JUSTUS LIEBIG UNIVERSITY GIESSEN

DISSERTATION

MEASURING TECHNOLOGICAL CHANGE WITH WEB-BASED DATA AND MACHINE LEARNING

Submitted in fulfilment of the requirements for the degree of

DOCTOR RERUM POLITICARUM (Dr. rer. pol.)

in the

Faculty of Economics and Business Studies
Chair of Economics of Digitalisation

by

Patrick Breithaupt

on

March 8, 2024

Supervisors:

Prof. Dr. Irene Bertschek (first supervisor), Chair of Economics of Digitalisation, Faculty of Economics and Business Studies, Justus Liebig University Giessen.

Prof. Dr. Peter Winker (second supervisor), Chair of Statistics and Econometrics, Faculty of Economics and Business Studies, Justus Liebig University Giessen.

Declaration of Authorship

Ich erkläre hiermit, dass ich die vorgelegten und nachfolgend aufgelisteten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die im jeweiligen Aufsatz angegeben oder zusätzlich in der nachfolgenden Liste aufgeführt sind. In der Zusammenarbeit mit den angeführten Koautoren war ich mindestens anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind, eingehalten.

Ort, Datum	Unterschrift

Table 1: **Contribution table.**

Title	Innovation Indicators Based on Firm Websites – Which Website Characteristics Predict Firm-Level Innovation Activity?
Funding	The paper was written as part of the research project 'TOBI - Text Data Based Output Indicators as Base of a New Innovation Metric' (funding ID: 16IFI001; BMBF).
Co-authors	Janna Axenbeck.
Status	Published.
	Axenbeck, J., & Breithaupt, P. (2021), Innovation Indicators Based on Firm Websites: Which Website Characteristics Predict Firm-Level Innovation Activity?, PLOS ONE Volume 16, Issue 4.
Own key contributions	My contribution is 50%: • I conceptualised the paper jointly with Janna Axenbeck. • I was responsible for the data curation and data analysis jointly with Janna Axenbeck. • I wrote the manuscript jointly with Janna Axenbeck.
Notes	Chapter 2 is a revised version of the PLOS ONE publication.
Title	Measuring Technological Change - A Novel Text Mining Approach Previous title: Measuring the Digitalisation of Firms – A Novel Text Mining Approach
Funding	The paper was written as part of the research project 'Taxation in the Era of Digital Transformation' (funding: Leibniz Association).
Co-authors	Janna Axenbeck.
Status	Working paper.
	Axenbeck, J., & Breithaupt, P. (2022), Measuring the Digitalisation of Firms - A Novel Text Mining Approach, ZEW Discussion Paper No. 22-065, Mannheim.
Own key contributions	My contribution is *70%*: • I conceptualised the paper jointly with Janna Axenbeck. • I was in major parts responsible for the data curation and analysis. • I wrote major parts of the manuscript.
Notes	Chapter 3 is a revised version of the ZEW Discussion Paper No. 22-065.
Title	Intangible Capital Indicators Based on Web Scraping of Social Media
Funding	The paper was written as part of the research project 'INFOWIK - Investments in New Forms of Knowledge-Based Capital' (funding ID: 161FI007; BMBF).
Co-authors	Reinhold Kesler, Thomas Niebel, and Christian Rammer.
Status	Working paper.
	Breithaupt, P., Kesler, R., Niebel, T., & Rammer, C. (2020), Intangible Capital Indicators Based on Web Scraping of Social Media, ZEW Discussion Paper No. 20-046, Mannheim.
Own key contributions	My contribution is 35%: • I conceptualised the paper with Reinhold Kesler, Thomas Niebel, and Christian Rammer. • I was partly responsible for the data curation and analysis. • I wrote the manuscript jointly with Reinhold Kesler, Thomas Niebel, and Christian Rammer.
Notes	Chapter 4 is a revised version of the ZEW Discussion Paper No. 20-046.
Title	Mapping Employee Mobility and Employer Networks using Professional Network Data
Funding	The paper was written as part of the research project 'Networks of Innovative Firms (NETINU)' (funding ID: 16 I FI105 and 16 I FI006; BMBF).
Co-authors	Hanna Hottenrott, Christian Rammer, and Konstantin Römer.
Status	Working paper.
	Breithaupt, P., Hottenrott, H., Rammer, C., & Römer, K. (2023), Mapping Employee Mobility and Employer Networks Using Professional Network Data, ZEW Discussion Paper No. 23-041, Mannheim.
Own key contributions	My contribution is 70%: • I conceptualised the paper with Hanna Hottenrott, Christian Rammer, and Konstantin Römer. • I was in major parts responsible for the data curation and analysis. • I wrote major parts of the manuscript.
Notes	Chapter 5 is a revised version of the ZEW Discussion Paper No. 23-041.

Acknowledgements

Special thanks are owed to my supervisors Prof. Dr. Irene Bertschek and Prof. Dr. Peter Winker. Without their expertise and continuous guidance over the last few years, this dissertation project would not have been completed. I would also like to thank the ZEW – Leibniz Centre for European Economic Research and the Justus Liebig University Giessen for providing an inspiring research environment.

