelegt, da diese Charakteristiken liefert, die vermutlich die Entscheidung beeinflussen (Aguinis & Bradley, 2014, S. 359). In dem untersuchten Zusammenhang werden drei Charakteristiken des Jobangebots miteinbezogen, das Einstiegsgehalt und in Bezug auf die Arbeitsplatzform, ob hybrides Arbeiten oder non-territoriales Arbeiten angewendet werden. Dabei sollte die Anzahl möglichst gering gehalten werden, sodass nicht zu viele Vignetten gebildet werden können, aber die wichtigsten Variablen enthalten sind (Aiman-Smith et al., 2002, S. 393; Atzmüller und Steiner, 2010, S. 136). In der Regel gibt es pro Variable zwei bis drei Ausprägungen, die numerisch oder kategorial, häufig in Form von Dummy-Variablen, daherkommen (Aiman-Smith et al., 2002, S. 396). Das festgelegte jährliche *Einstiegsgehalt* orientiert sich dabei an wahren Durchschnittswerten (Aguinis & Bradley, 2014, S. 362) von Einsteiger*innen im Controlling mit den Ausprägungen 44.000 €, 45.000 € und 46.000 €. [2]. Bei der kategorialen Variable *hybrides Arbeiten* wird zwischen „Sie arbeiten hybrid, 2 Tage pro Woche im Büro und 3 im Homeoffice." und nicht-hybridem Arbeiten „Sie arbeiten nur im Büro." unterschieden. Bei der kategorialen Variable *non-territoriales Arbeiten* wird zwischen non-territorialem Arbeiten „Im Büro suchen Sie sich täglich einen freien Sitzplatz, den Sie am Ende des Tages verlassen müssen." und territorialem Arbeiten „Sie haben einen festen Sitzplatz an einem Ihnen zugeordneten Schreibtisch im Büro." unterschieden. Dadurch sollen auch Befragte, welche die Begriffe nicht kennen, verstehen, was gemeint ist. Außerdem wurde entschieden, dass jedes Jobangebot alle Charakteristiken enthält, wie Aiman-Smith et al. (2002) beschreibt. Das Format der abhängigen Variable des Vignetten-Experiments sollte dem Format einer realen Entscheidungssituation entsprechen (Aiman-Smith et al., 2002, S. 393). Befragte würden in der Realität vermutlich eine Rangfolge ihrer Präferenz bilden, um zu entscheiden, welches Jobangebot sie annehmen möchten. Demnach wurde als Format der abhängigen Variable die Bildung einer Rangfolge entsprechend der Präferenzen festgelegt.

Anschließend muss in *Entscheidungspunkt 6* die Anzahl der Vignetten festgelegt werden, wobei nicht zu viele präsentiert werden sollten, um die Befragten nicht zu überfordern (Aguinis & Bradley, 2014, S. 362). Bei den vorliegenden Charakteristiken und Ausprägungen ergeben sich $3 \times 2 \times 2 = 12$ mögliche Kombinationen. Um die Bildung einer Rangfolge zu vereinfachen, wurde auf ein *fractional factorial* Design gesetzt (Atzmüller & Steiner, 2010, S. 131), indem in SPSS mittels orthogonalem Verfahren die Vignetten auf 8 reduziert wurden (Radermacher et al., 2017, S. 81), welche aus Ta-

[2] Basierend auf dem Gehaltscheck nach „Gehaltscheck: Gehalt für Controller:in" (2023) für Controller*innen mit weniger als 3 Jahren Berufserfahrung und ohne Personalverantwortung unter https://www.kununu.com/de/gehalt/controller-in-30993

belle 1 entnommen werden können.

Nach Atzmüller und Steiner (2010) wurde neben dem Vignetten-Experiment eine *traditionelle Befragung* zur Erfassung der den Entscheidungen zugrundeliegenden Mechanismen und persönlichen Kontrollvariablen erstellt. Dazu wurde ein Fragebogen als Instrument der standardisierten Befragung gewählt (Kaya, 2009, S. 51), bei dessen Erstellung darauf geachtet wurde, dass die Fragen eindeutig zu verstehen, einfach formuliert und für den Untersuchungsgegenstand relevant sind (Kaya, 2009, S. 54).

Um die Wirkung der verschiedenen Kombinationen der Arbeitsplatzformen auf die Arbeitsbedingungen abzufragen, wurde dieselbe Operationalisierung der unabhängigen kategorialen Variablen *hybrides Arbeiten* und *non-territoriales Arbeiten* wie in dem Vignetten-Experiment verwendet. Daraus wurden die vier möglichen Kombinationen der Arbeitsplatzform gebildet und die Befragten gebeten sich vorzustellen, in dieser Arbeitsplatzform zu arbeiten. Anschließend sollte erfasst werden, inwieweit die jeweilige Arbeitsplatzform die Arbeitsbedingungen Territorialität, Privatsphäre, Zusammenarbeit und Autonomie gewährt. Da es sich bei diesen Arbeitsbedingungen allerdings um abstrakte latente Konstrukte handelt, sind diese nicht direkt beobachtbar und messbar und es sollte nach El-Den et al. (2020) wenn möglich auf bereits validierte Messinstrumente zurückgegriffen werden. Bei diesen erfolgt die Messung des latenten Konstrukts indirekt über beobachtbare Variablen, die beispielsweise in Form von Antworten auf einer Skala bezüglich verschiedener Aussagen, welche auch als Items bezeichnet werden, erfasst werden (El-Den et al., 2020, S. 327).

Obwohl es in der Literatur eine Vielzahl an validierten Messinstrumenten gibt, ist es möglich, dass neue entwickelt werden müssen, um noch nicht messbare Konstrukte erfassen zu können oder um Konstrukte in neuen, wenig erforschten Zusammenhängen untersuchen zu können (El-Den et al., 2020, S. 326). Zur Messung der Wirkung der vorliegenden Arbeitsplatzformen auf die Arbeitsbedingungen konnten keine geeigneten Instrumente identifiziert werden. Dies bestätigt auch G. Brown (2009), welcher herausstellt, dass es in dem Forschungsgebiet des Arbeitsplatzes eine Vielzahl an Konstrukten wie Territorialität und Privatsphäre gibt, für die noch keine etablierten Messinstrumente existieren. So messen E. Sundstrom et al. (1980) die Privatsphäre etwa nur dichotom, indem Befragte angeben müssen, ob der Arbeitsplatz „privat“ oder „nicht privat“ ist. Darüber hinaus fallen für den hier untersuchten Einfluss durch die Arbeitsplatzform einige Messinstrumente raus, da andere Arbeitsbedingungen betrachtet werden. In der BOSSA Umfrage wird beispielsweise auch die visuelle Ästhetik der Umgebung abgefragt (Candido et al., 2016, S. 216) und in dem Work Design Questionnaire die Temperatur als Arbeitsbedingung und die Anforderungsvielfalt gemessen (Stegmann et al., 2010, S. 27), welche in dieser Arbeit nicht untersucht werden. Bei anderen Messinstrumenten werden für die hier untersuchten Arbeitsbedingungen Variablen beobachtet, die für hybride und non-territoriale Arbeitsplatzkonzepte keine oder eine untergeordnete Rolle spielen. So gibt es für das latente Konstrukt Autonomie zwar viele etablierte Messinstrumente, diese zielen aber vornehmlich auf die Autonomie hinsichtlich der Dimensionen Arbeitsmethode, Teamdesign oder strategische Entscheidungen ab (Lumpkin et al., 2009, S. 53), welche von hybriden und non-territorialen Arbeitsplatzkonzepten wenig beeinflusst werden.

Entsprechend mussten für diese Arbeit *neue Messinstrumente* entwickelt werden. Dazu wurden nach El-Den et al. (2020) aus der Literatur einzelne Items aus bestehenden Messinstrumenten verwendet oder neue Items anhand von Definitionen der Konstrukte entwickelt. Dabei wurde darauf geachtet diese unmissverständlich, klar und neutral zu formulieren und jeweils nur auf einen Aspekt abzuzielen (El-Den et al., 2020, S. 328). Da das Auffinden von exakt dem Konstrukt entsprechenden Items unmöglich ist, dürfen Items verwendet werden, welche Indikatoren für das untersuchte latente Konstrukt sind (Greving, 2009, S. 74). Pro Konstrukt wurden schließlich drei Items festgelegt. Um abzufragen, inwieweit die Arbeitsplatzform Territorialität ermöglicht, wurde etwa das Item „Ich habe ein starkes Gefühl von persönlichem Eigentum für meinen Arbeitsplatz.“ verwendet. In Bezug auf die Arbeitsbedingung Privatsphäre wurde beispielsweise das Item „Ich habe genügend Privatsphäre zum Arbeiten.“ verwendet und für Zusammenarbeit das Item „Mein Arbeitsplatz erlaubt es mir mit Kollegen zu interagieren.“. Die Wirkung der Arbeitsplatzform auf die Autonomie wurde beispielsweise über das Item „Mein Arbeitsplatz bietet einen hohen Grad der Selbstbestimmung.“ erfasst. In Tabelle 6 im Anhang ist eine vollständige Übersicht der verwendeten Items zur Erfassung der Wirkung der vier Kombinationen der Arbeitsplatzformen auf die vier Arbeitsbedingungen zu finden, in der auch Literaturnachweise hinterlegt sind.

Als Antwortformat wurde für alle latenten Konstrukte die meistverwendete 5-stufige *Likert-Skala* festgelegt, um die Einstellung der Befragten gegenüber den Items in Bezug auf die jeweils vorgestellte Arbeitsplatzform zu messen (Greving, 2009, S. 73). Die Likert-Skala ist von Vorteil, da sie im Allgemeinen als intervallskaliert angesehen werden kann und Einstellungen damit einfach und einheitlich gemessen werden können (Greving, 2009, S. 72; Radermacher, 2019, S. 94). Es werden 5 Stufen festgelegt, da die Reliabilität tendenziell bei 5 bis 7 Stufen am größten ist und diese Anzahl bei nicht allein stehenden Items empfohlen wird (Aiman-Smith et al., 2002, S. 394; Greving, 2009, S. 70). Die Antwortmöglichkeiten reichen dabei von „1 Stimme überhaupt nicht zu.“ bis zu „5 Stimme voll und ganz zu.“.

Neben den Mechanismen werden mittels der traditionellen Befragung auch *Kontrollvariablen* erhoben, um persönliche Merkmale der Befragten zu erfassen. Da eine Studie nach Volker und van der Voordt (2005) ergeben hat, dass das Alter und das Geschlecht die Bewertung von non-territorialen Arbeitsplätzen beeinflussen kann, werden diese erfasst. Das *Alter* wird als metrische Variable mittels Freitextfeld erfasst, das *Geschlecht* als nominale Variable durch Single-Choice (weiblich, männlich oder divers). In einer weiteren Studie konnte außerdem ein Einfluss durch den Bildungsstand und die Arbeitszeit belegt werden (Appel-Meulenbroek et al., 2022,

Tabelle 1: Die 8 verwendeten Vignetten in der Conjoint-Analyse.

		1	2	3	4	5	6	7	8
Hybrides Arbeiten	Sie arbeiten **hybrid**, 2 Tage pro Woche im Büro und 3 im Homeoffice.		X		X	X			X
	Sie arbeiten **nur im Büro**.	X		X			X	X	
Non-territoriales Arbeiten	Im Büro suchen Sie sich täglich einen **freien Sitzplatz**, den Sie am Ende des Tages verlassen müssen.			X	X	X		X	
	Sie haben einen **festen Sitzplatz** an einem Ihnen zugeordneten Schreibtisch im Büro.	X	X				X		X
Einstiegsgehalt	**44.000 €**	X			X			X	X
	45.000 €					X	X		
	46.000 €		X	X					

Hinweis: X kennzeichnet, dass die Ausprägung in der Vignette verwendet wurde.
Eigene Darstellung in Anlehnung an Radermacher et al. (2017, S. 93)

S. 1), sodass diese kategorial erhoben werden, wobei per Single-Choice der höchste *Bildungsabschluss* erfasst wird und per Multiple-Choice der derzeitige *Beschäftigungsstatus*. Zusätzlich wird abgefragt, ob die Befragten vorher schon einmal im Homeoffice gearbeitet haben oder in einem Büro ohne feste Sitzplätze, da vorherige Erfahrungen die Präferenz bezüglich dieser Arbeitsplatzformen beeinflussen könnten.

3.2. Datenerhebung

Vor der eigentlichen Datenerhebung wurde ein *Pretest* durchgeführt, um die Eignung und Länge der Umfrage zu prüfen (Aiman-Smith et al., 2002, S. 406; El-Den et al., 2020, S. 328; Kaya, 2009, S. 54). Dazu wurden acht per Convenience Sample ausgewählte Personen gebeten, an der Umfrage teilzunehmen und anschließend Feedback zur Verständlichkeit zu geben. Die Befragten haben durchschnittlich 9:45 Minuten zur Bearbeitung gebraucht. Entsprechend des Feedbacks wurde die Schriftstärke einzelner Wörter in den Beschreibungen der Ausprägungen in dem Vignetten-Experiment erhöht, um die Unterschiede zwischen den Jobangeboten plakativer zu gestalten und den Vergleich zu erleichtern. Die Abfrage der Mechanismen wurde angepasst. Diese sollte zunächst ordinal erfolgen, indem die vier Kombinationen der Arbeitsplatzformen nach Stärke der Assoziation mit der jeweiligen Arbeitsbedingung sortiert werden, da dafür Definitionen der Arbeitsbedingungen verwendet werden können und keine geeigneten Messinstrumente zur Verfügung standen. Die Definitionen wurden allerdings als zu abstrakt empfunden und die ordinale Abfrage als zu restriktiv, da dabei verschiedene Arbeitsplatzformen hinsichtlich der Arbeitsbedingungen nicht gleich bewertet werden können. Entsprechend wurden, wie zuvor beschrieben, Messinstrumente erstellt, sodass eine intervallskalierte Abfrage erfolgen konnte. Außerdem wurde auf die Erhebung des Einkommens als Kontrollvariable verzichtet, da Befragte nicht bereit sein könnten dieses preiszugeben.

Bevor die Umfrage durchgeführt werden konnte, wurde nach Aguinis und Bradley (2014) in *Entscheidungspunkt 7* zunächst das Sample an Befragten festgelegt, da die Stichprobe an die interessierende Grundgesamtheit angepasst werden sollte, um die externe Validität zu erhöhen. Dabei ist wichtig, dass das Thema der Umfrage und das Szenario der EVM den Befragten bekannt ist (Aguinis & Bradley, 2014, S. 363). Da das Thema Jobsuche weit verbreitet ist und die Verwendung von Studierendensamples hinsichtlich der Generalisierbarkeit diskutiert wird, wurde sich gegen eine Beschränkung des Samples entschieden und das *Convenience Sampling* angewendet (Kaya & Himme, 2009, S. 83). Um die Validität der Ergebnisse dennoch sicherstellen zu können, dienen die Kontrollvariablen wie Beschäftigungsstatus und Alter.

Die Umfrage wurde mit LimeSurvey erstellt und in Form einer *Internet-Befragung* durchgeführt. Diese stellt eine hohe Datengenauigkeit bei der Erhebung sicher, es tritt ein geringer Interviewer-Bias auf und der Zeitbedarf pro Erhebungsfall ist dabei gering (Kaya, 2009, S. 54). Die Umfrage wurde per Link, ergänzt durch einen kurzen Text über Thema, Zielgruppe, Dauer und Verlosung, versendet. Dadurch können in *Entscheidungspunkt 8* die Konditionen, in denen Befragte an der Umfrage teilnehmen, nicht kontrolliert werden (Aguinis & Bradley, 2014, S. 363). Da aber mehr Personen erreicht werden können und zu einem flexiblen Zeitpunkt daran teilnehmen können, kann eine größere Stichprobe erhoben werden. Die Umfrage wurde auf Deutsch erstellt und 3 Wochen, vom 24.07.2023 bis zum 15.08.2023, veröffentlicht.

In der *finalen Umfrage* (Abbildung 4 bis Abbildung 7 im Anhang) wurden die Teilnehmenden zunächst per Willkommensnachricht über den Zweck, die Dauer und den Umgang mit den erhobenen Daten informiert. Des Weiteren wurde bereits zu Beginn als Anreiz die Teilnahme an einer Verlosung von 3 × 20 € in Aussicht gestellt und eine E-Mail-Adresse für etwaige Rückfragen hinterlegt. Auf der ersten Seite der Umfrage wurde zu Beginn das hypothetische Szenarium kurz vorgestellt, sodass das Thema zwar verständlich wird, Befragte aber nicht ermüden (Aiman-Smith et al., 2002, S. 401). Darin wird ein bestimmter Beruf genannt, da die Art der Tätigkeit einen Einfluss auf die Zufriedenheit mit einer

Arbeitsplatzform haben kann (Volker & van der Voordt, 2005, S. 247). Controlling wurde als Arbeitsfeld festgelegt, da dieser Beruf in der Regel hybrides und non-territoriales Arbeiten erlaubt. Außerdem wird der Arbeitsplatz im Unternehmen näher beschrieben, wobei es sich um ein Großraumbüro mit 20 Sitzplätzen handelt. Um Reihenfolgeeffekte zu verhindern, wurden die 8 Vignetten mit unterschiedlichen Jobangeboten zufällig je Individuum angeordnet (Bustelo et al., 2020, S. 9; Thompson et al., 2015, S. 735). Anschließend wurde auf der zweiten Seite der Umfrage über die Items der Arbeitsbedingungen abgefragt, inwieweit die jeweiligen Kombinationen der Arbeitsplatzformen mit den Arbeitsbedingungen assoziiert werden. Im letzten Teil der Umfrage wurden die demografischen Daten erfasst und es konnte freiwillig die Angabe einer E-Mail-Adresse zur Teilnahme an der Verlosung erfolgen. Bei allen anderen Feldern handelte es sich um Pflichtfelder, um eine vollständige Datenerhebung für die anschließende Datenanalyse sicher zu stellen.

3.3. Datenanalyse

Um die Datenanalyse durchzuführen, wurden zunächst die mittels der Umfrage in LimeSurvey erhobenen Daten nach Excel exportiert und für die nachfolgenden Analysen in Stata 17.0 und SPSS aufbereitet.

Entsprechend dem *Entscheidungspunkt 9* muss anschließend eine geeignete Methode zur Analyse der Daten festgelegt werden (Aguinis & Bradley, 2014, S. 364). Die Datenanalyse umfasst insgesamt acht Schritte, wie in Tabelle 2 zu sehen ist. Im ersten Teil der Datenanalyse wird die vorliegende **Stichprobe** analysiert, indem in **Schritt 1** deskriptive Statistiken der finalen Stichprobenmerkmale und demografischen Daten erstellt werden.

Anschließend wird im zweiten Teil die **Forschungsfrage 1** beantwortet, indem die mittels des Vignetten-Experiments erhobenen Daten ausgewertet werden. Da durch das Studiendesign dieselbe Anzahl an Messungen pro Vignette erhoben werden, liegen ausgewogene Daten vor, welche die statistische Analyse und Interpretation der Effekte erleichtern (Atzmüller & Steiner, 2010, S. 133). Diese werden aufbereitet, indem für jeden Befragten für alle acht Ränge je eine Zeile erstellt wird. Um beantworten zu können, welche Arbeitsplatzform Arbeitnehmende präferieren und ob sie bereit sind, dafür auf Gehalt zu verzichten, wird zunächst in **Schritt 2** eine *deskriptive Statistik* des Rangmittelwertes je Vignette erstellt. Der Zusammenhang zwischen Arbeitsplatzform, Gehalt und Präferenz soll anschließend statistisch untersucht werden, sodass Annahme 1 getestet wird, welche besagt, dass die vier Arbeitsplatzformen die Präferenz der Arbeitnehmenden für ein Jobangebot beeinflussen. Dazu wird in **Schritt 3** zunächst ein *nicht-parametrischer Test* durchgeführt. Diese machen weniger strikte Annahmen über die verwendeten Daten und sind auch für kategoriale Daten wie die Arbeitsplatzform geeignet (Siegel, 1957, S. 13). Der verteilungsfreie Chi-Quadrat Test ist besonders nützlich, um Hypothesen über die Zusammenhänge von kategorialen Variablen zu testen, kann aber auch für ordinale Daten wie die Rangfolge verwendet werden (McHugh, 2013, S. 143–144). Dazu wird eine Kreuztabelle erstellt, die Zellen enthalten die absoluten Häufigkeiten des gewählten Ranges. Der Chi-Quadrat-Test wird durchgeführt, welcher testet, ob die Häufigkeiten der erwarteten Verteilung entsprechen oder nicht auf den Zufall zurückzuführen sind und entsprechend von einem Zusammenhang der Variablen auszugehen ist (McHugh, 2013, S. 146).

Da nicht-parametrische Tests zwar hinsichtlich der Generalität zu bevorzugen sind, aber eine geringere Stärke als parametrische Tests bei Erfüllung deren Annahmen aufweisen (Siegel, 1957, S. 14) und Kontrollvariablen bei letzteren hinzugezogen werden können, wird zusätzlich ein *parametrischer Test* in **Schritt 4** durchgeführt. Dadurch kann mehr über den in Annahme 1 vorausgesagten Zusammenhang zwischen Arbeitsplatzform und Präferenz der Befragten erfahren werden. Außerdem soll dadurch Annahme 2 überprüft werden, welche die Bereitschaft für eine Arbeitsplatzform auf Gehalt zu verzichten vorhersagt. Auch welche Arbeitsform präferiert wird und für welche die größte Bereitschaft besteht auf Gehalt zu verzichten, also Hypothese 1 und 2, sollen dadurch getestet werden. Dazu wird zunächst hergeleitet, welcher parametrische Test geeignet ist, um die Präferenz und den Wert der Attribute zu analysieren.

Der Conjoint-Analyse liegt ein multiattributives Präferenzstrukturmodell zugrunde, welches annimmt, dass sich der Gesamtnutzen einer Option aus den Teilnutzen der einzelnen Attribute ergibt (Klein, 2002, S. 10). Nach R. Singh (1975) kann dieses Modell auch auf Jobangebote übertragen werden. Entsprechend wird die Präferenz bezüglich unterschiedlicher Jobangebote durch deren Job-Charakteristiken bestimmt. Die wechselnden Charakteristiken bringen den Befragten dabei einen unterschiedlich großen Nutzen und der Nutzen des Jobangebotes ist entsprechend von der Kombination der Charakteristiken abhängig (Radermacher et al., 2017, S. 78). Die Beurteilung der unterschiedlichen Jobangebote und daraus folgende Präferenzbildung erfolgt letztlich mittels eines Nutzenvergleichs zwischen den Jobangeboten (Klein, 2002, S. 8). Weiterhin wird angenommen, dass die Teilnutzen der einzelnen Attribute additiv und linear integriert werden (R. Singh, 1975, S. 623). Dies kann damit begründet werden, dass die Entscheidungsfindung zum Großteil mathematischen Regeln folgt und Charakteristiken und deren Kombinationen je nach Nutzen bewertet werden, da Individuen als rational betrachtet werden können (Dalal & Singh, 1986). Entsprechend kann folgende allgemeine Nutzenfunktion von Person i für ein Jobangebot j aufgestellt werden, welches aus n Attributen besteht, und den zufälligen Fehlerterm ε enthält: $U_{ji} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n + \varepsilon_{ji}$

In Bezug auf die vorliegende Untersuchung ergibt sich folgende Nutzenfunktion:

$$U_{ji} = \beta_0 + \beta_1 \cdot Arbeitsplatzform + \beta_2 \cdot Einstiegsgehalt + \varepsilon_{ji}$$

Der Gesamtnutzen U_{ji} eines Jobangebotes wird durch die abhängige ordinale Variable *Rang* erfasst. Die Attribute des Jobangebotes, welche den Nutzen beeinflussen, werden als

Tabelle 2: Gliederung der Datenanalyse

Stichprobe		
Schritt 1:	Deskriptive Statistiken der finalen Stichprobe	
Forschungsfrage 1		
Schritt 2:	Deskriptive Statistik – Rangmittelwerte der Jobangebote	A1
Schritt 3:	Nicht-parametrischer Test – Chi-Quadrat-Test	A1
Schritt 4:	Parametrischer Test – Regression	
Schritt 4.1:	*Rank Ordered Logit*-Modell	A1, H1
Schritt 4.2:	MWTP	A2, H2
Schritt 5:	Berücksichtigung von Subgruppen	
Forschungsfrage 2		
Schritt 6:	Deskriptive Statistik – Mittelwerte der Arbeitsbedingung je Arbeitsplatzform	A3-A6
Schritt 7:	Items der erstellten Skalen überprüfen	
Schritt 7.1:	Korrelation zwischen den Items	
Schritt 7.2:	Interne Konsistenz der Items – Cronbachs Alpha	
Schritt 7.3:	Konstruktvalidität – Konfirmatorische Faktorenanalyse	
Schritt 8:	Nicht-parametrischer Test – Kruskal-Wallis Test	A3-A6
	Post hoc Dunn's Test	H3-H6

Quelle: Eigene Darstellung.

unabhängige Variablen *Arbeitsplatzform* und *Einstiegsgehalt* erfasst.

In **Schritt 4.1** wird die zuvor beschriebene Nutzenfunktion durch Schätzung eines Modells bestimmt, um Annahme 1 und Hypothese 1 in Kombination mit dem nicht-parametrischen Test zu überprüfen. Auf Basis der Annahme, dass der Nutzen eines Jobs von einer linearen Funktion der einzelnen Attribute abhängt, wird in der Literatur oftmals die Datenanalyse in Form einer linearen Regression durchgeführt (He et al., 2021, S. 722; Pouliakas und Theodossiou, 2010, S. 694; Radermacher et al., 2017, S. 81). Nach Cameron und Trivedi (2005) können für die Schätzung der Koeffizienten der Nutzenfunktion verschiedene Modelle hinzugezogen werden. Da es sich bei der abhängigen Variable um eine ordinale Variable handelt, welche eine natürliche Reihenfolge aufweist, können *Ordered Multinomial*-Modelle in Form des *Ordered Logit*- oder *Ordered Probit*-Modells verwendet werden. Die Vorzeichen der Koeffizienten β_n können bei dieser Methode hinsichtlich der Wirkung auf die abhängige Variable interpretiert werden (Cameron & Trivedi, 2005, S. 519–520). Bei diesem Modell werden nicht alle Beobachtungen, die sich auf eine Person beziehen, miteinander verknüpft. Da durch die Rangfolge allerdings mehrere Wahlen pro Individuum bekannt sind, sollte die Verknüpfung erfolgen. Auch bei der *Pooled-OLS*-Regression werden die Verknüpfungen nicht berücksichtigt, diese kann dennoch genutzt werden um Panel-Modelle zu schätzen (Cameron & Trivedi, 2005, S. 720). *Rank-Data*-Modelle berücksichtigen hingegen, dass eine Person voneinander abhängige Entscheidungen trifft (Cameron & Trivedi, 2005, S. 521)). Da Befragte eine Rangfolge der Alternativen bilden, eignet sich besonders das *Rank Ordered Logit*- Modell nach Beggs et al. (1981) als Spezialform des *Ordered Logit*-Modells. Das Modell lässt sich für Conjoint-Experimente verwenden (Herrmann et al., 2006, S. 133), da die darin erfasste Rangfolge eine natürliche Ordnung aufweist (Calfee et al., 2001, S. 699). Die Regression wird mit allen zuvor aufgeführten Schätzmodellen durchgeführt, um die Robustheit der Ergebnisse zu überprüfen.

Die Regressionsergebnisse werden in **Schritt 4.2** verwendet, um die Annahme 2 und Hypothese 2 zu testen, indem damit die *Marginal Willingness to Pay (MWTP)* berechnet wird. Die MWTP schreibt den Attributen einen impliziten monetären Wert zu, da diese angibt, auf wie viel Gehalt die Befragten bereit sind zu verzichten, um eine Vignette mit einem speziellen Attribut zu wählen, wenn alle anderen Attribute unverändert bleiben (Hirogaki, 2013, S. 543; Radermacher et al., 2017, S. 83). Die MWTP wird durch das Verhältnis der Koeffizienten des interessierenden Attributes und des Einstiegsgehaltes berechnet: $MWTP = \beta_n/\beta_2$

Dies ist möglich, da durch das *Rank Ordered Logit*-Modell eine Nutzenfunktion berechnet wird, in der die Koeffizienten jeder Variable deren Effekt bei Änderung des entsprechenden Attributes auf den Nutzen anzeigt (Calfee et al., 2001, S. 702). In Bezug auf die Jobangebote lässt sich dadurch bestimmen, auf wie viel Gehalt die Befragten bereit sind, für eine bestimmte Arbeitsplatzform zu verzichten.

In **Schritt 5** werden Schritt 4.1 und Schritt 4.2 außerdem für unterschiedliche Gruppen von Befragten durchgeführt, um für demografische Daten kontrollieren zu können.

Teil 3 der Datenanalyse bezieht sich schließlich auf **Forschungsfrage 2**, in der die dem Zusammenhang zwischen Arbeitsplatzform und Präferenz zugrundeliegenden Mechanismen untersucht werden und die Annahmen und Hypothesen 3-6 getestet werden. Dafür werden die Daten der standardisierten Umfrage hinzugezogen und so aufbereitet, dass die Bewertungen der Items je Arbeitsplatzform eine Zeile ergeben.

Um in **Schritt 6** eine aussagekräftige deskriptive Statistik der Mittelwerte der vier Arbeitsbedingungen je Arbeitsplatzform erstellen zu können, werden aus den 12 Items vier Skalen gebildet, welche je eine Arbeitsbedingung abbilden. Die Skalen entsprechen dem Mittelwert der drei Items der entsprechenden Arbeitsbedingung pro Zeile und geben damit Auskunft über die durchschnittliche Bewertung der jeweiligen Arbeitsbedingung in einer Arbeitsplatzform. Durch dieses Vorgehen ist es möglich, direkt abzulesen, ob die Werte hoch oder niedrig sind.

Da die Forschungsinstrumente für weitere Tests reliabel und valide sein müssen, um verlässliche Ergebnisse zu liefern (A. S. Singh, 2017, S. 791), werden in **Schritt 7** die verwendeten Items der Skalen über den inhaltlichen Aspekt hinaus statistisch überprüft. Dazu wird in **Schritt 7.1** zunächst der Zusammenhang zwischen den Items betrachtet, indem die Korrelation der Items nach Pearson und zusätzlich nach Spearman berechnet wird. Dabei sollten die Korrelationen der Items derselben Arbeitsbedingung größer sein als die Korrelationen der Items verschiedener Konstrukte (Radermacher, 2019, S. 119–120). In **Schritt 7.2** wird außerdem die Reliabilität der Skalen durch Cronbachs Alpha berechnet (El-Den et al., 2020, S. 329). Dafür wird die interne Konsistenz der in den Skalen enthaltenen Items getestet, welche Aufschluss darüber gibt, ob Items dasselbe Konstrukt messen (A. S. Singh, 2017, S. 797; Tavakol und Dennick, 2011, S. 53). Da der Fragebogen Items für die vier verschiedenen Arbeitsbedingungen enthält, wird Cronbachs Alpha je Arbeitsbedingung einzeln berechnet Tavakol und Dennick, 2011, S. 54. In Schritt 6.3 wird darüber hinaus eine konfirmatorische Faktorenanalyse (KFA) durchgeführt, welche als Spezialfall des Strukturgleichungsmodells (SGM) zur Analyse der Operationalisierung latenter Variablen genutzt wird (Backhaus et al., 2015, S. 13; T. A. Brown und Moore, 2012, S. 361). Diese gibt Aufschluss über die Konstruktvalidität, die Fähigkeit der Items das Konstrukt zu messen (El-Den et al., 2020, S. 330). Um zu bestimmen, welche Schätzmethode dabei geeignet ist, wird die Verteilung der Daten geprüft. Der standardmäßig verwendeten Maximum-Likelihood-Methode (ML-Methode) liegen die Annahmen zugrunde, dass die Daten normalverteilt sind und die Items intervallskaliert vorliegen (T. A. Brown & Moore, 2012, S. 368). Wenn die Daten nicht normalverteilt sind und die beobachteten Variablen (annähernd) intervallskaliert vorliegen, sollte die robuste ML verwendet werden (Morata-Ramírez & Holgado-Tello, 2013, S. 55). Zur Prüfung auf Normalverteilung werden nach Radermacher (2019, S. 119) verschiedene Verfahren kombiniert, zunächst werden Histogramme der Datenverteilungen erstellt. Außerdem wird der Shapiro-Wilk-Test durchgeführt und der Doornik-Hansen Test zur Prüfung multivariater Normalverteilung, sowie ein Test auf Schiefe und Wölbung der Daten. Nachdem ein geeignetes Schätzverfahren identifiziert wurde, wird die KFA zunächst für jedes Konstrukt einzeln durchgeführt (Hildebrandt & Temme, 2006, S. 19). Dadurch können zunächst die einzelnen Skalen geprüft werden, wobei die Items signifikante Faktorladungen größer als 0,7 aufweisen sollten und eine Varianz der latenten Variablen von mindestens 0,5 erfassen sollen (Hildebrandt & Temme, 2006, S. 16). Anhand dessen wird geprüft, ob Modifizierungen notwendig sind. Anschließend wird eine KFA aller Konstrukte gemeinsam durchgeführt, bei der Modifizierungen vorgenommen werden. Außerdem wird die konvergente und diskriminierende Validität überprüft (Fornell & Larcker, 1981, S. 12). Die konvergente Validität gibt an, wie stark die Items desselben Konstruktes korrelieren (Mehmetoglu, 2015). Diese wird als durchschnittlich erfasste Varianz (engl.: average variance extracted, AVE) bezeichnet und sollte über 0,5 sein (Hildebrandt & Temme, 2006, S. 14). Die diskriminierende Validität gibt an, inwiefern sich ein Konstrukt von anderen unterscheidet, dazu werden die durchschnittlich erfassten Varianzen verschiedener Konstrukte verglichen (Morrow et al., 2012, S. 106).

Im nächsten Schritt werden erneut Mittelwerte der Items je Arbeitsbedingung gebildet, wobei aufgrund der Ergebnisse der KFA das Item A3 weggelassen wird. Da Likert-Daten häufig die Annahmen der parametrischen Analysen verletzen und die Prüfung der Daten auf Normalverteilung in Schritt 6.3 dies unterstützt, wird in **Schritt 8** ein nichtparametrischer Kruskal-Wallis-Test nach Kruskal und Wallis (1952) durchgeführt. Dieser ist für einen Vergleich der Mittelwerte zwischen mehr als zwei Gruppen geeignet, indem getestet wird ob dieselbe Verteilung den Daten zugrunde liegen kann, oder es Unterschiede der zentralen Tendenzen gibt (Ostertagová et al., 2014, S. 115; Vargha und Delaney, 1998, S. 174). Der Kruskal-Wallis Test weist bei nichtsymmetrisch verteilten Daten eine bessere Aussagekraft auf als die ANOVA-Analyse als parametrisches Pendant (van Hecke, 2012, S. 247). Der Test wird für die vier Arbeitsbedingungen separat durchgeführt, um zu testen, ob mindestens eine Arbeitsplatzform mit einer signifikant geringeren oder größeren Ausprägung der Arbeitsbedingung einhergeht (Vargha & Delaney, 1998, S. 187) und damit die Annahmen A3-A6 belegt werden können. Da ein signifikantes Ergebnis des Kruskal-Wallis-Tests lediglich aussagt, dass mindestens eine Gruppe von einer anderen abweicht und nicht angibt, welche Gruppen voneinander abweichen, wird ein Post-Hoc-Test durchgeführt, der paarweise Vergleiche anstellt (Ostertagová et al., 2014, S. 117), um die Hypothesen 3-6 zu testen. Dazu wird der Dunn's Test (Dunn, 1964) in Stata mithilfe des *dunntest* Befehls nach Dinno (2015) durchgeführt. Dabei wird die Bonferroni-Methode nach Dunn (1961) zur Anpassung der Signifikanzlevel bei multiplen Vergleichen verwendet, bei der die p-Werte der einzelnen Vergleiche mit der Anzahl der paarweisen Tests multipliziert werden, um die Fehleranfälligkeit zu verringern (Dinno, 2015, S. 295).

Die Ergebnisse der zuvor beschriebenen Datenanalyse werden anschließend dargelegt und genutzt, um die anhand der Theorie aufgestellten Hypothesen zu testen und damit die zwei Forschungsfragen zu beantworten.

4. Ergebnisse

Zunächst werden die in **Schritt 1** ermittelten Merkmale der finalen Stichprobe vorgestellt, welche den anschließend dargelegten Ergebnissen der Datenanalyse zugrunde liegt.

Von 410 erhobenen Antworten wurden 164 unvollständige gelöscht, zwar wurde in 52 der unvollständigen Antworten das Vignetten-Experiment vollständig durchgeführt, aber es wurden keine demografischen Daten erhoben. Da diese als Kontrollvariablen fungieren, wurden als *Ausschlusskriterium* alle unvollständigen Antworten von der Analyse ausgeschlossen. Somit wurden für die nachfolgende Analyse 246 vollständige Antworten ohne fehlende Werte hinzugezogen. Die Befragten haben durchschnittlich 9:33 Minuten zur Beantwortung der Fragen gebraucht. Von den Befragten haben 203 angegeben, bereits teilweise im Homeoffice gearbeitet zu haben, 43 haben noch nicht hybrid gearbeitet. 108 Befragte haben bereits in einem non-territorialen Büro ohne feste Sitzplätze gearbeitet und 138 Befragte noch nicht. Von den Befragten waren 69 männlich, 174 weiblich und 3 divers, sodass die Geschlechterverteilung nicht ausgeglichen ist. Das Alter reicht von 18 bis 61 Jahren und wurde entsprechend der Generationen Z (18-27 Jahre), Y (28-43 Jahre) und X/-Babyboomer (44-61 Jahre) (Mangelsdorf, 2015, S. 13) in drei Kategorien aufgeteilt, wobei die Generation Z mit 58% am stärksten vertreten ist. Der Altersdurchschnitt liegt bei 29,16 Jahren. Als höchsten Bildungsabschluss hat der Großteil der Befragten mit einer Anzahl von 154 Personen einen Hochschulabschluss angegeben. Als Beschäftigungsstatus geben 139 Befragte an, mehr als 20 Stunden pro Woche zu arbeiten und 100 Befragte sind Studenten. Weitere Details und prozentuale Verteilungen der Stichprobe können aus der Tabelle 7 im Anhang entnommen werden.

Die Stichprobe wird zur Beantwortung der Forschungsfragen hinzugezogen, indem separat zunächst die Ergebnisse der Datenanalyse in Bezug auf die erste Forschungsfrage vorgestellt werden, und anschließend auf die Ergebnisse eingegangen wird, welche zur Beantwortung der zweiten Forschungsfrage beitragen.

4.1. Präferenz der Arbeitsplatzformen und MWTP

Für die Beantwortung der ersten Forschungsfrage, welche Arbeitsplatzform Arbeitnehmende präferieren und ob sie bereit sind, dafür auf Gehalt zu verzichten, werden die Daten aus dem Vignetten-Experiment und die demografischen Daten hinzugezogen. Da die Befragten jeweils allen acht verwendeten Vignetten Ränge entsprechend der Präferenz zugeordnet haben, werden aus den 246 vollständigen Antworten $8 \times 246 = 1968$ erhobene Messungen, wobei der Rang die abhängige Variable darstellt und die unabhängigen Variablen durch die in den Vignetten enthaltenen Informationen abgebildet werden.