I am very grateful for the great colleagues and co-authors who have accompanied and supported me along the way. Janna, thank you for all the discussions about text mining, machine learning, and digital economics. I learned a lot about economic research from you in our joint projects. Thomas, thank you for your continuous feedback and fruitful collaboration on a variety of research and policy advice projects. Many thanks also go to my co-authors Hanna, Christian, and Sandra of the ZEW department Economics of Innovation and Industrial Dynamics. In our collaborations, I was able to broaden my perspective and learn about neighbouring research areas. Furthermore, I would like to thank Dominik, Daniel, Robin, and the ZEW department Digital Economy for many discussions about our research and projects. My appreciation also goes to the IT department of ZEW, which provided me with the technical infrastructure needed to collect and process large data sets.

Finally, I would like to thank my family and friends for their support and understanding during this time. Particular appreciation goes to my fiancée Louisa, my parents Werner and Nicole, my siblings, and my godchild Edda.

Preface

This dissertation consists of four essays and was written between July 2019 and March 2024. During that time, I was employed as a researcher at ZEW – Leibniz Centre for European Economic Research. The dissertation is submitted in partial fulfilment of the requirements for the academic degree of *doctor rerum politicarum* (Dr. rer. pol.) at the Justus Liebig University Giessen, where I was an external doctoral candidate at the Chair of Economics of Digitalisation.

Chapter 1 is an introduction to the dissertation. In this part, the four essays get embedded in the economic, machine learning, and data mining literature. Furthermore, I provide definitions and present research gaps that motivate the essays.

The first two essays introduce approaches for measuring technological change in firms. In both essays, we use firm website data and machine learning techniques.

Chapter 2 (Essay 1) is concerned with the measurement of innovation, one of the main drivers of economic growth. In co-authorship with Janna Axenbeck, I use natural language processing and machine learning techniques to analyse firm websites. We find that firm websites contain information that can be used to detect and measure innovation activity. Compared to traditional survey-based measures, our model is cost-effective, can be updated quickly, and is available for a large number of German firms. My relative contribution to the essay is 50%.

Chapter 3 (Essay 2) deals with measuring the case of digital technologies as one prominent example of general purpose technologies (GPTs). Co-authored with Janna Axenbeck, I use natural language processing and machine learning techniques to analyse newspaper articles as well as firm websites to estimate a firm-level digitalisation score. We present a methodology for producing a digitalisation indicator that is in line with several traditional measures. Finally, we analyse the link of this indicator with firm resilience during the COVID-19 crisis as an example of a policy-relevant application. My relative contribution to the essay is 70%.

In the last two essays, we measure intangible and human capital that is linked to technological change. In both essays, we use data on firms from digital platforms.

Chapter 4 (Essay 3) focuses on firm-level indicators for intangible capital. In co-authorship with Reinhold Kesler, Thomas Niebel, and Christian Rammer, I use data from the platforms Facebook and Kununu. By using this public data, we introduce two indicators for marketing and on-the-job training that are positively related to

corresponding data from the Community Innovation Survey (CIS). The proposed approach is more cost-effective and can be updated more often than a survey-based measure. My relative contribution to the essay is 35%.

Chapter 5 (Essay 4) is concerned with measuring human capital. Co-authored with Hanna Hottenrott, Christian Rammer, and Konstantin Römer, I create a Linked Employer-Employee (LEE) data set using public data from the career-oriented social networking platform XING. The data contain information on employers, employees, and employee flows. By using independent data, we show that the XING data are plausible, can be used for subsequent economic research, and represent a viable alternative to official data. My relative contribution to the essay is 70%.

Chapter 6 summarises the findings of the four essays, provides concluding remarks, and points out possible directions for future research.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ARGUS	Automated Robot for Generic Universal Scraping
AUC	Area Under the Curve
CIS	Community Innovation Survey
COVID-19	Coronavirus Disease 2019
DACH	Germany (D), Austria (A), and Switzerland (CH)
DESI	Digital Economy and Society Index
Destatis	German Federal Statistical Office (Statistisches Bundesamt)
ERP	Enterprise Resource Planning
EU	European Union
Eurostat	Statistical Office of the European Communities
FE	Fixed Effects
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GFLOPS	Giga Floating Point Operations per Second
GPT	General Purpose Technology
HTML	Hypertext Markup Language
IAB	Institute for Employment Research
ICT	Information and Communications Technologies
IT	Information Technologies
INFOWIK	Investments in New Forms of Knowledge-Based Capital
LDA	Latent Dirichlet Allocation
KNN	K-Nearest Neighbour
LEE	Linked Employer-Employee
ln	Natural Logarithm
MAE	Mean Absolute Error
Mbit/s	Megabits per Second
MDI	Mean Decrease in Impurity
MEUR	Million Euros
MIP	Mannheim Innovation Panel

ML	Machine Learning
MUP	Mannheim Enterprise Panel
MWP	Mannheim Web Panel
NACE	Statistical Classification of Economic Activities in the European Community
NETINU	Networks of Innovative Firms
NLP	Natural Language Processing
NN	Neural Network
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PATSTAT	Worldwide Patent Statistical Database
PC	Personal Computer
R&D	Research and Development
RF	Random Forest
ROC	Receiver Operating Characteristic
SEO	Search Engine Optimisation
SME	Small and Medium-sized Enterprise
SNA	Social Network Analysis
SOEP	Socio-Economic Panel
S&P 500	Standard & Poor's 500
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TN	True Negative
TOBI	Text Data Based Output Indicators as Base of a New Innovation Metric
TP	True Positive
TPR	True Positive Rate
UK	United Kingdom
UrhG	Act on Copyright and Related Rights (Urheberrechtsgesetz)
URL	Uniform Resource Locator
US	United States
USA	United States of America
VPN	Virtual Private Network
ZEW	ZEW – Leibniz Centre for European Economic Research

Chapter 1

Introduction

1.1 Measuring Technological Change

Technological change is a key factor of economic growth (Solow 1956, 1957). Therefore, effective measuring and tracking methods are of great economic and political importance. However, traditional measurement approaches have several disadvantages, such as high costs, low case numbers, and time lags. In this dissertation, the existing measurement approaches are enhanced by new methods and data.

Key factors of technological change include *inventions*, *innovations*, and their *diffusion* (Jaffe et al. 2002). Inventions such as the printing press, steam engines, telephones, computers, and the World Wide Web replaced existing technologies or introduced new ones.[1] Innovations represent an important part of technological change as well. They are defined as the implementation of either new or significantly improved products or processes, as well as combinations thereof (OECD/Eurostat 2018).[2] Recent examples of innovations are SpaceX's reusable rockets and personalised movie recommendations by the streaming platform Netflix.

In this dissertation, I focus on technological change within the boundaries of corporate firms. In this case, measuring technological change is particularly difficult, as most firms do not allow certain internal information to be accessed by the public, e.g. to hide innovations or trade secrets from competitors. Traditional measurement approaches tackle this problem via firm surveys (e.g. Mohnen 2019), which usually guarantee anonymity for the firms. A large disadvantage of surveys is the limited number of questions that can be asked in questionnaires. Further, the observation count is usually low due to the associated costs and low willingness to participate in the survey. In contrast, with the increase in computing power and the amount

[1] General Purpose Technologies (GPTs) are closely related to technological change (Petralia 2020) and are "characterized by pervasiveness, inherent potential for technological improvements, and 'innovational complementaries'" (Bresnahan & Trajtenberg 1995, p.83). Electricity, computers, and the World Wide Web are prime examples of such technologies.

[2] Other definitions are more distinctive as they consider, for example, innovations at the organisational level as well (e.g. Johannessen 2008).

of public data on firms, novel and more efficient methods of measurement have emerged in recent decades.

Technological advancements have not stopped with personal computers and their computing power. Until the end of the 1990s, Intel Pentium processors provided less than 3 GFLOPS.[3] In comparison, the Intel Core i9-12900K processor released in 2021 provides up to 819 GFLOPS[4]. The number of FLOPS[5], used as an approximation for computing power, has therefore considerably increased in this period. Research into quantum computers promises the continued increase of computing power (at least for certain computing tasks) in the future. In addition to higher computing power, there has also been progress in other areas, such as memory and storage. This is a necessity, as the amount of data available worldwide has multiplied over the last few decades.[6] The technological progress of personal computers and servers as well as their on-demand availability, e.g. storage and computing on Amazon Web Services or high-performance computing clusters that are dedicated to research and teaching[7], has enabled the storage of large data sets and the use of computationally intensive methods to measure technological change. This would not have been possible without advances in computing power, memory, and storage.

In addition to technological advancements, there have also been major methodological advances. These include research into natural language processing (NLP) as well as machine learning (ML). The number of publications on these topics has increased drastically in recent years (e.g. Rahal et al. 2022, Su et al. 2023), approximating the rising importance of these topics for various stakeholders, such as scientists, governments, and firms. Besides NLP, the main focus of this dissertation is on the area of supervised machine learning (Hastie et al. 2001). This field involves learning from existing labelled data to be able to make a prediction for unseen data. Progress has also been made in recent years, particularly in the fields of ensemble methods and artificial neural networks (e.g. Breiman et al. 1984, LeCun et al. 2015), which are now able to complete a large number of routine tasks very quickly and with high accuracy. Some well-known examples of these are *AlphaGo*, *DeepL*, and *ChatGPT*. The possible applications are immense, e.g. processing of publicly accessible data

[3]Intel Pentium processors: `https://www.intel.de/content/www/de/de/support/articles/000007250/processors.html` [Last accessed: 24.02.2024].

[4]Intel i9 processors: `https://www.intel.com/content/dam/support/us/en/documents/processors/APP-for-Intel-Core-Processors.pdf` [Last accessed: 24.02.2024].

[5]Floating-point operations per second (FLOPS) is a performance measure of processors referring to the number of floating-point operations (i.e. additions or multiplications) that can be performed per second. One GFLOPS corresponds to one billion floating-point operations per second.