Wie anhand der in **Schritt 2** erstellten deskriptiven Statistik in Abbildung 2 zu erkennen ist, weichen die Rangmittelwerte der Vignetten stark voneinander ab. Weiterhin ist zu erkennen, dass der Rangmittelwert von Vignetten mit gleicher Arbeitsplatzform und geringerem Gehalt höher ist, Befragte also im Durchschnitt bei gleicher Arbeitsplatzform rational das Jobangebot mit höherem Gehalt präferieren. Es wird jedoch auch deutlich, dass Befragte im Durchschnitt bereit sind aufgrund einer präferierten Arbeitsplatzform auf ein höheres Gehalt zu verzichten. Jobangebote mit geringerem Gehalt weisen teilweise geringere Rangmittelwerte auf und werden gegenüber Angeboten mit höherem Gehalt und anderer Arbeitsplatzform bevorzugt.

Um die Annahme 1 zu testen, ob die vier Arbeitsplatzformen die Präferenz der Arbeitnehmenden beeinflussen, wird in **Schritt 3** der nicht-parametrische Chi-Quadrat Test durchgeführt. Die absoluten Häufigkeiten der Ränge je Vignette können aus Tabelle 3 entnommen werden. Die Berechnung in SPSS belegt, dass keine Zelle eine erwartete Häufigkeit von fünf oder kleiner hat, sodass der Test geeignet ist und nicht auf den rechenintensiveren Fisher's-Exact Test zurückgegriffen werden muss (McHugh, 2013, S. 144). Der Chi-Quadrat Test kann die Nullhypothese, dass kein Zusammenhang zwischen den Vignetten und dem Rang besteht, widerlegen. Es gibt einen statistisch signifikanten Zusammenhang, mit $\chi^2(49) = 2.800$ und $p<0,001$.

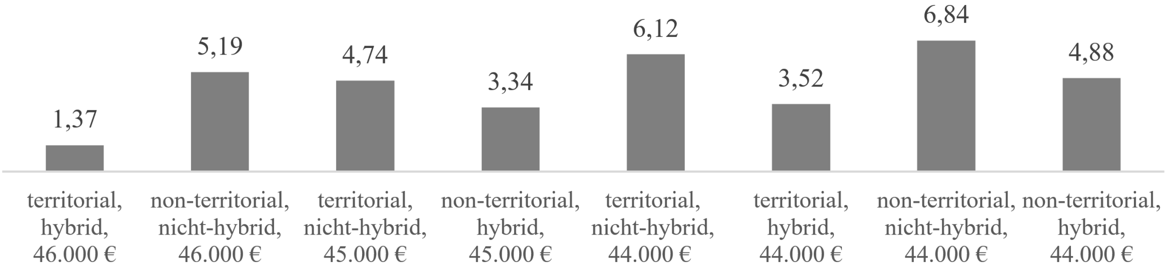
Da die Vignetten allerdings nicht nur die Arbeitsplatzform, sondern auch die verschiedenen Einstiegsgehälter beinhalten und durch das *fractional factorial* Design nur 8 der 12 möglichen Kombinationen beinhalten, wird der Test außerdem separat für die Arbeitsplatzform und den Rang und das Gehalt und den Rang durchgeführt. Dabei wird auch ein statistisch signifikanter Zusammenhang belegt, wenn nur die Arbeitsplatzform und der Rang ($\chi^2(21)$= 1.100, p<0,001), oder nur das Gehalt und der Rang ($\chi^2(14)$= 847, p<0,001), betrachtet werden. Der Chi-Quadrat-Test zeigt somit einen statistisch signifikanten Zusammenhang zwischen den gezeigten Vignetten (sowohl für Arbeitsplatzform, Gehalt und beide Variablen gemeinsam) und dem Rang. Da es sich bei dem Chi-Quadrat Test um einen Signifikanztest handelt, sollte dieser mit einem Test der Stärke ergänzt werden (McHugh, 2013, S. 143). Dazu kann nach J. Cohen (1988) Cramérs V berechnet werden. Dieses belegt mit den hier berechneten Werten von knapp über V = 0,4 einen mittelstarken bis starken Effekt.

Um mehr über den mittels des nicht-parametrischen Tests belegten Zusammenhang zwischen der Arbeitsplatzform und dem Rang zu erfahren, wird in **Schritt 4** zusätzlich ein parametrischer Test durchgeführt. Da die Ergebnisse der unterschiedlichen Schätzmodelle wenig voneinander abweichen und das *Rank Ordered Logit*-Modell aufgrund der Datenstruktur am geeignetsten erscheint, werden lediglich die Regressionsergebnisse dieses Modelles in Tabelle 4 berichtet, wobei Koeffizienten folgender Nutzenfunktion geschätzt werden:

$$Rang_{ij} = \beta_1 \cdot Arbeitsplatzform + \beta_2 \cdot Einstiegsgehalt$$

Die Regressionsergebnisse der anderen Schätzmodelle sind in Tabelle 8 im Anhang zu finden.

Die in **Schritt 4.1** durchgeführte Schätzung per *Rank Ordered Logit*-Modell zeigt, dass sowohl der Koeffizient des Einstiegsgehaltes als auch die Koeffizienten der unterschiedlichen Kombinationen der kategorialen Variable Arbeitsplatzform statistisch signifikant sind mit p<0,01. Das negative Vorzeichen des Koeffizienten des Einstiegsgehaltes weist dar-



Hinweis: Von links nach rechts nach absteigendem Gehalt sortiert. Je kleiner der Rangmittelwert, desto besser wurde die Vignette bewertet, da das am meisten präferierte Jobangebot den Rang 1 bekommt.

Abbildung 2: Rangmittelwert je Vignette (Quelle: eigene Darstellung)

Tabelle 3: Häufigkeit des gewählten Ranges je Vignette & Chi-Quadrat-Test

	Rang								
Vignette	1	2	3	4	5	6	7	8	Gesamt
territorial, nicht-hybrid, 45.000 €	13	18	36	30	45	72	25	7	246
territorial, nicht-hybrid, 44.000 €	2	8	13	32	17	29	98	47	246
territorial, hybrid, 46.000 €	212	11	9	4	4	3	1	2	246
territorial, hybrid, 44.000 €	3	74	71	32	35	19	11	1	246
non-territorial, nicht-hybrid, 46.000 €	3	21	28	14	70	43	53	14	246
non-territorial, nicht-hybrid, 44.000 €	3	4	15	19	8	20	25	152	246
non-territorial, hybrid, 45.000 €	9	97	52	25	37	10	10	6	246
non-territorial, hybrid, 44.000 €	1	13	22	90	30	50	23	17	246
Gesamt	246	246	246	246	246	246	246	246	1968
Arbeitsplatzform & Gehalt × Rang:	Chi-Quadrat (49 Freiheitsgraden) = 2.800, Pr < 0,001, V = 0,451								
Arbeitsplatzform × Rang:	Chi-Quadrat (21 Freiheitsgraden) = 1.100, Pr < 0,001, V = 0,437								
Gehalt × Rang:	Chi-Quadrat (14 Freiheitsgraden) = 847, Pr < 0,001, V = 0,464								

Quelle: Eigene Darstellung.

auf hin, dass ein höheres Einstiegsgehalt mit einer geringeren Wahrscheinlichkeit einhergeht, dass ein Jobangebot mit einem höheren Rang bewertet wird. Das bedeutet, dass ein Jobangebot mit höherem Einstiegsgehalt eher präferiert wird und es darum einen geringeren Rang aufweist. Der Effekt der unterschiedlichen Arbeitsplatzformen kann im Vergleich zu der Basiskategorie interpretiert werden: Arbeitsplatzform 2 (territorial & hybrid) und Arbeitsplatzform 4 (non-territorial & hybrid) weisen im Vergleich zu Arbeitsplatzform 1 (territorial & nicht-hybrid) eine geringere Wahrscheinlichkeit auf, mit einem höheren Rang einherzugehen, werden also besser bewertet. Arbeitsplatzform 3 (non-territorial & nicht-hybrid) geht hingegen mit einer höheren Wahrscheinlichkeit einher, schlechter bewertet zu werden als Arbeitsplatzform 1. Damit kann in Kombination mit den Ergebnissen des nicht-parametrischen Tests Annahme 1 belegt werden, welche besagt, dass die Arbeitsplatzform die Präferenz von Arbeitnehmenden beeinflusst. Hypothese 1 wird ebenfalls unterstützt, da ein territorialer & hybrider Arbeitsplatz im Vergleich zu den anderen Arbeitsplatzformen mit dem geringsten Rang einhergeht und ein non-territorialer & nicht-hybrider Arbeitsplatz mit dem höchsten Rang. Dies kann den Ergebnissen der zur Robustheitsprüfung durchgeführten *Pooled OLS*-Regression in Tabelle 8 im Anhang entnommen werden, bei der die ordinale abhängige Variable als intervallskaliert angenommen wird und damit eine quantitative Interpretation möglich ist.

Um Annahme 2 und Hypothese 2 zu untersuchen, sind neben den Regressionsergebnissen auch die in **Schritt 4.2** berechnete MWTP und MWTP der unteren und oberen Grenze des 95 % Konfidenzintervalls in Tabelle 4 enthalten. Die Befragten sind mit durchschnittlich 2.175 € bereit auf am meisten Einstiegsgehalt zu verzichten, um ein Jobangebot mit territorialer & hybrider Arbeitsplatzform anzunehmen, im Vergleich zu einer territorialen & nicht-hybriden Arbeitsplatzform. Für einen non-territorialen & nicht-hybriden Arbeitsplatz sind sie hingegen nicht bereit auf Gehalt zu verzichten, da dieser den Befragten durchschnittlich 937 € weniger Wert ist als ein territorialer & nicht-hybrider Arbeitsplatz, wobei die Spanne der MWTP nach dem 95 % Konfidenzintervall im Vergleich zu den anderen deutlich größer ist. Die Bereitschaft für einen non-territorialen & hybriden Arbeitsplatz auf

Tabelle 4: Regressionsergebnisse der *Rank Ordered Logit*-Regression

Abhängige Variable: Rang						
Unabhängige Variablen	Koeffizient	95% Konf. Intervall		MWTP	MWTP (95% Konf. Intervall)	
Arbeitsplatzform						
Basiskategorie: 1 (territorial & nicht-hybrid)	.	.	.	.	.	.
2 (territorial & hybrid)	-1,7552* (0,0959)	-1,9431	-1,5674	2,1755	2,1914	2,1563
3 (non-territorial & nicht-hybrid)	0,7562* (0,0812)	0,5971	0,9152	-0,9373	-0,6734	-1,2590
4 (non-territorial & hybrid)	-0,8633* (0,0822)	-1,0244	-0,7023	1,0700	1,1553	0,9662
Einstiegsgehalt	-0,8068* (0,0408)	-0,8867	-0,7269	.	.	.
LR chi2(4) = 1177,04						
Prob > chi2 = 0,0000						

Hinweis: Werte gerundet auf 4 Nachkommastellen. Einstiegsgehalt in 1.000 €. Alle Koeffizienten sind statistisch signifikant (* p<0,01).
Quelle: Eigene Darstellung.

Einstiegsgehalt im Vergleich zu einem territorialen & nicht-hybriden Arbeitsplatz zu verzichten, lässt sich mit dem impliziten monetären Wert von 1.070 € beziffern. Damit kann sowohl Annahme 2 als auch Hypothese 2 unterstützt werden, welche besagen, dass die Arbeitsplatzform die Bereitschaft auf Gehalt zu verzichten beeinflusst und die Bereitschaft für die territoriale & hybride Arbeitsplatzform auf Gehalt zu verzichten am größten ist und für die non-territoriale & nicht-hybride Arbeitsplatzform am geringsten.

Da die Ergebnisse der Regression Auskunft über die durchschnittliche MWTP liefern sollen, werden neben der Regression mit allen Befragten in **Schritt 5** auch Regressionen für Subgruppen als Robustheitsprüfung durchgeführt (Radermacher et al., 2017, S. 78). Dieses Vorgehen empfiehlt sich, um zu verstehen, inwieweit sich die durchschnittlichen Präferenzen der Subgruppen voneinander unterscheiden (Leeper et al., 2019, S. 219–220). Dadurch kann für den Effekt von möglichen Moderatoren auf den Zusammenhang zwischen der Arbeitsplatzform und der Präferenz, beziehungsweise der Bereitschaft auf Einstiegsgehalt zu verzichten, kontrolliert werden. Dazu werden 16 weitere Regressionen für Subgruppen durchgeführt, in denen die vorherigen Erfahrungen mit der Arbeitsplatzform (Regression 2-5), das Geschlecht der Befragten (Regression 6-7), das Alter entsprechend den Generationen (Regression 8-10), der höchste erreichte Bildungsabschluss (Regression 11-14) und der Beschäftigungsstatus (Regression 15-17) berücksichtigt werden. Die Regressionsergebnisse und die berechneten MWTP sind in Tabelle 9 im Anhang zu finden. Lediglich zwei Koeffizienten der Arbeitsplatzform 4 (non-territorial & hybrid) in Regression 3 und 14 weisen ein von den vorherigen Ergebnissen abweichendes Vorzeichen auf, diese sind allerdings nicht statistisch signifikant und die Regressionen beinhalten nur geringe Stichprobengrößen. Die Richtung der Wirkung der Arbeitsplatzformen und des Einstiegsgehaltes auf den Rang des Jobangebotes unterscheidet sich ansonsten für keine Subgruppe. Alle Regressionen zeigen, dass territoriales & hybrides und non-territoriales & hybrides Arbeiten im Vergleich zu dem territorialen & nicht-hybriden Arbeiten bevorzugt werden und non-territoriales & nicht-hybrides Arbeiten schlechter bewertet wird. Befragte, welche vorher nicht non-territorial gearbeitet haben, verlangen für non-territoriales & nicht-hybrides Arbeiten mehr Gehalt (1.298 € vs. 587 €) und sind für einen non-territorialen & hybriden Job nur bereit weniger zu zahlen als Befragte, die es vorher getestet haben (759 € vs. 1.368 €). Bei Betrachtung der Geschlechter wird deutlich, das territoriales & hybrides Arbeiten Frauen mehr wert ist als Männern (2.343 € vs. 1.836 €), bei den anderen Arbeitsformen sind die Unterschiede allerdings geringer. Befragte der ältesten Generation verlangen für non-territoriales & hybrides Arbeiten mit 2.154 € deutlich mehr als jüngere Generationen mit 778 € und 979 €. Befragte, welche als höchsten Bildungsabschluss eine Ausbildung aufweisen, bewerten den territorialen & nicht-hybriden Arbeitsplatz besser, sie sind nur bereit weniger für hybride Arbeitsplätze zu zahlen (1.559 € vs. 2.393 € und 608 € vs. 1.340 €) und verlangen mehr für non-territoriale & nicht-hybride Arbeitsplätze (1.596 € vs. 715 €) als Hochschulabsolvent*innen. Außerdem verlangen Beschäftigte, die mehr als 20 Stunden pro Woche arbeiten, mehr Gehalt für non-territoriales & nicht-hybrides Arbeiten als Beschäftigte, die bis zu 20 Stunden arbeiten (1.150 € vs. 414 €).

In Bezug auf die erste Forschungsfrage lässt sich zusammenfassen, dass die vier betrachteten Arbeitsplatzformen die Präferenz der Befragten signifikant beeinflussen und auch die MWTP deutlich davon abhängt. Die territoriale & hybride Arbeitsplatzform wird durchschnittlich am meisten präferiert, die non-territoriale & nicht-hybride Arbeitsplatzform ist hingegen am unbeliebtesten. Um mehr darüber zu erfahren, warum die verschiedenen Arbeitsplatzformen sich hinsichtlich der Präferenz unterscheiden, werden nachfolgend die Ergebnisse in Bezug auf die zweite Forschungsfrage vorgestellt.

4.2. Zugrundeliegende Mechanismen

Forschungsfrage zwei beschäftigt sich mit den Mechanismen, welche den Auswirkungen der Arbeitsplatzform auf die Präferenz der Arbeitnehmenden zugrunde liegen. Dazu werden die Daten der standardisierten Umfrage hinzugezogen, welche Informationen über durch Arbeitsplatzformen veränderte Arbeitsbedingungen beinhalten. Die Daten werden ebenfalls aufbereitet, indem pro Person je Arbeitsplatzform eine Zeile generiert wird, um die Bewertung der Items je Arbeitsplatzform auswerten zu können, es ergeben sich somit 246 × 4 = 984 Messungen.

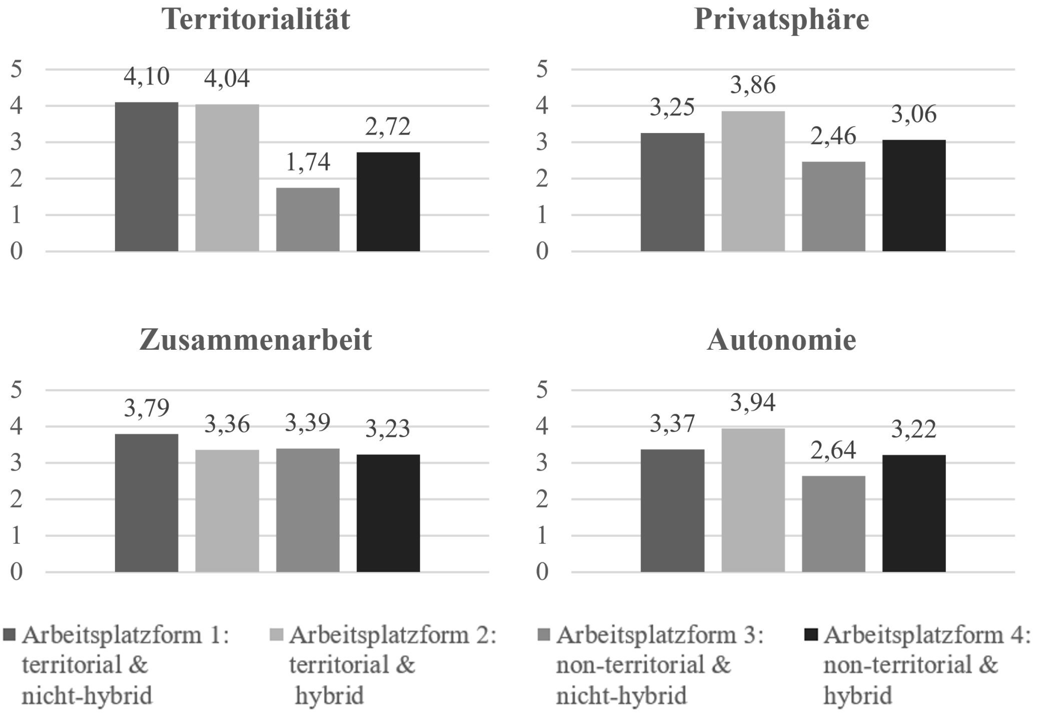
Zunächst wird die in **Schritt 6** erstellte deskriptive Statistik der Daten ausgewertet. Wie anhand der Abbildung 3 zu erkennen ist, unterscheiden sich die durchschnittlichen Bewertungen der Arbeitsbedingungen je Arbeitsplatzform.

Besonders deutliche Unterschiede sind bei der Arbeitsbedingung Territorialität zu erkennen, während ein non-territorialer & nicht-hybrider Arbeitsplatz durchschnittlich nur mit 1,74 auf der 5-stufigen Likert Skala bewertet wird, werden Arbeitsplatzform 1 und 2 mehr als doppelt so gut bewertet. Hinsichtlich der Arbeitsbedingungen Privatsphäre und Autonomie sind ebenfalls deutliche Unterschiede je Arbeitsplatzform zu sehen, die Bewertungen der Arbeitsbedingung Zusammenarbeit liegen näher beieinander. Außerdem ist ersichtlich, dass die Arbeitsplatzform 2, territoriales & hybrides Arbeiten, insgesamt mit der besten Bewertung der Arbeitsbedingungen einhergeht, non-territoriales & nicht-hybrides Arbeiten wird hinsichtlich der Arbeitsbedingungen insgesamt am schlechtesten bewertet. Da dieses Ergebnis mit den Ergebnissen der im Rahmen der Forschungsfrage 1 durchgeführten Regressionen einhergeht, kann es als erstes Indiz angesehen werden, dass die vier hier betrachteten Arbeitsbedingungen tatsächlich Auskunft über die zugrundeliegenden Mechanismen geben. Arbeitsplatzform 1 weist allerdings die zweitbeste Bewertung der Arbeitsbedingungen auf, wird laut Regression aber nur auf dritter Position präferiert und Arbeitsplatzform 4 die drittbeste Bewertung, ist laut Regression allerdings auf zweiter Position, sodass die hypothetisierten Mechanismen durch statistische Tests genauer untersucht werden müssen.

Zuvor werden allerdings in **Schritt 7** die Items der erstellten Skalen geprüft. Die dazu in **Schritt 7.1** berechneten Korrelationskoeffizienten der Items nach Pearson und Spearman sind der Tabelle 10 und Tabelle 11 im Anhang zu entnehmen, wobei jeweils die Koeffizienten der drei Items markiert sind, welche laut Theorie dieselbe Arbeitsbedingung abfragen. Die Ergebnisse der beiden Berechnungen weisen keine relevanten Unterschiede auf, sodass im Folgenden die Ergebnisse der Berechnung nach Pearson betrachtet werden. Die internen Korrelationen der Items je Konstrukt sind statistisch signifikant auf dem 5 % Level und mittel bis groß, da sie von 0,56 bis 0,86 reichen. Die Korrelationskoeffizienten zeigen außerdem, dass die Korrelationen zwischen Items desselben Konstruktes größer sind als die Korrelationen zwischen dem jeweiligen Item und den Items anderer Konstrukte. Die Ergebnisse zeigen allerdings auch, dass es mittlere Korrelationen zwischen einzelnen Items unterschiedlicher Konstrukte gibt, sodass der Grad der Differenzierung eher gering zu sein scheint. Die Ergebnisse der Berechnung von Cronbachs Alpha für die vier unterschiedlichen Arbeitsbedingungen in **Schritt 7.2** sind in Tabelle 12 im Anhang zu finden. Da Cronbachs Alpha Werte zwischen 0,833 und 0,929 annimmt, deuten diese auf eine sehr gute interne Konsistenz der Items je Arbeitsbedingung hin. Der Wert würde bei der Arbeitsbedingung Autonomie durch Weglassen des Items A3 leicht von 0,833 auf 0,846 steigen, dies wird zusätzlich durch die KFA überprüft.

Bevor in **Schritt 7.3** die KFA zur weiteren Testung der Items und Skalen durchgeführt werden kann, wird jedoch die Verteilung der Daten betrachtet, um die Eignung der Schätzmethoden zu bestimmen. Zur Prüfung der Daten auf Normalverteilung werden zunächst Histogramme der Verteilung der Daten der Items und der Mittelwerte der jeweiligen Items erstellt. Diese können der Abbildung 8 im Anhang entnommen werden und weichen allesamt von den entsprechenden Normalverteilungskurven ab. Die Ergebnisse des Shapiro-Wilk Tests auf dem 5 % Signifikanzniveau sind der Tabelle 13 im Anhang zu entnehmen. Lediglich das Item A3 weist ein p>0,05 auf, was darauf hinweist, dass alle anderen Variablen nicht normalverteilt sind. Auch der Doornik-Hansen Test belegt mit einem $\chi^2(24)= 212{,}984$, p<0,001, dass die Items, welche Arbeitsbedingungen einfangen, zusammen nicht normalverteilt sind. Dies belegt auch der Test auf Schiefe und Wölbung der Daten, der zeigt, dass Schiefe und Wölbung für keine Variable einer Normalverteilung entsprechen, die Ergebnisse können der Tabelle 14 im Anhang entnommen werden. Somit lässt sich zusammenfassen, dass die Daten nicht normalverteilt sind. Entsprechend muss die ML-Methode als Schätzverfahren kritisch betrachtet werden. Da die Likert-Skala als intervallskaliert angenommen wird, wird hauptsächlich eine ML mit robusten Standardfehlern durchgeführt.

Zunächst wird die KFA per robuster ML Schätzung für jedes Konstrukt einzeln durchgeführt, die Ergebnisse werden in Tabelle 15 im Anhang zusammengefasst. Bei keinem Konstrukt werden von Stata Modifizierungen vorgeschlagen. Die Ergebnisse der KFA des Konstruktes *Territorialität* zeigen, dass alle drei Items eine signifikante Faktorladung von >0,7 aufweisen und jeweils mehr als 50 % der Varianz erklären können. Die durchschnittlich erfasste Varianz beträgt 0,817 und liegt entsprechend ebenso über dem geforderten Wert von 0,5, um konvergente Validität zu zeigen. Für die Konstrukte *Privatsphäre* und *Zusammenarbeit* sind ebenfalls alle Items signifikant mit Faktorladungen von >0,7. Die Items erklären außerdem alle jeweils mindestens 50 % der Varianz. Die durchschnittlich erfassten Varianzen zeigen mit Werten von 0,731 und 0,724 ebenso konvergente Validität der Skalen. Entsprechend werden keine Items der drei Konstrukte



Hinweis: Es ist der Mittelwert der jeweiligen 3 Items je Arbeitsbedingung dargestellt, welche auf einer 5-stufigen Likert-Skala bewertet wurden, getrennt nach den vier verschiedenen Arbeitsplatzformen.

Abbildung 3: Durchschnittliche Bewertung der Arbeitsbedingungen je Arbeitsplatzform (Quelle: eigene Darstellung)

entfernt. Die Faktorladungen der Items des Konstruktes *Autonomie* sind signifikant und bis auf das Item A3 ebenfalls >0,7 und erklären jeweils mehr als 50 % der Varianz. Item A3 weist hingegen nur eine Faktorladung von 0,66 auf, außerdem erklärt das Item nur 44 % der Varianz. Da bei der Berechnung von Cronbachs Alpha ebenfalls deutlich wurde, dass das Weglassen des Items vorteilhaft ist, wird es inhaltlich geprüft. Da das Item A3 „Ich habe einen großen Einfluss auf Entscheidungen, die meine Arbeit betreffen.“ die Autonomie in Bezug auf Entscheidungen abfragt, welche möglicherweise durch die Arbeitsplatzform weniger beeinflusst wird als der Ort und der Grad der Selbstbestimmung, welche in A1 und A2 abgefragt werden, wird dieses weggelassen. Da die Skala entsprechend aus nur zwei Items besteht, wird die Faktorladung des ersten Items bei Betrachtung des einzelnen Konstruktes auf 1 festgelegt, die Faktorladungen und erklärten Varianzen der zwei verbleibenden Items sind höher und die durchschnittlich erfasste Varianz steigt auf 0,771. Da eine Schätzung per robuster ML durchgeführt wurde, können nur Statistiken bezüglich der Residuen angegeben werden, diese weisen allerdings auf gute Passung hin.

Anschließend wird die KFA aller Konstrukte durchgeführt, die Ergebnisse können der Tabelle 16 im Anhang entnommen werden. Die Daten lassen auf eine starke Kovarianz zwischen den Konstrukten Territorialität und Privatsphäre, Territorialität und Autonomie, sowie Privatsphäre und Autonomie schließen. Da die Konstrukte auch inhaltlich nah beieinander liegen und in der Literatur Abhängigkeiten aufgezeigt werden (Brunia und Hartjes-Gosselink, 2009, S. 176; Frankó et al., 2022, S. 241; Gove, 1978, S. 638), werden Kovarianzen zwischen diesen hinzugefügt. Dadurch wird die Passung des Modells weiter verbessert. In dem finalen, per robuster ML-Methode geschätzten Modell sind ebenfalls alle Faktorladungen signifikant >0,7, auch die Kovarianzen zwischen den Konstrukten sind signifikant. Die erklärten Varianzen aller Items betragen mehr als 62 %, sodass R^2 insgesamt 0,9997 beträgt. Weitere kleinere von Stata vorgeschlagene Modifizierungen werden nicht durchgeführt. Durchschnittlich erfasste Varianzen von >0,5 deuten darauf hin, dass diskriminierende und konvergierende Validität vorliegen. Eine ergänzende Schätzung per ML-Methode, wodurch weitere Statistiken über die Passung des Modells erhalten werden können, weist folgende Indizes auf: $\chi^2(38) = 178{,}218$, p<0,001, einen *Root mean square error of approximation* (RMSEA) von 0,061, einen *Comparative fit index* (CFI) von 0,983, einen *Tucker–Lewis index* (TLI) von 0,975, ein *Standardized root mean squared residual* (SRMR) von 0,031 und einen *Coefficient of determination* (CD) von 1,00. Diese weisen insgesamt auf eine gute Passung des globalen Modells hin, da die Cutoffs nach Hu und Bentler (1999, S. 27) für RMSEA bei etwa 0,06, für CFI und TLI jeweils bei $\geq$ 0,95 und für SRMR bei $\leq$ 0,08 liegen. Die Schätzung kann aufgrund der weniger geeigneten Methode aber abweichende Ergebnisse liefern.

In **Schritt 8** wird je Arbeitsbedingung ein nicht-parametrischer Kruskal-Wallis-Test durchgeführt, um die Annahmen

A3-A6 zu testen, welche besagen, dass die vier unterschiedlichen Arbeitsplatzformen einen Einfluss auf die jeweilige Arbeitsbedingung haben. Alle vier Kruskal-Wallis-Tests weisen ein p<0,0001 auf (mit $\chi^2(3)= 529{,}302$ für Territorialität, $\chi^2(3)= 230{,}437$ für Privatsphäre, $\chi^2(3)= 63{,}015$ für Zusammenarbeit und $\chi^2(3)= 239{,}485$ für Autonomie), sodass die Nullhypothese, dass es keinen Unterschied der Arbeitsbedingungen je Arbeitsplatzform gibt, abgelehnt werden kann. Damit kann belegt werden, dass es einen signifikanten Unterschied zwischen mindestens zwei der Arbeitsplatzformen hinsichtlich des Effektes auf die Arbeitsbedingungen gibt. Um mehr darüber zu erfahren, zwischen welchen Arbeitsplatzformen der Unterschied signifikant ist und unter welcher Arbeitsplatzform die jeweilige Arbeitsbedingung am stärksten ausgeprägt ist, wird der Dunn's Post-Hoc-Test durchgeführt, die Ergebnisse können der Tabelle 5 entnommen werden.

In Bezug auf die Arbeitsbedingung Territorialität zeigt der Test signifikante Unterschiede zwischen allen vier Arbeitsplatzformen mit p<0,001, außer zwischen Arbeitsplatzform 1 und 2, sodass Annahme 3 belegt werden kann, da die Arbeitsplatzform die Territorialität beeinflusst. Zwar ist die Rangsumme und der Mittelwert bei der ersten Arbeitsplatzform am höchsten, da es aber keinen signifikanten Unterschied zu Arbeitsplatzform 2 gibt, kann belegt werden, dass sowohl territoriales & nicht-hybrides als auch territoriales & hybrides Arbeiten mit der größten Territorialität einhergehen. Damit kann Hypothese H3a teilweise unterstützt werden, welche vorausgesagt hatte, dass eine territoriale & hybride Arbeitsplatzform die Territorialität am positivsten beeinflusst. Hypothese H3b, welche besagt, dass ein non-territorialer & nicht-hybrider Arbeitsplatz die Territorialität am negativsten beeinflusst, wird durch die Ergebnisse belegt, da die Arbeitsplatzform die geringste Rangsumme aufweist.

Alle Arbeitsplatzformen außer territorial & nicht-hybrid und non-territorial & nicht-hybrid unterscheiden sich signifikant hinsichtlich der Arbeitsbedingung Privatsphäre mit p<0,05, sodass auch Annahme 4 unterstützt werden kann, die Arbeitsplatzform beeinflusst die Privatsphäre. Ebenso können die Hypothesen H4a und H4b unterstützt werden, da laut den Rangsummen und Mittelwerten ein territorialer & hybrider Arbeitsplatz die Privatsphäre am positivsten und ein non-territorialer & nicht-hybrider Arbeitsplatz die Privatsphäre am negativsten beeinflusst.

In Bezug auf die Arbeitsbedingung Zusammenarbeit gibt es ebenso signifikante Unterschiede zwischen den Arbeitsplatzformen, sodass auch Annahme 5 belegt werden kann, die Arbeitsplatzform beeinflusst die Zusammenarbeit. Der Dunn's Test zeigt allerdings, dass die Unterschiede zwischen territorialen & hybriden Arbeitsformen und non-territorialen & nicht-hybriden Arbeitsformen nicht signifikant sind, ebenso wie zwischen territorialen & hybriden Arbeitsformen und non-territorialen & hybriden Arbeitsformen, da p>0,05. Hypothese H5a kann nicht unterstützt werden, da ein territorialer & nicht-hybrider Arbeitsplatz die Zusammenarbeit am positivsten beeinflusst und nicht der non-territoriale & nicht-hybride Arbeitsplatz. Hypothese H5b kann weder widerlegt noch unterstützt werden, da Rangsumme und Mittelwert eines non-territorialen & hybriden Arbeitsplatzes zwar am geringsten sind, aber kein signifikanter Unterschied zu einem territorialen & hybriden Arbeitsplatz besteht.

Zwischen allen vier Arbeitsplatzformen, außer zwischen territorial & nicht-hybrid und non-territorial & hybrid, gibt es hinsichtlich der Arbeitsbedingung Autonomie signifikante Unterschiede mit p<0,01, sodass Annahme 6 belegt wird, welche besagt, dass die Arbeitsplatzform die Autonomie beeinflusst. Die Hypothese H6a wird allerdings widerlegt, da ein territorialer & hybrider Arbeitsplatz die höchste Rangsumme und den höchsten Mittelwert aufweist und damit die Autonomie am positivsten beeinflusst, anstatt dem non-territorialen & hybriden Arbeitsplatz. Hypothese H6b wird ebenfalls widerlegt, da ein non-territorialer & nicht-hybrider Arbeitsplatz hinsichtlich der Autonomie am geringsten bewertet wird und nicht der territoriale & nicht-hybride Arbeitsplatz, welcher die zweitbeste Bewertung erhielt.

Die Ergebnisse zeigen insgesamt, dass die vier Arbeitsplatzformen die Arbeitsbedingungen Territorialität, Privatsphäre, Zusammenarbeit und Autonomie beeinflussen, die Bewertungen der letzteren beiden allerdings nicht den hypothetisierten Zusammenhängen entsprechen.

Abbildung 9 im Anhang liefert eine Übersicht der Ergebnisse in Bezug auf die Hypothesen, welche im Anschluss interpretiert werden.

5. Interpretation

Die zuvor präsentierten Ergebnisse werden folgend in Rückbezug zur Literatur und der zugrundeliegenden Theorie diskutiert. Außerdem wird deren Relevanz herausgestellt, indem neue Erkenntnisse durch die Arbeit zusammengefasst werden und auf die Verhinderung bisheriger Limitationen der Forschung eingegangen wird. Anschließend werden dennoch bestehende Limitationen aufgezeigt, aus denen Implikationen für die weitere Forschung abgeleitet werden. Letztlich werden Handlungsempfehlungen für die Praxis aufgrund der vorliegenden Ergebnisse gegeben.

5.1. Diskussion

Die vorliegenden Ergebnisse in Bezug auf die Forschungsfrage 1 belegen den durch das *Job-Demands-Resources*-Modell nach Bakker und Demerouti (2007) vorgestellten grundlegenden Zusammenhang, dass die Arbeitsplatzform und deren Merkmale arbeitsbezogene Konsequenzen beeinflussen, da die vier hier betrachteten Arbeitsplatzformen die Präferenz der Arbeitnehmenden für ein Jobangebot signifikant beeinflussen. Das Ergebnis geht mit dem einer Vielzahl an Studien einher, welche für andere Arbeitsplatzformen ebenfalls die grundlegende Annahme des theoretischen Modells, dass die Arbeitsplatzform die arbeitsbezogenen Konsequenzen kausal beeinflusst, bestätigen (Lesener et al., 2019, S. 93). Damit wird erneut deutlich, dass das Modell für eine Vielzahl an Untersuchungskontexten genutzt werden kann. Die MWTP verdeutlicht außerdem, dass die Ergebnisse auch ökonomisch signifikant sind, da die Befragten durchschnittlich

Tabelle 5: Ergebnisse des Kruskal-Wallis und Post-Hoc Dunn's Test

	Kruskal-Wallis Test					**Dunn's Test**		
	Mittelwert	Beobachtungen	Rangsumme	chi2(3)	Prob	Arbeitsplatzform		
						1	2	3
Territorialität								
Arbeitsplatzform 1	4,10	246	171.485,50	529,30	0,0001	.	.	.
Arbeitsplatzform 2	4,04	246	168.129,00			0,53	.	.
Arbeitsplatzform 3	1,74	246	49.665,00			19,42*	18,88*	.
Arbeitsplatzform 4	2,72	246	95.340,50			12,14*	11,60*	-7,28*
Privatsphäre								
Arbeitsplatzform 1	3,25	246	126.269,00	230,44	0,0001	.	.	.
Arbeitsplatzform 2	3,86	246	169.845,50			-6,95*	.	.
Arbeitsplatzform 3	2,46	246	75.092,00			8,16*	15,11*	.
Arbeitsplatzform 4	3,06	246	113.413,50			2,05	9,00*	-6,11*
Zusammenarbeit								
Arbeitsplatzform 1	3,79	246	150.395,00	63,02	0,0001	.	.	.
Arbeitsplatzform 2	3,36	246	112.346,50			6,13*	.	.
Arbeitsplatzform 3	3,39	246	118.381,50			5,15*	-0,97	.
Arbeitsplatzform 4	3,23	246	103.497,00			7,55*	1,42	2,40*
Autonomie								
Arbeitsplatzform 1	3,43	246	127.189,50	239,49	0,0001	.	.	.
Arbeitsplatzform 2	4,10	246	170.399,50			-6,94*	.	.
Arbeitsplatzform 3	2,54	246	73.905,50			8,56*	15,50*	.
Arbeitsplatzform 4	3,24	246	113.125,50			2,26	9,20*	-6,30*

Hinweis: *p<0,05, Werte gerundet auf 2 Nachkommastellen.
Quelle: Eigene Darstellung.

bereit sind auf 2.175 € für ein Jobangebot mit territorialer & hybrider Arbeitsplatzform im Vergleich zu einer territorialen & nicht-hybriden Arbeitsplatzform zu verzichten, was bei einem Einstiegsgehalt von 46.000 € fast 5% ausmacht. Zwar gibt es hinsichtlich verschiedener Subgruppen Unterschiede in Bezug auf die MWTP, sodass die berechneten Beträge lediglich als Annäherungen gesehen werden sollten, die Tendenz der Ergebnisse ist aber über alle Gruppen hinweg konstant.

Weiterhin konnten durch die Betrachtung der vier Arbeitsbedingungen Territorialität, Privatsphäre, Zusammenarbeit und Autonomie als Arbeitsressourcen die Ergebnisse in Bezug auf die erste Forschungsfrage unterstrichen werden. Außerdem konnten Informationen bezüglich der in Forschungsfrage 2 betrachteten Mechanismen gesammelt werden, welche dem Effekt der Arbeitsplatzform auf die Präferenz der Arbeitnehmenden zugrunde liegen. Es wird allerdings lediglich eine begrenzte Anzahl an zugrundeliegenden Mechanismen betrachtet, indem sich auf die vier in der Literatur identifizierten Arbeitsbedingungen festgelegt wird, welche vermutlich im Kontext des non-territorialen und hybriden Arbeitens relevant sind. Die Ergebnisse dürfen entsprechend nicht als vollständige Erklärungen der zugrundeliegenden Mechanismen angenommen werden, sondern zeigen lediglich, dass die vier Arbeitsbedingungen von der untersuchten Arbeitsplatzform beeinflusst werden. Für weitere Arbeitsplatzformen und Merkmale können auch andere Mechanismen entscheidend sein.

Außerdem werden in dieser Studie nicht die im JDR-Modell inkludierten Job-Anforderungen berücksichtigt, da diese bei der Jobauswahl nicht bekannt sind. Es kann aber angenommen werden, dass die Anforderungen mit den Job-Ressourcen interagieren und in einem dualen Prozess Stress und Motivation von Arbeitnehmenden beeinflussen (Bakker und Demerouti, 2007, S. 314; Gatt und Jiang, 2021, S .977), sodass in dieser Studie nur ein Teil der zugrundeliegenden Mechanismen erforscht wird. In dem Kontext der Präferenz bezüglich verschiedener Jobangebote werden lediglich die psychologischen und sozialen Ressourcen betrachtet.

Die in der Literatur häufig untersuchten Arbeitsbedingungen als Ressourcen werden auch von den Kombinationen hybrider oder nicht-hybrider und non-territorialer oder territorialer Arbeitsplatzformen beeinflusst, wie der Kruskal-Wallis und Dunn's Test belegen. Die Ergebnisse belegen allerdings nicht alle anhand der Literatur und einer Vielzahl an Theorien bezüglich der einzelnen Arbeitsbedingungen getroffenen Voraussagen. Es gibt Abweichungen in Bezug darauf, unter welcher der vier Kombinationen die jeweilige Arbeitsbedingung als am stärksten oder geringsten vorhanden bewertet wird. Dies könnte darauf zurückgeführt werden, dass bereits bei der Betrachtung der beiden Merkmale getrennt voneinander teilweise uneinige Ergebnisse in der Literatur vorliegen oder verschiedene Aspekte betrachtet werden. Ein von den Hypothesen in Bezug auf die Zusammenarbeit abweichendes Ergebnis könnte etwa dadurch erklärt werden, dass keine klare Unterscheidung vorgenommen wurde, ob es sich um die Zusammenarbeit in einem Team oder zwischen Arbeitnehmenden unterschiedlicher Teams handelt. Ein non-territorialer Arbeitsplatz scheint die Autonomie entgegen den Voraussagen zu verringern, was darauf zurückzuführen sein könnte, dass non-territoriales Arbeiten in realen Situationen in abgewandelter Form angewendet werden kann. Teammitglieder können sich beispielsweise abstimmen, sich trotz non-territorialer Arbeitsplätze einen gemeinsamen Bereich zu suchen. Es ist möglich, dass solche oder weitere Strategien von Befragten nicht berücksichtigt wurden.

In der vorliegenden Arbeit wird darüber hinaus die Kombination von zwei Merkmalen betrachtet, sodass fraglich ist, ob diese bei der Bewertung durch die Befragten dieselbe Gewichtung erhalten. Dies wird auch durch das Ergebnis unterstrichen, dass territoriale & hybride Arbeitsplätze am stärksten präferiert werden, gefolgt von non-territorialen & hybriden, territorialen & nicht-hybriden und den am wenigsten präferierten non-territorialen & nicht-hybriden Arbeitsplätzen. Hybrides Arbeiten scheint den Befragten wichtiger zu sein als die Unterscheidung zwischen territorialem oder non-territorialem Arbeiten. Die Ergebnisse spiegeln dennoch die Tendenzen der Literatur wider, dass hybrides Arbeiten gegenüber nicht-hybridem Arbeiten bevorzugt wird und territoriales Arbeiten gegenüber dem non-territorialem Arbeiten.

Die mittlere Korrelation zwischen Items verschiedener Arbeitsbedingungen und der Modifizierungsvorschlag der KFA, Kovarianzen zwischen Arbeitsbedingungen aufzunehmen, kann außerdem darauf hinweisen, dass die Arbeitsbedingungen nicht trennscharf abgefragt werden oder nicht trennscharf sind. In der Literatur herrscht ebenfalls Uneinigkeit und Unwissen über deren Zusammenhang. Nach Altman (1975) ist etwa Privatsphäre das wichtigste Konzept, um zu verstehen, welche Auswirkungen ein physisches Umfeld hat und eng mit weiteren Konzepten wie der Territorialität verbunden. Die Beziehungen zwischen diesen Arbeitsbedingungen werden aber auch in der Literatur nicht klar herausgestellt (Gove, 1978, S. 638). Nach Frankó et al. (2022, S. 231) ist hingegen Territorialität besonders wichtig, da durch territoriales Verhalten die Privatsphäre gewahrt wird. Die Ergebnisse dieser Analyse können lediglich einen Hinweis darauf geben, dass zwischen den Konstrukten Territorialität und Privatsphäre, Territorialität und Autonomie, sowie Privatsphäre und Autonomie Zusammenhänge bestehen, es wird aber nicht deutlich, wie diese zusammenhängen oder welcher Mechanismus besonders wichtig ist.

5.2. Relevanz

Diese Arbeit liefert dennoch einen wichtigen Beitrag, indem die Auswirkungen einer neuen Arbeitsplatzform, dem non-territorialen und hybriden Arbeiten, untersucht werden, die durch die Weiterentwicklung von Technologien, der Forderung von Arbeitnehmenden nach flexibleren Arbeitsmodellen und der Absicht der Arbeitgebenden, die Kosten für Büroräume zu reduzieren, immer relevanter wird (Bhave et al., 2020, S. 142). Da non-territoriale und hybride Arbeitsplatzformen in Kombination besonders naheliegend sind, ist es wichtig die gemeinsamen Auswirkungen zu untersuchen. Neben dem Ergebnis, welche Kombination präferiert wird, wird durch die Berechnung der MWTP auch ein impliziter monetärer Wert bestimmt, sodass deutlich wird, in welchem Ausmaß eine Arbeitsplatzform präferiert wird.

Die Betrachtung der Arbeitsbedingungen liefert einen wichtigen Beitrag, da dadurch der Beziehung zwischen der Arbeitsplatzform und den organisationalen Konsequenzen zugrundeliegende Mechanismen untersucht werden, wie von Wohlers und Hertel (2017, S. 480) gefordert und in dieser Studie auf non-territoriale und hybride Arbeitsplätze angewendet wird. Dadurch kann zumindest teilweise aufgedeckt werden, wie die Präferenzen zustande kommen und warum eine Arbeitsplatzform gegenüber einer anderen präferiert wird. Damit liefert die Studie auch einen Beitrag zur Theorie, da das JDR-Modell zwar den grundlegenden Zusammenhang zwischen Arbeitsplatzformen und organisationalen Konsequenzen beschreibt, aber keine spezifischen Theorien angibt, welche den Zusammenhang zwischen bestimmten Ressourcen oder Anforderungen und den Ergebnissen erklären (Schaufeli & Taris, 2014, S. 60). Diese Studie vereint hingegen im JDR-Modell weitere bereits bestehende Theorien, wie etwa die *Self-determination* Theorie nach Deci und Ryan (2000) oder die *Privacy Regulation* Theorie nach Altman (1975), welche vorhersagen, warum die vier Arbeitsbedingungen als Ressourcen den Arbeitnehmenden einen Nutzen bringen. Da belegt wird, dass die betrachteten Arbeitsbedingungen von der Arbeitsplatzform beeinflusst werden, können diese auch in der weiteren Forschung berücksichtigt werden, um Mechanismen hinter dem übergreifenden Modell zu verstehen.

Darüber hinaus werden Limitationen bisheriger Forschung verhindert. Durch die Verwendung einer Online-Befragung kann das Risiko sozial erwünschter Antworten reduziert werden (G. Brown, 2009, S. 68; Kaya, 2009, S. 52), was insbesondere bei dem Thema Gehalt relevant sein kann. Aufgrund der Wahl einer Vignetten-Studie in Form der Conjoint-Analyse ist davon auszugehen, dass die hier vorliegenden Ergebnisse Aufschluss über kausale Zusammenhänge liefern können, da systematische Unterschiede aufgrund der veränderten Komponenten aufgedeckt werden können (Atzmüller und Steiner, 2010, S. 129; Hainmueller et al., 2014, S. 1–2). Studien mit ähnlichem Design wurden in der Management-Literatur insgesamt bisher nur in der Minderheit durchgeführt (Aguinis & Bradley, 2014, S. 352). Insbesondere die Präferenz von Arbeitnehmenden bezüglich diverser Jobangebote mit verschiedenen Attributen wurde bisher nur in einem sehr geringen Ausmaß mittels eines Conjoint-Experiments untersucht (Pouliakas & Theodossiou, 2010, S. 691). Dabei kann dieses Studiendesign das Dilemma zwischen interner und externer Validität lösen (Aguinis & Bradley, 2014, S. 366), da neben den beabsichtigten Manipulationen der Attribute weitere Veränderungen ausgeschlossen werden können und angegebene Präferenzen den wahren Präferenzen sehr nah kommen (Hainmueller et al., 2015, S. 2400; Pouliakas und Theodossiou, 2010, S. 691). Außerdem kann die Erstellung einer Rangfolge in dem Kontext der Jobsuche als realitätsnah angesehen werden, da sich Arbeitssuchende für ein Jobangebot entscheiden müssen und entsprechend zwischen den Angeboten abwägen müssen (Klein, 2002, S. 22).

Die Ergänzung der Vignetten-Studie durch die standardisierte Umfrage ermöglicht es außerdem die zugrundeliegenden Mechanismen und Moderationseffekte abzufragen, auch wenn in der vorliegenden Studie hinsichtlich persönlicher Charakteristika als Moderatoren keine drastisch abweichenden Tendenzen festgestellt wurden.

5.3. Limitationen

Trotz der Relevanz und des zuvor herausgestellten Beitrages weist die vorliegende Studie auch Limitationen in Bezug auf die interne und externe Validität auf, welche im Folgenden beschrieben werden. Hinsichtlich der *internen Validität* lassen sich zunächst einige Aspekte der gewählten Methode kritisieren. Dabei wird zunächst auf die Vignetten-Studie für Forschungsfrage 1 in Form der Conjoint-Analyse eingegangen und anschließend auf die standardisierte Umfrage, welche für die Beantwortung der Forschungsfrage 2 verwendet wird.

In der Literatur gibt es Hinweise darauf, dass es für Befragte schwierig sein kann ein vollständiges Ranking aller präsentierten Vignetten zu erstellen und die Regression mittels *Rank Ordered Logit*-Modell folglich verzerrte Ergebnisse liefern kann (Fok et al., 2012, S. 843). Um einen Bias zu verhindern, könnte bei jedem Befragten überprüft werden, ob die Präferenzen stringent sind und solche von der Auswertung ausgeschlossen werden, die nicht der Annahme des Homo oeconomicus entsprechen und beispielsweise ein Jobangebot mit gleicher Arbeitsplatzform und geringerem Gehalt besser bewertet haben. Alternativ könnte festgelegt werden, dass nur eine gewisse Anzahl der obersten Präferenzen in die Analyse miteinbezogen wird. Da dabei allerdings weitere Annahmen getroffen werden und wertvolle Informationen wegfallen können, wurde hier darauf verzichtet. Dennoch wäre dieses Vorgehen als Ergänzung sinnvoll, um zu überprüfen, ob die Ergebnisse verzerrt sein könnten. Die Conjoint-Analyse und die vollständige Erstellung eines Rankings werden insgesamt dennoch als sinnvoll erachtet, da diese realitätsnäher sind und viele Informationen liefern (Hausman & Ruud, 1987, S. 83).

Die Reduktion der Vignetten von 12 auf 8 geht zwar mit einem Informationsverlust einher, reduziert aber die Komplexität der Aufgabe (Aiman-Smith et al., 2002, S. 399; Atzmüller und Steiner, 2010, S. 130). Um die Komplexität weiter zu reduzieren, wurden die betrachteten Attribute streng limitiert. Es gibt allerdings Variablen, welche die Entscheidung beeinflussen könnten und nicht genau definiert wurden. Ein Beispiel ist der Fahrtweg, den Befragte bei der Beurteilung angenommen haben. Hybrides Arbeiten könnte etwa auch einen geldlichen Vorteil darstellen, wenn ein langer Fahrtweg wegfällt, dafür wurde allerdings nicht kontrolliert. Die Bürosituation wurde ebenfalls nicht im Detail erklärt (De Croon et al., 2005, S. 130), eine bessere Ausstattung mit höhenverstellbaren Schreibtischen könnte non-territoriales Arbeiten beispielsweise attraktiver machen. Eine Ergänzung der kurzen Beschreibung durch ein realitätsnahes Foto hätte außerdem hilfreich sein können, um für die Arbeitsbedingung vor Ort zu kontrollieren. Ebenso wurde nicht näher auf die Arbeitsaufgaben im Bereich Controlling eingegangen, obwohl die Aufgaben die Wahl der Arbeitsplatzform beeinflussen können (Volker & van der Voordt, 2005, S. 247). Insgesamt hätten folglich organisationale, personenbezogene und aufgabenbezogene Moderatoren genauer erhoben werden können (Wohlers & Hertel, 2017, S. 482), um die interne Validität weiter zu erhöhen. Für frühere Erfahrungen mit den Arbeitsplatzformen wird allerdings kontrolliert (Pouliakas & Theodossiou, 2010, S. 704) und diese weisen keine drastischen Unterschiede auf.

Das *Rank Ordered Logit*-Modell beruht auf der Annahme der Unabhängigkeit irrelevanter Alternativen (IIA), welche durch den Hausman-Test überprüft werden kann (Hausman & Ruud, 1987, S. 83). Dieser Test ist durch das Design des Conjoint-Experiments aber nicht aufschlussreich (Cheng & Long, 2007, S. 598), da es durch die Vignetten eine eingeschränkte Auswahl gibt, sodass angenommen werden muss, dass die IIA hält (Fok et al., 2012, S. 844). Eine Studie nach Calfee et al. (2001) hat allerdings ergeben, dass die Ergebnisse des herkömmlichen *Rank Ordered Logit*-Modells nicht wesentlich von Ergebnissen des Mixed Logit-Modells abweichen, bei dem Fehlerterme korreliert sein dürfen, sodass auch herkömmliche Modelle verwendet werden können (Calfee et al., 2001, S. 703–705).

Weiterhin wird bei der Conjoint-Analyse eine lineare und additive Nutzenfunktion unterstellt. Bei der Analyse wurden die Kombinationen aus non-territorialem und hybridem Arbeiten als eine Arbeitsplatzform angenommen und die Merkmale nicht getrennt voneinander betrachtet, sodass auch mögliche Interaktionseffekte nicht berücksichtigt werden konnten. Dies entspricht zwar einer bei Conjoint-Analysen häufig getroffenen Annahme (Klein, 2002, S. 11–12), es stellt sich dennoch die Frage, ob die Kombination der Merkmale relevant ist.

In Bezug auf die standardisierte Umfrage und Forschungsfrage 2 können die zur Erfassung der Arbeitsbedingungen entwickelten Skalen bestehend aus jeweils drei Items kritisiert werden, da in der Literatur keine geeigneten bereits validierten Skalen identifiziert werden konnten und entsprechend neue Skalen gebildet werden mussten. Diese wurden zwar anhand der Literatur mit größter Sorgfalt erstellt, die Güte konnte vorher allerdings nicht durch diverse Datenerhebungen belegt werden. Die Einholung zusätzlichen Expertenfeedbacks wäre hilfreich, um die Skalen zu verbessern und die Inhaltsvalidität sicher zu stellen. Es wurden allerdings statistische Tests durchgeführt, um die Konsistenz und Konstruktvalidität zu testen.

Cronbachs Alpha zeigt zwar eine starke interne Konsistenz der Items an, bei Werten über 0,9 kann allerdings hinterfragt werden, ob redundante Items inkludiert sind (Streiner, 2003, S. 103). So könnte etwa T1 bei zukünftigen Erhebungen weggelassen werden, da Cronbachs Alpha nur marginal von 0,929 auf 0,921 sinken würde. Außerdem steigt der Wert mit der Anzahl der Items, das Risiko einer Verzerrung durch die Anzahl der Items ist jedoch gering, da nur 3 je Konstrukt inkludiert sind (Tavakol & Dennick, 2011, S. 53). Es muss allerdings beachtet werden, dass der Wert keinen Aufschluss über die inhaltliche Passung angibt.

Da für die hier betrachteten Arbeitsbedingungen keine etablierten Skalen identifiziert werden konnten, wäre eine explorative Faktorenanalyse (EFA) ebenfalls hilfreich gewesen, um zunächst die Struktur der Skalen kennenzulernen (T. A. Brown & Moore, 2012, S. 362). Darauf wurde allerdings verzichtet, da die Skalen theoriegeleitet erstellt wurden und direkt die Passung des Modells durch die KFA ermittelt wurde (El-Den et al., 2020, S. 330). Da die Items mittels Likert-Skala erhoben wurden, lässt sich allerdings auch die Schätzmethode der KFA kritisieren. Nach Morata-Ramírez und Holgado-Tello (2013, S. 54) wird die Verwendung der robusten ML gegenüber der ML zwar bevorzugt, es kann dennoch hinterfragt werden, ob Likert-Skalen intervallskalierte Daten liefern. Wenn diese als ordinal angesehen werden, sollte die robuste ungewichtete Kleinste-Quadrate-Methode verwendet werden, sodass eine zusätzliche Schätzung zur Absicherung sinnvoll sein könnte.

Nach El-Den et al. (2020) gibt es neben der Bildung der latenten Variablen aus dem Mittelwert der entsprechenden Items anspruchsvollere und durchdachtere Möglichkeiten. Die Items, welche eine Skala bilden, können anhand der KFA unterschiedlich gewichtet werden, sodass Items mit hoher Ladung stärker in das Gesamtergebnis einbezogen werden oder Items, welche unterschiedliche Dimensionen beinhalten, getrennt betrachtet werden (El-Den et al., 2020, S. 333).

Insgesamt wäre auch die Ergänzung der Umfrage durch einen Attention-Check sinnvoll gewesen, sodass dieser als Ausschlusskriterium hätte herangezogen werden können, um Datensätze ausschließen zu können, wenn Befragte nicht aufmerksam antworten. Außerdem könnten die Items je Arbeitsplatzform in zufälliger Reihenfolge erscheinen, sodass Reihenfolgeeffekte wie bei der Vignetten-Studie verhindert werden. Dies wurde allerdings nicht umgesetzt, da die Bewertung der Arbeitsplatzformen hinsichtlich der Arbeitsbedingungen schneller erfolgen kann, wenn die Items in derselben Reihenfolge präsentiert werden.

Die *externe Validität* der Studienergebnisse muss ebenfalls berücksichtigt werden. Dabei kann zunächst die gewählte Stichprobe per Convenience Sample kritisiert werden. Dieses geht zwar mit den geringsten Kosten einher, sollte aber hinsichtlich der Repräsentativität der Stichprobe für die Grundgesamtheit hinterfragt werden (Kaya & Himme, 2009, S. 83). Da es sich bei der Wahl eines Jobangebotes allerdings um eine Situation handelt, welche vielen Menschen bekannt ist, wurde zu Gunsten einer größeren Stichprobe auf strengere Selektionskriterien verzichtet (Aguinis & Bradley, 2014, S. 363).

Außerdem können als letzter Aspekt die Auswirkungen des im Vignetten-Experiment festgelegten Szenariums auf die externe Validität kritisch betrachtet werden. In dem Szenarium wurde festgelegt, dass die Arbeit im Betrieb in einem Großraumbüro mit 20 Personen verrichtet wird. Studien belegen allerdings, dass schon Großraumbüros mit mehr als 15 Personen vermieden werden sollten (Brunia et al., 2016, S. 43), sodass Befragte in diesem Kontext etwa eher hybrides Arbeiten bevorzugen könnten als wenn es sich um ein Großraumbüro mit 10 Personen handeln würde. Dass räumliche Faktoren neben der Arbeitsplatzform eine wesentliche Rolle spielen, belegen auch Kim et al. (2016, S. 203), da diese einen signifikanteren Einfluss auf die Zufriedenheit von Arbeitnehmenden haben als das dort untersuchte non-territoriale Arbeiten selbst. Außerdem werden non-territoriales und hybrides Arbeiten nicht nuanciert betrachtet, obwohl einzigartige Effekte durch unterschiedliche Formen möglich sind (Gatt & Jiang, 2021, S. 957). Es kann beispielsweise einen Unterschied machen, ob hybrides Arbeiten bedeutet, dass Arbeitnehmende frei entscheiden dürfen, wie viele Tage sie von zu Hause aus arbeiten, oder dass eine gewisse Anzahl an Tagen vorgeschrieben wird, da eine höhere Intensität etwa die Beziehung zu dem Kollegium verschlechtern kann (Gajendran & Harrison, 2007, S. 1538). Die Studie nach Brunia et al. (2016, S. 30) zeigt ebenfalls, dass eine nuancierte Betrachtung wichtig ist, da die Zufriedenheit von Arbeitnehmenden mit gleichen Bürokonzepten und lediglich vermeintlich kleinen Unterschieden wie dem Design der Büroräume oder der Raumaufteilung signifikant beeinflusst wird.

Da in der vorliegenden Studie die Präferenz bei der Wahl eines Jobangebotes untersucht wird, können einige dieser Aspekte weniger kritisch betrachtet werden, da Arbeitssuchenden in der Realität häufig auch nicht alle Informationen über den zukünftigen Arbeitsplatz zur Verfügung stehen. Dennoch sollte berücksichtigt werden, dass die Ergebnisse der Studie nicht ausnahmslos auf Arbeitsangebote mit Kombinationen dieser Arbeitsplatzformen und abweichenden Ausprägungen übertragen werden können.

5.4. Implikationen

Trotz der zuvor beschriebenen Limitationen wurden entsprechend des *Entscheidungspunktes 10* nach Aguinis und Bradley (2014, S. 364) die angewendete Methode und die Ergebnisse transparent dokumentiert und erläutert, sodass die Forschung reproduzierbar ist und als Grundlage für weitere Untersuchungen in Bezug auf non-territoriale und hybride Arbeitsplatzformen genutzt werden kann und Ergebnisse hinsichtlich der Relevanz für die Praxis eingeordnet werden können. Daraus ergeben sich eine Reihe an Implikationen für die Forschung und Praxis, auf die nachfolgend eingegangen wird.

In Bezug auf die *Forschung* könnte eine nuancierte Betrachtung der hier untersuchten Arbeitsplatzformen sinnvoll sein, da das in dem Vignetten-Experiment gewählte Szenarium und die Ausprägung der Arbeitsplatzform einen Einfluss auf die Ergebnisse haben könnten, wie im Rahmen der Limitationen beschrieben. In Bezug auf hybrides Arbeiten sollte untersucht werden, wie die Intensität und die Flexibilität hinsichtlich der Homeoffice-Tage die Ergebnisse beeinflusst, um mehr darüber zu erfahren, welcher Typ von hybridem Arbeiten präferiert wird (Gajendran & Harrison, 2007,

S. 1538). Außerdem sollte zur Kontrolle die Situation im Homeoffice abgefragt werden, da diese die Entscheidung beeinflussen könnte. Ebenso sollten unterschiedliche Formen des non-territorialen Arbeitens hinzugezogen werden, indem unterschieden wird in welchen Bereichen Arbeitsplätze gesucht werden, ob in dem gesamten Unternehmen oder in für die Abteilung oder das Team festgelegten Bereichen, und wie es umgesetzt wird, ob es ein Reservierungssystem gibt oder nicht (Frankó et al., 2022, S. 241). Um die externe Validität zu erhöhen, sollten weitere Studien durchgeführt werden, in denen andere Berufe untersucht werden, da die Aufgaben den Effekt der Arbeitsplatzform beeinflussen können (Ashkanasy et al., 2014, S. 1174). Andere Gegebenheiten im Büro sollten ebenfalls präsentiert werden, da Großraumbüros die Auswirkungen ebenfalls verändern können (Brunia et al., 2016, S. 43) und deren Effekt somit kontrolliert werden sollte. Die Literatur liefert erste Hinweise darauf, dass die Betrachtung weiterer Moderatoren ebenfalls relevant sein könnte, darunter fallen die Nationalität und persönliche Charaktermerkmale (G. Brown et al., 2005, S. 588) und persönliche Ressourcen (Xanthopoulou et al., 2007, S. 137).

In dieser Studie wird die Kombination aus non-territorialen und hybriden Arbeitsplatzformen untersucht, dabei werden die zwei Merkmale allerdings nicht getrennt voneinander berücksichtigt, sondern die Präferenz hinsichtlich der gesamten Arbeitsplatzform erfasst. Da die Regressionsergebnisse zur Beantwortung der Forschungsfrage 1 allerdings zeigen, dass die beiden Kombinationen mit hybrider Arbeitsplatzform am stärksten präferiert werden, kann vermutet werden, dass hybrides Arbeiten gegenüber nicht-hybridem Arbeiten den Befragten wichtiger ist als die Unterscheidung zwischen territorialem oder non-territorialem Arbeiten. Um genauer aufschlüsseln zu können, warum eine Arbeitsplatzform präferiert wird, sollte untersucht werden, ob die Merkmale wirklich additiv wirken, ob sie gleich stark gewichtet werden und ob Wechselwirkungen auftreten, indem Interaktionseffekte zwischen diesen Merkmalen untersucht werden (Hainmueller et al., 2014, S. 12).

Neben der genaueren Betrachtung der einzelnen Merkmale sollte eine Strukturgleichungsanalyse durchgeführt werden, da dabei komplexe Kausalbeziehungen berücksichtigt werden können (Backhaus et al., 2015, S. 12). Dadurch könnten sowohl die übergreifende Beziehung zwischen der Arbeitsplatzform und der Präferenz als auch die zugrundeliegenden Mechanismen untersucht werden. Mithilfe der Strukturgleichungsanalyse könnte eine richtige Mediationsanalyse durchgeführt werden, da dabei auch latente Konstrukte wie die Arbeitsbedingungen berücksichtigt werden können. Zwar kann ein Modell niemals alle relevanten Variablen beinhalten, welche die Beziehung beeinflussen (Aiman-Smith et al., 2002, S. 406), dennoch würde diese Analyse näher an reale Prozesse herankommen als die getrennte Betrachtung, da dabei hypothetische kausale Pfade gleichzeitig geschätzt werden (Radermacher, 2019, S. 101).

Zusätzlich sollten die zugrundeliegenden Mechanismen weiter untersucht werden. Um sicherzustellen, dass neben den vier betrachteten Arbeitsbedingungen keine relevanten Aspekte bei non-territorialem und hybridem Arbeiten außer Acht gelassen wurden, könnte eine explorative Studie durchgeführt werden, um mehr über Hintergründe der Entscheidung zu erfahren. Außerdem sollten Messinstrumente für die Arbeitsbedingungen weiter verbessert werden, da es teilweise noch keine etablierten Skalen gibt, wie für Territorialität (G. Brown, 2009, S. 16), oder die vorhandenen Skalen nicht auf andere Kontexte anwendbar sind, wie die der Autonomie, da sich Messinstrumente häufig auf die Methode und Entscheidungen beziehen (Lumpkin et al., 2009, S. 53) und für den Kontext der Arbeitsplatzform damit nicht geeignet sind.

Als letzte Implikation für die Forschung kann die Betrachtung von non-territorialen und hybriden Arbeitsplatzformen anhand weiterer abhängiger Variablen angegeben werden, sowie eine Untersuchung der langfristigen Wirkung der Arbeitsplatzform auf diese. Die vorliegende Studie in Form der Conjoint-Analyse betrachtet lediglich die Präferenz bei der hypothetischen Jobwahl. Dabei ist allerdings noch ungeklärt, inwieweit die Arbeitsplatzform als Symbol die Entscheidung von Jobsuchenden beeinflusst (Maier et al., 2022, S. 2), da diese über die bloße Gestaltung der physischen Arbeitsumgebung hinausgeht, welche als Symbol bereits einen Einfluss auf das Verhalten und die Leistung der Individuen haben kann (Berg & Kreiner, 1990, S. 65). Weiterhin gibt es Hinweise darauf, dass die Arbeitsplatzform auch die Produktivität (Shobe, 2018, S. 4), Gesundheit (Lesener et al., 2019, S. 94) und Arbeitszufriedenheit (Bodin Danielsson & Theorell, 2019, S. 1021) beeinflussen kann, sodass diese im Zusammenhang mit non-territorialen und hybriden Arbeitsplatzformen untersucht werden können. Bei diesen abhängigen Variablen können Langzeitstudien relevant sein, um kausale Ergebnisse zu validieren (Xanthopoulou et al., 2007, S. 138), da kurzfristige Änderungen nach Einführung neuer Arbeitsplatzkonzepte möglich sind (Allen & Gerstberger, 1973, S. 495) und erst Langzeitstudien zeigen, dass sich Anforderungen, Ressourcen und Ergebnisse auch wechselseitig beeinflussen können (Lesener et al., 2019, S. 76; Schaufeli und Taris, 2014, S. 57–58). Dabei sollte berücksichtigt werden, dass neue Strategien aufkommen können. Eine Person hat im Fragebogen etwa angegeben, dass sie bereits in einem non-territorialen Büro gearbeitet hat, sich aber einen festen Platz geschaffen hat, indem immer der gleiche Platz gewählt wurde. Dieses Verhalten bestätigt auch eine Studie nach Volker und van der Voordt (2005, S. 248–249). Langzeitstudien mit veränderten abhängigen Variablen könnten diese Folgen von non-territorialen und hybriden Arbeitsplätzen erfassen. Durch Feldexperimente kann untersucht werden, ob sich die Präferenzen verändern, wenn Arbeitnehmende langfristig in einem Job mit der Arbeitsplatzform arbeiten.

Neben Implikationen für die Forschung können anhand der Studie auch Implikationen für die *Praxis* abgeleitet werden. Die Studie hat gezeigt, dass territoriale und hybride Arbeitsplätze im Vergleich mit den anderen Kombinationen am beliebtesten sind und Arbeitnehmende sogar bereit sind dafür mit 2.175 € auf fast 5 % des Gehalts zu verzichten. Außerdem wird belegt, dass die Arbeitsplatzform Auswirkungen auf die Privatsphäre, Territorialität, Zusammenarbeit und Autonomie hat.

Durch die Ergebnisse wird deutlich, dass *Arbeitgebende* die Arbeitsplatzform nutzen können, um attraktivere Angebote für Arbeitnehmende zu schaffen, falls es sich mit der Arbeit vereinbaren lässt. Darüber hinaus ist es möglich, dass es sich wirtschaftlich lohnt eine bestimmte Arbeitsplatzform anzubieten, da Arbeitnehmende sogar bereit sein können für die präferierte Arbeitsform auf einen Teil des Gehalts zu verzichten. Dann könnte Geld für weitere Benefits bereitgestellt werden, um als noch attraktiver wahrgenommen zu werden. Entsprechend sollte auch wirtschaftlich betrachtet nicht nur die Reduzierung der Kosten durch Wegfall eines Gebäudes dank einer neuen Arbeitsplatzform berücksichtigt werden, sondern auch die Präferenz der Arbeitnehmenden abgefragt werden. Weiterhin sollten Arbeitgebende die vier untersuchten Arbeitsbedingungen bei der Gestaltung der Arbeitsräume beachten, um sicher zu stellen, dass die Arbeitnehmenden zufrieden sind. Es können beispielsweise abgetrennte Arbeitsbereiche in große Büroräume integriert werden, um mehr Privatsphäre sicher zu stellen, und offene Bereiche, um die Zusammenarbeit zu fördern. Insgesamt zeigt die Studie Arbeitgebenden, dass die Arbeitsplatzform für Arbeitnehmende relevant ist und damit letztlich auch auf organisationale Ergebnisse wirken kann.

Vorgesetzte sollten besonders die Arbeitsbedingungen im Blick behalten und mit den Arbeitnehmenden in den Austausch gehen, ob sich diese Verbesserungen wünschen. Außerdem weist das JDR-Modell darauf hin, dass auch Job-Anforderungen beachtet werden sollten, da eine den Präferenzen der Arbeitnehmenden entsprechende Arbeitsplatzform kein Allheilmittel ist, sondern nur einen positiven Beitrag leistet. Zu hohe Anforderungen trotz vorhandener Ressourcen sind die wichtigsten Prädikatoren von Erschöpfung (Xanthopoulou et al., 2007, S. 138).

Arbeitnehmende sollten ihre eigenen Bedürfnisse klar kommunizieren und mit Vorgesetzten abstimmen, ob eine Anpassung der Arbeitsplatzform entsprechend den individuellen Bedürfnissen möglich ist oder eine Verbesserung der Ressourcen, falls gewünscht. Wenn sogar eine Bereitschaft besteht auf einen Teil des Gehalts für die entsprechende Arbeitsplatzform zu verzichten, beispielsweise weil durch hybrides Arbeiten ein langer und kostenreicher Arbeitsweg teilweise wegfällt, kann ein Verzicht auf einen Teil des Gehaltes angeboten werden, wenn der Arbeitgebende vorher nicht dem Wunsch entgegenkommen möchte, obwohl die Arbeitsplatzform mit den Aufgaben vereinbar ist, dies sollte individuell abgewogen werden.

Für die *Gesellschaft* können die Ergebnisse außerdem relevant sein, da die Studie einen ersten Hinweis liefert, dass insbesondere hybrides Arbeiten von Arbeitnehmenden präferiert wird und neue Arbeitsplatzformen wie hybrides Arbeiten durch den Wegfall des Arbeitsweges positive Auswirkungen auf die Umwelt haben könnten. Weiterhin sollte eine digitale Infrastruktur sichergestellt werden, welche die Voraussetzung für die Umsetzung neuer Arbeitsplatzformen ist. Da neue Arbeitsplatzformen an Relevanz gewinnen, sollten Arbeitsschutzmaßnahmen darauf ausgeweitet werden, sodass Unternehmen beispielsweise verpflichtet werden könnten höhenverstellbare Schreibtische bei non-territorialen Arbeitsplatzformen bereitzustellen, sodass die Gesundheit der Arbeitnehmenden nicht gefährdet wird.

Insgesamt zeigt die vorliegende Arbeit, dass die Kombination der non-territorialen und hybriden Arbeitsplatzform bisher eine Forschungslücke darstellt, obwohl diese die Präferenz von Arbeitnehmenden und die Bereitschaft auf Gehalt zu verzichten maßgeblich beeinflusst. Diese und weitere neu entstehende Arbeitsplatzformen können vielfältige Auswirkungen auf Individuen, Organisationen und die Gesellschaft haben, sodass weitere Forschung in diesem Gebiet relevant ist und zugrundeliegende Mechanismen vollständig aufgedeckt werden sollten.

Literatur

Aguinis, H., & Bradley, K. J. (2014). Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods*, *17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Aiman-Smith, L., Scullen, S. E., & Barr, S. H. (2002). Conducting Studies of Decision Making in Organizational Contexts: A Tutorial for Policy-Capturing and Other Regression-Based Techniques. *Organizational Research Methods*, *5*(4), 388–414. https://doi.org/10.1177/109442802237117

Allen, T. J., & Gerstberger, P. G. (1973). A Field Experiment to Improve Communications in a Product Engineering Department: The Nonterritorial Office. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *15*(5), 487–498. https://doi.org/10.1177/001872087301500505

Altman, I. (1975). *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. Brooks/Cole Publishing Company.

Appel-Meulenbroek, R., Kemperman, A., van de Water, A., Weijs-Perrée, M., & Verhaegh, J. (2022). How to attract employees back to the office? A stated choice study on hybrid working preferences. *Journal of Environmental Psychology*, *81*, 1–12. https://doi.org/10.1016/j.jenvp.2022.101784

Ashforth, B. E., & Mael, F. (1989). Social Identity Theory and the Organization. *The Academy of Management Review*, *14*(1), 20–39. https://doi.org/10.2307/258189

Ashkanasy, N. M., Ayoko, O. B., & Jehn, K. A. (2014). Understanding the physical environment of work and employee behavior: An affective events perspective. *Journal of Organizational Behavior*, *35*(8), 1169–1184. https://doi.org/10.1002/job.1973

Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology*, *6*(3), 128–138. https://doi.org/10.1027/1614-2241/a000014

Backhaus, K., Erichson, B., & Weiber, R. (2015). *Fortgeschrittene Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (3. Aufl.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-46087-0

Bakker, A. B., & Demerouti, E. (2007). The Job Demands-Resources model: state of the art. *Journal of Managerial Psychology*, *22*(3), 309–328. https://doi.org/10.1108/02683940710733115

Bakker, A. B., & Demerouti, E. (2017). Job demands-resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, *22*(3), 273–285. https://doi.org/10.1037/ocp0000056

Bartz, M., & Schwand, C. (2017). Preis der Freiheit - Nutzen von Spielregeln für Mobil-Flexibles Arbeiten. http://ffhoarep.fh-ooe.at/bitstream/123456789/1024/1/Panel_120_ID_180.pdf

Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, *17*(1), 1–19. https://doi.org/10.1016/0304-4076(81)90056-7

Bencivenga, M., & Camocini, B. (2022). Post-pandemic scenarios of office workplace: new purposes of the physical spaces to enhance social and individual well-being. In A. Dominoni & F. Scullica (Hrsg.), *Designing Behaviours For Well-Being Spaces: How disruptive approaches can improve living conditions* (S. 90–111). FrancoAngeli.

Berg, P. O., & Kreiner, K. (1990). Corporate Architecture: Turning Physical Settings into Symbolic Resources. In P. Gagliardi (Hrsg.), *Symbols and Artifacts* (S. 41–67, Bd. 24). De Gruyter. https://doi.org/10.4324/9781315130538-2

Bhave, D. P., Teo, L. H., & Dalal, R. S. (2020). Privacy at Work: A Review and a Research Agenda for a Contested Terrain. *Journal of Management*, *46*(1), 127–164. https://doi.org/10.1177/0149206319878254

Bodin Danielsson, C., & Theorell, T. (2019). Office Employees' Perception of Workspace Contribution: A Gender and Office Design Perspective. *Environment and Behavior*, *51*(9-10), 995–1026. https://doi.org/10.1177/0013916518759146

Brown, G. (2009). Claiming a corner at work: Measuring employee territoriality in their work-spaces. *Journal of Environmental Psychology*, *29*(1), 44–52. https://doi.org/10.1016/j.jenvp.2008.05.004

Brown, G., Lawrence, T. B., & Robinson, S. L. (2005). Territoriality in Organizations. *Academy of Management Review*, *30*(3), 577–594. https://doi.org/10.5465/amr.2005.17293710

Brown, T. A., & Moore, M. T. (2012). Confirmatory Factor Analysis. In R. H. Hoyle (Hrsg.), *Handbook of Structural Equation Modeling* (2. Aufl., S. 361–379). Guilford Press.