[6]`https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/` [Last accessed: 24.02.2024]. However, the estimates for the data volume should be used with caution.

[7]For example: "bwUniCluster 2.0" by the German federal state of Baden-Württemberg.

(in near real-time) to gain insights into firms' technological advancements.

The increase in the amount of data, technological advancements, and advances in NLP and ML methods enable the use of novel approaches to measure technological change and related indicators. My thesis makes use of the rise in publicly available data sources. In particular, this includes data from firm websites and digital platforms. These data sets have the great advantage that they usually contain information on a large number of firms. The published information is provided by the firm or its employees and offers an insight into the firm for people on the outside. Obtaining public data on firms is the bottleneck in this type of research. The data are often collected using web scraping[8] or sometimes based on industry cooperations, e.g. with a platform provider. Platform data can be very diverse: The data sets range from marketing, e.g. Google and Meta advertisements, to social networks for employees, e.g. XING and LinkedIn, to feedback platforms, e.g. Kununu and Glassdoor. Using this data also introduces some new challenges: for example, their unclear authenticity and the necessary linking of public data to standardised firm databases. Linking the firm data with other data sets is a prerequisite for gaining access to additional firm attributes.

In short, I propose the approach shown in Equation 1.1 for the measurement of firm-level technological change. The input to the function f is publicly available data from online sources on firm i.

$$y_i = f(\text{public data}_i) \tag{1.1}$$

The indicator y_i is calculated using the function f, which ranges from simple mathematical operations to complex machine learning. In some cases, the public data can be supplemented with proprietary data, e.g. from firm databases, to improve the performance of the function f. The web-based approach has important advantages. Unlike traditional measures, the indicator can be updated for many firms frequently, on demand, and at a low cost. However, the resulting web indicators must always be evaluated with some caution. They should be compared with at least one established measure to assess their quality and external validity.

The proposed class of indicators can be used to measure technological change. This class includes, for example, different types of innovations and the digitalisation of firms. In addition, related variables that are often prerequisites for technological change can also be captured. These variables include intangible assets, such as human capital, on-the-job training, and marketing expenditures. The transfer of human capital (e.g. employees switching employers) and on-the-job training, for instance,

[8] Researchers have special rights for scraping web data and text/data mining (e.g. UrhG §60d). The Mannheim Web Panel (MWP) has been collected at ZEW at regular intervals since 2018. Among other things, the data set contains the website texts from a large number of German firms (`https://kooperationen.zew.de/en/zew-fdz/provided-data/mannheim-webpanel` [Last accessed: 24.02.2024]).

foster technology diffusion. Existing contributions in the literature use similar data, e.g. for websites (Choi & Varian 2012, Lenz & Winker 2020, Kinne & Lenz 2021), and platforms (Glaeser et al. 2018, Bertschek & Kesler 2022).

Further, the value of web data and indicators has increased for researchers, firms, and government institutions in recent years. The data can not only be used as a substitute but also as a complement to existing approaches and, thus, provides value for economic and political decision-making and for shedding light on poorly-researched areas. The availability of large-scale firm-level web data enables a variety of research and policy advice (e.g. Dörr et al. 2022, Axenbeck et al. 2023).

1.2 Contribution

Table 1 lists the four essays, my co-authors, the publication status, and my relative contribution. The remainder of the thesis is structured as follows[9]: In Chapter 2, the essay "Innovation Indicators Based on Firm Websites – Which Website Characteristics Predict Firm-Level Innovation Activity?" is presented. My co-authors and I contribute to the discussion on whether web-based innovation indicators are a feasible alternative to survey-based innovation indicators. Our results show that website characteristics provide valuable information about firm-level innovation activities. The novel indicators can be quickly updated, are on a very granular level, and are less expensive than questionnaire-based surveys. Chapter 3 contains the essay "Measuring Technological Change - A Novel Text Mining Approach". Our contribution to the literature employs an up-to-date, low-cost, and quality-tested methodology to measure digital technology adoption, covering all firm-size classes, regions, and economic sectors in Germany. The method covers a broad definition of digital technologies and requires no (or only very little) human assessment of what digitalisation is. Chapter 4 contains the essay "Intangible Capital Indicators Based on Web Scraping of Social Media". My co-authors and I contribute to the literature in two ways. First, we develop a method for matching and linking firm-level survey data with platform-based data. Second, using publicly available data from social media, we derive new indicators of firm-specific human capital and brand equity that can complement firm surveys, thus improving the measurement of knowledge-based capital. Chapter 5 presents the essay "Mapping Employee Mobility and Employer Networks using Professional Network Data". The paper contributes to recent developments in web-based indicators by analysing a little-explored source of large-scale data on employee and employer activities. By doing so, we contribute to the *Linked*

[9]The remainder of the paragraph is partly based on the essays.

Employer-Employee literature, i.e. by combining platform and proprietary data. Our LEE data have the advantage of using up-to-date and public data. In Chapter 6, I summarise and discuss our findings and provide directions for future research.