Brunia, S., De Been, I., & van der Voordt, T. J. M. (2016). Accommodating new ways of working: lessons from best practices and worst cases. *Journal of Corporate Real Estate*, *18*(1), 30–47. https://doi.org/10.1108/JCRE-10-2015-0028

Brunia, S., & Hartjes-Gosselink, A. (2009). Personalization in non-territorial offices: a study of a human need. *Journal of Corporate Real Estate*, *11*(3), 169–182. https://doi.org/10.1108/14630010910985922

Bustelo, M., Díaz, A. M., Lafortune, J., Piras, C., Salas Bahamón, L., & Tessada, J. (2020). What is The Price of Freedom? Estimating Women's Willingness to Pay for Job Schedule Flexibility. https://doi.org/10.18235/0002286

Calfee, J., Winston, C., & Stempski, R. (2001). Econometric Issues in Estimating Consumer Preferences from Stated Preference Data: A Case Study of the Value of Automobile Travel Time. *The Review of Economics and Statistics*, *83*(4), 699–707. https://doi.org/10.1162/003465301753237777

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press. https://doi.org/10.1017/cbo9780511811241

Candido, C., Kim, J., de Dear, R., & Thomas, L. (2016). BOSSA: a multidimensional post-occupancy evaluation tool. *Building Research & Information*, *44*(2), 214–228. https://doi.org/10.1080/09613218.2015.1072298

Castrillon, C. (2022, Dezember). The Top Workplace Trends For 2023. https://www.forbes.com/sites/carolinecastrillon/2022/12/11/the-top-workplace-trends-for-2023/

Cheng, S., & Long, J. S. (2007). Testing for IIA in the Multinomial Logit Model. *Sociological Methods & Research*, *35*(4), 583–600. https://doi.org/10.1177/0049124106292361

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203771587

Cohen, S. (1978). Environmental load and the allocation of attention. In A. Baum, J. E. Singer & S. Valins (Hrsg.), *Advances in Environmental Psychology* (1. Aufl., S. 1–29). Lawrence Erlbaum Associates, Publishers.

Dalal, A. K., & Singh, R. (1986). An Integration Theoretical Analysis of Expected Job Attractiveness and Satisfaction. *International Journal of Psychology*, *21*(1-4), 555–564. https://doi.org/10.1080/00207598608247606

Davis, M. C., Leach, D. J., & Clegg, C. W. (2020). Breaking Out of Open-Plan: Extending Social Interference Theory Through an Evaluation of Contemporary Offices. *Environment and Behavior*, *52*(9), 945–978. https://doi.org/10.1177/0013916519878211

De Been, I., & Beijer, M. (2014). The influence of office type on satisfaction and perceived productivity support. *Journal of Facilities Management*, *12*(2), 142–157. https://doi.org/10.1108/JFM-02-2013-0011

De Croon, E. M., Sluiter, J. K., Kuijer, P. P. F. M., & Frings-Dresen, M. H. W. (2005). The effect of office concepts on worker health and performance: a systematic review of the literature. *Ergonomics*, *48*(2), 119–134. https://doi.org/10.1080/00140130512331319409

Deci, E. L., Olafsen, A. H., & Ryan, R. M. (2017). Self-Determination Theory in Work Organizations: The State of a Science. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 19–43. https://doi.org/10.1146/annurev-orgpsych-032516-113108

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology*, *53*(6), 1024–1037. https://doi.org/10.1037/0022-3514.53.6.1024

Deci, E. L., & Ryan, R. M. (2000). The "Whatänd "Whyöf Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology*, *86*(3), 499–512. https://doi.org/10.1037/0021-9010.86.3.499

Demerouti, E., Derks, D., Brummelhuis, L. L. t., & Bakker, A. B. (2014). New Ways of Working: Impact on Working Conditions, Work-Family Balance, and Well-Being. In C. Korunka & P. Hoonakker (Hrsg.), *The Impact of ICT on Quality of Working Life* (S. 123–141). Springer. https://doi.org/10.1007/978-94-017-8854-0_8

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal*, *15*(1), 292–300. https://doi.org/10.1177/1536867X1501500117

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Dunn, O. J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, *6*(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181

El-Den, S., Schneider, C., Mirzaei, A., & Carter, S. (2020). How to measure a latent construct: Psychometric principles for the development and validation of measurement instruments. *International Journal of Pharmacy Practice*, *28*(4), 326–336. https://doi.org/10.1111/ijpp.12600

Elsbach, K. D. (2003). Relating Physical Environment to Self-Categorizations: Identity Threat and Affirmation in a Non-Territorial Office Space. *Administrative Science Quarterly*, *48*(4), 622–654. https://doi.org/10.2307/3556639

Elsbach, K. D., & Pratt, M. G. (2007). The Physical Environment in Organizations. *Academy of Management Annals*, *1*(1), 181–224. https://doi.org/10.5465/078559809

Fok, D., Paap, R., & van Dijk, B. (2012). A Rank-Ordered Logit Model with Unobserved Heterogeneity in Ranking Capabilities. *Journal of Applied Econometrics*, *27*(5), 831–846. https://doi.org/10.1002/jae.1223

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Frankó, L., Erdélyi, A., & Dúll, A. (2022). Transformation of the office: territorial behaviour and place attachment in shared desk design. *Journal of Corporate Real Estate*, *25*(3), 229–245. https://doi.org/10.1108/JCRE-12-2021-0043

Gajendran, R. S., & Harrison, D. A. (2007). The good, the bad, and the unknown about telecommuting: meta-analysis of psychological mediators and individual consequences. *Journal of Applied Psychology*, *92*(6), 1524–1541. https://doi.org/10.1037/0021-9010.92.6.1524

Gatt, G., & Jiang, L. (2021). Can Different Types of Non-Territorial Working Satisfy Employees' Needs for Autonomy and Belongingness? Insights From Self-Determination Theory. *Environment and Behavior*, *53*(9), 953–986. https://doi.org/10.1177/0013916520942603

Gehaltscheck: Gehalt für Controller:in. (2023). https://www.kununu.com/de/gehalt/controller-in-30993

Gonsalves, L. (2023). Work Un(Interrupted): How Non-territorial Space Shapes Worker Control over Social Interaction. *Organization Science*, *34*(5), 1–21. https://doi.org/10.1287/orsc.2022.1649

Gordon Brown, M. (2008). Proximity and collaboration: measuring workplace configuration. *Journal of Corporate Real Estate*, *10*(1), 5–26. https://doi.org/10.1108/14630010810881630

Gove, W. R. (1978). Social Psychology. Reviewed Work: "The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding." by Irwin Altman. *Contemporary Sociology*, *7*(5), 638. https://doi.org/10.2307/2065073

Greving, B. (2009). Messen und Skalieren von Sachverhalten. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (2. Aufl., S. 65–78). Gabler. https://doi.org/10.1007/978-3-8349-9121-8_5

Haapakangas, A., Hongisto, V., Varjo, J., & Lahtinen, M. (2018). Benefits of quiet workspaces in open-plan offices – Evidence from two office relocations. *Journal of Environmental Psychology*, *56*, 63–75. https://doi.org/10.1016/j.jenvp.2018.03.003

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, *60*(2), 159–170. https://doi.org/10.1037/h0076546

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395–2400. https://doi.org/10.1073/pnas.1416587112

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Halford, S. (2005). Hybrid workspace: re-spatialisations of work, organisation and management. *New Technology, Work and Employment*, *20*(1), 19–33. https://doi.org/10.1111/j.1468-005X.2005.00141.x

Hausman, J. A., & Ruud, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, *34*(1-2), 83–104. https://doi.org/10.1016/0304-4076(87)90068-6

He, H., Neumark, D., & Weng, Q. (2021). Do Workers Value Flexible Jobs? A Field Experiment. *Journal of Labor Economics*, *39*(3), 709–738. https://doi.org/10.1086/711226

Herrmann, M., Shikano, S., Thurner, P. W., & Becker, A. (2006). Die Analyse von Wählerpräferenzen mit Rank Ordered Logit. In U. Druwe, V. Kunz & T. Plümper (Hrsg.), *Jahrbuch für Handlungs- und Entscheidungstheorie* (S. 113–134). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90465-8_4

Hildebrandt, L., & Temme, D. (2006). *Probleme der Validierung mit Strukturgleichungsmodellen* (Discussion Paper Nr. 82). Humboldt-Universität zu Berlin, Collaborative Research Center 649 - Economic Risk.

Hirogaki, M. (2013). Estimating Consumers' Willingness to Pay for Health Food Claims: A Conjoint Analysis. *International Journal of Innovation, Management and Technology*, *4*(6), 541–546. https://doi.org/10.7763/ijimt.2013.v4.458

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multi-disciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Inamizuu, N. (2013). Positive Effect of Non-territorial Office on Privacy: Allen's Experiment Secret. *Annals of Business Administrative Science*, *12*(3), 111–121. https://doi.org/10.7880/abas.12.111

Jassawalla, A. R., & Sashittal, H. C. (1998). An Examination of Collaboration in High-Technology New Product Development Processes. *Journal of Product Innovation Management*, *15*(3), 237–254. https://doi.org/10.1111/1540-5885.1530237

Karasek, R. A. (1979). Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign. *Administrative Science Quarterly*, *24*(2), 285–308. https://doi.org/10.2307/2392498

Kaya, M. (2009). Verfahren der Datenerhebung. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (2. Aufl., S. 49–64). Gabler. https://doi.org/10.1007/978-3-322-96406-9_4

Kaya, M., & Himme, A. (2009). Möglichkeiten der Stichprobenbildung. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (2. Aufl., S. 79–88). Gabler. https://doi.org/10.1007/978-3-8349-9121-8_6

Khazanchi, S., Sprinkle, T. A., Masterson, S. S., & Tong, N. (2018). A Spatial Model of Work Relationships: The Relationship-Building and Relationship-Straining Effects of Workspace Design. *Academy of Management Review*, *43*(4), 590–609. https://doi.org/10.5465/amr.2016.0240

Kim, J., Candido, C., Thomas, L., & de Dear, R. (2016). Desk ownership in the workplace: The effect of non-territorial working on employee workplace satisfaction, perceived productivity and health. *Building and Environment*, *103*, 203–214. https://doi.org/10.1016/j.buildenv.2016.04.015

Klein, M. (2002). Die Conjoint-Analyse: eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, (50), 7–45. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-199069

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Leeper, T. J., Hobolt, S. B., & Tilley, J. (2019). Measuring Subgroup Preferences in Conjoint Experiments. *Political Analysis*, *28*(2), 207–221. https://doi.org/10.1017/pan.2019.30

Lesener, T., Gusy, B., & Wolter, C. (2019). The job demands-resources model: A meta-analytic review of longitudinal studies. *Work & Stress*, *33*(1), 76–103. https://doi.org/10.1080/02678373.2018.1529065

Lumpkin, G. T., Cogliser, C. C., & Schneider, D. R. (2009). Understanding and Measuring Autonomy: An Entrepreneurial Orientation Perspective. *Entrepreneurship Theory and Practice*, *33*(1), 47–69. https://doi.org/10.1111/j.1540-6520.2008.00280.x

Maier, L., Baccarella, C. V., Wagner, T. F., Meinel, M., Eismann, T., & Voigt, K.-I. (2022). Saw the office, want the job: The effect of creative workspace design on organization-al attractiveness. *Journal of Environmental Psychology*, *80*, 1–15. https://doi.org/10.1016/j.jenvp.2022.101773

Mangelsdorf, M. (2015). *Von Babyboomer bis Generation Z: Der richtige Umgang mit unterschiedlichen Generationen im Unternehmen*. Whitebooks. Gabal.

Mas, A., & Pallais, A. (2017). Valuing Alternative Work Arrangements. *American Economic Review*, *107*(12), 3722–3759. https://doi.org/10.1257/aer.20161500

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, *23*(2), 143–149. https://doi.org/10.11613/bm.2013.018

Mehmetoglu, M. (2015, April). 'CONDISC': Stata module to perform convergent and discriminant validity assessment in CFA. http://fmwww.bc.edu/repec/bocode/c/condisc.sthlp

Morata-Ramírez, M. D. L., & Holgado-Tello, F. P. (2013). Construct validity of Likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies*, *1*(1), 54–61. https://doi.org/10.11114/ijsss.v1i1.27

Morrison, R. L., & Macky, K. A. (2017). The demands and resources arising from shared office spaces. *Applied Ergonomics*, *60*, 103–115. https://doi.org/10.1016/j.apergo.2016.11.007

Morrow, P. C., McElroy, J. C., & Scheibe, K. P. (2012). Influencing organizational commitment through office redesign. *Journal of Vocational Behavior*, *81*(1), 99–111. https://doi.org/10.1016/j.jvb.2012.05.004

Müller, T., Schuberth, F., Bergsiek, M., & Henseler, J. (2022). How can the transition from office to telework be managed? The impact of tasks and workplace suitability on collaboration and work performance. *Frontiers in Psychology*, *13*, 1–18. https://doi.org/10.3389/fpsyg.2022.987530

Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and Application of the Kruskal-Wallis Test. *Applied Mechanics and Materials*, *611*, 115–120. https://doi.org/10.4028/www.scientific.net/amm.611.115

Pouliakas, K., & Theodossiou, I. (2010). Measuring the Utility Cost of Temporary Employment Contracts Before Adaptation: A Conjoint Analysis Approach. *Economica*, *77*(308), 688–709. https://doi.org/10.1111/j.1468-0335.2009.00786.x

Rack, O., & Christophersen, T. (2009). Experimente. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (2nd, S. 17–32). Gabler. https://doi.org/10.1007/978-3-322-96406-9_2

Radermacher, K. (2019). *How corporate architecture affects job seekers. Experimental evidence of signal-based mechanisms* [Diss., Universität Paderborn]. https://doi.org/10.17619/UNIPB/1-797

Radermacher, K., Schneider, M. R., Iseke, A., & Tebbe, T. (2017). Signalling to young knowledge workers through architecture? A conjoint analysis. *German Journal of Human Resource Management: Zeitschrift für Personalforschung*, *31*(1), 71–93. https://doi.org/10.1177/2397002216676038

Sack, R. D. (1986). *Human territoriality: Its theory and history*. Cambridge University Press.

Schaufeli, W. B., & Taris, T. W. (2014). A Critical Review of the Job Demands-Resources Model: Implications for Improving Work and Health. In G. F. Bauer & O. Hämmig (Hrsg.), *Bridging Occupational, Organizational and Public Health* (S. 43–68). Springer. https://doi.org/10.1007/978-94-007-5640-3_4

Sewell, G., & Taskin, L. (2015). Out of Sight, Out of Mind in a New World of Work? Autonomy, Control, and Spatiotemporal Scaling in Telework. *Organization Studies*, *36*(11), 1507–1529. https://doi.org/10.1177/0170840615593587

Shobe, K. (2018). Productivity Driven by Job Satisfaction, Physical Work Environment, Management Support and Job Autonomy. *Business and Economics Journal*, *9*(2), 1–9. https://doi.org/10.4172/2151-6219.1000351

Siegel, S. (1957). Nonparametric Statistics. *The American Statistician*, *11*(3), 13–19. https://doi.org/10.1080/00031305.1957.10501091

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology*, *1*(1), 27–41. https://doi.org/10.1037/1076-8998.1.1.27

Singh, A. S. (2017). Common procedures for development, validity and reliability of a questionnaire. *International Journal of Economics, Commerce and Management*, *5*(5), 790–801.

Singh, R. (1975). Information integration theory applied to expected job attractiveness and satisfaction. *Journal of Applied Psychology*, *60*(6), 621–623. https://doi.org/10.1037/0021-9010.60.5.621

Statista Research Department. (2022). Wie würdest du in Zukunft am liebsten arbeiten? Appinio-Umfrage zum Arbeitsplatz der Zukunft in Deutschland 2022. https://de.statista.com/statistik/daten/studie/1296962/umfrage/umfrage-arbeitsplatzder-zukunft/

Stegmann, S., van Dick, R., Ullrich, J., Charalambous, J., Menzel, B., Egold, N., & Wu, T. T.-C. (2010). Der Work Design Questionnaire. *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, *54*(1), 1–28. https://doi.org/10.1026/0932-4089/a000002

Streiner, D. L. (2003). Starting at the Beginning: an Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/s15327752jpa8001_18

Sundstrom, E., Burt, R. E., & Kamp, D. (1980). Privacy at Work: Architectural Correlates of Job Satisfaction and Job Performance. *Academy of Management Journal*, *23*(1), 101–117. https://doi.org/10.5465/255498

Sundstrom, E. D., & Sundstrom, M. G. (1986). *Work places: The psychology of the physical environment in offices and factories*. Cambridge University Press.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thatcher, S. M. B., & Zhu, X. (2006). Changing Identities in a Changing Workplace: Identification, Identity Enactment, Self-Verification, and Telecommuting. *Academy of Management Review*, *31*(4), 1076–1088. https://doi.org/10.5465/amr.2006.22528174

Thompson, R. J., Payne, S. C., & Taylor, A. B. (2015). Applicant attraction to flexible work arrangements: Separating the influence of flextime and flexplace. *Journal of Occupational and Organizational Psychology*, *88*(4), 726–749. https://doi.org/10.1111/joop.12095

Turner, G., & Myerson, J. (1998). *New workspace, new culture: Office design as a catalyst for change*. Routledge. https://doi.org/10.4324/9781315247977

van Hecke, T. (2012). Power study of ANOVA versus Kruskal-Wallis test. *Journal of Statistics and Management Systems*, *15*(2-3), 241–247. https://doi.org/10.1080/09720510.2012.10701623

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, *23*(2), 170–192. https://doi.org/10.3102/10769986023002170

Volker, L., & van der Voordt, T. (2005). An integral tool for the diagnostic evaluation of non-territorial offices. In B. Martens & A. G. Keul (Hrsg.), *Designing Social Innovation: Planning, Building, Evaluating* (S. 241–250). Hogrefe Publishing Group.

Wells, M. M. (2000). Office clutter or meaningful personal displays: The role of office personalization in employee and organizational wellbeing. *Journal of Environmental Psychology*, *20*(3), 239–255. https://doi.org/10.1006/jevp.1999.0166

Wohlers, C., & Hertel, G. (2017). Choosing where to work at work - towards a theoretical model of benefits and risks of activity-based flexible offices. *Ergonomics*, *60*(4), 467–486. https://doi.org/10.1080/00140139.2016.1188220

Xanthopoulou, D., Bakker, A. B., Demerouti, E., & Schaufeli, W. B. (2007). The role of personal resources in the job demands-resources model. *International Journal of Stress Management*, *14*(2), 121–141. https://doi.org/10.1037/1072-5245.14.2.121

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Exploring Discrepancies in Energy Performance Certificates: Analyzing Energy Efficiency Premiums for Buildings Based on Theoretical Energy Requirements Versus Actual Energy Consumption

Timo Andreas Deller

Technical University of Munich

Abstract

The building sector is lagging its needed decarbonization pathway. This paper examines EPC policy impacts on building economics in the Rhein-Main Region in Germany. Energy efficiency premiums for rents and sales prices and the effects of the EPC type are investigated using data from 01/2015 - 06/2023 (N = 212 167 rent sample; N = 159 573 sales sample) and hedonic price models. Energy efficiency premiums are present and range up to 7.0%, 4.6% and 6.9% for cold and warm rents and sales prices, respectively, when comparing an A+ to a D rated building. Consumption certificates reflect warm rents better but have a limited sales price impact. Results are rent efficiency premiums of up to 7.1% (A+), no rent discounts for energy inefficiency and a general sales price discount of about 3%. Requirement certificates are viewed as objective, yet less consumption-indicative, especially in the sales market. Rent efficiency premiums of up to 8.8% (A+) and no rent discounts for energy inefficiency are estimated for a building with a requirement certificate. Sales price efficiency premiums of up to 7.4% (A+) and sales price inefficiency discounts of up to -10.2% (H) exist. Overall, current German EPC policy does not address imperfect information, and it is recommended to revise its implementation.

Keywords: energy efficiency; energy performance certificate; EPC; hedonic price model; real estate investments; real estate valuation

1. Introduction & overview

The green transformation of the global economy continues to dominate key policy discussion points within national governments and international institutions. At the beginning of 2023, the Intergovernmental Panel on Climate Change (2023, pp. 4-11) presented the alarming current course of climate change and the insufficient actions taken by governments worldwide. Worryingly, the gap between specific sectors and their defined 2050 decarbonization pathways is widening (International Energy Agency and the United Nations Environment Program, 2022, p. 32). One significant driver of emissions is the building sector. The Global Alliance for Buildings and Constructions has indicated that the building sector is responsible for 37% of all energy-related emissions worldwide (International Energy Agency and the United Nations Environment Program, 2022, p. 37). Major industrialized countries, i.e., Germany, have failed and continue to fail to reach their building sector emission targets (Umweltbundesamt, 2023). The EU Commission has made closing the gap between the 2050 pathway and the status quo in the EU one of their key targets (Directorate-General for Climate Action, 2019, p. 6 & p. 9.). The latest policy changes reflect the importance of transforming the building sector. The EU policy, Directive (EU) 2023/1791, requires 3% of the total area of all publicly owned buildings to be renovated each year. National laws such as the Gebäudenergiegesetz (GEG) in Germany define specific types of energy sources and thermal transmittance values for building components when renovating existing buildings. At the same time, the question of climate change is accompanied by the economic demand and social need for appropriate residential real estate for residents in terms of quantity and quality. In Germany, this has

DOI: https://doi.org/10.5282/jums/v10i2pp522-560

led to a federal initiative focusing on affordable housing with the goal of building 400 000 new residential buildings per year (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2022, p. 4). This goal starkly contrasts the reality of only 306 000, 293 000, and 295 000 new buildings built in 2020, 2021, and 2022, respectively (Statistisches Bundesamt, 2023c).

The need for a radical transformation of the building sector and the differences between theoretical demand for and actual supply of living space warrant an in-depth analysis of the current state of residential real estate economics. It seems crucial to understand the regional implications of the energy efficiency policies passed by the EU institutions for the building sector in Germany, the largest EU member state by population and size of the economy (Eurostat, 2023b, 2023c). Evidence is needed to guide the discussions around energy efficiency of buildings on a political level and to understand the incentive structures for all building sector stakeholders (i.e., building owners, tenants, and industry service providers). Both can be achieved by looking at the energy performance certificates (EPCs) of residential buildings. EPCs were first introduced by a key EU policy targeting energy efficiency of buildings, the Energy Performance of Buildings Directive (EPBD) 2002/91/EC. The question arises whether differences in energy efficiency presented in these EPCs impact the rent and sales prices of buildings. This could be caused by the capitalization of energy savings or changes in building specific risk. Past research has provided evidence that differences in energy efficiency of buildings are directly correlated with differences in their rents and sales prices (see Brounen and Kok, 2011; Cajias et al., 2019; Deller, 2022; Högberg, 2013; Hyland et al., 2013). These studies show regional differences across different EU countries for rental and sales price premiums. For the German market, research has shown rental premiums between 0.9% and 5.8% and sales price premiums between and 5.0% and 6.8% when comparing the most efficient buildings to the average building stock (Kholodilin et al., 2017, p. 3231; Cajias et al., 2019, p. 183; Deller, 2022, p. 802). While this research provided first evidence on the topic, distinctive implementation details of the EU policy in Germany have not been considered. This particularly concerns the EPC type used by the building owner. The analysis presented in this paper is one of the first to consider the differences in energy efficiency premiums based on the EPC type used by the building owner. Hedonic price models are specified to analyze a rent data sample and a sales data sample with observations in the Rhein-Main Region in Germany. The results provide evidence for general energy efficiency premiums for the rental and sales market of up to 7.0% and 6.9% respectively. Additionally, evidence is presented that shows significant differences between the used EPC types. While the consumption certificate more accurately encompasses operational costs for buildings in the rental market, the requirement certificate is the one trusted by prospective buyers and crucial in determining sales prices. Controlling for the EPC types increases the premiums of the most efficient buildings with a requirement certificate to 8.8% and 7.4% for the rental and sales market, respectively. The findings of this paper have implications for policy makers and other stakeholders and recommend a revision or at least a re-evaluation of the EPC policy in the German market. Further, it becomes clear that the costs of transforming the building sector must be shared between asset owners, tenants and regulators. While owners might have to accept lower profitability, tenants must support modernizations via increased rents. Regulators need to support the sector transformation with non-financial processual adjustments. Whether financial subsidies by regulators are needed, too, is beyond the scope of this analysis. But evidence for this exists in the literature (Groh et al., 2022, pp. 105-107).

The remainder of this paper is structured as follows: In section 2, the extant literature is reviewed, and the hypotheses of this paper are derived. The review includes normative residential real estate valuation theory, EU and German policy and empirical literature. Next, in section 3, the methodology, the sample statistics, and the model specifications are presented. Section 4 reports the empirical results. They are subsequently discussed in section 5. In the final section 6, a conclusion and outlook on future research opportunities are given.

2. Review of the extant literature

The purpose of this paper is to provide relevant insights based on a rigorous quantitative analysis that can help policymakers and private and institutional investors make informed decisions when it comes to the economic meaning of the energy efficiency of residential real estate. To enable readers from a non-real estate background to better understand the results and implications of the analysis, a short introduction to the characteristics and economics of real estate as an asset class is given. Throughout the paper it will be referred to this normative theory.

2.1. Theoretical background & basic concepts

Real estate belongs to the field of the alternative asset classes. The variety of available investment opportunities within this particular asset class and across other alternative asset classes is vast. Further, the capital invested in real estate is significant. Estimates suggest that fifty percent of global wealth is invested into real estate (Baum & Hartzell, 2021, p. 4). The actual value of real estate as an asset class remains unclear. Overall, the global real estate market is made up of different national and regional markets shaped by their own regulations and characteristics. At the same time, there are characteristics and valuation methodologies regarding real estate as an asset class that remain valid across markets. The core characteristics of real estate are summarized below (Baum & Hartzell, 2021, p. 12–26):

- Depreciation: Properties are real assets that are affected by physical deterioration over time, leading to depreciation of their overall value (Baum & Hartzell, 2021, p. 12).

- Lease and cash flows: The lease agreement of a property is the main determinant of the cash flow generated by the asset over time (Baum & Hartzell, 2021, p. 13).

- Inelasticity of supply: Rise in demand is met by the supply side with a significant time lag that is caused by the long processes of acquiring permits and the subsequent construction of properties (Baum & Hartzell, 2021, p. 13).

- Valuation and investment performance: The appraisal value of a property sets the anchor for any future transactions and results in the issue of valuation smoothing (Baum & Hartzell, 2021, p. 14–15).

- Illiquidity of properties: High transaction costs, long sales processes and large bid-offer spreads are causes for the illiquidity of properties (Baum & Hartzell, 2021, p. 15).

- Asset specific risks: High capital values of properties create specific risks and make reducing risk to the systematic level difficult (Baum & Hartzell, 2021, p. 16–17).

- Leverage: Most investments are accompanied by loans against the property as collateral, resulting in a different risk–return profile (Baum & Hartzell, 2021, p. 18–19).

- Inflation correlation: Evidence in form of strong correlation exists that shows that properties might be an inflation hedge in the long-run (Baum & Hartzell, 2021, p. 19–21).

- Medium risk-return profile: Historical values suggest that while risk appears low, existing illiquidity and income uncertainty of properties result in an overall medium risk-return profile (Baum & Hartzell, 2021, p. 21–22).

- Return impact of real estate cycles: Inelasticity of supply and the caused lag in adaptive behavior leads to cyclical returns of investments (Baum & Hartzell, 2021, p. 22–24).

- Diversification impact of properties: Based on modern portfolio theory, real estate can be used for diversification because it shows low correlation with returns on equities and bonds (Baum & Hartzell, 2021, p. 24–26).

The amount and conditions of leverage used, the potential tax benefits attributed to a property, or the diversifying impact of the property on an investor's portfolio are crucial aspects when determining how much an investor is willing to pay for a property (Baum & Hartzell, 2021, p. 145–156). However, before these factors come into play, an appraisal of a property based on the status quo of market characteristics is calculated. The valuation methodology used for this is explained below. Tax benefits, the impact of leverage or the strategic relevance of an investment are out of scope for the present paper.

Within the German market, the official appraisal process of a property is regulated by the Immobilienwertermittlungsverordnung (ImmoWertV). This act is the basis for an objective valuation of a property and its methodologies are used by surveyors for calculating the valuation of a property for a forced sale. Additional information can be found in the "Muster- Anwendungshinweise zur Immobilienwertermittlungsverordnung," a set of instructions on how to implement the ImmoWertV (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023). Each property is an individual case and object-specific characteristics make each appraisal different. This gives some leeway to the individual performing an appraisal.

The appraisal methodologies described in the ImmoWertV are the "Vergleichswertverfahren" (§§ 24-26 ImmoWertV), the "Ertragswertverfahren" (§§ 27-34 ImmoWertV) and the "Sachwertverfahren" (§§ 35-39 ImmoWertV). All methodologies follow the three-step process that is shown in Figure 1 (§ 6 ImmoWertV). First, a methodology is used to provide a preliminary appraisal result. Next, local market characteristics are assessed, and the appraisal is adjusted. This is done regardless of which appraisal methodology was chosen and is the same proceeding for all the methodologies. As a final step, property specific characteristics are valued and included in the appraisal. This can include, for example, rights of special use. The final appraisal value is based on one or several values of the appraisal methodologies.

Internationally, the International Valuation Standards Council (IVSC) is recognized as the leading institution regarding standards in property valuation. This is underlined by the great number of national valuation associations that are members of the IVSC (International Valuation Standards Council, 2023). This includes, for example, the British national valuation association called Royal Institute of Chartered Surveyors. The latest standard on property valuation was published by the IVSC in January 2022 (International Valuation Standards Council, 2022). It describes the same three valuation methods for properties that were presented above: market approach, income approach, cost approach (International Valuation Standards Council, 2022, pp. 33-53). Thus, this general comparability across markets is established. The remainder of the paper focuses on the proceedings described in the ImmoWertV as the regional market of the Rhein-Main Region is under its jurisdiction. The English translation of the methodologies is used to improve the flow of reading. The English terms refer to the implementation in the German ImmoWertV. The market approach refers to the "Vergleichswertverfahren," the income approach refers to the "Ertragswertverfahren" and the cost approach refers to the "Sachwertverfahren."

The market approach is a comparably simple valuation methodology that uses similar transactions in the recent past to value a property (§§ 24-26 ImmoWertV). An overview of the market approach methodology is given in Figure 2.

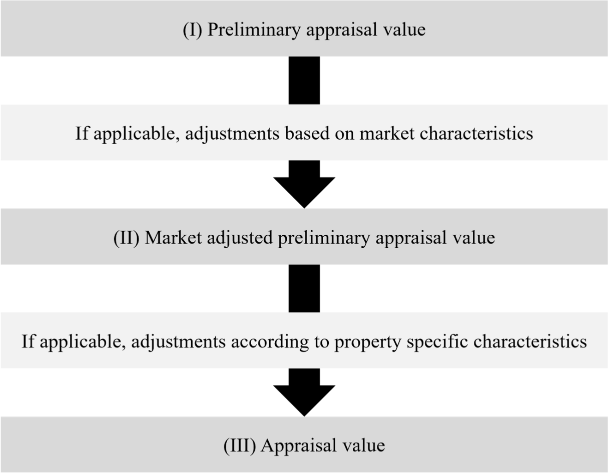

(I) Preliminary appraisal value
If applicable, adjustments based on market characteristics
(II) Market adjusted preliminary appraisal value
If applicable, adjustments according to property specific characteristics
(III) Appraisal value

Figure 1: Three-step appraisal process
(source: translated from Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen (2023, p. 14))

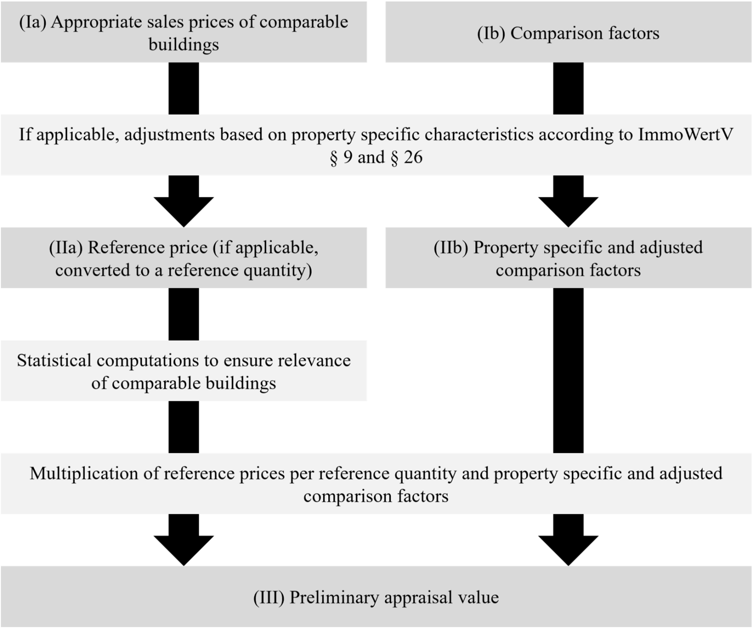

(Ia) Appropriate sales prices of comparable buildings
(Ib) Comparison factors
If applicable, adjustments based on property specific characteristics according to ImmoWertV § 9 and § 26
(IIa) Reference price (if applicable, converted to a reference quantity)
(IIb) Property specific and adjusted comparison factors
Statistical computations to ensure relevance of comparable buildings
Multiplication of reference prices per reference quantity and property specific and adjusted comparison factors
(III) Preliminary appraisal value

Figure 2: Market approach
(source: translated from Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen (2023, p. 28))

First, a statistically relevant number of transactions are identified. Next, benchmark values for relevant hedonic characteristics are calculated and comparison factors for the property to be valued are defined. These are multiplied and result in a preliminary appraisal value of the property. The preliminary appraisal value is equal to the one mentioned under (I) in Figure 1. It is then adjusted according to the process described above. The market approach is a more qualitative and simplified version of the statistical analysis performed in this paper. In its core, the idea used by the market approach and in this paper remains the same.

The income approach (§§ 27-34 ImmoWertV) is a version of a discounted cash flow (DCF) analysis. The DCF analysis is an internationally used standard valuation methodology for properties and other asset classes. Three different methods are defined for the income approach: a) general income approach b) simplified income approach c) periodic income approach. An overview of those three methods is given in Figure 3.

The preliminary appraisal value of the general income approach is calculated using the following equation (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023, p. 30):

$$pAV = (NOI - LV \times cr) \times CF + LV \tag{1}$$

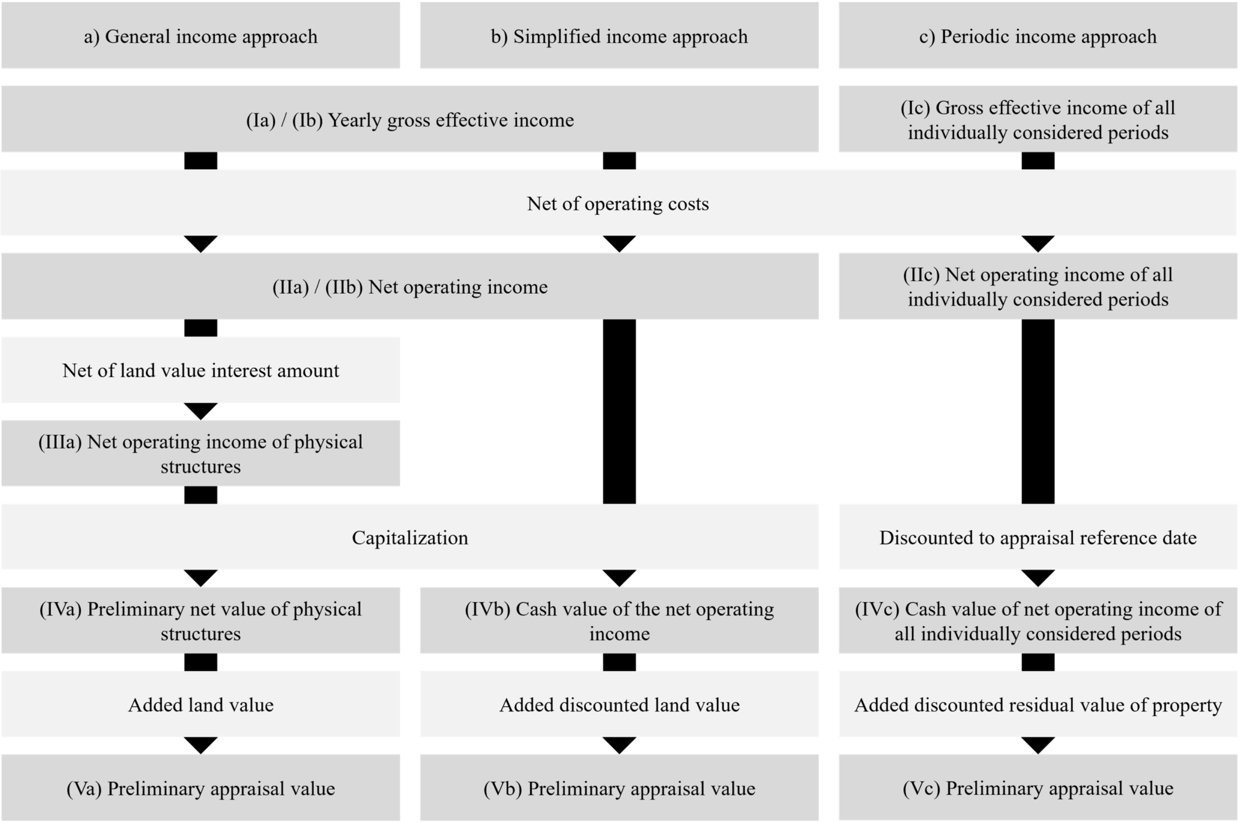

a) General income approach
b) Simplified income approach
c) Periodic income approach
(Ia) / (Ib) Yearly gross effective income
(Ic) Gross effective income of all individually considered periods
Net of operating costs
(IIa) / (IIb) Net operating income
(IIc) Net operating income of all individually considered periods
Net of land value interest amount
(IIIa) Net operating income of physical structures
Capitalization
Discounted to appraisal reference date
(IVa) Preliminary net value of physical structures
(IVb) Cash value of the net operating income
(IVc) Cash value of net operating income of all individually considered periods
Added land value
Added discounted land value
Added discounted residual value of property
(Va) Preliminary appraisal value
(Vb) Preliminary appraisal value
(Vc) Preliminary appraisal value

Figure 3: Income approach
(source: translated from Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen (2023, p. 30))

pAV stands for the preliminary appraisal value. NOI is the current yearly net operating income (NOI). *LV* stands for the land value and cr for the capitalization rate (cap rate). CF is the capitalization factor. To calculate the preliminary appraisal value, first, the NOI is calculated. This is done by estimating the yearly gross effective income (GEI) of the property based on current market values or the current lease. Next, operating expenses are deducted. This includes administrative costs, maintenance costs, risk of loss of rental income and running costs. Generally, the running costs are covered by the tenant in Germany following § 556 Bürgerliches Gesetzbuch (BGB). If this is the case, they are not deducted and not included in GEI received by the property owner. The GEI or in Germany the GEI plus the operating costs covered by the tenant is equal to the so-called warm rent. The NOI is equal to the so-called cold rent in Germany. Before multiplying the NOI with the capitalization factor, the land value interest amount is deducted to separate the land value and the income produced by the building. The land value interest amount is the land value multiplied with the cap rate. The land value is calculated based on defined standard land values for the location of the property. The cap rate is calculated iteratively based on past transactions in the local market. Each year, rating committees calculate official cap rates for local markets. One example is the Gutachterausschuss Frankfurt am Main that published the Immobilienmarktbericht 2023 (Debus, 2022). The calculations of the rating committee are based on § 21 ImmoWertV and readers are referred to this regulation for details. Following this, the multiplication with the capitalization factor results in the income value of the building of the considered property. The capitalization factor is calculated using the following equation (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023, p. 30):

$$CF = \frac{(1+cr)^n - 1}{(1+cr)^n \times cr} \tag{2}$$

The n in the equation above stands for the number of residual years of usage of the property. This value depends on the construction year, building type and modernizations done to improve the property. As a final step, the land value is added to the value of the property. The resulting preliminary appraisal value is adjusted using the steps of the process described in Figure 1.

The simplified income approach is applied in a similar way. The result is again the preliminary appraisal value that is subsequently adjusted. The meaning of the variables mentioned above stays the same. It is calculated using the following equation (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023, p. 31):

$$pAV = NOI \times CF + LV \times DF \tag{3}$$

DF stands for the discount factor applied to the land value. The discount factor is calculated as follows (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023, p. 31):

$$DF = (1 + cr)^{-n} \tag{4}$$

Compared to the general income approach, the only differences are that the land value interest amount is not deducted from the NOI and in turn the land value is multiplied with the discount factor when added to the appraisal value. Finally, the periodic income approach is calculated as follows (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023, p. 31):

$$\begin{aligned} pAV = {} & NOI_1 \times DF_1 + NOI_2 \times DF_2 \\ & + \ldots NOI_i \times DF_i + RV \times DF_b \end{aligned} \tag{5}$$

RV stands for the residual value of the property. The indexes indicate the specific period of the holding term considered. It starts with the first period and goes up until i, the index of the last period. Index b indicates the overall holding term. This variant of the income approach adds up the discounted future cashflows of the property and allows for changes in rental income and cap rate and thus capitalization factor. The length of the term can differ for each investment and depends on the number of periods considered. If the calculation does not include all periods until the final year of usage of a property, the residual value is added as a final value. The residual value is calculated using the simplified income approach. When considering long term projections, the estimations of these values become more difficult. When considering short term projections, the investment return strongly depends on the residual value of the property.

The final method considered here is the cost approach (§§ 35-39 ImmoWertV). The approach is summarized in Figure 4. The appraisal value is calculated by adding together the production costs of the usable main buildings, additional material assets and the land value. The production costs are equal to the calculated building costs needed in the current market conditions to construct a building that is comparable in kind and size to the property. Usually, these costs are based on industry modelling costs that are then multiplied with the respective reference units of the property. Additional adjustments are done based on the current price index published by the Statistisches Bundesamt in Germany. The average productions costs of the property are subsequently adjusted based on the age and the location of the property. This leads to the preliminary appraisal value mentioned as (I) in Figure 1. Next, this value is adjusted using the respective local market factors as was explained above.

Together, the different methods offer insights into the current appraisal value of the property. The final appraisal value is usually computed by taking the mean value of the different outcomes. In the context of this paper, the question that arises is how an increase or decrease in energy efficiency of a building would affect the outcome of the computation of the appraisal value when applying the presented methodologies. The following paragraphs describe the normative reasoning on why there should be energy efficiency premiums and discounts present in the residential real estate market of Germany. Of note, the market approach is not discussed because, as explained above, it represents a simplified version of the analysis in this paper.

When looking at the variants of the income approach, several variables in the equation could be affected by an increase or decrease in the energy efficiency of a building: gross effective income, net operating income, cap rate and residual years of usage. These variables represent the cashflows generated as well as the discount rate applied to these cash flows. The cash flows are considered first under the aspect of an increase in energy efficiency:

An increase in energy efficiency of a building leads to a decrease of energy usage and thus energy costs. The energy costs are the running costs of the building and are included in the operational expenses. In Germany, the operational expenses are mostly covered by the tenant and not the landlord. Thus, any investment into energy efficiency improvements is paid by the landlord and the decrease in operational expenses benefits the tenant. This is the so-called landlord-tenant dilemma and a key non-technical barrier to improving energy efficiency of real estate (Hirst and Brown, 1990, p. 276; Jaffe and Stavins, 1994, p. 805). Based on this argument alone, no changes to the net operating income of the building would occur. This, however, does not seem plausible in the case of a market environment. The landlord would try to recoup the investments and participate in the energy savings by increasing the cold rent for the tenant. This leads to a capitalization of the investment and an increase in the NOI for the owner. How much of the investment and how fast it can be capitalized remains a discussion topic in current research (see e.g., März et al., 2022, p. 20). It might depend on the kind of investment made and the local policy restrictions on rent increases. Often, energy efficiency improvements are prohibitively expensive compared to the achieved energy savings in monetary terms (März et al., 2022, p. 20). This leads to long recouping periods. Only recouping energy savings in the cold rent seems to make the investments unattractive. One additional option for landlords is to increase the cold rent more than the energy savings achieved by the improvements. This would lead to a disproportionate increase in cold rent and, subsequently, for the tenant to an increase in warm rent compared to the situation before the investment. In conclusion, an increase in energy efficiency would likely lead to an increase in the cold rent and warm rent of a building.

Besides the cash flows, the cap rate is crucial for the appraisal value calculation. As mentioned above, average cap rates for a local market are published each year. The cap rate is an indicative value for the sum of the property market risk premium, a location premium and a mean value for the asset

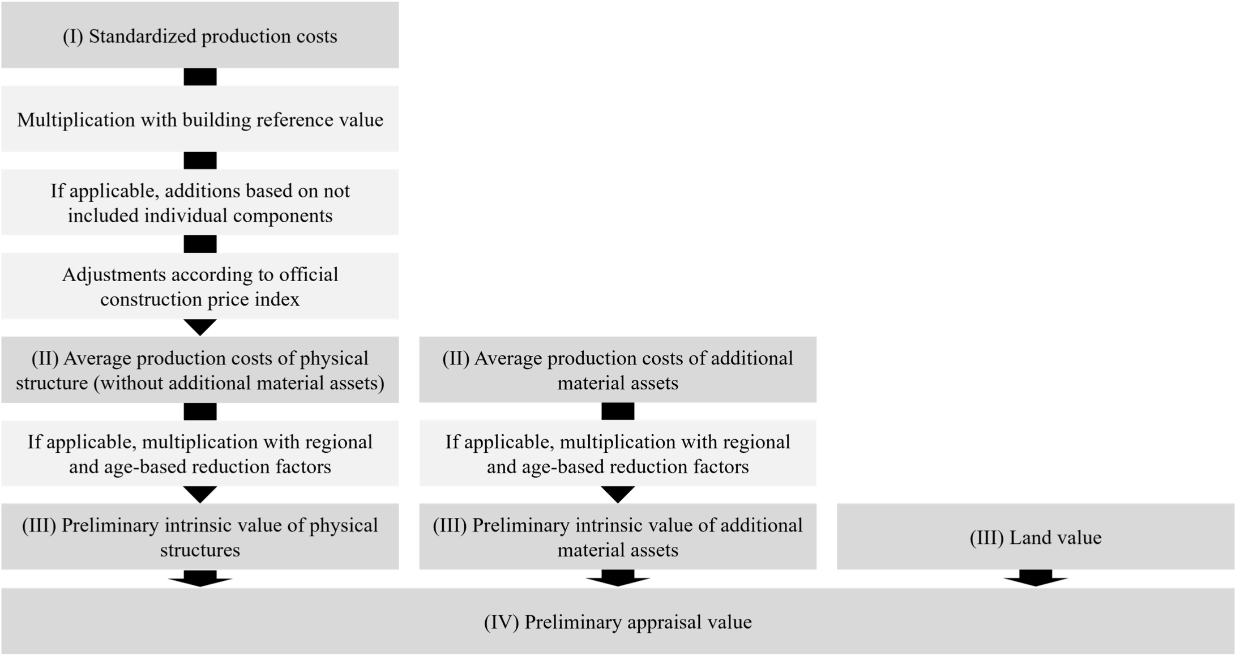

(I) Standardized production costs
Multiplication with building reference value
If applicable, additions based on not included individual components
Adjustments according to official construction price index
(II) Average production costs of physical structure (without additional material assets)
(II) Average production costs of additional material assets
If applicable, multiplication with regional and age-based reduction factors
If applicable, multiplication with regional and age-based reduction factors
(III) Preliminary intrinsic value of physical structures
(III) Preliminary intrinsic value of additional material assets
(III) Land value
(IV) Preliminary appraisal value

Figure 4: Cost approach
(source: translated from Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen (2023, p. 34))

specific premium (Baum & Hartzell, 2021, p. 153). Which exact cap rate is chosen depends on the decision made by the appraiser and is based on the property specific characteristics (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen, 2023, p. 33). The question arises how an increase in energy efficiency of a property would affect the risk associated with it. As argued by Deller (2022, p. 806) and underlined by the most recent public discussions regarding energy efficiency policies for buildings (Nieskes, 2023), high energy consumption could be associated with uncertainty and thus higher building specific risk. This leads to a decrease in risk when an increase in energy efficiency can be achieved for a property and thus a lower cap rate. A decrease in the cap rate leads to an increase in the appraisal value.

The last variable that is directly affected is the residual years of usage variable. An increase in energy efficiency can be achieved by modernizing a building. Such a modernization can range from, e.g., improving isolation to changing the heating system or windows. Each of these modernizations improves the energy efficiency of the building and influences the residual years of usage according to Appendix 2 of the ImmoWertV. Appendix 2 of the ImmoWertV provides a list of improvements and a table showing their impact on the residual years of usage. Based on the improvements made, an adjusted value for the residual years of usage is computed. As an example, the following impact could be achieved for an apartment in a multi-family apartment building that is 50 years old with 30 residual years of usage: If modernizations are performed that improve the isolation of the walls and the heating system is replaced, the residual years of usage increase to 37. Recalculating the cap factor of this building now results in 25.97 compared to the cap factor of 22.40 before. A 2% cap rate is applied. The increase in the cap factor results in an increase of 64 260 € in the appraisal value when using a 100 m^2 building and a cold rent of 15 € / m^2. The assumed cold rent is in line with current market prices for the city of Frankfurt a.M., the center of the Rhein-Main Region (Immowelt, 2023).

The other appraisal approach considered here is the cost approach. The appraisal value is based on the production costs of a comparable building in kind and size and adjusted based on age and current market conditions. The question is how an increase or decrease in energy efficiency would affect the appraisal value computed. As a first step, the construction costs are computed. The first preliminary production costs per m^2 to be used can be found in Appendix 4 of the ImmoWertV. They depend on the type and size of the building being constructed. A further differentiating factor is the standard of the building. The standard of the building is defined by the materials and technique used to construct it. A detailed description can be found in Appendix 4 of the ImmoWertV. When looking at specific elements like the outer walls or roof, it becomes clear that a higher level of energy efficiency is associated with a higher level of the standard of the building. The highest standard, level five, requires isolation of the walls and rooftop according to the Passivhaus-Standard for example. This impact of building components associated with energy efficiency on the standard of the building has a direct impact on the construction costs associated with it. For a multi-family home of up to six apartments, a construction cost of 825 € / m^2 is set for the standard level 3. A standard level 5 has a preliminary construction cost of 1190 € /

m^2. For a 100 m^2 apartment, this would change the appraisal value by 36 500 € without including considerations regarding the current construction price index or regional market adjustments. Evidently, it is not only energy efficiency that plays a role when it comes to determining the standard of the building. However, this quick estimation does show its relevance and impact on the appraisal value. Thus, buildings built with a higher level of energy efficiency should have a higher appraisal value when using the cost approach.

Based on the normative analysis above, there is a strong link between the energy efficiency of a building and its appraisal value. The question remains whether this strong link can be found in market prices, too. Further, the magnitude of this link is unclear. Lastly, the appraisal approaches outlined above might not correctly reflect the status quo of methods used by most market participants. For the market to reflect the importance of energy efficiency in market prices, an effective communication of energy efficiency of buildings is necessary. More importantly, differences in energy efficiency between buildings need to be clearly visible. How this communication is implemented is crucial and impacts the level of transparency for market participants. The implementation in the EU market is summarized below.

Depending on the region and type of real estate market, different certificates have been defined to communicate energy efficiency values. They range from certificates with a wide coverage of sustainability characteristics of a property (e.g., LEED, BREEAM) to others focusing primarily on the energy usage of a building (EU EPC) (BRE Group, 2023; Directorate-General for Energy, 2023; U.S. Green Building Council, 2023). When it comes to the residential real estate market in the European Union, the EPC was introduced as a mandatory certificate for communicating the energy efficiency of a building (Directorate-General for Energy, 2023). This purpose of this policy is to address the market failure of imperfect information.

The EPC was first introduced by the EU Energy Performance of Buildings Directive (EPBD) 2002/91/EC in 2002 and has since been replaced or amended multiple times. In 2010, the directive 2010/31/EU was approved, replacing the previous one. It included various changes such as making an EPC a requirement by law when leasing or selling a property and encouraging member states to provide financial incentives for energy efficiency improvements. In 2018 an amendment was approved: EU directive 2018/844/EU. It included improvements regarding the comparability of the calculation methodology used for EPCs and requires member states to formulate long-term renovation strategies. Another objective was to better align the EPBD with other directives such as the Energy Efficiency Directive and the Renewable Energy Directive. In December 2021, the EU Commission proposed a revision of the EPBD. The legislative procedure that can be found under "Procedure 2021/0426/COD" is currently in the stage of the trilogue negotiations (Dulian, 2023). It is likely to lay out compulsory 2030 targets for lowering energy consumption of the building sector overall. It will, however, not change the EPC framework and will not affect the quality, harmonization or accessibility of the EPCs.

As the EPCs on an EU level are regulated by a directive only, each member state has the responsibility to implement them on a national level (Directorate-General for Communication, 2023). The freedom in doing so limits the comparability of EPCs across member states. However, within one member state, the legislation is the same and in the context of this paper this is sufficient. In Germany, the EPC or in German "Energieausweis" is regulated by the Gebäudeenergiegesetz (GEG) in §§ 79-88.

The GEG lays the foundation for the EPC to work as a proxy in this paper. An EPC must be presented to any prospective tenant or buyer and information about the energy efficiency must be included in an online listing (§ 87 GEG). The EPC is valid for ten years if no building modernizations took place during that time (§ 79 GEG). It shows, e.g., the final energy consumption measured in kWh / (m^2 * a), the type of heating system and recommendations for possible energy efficiency improvements (§§ 84-85 GEG). The assessment and issuance of an EPC is only allowed to be performed by people with specific training and professional experience (§ 88 GEG). There are two different EPC types in Germany. One states the final energy consumption in the EPC that is computed using the energy consumption over the last 36 months. This EPC type is called "Verbrauchsausweis" (§ 82 GEG). This is translated to "consumption certificate" and used as such in the remainder of this paper. The second type states the final energy consumption that is computed using the theoretical consumption needs of the building based on material and construction information. It is called "Bedarfsausweis" (§ 81 GEG). This is translated to "requirement certificate" and is used as such in the remainder of this paper. Not every building can be issued either certificate. Certain property characteristics must be fulfilled to be issued a consumption certificate. These characteristics include the construction year, its coherence with energy efficiency laws introduced in 1977 and energy consumption data availability(§ 80 GEG). The different cases are presented in Appendix 1.

The calculations needed for the consumption certificate are easier and thus, this certificate is cheaper to acquire than a requirement certificate (Verbraucherzentrale NRW e.V., 2023b). Critics state that the consumption certificate lacks detailed information on the actual energy efficiency quality of the building and heavily depends on the behavior of the tenants in the past three years (Verbraucherzentrale NRW e.V., 2023b). The difference between theoretical and actual energy consumption has led to the terms of the "prebound" and "rebound" effect (Galvin, 2023, p. 502). The prebound effect describes the phenomenon where the actual energy consumption of an energy inefficient building is on average much lower than its theoretical energy consumption stated in the requirement certificate (Galvin, 2023, p. 502). The rebound effect describes the phenomenon where the actual energy consumption of a highly energy efficient building is on average higher than its theoretical energy consumption stated in the requirement certificate (Galvin, 2023, p. 502).

Independent of the limited comparability across EPC types, reliability and accuracy of the results presented can be assumed across one EPC type because only trained individuals with government issued certifications are allowed to offer certification services and can be held accountable if falsified or incorrect data were used (§ 83 GEG; § 88 GEG). The requirement of including information on the energy efficiency of a building on listings has improved data availability and market transparency. This data has been analyzed by academics since the EU-wide introduction of the EPC to provide evidence for the evaluation of the effectiveness of the policy and its impact on building economics. In the next subsection, the relevant empirical literature in the context of this paper is presented.

2.2. Empirical literature

The field of empirical literature analyzing the connection between sustainability characteristics such as energy efficiency and property valuation is vast. One approach to categorizing the literature is to sort the publications by some of the building sector characteristics that were mentioned above. This leads to the following classification criteria:

I. Year of publication

II. Region of real estate market (i.e., Americas, Europe, Asia, etc.)

III. Type of real estate market (i.e., industrial, commercial, residential)

IV. Type of certificate (i.e., LEED, BREEAM, EPC)

V. Type of transaction (i.e., rent or sale)

VI. Type of property (i.e., apartment, semi-detached, detached)

VII. Methodology applied (i.e., qualitative vs. quantitative)

VIII. Data sample analyzed (i.e., data sources, time span of data)

The first publications in this field of literature regarding residential real estate were performed using data from the USA but were limited by computing power and data availability. Those include for example Dinan and Miranowski (1989), Johnson and Kaserman (1983), and Nevin and Watson (1998). They represent some of the first studies analyzing the impact of sustainability characteristics such as energy efficiency on the valuation of properties. All find evidence that savings on energy costs are capitalized into the value of properties. Since then, data availability and access to computing power have considerably increased, making it possible to analyze different combinations of the above-mentioned classification criteria.

Meta-analyses of literature findings exist but remain limited. One reason why they remain limited could be the notable heterogeneity across markets. Relevant reviews include Ankamah-Yeboah and Rehdanz (2014), Brown and Watkins (2016), Cespedes-Lopez et al. (2019), Dalton and Fuerst (2018), Fizaine et al. (2018), and Kim et al. (2016). Of note, only the reviews by Cespedes-Lopez et al. (2019), Dalton and Fuerst (2018), and Fizaine et al. (2018) were published. The two other reviews are a conference paper and a working paper. Overall, the authors of the literature reviews agree that significant energy efficiency premiums exist in the sales and rental market. However, Dalton and Fuerst state that the values found for the confidence interval of the EPC include zero and thus they cannot state that there exists a significant premium for energy efficiency (Dalton & Fuerst, 2018). They argue that this might be caused by a strong heterogeneity in different EU markets. Further, caution is raised regarding the generalizability of the findings (Cespedes-Lopez et al., 2019, p. 54). The authors argue that findings in literature should only be considered having the respective analyzed market characteristics in mind. The meta-analyses focused on energy efficiency premiums in the sales market. They found average premiums of a magnitude between 3.5% to 7.6% (Fizaine et al., 2018, p. 1017 ; Ankamah-Yeboah and Rehdanz, 2014, p. 20). Dalton and Fuerst (2018, p. 18) also found evidence for the rental market with an overall premium of 8.2%. The lowest value of the sales price premiums was found by Fizaine et al. (2018). They raise the issue of publication bias and argue that considering this bias reduces the magnitude of the energy efficiency premiums on the valuation of buildings from 8% to 3.5-4.5% (Fizaine et al., 2018, p. 1013). Besides evidence for a publication bias, the usability of insights from older studies regarding the current magnitude of premiums or discounts might be limited. The reasons for that are changes in property valuation methods that now explicitly include sustainability characteristics and public and institutional investor green awareness that might influence their willingness to pay. The empirical studies presented help to understand an overarching trend but should not be considered as evidence of causality across markets or outside of the context of their study characteristics (Cespedes-Lopez et al., 2019, p. 54). This includes the analysis presented in this paper that can be categorized using the characteristics from above as follows:

I. Year of publication: not applicable

II. Region of real estate market: Rhein-Main Region, Germany

III. Type of real estate market: residential

IV. Type of certificate: EPC according to GEG in Germany

V. Type of transaction: rent and sale

VI. Type of property: no constraints but included as control variable

VII. Methodology applied: hedonic price model (multivariate regression analysis)

VIII. Data sample analyzed: various online multiple listing sources; 01/2015-06/2023

Keeping the remarks concerning generalizability of results and the specifications of this paper in mind, the remainder of the empirical literature review focuses on similar analyses of the residential real estate market of the EU to increase the comparability and relevance of the findings. An additional constraint for papers to be included is the use of the EPC as a proxy indicator. As mentioned above, it is the widely used certificate for residential buildings across the EU. The earliest study found matching the search criteria was written by Brounen and Kok (2011). The authors analyzed whether an overall energy efficiency premium can be detected in the Dutch market. Over the years, studies have included more complex relationships in their models such as interacting variables that can have a mediating impact on the studied effects. These variables include e.g., signaling effects, purchasing power and environmental awareness (see e.g., Fuerst, Oikarinen, and Harjunen, 2016; Pommeranz and Steininger, 2021). Further, economic theory on scarcity, willingness to pay and other factors has been employed to explain the findings of and understand the mechanisms behind the hedonic price models (Geske, 2022, p. 5-8).

Brounen and Kok (2011) used sales transaction data from the Netherlands and available EPC certifications. They reported an overall premium of 3.7% for a building with an A, B or C label compared to all other labels (Brounen & Kok, 2011, p. 175). Additionally, when analyzing each level individually and comparing it to the D label, they found stepwise premiums ranging from +10.2% for an A rated building to discounts of up to -5.1% for a G rated building (Brounen & Kok, 2011, p. 175). The findings were the first for the EU market and were robust when including more thermal and quality characteristics in the model (Brounen & Kok, 2011, p. 177).

In subsequent years, other analyses were published. The EU Commission initiated their own policy assessment in selected real estate markets that included Austria, Belgium, France and the UK (Bio Intelligence Service et al., 2013). Significant price premiums were found for more energy efficient buildings across all markets except for the regional market of Oxford, UK (Bio Intelligence Service et al., 2013, p. 12). In this market, the sample size and available explanatory variables were insufficient, limiting the explanatory power of this result (Bio Intelligence Service et al., 2013, p. 12). The problem of omitted variables cannot be excluded for the other models either as the age of the building for example was not included in all of them (Bio Intelligence Service et al., 2013, pp. 61-63). Thus, i.e., strong effects of 8% increases in sales prices per one-letter improvement in energy efficiency in Austria should be considered with caution. The study also found evidence of energy efficiency premiums in the rental market ranging from 4.4% in Austria to 1.4% in Ireland per one-letter improvement (Bio Intelligence Service et al., 2013, pp. 12-13). It is pointed out that a difference in magnitude between rural and city areas exists. This can be caused by the relative size of the energy savings compared to the € /m^2 costs of a building. This finding is supported by Hyland et al. (2013, pp. 948-949) who detected stronger effects for rural buildings compared to buildings in bigger cities in Ireland. Besides overall price premiums of 9.3% for sales prices for buildings with an A label compared to a D label and rental premiums of 1.8% for A rated buildings compared to D rated buildings, they identified that when market conditions are worse, the effect of energy efficiency on prices increases (Hyland et al., 2013, pp. 948-949). One reason could be that a broader supply and limited demand allows for more price differentiation in the market. Published in the same year, Högberg (2013, p. 256) showed for the market in Stockholm, Sweden, that a one percent decrease in energy consumption in kWh leads to an increase in sales prices of 0.04%, while the recommendation for specific energy efficiency improvements for a building stated in the EPC reduce sales prices by 2.4%. This shows that the potential need for retrofitting is seen as a hustle and decreases sales prices in the Stockholm market. Other studies across the EU also found significant energy efficiency premiums ranging from 11.3% for A or B rated buildings when compared to D rated buildings in Wales (Fuerst, McAllister, et al., 2016, p. 26) to 9.8% in Spain (de Ayala et al., 2016, pp. 21-22). At the same time, inefficient and G rated buildings are discounted in the market. Fuerst, McAllister, et al. (2016, p. 26) found discounts of -7.17% in Wales and Jensen et al. (2016, pp. 233-234) identified discounts of up to -24.3% in the Danish market.

However, not all evidence found shows positive price premiums in the EU market. Wahlström (2016, p. 197) analyzed single-family buildings in the Swedish residential sector and could not find evidence of price premiums for lower levels of energy consumption. However, Wahlström (2016, p. 204) did find that buyers are willing to pay for specific building characteristics that reduce energy usage. It seems questionable whether a potential multicollinearity problem could exist between these explanatory variables. The author, however, states that this is not the case (Wahlström, 2016, p. 201). Fregonara et al. (2017, p. 165) found no premiums for the Italian market when looking at transaction data in Turin. The sample, however, does not seem sufficiently large and the lack of findings could be caused by missing comparable transactions (Fregonara et al., 2017, pp. 156–158). For the category of A rated buildings, there exists only one observation and for B rated buildings only four (Fregonara et al., 2017, pp. 156–158). Marmolejo-Duarte and Chen (2022, p. 11) found evidence that when including a set of control variables regarding architectural quality, the effect of energy efficiency disappears. The sample used in their study shows a highly skewed distribution of energy efficiency levels with around half of all observations ranked as E (Marmolejo-Duarte & Chen, 2022, p. 9). Overall, the publications questioning the existence of energy efficiency premiums remain limited so far.

When it comes to the German residential real estate market, the first analysis was published by Cajias and Piazolo (2013). They found evidence for an existing positive price premium regarding total return of investments, rents and sales prices for residential buildings after controlling for regional, geographical and building-specific factors using hedonic models and data from 2008-2010 (Cajias & Piazolo, 2013, p. 57). A one percent increase in energy usage decreases on average the total return by -0.015%, the rent by -0.08% and the market value by -0.45% (Cajias & Piazolo, 2013, p. 53). Further, they recommend an asymmetric treatment for analyses of energy efficiency in general (Cajias & Piazolo, 2013, p. 67). When it comes to the generalizability of this paper, some limitations must be mentioned: The data used were collected right after the housing crisis that reached a climax in 2008 and completely dried up funding, strongly changing the financing conditions (Baum & Hartzell, 2021, pp. 60–62). The sample size used for analysis was small with 2630 building observations (Cajias & Piazolo, 2013, p. 57). The energy efficiency categories used were based on the Swiss Norm SIA 2031 and not the German EPC (Cajias & Piazolo, 2013, p. 58). Additionally, the maximum values for the categories have changed and buildings would be categorized differently today. Thus, the paper provided first evidence, but its values should not be generalized to the German market.

In the following years, more papers focusing on the German residential market were published. Kholodilin et al. (2017, p. 3224) analyzed the Berlin market with data from June 2011 to December 2014 that they collected from German multiple listing websites. They found energy efficiency improvements for the sales and rental market. Each additional kWh / (m^2 * a) needed decreases the sales price by -0.05% and the rent by -0.02% (Kholodilin et al., 2017, p. 3231). They further provided evidence on the landlord-tenant dilemma by showing that the energy savings are capitalized well in sales prices but exceed tenants' willingness to pay by a factor of 2.5 (Kholodilin et al., 2017, p. 3232). This seems reasonable because of the strong tenant rights existing in the German real estate market. Overall, the values found seem plausible and are in line with evidence from other markets across Europe.

The capitalization of energy savings in the rental market was further investigated by Cajias et al. (2019, p. 177) using a sample of almost 760 000 observations across all of Germany. They found evidence of energy efficiency premiums in the rental market of 0.9% for A+ rated buildings and discounts of up to -0.5% for H rated buildings when compared to the reference category D (Cajias et al., 2019, pp. 186–187). These premiums differ strongly when comparing secondary markets to the metropolitan regions. In secondary markets green premiums increase to 2.3% for A+ rated buildings and discounts increase to -1.8% for H rated buildings, respectively (Cajias et al., 2019, pp. 186–187). For the top markets, the results were mixed with no clear indications. This might have been caused by the high demand and inelastic supply in these regions (Cajias et al., 2019, p. 186). The reason for these differences remains unclear.

März et al. (2022, pp. 17–18) also provided evidence that energy efficiency premiums for rental apartments exist, but that they differ based on market conditions. This is even the case on a neighborhood level within a city (März et al., 2022, p. 18). Further, they showed that needed investments in energy efficiency improvements are currently not reasonable from a landlord perspective with payback periods of up to 100 years when only considering increases in rent (März et al., 2022, p. 20). However, they measured the effects using a linearly coded explanatory variable for energy efficiency rather than EPC categories (März et al., 2022, p. 14).

The question of whether energy efficiency premiums are big enough to incentivize investments was further analyzed by Groh et al. (2022, p. 95) for the German residential market. The positive energy efficiency premiums that they identified in the German rental market (+3.98% for A+ rated buildings compared to G and H rated buildings) are by far not sufficient to provide enough benefits for investors to accept the investment costs (Groh et al., 2022, pp. 104–107). Even when government subsidies of up to 45% were considered and a potential CO_2 tax split between the landlord and the tenant was included in calculations, the marginal benefits remained below marginal costs for retrofitting the average building (Groh et al., 2022, pp. 104–107). They argue that owner-occupiers have greater benefits and can more easily achieve economically reasonable energy efficiency improvements (Groh et al., 2022, p. 109). The analysis, however, only focused on increases in rent and not increases in building valuation. Such a consideration could change the assessment of the profitability of investments. Increases in building valuation because of higher energy efficiency have been shown by several analyses for the German market (see for example Cajias and Piazolo (2013), Deller (2022), and Kholodilin et al. (2017)). While Cajias and Piazolo (2013) and Kholodilin et al. (2017) used a continuous variable specification, Deller (2022, p. 815) made a categorical comparison. With a study scope similar to the one in this paper, Deller (2022, p. 817) identified premiums of 6.81% for A+ rated buildings and discounts of up to -8.8% for H rated buildings when compared to the reference category of D. The impact of an increase in property valuation on retrofitting profitability has since then been analyzed by Taruttis and Weber (2022). They used data from 2014 – 2018 on single-family homes across all over Germany (Taruttis & Weber, 2022, p. 1). The provided evidence is primarily valid for owner-occupied buildings, for which it seems to be more profitable when it comes to retrofitting (Groh et al., 2022, p. 109). Taruttis and Weber (2022, p. 6) provided evidence of significant energy efficiency premiums with a 100 kWh /(m^2 * a) decrease in energy consumption leading to a 6.9% increase in valuation. They further detected differences between rural and urban areas (Taruttis & Weber, 2022, p. 8). Rural areas experience a higher relative impact, but the absolute impact is comparatively lower when compared to urban buildings (Taruttis & Weber, 2022, p. 11). This is in line with evidence found for other European markets. Further, their study was one of the first to show that the EPC type of a building can have an impact on its valuation (Taruttis & Weber, 2022, pp. 8–9). While the authors considered the effect using subsamples, they did not consider interaction effects. Finally, they looked at investments in energy efficiency. They showed that energy savings are capitalized in building valuations and that the increases in valuation correspond to the capitalization of energy savings but that the investment costs are still higher (Taruttis & Weber, 2022, pp. 12–13). Several limitations regarding the retrofitting computations exist that should be kept in mind: Each analysis focuses on specific mean energy prices, construction costs and in some cases tax incentives and investor interest rates. All these variables are relatively volatile and could change significantly in the coming years, making new computations necessary. Taruttis and Weber (2022, p. 11) state that their computations strongly depend on the assumptions made.

Some studies have addressed the heterogeneity of energy efficiency premiums using more complex hedonic models with interaction effects. Pommeranz and Steininger (2021, p. 220) used German rental apartment data from Q1 2007 until Q1 2019 and identified overall energy efficiency premiums for rents. They analyzed the interaction effect of purchasing power as well as green awareness of inhabitants with these premiums. They found differences of 8.6% in rents when comparing the worst to the best level of energy efficiency. Of note, the threshold value for the EPCs is incorrect in this paper (Pommeranz & Steininger, 2021, p. 228). Since this error affected the descriptive statistics only, it does not impact the main conclusions of the paper: They found evidence that both factors drive the magnitude of energy efficiency premiums (Pommeranz & Steininger, 2021, p. 234). Purchasing power has a stronger effect and outweighs the effect of the green awareness of inhabitants (Pommeranz & Steininger, 2021, p. 234). Pommeranz and Steininger (2021, p. 239) acknowledge the heterogeneity of energy efficiency premiums and suggest further research on this topic to better understand the specific effects.

Galvin (2023) analyzed the topic of the prebound effect in the German residential sales market. Using data from semi-detached houses built before 1980 and sold between 2019 and 2021, the paper provides evidence that purchasers systematically overpay if they base their decision on theoretical energy savings as shown in the requirement certificate compared to actual energy savings (Galvin, 2023, p. 501). Galvin (2023, p. 511) states that the difference between theoretically needed consumption and actual consumption differs based on the energy efficiency level and presents an equation for an adjusted estimation. The results measuring the impact of the consumption certificate are seen as inconclusive as they depend too much on the behavior of the current owners or tenants (Galvin, 2023, p. 511). This seems unlikely as such a conclusion would render the consumption certificate unusable and warrants further analysis (this paper). Galvin (2023, p. 510) only considered the continuous values of energy consumption and did not analyze discrete levels of energy efficiency as shown in the German EPCs. The categorical analysis, however, is recommended to account for a non-linear functional form (Cespedes-Lopez et al., 2019, p. 53). Additionally, only effects on sales prices were analyzed. As is suggested in the paper, data on the rental market and the impact that can be identified in this market should be the subject of future research (Galvin, 2023, p. 505).

In conclusion, the extant empirical literature has demonstrated energy efficiency premiums exist in the German residential rental and sales market. These premiums experience strong heterogeneity associated with characteristics of the inhabitants, market conditions and the EPC type used. The analysis in this paper adds to these findings in three ways: First, by investigating general energy efficiency premiums using data that might reflect the most recent impacts of rises in energy prices and interest rates. Second and third, by addressing the identified gaps in the literature regarding the EPC type analysis for the rental and the sales market. Interaction effects between the EPC type and EPC rent premiums are investigated. The same is done for the EPC sales price premiums. The results show evidence for an important market within Germany while considering heterogeneity aspects that have not been analyzed in detail so far.

2.3. Hypotheses

In the two subsections above, the core concepts of real estate valuation and the current state of the empirical literature were discussed. It was shown that EU and German policy around energy efficiency was adjusted in the last years and that specific aspects regarding the heterogeneity of energy efficiency premiums need to be researched in more detail. These aspects help to define the different hypotheses for this paper. Following the normative approach of real estate valuation theory and evidence found by related literature, significant energy efficiency premiums should exist for the residential real estate market in the Rhein-Main Region of Germany. This leads to the following hypotheses:

> *Hypothesis 1 a): An increase in the energy efficiency of a residential building leads to an increase in its cold rent.*
>
> *Hypothesis 1 b): An increase in the energy efficiency of a residential building leads to an increase in its warm rent.*
>
> *Hypothesis 1 c): An increase in the energy efficiency of a residential building leads to an increase in its sales price.*

Higher energy efficiency results in energy cost savings. The decrease in energy costs is capitalized via increased cold rents. Additionally, signaling effects, prestige factors or the need to recoup investment costs might lead to an increase in warm rent overall. Including the increases in rent in the valuation of a building increases its sales price. Further, a reduction in property-specific risk that affects the cap rate and capitalization factor can lead to an additional value add and increase the sales price even further.

Looking at the heterogeneity of energy efficiency premiums identified by the empirical literature, a gap regarding the impact of the EPC type used by the seller or landlord was identified. First regional evidence regarding the German sales market exists (Galvin, 2023, p. 510), but so far, the rental market has not been analyzed. Further, more evidence for the sales market within the German real estate market needs to be provided. Evidence regarding the effects of the EPC type on the rental and sales markets will present valuable insights. The hypotheses that are to be tested regarding the EPC type used are the following:

> *Hypothesis 2 a): Based on the EPC type, an increase in the energy efficiency of a residential building leads to an increase in its cold rent.*
>
> *Hypothesis 2 b): Based on the EPC type, an increase in the energy efficiency of a residential building leads to an increase in its warm rent.*
>
> *Hypothesis 2 c): Based on the EPC type, an increase in the energy efficiency of a residential building leads to an increase in its sales price.*

On average, the requirement certificate and the consumption certificate do not show the same level of energy consumption for the same building (Verbraucherzentrale NRW e.V., 2023a). The impact of this discrepancy on the sales price of a building is currently unclear. The market should be efficient and include the difference in rents and overall building valuation. Market participants are likely to differentiate between the different EPC types and their economic meaning. Analyzing the market will help with understanding the structure of energy efficiency premiums in Germany and the impact of the two different certificates.

3. Methodology & data

In the section above, a normative reasoning for the existence of energy efficiency premiums in the Rhein-Main Region was developed and a gap in the literature was identified. The subsequently formulated hypotheses are tested in the remainder of this paper using a hedonic price model. The first theoretical foundations in the field of hedonic price models were developed by Lancaster (1966) and Rosen (1974). Since then, this methodology has been used in several studies analyzing energy efficiency premiums in real estate markets (see for example, Brounen and Kok, 2011; Deller, 2022; Hyland et al., 2013; Wahlström, 2016). Data used for the analysis in this paper are provided by the Real Estate Pilot AG (www.realestatepilot.com). They were collected from various digital sources. The dates of the observations range from 01/2015 – 06/2023. Next, the used methodology is introduced in more detail and the data generating process, descriptive statistics and model specifications are presented.

3.1. The hedonic price model

Compared to mass-produced products that are nearly identical with regards to function and form, real estate is very heterogeneous. Homogeneous products are traded in an explicit marketplace and their prices can be observed. The price must be paid to access specific characteristics. This observation of prices is not possible when it comes to the characteristics of heterogeneous real estate. They are traded on implicit markets, where the prices of the building-specific characteristics cannot be observed directly. To estimate the implicit prices of individual characteristics, a hedonic price model can be used. In its most basic form, it represents a multiple linear regression of sales prices or rents on real estate characteristics. The regression results, i.e., the coefficients of the independent variables, represent the estimated prices for the individual characteristics. Depending on the functional form, these can be absolute values or price elasticity values. This is the so called "first stage" hedonic model. Its results can be used to define "second stage" hedonic models that identify structural demand and supply parameters when certain assumptions are met. Mathematical theory includes the assumption of perfect elasticity for all characteristics. This is rarely the case in real-life settings and goes beyond the scope of the analysis in this paper. (Malpezzi, 2002, p. 68 - 71)

In today's real estate literature, hedonic price models have become a core research methodology. One could argue that the application today is in line with the first known applications of this methodology to estimate farmland values in Minnesota and Iowa (Haas, 1922; Wallace, 1926). Whether these applications, that took place before the theoretical developments of Lancaster (1966) and Rosen (1974), can be seen as hedonic models has since been discussed in the literature (Colwell & Dilmore, 1999, p. 620). Other early applications can be found in the automobile industry (Court, 1939; Griliches, 1961).

It is undisputed that Lancaster's work on consumer theory and Rosen's publication on hedonic prices and implicit markets paved the way for the theoretical development of this methodology. Lancaster (1966, p. 133) demonstrates how consumers maximize their utility based on product characteristics. He writes that "goods possess, or give rise to, multiple characteristics in fixed proportions and that it is these characteristics, not goods themselves, on which the consumer's preferences are exercised." (Lancaster, 1966, p. 154)

When combining the utility of the characteristics with their implicit prices, a market between buyers and sellers is established. The buyers receive utility from the characteristics of the product that they acquire. They are constrained by their budget. The sellers receive returns from specialized production of goods. The produced goods possess characteristics that are desired by the market. The clearing price for each product characteristic is determined by the distributions of consumer tastes and producer costs. These implicit clearing prices for product characteristics are identified with hedonic price models. (Rosen, 1974, p. 35 -36)

The theoretical foundation as well as empirical application of hedonic price models has since been extended through a variety of publications. An excellent general review of the literature on hedonic price models is given by Malpezzi (2002) and a thorough discussion on theoretical and econometric constraints can be found in Follain and Jimenez (1985) and Sheppard (1999). Further reading includes Bajari and Benkard (2005), Bartik (1987), Blomquist et al. (1988), Edlefsen (1981), Epple (1987), and Roback

(1982). Amemiya (1980), Hocking (1976), and Leamer (1978) comment on the selection process of independent variables.