Chapter 2

Innovation Indicators Based on Firm Websites – Which Website Characteristics Predict Firm-Level Innovation Activity?

Joint work with Janna Axenbeck.

2.1 Introduction

Innovation, defined as the implementation of either new or significantly improved products or processes, as well as combinations thereof (OECD/Eurostat 2018), brings vast benefits to consumers and businesses. Moreover, technological progress is considered a main driver of economic growth (Solow 1957). It is therefore a matter of public interest to analyse and understand innovation dynamics, as conducted in several studies (e.g. Crepon et al. 1998, Klomp & Van Leeuwen 2001, Belderbos et al. 2004, Hall et al. 2005, Griffith et al. 2006, Frenz & Ietto-Gillies 2009, Kogan et al. 2017).

A prerequisite for the analysis of innovation-related questions is to correctly measure firm-level innovation activities. However, it should be noted that no universally accepted measurement approach exists. For example, firm-level innovation indicators are traditionally constructed with data from large-scale questionnaire-based surveys like the biennial European Community Innovation Survey (CIS) or the annual Mannheim Innovation Panel (MIP) (see Peters & Rammer 2013, Rammer et al. 2019), which is the German contribution to the CIS. However, these innovation indicators suffer from some major drawbacks (e.g. Mairesse & Mohnen 2010, Pukelis & Stanciauskas 2019, Kinne & Axenbeck 2020). For instance, the MIP covers around 18,000 firms annually, which correspond to only a fractional share of the total number of German firms. As a result, the survey may lack regional granularity and comprehensive coverage. In addition, questionnaire-based surveys, especially on a large

scale, have the added disadvantage of being costly and lacking timeliness. Also, most surveys require firm participation, and as a consequence, surveys such as the MIP suffer from low response rates (Mairesse & Mohnen 2010). Besides, firm-level innovation can also be studied through patent or publication analysis. However, respective indicators only cover technological progress for which legal protection is sought (Archibugi & Planta 1996, Arundel & Kabla 1998) and not every innovation can be patented. For example, due to the German regulatory framework, it is quite difficult to patent software, i.e. digital innovations.

Issues, however, could be solved by adding web-based data. Advances in computing power, statistical learning methods, and natural language processing tools enable the extraction of website information on a large scale. Through this, it is technically possible to complement traditional innovation indicators with information from scraped firm websites. Nowadays, almost every firm has an online presence. Firm websites can include information about new products, key personnel decisions, firm strategies, and relationships with other firms (Gök et al. 2015). Those pieces of information might be directly or indirectly related to a firm's innovation status. By using this information, it is possible to conduct an automatic, timely, and comprehensive analysis of firm-level innovation activities, as measurements can be carried out faster and at shorter intervals in comparison to traditional indicators.

The contribution of this paper to the question of whether web-based innovation indicators are feasible is threefold. First, we analyse to what extent firm websites improve predictions of firm-level innovation activity. Second, we assess which characteristics of a website relate most to a firm's innovation status. Third, we examine which characteristics are appropriate for predicting different forms of innovation activity. We test the latter by additionally comparing the predictive power of different innovation indicators related either to *product innovations*, *process innovations*, or *innovation expenditures*. We assume differences between indicators, for example, because firms with process innovations may have a smaller incentive to announce their innovation activities on the websites. The reason for this may be that new processes are less relevant for most website visitors.

For our analysis, data on 4,487 German firms from the MIP 2019 are used. We extract their websites' text and hyperlink structure by applying the ARGUS webscraper (Kinne & Axenbeck 2020). Several methods, including topic modelling and natural language processing tools, are applied to generate features that potentially relate to the firm-level innovation status. Furthermore, we extract information related to a website's technical maturity, such as how fast it is responding and whether a version for mobile end-user devices is available. After extracting and calculating a wide variety of features, we divide them into three feature sets: I) *text-based features*, including, e.g. words, document-topic probabilities derived from a topic modelling

algorithm, and the share of the English language, II) *meta-information features*, including, e.g. website size-related features, availability of a mobile version, and loading time, and III) *network features*, including, e.g. hyperlinks to social networks, as well as incoming and outgoing hyperlinks. Based on these three feature sets, we analyse which website characteristics best predict a firm's innovation status reported in the MIP 2019 by using a *random forest* classifier.

Our results show that predictions based on website characteristics can perform significantly better than a random prediction based on the sample mean. Consequently, firm websites include information that relates to firm-level innovation activity. In addition, our website characteristics better predict firms with product innovations and innovation expenditures than with process innovations. Moreover, text features make the biggest contribution to our prediction performance.

Evaluating the predictive power of single variables across feature sets using the mean decrease in impurity (MDI) reveals that the language and size of websites, measured by the number of subpages as well as the total number of characters, are always relevant in the models with the highest predictive power for all considered innovation indicators. Moreover, some characteristics are highly important only for specific indicators; e.g. the verb *'to develop'* is more important for innovation expenditures and product innovators than for process innovators.