As stated above, the statistical core of the hedonic price model is a multiple linear regression. In this paper, an ordinary least squares (OLS) estimator is used for estimating the coefficients of the independent variables. The functional form as well as the selection of model parameters are further elaborated on in the remainder of this section. When applying the hedonic price model, certain assumptions must be fulfilled (see Appendix 2). Model assumption tests were performed for all presented models and the results are discussed at the end of section 3.4.

3.2. Review of the data generating process

For quantitively testing the above presented hypotheses by applying a hedonic price model, micro-level data on a building level are required. Ease of data collection depends on the jurisdiction of the analyzed market. In some EU countries, public registries with detailed sales transaction data exist. Other countries offer sales transaction data but cannot provide detailed data on the characteristics of the real estate sold. In the remaining EU countries, there is no publicly available information on market transactions. When it comes to rent data, the challenge of collecting real transaction data is even greater. Within the German market, there is no governmental agency that offers transaction data on a micro-level. The house price index that is published by the Statistische Bundesamt is computed using micro-level sales data provided by the regional committees on real estate (Statistisches Bundesamt, 2018, p. 5-6). However, the raw data are not published. When it comes to rental data in Germany, official rent indices are computed using data from the "Mikrozensus," a yearly survey on working and living conditions in Germany (Statistisches Bundesamt, 2023b). The survey is answered by around 1% of the population in Germany and used for different analyses (Statistisches Bundesamt, 2023b). Because of the difficulty of accessing transaction data in Germany, most empirical literature use listing data collected from digital sources such as multiple listing service providers (see Deller, 2022; Kholodilin et al., 2017; März et al., 2022; Taruttis and Weber, 2022). This use of listing data comes with limitations as discussed by Kholodilin et al. (2017, pp. 3224–3225):

- Duplications of listings across the different digital and analogue listing platforms exist.
- Listings are used for marketing purposes by developers or construction companies.
- Owners / landlords might leave out information on energy efficiency on purpose.
- The final transaction price or rent paid differs from the values stated.

To limit the impact of these issues the raw datasets were prepared prior to analysis (see below).

The data analyzed in this paper are made up of two samples of listing data provided by the Real Estate Pilot AG. The observations for both samples were collected from multiple listing service providers from January 2015 until June 2023. More specifically, for listings to be included in the analysis, the first date of the listing had to be between the 1^{st} of January 2015 and the 30^{th} of June 2023. Data were updated once every day during that period. Of note, the Real Estate Pilot AG filters for and deletes duplicates, addressing one of the issues mentioned by Kholodilin et al. (2017, pp. 3224–3225). Further details regarding the data collection process are published in Deller (2022, p. 811). The cities and counties making up the market of the Rhein-Main Region are defined following theRegionalverband FrankfurtRheinMain (2022). The first sample comprises 917 213 micro-level observations of the rent market in the Rhein-Main Region. The second sample comprises 556 791 micro-level observations of the sales market in the Rhein-Main Region. The data includes information on various hedonic characteristics such as living space, energy consumption in kWh, year of construction, postal code and others.

All the data preparation, cleaning and analysis described in this paper were done using R (www.r-project.org). The R code used is available upon reasonable request. The data preparation was performed before starting with the analysis of the two data samples. This helps address several of the issues mentioned by Kholodilin et al. (2017, pp. 3224–3225). First, the attributes were selected and encoded. The attribute inclusion process was based on a hierarchical method of attribute selection that includes the most relevant explanatory attributes first. Further, the suggestions made by Malpezzi (2002, pp. 78–79) regarding hedonic characteristics were followed as best as possible with the available data sample attributes. To avoid large drops in sample size because of missing values, 19 hedonic characteristics were included in the final selection across both data samples. Appendix 3 shows an overview and a description of the included attributes. The encoding of attributes mainly concerned the binary control attributes that state for example whether a building is refurbished, comes with an elevator or is a landmarked building. Next, the observations with missing data were removed. Observations with complete information are necessary for the hedonic price model. In line with literature, trimming of values and not imputation was performed to manage outliers (see e.g., Taruttis and Weber, 2022, p. 4). This approach also seemed reasonable with regard to the issue of listings being a marketing tool for developers and construction companies. It is likely that they either did not fill out all the information of the 19 attributes or put placeholders in with values that are not plausible and could reliably be identified as erroneous. For example, a value of "9999" entered for energy efficiency is meaningless and must be removed. Trimming the samples helped to filter out such outliers or observations where specific values are placeholders. For the trimming of the data samples, metric independent and dependent attributes were used. Additionally, relative values such as living space per room or the ratio between operating costs to cold rent were computed. When it comes to the sales price sample, the sales price per m^2, the living space in m^2, the living space per room in m^2 and the energy consumption in kWh / (m^2 * a) were considered. The bottom and top half percentile of these absolute and relative metric attributes were computed and any observations exceeding those values were deleted. This is a more conservative approach than others have applied (see e.g., Taruttis and Weber, 2022, p. 4). The same was done with the rent sample. The metric attributes used were the cold rent per m^2, the operating costs in € , the ratio between operating costs to cold rent, the living space in m^2, the living space per room in m^2 and the energy consumption kWh / (m^2 * a). A final plausibility check was performed regarding the construction year of observations following Deller (2022, Appendix 15). To filter out advertisements, all observations with a construction year greater than 2023 were deleted. Observations with a construction year smaller than 1871 were deleted, too, to account for the beginning of the "Gründerzeit" in Germany. The start of the "Gründerzeit" in Germany lead to changes in construction technology. Further, deleting these observations helped to avoid the risk of unobserved refurbishment of the historical building stock as mentioned by Cajias et al. (2019, p. 184).

The two issues of consciously not including specific information and differences between actual transaction and listing values remain to be discussed. Not including specific information in an online listing is a very general limitation of empirical research using online platform data. For buildings, it can be argued that the information not included in the listing itself would have nonetheless been included in the valuation of the building. Thus, while making the usable sample smaller, it should, on average, not bias the results. When it comes to the issue of potential differences between listing and transaction values, two arguments can be made: First, past research has shown that listing prices can be a reasonable representation of transaction prices in a metropolitan region of Germany (Henger & Voigtländer, 2014, p. 15). This is true for time periods of a strong and upward moving real estate market. This was the case for most of the period considered in this paper. Only the past 18 months show the first effects of the rising interest rates and macroeconomic developments such as the energy crisis (Statistisches Bundesamt, 2023a). Second, it can be argued that an above-market listing price increases the time a building remains on the market (Knight, 2002, p. 213). This serves as an incentive for owners and real estate agents to price buildings in line with current market conditions.

The results of the described data generating process and the subsequent data preparation and cleaning were the final data samples used for the remainder of this paper. The rent data sample includes 212 167 observations. The sales data sample includes 159 573 observations. Next, the descriptive sample statistics of both are displayed. This is followed by the specification of the hedonic price models used to test the hypotheses of this paper.

3.3. Descriptive sample statistics

Table 1 gives an overview of the descriptive sample statistics for the rent price sample and Table 2 for the sales price sample. The variables include the dependent variables (i.e., cold rent, warm rent, sales price), the key explanatory variable (i.e., energy efficiency) and additional control variables (e.g., living space, construction year, EPC type). Additional control variables not shown in the tables of the descriptive sample statistics are the nominal variables "type of building," "postal code" and "upload date." The type of building is shortly discussed below, while the definition of postal code and upload date can be found in Appendix 3. All sample statistics of the metric variables are presented without the usage of log values. In the next subsection, log transformation is applied to account for heteroscedasticity of residuals.

The rent price sample is made up of 212 167 observations. The average building for rent in the Rhein-Main Region between the dates 01/2015 – 06/2023 was built in 1979, offers 2.74 rooms stretching across 78.76 m^2 and is equipped with a parking space. Its energy consumption is 119.90 kwh / (m^2 * a), while its cold rent is 855.00 € . Its warm rent amounts to 1 039. 40 € . The energy consumption is equal to the energy efficiency level D. The cold rent is equivalent to 10.86 € / m^2 and the warm rent is equivalent to 13.20 € / m^2. When it comes to the overall sample, 41% are issued a requirement certificate and 59% are issued a consumption certificate, 3% of all buildings are furnished, 28% are refurbished, 13% have not been lived in before, less than 0.5% are landmarked buildings and 26% are equipped with an elevator. The sample statistics show that the dependent variables of cold rent and warm rent and the control variable living space are slightly positively skewed. Their mean is greater than their median. The key explanatory variable energy efficiency is not skewed to a relevant degree.

The distribution of EPC levels of all the observations in the sample is shown in Figure 5. The EPC level with the greatest number of observations is D, which is in line with the mean of the energy consumption of the sample. The number of observations with energy efficiency level B, however, is surprisingly high. One explanation could be the subsidies in the past that were provided by the German federal government for building energy efficient homes or refurbishing existing buildings. Higher construction costs were potentially compensated for and thus the buildings were built with higher levels of energy efficiency. This explanation remains speculative. The distribution should be kept in mind when interpreting the empirical results of the hedonic price models later. One other interesting aspect is the distribution of the EPC type used. The consumption certificate is used more often than the requirement certificate. This is, however, not surprising as it is cheaper to obtain a consumption certificate.

The sales price sample is made up of 159 573 observations. The average building for sale in the Rhein-Main Region between the dates 01/2015 – 06/2023 was built in 1980, offers 4.73 rooms across 134.84 m^2. It is equipped with a parking space and has an energy consumption of 133.07 kWh / (m^2 * a). This is equivalent to an EPC rating of E. The

Table 1: Summary statistics of the rent data sample

Variable	Unit	Mean	P 25	Median	P 75	St. Dev.	Minimum	Maximum
Dependent variables								
Cold rent	Price in Euros (€)	855.00	550.00	740.00	1035.00	451.50	130.00	5700.00
Warm rent	Price in Euros (€)	1039.40	695.90	910.00	1250.00	506.34	198.80	6200.00
Building-specific independent variables								
Energy consumption	kwh / (m * annum)	119.90	76.60	118.00	154.00	56.86	4.40	334.80
Living space	m^2	78.76	56.00	73.07	95.00	33.15	19.53	230.82
Number of rooms	Numeric	2.74	2.00	3.00	3.00	1.09	1.00	12.00
Construction year	Numeric	1979	1963	1982	2003	30.89	1871	2023
EPC type	Binary; reference = requirement	0.59	0.00	1.00	1.00	0.49	0.00	1.00
Furnished	Binary; reference = 0; true = 1	0.03	0.00	0.00	0.00	0.17	0.00	1.00
Refurbished	Binary; reference = 0; true = 1	0.28	0.00	0.00	1.00	0.45	0.00	1.00
First occupancy	Binary; reference = 0; true = 1	0.13	0.00	0.00	0.00	0.33	0.00	1.00
Landmarked building	Binary; reference = 0; true = 1	0.00	0.00	0.00	0.00	0.02	0.00	1.00
Elevator	Binary; reference = 0; true = 1	0.26	0.00	0.00	1.00	0.44	0.00	1.00
Parking space	Binary; reference = 0; true = 1	0.50	0.00	1.00	1.00	0.50	0.00	1.00

Number of observations in the rent data sample: 212167

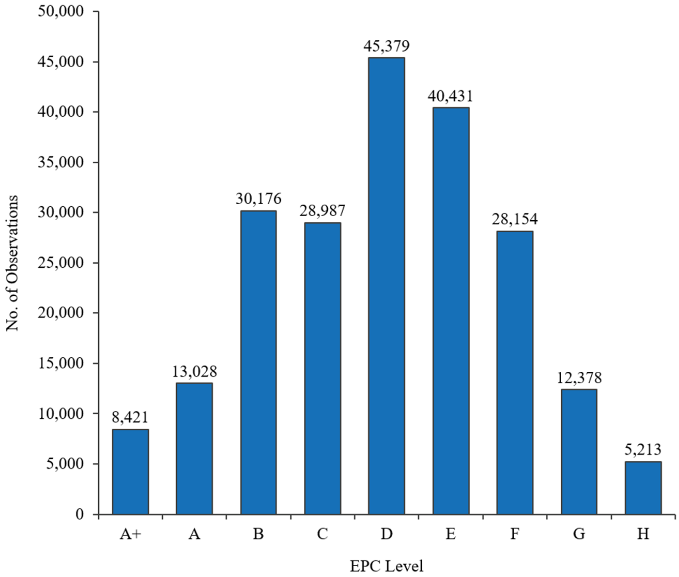


Figure 5: Distribution of EPC levels in the rent sample

sales price of the average building is 419 012.00 € , leading to a price of 3 107.48 € / m^2. Across the overall sample, 46% are issued a requirement certificate and 54% are issued a consumption certificate, 13% of all buildings are refurbished, 11% have not been lived in before, around 1% are landmarked buildings and 18% are equipped with an elevator. Additionally, 12% of all buildings have an active lease and 22% are sold without a commission fee for the buyer. When comparing the mean and median of the variables, the values of the sales price, the energy consumption, the living space and number of rooms point to a positively skewed distribution.

The EPC rating distribution of the key explanatory variable energy consumption is shown in Figure 6. The EPC rating D includes the greatest number of observations, followed by E. This is in line with the mean and median of the variable energy consumption. The mean is at the lower end of the EPC rating E and the median is at the upper end of the EPC rating D. Thus, using different cut off points could have resulted in a histogram that better reflects a normal distribution. The number of observations that have an EPC rating of A+ or H are greater than expected. This might be indicating that owners want to decrease risk of their real estate portfolios by selling very energy inefficient assets and developers that can increase profitability when increasing the energy efficiency of refurbishment or building projects. This again, however, remains speculative and is a topic beyond the scope of the analysis in this paper.

When comparing the rent data sample with the sales data sample, several differences can be noted and should be discussed: A building that is sold is on average bigger in floor size and has more rooms than a building available for rent. It is also less energy efficient and more likely than a building for rent to have a requirement certificate. The higher energy consumption is in line with the significantly lower percentage of buildings that are refurbished in the sales sample compared to the rent sample (13% vs. 28%). A building that is up for sale is also much less likely to have an elevator, but more likely to have a parking space. It is likely that an underlying data sample characteristic is the cause of these differences in distribution. Looking at the data, the most likely one is the building type. Apartments are more likely to be located in the bigger cities and city centers. At these locations, less space is available, making individual properties smaller. At the same time, apartments are part of large multi-family homes that are bigger than the ones in rural areas. This might lead to more buildings that need an elevator and can get a consumption certificate issued. The considerations are supported by the data: the most common building type in the rent data sample is an “apartment” while the most common building type in the sales data sample is a “detached single or dual family home.” A more detailed analysis focusing on the impact of energy efficiency based on specific building types would likely result in interesting new insights. This provides an opportunity for research in the future.

Before the specification of the hedonic price models, correlation between the different attributes needs to be assessed within both samples. High correlation between two explanatory variables in a linear model can lead to a decrease in significance for both and should be avoided. Appendix 4 shows the correlation matrix for the rent data sample. The correlation matrix for the sales data sample can be found in Appendix 5. When looking at the correlation of the rent data sample, the attributes “living space” and “number of rooms” show a high value of correlation (0.87). These two attributes show a similarly high value of correlation in the sales data sample (0.91). The question arises whether both attributes should be used as explanatory variables for the hedonic price models. This would be the case if they measured different effects. It can be argued that up to a certain size of a building

Table 2: Summary statistics of the sales data sample

Variable	Unit	Mean	P 25	Median	P 75	St. Dev.	Minimum	Maximum
Dependent variable								
Sales price	Price in Euros (€)	419012.00	219500.00	349000.00	533270.00	300919.80	21000.00	5350000.00
Building-specific independent variables								
Energy consumption	kwh / (m ^{2} annum)	133.07	79.80	123.00	170.00	77.26	0.01	465.39
Living area	m^2	134.84	80.90	119.00	165.00	75.95	26.52	569.00
Number of rooms	Numeric	4.73	3.00	4.00	6.00	2.65	1.00	30.00
Construction year	Numeric	1980	1966	1981	2000	29.22	1871	2023
EPC type	Binary; reference = requirement	0.54	0.00	1.00	1.00	0.50	0.00	1.00
Refurbished	Binary; reference = 0; true = 1	0.13	0.00	0.00	0.00	0.34	0.00	1.00
First occupancy	Binary; reference = 0; true = 1	0.11	0.00	0.00	0.00	0.30	0.00	1.00
Landmarked building	Binary; reference = 0; true = 1	0.01	0.00	0.00	0.00	0.07	0.00	1.00
Elevator	Binary; reference = 0; true = 1	0.18	0.00	0.00	0.00	0.39	0.00	1.00
Parking space	Binary; reference = 0; true = 1	0.66	0.00	1.00	1.00	0.47	0.00	1.00
Active lease	Binary; reference = 0; true = 1	0.12	0.00	0.00	0.00	0.32	0.00	1.00
Contract-specific independent variable								
Commission free	Binary; reference = 0; true = 1	0.22	0.00	0.00	0.00	0.41	0.00	1.00

Number of observations in the sales data sample: 159573

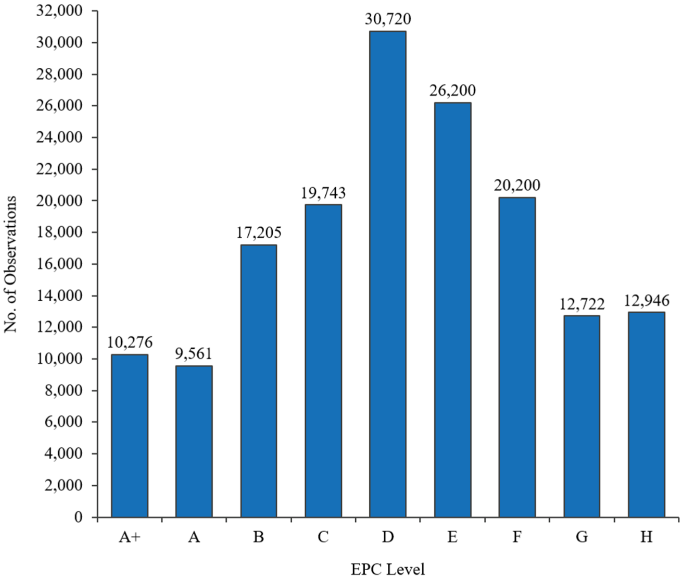


Figure 6: Distribution of EPC levels in the sales sample

an additional room is valued more than having bigger rooms overall. To measure this effect, it is necessary to use both variables for the hedonic price models. This argument alone seems rather weak. However, the interpretation of both attributes is not the purpose of this paper and, thus, it might be beneficial to account for more heterogeneity in the data. Including the attribute as an explanatory variable is also in line with the empirical literature and recommended (Malpezzi, 2002, p. 78). The key explanatory variable energy efficiency is not affected by this potential problem of correlation. The potential problem of multicollinearity is again checked when specifying the hedonic price models. Besides the correlation between these two variables, there is also moderate correlation found between the key explanatory variable of energy efficiency and construction year. This seems plausible as the building standards have increasingly required developers to improve the energy efficiency of newly built buildings. The correlation coefficient for the rent data is -0.53 and therefore smaller than the correlation of -0.63 that can be found in the sales data sample. This is in line with the higher rate of refurbished buildings that was identified above. The potential problem arising from the moderate correlation is checked during the specification of the hedonic price models.

3.4. Specification of the hedonic price models and testing of model assumptions

The two final samples from above are used to estimate the models that measure the impact of energy efficiency on the cold and warm rent and on the sales prices of residential real estate. Different models are specified for each of the dependent variables *"cold_rent," "warm_rent,"* and *"sales_price."* The cold rent models are used to test hypotheses 1 a) and 2 a). The warm rent models are used to test hypotheses 1 b) and 2 b). The sales price models are used to test hypotheses 1 c) and 2 c). The first model specified for all dependent variables is a basic hedonic price model that includes the key explanatory variable energy efficiency and the most relevant control variables. Next, a full categorical model that includes all control variables is specified. The term "categorical" refers to the way the key explanatory variable energy consumption is coded. In this model and the basic model, it is specified using the EPC rating of the observation. Robustness of those models is tested by exchanging the EPC rating with the continuous value of the energy consumption of the buildings. This model is the full continuous model. The fourth and final model includes all variables from the full categorical model, the EPC type used and an interaction term between the EPC type and the EPC rating. This is the full interaction model. All independent variables are either building, location, time or contract specific characteristics. As stated, a hierarchical method of variable selection is applied. The definitions and the indexes remain the same across all models. An overview of all definitions can be found in Appendix 6. The results of all models are presented in the next section. In total, twelve models are specified. All of the full categorical models are based on Deller (2022, pp. 815–816). The cold rent models are defined first. This is followed by the warm rent models. The section concludes by presenting the sales price models. The basic cold rent model is specified in the following way:

$$\begin{aligned}\ln(cold\_rent_{ilt}) &= \alpha + \beta_1 epc\_level_i + \beta_2 living\_space_i \\ &+ \beta_3 \ln(living\_space_i) + \beta_4 construction\_year_i \\ &+ \gamma_l + \varepsilon_{ilt}\end{aligned} \tag{6}$$

A log-linear functional form is chosen to account for heteroscedasticity (Malpezzi, 2002, p. 80). This is done by transforming the dependent variable *"cold_rent"* using the natural logarithm. The indexes *"i," "l"* and *"t"* stand for the observation *"i,"* the location *"l"* and the time *"t."* The basic cold rent model includes the intercept *"α"*, the independent variables and the error term *"ε."* The first independent variable is the *"epc_level."* It is coded as a categorical variable ranging from A+ to H. These categories are based on the definition of the EPC ratings in Appendix 10 of the GEG. Including the energy efficiency of a building as a categorical variable and not a continuous variable is done to account for potential non-linearity. It is also recommended in the literature (Cespedes-Lopez et al., 2019, p. 53). Additionally, the categories ranging from A+ to H are well known and often used to communicate the energy efficiency of a building. The impact of the letter grade rating might be more significant for the valuation of the energy efficiency of a building than the value of energy consumption measured as kWh / (m^2 * a). Nevertheless, the impact of the continuous variable is investigated in the third cold rent model as a robustness check. The reference value of *"epc_level"* is defined as the EPC rating D. This includes the general EPC rating and does not consider whether the EPC type used is a consumption certificate or a requirement certificate. Choosing the EPC level D as the reference value serves two goals: It helps to make a comparison with the average building in the data sample more intuitive and is in line with the recommendations made in literature (Cespedes-Lopez et al., 2019, p. 53). Next, the *"living_space"* of the building is included. This metric variable measures living space in m^2. The living space is included as a continuous variable that is not transformed and in a second variable that is transformed using the natural logarithm. Including both terms was done to account for non-linearity in the data and increase compliance with this assumption of the hedonic price model. This decision, however, increases multicollinearity between the two independent variables. As living space is not the key explanatory variable that is to be interpreted in this analysis, this is accepted. When interpreting the results, this must be kept in mind and thus the coefficients of the two variables that include the living space can only be interpreted with caution. The *"construction_year"* variable is added as a categorical variable that controls for the year a building was constructed using 10-year intervals. This follows literature such as Cajias et al. (2019, p. 184). The age of a building has a significant impact on its value and cold rent. It is the key reference point when accounting for the depreciation of a real estate asset. The final control variable included in the basic model is the categorical variable *"γ"* that controls for the location of the building on a postal code level. Location, too, is a key characteristic of a building and has a significant impact on its rent and sales price. Limitations exist regarding the location control using postal codes as they can stretch across micro-locations. However, using postal codes is the most appropriate measure as usage of streets as control variables would lead to a significant decrease in sample size.

In the remainder of this sub-section, only the additional or changed explanatory variables are elaborated on when presenting the specifications of the hedonic price models. All other variable definitions and indexes remain the same. The full categorical cold rent model that includes all control variables is specified in the following way:

$$\begin{aligned}\ln(cold\_rent_{ilt}) &= \alpha + \beta_1 epc\_level_i \\ &+ \beta_2 living\_space_i + \beta_3 \ln(living\_space_i) \\ &+ \beta_4 no\_rooms_i + \beta_5 furnished_i \\ &+ \beta_6 refurbished_i + \beta_7 first\_occupancy_i \\ &+ \beta_8 landmarked\_building_i + \beta_9 elevator_i \\ &+ \beta_{10} parking\_space_i + \beta_{11} building\_type_i \\ &+ \beta_{12} construction\_year_i + \gamma_l + \delta_t + \varepsilon_{ilt}\end{aligned} \tag{7}$$

Additional control variables that are added are *"no_rooms," "furnished," "refurbished," "first_occupancy," "landmarked_building," "elevator," "parking_space," "building_type"* and *"δ"*. The *"no_rooms"* variable is a categorical variable that indicates the number of rooms of the property. It ranges from one to twelve rooms (see also section 3.3) and its reference value is set to one. It is included as a categorical variable and not a continuous variable to allow for a more flexible functional form following Malpezzi (2002, p. 81). The variable *"building_type"* is controlling for the building type. It is coded as a categorical variable and controls for differences in valuation between e.g., an apartment and a detached single-family building. The *"δ"* variable stands for the quarter and the year an observation was first seen online. It controls for differences in valuation that are caused by the real estate market cycle. The models were also tested with monthly control variables, leading to no relevant changes of the coefficients regarding magnitude or significance. The remaining variables are binary control variables that either indicate that a building has a certain characteristic or does not have a certain characteristic. The property is either furnished, has been refurbished in the past, has not been occupied before, falls under the "Denkmalschutz" in Germany, is equipped with an elevator, comes with a parking space or the respective opposite. As a robustness check, a full continuous cold rent model using the explanatory variable *"energy_consumption"* is also specified:

$$\begin{aligned}\ln(cold\_rent_{ilt}) &= \alpha + \beta_1 energy\_consumption_i \\ &+ \beta_2 living\_space_i + \beta_3 \ln(living\_space_i) \\ &+ \beta_4 no\_rooms_i + \beta_5 furnished_i \\ &+ \beta_6 refurbished_i + \beta_7 first\_occupancy_i \\ &+ \beta_8 landmarked\_building_i + \beta_9 elevator_i \\ &+ \beta_{10} parking\_space_i + \beta_{11} building\_type_i \\ &+ \beta_{12} construction\_year_i + \gamma_l + \delta_t + \varepsilon_{ilt}\end{aligned} \tag{8}$$

The only difference between this model and the previous model is the key explanatory variable of *"energy_consumption"* that is now included. The variable *"epc_level"* is not included in this model. The variable *"energy_consumption"* is a metric variable indicating the energy consumption of a building in kWh / (m^2 * a). The final model for estimating the impact of energy efficiency on the cold rent of a building includes the EPC type used to communicate the energy efficiency of the building and an interaction term between the EPC type used and the EPC rating. The full interaction cold rent model is defined in the following way:

$$
\begin{aligned}
\ln(cold\_rent_{ilt}) &= \alpha + \beta_1 epc\_level_i \\
&+ \beta_2 epc\_type_i + \beta_3 epc\_type_i * epc\_level_i \\
&+ \beta_4 living\_space_i + \beta_5 \ln(living\_space_i) \\
&+ \beta_6 no\_rooms_i + \beta_7 furnished_i \\
&+ \beta_8 refurbished_i + \beta_9 first\_occupancy_i \\
&+ \beta_{10} landmarked\_building_i + \beta_{11} elevator_i \\
&+ \beta_{12} parking\_space_i + \beta_{13} building\_type_i \\
&+ \beta_{14} construction\_year_i + \gamma_l + \delta_t + \varepsilon_{ilt}
\end{aligned} \tag{9}
$$

The *“epc_type”* variable is a binary control variable that indicates whether the building is issued a requirement certificate or a consumption certificate. The coefficient of this binary variable measures the magnitude of the difference between the two EPC types when a building has the *“epc_level”* that is equal to D. The reference value is set to requirement certificate. The term *“epc_type * epc_level”* is the interaction term between the EPC type issued and the EPC rating. If the building is issued a requirement certificate, the variable *“epc_type”* is equal to zero because this is the reference value of the binary control variable. In this case, the whole term *“epc_type * epc_level”* is equal to zero. If the building is issued a consumption certificate, the control variable *“epc_type”* is equal to one. Then, the coefficient of the interaction term is added to the coefficient of the *“epc_type”* variable. Together, the terms measure the premiums or discounts for a respective EPC level when a consumption certificate is used compared to the reference value of a building with an EPC level of D and a requirement certificate.

Next, the warm rent models are specified. They help to test the hypotheses 1 b) and 2 b). While the specification of the warm rent models remains similar to the cold rent models, their interpretation is more complex. The reason for this is the added layer of costs that can vary significantly between around 6% and 60% of the cold rent in the data sample. The basic warm rent model is defined in the following way:

$$
\begin{aligned}
\ln(warm\_rent_{ilt}) &= \alpha + \beta_1 epc\_level_i \\
&+ \beta_2 living\_space_i + \beta_3 \ln(living\_space_i) \\
&+ \beta_4 construction\_year_i + \gamma_l + \varepsilon_{ilt}
\end{aligned} \tag{10}
$$

Compared to the basic cold rent model, only the dependent variable is changed. The dependent variable is now equal to the warm rent of the building. Again, this variable is transformed using the natural logarithm to account for potential heteroscedasticity. All other explanatory variables remain the same as in the basic cold rent model. This is also the case for the full categorical warm rent model, the full continuous warm rent model and the full interaction warm rent model. Thus, these three models are specified in the following way:

$$
\begin{aligned}
\ln(warm\_rent_{ilt}) &= \alpha + \beta_1 epc\_level_i \\
&+ \beta_2 living\_space_i + \beta_3 \ln(living\_space_i) \\
&+ \beta_4 no\_rooms_i + \beta_5 furnished_i \\
&+ \beta_6 refurbished_i + \beta_7 first\_occupancy_i \\
&+ \beta_8 landmarked\_building_i + \beta_9 elevator_i \\
&+ \beta_{10} parking\_space_i + \beta_{11} building\_type_i \\
&+ \beta_{12} construction\_year_i + \gamma_l + \delta_t + \varepsilon_{ilt}
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
\ln(warm\_rent_{ilt}) &= \alpha + \beta_1 energy\_consumption_i \\
&+ \beta_2 living\_space_i + \beta_3 \ln(living\_space_i) \\
&+ \beta_4 no\_rooms_i + \beta_5 furnished_i \\
&+ \beta_6 refurbished_i + \beta_7 first\_occupancy_i \\
&+ \beta_8 landmarked\_building_i + \beta_9 elevator_i \\
&+ \beta_{10} parking\_space_i + \beta_{11} building\_type_i \\
&+ \beta_{12} construction\_year_i + \gamma_l + \delta_t + \varepsilon_{ilt}
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
\ln(warm\_rent_{ilt}) &= \alpha + \beta_1 epc\_level_i \\
&+ \beta_2 epc\_type_i + \beta_3 epc\_type_i * epc\_level_i \\
&+ \beta_4 living\_space_i + \beta_5 \ln(living\_space_i) \\
&+ \beta_6 no\_rooms_i + \beta_7 furnished_i \\
&+ \beta_8 refurbished_i + \beta_9 first\_occupancy_i \\
&+ \beta_{10} landmarked\_building_i + \beta_{11} elevator_i \\
&+ \beta_{12} parking\_space_i + \beta_{13} building\_type_i \\
&+ \beta_{14} construction\_year_i + \gamma_l + \delta_t + \varepsilon_{ilt}
\end{aligned} \tag{13}
$$

The six different models specified above all concern the rental market and use the same data sample to estimate the coefficients. For analyzing the effect of energy efficiency on the sales prices of a building, a different data sample is used. While the underlying data is different, the models themselves are only partially adjusted. The variable *“furnished”* is dropped as an explanatory variable. On the other hand, additional relevant explanatory variables are added to control for heterogeneity in the data. These include the *“active_lease”* and *“commission_free”* variables. Starting again with the basic model, it is specified in the following way:

$$
\begin{aligned}
\ln(sales\_price_{ilt}) &= \alpha + \beta_1 epc\_level_i + \\
&+ \beta_2 \ln(living\_space_i) + \beta_3 construction\_year_i \\
&+ \gamma_l + \varepsilon_{ilt}
\end{aligned} \tag{14}
$$

The dependent variable *“sales_price”* is equal to the listing sales price of the observation. The dependent variable is transformed using the natural logarithm to account for potential heteroscedasticity. Compared to the basic cold rent model, the linear non-transformed term of the *“living_space”* variable is not included in the model. The residual plot of the model was considered, and no relevant non-linearity was found that would require an additional term. The key explanatory variable *"epc_level"* and the other control variables remain the same as in the basic cold rent model and basic warm rent model. The full categorical sales price model is specified in the following way:

$$\begin{aligned}\ln(sales\_price_{ilt}) &= \alpha + \beta_1 epc\_level_i \\ &+ \beta_2 \ln(living\_space_i) + \beta_3 no\_rooms_i \\ &+ \beta_4 active\_lease_i + \beta_5 refurbished_i \\ &+ \beta_6 first\_occupancy_i + \beta_7 landmarked\_building_i \\ &+ \beta_8 elevator_i + \beta_9 parking\_space_i \\ &+ \beta_{10} building\_type_i + \beta_{11} construction\_year_i \\ &+ \beta_{12} commission\_free_i + \gamma_l + \delta_t + \epsilon_{ilt}\end{aligned} \tag{15}$$

The added variables are *"no_rooms," "active_lease," "refurbished," "first_occupancy," "landmarked_building," "elevator," "parking_space," "building_type," "commission_free"* and *"δ."* The variable *"furnished"* is not added to the model as it is generally not relevant for properties that are for sale. Like the cold rent model, the *"no_rooms"* variable is a categorical variable indicating the number of rooms of the building. For the sales price data sample, it ranges from 1 to 30 rooms. All the other explanatory variables that were included in the rent models have the same meaning. This includes the *"δ"* variable controlling for the upload date of the building. Two variables, however, were not included before: *"active_lease"* and *"commission_free."* The *"active_lease"* variable indicates whether the building for sale has an active lease. This limits the buyer's possibilities of leasing the property or occupying the building. One implication can be the limitation of the agreed upon rent between the previous owner and the tenant. If it is below current rents in the market, the income generated by the property is smaller than the potential one of a comparable property without an active lease. The variable *"commission_free"* indicates whether the buyer must pay a commission to a real estate agent managing the sale or not. If no commission must be paid, this reduces the additional transaction costs that a buyer needs to pay for when becoming the new owner of a property. The third model, the full continuous sales price model, is specified in the following way:

$$\begin{aligned}\ln(sales\_price_{ilt}) &= \alpha + \beta_1 energy\_consumption_i \\ &+ \beta_2 ln(living_space_i) + \beta_3 no\_rooms_i \\ &+ \beta_4 active\_lease_i + \beta_5 refurbished_i \\ &+ \beta_6 first\_occupancy_i + \beta_7 landmarked\_building_i \\ &+ \beta_8 elevator_i + \beta_9 parking\_space_i \\ &+ \beta_{10} building\_type_i + \beta_{11} construction\_year_i \\ &+ \beta_{12} commission\_free_i + \gamma_l + \delta_t + \varepsilon_{ilt}\end{aligned} \tag{16}$$

As explained above for the rent models, the only difference between this model and the full categorical sales price model is the key explanatory variable of *"energy_consumption"* that is now included. The variable *"epc_level"* is not included in this model. The variable *"energy_consumption"* is a metric variable indicating the energy consumption of a building in kWh / (m^2 * a). The final model for estimating the impact of energy efficiency on the sales price of a building includes the EPC type and an interaction term between the EPC type used and the EPC rating. The full interaction sales price model is defined in the following way:

$$\begin{aligned}\ln(sales\_price_{ilt}) &= \alpha + \beta_1 epc\_level_i + \beta_2 epc\_type_i \\ &+ \beta_3 epc\_type_i * epc\_level_i + \beta_4 \ln(living\_space_i) \\ &+ \beta_5 no\_rooms_i + \beta_6 active\_lease_i \\ &+ \beta_7 refurbished_i + \beta_8 first\_occupancy_i \\ &+ \beta_9 landmarked\_building_i + \beta_{10} elevator_i \\ &+ \beta_{11} parking\_space_i + \beta_{12} building\_type_i \\ &+ \beta_{13} construction\_year_i + \beta_{14} commission\_free_i \\ &+ \gamma_l + \delta_t + \varepsilon_{ilt}\end{aligned} \tag{17}$$

The variable *"epc_type"* and the interaction term *"epc_type * epc_level"* are both added to the full categorical sales price model. The interpretation of both is the same as for the full cold rent interaction model and the full warm rent interaction model with the difference that the impact on the sales price is measured.

Finally, the specified models are run, and their explanatory power is assessed by validating the model assumptions mentioned in section 3.1. The assumptions, their definitions and the respective tests that are run can be found in Appendix 2. The results of these assumptions tests are shortly summarized now before presenting the empirical results in the next section.

Linearity assumption: Graphical plots showing the fitted values on the x-axis and the magnitude of their residuals on the y-axis are used to assess this assumption. Overall, no relevant deviation from linearity can be detected. While there is some non-linearity present for the basic models, the full categorical, continuous and interaction models show almost perfect linearity of residuals. This supports the conclusion that the full models have a greater explanatory power and should be the ones interpreted rather than the basic models.

No multicollinearity assumption: Strong multicollinearity impacts the significance of the correlated variables. First indicative values for correlation in the data were presented with the correlation matrixes and discussed in section 3.3. To assess multicollinearity of explanatory variables, the variance inflation factor (VIF) is calculated. As the models include multiple categorical explanatory variables with more than one degree of freedom, the values of the variables are likely to be artificially inflated. Thus, the generalized variance inflation factor (GVIF) is calculated following Fox and Monette (1992, p. 140). Overall, no problematic multicollinearity is found for the models. It should be mentioned that there exist high values for the explanatory variables that are based on the amount of living space. However, this was expected and considered when the models were specified. The moderate correlation found in the data in section 3.3 for energy efficiency and construction year does not result in problematic multicollinearity. The same is true for the use of the number of rooms as categorically coded variable. The values of the GVIF for all variables except the ones based on living space models range between 1.0 and 1.4. The terms based on living space show high values of up to 5.9, potentially impacting their significance values.

Homoscedasticity assumption: Heteroscedasticity impacts the significance of coefficients. To address this problem, the dependent variable is log-transformed, a graphical plot of residuals is used for diagnosis and robust standard errors , also known as White standard errors, that account for heteroscedasticity are calculated (see White, 1980). In the plots, there is only a slight deviation from the perfect value of one for the square root of the absolute values of the standardized residuals present. Such a deviation is common for multiple linear regression applications and the reason why additionally White standard errors are computed. They are presented as part of the empirical results in the next section. Thus, the homoscedasticity assumption can be seen as sufficiently fulfilled.

No autocorrelation assumption: The Durbin-Watson statistic is used to test the assumption of no autocorrelation. The test can result in values that are between zero and four. No autocorrelation is present when the test result is equal to two. A rather conservative approach is to say that values below 1 and above 3 are of concern and problematic (Field et al., 2012, p. 917). The values found for the cold rent, warm rent and sales price models range between 1.72 and 1.96. This indicates weak autocorrelation for some models and almost none for others. However, all values are within the unproblematic area of 1 - 3, resulting in the no autocorrelation assumption being sufficiently fulfilled.

Exogeneity assumption: The exogeneity assumption requires the expected value of the residual vector to be zero. This is always technically fulfilled for the data sample used to estimate the models and caused by the mathematical process of minimizing the squared residuals of the model. However, no statement can be made regarding the overall population. Further, the potential problem of omitted variable bias is not addressed. These are general limitations of any empirical study that employs hedonic price models.

4. Presentation of the empirical results

The economic analysis in this paper consists of two parts. The first part addresses hypotheses 1 a) – 1 c) regarding the general impact of energy efficiency on cold rent, warm rent and sales prices. The second part addresses hypotheses 2 a) – 2 c) regarding the impact of energy efficiency based on the EPC type used on the cold rent, warm rent and sales prices. Thus, the empirical results of the models that do not include the interaction term are presented first. This is followed by the presentation of the empirical results of the models that include the interaction term. The results in the tables are rounded to five decimal places. Significance computations were done using the non-rounded values. All coefficients presented in the tables show the impact on the log-transformed dependent variable. A one unit increase in a non-log transformed explanatory variable increases the log value of the dependent variable by the magnitude of the coefficient. To make the understanding of these values and their economic meaning more intuitive, they are converted into percentage values in the text. All standard errors presented are robust White standard errors. They can be found below the coefficient values in the tables. The focus is on the key explanatory variable of energy efficiency. The other explanatory variables are shortly presented at the end of the section.

Table 3 shows the empirical results of the basic, full categorical and full continuous cold rent model. The basic cold rent model shows a strong overall statistical significance. Its F-statistic is equal to 2 936 (p-value: < 2.2e-16) with 499 and 211 667 degrees of freedom for the regression and error, respectively. The R-squared is equal to 0.8737 and the adjusted R-squared is equal to 0.8734. These values indicate that a significant proportion of the variance can be explained by this model while the high R squared value is not caused by a high number of explanatory variables. For the rest of the models, the R-squared and adjusted R-squared values are presented, too. Their meaning remains the same if no large difference in magnitude exists. All explanatory variables are highly significant at the 0.1% level. Significant cold rent premiums are present for an above average energy efficient building. The discounts for a very inefficient building are also significant but comparatively smaller. The magnitude of the coefficients of the EPC levels ranges from 9.9% for an A+ rated building to -1.9% and -1.1% for a G and H rated building when compared to a D rated building. Looking at the magnitude of these values, no continuous linear decrease in cold rent is present.

When looking at the empirical results of the full categorical cold rent model, this non-linearity becomes more evident: The full categorical cold rent model shows a strong overall statistical significance, too. Its F-statistic is equal to 3 505 (p-value: < 2.2e-16) with 581 and 211 585 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.9059 and the adjusted R squared is equal to 0.9056. When looking at the coefficients of the different EPC levels, highly significant premiums for a building that is rated higher than D are estimated. An A+ rated building is estimated to have a cold rent 7.0% higher than a D rated building. An A, B, and C rated building is estimated to have 3.5%, 3.2% and 1.0% higher cold rent, respectively. The estimates for the EPC level E, F, G, and H do not show the same significance level as before. Their coefficients are equal to -0.1%, 0.2%, -0.8% and -0.4% respectively, but only the coefficient of the EPC level G is still significant at the 5% level. Thus, it cannot be stated that the impact of the EPC levels E, F and H on cold rent is different from zero. Similarly, to the basic cold rent model, there are strong premiums present for a very energy efficient building and only limited or no discounts present for an energy inefficient building.

Table 3: Cold rent hedonic regression results

Independent variables	**(1) Basic model**	**(2) Full categorical model**	**(3) Full continuous model**	**(4) Full interaction model**
EPC - A+ (Ref: D)	0.09489 **** 0.00239	0.06783 *** 0.00210	-	0.08458 *** 0.00264
EPC - A (Ref: D)	0.04815 *** 0.00207	0.03395 *** 0.00182	-	0.04962 *** 0.00250
EPC - B (Ref: D)	0.04488 *** 0.00160	0.03194 *** 0.00141	-	0.05191 *** 0.00222
EPC - C (Ref: D)	0.01766 *** 0.00130	0.00994 *** 0.00113	-	0.02902 *** 0.00232
EPC - E (Ref: D)	-0.00707 *** 0.00117	-0.00053 0.00100	-	0.00269 0.00231
EPC - F (Ref: D)	-0.00798 *** 0.00133	0.00199 0.00114	-	0.00315 0.00242
EPC - G (Ref: D)	-0.01925 *** 0.00186	-0.00758 *** 0.00162	-	-0.01478 ** 0.00277
EPC - H (Ref: D)	-0.01136 *** 0.00276	-0.00402 0.00234	-	-0.00412 0.00323
EPC type (consumption)	-	-	-	0.01094 *** 0.00179
EPC - A+ Int.	-	-	-	-0.02727 *** 0.00586
EPC - A Int.	-	-	-	-0.01818 *** 0.00377
EPC - B Int.	-	-	-	-0.03527 *** 0.00263
EPC - C Int.	-	-	-	-0.02579 *** 0.00262
EPC - E Int.	-	-	-	-0.00411 0.00254
EPC - F Int.	-	-	-	-0.00049 0.00272
EPC - G Int.	-	-	-	0.01872 *** 0.00339
EPC - H Int.	-	-	-	0.01146 * 0.00489
Energy consumption	-	-	-0.00017 *** 0.00001	-
Living space	0.00401 *** 0.00005	0.00288 *** 0.00006	0.00291 *** 0.00006	0.00286 *** 0.00006
Ln (living space)	0.57596 *** 0.00350	0.61415 *** 0.00513	0.61204 *** 0.00514	0.61602 *** 0.00513
Refurbished	-	0.04216 *** 0.00077	0.04167 *** 0.00078	0.04246 *** 0.00077
First occupancy	-	0.06711 *** 0.00109	0.06900 *** 0.00109	0.06664 *** 0.00109
Landmarked building	-	0.05402 *** 0.01213	0.05387 *** 0.01223	0.05241 *** 0.01198
Elevator	-	0.01906 *** 0.00089	0.01969 *** 0.00089	0.01866 *** 0.00089
Parking space	-	0.03820 *** 0.00075	0.03811 *** 0.00075	0.03796 *** 0.00075
Furnished	-	0.17042 *** 0.00259	0.17062 *** 0.00259	0.17074 *** 0.00259
Intercept	3.40426 *** 0.03643	3.13110 *** 0.03349	3.16666 *** 0.03304	3.11788 *** 0.03322
Categorical control variables				
No. of rooms	No	Yes	Yes	Yes
Building type	No	Yes	Yes	Yes
Construction year	Yes	Yes	Yes	Yes
Location	Yes	Yes	Yes	Yes
Upload date	No	Yes	Yes	Yes
Model statistics				
R squared	0.8737	0.9059	0.9055	0.9061
Adjusted R squared	0.8734	0.9056	0.9052	0.9058
No. of observations	212 167	212 167	212 167	212 167

Significance Levels: (*) $p < 0.05$; (**) $p < 0.01$; (***) $p < 0.001$ Of note: The coefficients show the impact of the log-transformed dependent variable. In the text, the converted valves in percent are used. Thus there might exist differences between the valves.

When looking at the full continuous model, the overall significance of the model does not change much compared to the other two models. Its F-statistic is equal to 3 532 (p-value: < 2.2e-16) with 574 and 211 585 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.9055 and the adjusted R squared is equal to 0.9052. When looking at the value of the coefficient of the explanatory variable *"energy_consumption,"* the estimate is highly significant at the 0.1% level. The magnitude of the coefficient is small with -0.02%. This means that each additional kWh / (m^2 * a) decreases the cold rent by 0.02%. The direction of the coefficient is in line with the full categorical model, but its magnitude is much smaller.

Table 4 displays the empirical results of the basic, full categorical and full continuous warm rent model. The basic warm rent model shows a strong overall statistical significance. Its F-statistic is equal to 2 925 (p-value: < 2.2e-16) with 499 and 211 667 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.8733 and the adjusted R squared is equal to 0.8730. Again, these values indicate that a significant proportion of the variance can be explained by this model while the high R squared value is not caused by a high number of explanatory variables. Most explanatory variables are highly significant at a 0.1% level. Significant at the 1% level is the EPC level H. Not significant are the EPC levels E and F. Similarly, to the basic cold rent model, significant warm rent premiums are present for an above average energy efficient building. The discounts for a very inefficient building are less significant and comparatively smaller. The magnitude of the coefficients of the EPC levels ranges from 7.4% for an A+ rated building to -1.1% and -0.7% for a G and H rated building when compared to a D rated building. Looking at the magnitude of the values, they are showing a non-linear decrease and premiums for an energy efficient building are much greater than the discounts for an energy inefficient building.

The full categorical warm rent model shows a strong overall statistical significance, too. Its F-statistic is equal to 3 521 (p-value: <2.2e-16) with 581 and 211 585 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.9063 and the adjusted R squared is equal to 0.9060. When looking at the coefficients of the different EPC levels, highly significant premiums for a building that is rated higher than D are estimated. The EPC levels above D are all significant at the 0.1% level. The magnitude of the coefficients is smaller than the ones found in the basic warm rent model and smaller than the ones found for the full categorical cold rent model. An A+ rated building is estimated to have a warm rent 4.6% higher than a D rated building. An A, B, and C rated building is estimated to have 2.0%, 2.1% and 0.4% higher warm rents, respectively. The estimates for the EPC level E and F are significant at the 0.1% level with estimates indicating warm rents that are 0.6% and 1.1% higher, respectively. EPC level G and H show lower significance levels at 5%. Their coefficients are equal to 0.3% and 0.4% respectively. Similarly, to the basic and full categorical cold rent models, there are strong premiums present for a very energy efficient building. No discounts are present for an energy inefficient building. The premiums for an energy efficient building above the D level are smaller compared to the cold rent model. There are small premiums for an energy inefficient building below the D level when compared to the cold rent model.

When looking at the full continuous model, the overall significance of the model does not change much compared to the other two models. Its F-statistic is equal to 3 553 (p-value: <2.2e-16) with 574 and 211 592 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.9060 and the adjusted R squared is equal to 0.9057. When looking at the value of the coefficient of the explanatory variable *"energy_consumption,"* the estimate is highly significant at the 0.1% level. The magnitude of the coefficient is small with -0.004%. This means that each additional kWh / (m^2 * a) decreases the warm rent by 0.004%. The direction of the coefficient is the same as the above D rated buildings in the full categorical warm rent model and in the full categorical cold rent model, but its magnitude is much smaller.

Table 5 displays the empirical results of the basic, full categorical and full continuous sales price model. The basic sales price model shows a strong overall statistical significance. Its F-statistic is equal to 1 001 (p-value: < 2.2e-16) with 508 and 159 064 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.7617 and the adjusted R squared is equal to 0.7609. Most explanatory variables are highly significant at the 0.1% level. Significant at the 1% level is the EPC level H. Significant at the 5% level is the EPC level G. Not significant is the EPC level F. Significant sales price premiums are present for an above average energy efficient building. The discounts for an inefficient building are less significant and smaller. The magnitude of the coefficients of the EPC levels ranges from 16.4% for an A+ rated building to -1.1% for an H rated building when compared to a D rated building. Looking at the form of the values, it is comparable with the one found for the basic cold rent model and basic warm rent model. It is also non-linear, but premiums for an energy efficient building are much greater. At the same time, the discounts are not as large and oscillating around zero.

The full categorical sales price model shows a clearer picture regarding the premiums for an energy efficient building and the discounts for an energy inefficient building. Overall, the model shows a strong statistical significance. Its F-statistic is equal to 1 761 (p-value: < 2.2e-16) with 609 and 158 963 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.8709 and the adjusted R squared is equal to 0.8704. When looking at the coefficients of the different EPC levels, highly significant premiums for a building that is rated higher than D are estimated. The EPC levels A+, A and B are all significant at the 0.1% level. The EPC level C is significant at the 1% level. The magnitude of the coefficients is lower than the ones found in the basic sales price model. An A+ rated building is estimated to have a sales price 6.9% higher than a D rated building. An A, B, and C rated building is estimated to have a 2.6%, 3.1% and

Table 4: Warm rent hedonic regression results

Independent variables	**(1) Basic model**	**(2) Full categorical model**	**(3) Full continuous model**	**(4) Full interaction model**
EPC - A+ (Ref: D)	0.07131 *** 0.00226	0.04496 *** 0.00195	-	0.06521 *** 0.00246
EPC - A (Ref: D)	0.03419 *** 0.00196	0.02005 *** 0.00170	-	0.03947 *** 0.00233
EPC - B (Ref: D)	0.03434 *** 0.00151	0.02047 *** 0.00132	-	0.04319 *** 0.00208
EPC - C (Ref: D)	0.01233 *** 0.00122	0.00392 *** 0.00105	-	0.02243 *** 0.00219
EPC - E (Ref: D)	-0.00141 0.00109	0.00557 *** 0.00094	-	0.00561 *** 0.00218
EPC - F (Ref: D)	0.00015 0.00124	0.01142 *** 0.00107	-	0.00991 ** 0.00230
EPC - G (Ref: D)	-0.01062 *** 0.00172	0.00295 * 0.00150	-	-0.00372 0.00258
EPC - H (Ref: D)	-0.00671 ** 0.00256	0.00426 * 0.00221	-	0.00614 * 0.00306
EPC type (consumption)	-	-	-	0.01863 *** 0.00168
EPC - A+ Int.	-	-	-	-0.03552 *** 0.00538
EPC - A Int.	-	-	-	-0.02639 *** 0.00357
EPC - B Int.	-	-	-	-0.03782 *** 0.00247
EPC - C Int.	-	-	-	-0.02417 *** 0.00246
EPC - E Int.	-	-	-	0.00006 0.00240
EPC - F Int.	-	-	-	0.00383 0.00257
EPC - G Int.	-	-	-	0.02047 *** 0.00315
EPC - H Int.	-	-	-	0.01262 ** 0.00458
Energy consumption	-	-	-0.00004 *** 0.00001	-
Living space	0.00320 *** 0.00004	0.00282 *** 0.00006	0.00284 *** 0.00006	0.00280 *** 0.00006
Ln (living space)	0.59321 *** 0.00327	0.58261 *** 0.00476	0.58098 *** 0.00477	0.58516 *** 0.00476
Refurbished	-	0.03619 *** 0.00072	0.03577 *** 0.00072	0.03640 *** 0.00072
First occupancy	-	0.05764 *** 0.00101	0.05910 *** 0.00101	0.05739 *** 0.00101
Landmarked building	-	0.06186 *** 0.01257	0.06235 *** 0.01263	0.05980 *** 0.01247
Elevator	-	0.03560 *** 0.00082	0.03591 *** 0.00082	0.03533 *** 0.00082
Parking space	-	0.03725 *** 0.00069	0.03720 *** 0.00069	0.03673 *** 0.00069
Furnished	-	0.16025 *** 0.00241	0.16035 *** 0.00241	0.16058 *** 0.00241
Intercept	3.66727 *** 0.03246	3.54800 *** 0.02907	3.56467 *** 0.02885	3.52455 *** 0.02866
Categorical control variables				
No. of rooms	No	Yes	Yes	Yes
Building type	No	Yes	Yes	Yes
Construction year	Yes	Yes	Yes	Yes
Location	Yes	Yes	Yes	Yes
Upload date	No	Yes	Yes	Yes
Model statistics				
R squared	0.8733	0.9063	0.9060	0.9067
Adjusted R squared	0.8730	0.9060	0.9057	0.9064
No. of observations	212 167	212 167	212 167	212 167

Significance Levels: (*) $p < 0.05$; (**) $p < 0.01$; (***) $p < 0.001$ Of note: The coefficients show the impact of the log-transformed dependent variable. In the text, the converted valves in percent are used. Thus there might exist differences between the valves.

0.6% higher sales price, respectively. The estimates for the EPC level E and F are not significant. The coefficients are estimated at 0.6% and 1.1%, respectively. However, it cannot be excluded that these EPC levels have no impact on the sales price. EPC level G and H are highly significant again at the 0.1% level. Their coefficients are equal to -1.7% and -7.5% respectively. Overall, there are significant and large premiums present for energy efficient buildings and significant and large discounts present for energy inefficient buildings.

When looking at the full continuous model, the overall significance of the model does not change much compared to the other two sales price models. Its F-statistic is equal to 1 779 (p-value: < 2.2e-16) with 602 and 158 970 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.8707 and the adjusted R squared is equal to 0.8702. When looking at the coefficient of the explanatory variable *"energy_consumption,"* the estimate is highly significant at the 0.1% level. The magnitude of the coefficient is small with -0.04%. This means that each additional kWh / (m^2 * a) decreases the cold rent by 0.04%. Compared to the rent models, the estimate found is larger in magnitude and comes closer to the coefficients of the full categorical sales price model.

Based on the findings above, it is assessed in section 5 whether hypotheses 1 a) – 1 c) are supported by market data. To assess hypotheses 2 a) – 2 c), the more complex full interaction hedonic price models are used. The interaction term *"epc_type * epc_level"* and the binary explanatory variable *"epc_type"* are included in the models to capture the difference in valuation between a building that uses a requirement certificate compared to a consumption certificate. The findings for the interaction models can be found in the last column of Tables 3, 4 and 5. To increase the intuitive understanding of the values, they are converted into percentage values. Further, they are added together to show the overall and direct comparison between the two EPC types. The reference building is a building that is issued a requirement certificate and has an EPC rating D. Any comparison is made to this reference building. Table 6 displays the computed percentage values of the premiums and discounts of buildings with a requirement certificate and a consumption certificate based on the full interaction cold rent, warm rent and sales price model. Figures 7, 8 and 9 visualize the findings.

The full interaction cold rent model shows a strong overall statistical significance. Its F-statistic is equal to 2 936 (p-value: <2.2e-16) with 590 and 211 576 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.9061 and the adjusted R squared is equal to 0.9058. As before, these values indicate that a significant proportion of the variance can be explained by this model while the high R squared value is not caused by a high number of explanatory variables. When looking at the coefficients of the different EPC levels, highly significant premiums for a building that is rated higher than D are estimated for both EPC types. The EPC levels A+, A, B and C are all significant at the 0.1% level. An A+ rated building with a requirement certificate is estimated to have a cold rent 8.82% higher than a D rated building. An A, B, and C rated building is estimated to have 5.09%, 5.33% and 2.94% higher cold rent, respectively. The coefficients of a building with a consumption certificate are lower. An A+ rated building is estimated to have a 7.08% higher cold rent while an A, B and C rated building is estimated to have a 4.34%, 2.82% and 1.50% higher cold rent, respectively. A D rated building with a consumption certificate has a 1.1% higher cold rent than a building with a requirement certificate. The estimates for the EPC level E and F are not significant for either EPC type. The coefficients are estimated with 0.26% and 0.31% for the requirement certificate and 0.95% and 1.41% for the consumption certificate, respectively. However, it cannot be excluded that these EPC levels have no impact on the cold rent. EPC level G is highly significant again at the 0.1% level. The coefficients are equal to -1.39% for the requirement certificate and 1.57% for the consumption certificate. While the EPC level H is not significant for the requirement certificate (with a coefficient of -0.40%), it is significant at the 5% level for the consumption certificate with a coefficient value of 1.85%. Overall, there are significant and large premiums present for a very energy efficient building. They are larger for the requirement certificate. For a very energy inefficient building there are no large discounts visible. There are even higher cold rents estimated for an inefficient building with a consumption certificate.

The full interaction warm rent model shows a strong overall statistical significance. Its F-statistic is equal to 3 483 (p-value: <2.2e-16) with 590 and 211 576 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.9067 and the adjusted R squared is equal to 0.9064. When looking at the coefficients of the different EPC levels, highly significant premiums for a building that is rated higher than D are estimated for both EPC types. The EPC levels A+, A, B and C are all significant at the 0.1% level. An A+ rated building with a requirement certificate issued is estimated to have a warm rent 6.74% higher than a D rated building. An A, B, and C rated building is estimated to have 4.02%, 4.40% and 2.27% higher warm rent, respectively. The coefficients of a building with a consumption certificate are a bit lower: An A+ rated building is estimated to have a 5.00% higher warm rent while an A, B and C rated building are estimated to have 3.25%, 2.5% and 1.71% higher warm rent, respectively. A D rated building with a consumption certificate has a 1.88% higher warm rent than a building with a requirement certificate. The estimates for a building with a requirement certificate and an EPC level E (0.56%) are significant at the 1% level and significant at the 0.1% level for an EPC level F (0.99%). EPC level G (-0.30%) is not significant while the EPC level H (0.61%) is significant at the 5% level. When it comes to a building with a consumption certificate, the EPC levels E (2.45%) and F (3.28%) are not significant while the EPC level G (3.67%) is significant at the 0.1% level and the EPC level H (3.80%) is significant at the 1% level. Overall, there are still significant and large premiums present for a very energy efficient building. They are larger for the requirement certificate. A very energy inefficient building shows a higher warm rent compared to a D

Table 5: Sales price hedonic regression results

Independent variables	**(1) Basic model**	**(2) Full categorical model**	**(3) Full continuous model**	**(4) Full interaction model**
EPC - A+ (Ref: D)	0.15196 *** 0.00478	0.06730 *** 0.00374	-	0.07137 *** 0.00498
EPC - A (Ref: D)	0.05853 *** 0.00458	0.02614 *** 0.00357	-	0.04167 *** 0.00509
EPC - B (Ref: D)	0.04600 *** 0.00358	0.03024 *** 0.00272	-	0.04557 *** 0.00474
EPC - C (Ref: D)	0.01832 *** 0.00290	0.00602 ** 0.00208	-	0.01705 *** 0.00465
EPC - E (Ref: D)	-0.00929 *** 0.00274	-0.00319 0.00193	-	-0.01666 *** 0.00412
EPC - F (Ref: D)	-0.00050 0.00311	-0.00370 0.00227	-	-0.01880 *** 0.00414
EPC - G (Ref: D)	0.00852 * 0.00393	-0.01745 *** 0.00295	-	-0.04447 *** 0.00437
EPC - H (Ref: D)	-0.01105 ** 0.00426	-0.07799 *** 0.00334	-	-0.10711 *** 0.00439
EPC type (consumption)	-	-	-	-0.02981 *** 0.00333
EPC - A+ Int.	-	-	-	0.01070 0.00821
EPC - A Int.	-	-	-	-0.02983 *** 0.00636
EPC - B Int.	-	-	-	-0.02201 *** 0.00508
EPC - C Int.	-	-	-	-0.01297 * 0.00513
EPC - E Int.	-	-	-	0.01617 *** 0.00465
EPC - F Int.	-	-	-	0.01588 ** 0.00492
EPC - G Int.	-	-	-	0.04499 *** 0.00629
EPC - H Int.	-	-	-	0.08648 *** 0.00884
Energy consumption	-	-	-0.00038 *** 0.00001	-
Ln (living space)	1.00805 *** 0.00179	0.90813 *** 0.00368	0.90783 *** 0.00368	0.90461 *** 0.00368
Refurbished	-	0.04945 *** 0.00192	0.04885 *** 0.00192	0.04884 *** 0.00191
First occupancy	-	0.05171 *** 0.00281	0.05197 *** 0.00282	0.04724 *** 0.00280
Landmarked building	-	0.01774 0.01200	0.01795 0.01198	0.01632 0.01192
Elevator	-	-0.03396 *** 0.00182	-0.03511 *** 0.00181	-0.03478 *** 0.00181
Parking space	-	0.01745 *** 0.00137	0.01692 *** 0.00137	0.01815 *** 0.00137
Existing lease	-	-0.05869 *** 0.00185	-0.05858 *** 0.00185	-0.05827 *** 0.00185
Commission free	-	0.02027 *** 0.00157	0.02047 *** 0.00157	0.01836 *** 0.00157
Intercept	7.24672 *** 0.11175	7.21082 *** 0.08352	7.26770 *** 0.08368	7.24219 *** 0.08315
Categorical control variables				
No. of rooms	No	Yes	Yes	Yes
Building type	No	Yes	Yes	Yes
Construction year	Yes	Yes	Yes	Yes
Location	Yes	Yes	Yes	Yes
Upload date	No	Yes	Yes	Yes
		Model statistics		
R squared	0.7617	0.8709	0.8707	0.8713
Adjusted R squared	0.7609	0.8704	0.8702	0.8708
No. of observations	159 573	159 573	159 573	159 573

Significance Levels: (*) $p < 0.05$; (**) $p < 0.01$; (***) $p < 0.001$ Of note: The coefficients show the impact of the log-transformed dependent variable. In the text, the converted valves in percent are used. Thus there might exist differences between the valves.

Table 6: Full interaction model results in percentage values

	Cold rent interaction effects		Warm rent interaction effects		Sales price interaction effects	
EPC level	Requirement certificate	Consumption certificate	Requirement certificate	Consumption certificate	Requirement certificate	Consumption certificate
A+	8.82%	7.08%	6.74%	5.00%	7.40%	5.36%
A	5.09%	4.34%	4.02%	3.25%	4.26%	-1.78%
B	5.33%	2.82%	4.40%	2.50%	4.66%	-0.62%
C	2.94%	1.50%	2.27%	1.71%	1.72%	-2.54%
D	0.00%	1.10%	0.00%	1.88%	0.00%	-2.94%
E	0.26%	0.95%	0.56%	2.45%	-1.65%	-2.98%
F	0.31%	1.41%	0.99%	3.28%	-1.86%	-3.22%
G	-1.39%	1.57%	-0.30%	3.67%	-4.35%	-2.89%
H	-0.40%	1.85%	0.61%	3.80%	-10.16%	-4.92%

Of note: The grey font color is used for non-significant values.

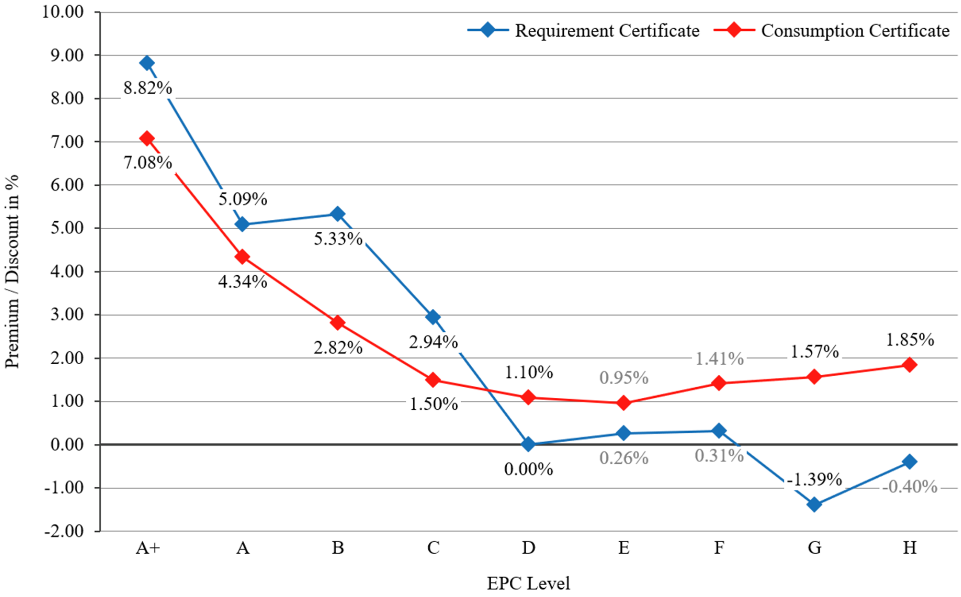


Figure 7: Full interaction cold rent effects (font color grey for non-significant values)

rated building, mainly when it is issued a consumption certificate.

The sales price interaction model shows a strong overall statistical significance. Its F-statistic is equal to 1 742 (p-value: <2.2e-16) with 618 and 158 954 degrees of freedom for the regression and error, respectively. The R squared is equal to 0.8713 and the adjusted R squared is equal to 0.8708. When looking at the coefficients of the different EPC levels, all are highly significant at the 0.1% level for a building that is issued a requirement certificate. For a building issued with a consumption certificate, EPC levels A, B, D, E, G and H are significant at the 0.1% level, EPC level F is significant at the 1% level and EPC level C is significant at the 5% level. Only EPC level A+ is not significant. The coefficients for a building with a requirement certificate are estimated at 7.4% for EPC level A+, 4.26% for A, 4.66% for B, 1.72% for C, -1.65% for E, -1.86% for F, -4.35% for G and -10.16% for H. These values are higher compared to the ones found for the full categorical sales price model. The coefficients for a building with a consumption certificate do not show premiums for above-average energy efficient buildings. Only the non-significant EPC level A+ has a premium of 5.36%. The other EPC levels show discounts. An A rated building with a consumption certificate is estimated to have a discount of -1.78%, a B rated building -0.62%, a C rated building -2.54%, a D rated building -2.94%, an E rated building -2.98%, a F rated building -3.22%, a G rated building -2.89% and an H rated building -4.92%. Overall, there are significant and large premiums and discounts present for a building with a requirement certificate. Buildings with a consumption certificate are estimated to have significantly lower sales price compared to buildings with a requirement certificate across all EPC levels except for G and H.

In the full categorical, continuous and interaction model, various control variables are included. The estimates for their coefficients differ between the cold rent, warm rent and sales

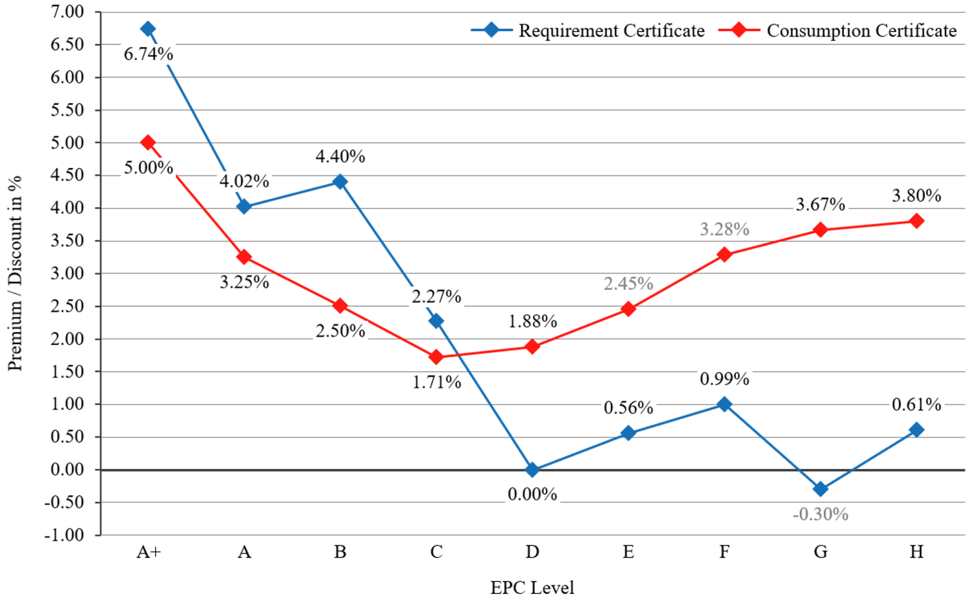


Figure 8: Full interaction warm rent effects (font color grey for non-significant values)

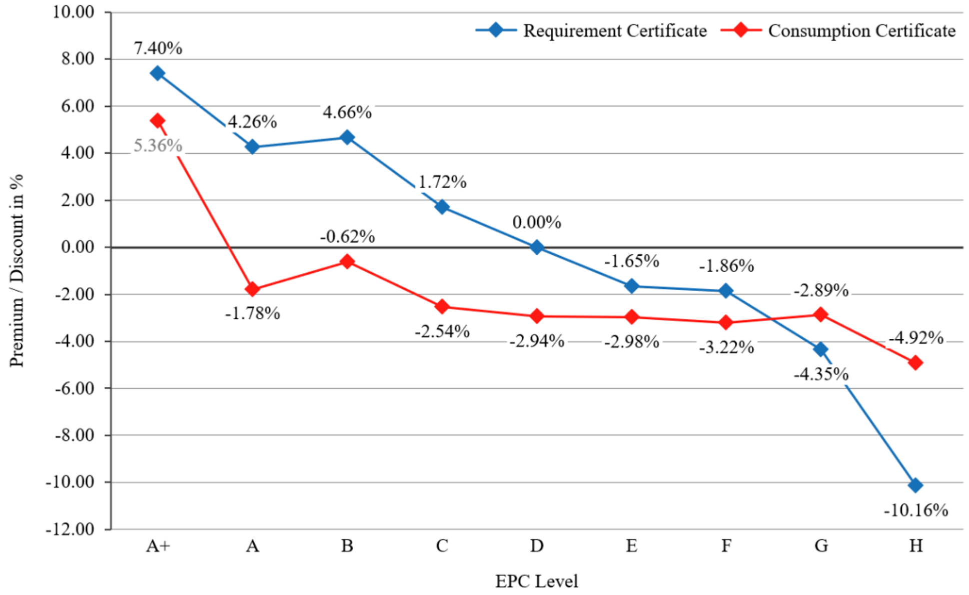


Figure 9: Full interaction sales price effects (font color grey for non-significant values)

price models. When considering the different models for only one of the dependent variables, the significance and magnitude of the coefficients of the control variables do not change much. The values presented below are taken from the full categorical models found in Tables 3, 4 and 5. The control variables of the cold rent model are all significant at the 0.1% level. An increase in living space increases the cold rent by 0.3% for every additional m^2. It also increases relatively by 84.8% for every 100% increase in living space. A furnished building has an 18.6% higher cold rent, the cold rent for a newly built building increases by 6.9% and for a refurbished building by 4.3%. A landmarked building has a 5.5% higher cold rent. Equipping a building with an elevator increases the cold rent by 1.9% and an existing parking space by 3.9%. The coefficients for the warm rent model are similar. They are overall a bit smaller, except for the elevator. This seems logical as the operating costs are not directly affected by most, except for the elevator, and thus their impact is proportionally smaller. Each additional m^2 of space increases the warm rent by 0.3% and the relative increase is 79.1% for every 100% increase in living space. Furnishing an apartment leads to a 17.4% increase in warm rent. A newly built building can achieve a 5.9% higher warm rent and a refurbished building a 3.7% higher warm rent. A landmarked building has a 6.4%

higher warm rent. Adding an elevator to a building increases the warm rent by 3.6%, while a parking space increases the warm rent by 3.8%. Most of the hedonic characteristics of the full categorical sales price model are also highly significant. Except for the control variable *"landmarked_building,"* which is not significant, all are significant at the 0.1% level. A 100% increase in living space increases the sales price by 147.9%. A newly built building is valued at 5.3% more and a refurbished building is estimated at a 5.1% higher sales price. Selling a building without a commission increases the price by 2.0%, while a parking space increases the building price by 1.8%. An elevator decreases the sales price by -3.3% and an existing lease decreases the sales price by -5.6%.

5. Discussion

This study aims to analyze how the energy efficiency of a building impacts its cold rent, warm rent and sales price in the residential real estate market in the Rhein-Main Region in Germany. Based on the normative approach of real estate valuation theory and evidence found in the extant literature, significant energy efficiency premiums should exist. For the rental market it was hypothesized that the energy efficiency of a building influences both its cold rent and warm rent (hypotheses 1a, 1b) and that these effects depend on the EPC type of the building (hypotheses 2a, 2b). For the sales market it was hypothesized that an increase in the energy efficiency of a building increases its sales price (hypothesis 1c), and that the EPC type modifies this effect (hypothesis 2c). These hypotheses were tested using real world data and hedonic price models. Next, the results of these estimations are discussed, and it is assessed whether they corroborate the hypotheses. Finally, possibilities for future research are outlined.

The impact of the energy efficiency of a building on its cold and warm rent

Hypothesis 1 a) states that an increase in the energy efficiency of a residential building leads to an increase in its cold rent. To test this hypothesis, the different cold rent models ranging from the basic cold rent model to the full continuous cold rent model were specified. Following the normative logic that led to the formulation of this hypothesis, there should be a clear linear trend from A+ to H. The reason for this is that energy savings or additional energy costs are reflected in the cold rent of residential real estate in a market environment. All market participants should be price takers in perfect market conditions. When looking at the results of the full categorical cold rent model, this is only partially supported. Above-average energy efficient buildings achieve significant cold rent premiums of up to 7.0% (A+ rating) when compared to a D-rated building. Even though C rated buildings can achieve a 1.0% higher cold rent, the trend towards decreasing cold rents does not continue for buildings with higher energy consumption. Significance is low and there is no clear indication of large discounts. This shows a non-linear impact of energy efficiency on the cold rent of residential buildings. The existence of this non-linear impact demonstrates that while the full continuous model serves its purpose as a robustness check, its coefficients should be interpreted with caution. A reason for this non-linear impact could be market conditions. For example, in a market in which housing is scarce, the negotiation power of prospective tenants is not strong enough to achieve discounts for energy inefficient buildings compared to the average. At the same time, owners can ask for premiums for buildings that are better than the average building stock. While this might be beneficial for owners of highly energy inefficient buildings, this inhibits the renovation of the current building stock. The incentive of improving the energy efficiency of a G rated building for example is limited because it must at least be increased to an energy efficiency level of C. Achieving an EPC level of A or B would be better as it seems unlikely that a 1.0% increase in cold rent achieved with EPC level C would be enough to recoup the investment costs in a reasonable way. This is in line with literature that argues that refurbishment of the building stock is currently not profitable for building owners (März, Stelk, & Stelzer 2022, p. 20). In conclusion, the hypothesis 1 a) is supported for above-average energy efficient buildings only.

An additional layer of complexity is considered with the full interaction cold rent model. It is used to test hypothesis 2 a), which states that based on the EPC type, an increase in the energy efficiency of a residential building leads to an increase in its cold rent. It seems plausible that the requirement certificate is trusted more by prospective tenants than the consumption certificate. The reason for this is that it is based on the building characteristics and thus more objective than the consumption values that are dependent on the behavior of past and current tenants. This leads to the assumption that the requirement certificate should result in clear premiums and discounts. The discounts could now be present with this variable included in the model that addresses a crucial source of heterogeneity in the data. Whether premiums and discounts are present for the consumption certificate is difficult to predict because it depends on the perception of this EPC type by prospective tenants. When looking at buildings with a requirement certificate, significant premiums for energy efficient buildings exist and range up to 8.8% (A+ rating). The premiums found are greater than the ones for the full categorical model for all levels above D. However, there are still no continuous discounts found for energy inefficient buildings. When looking at the buildings with a consumption certificate, significant premiums for above-average energy efficient buildings exist, too. This means that prospective tenants do find some value in the information communicated via the consumption certificate. The magnitude of some coefficients is smaller than for the requirement certificate. Further, a continuously decreasing trend is visible until the D rated building. Then, counterintuitively, the cold rent shows an increase again for a building with a consumption certificate. The findings for the buildings with a requirement certificate of the full interaction model can be explained in the same way as for the full categorical model: differences in negotiation power and a minimum cold rent as a floor value that can be achieved in a competitive market environment. However, this cannot explain the findings for the buildings with a consumption certificate. Here, EPC levels G and H are significant and show positive cold rent premiums of 1.6% and 1.9%. Two explanations come to mind: First, there might be an unobserved variable that is impacting the model results (e.g., an architectural design premium predominantly found in G and H rated buildings that have a consumption certificate). However, no indication for this exists and this explanation remains speculative. Secondly, prospective tenants might be more sensitive to the operational costs that must be paid rather than a small premium on the cold rent of a building. Thus, an owner could choose to decrease the operational costs allocated to the tenant while increasing the cold rent of the building. This could be part of a sales strategy that utilizes the price sensitivity of prospective tenants. While the higher cold rent might be achieved, it is conceivable that such a building will remain on the market for a longer time. Evidence for this exists for the German market (Cajias et al., 2019, p. 177). However, this is speculative and remains unclear with the data used in this analysis. Whether the assumption regarding the sales strategy holds any merit is discussed again when looking at the full interaction warm rent model results. It presents an interesting opportunity for future research. In conclusion, hypothesis 2 a) is partially supported as differences between EPC types exist. Evidence for this is the existence of significant premiums for above-average energy efficient buildings. The premiums exist for both EPC types but are larger for a building with a requirement certificate. No continuous discounts exist for energy inefficient buildings. Highly inefficient buildings (G and H rated) with a consumption certificate might even be priced higher to compete in a market with tenants that have become increasingly more green-aware and sensitive to operational costs.