The remainder of this paper is structured as follows: Previous literature is reviewed in Section 2.2. In Section 2.3, we present our data, and in Section 2.4 the descriptive statistics. Section 2.5 describes the methodology, and Section 2.6 shows the results, which are discussed in Section 2.7. This paper concludes in Section 2.8.

2.2 Literature Review

The use of text data to generate innovation-related indicators has already been tested in previous studies. For example, Kelly et al. (2021) show that the significance, i.e. relevance, of a patent is higher when its textual content is very distinct from previous patents but similar to subsequent ones. Lenz & Winker (2020) generate innovation-related topics from 170,000 technology news articles using a Paragraph Vector Topic Model. They analyse the diffusion of the identified topics within the text corpus. Their results suggest that technology trends can be assessed by measuring the importance of topics over time. Using PATSTAT data, Tacchella et al. (2020) show that the contextual similarity of technological codes relates to innovative events. The probability that new combinations of technological codes appear in one patent can be predicted by their contextual similarity in patents where they have been used before.

Remarkable work is also conducted by Bellstam et al. (2020). In this study, a Latent Dirichlet Allocation (LDA) model is fitted with analyst reports of firms included in the S&P 500 index. The LDA topic that has the lowest Kullback-Leibler divergence from the wording of a mainstream economic textbook on innovation is chosen as an innovation indicator. The authors show that firms have patents with greater impact (i.e. more citations per patent) if the innovation topic has a larger share in their analyst report. However, analyst (or annual) reports are not available for every firm, and smaller firms are particularly underrepresented. In contrast, firm websites are available for a large share of small and medium-sized firms.

Furthermore, previous literature shows that information produced online can be used to construct frequent real-time estimates (Gentzkow et al. 2019). Famous *now-casting* examples that utilise web-based information are: Ginsberg et al. (2009), who use Google search queries to accurately predict influenza activity in the United States (US). Choi & Varian (2012) claim that search engine query indices are often correlated with economic activities and enable the generation of frequent indicators. They show that forecasts concerning, for example, automobile sales and unemployment can be significantly improved by including search term indices in prediction models.

Not only information from online searches but also firm website information can be used to generate economic indicators. As firm websites provide detailed information about the firm as well as its products, they appear to be suitable for measuring firm-level innovation activities (Gök et al. 2015). Kinne & Axenbeck (2020) summarise previous studies that analyse the possibility of firm website-based innovation indicators (e.g. Katz & Cothey 2006, Ackland et al. 2010, Arora et al. 2013, Gök et al. 2015, Rietsch et al. 2016, Nathan & Rosso 2022). Most studies solely focus on the hyperlink structure of websites, only conduct a simple keyword search, or are limited to small numbers of firms from a particular economic sector.

Firstly applying advances in statistical learning, Kinne & Lenz (2021) attempt to predict innovation at the firm level using textual information on websites and novel machine learning tools. They use a questionnaire-based firm-level product innovation indicator (innovative/non-innovative) from the MIP (covering the time period from 2015 until 2017) as a target variable to train an artificial neural network classification model based on website texts. The authors only consider stable product innovators in their main analysis. Firms that switch between innovation statuses, which is a phenomenon that is highly relevant in the field of innovation economics, are only observed in a secondary analysis. The average F1-score for the respective prediction is 0.68%. Moreover, Pukelis & Stanciauskas (2019) fit several machine learning models to develop a firm website-based innovation indicator, with their annotated data set being limited to 500 firms. One important characteristic of their

work is the individual analysis of websites' subpages instead of predicting the innovation status of an entire website, i.e. a firm. Additionally, their subpages are manually labelled as either innovation- or non-innovation-related messages instead of using survey or patent data as target variables. The best performance is achieved with an artificial neural network. Even though the predictive performance is very high, the authors cannot show the external validity of their indicator. Furthermore, another issue with both approaches is that neural networks do not reveal any decision rules that can be easily interpreted by humans, which is why they are often called *black box* models. Both studies only consider text, but previous results show that there must be distinct website characteristics that relate to a firm's innovation status. However, the particular website characteristics have not been identified yet.

Gandin & Cozza (2019) analyse whether firms' expenditures on innovation can be predicted using administrative records and balance sheet data. Applying a random forest regression approach, the authors identified firm size, industry affiliation, and investment in intangible assets as the most important predictors. Random forests usually provide better predictive performance than linear methods while retaining the interpretability of feature relevance.

By applying a random forest approach to large-scale firm-level web data, we analyse which website characteristics are linked to firms' innovation activity.

2.3 Data

Based on the Oslo Manual, we define an innovation as "a new or improved product or process (or combination thereof) that differs significantly from the unit's previous products or processes and that has been made available to potential users (product) or brought into use by the unit (process)" (OECD/Eurostat 2018, p. 20). Furthermore, we consider all expenditures spent for innovation purposes as innovation expenditures and summarise firm-level product or process innovation as well as innovation expenditures as innovation activity.

We use data from the MIP 2019 to classify firms as either innovative or non-innovative.[10] The MIP is an annual survey conducted by the ZEW – Leibniz Centre for European Economic Research. The survey covers firms from the manufacturing and service sectors and is conducted as a mail survey with the option to respond online.