Hypothesis 1 b) states that an increase in the energy efficiency of a residential building leads to an increase in its warm rent. To test this hypothesis, the different warm rent models ranging from the basic warm rent model to the full continuous warm rent model were specified. Following the normative logic that led to the formulation of this hypothesis, there should be no premiums or discounts present if only the energy savings or additional expenditures are capitalized into cold rents. If only the energy savings and additional expenditures are capitalized, then adding the respective operating costs to the cold rent would result in the same warm rent. If, additionally, investment costs are recouped to make the construction or the modernization of the building viable, there might still be premiums present for the above-energy efficient buildings. Signaling and prestige effects might be present that affect above-average and below-average energy efficient buildings. When looking at the results of the full categorical warm rent model, this is only partially supported. All EPC level coefficients are significant. For above-average energy efficient buildings, premiums exist that go up to 4.6%. This indicates that while a more energy efficient building has lower energy costs, its cold rent is increased by more than the savings achieved with the better energy efficiency. Two explanations appear plausible: First, highly energy efficient buildings might be able to utilize their image to generate prestige premiums. This might be caused by a subgroup of prospective tenants that value energy efficiency more than others, leading to scarcity and higher cold rents on the market. This is in line with literature that has found an energy efficiency premium connected to green awareness (Pommeranz & Steininger, 2021, p. 234). Secondly, the effect might be caused by the mentioned need of investors to recoup their investments more quickly. When building modernizations have taken place, it is allowed to raise rents to achieve this (§ 555 BGB). While this explains the results for the above-average energy efficient buildings, it does not explain the findings for the below-average energy efficient buildings. Here, small premiums are found, too. Based on the normative reasoning, they should not be present. This needs to be interpreted in the context of the full categorical cold rent model. There, the normative reasoning was not supported either. No continuous discounts were found. Thus, it seems logical that when no cold rent discounts for energy inefficient buildings are found that the warm rent is comparably higher because of an increase in energy costs. This increase should be linear, though, which is not found here. It levels off for the G and H rated buildings. Further interpretation requires the additional control for heterogeneity by the full interaction warm rent model. Before going into detail regarding this, it can be said in conclusion that hypothesis 1 b) is only partially supported. While significant warm rent premiums exist, there are no continuous discounts present. Counterintuitively, there are small premiums present for energy inefficient buildings.

The hypothesis 2b) states that based on the EPC type, an increase in the energy efficiency of a residential building leads to an increase in its warm rent. As already explained for hypothesis 2 a), there should be a difference visible between the requirement certificate and the consumption certificate caused by the difference in objectivity and perception of the calculated values. This is combined with the normative reasoning for the warm rent presented above. The resulting prediction is that there are warm rent premiums present for buildings with a requirement certificate, while it remains difficult to predict the results for a building with a consumption certificate. Looking at the model results, this is partially supported. Significant premiums greater in magnitude than in the full categorical model are found for a building with a requirement certificate and a building with a consumption certificate. They range up to 6.7% (A+ rating) for a requirement certificate and up to 5.0% (A+ rating) for a consumption certificate. The results for below-average energy efficient buildings are less clear. They are similar to the findings for the cold rent models. Significant premiums that remain small in magnitude are found for the requirement certificate for buildings rated E, F or H. This is the case even though there were no discounts for the cold rent. Based on the cold rent findings, there should be a linear increase in warm rents present for a building with a requirement certificate that shows the rising energy costs. However, this is not the case. The rents are a bit higher than the cold rents, but the findings seem inconclusive regarding this aspect. Three explanations exist: First, this might be additional evidence for the existence of the prebound effect that was mentioned in section 2.1 (Galvin, 2023, p. 502). Even though the objective energy requirements of a building are high, the actual consumption is much lower, not leading to higher energy costs and thus not leading to a higher warm rent. Second, the owner might financially offset some of the heating costs by not allocating all the other types of operating costs to the tenant to stay competitive in the market. Third: There is a systematic bias in the data because owners enter operating costs in the listings that do not include heating costs. The last one is a possible but rather speculative explanation. The most likely explanation is the first, as evidence for this effect has already been presented in literature. However, this explanation of the prebound effect does not hold true for a building with a consumption certificate. This is the case because the energy efficiency is calculated based on past consumption values that already include the occupant behavior that can cause the prebound effect. A linear increase in warm rent should be visible. Further, the premiums already present for energy inefficient buildings in the cold rent model should increase the magnitude of the warm rent premiums further. Thus, G and H rated buildings should see a disproportionate increase in warm rent. While there is an increase present, this increase flattens significantly for the G and H rated buildings. These values are significant, while the ones for E and F rated buildings are not significant. The findings are difficult to explain at first. However, the flattening of the values for G or H rated buildings is in line with the speculative explanation given for the cold rent results: The owner might be waiving specific operating costs while increasing the cold rent of the building to make up for the additional costs. These operating costs cannot be heating costs, as they are paid for by the tenant, but must be other operating costs. Whether this is true and whether the sensitivity of prospective tenants regarding cold rents is lower than for operating costs should be considered in future research. This would improve the understanding of pricing decisions made by market participants. In conclusion, hypothesis 2 b) is supported by the evidence as significant differences exist between the EPC types. This is shown by the difference in magnitude for above-average and below-average energy efficient buildings.

The impact of the energy efficiency of a building on its sales price

The final dependent variable that is considered is the sales price. Hypothesis 1 c) states that an increase in the energy efficiency of a residential building leads to an increase in its sales price. Following the reasoning from section 2.1, this is caused by an increase in cash flow and a decrease in the cap rate because of lower building-specific risk. When looking at the results of the full categorical sales price model, this is almost fully supported by the findings. Above-average energy efficient buildings show significant and large premiums of up to 7.0% (A+ rating). Further, inefficient buildings show significant discounts of up to -7.5% (H rating). Only EPC levels E and F are not significant, and the impact could be equal to zero. This indicates that an average level of energy efficiency (i.e., buildings rated D-F) does not significantly influence the sales price. Other hedonic characteristics seem to be more important for the buying decision. However, when a building is highly inefficient, prospective buyers realize that additional investments might be necessary shortly after purchasing the building, requiring discounts to account for this. In conclusion, hypothesis 1 c) is mainly supported by the evidence, but exceptions exist for the average of the current building stock where other hedonic characteristics might outweigh the impact of energy efficiency regarding purchasing decisions.

The final hypothesis to be evaluated is hypothesis 2 c) that states that based on the EPC type, an increase in the energy efficiency of a residential building leads to an increase in its sales price. The prediction of the hypothesis focuses on the perception of the different EPC types and the extent to which the values presented in them are trusted. It seems plausible that the requirement certificate is trusted more and thus shows a stronger relationship between sales price and energy efficiency while the consumption certificate is seen as not reliable because its values are influenced too much by previous occupants. Looking at the results of the full interaction sales price model, this hypothesis is supported by the evidence. The coefficients of all the EPC levels of the requirement certificate are highly significant. This is also the case for the ones that were not significant in the full categorical model. Further, the values show larger premiums for energy efficient buildings and larger discounts for energy inefficient buildings. Looking at the graph in Figure 8, the sales price of a building continuously decreases with the EPC categories that indicate lower energy efficiency. This shows that prospective buyers rationally account for differences in energy consumption and energy costs. Disproportionally large premiums and discounts are present for the A+ and H rated buildings. The additional premiums existing for A+ rated buildings might be caused by scarcity of such buildings on the market or by disproportionately high construction costs for such buildings. They thus achieve higher sales prices. The additional discounts for H rated buildings show that the building specific risk increases with higher energy consumption. This indicates that the risk of these buildings becoming a stranded asset might be relevant to the formation of their sales prices. Finally, the values found for the consumption certificate are considered. All coefficients, except the one for the A+ rating, are significant. Their magnitude is different from the ones found for a building with a requirement certificate. While it might be possible to detect a linear relationship when excluding the EPC level A+, the slope of this linear relationship is very small. The results indicate three things: First, buildings with a consumption certificate are generally valued lower than buildings with a requirement certificate. This is the case for all except for highly energy inefficient buildings, which show smaller discounts. One explanation could be as follows: Compared to a rental contract, the obligations agreed upon in a sales contract are several factors greater. Additionally, the buyers might become the occupiers of the building. Combining both aspects makes uncertainty become an important factor. It seems plausible that prospective buyers see greater uncertainty in a consumption certificate than in a requirement certificate. Thus, the values are not trusted and the impact on sales prices limited. The general uncertainty is compensated for by a discount when compared to the average D rated building with a requirement certificate. Second, the values presented in a consumption certificate are generally seen as less objective and more dependent on occupant behavior. This is also relevant when comparing two buildings that both have a consumption certificate. They are both priced similarly in the middle of the scale (C-G). Prospective buyers are probably of the opinion that any shift in this area might be caused by behavior. This leads to the situation that potential differences in energy consumption and energy cost are not included in the sales price. Third, the top and bottom end of the energy efficiency scale show disproportionally large premiums and discounts that were already found for buildings with a requirement certificate. The explanation behind these values remains the same: scarcity on the one hand and building specific risk on the other. Even though these extreme values are measured using a consumption certificate, they seem to be indicative for the energy efficiency of the building. Thus, some informational value is provided by the consumption certificate when extreme energy consumption values are displayed. Extreme values could indicate building characteristics beyond the impact of the occupant's behavior. In conclusion, the hypothesis 2 c) is supported by the results of the analysis. Significant differences exist between the impact of energy efficiency on the sales price of a building with a requirement certificate compared to a building with a consumption certificate. There are large and significant premiums and discounts present for buildings with a requirement certificate. Uncertainty and impacts of occupant behavior make buildings with a consumption certificate difficult to value, leading to moderate discounts for most buildings and limited impacts at the upper and lower end of the scale. While there is a clear linear relationship with extreme values at both ends of the scale for a building with a requirement certificate, there is a much weaker relationship found for a building with a consumption certificate.

Limitations of the present study

When it comes to the generalizability of the results found in this paper, there are several limitations that need to be considered: First, the data used for the analysis only consists of observations from the Rhein-Main Region. This means that the validity of results is greatest for the Rhein-Main Region. For other regions in Germany, the results can be seen as a benchmark value to consider but should not be used as is for quantitative assessments such as profitability computations. The same is true for the usage or comparison of these values on an international level: Within the EU, the implementation of the EPC is different. Beyond the EU, there are other proxies used to measure energy efficiency. Thus, as explained in the literature review, the findings do not establish causality, but are guidance regarding the development of the industry. Second, the limitation of using listings and not actual transaction data should be mentioned again. This was discussed in section 3.2. Third, the data sample is limited by the time period considered. Data were collected starting from 01/2015. Thus, a long time period is considered. Using data only from the first or last years might lead to different estimates. An analysis focusing on the development of the impact of energy efficiency on residential buildings over the years is an interesting topic for future research. Fourth, a further data sample limitation is the information on control variables. This includes the level of detail for control variables (i.e., operating costs) and the problem of missing information resulting in a reduction of the sample size. Finally, the EPC type allocation might not be truly random. This could be caused by the relevant characteristics for allocation according to § 80 GEG. Identifying a subsample with a truly random allocation and no selection bias is another opportunity for future research.

Comparison of the residential markets of the Rhein-Main Region, Germany and the EU

Keeping the limitations above in mind, the results of this paper are now considered in the wider context of empirical literature. First, the results are compared to publications focusing on the German residential market. Next, they are compared to findings in different EU markets. The first evidence that was provided for the German market by Cajias and Piazolo (2013, p. 58) is barely comparable because of their elasticity measurement in percent and usage of other norms. If the values are transferred to a comparison between a D (115 kWh / (m^2 * a)) and an A+ (15 kWh / (m^2 * a)) rated building, no meaningful results are computed (Cajias & Piazolo, 2013, p. 53). Kholodilin et al. (2017) presented the first comparable evidence. The authors measured the linear impact of an increase in kWh / (m^2 * a). Cold rent decreases by -0.02% and the sales price by -0.05% for each additional kWh / (m^2 * a) of energy consumption (Kholodilin et al., 2017, p. 3231) . This is close to the findings of the continuous models in this paper. Each additional kWh / (m^2 * a) is estimated to decrease cold rent by -0.017% and the sales price by -0.038%. This is also in line with the decrease in rent of -0.017% per kWh / (m^2 * a) found by März et al. (2022, pp. 17–18). However, as already discussed, the continuous measurement understates the impact of energy efficiency because of its mainly non-linear impact on rents and sales prices. Cajias et al. (2019) analyzed the impact of energy efficiency using the EPC rating as a categorical variable. Their analysis using a large sample for all of Germany found much lower premiums for the rental market than the ones in this paper. The 0.9% found by Cajias et al. (2019, pp. 186-187) are much lower than the cold rent premiums of 8.8% and 7.1% found by this paper when controlling for the EPC type. This may be explained by the different time periods investigated and the competitive residential market of the Rhein-Main region. However, more recent literature for the rental market presented by Groh et al. (2022, pp. 104–107) also found lower values with premiums of 3.98% for an A+ building compared to a G or H rated building. Sales price values found by Taruttis and Weber (2022) are closer to the ones in this paper. They estimated a premium of 6.9% for a decrease of 100 kWh / (m^2 * a) in energy usage (Taruttis & Weber, 2022, p. 6). The 6.9% are close to the sales price premiums of 5.4% and 7.4% found in this paper when comparing an A+ rated building to a D rated building and controlling for the EPC type. The remaining papers regarding the German residential real estate market focus on integrating additional factors in their models. Pommeranz and Steininger (2021) analyzed interaction effects with data on purchasing power and green awareness of inhabitants. They found a difference of 8.6% for the rental market between an A+ and H rated building (Pommeranz & Steininger, 2021, p. 235). Galvin (2023, p. 501) introduced the differences between EPC types in a more elaborate way than before by considering the prebound effect and comparing the theoretical savings to actual savings. However, here, the continuous encoding of the energy efficiency variable was used in the model (Galvin, 2023, p. 510). The variable was log-transformed, and the results are similar to the ones found for the full continuous model in this paper. When transforming the absolute values found by Galvin (2023) to percentage values using his descriptive statistics of the average building, a decrease of -0.035% is found for each additional kWh / (m^2 * a) while the estimate of the continuous model in this paper is -0.038%. Finally, the results in this paper are compared to Deller (2022, p. 802), a study that analyzed energy efficiency premiums within the same region. However, only general energy efficiency premiums were presented, and the data samples used were much smaller than the ones in this paper (Deller, 2022, p. 802). When looking at the findings, they are comparable to the ones found for the full categorical models in this paper. While the functional form of the coefficients is similar, the magnitude of the premiums for an A+ rated building is greater in this paper for the cold rent model (7.0% compared to 5.8%) and warm rent model (4.6% compared to 3.9%) (Deller, 2022, p. 802). While almost identical premiums for an A+ rated building in the sales market exist (6.9% compared to 6.8%), larger premiums for a B rated building are found in the present study (B: 3.1% compared to 1.5%) (Deller, 2022, p. 802). Additionally, the findings in this paper show greater significance, which is likely caused by the increase in the data sample sizes (Deller, 2022, p. 818). In sum, the results of the present analysis extend and confirm the earlier results of Deller (2022).

Overall, the comparison with literature for the German market indicates that the values found for the continuous models are similar, while the ones found for the full categorical models and the full interaction models are comparably high. Especially the full interaction model shows larger premiums and discounts in comparison. This might point to the aspect that controlling for heterogeneity in the data is crucial. Further, future research should use the categorical variable of the EPC rating to capture the non-linear form of the impact of energy efficiency on the dependent variables. It should follow the literature by setting the reference category to D.

While the estimates found in this paper are comparatively large for the German market, this is not the case when compared to other EU countries. Higher sales price premiums have been found for the most energy efficient buildings when compared to D rated buildings in the Netherlands (Brounen & Kok, 2011, p. 175), Ireland (Hyland et al., 2013, p. 948 – 949) and Wales (Fuerst, McAllister, et al., 2016, p. 26). Evidence of large sales price discounts for highly inefficient buildings was also found in Wales (Fuerst, McAllister, et al., 2016, p. 26). Still, the discounts found using the full interaction sales price model in this paper are even larger for the worst buildings with a requirement certificate (H rating: -10.2%). When comparing the full continuous sales price model with findings on the EU level, they are quite similar. Högberg (2013, p. 256), for example, found a decrease of -0.04% for each additional kWh / (m^2 * a) in Stockholm, Sweden.

It becomes evident that differences between EU countries might exist. Several reasons make this seem plausible: First, the EPC is implemented in a different way in each country. Technical differences in computation methods or cut-off points for the EPC ratings could lead to differing results. Sweden, for example, does not have the EPC rating A+ and starts with A (Boverket, 2023). Second, energy costs are significantly different in the EU countries (Eurostat, 2023a). This would, based on the normative reasoning of capitalizing energy savings, lead to different impacts of the energy efficiency level on the valuation of a residential property. Third, average house prices are significantly different (Deloitte, 2023, p. 32). This is especially true for metropolitan regions such as Paris for example (Deloitte, 2023, p. 19-20). As the impact is mostly measured in percentage values, this leads to different outcomes. Fourth, refurbishment costs are not the same in different EU countries (CBRE, 2021, p. 24). The need to invest into energy efficiency improvements can decrease the value of a property (Högberg, 2013, p. 256). Thus, if these improvements are relatively more or less expensive, this leads to differences in discounts. This might also be an explanation for the large discounts found in the German market compared to other markets. In conclusion, major differences between EU markets exist that might be caused by the implementation details of the EPC or differing local market conditions.

6. Conclusion & outlook

The analysis in this paper sets out to present new evidence that can help answer two overarching questions: How does energy efficiency impact residential real estate economics? What role does the EPC type of a building play? These two questions are considered within the regional scope of the

Rhein-Main Region in Germany. Six different hypotheses are defined and tested. These six hypotheses focus on the impact of energy efficiency of buildings on the three dependent variables cold rent, warm rent and sales price. They are analyzed using two different data samples. The first data sample contains 212 167 observations and is used to test the impact on the cold rent and warm rent, while the second data sample contains 159 573 observations and explores the impact on the sales price. Hedonic price models are defined using various building characteristics ranging from energy efficiency in kWh / (m^2 * a) to building size and availability of elevators. The results of the models show that above-average energy efficient buildings can achieve premiums for cold rents (EPC rating A+ compared to D: 7.0%), warm rents (EPC rating A+ compared to D: 4.6%) and sales prices (EPC rating A+ compared to D: 6.9%). When taking the EPC type into account, these effects become even stronger (EPC rating A+ compared to D: 8.8%, 6.7%, 7.4%, respectively). This does not hold true for a building with a consumption certificate that is for sale. Compared to an average building with a requirement certificate, such a building shows a significant discount. Considering a building that is below-average in energy efficiency, further differences exist. Below-average energy efficient buildings show no continuous discounts for cold rents leading to comparatively higher warm rents. This effect is even stronger for buildings with a consumption certificate. Sales prices for below average buildings with a requirement certificate show large discounts (EPC rating H compared to D: -10.2%). Buildings with a consumption certificate only show limited sales price discounts (EPC rating H compared to D: -4.9%).

The contribution of this paper to the literature is threefold: First, new evidence regarding the existence of energy efficiency premiums in the rent and sales market is presented. The evidence is based on data samples stretching across a time period of 8.5 years. This supports the conclusion that these energy efficiency premiums exist across longer time periods. Secondly, it presents the first detailed evidence of energy efficiency premiums based on EPC types for the German rental market. It shows the importance of occupant behavior regarding operating costs and warm rents. Third, it provides evidence on the difference in perception of EPC types in the sales market. It is shown that requirement certificates are seen as objective while consumption certificates bring only limited value to prospective buyers. Additionally, interesting new research opportunities are identified. Their exploration will help to better understand the heterogeneity of energy efficiency premiums in the future.

The question that remains is how the new evidence fits in with the challenges of the current status quo in the building sector. Two core challenges currently are i) the provision of enough suitable living space via new buildings or refurbishments and ii) getting the overall building stock in line with industry emission targets. While no holistic answer to these questions can be given, implications for different building sector stakeholders can be derived. These include current and prospective owners and tenants, service providers in the building sector and policy makers. The implications are as follows:

- Owners can achieve a higher cash flow by increasing the level of energy efficiency of a building (C or better) while a floor price for cold rents protects them from a potential cash-flow downside for an energy inefficient building (D-H).
- Owners can in general increase the sales price of a building by improving its energy efficiency. The energy efficiency level should be shown with a requirement certificate.
- Owners of highly inefficient buildings issued with a requirement certificate should consider improving their energy efficiency to protect them from large sales price discounts and future downside risk.
- Prospective tenants looking for the most economical choice should primarily consider a building with a requirement certificate that is rated D-G. They should be aware that their consumption behavior is relevant when renting a building with a requirement certificate. This is less the case for a building with a consumption certificate.
- Policy makers should rethink the way the EPC types in Germany are implemented. One suggestion would be to retire the consumption certificate while at the same time adding additional information on the requirement certificate that is indicative of actual consumption (i.e., including a 95% corridor based on average real-world values).
- Service providers in the real estate industry should focus on solutions that can create the most value at scale with the least investment. One suggestion would be to consider G and H rated properties with a requirement certificate and increase their efficiency to a C or D rating. This could present a substantial market opportunity.

The presented implications are formulated based on the evidence found in this paper. It should be noted that the evidence in itself is no proof of causality. The implications represent an academic perspective on real estate valuation but cannot give a general answer regarding the viability of business models or development projects. Thus, as a next step, the evidence should be used by the mentioned stakeholders to develop an overarching building sector transformation strategy. One approach could be along the following lines: First, the profitability analyses of new building projects and refurbishment initiatives should be adjusted by integrating the found values. Next, an indicative ranking should be drafted that shows which of the initiatives is most effective and cost efficient when it comes to increasing energy efficiency across the overall building stock. Finally, the mentioned stakeholders should focus on enabling these initiatives top-down together.

The importance of this transformation being led by a coalition of stakeholders seems to be difficult to overstate. Once the key levers are identified from a physical and construction perspective and their cost efficiency has been determined, it will be necessary to streamline administrative and approval processes and provide funding at a reasonable cost. Additionally, it will be crucial to communicate the importance of these initiatives effectively to the public as they will be directly affected as tenants or owner-occupants.

Together, the described implications and the next steps might support the progress needed to face the monumental challenge of creating energy efficient living space at socially acceptable costs. It seems plausible that energy efficiency will become an even more important aspect of residential real estate valuation in the future. The magnitude of this effect in the long term will be determined by the decisions of the stakeholders made in the next few years. The differentiation is likely to increase should factors such as scarcity of newly built energy efficient buildings and high costs of refurbishments last. Once these factors are improved or the levels of energy costs decrease, differentiation between buildings with different levels of energy efficiency will likely decrease. Further research on how this is achievable within the existing time frame defined by the building sector decarbonization pathway is urgently needed.

References

Amemiya, T. (1980). Selection of Regressors. *International Economic Review*, *21*(2), 331–354.

Ankamah-Yeboah, I., & Rehdanz, K. (2014). Explaining the Variation in the Value of Building Energy Efficiency Certificates: A Quantitative Meta-Analysis.

Bajari, P., & Benkard, L. (2005). Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach. *Journal of Political Economy*, *113*(6), 1239–1276.

Bartik, T. J. (1987). The Estimation of Demand Parameters in Hedonic Price Models. *Journal of Political Economy*, *95*(1), 81–88.

Baum, A. E., & Hartzell, D. J. (2021). *Real Estate Investment Strategies, Structures, Decisions* (second).

Bio Intelligence Service, Lyons, R., & IEEP. (2013). Energy performance certificates in buildings and their impact on transaction prices and rents in selected EU countries. *European Commission (DG Energy)*.

Blomquist, G. C., Berger, M. C., & Hoehn, J. P. (1988). New Estimates of Quality of Life in Urban Areas. *The American Economic Review*, *78*(1), 89–107.

Boverket. (2023). Energy Performance Certificate. Retrieved November 20, 2023, from https://www.boverket.se/en/start/building-in-sweden/contractor/inspection-delivery/energy-performance-certificate/

BRE Group. (2023). How BREEAM works: A guide to BREEAM sustainable building certification and third-party accreditation. Retrieved November 15, 2023, from https://bregroup.com/products/breeam/how-breeam-works/

Brounen, D., & Kok, N. (2011). On the economics of energy labels in the housing market. *Journal of Environmental Economics and Management*, *62*(2), 166–179.

Brown, M. J., & Watkins, T. (2016). The "green premium" for environmentally certified homes: a meta-analysis and exploration. *Oregon Department of Environmental Quality*.

Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen. (2022). Bündnis bezahlbarer Wohnraum, Maßnahmen für eine Bau-, Investitions- und Innovationsoffensive. *Federal Government of Germany*.

Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen. (2023). Muster- Anwendungshinweise zur Immobilienwertermittlungsverordnung (ImmoWertV-Anwendungshinweise – ImmoWertA). Retrieved October 20, 2023, from https://www.bmwsb.bund.de/SharedDocs/downloads/Webs/BMWSB/DE/veroeffentlichungen/wohnen/immowerta.pdf?__blob=publicationFile&v=3

Cajias, M., Fuerst, F., & Bienert, S. (2019). Tearing down the information barrier: the price impacts of energy efficiency ratings for buildings in the German rental market. *Energy Research & Social Science*, *47*, 177–191.

Cajias, M., & Piazolo, D. (2013). Green performs better: energy efficiency and financial return on buildings. *Journal of Corporate Real Estate*, *15*(1), 53–72.

CBRE. (2021). Investor Refurbishment Cost Guide, European Cities. Retrieved November 20, 2023, from https://mktgdocs.cbre.com/2299/f690af3e-8c63-4099-af4d-ba0fdbf876fb-410753402/European_20Investor_20Refurbis.pdf

Cespedes-Lopez, M.-F., Mora-Garcia, R.-T., Perez-Sanchez, V. R., & Perez-Sanchez, J.-C. (2019). Meta-Analysis of Price Premiums in Housing with Energy Performance Certificates (EPC). *Sustainability*, *11*(22), 1–59.

Colwell, P., & Dilmore, G. (1999). Who was first? An Examination of an Early Hedonic Study. *Land Economics*, *75*(4), 620–626.

Court, A. (1939). Hedonic Price Indexes With Automotive Examples. In *The Dynamics of Automobile Demand* (pp. 99–117). General Motors Corporation.

Dalton, B., & Fuerst, F. (2018). The 'green value' proposition in real estate: A meta-analysis. In S. Wilkinson, T. Dixon, N. Miller, & S. Sayce (Eds.), *Routledge Handbook of Sustainable Real Estate* (pp. 1–24). Routledge.

de Ayala, A., Galarraga, I., & Spadaro, J. V. (2016). The price of energy efficiency in the Spanish housing market. *Energy Policy*, *94*, 16–24.

Debus, M. (2022). *Immobilienmarktbericht Frankfurt am Main 2023. Yearly Report. Gutachterausschuss für Immobilienwerte für den Bereich der Stadt Frankfurt am Main.*

Deller, T. A. (2022). The Economic Upside of Green Real Estate Investments: Analyzing the Impact of Energy Efficiency on Building Valuation in the Residential Sector. *Junior Management Science*, *7*(3), 802–825.

Deloitte. (2023). Property Index: Overview of European Residential Markets. Retrieved November 22, 2023, from https://www2.deloitte.com/content/dam/Deloitte/at/Documents/presse/at-deloitte-property-index-2023.pdf

Dinan, T., & Miranowski, J. (1989). Estimating the implicit price of energy efficiency improvements in the residential housing market: A hedonic approach. *Journal of Urban Economics*, *25*(1), 52–67.

Directorate-General for Climate Action. (2019). Going climate-neutral by 2050: a strategic long-term vision for a prosperous, modern, competitive and climate-neutral EU economy. Retrieved November 22, 2023, from https://data.europa.eu/doi/10.2834/02074

Directorate-General for Communication. (2023). Types of legislation. Retrieved November 15, 2023, from https://european-union.europa.eu/institutions-law-budget/law/types-legislation_en

Directorate-General for Energy. (2023). Certificates and inspection: Energy performance certificates provide information on the energy efficiency of buildings and recommended improvements. Retrieved November 15, 2023, from https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/certificates-and-inspections_en

Dulian, M. (2023). Revision of the Energy Performance of Buildings Directive, In "A European Green Deal". Retrieved November 15, 2023, from https://www.europarl.europa.eu/legislative-train/package-fit-for-55/file-revision-of-the-energy-performance-of-buildings-directive

Edlefsen, L. E. (1981). The Comparative Statics of Hedonic Price Functions and Other Nonlinear Constraints. *Econometrica*, *49*(6), 1501–1520.

Epple, D. (1987). Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products. *Journal of Political Economy*, *95*(11), 59–80.

Eurostat. (2023a). Electricity prices for household consumers – bi-annual data (from 2007 onwards). Retrieved November 20, 2023, from https://ec.europa.eu/eurostat/databrowser/view/NRG_PC_204/bookmark/table?lang=en&bookmarkId=ba5340c2-e6bf-40c5-af4a-b0e978ac210c

Eurostat. (2023b). Gross domestic product at market prices. Retrieved November 25, 2023, from https://ec.europa.eu/eurostat/databrowser/view/tec00001/default/table?lang=en

Eurostat. (2023c). Population change - Demographic balance and crude rates at national level. Retrieved November 25, 2023, from https://ec.europa.eu/eurostat/databrowser/view/DEMO_GIND__custom_7127262/default/table

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London.

Fizaine, F., Voye, P., & Baumont, C. (2018). Does the Literature Support a High Willingness to Pay for Green Label Buildings? An Answer with Treatment of Publication Bias. *Revue d'économie politique*, *128*(5), 1013–1046.

Follain, J., & Jimenez, E. (1985). The Demand for Housing Characteristics in Developing Countries. *Urban Studies*, *22*(5), 421–432.

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183.

Fregonara, E., Rolando, D., & Semeraro, P. (2017). Energy performance certificates in the Turin real estate market. *Journal of European Real Estate Research*, *10*(2), 149–169.

Fuerst, F., McAllister, P., Nanda, A., & Wyatt, P. (2016). Energy performance ratings and house prices in Wales: An empirical study. *Energy Policy*, *92*, 20–33.

Fuerst, F., Oikarinen, E., & Harjunen, O. (2016). Green signalling effects in the market for energy-efficient residential buildings. *Applied Energy*, *180*, 560–571.

Galvin, R. (2023). How prebound effects compromise the market premium for energy efficiency in German house sales. *Building Research & Information*, *51*(5), 501–517.

Geske, J. (2022). The value of energy efficiency in residential buildings – a matter of heterogeneity?! *Energy Economics*, *113*, 1–15.

Griliches, Z. (1961). Hedonic Price Indexes for Automobiles: An econometric of quality change (P. S. R. Committee, Ed.), 173–196.

Groh, A., Kuhlwein, H., & Bienert, S. (2022). Does Retrofitting Pay Off? An Analysis of German Multifamily Building Data. *Journal of Sustainable Real Estate*, *14*(1), 95–112.

Haas, G. C. (1922). *A Statistical Analysis of Farm Sales in Blue Earth County, Minnesota, as a Basis for Farm Land Appraisal* [Master Thesis].

Henger, R., & Voigtländer, M. (2014). Transaktions- und Angebotsdaten von Wohnimmobilien – eine Analyse für Hamburg. *IW – Trends*, *41*(4), 1–16.

Hirst, E., & Brown, M. (1990). Closing the efficiency gap: barriers to the efficient use of energy. *Resources, Conservation and Recycling*, *3*(4), 267–281.

Hocking, R. R. (1976). A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, *32*(1), 1–49.

Högberg, L. (2013). The impact of energy performance on single-family home selling prices in Sweden. *Journal of European Real Estate Research*, *6*(3), 242–261.

Hyland, M., Lyons, R., & Lyons, S. (2013). The value of domestic building energy efficiency – evidence from Ireland. *Energy Economics*, *40*, 943–952.

Immowelt. (2023). Mietspiegel in Frankfurt am Main. Retrieved November 25, 2023, from https://www.immowelt.de/immobilienpreise/frankfurt-am-main/mietspiegel

Intergovernmental Panel on Climate Change. (2023). Synthesis Report of the IPCC Sixth Assessment Report (AR6). Retrieved March 20, 2023, from https://report.ipcc.ch/ar6syr/pdf/IPCC_AR6_SYR_SPM.pdf

International Energy Agency and the United Nations Environment Program. (2022). 2022 Global Status Report for Buildings and Construction. Retrieved March 20, 2023, from https://globalabc.org/our-work/tracking-progress-global-status-report

International Valuation Standards Council. (2022). International Valuation Standards (IVS). Retrieved November 14, 2023, from https://viewpoint.pwc.com/dt/gx/en/ivsc/international_valuat/assets/IVS-effective-31-Jan-2022.pdf

International Valuation Standards Council. (2023). Members, Membership of the IVSC. Retrieved November 14, 2023, from https://www.ivsc.org/members/

Jaffe, A., & Stavins, R. (1994). The energy-efficiency gap: What does it mean? *Energy Policy*, *22*(10), 804–810.

Jensen, O. M., Hansen, A. R., & Kragh, J. (2016). Market response to the public display of energy performance rating at property sales. *Energy Policy*, *93*, 229–235.

Johnson, R., & Kaserman, D. (1983). Housing Market Capitalization of Energy-Saving Durable Good Investments. *Economic Inquiry*, *21*(3), 374–386.

Kholodilin, K. A., Mense, A., & Michelsen, C. (2017). The market value of energy efficiency in buildings and the mode of tenure. *Urban Studies*, *54*(14), 3218–3238.

Kim, S., Lim, B. T. H., & Kim, J. (2016). Green features, symbolic values and rental premium: systematic review and meta-analysis. *Conference Paper*.

Knight, J. R. (2002). Listing Price, Time on Market, and Ultimate Selling Price: Causes and Effects of Listing Price Changes. *Real Estate Economics*, *30*(2), 213–237.

Lancaster, K. J. (1966). A New Approach To Consumer Theory. *Journal of Political Economy*, *74*(2), 132–157.

Leamer, E. E. (1978). Regression Selection Strategies and Revealed Priors. *Journal of the American Statistical Association*, *73*(363), 580–587.

Malpezzi, S. (2002). Hedonic Pricing Models: A Selective and Applied Review. In T. O'Sullivan & K. Gibb (Eds.), *Housing Economics and Public Policy* (pp. 67–89).

Marmolejo-Duarte, C., & Chen, A. (2022). The effect of energy performance ratings over residential prices or how an insufficient control of architectural-quality may render spurious conclusions. *Cities*, *126*, 1–15.

März, S., Stelk, I., & Stelzer, F. (2022). Are tenants willing to pay for energy efficiency? Evidence from a small-scale spatial analysis in Germany. *Energy Policy*, *161*, 1–16.

Nevin, R., & Watson, G. (1998). Evidence of rational market valuations for home energy efficiency. *The Appraisal Journal*, *4*(66), 401–409.

Nieskes, J. (2023). Der steinige Weg zum Heizungsgesetz, Chronologie des Ampel-Streits. Retrieved November 25, 2023, from https://www.zdf.de/nachrichten/politik/heizungsgesetz-chronologie-ampel-koalition-einigung-100.html

Pommeranz, C., & Steininger, B. (2021). What Drives the Premium for Energy-Efficient Apartments - Green Awareness or Purchasing Power? *The Journal of Real Estate Finance and Economics*, *62*(2), 220–241.

Regionalverband FrankfurtRheinMain. (2022). Einwohnerzahl der kreisfreien Städte und Landkreise der Metropolregion FrankfurtRhein-Main im Jahr 2021. Retrieved March 14, 2023, from https://service.region-frankfurt.de/ia/metropolregion/bevoelkerung/atlas.html

Roback, J. (1982). Wages, Rents, and the Quality of Life. *Journal of Political Economy*, *90*(6), 1257–1278.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55.

Sheppard, S. (1999). Hedonic analysis of housing markets. In P. Cheshire & E. S. Mills (Eds.), *Handbook of Regional and Urban Economics* (pp. 1595–1635, Vol. 3).

Statistisches Bundesamt. (2018). Preise, Häuserpreisindex. Retrieved November 17, 2023, from https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Preise/haeuserpreisindex.pdf?__blob=publicationFile

Statistisches Bundesamt. (2023a). Preise für Wohnimmobilien im 2. Quartal 2023: -9.9 % zum Vorjahresquartal, Wohnimmobilienpreise sinken nach Höchststand im 2. Quartal 2022 weiter. Retrieved December 5, 2023, from https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/09/PD23_379_61262.html

Statistisches Bundesamt. (2023b). Was ist der Mikrozensus? Retrieved November 17, 2023, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Haushalte-Familien/Methoden/mikrozensus.html

Statistisches Bundesamt. (2023c, May). Presse, 0,6 % mehr neue Wohnungen im Jahr 2022. Retrieved November 17, 2023, from https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/05/PD23_199_31121.html

Taruttis, L., & Weber, C. (2022). Estimating the impact of energy efficiency on housing prices in Germany: Does regional disparity matter? *Energy Economics*, *105*, 1–18.

Umweltbundesamt. (2023). UBA-Prognose: Treibhausgasemissionen sanken 2022 um 1,9 Prozent. Retrieved March 20, 2023, from https://www.umweltbundesamt.de/presse/pressemitteilungen/uba-prognose-treibhausgasemissionen-sanken-2022-um

U.S. Green Building Council. (2023). LEED certification for residential, LEED helps create living spaces where people can thrive. Retrieved November 15, 2023, from https://www.usgbc.org/leed/rating-systems/residential

Verbraucherzentrale NRW e.V. (2023a). Energieausweis: Was sagt dieser Steckbrief für Wohngebäude aus? Retrieved November 16, 2023, from https://www.verbraucherzentrale.de/wissen/energie/energetische-sanierung/energieausweis-was-sagt-dieser-steckbrief-fuer-wohngebaeude-aus-24074

Verbraucherzentrale NRW e.V. (2023b). So kommen Sie an einen Energieausweis für Ihre Immobilie. Retrieved November 15, 2023, from https://www.verbraucherzentrale.de/wissen/energie/energetische-sanierung/so-kommen-sie-an-einen-energieausweis-fuer-ihre-immobilie-24058

Wahlström, M. H. (2016). Doing good but not that well? A dilemma for energy conserving homeowners. *Energy Economics*, *60*, 197–205.

Wallace, H. A. (1926). Comparative Farm-Land Values in Iowa. *The Journal of Land & Public Utility Economics*, *2*(4), 385–392.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838.