In the MIP 2019, firms were asked whether they introduced a product or process innovation within the last three years (between 2016 and 2018) and for the total amount spent on innovation activities in the last year (2018). We consider a firm that stated that it introduced a product innovation within the considered time frame as

[10] See ZEW (2024*a*,*b*).

a product innovator and a firm that stated that it introduced a process innovation within the considered time frame as a process innovator. A firm is an innovator if it introduced at least one of both. Every firm that spent financial resources on innovation, independent of the magnitude, is regarded as a firm with innovation expenditures. Our initial sample consists of 13,747 firms from the MIP 2019. We merge these firms with the Mannheim Enterprise Panel (MUP), which consists of more than 3.2 million economically active firms (Bersch et al. 2014), to receive information about the firms' website addresses. The MUP serves as a sampling frame for surveys like the MIP and contains, for example, firm-level information on turnover, number of employees, and industry affiliation. Only 54% of the firms in our sample can be linked to website addresses, as we limit ourselves to quality-assured observations. In total, we have 6,368 firms with information on their website addresses and at least one innovation indicator. We extract website content by applying the ARGUS web-scraper, which allows us to collect texts as well as hyperlinks to other websites.[11] Firm websites were first scraped in September 2018 to collect texts, then again in January 2019 for adding hyperlinks. We scraped a third time in October 2019 to add information about technical features, e.g. information on the existence of firm websites for mobile end-user devices. The number of scraped subpages per website is limited to a maximum of 50. We consider this to be a sufficient number, as the median number of subpages in the MUP is 15 (see Kinne & Axenbeck 2020), and only 1.5% of all firms in our subsample have 50 or more subpages. Since only a few firms exceed the subpage limit, we assume this bias to be negligible. Moreover, the scraping programme is set to prefer subpages with shorter website addresses because we assume these subpages include more important information about the firm. Also, ARGUS is set to prefer websites in the German language. Hence, when we calculate the share of different languages on a website, we expect a small bias. While scraping the data, especially while collecting meta-information features, we received several error messages. Furthermore, we only use observations for which all features are not missing. If, for example, a meta-information feature is not available, the observation will not be used for training or testing with other feature sets as well. Therefore, after the entire data collection process, we have a sample of 4,487 firms for predicting product innovators and innovators, and 4,484 firms for predicting process innovators.[12] For predicting whether a firm has innovation expenditures, the sample size is 1,893 (Table 2.1).

Additionally, a random sample of approximately 32,000 website addresses of firms that are not included in the MIP is drawn from the MUP and scraped with the ARGUS web-scraper using the same settings as for the MIP sample. The sample

[11] For a detailed description of the web scraper, see Kinne & Axenbeck (2020) and Kinne (2018).

[12] There are three more observations for product innovators than for process innovators. Since we know that these firms have at least a product innovation, we consider them to be innovators.

Table 2.1: **Summary statistics for product innovators, process innovators, innovators, as well as firms with innovation expenditures.**

Variable	Definition	N	Mean	SD	Min	Max
Product innovators	1: If a firm is a product innovator 0: Otherwise	4,487	0.39	0.49	0	1
Process innovators	1: If a firm is a process innovator 0: Otherwise	4,484	0.52	0.50	0	1
Innovators	1: If a firm is a product or / and process innovator 0: Otherwise	4,487	0.61	0.49	0	1
Innovation expenditures	1: If firm innovation expenditures were reported 0: Otherwise	1,893	0.39	0.49	0	1

is used for topic modelling. We train a topic model on a separate sample for two reasons. First, it allows the inclusion of more data points. Second, it ensures that no observation used for calibrating topics is considered for evaluating the random forest models. Hence, it prevents data leakage. In the following, the data are referred to as the *LDA sample*.

As we need to exclude a large share of observations due to missing values in our MIP sample, we cannot rule out a selection bias. Also, firms from certain industries and smaller firms are less likely to have a website and may therefore be underrepresented. In machine learning, adverse selection might lead to two issues: It could cause our model to be better fitted for groups that are overrepresented in our sample, and it could induce that the class correlated with the overrepresented group is predicted more often. To identify whether a potential selection bias exists, we analyse how the sample distribution changes with respect to the number of employees and industries when excluding observations with missing information (see Figure A.1 and Figure A.2). Except for *transportation and post* (industry 15), we do not see a notable change in the distribution of firms that could be linked to a severe selection bias.

To capture website characteristics, we apply several methods to create features, like keyword search and natural language processing, as well as an analysis of hyperlinks (network analysis methods). We use Python as a programming language for calculating our features and for training our random forest models. For an overview of feature sets, see Table 2.2.

Table 2.2: **Features related to text, meta-information, and network measures.**

Text-based features	
1) Textual content	The term-document matrix with the 5,000 most frequent words (TF-IDF weighting is applied).
2) Emerging technologies	A dummy variable that measures whether a technology from Wikipedia's list of emerging technologies appears on a firm's website.
3) Latent patterns	The topic-document probabilities of 150 topics created by the LDA approach.
4) Topic popularity index	The sum of LDA topic probabilities per document. Each probability is weighted with the relative frequency of its appearance in the entire LDA sample.
5) International orientation	The share of subpages in the English language and the share of all other non-German subpages in all subpages.
6) Share of numbers	The share of numbers in the text of a website (measured in characters).
7) Flesch-reading-ease score	A numerical metric assessing the readability of texts.
Meta-information features	
8) Website size	The number of subpages on a website and the total number of characters on a website.
9) Loading time	The time from sending a request (http/https) to get the start page of a website until the arrival of the response (in ms).
10) Mobile version	A dummy variable that is one if a version for mobile end-user devices exists and zero otherwise.
11) Domain purchase year	The year of the first occurrence on web.archive.org.
Network features	
12) Centrality	The total number of incoming and outgoing hyperlinks, as well as the PageRank centrality.
13) Social media	The number of hyperlinks to Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub, Flickr, and Vimeo.
14) Bridges	The number of bridges a firm is part of in the hyperlink network.

2.3.1 Text-based Features

First, information from website texts is analysed (see Table 2.2) as it might be related to a firm's innovation status for the following reasons: Presumably, most firms are using their websites to inform customers about new products or services and might mention whether their product is new or innovative, i.e. it is likely that innovative firms use particular innovation-related words. Information about process innovations can also be detected and used if reported on the website.

Moreover, a firm might report that it uses a recently emerging technology like blockchain, 3D printing, or augmented reality (for an overview of recently emerging technologies, see Appendix A.2). Hence, an emerging technology term might appear

on a firm's website, and if so, it is likely that the firm can be considered innovative as it makes use of fairly new technologies.

Additionally, there might be latent patterns on a website that reveal a firm's innovation status; these latent patterns can be captured by the LDA topic modelling approach, as shown in Bellstam et al. (2020). Furthermore, innovative firms might follow some general technological trends, like digital transformation. As these technological trends are quite general, LDA topics related to these trends might appear quite often on firm websites. To capture this, we construct a topic popularity index that indicates the distribution of popular and less popular topics on a website.

We additionally analyse the following text-based metrics: Languages that are used on a website might relate to the export status of a firm and could provide information about a firm's innovation status (e.g. Lachenmaier & Wößmann 2006, Kirbach & Schmiedeberg 2008, Cassiman & Golovko 2011). Also, we test whether the share of numbers in all text strings and the text complexity, measured by the Flesch-reading-ease score (Flesch 1948), differ between innovative and non-innovative firms.

2.3.2 Meta Information Features

Second, the meta-information of firm websites (see Table 2.2) might allow us to distinguish innovative from non-innovative firms. For example, the size of the website might help to predict a firm's innovation status. Large firms are more likely to be innovative (Rammer et al. 2019). As the number of subpages of a website correlates with the number of employees of a firm (Kinne & Axenbeck 2020), the size of a website might provide information about whether a firm introduced an innovation. Also, the technological properties of a website could be relevant. Innovative firms might have better technical knowledge and can apply more technologically advanced features to their websites, e.g. dynamic elements. For example, the loading time of a website could be faster, and a mobile version might be more often available when firms are more technologically advanced. However, there might be some noise because the loading time may also be short if the website is relatively simple.

Another potentially relevant feature is the age of a website, i.e. the domain purchase year, as it might relate to the actual firm age.[13] One has to consider, however, that this relationship is unlikely to be linear. On the one hand, a fairly new website might indicate a start-up with an innovative idea. On the other hand, having a very old website means the firm has adopted this new technology very early. This could also relate to a more technologically advanced, hence innovative firm.

[13] We approximate a website's domain purchase year by the year of the first entry on web.archive.org.

2.3.3 Network Features

Third, hyperlinks between websites (see Table 2.2) might also help to identify the firm-level innovation status. Firms that have more business relationships with other firms or are more relevant, according to centrality measures, might be better informed and know earlier about new profitable applications. Hence, firms with more relationships with other firms could be more likely to innovate. Moreover, innovation projects are often realised in cooperation with other firms (e.g. Becker & Dietz 2004). Thus, patterns in firm-level cooperation are expected to be of interest. A firm that connects (or bridges) different network parts is usually relevant, and its removal will decompose the network. Lastly, Bertschek & Kesler (2022) show that a firm's use of the social network Facebook is linked to product innovations. Hence, the use of social media might reveal information about a firm's innovation status as well.

Our study analyses whether the three groups of features differ in their performance when predicting a firm's innovation status. A more detailed description of feature generation can be found in Appendix A.3.

2.4 Descriptive Analysis

The descriptive statistics for our predictor variables are presented in this section. Table 2.3 shows mean values and p-values for innovative and non-innovative firms, obtained from a t-test, regarding the difference of both means for selected features.

Differences exist for most variables. Looking at *text* features, innovative firms are more likely to mention an emerging technology term and have more subpages in the English language. The share of subpages in other languages, however, does not show any significant difference between both groups. Differences are also small for the share of numbers, our topic popularity index, and the Flesch-reading-ease score, but the deviation is statistically significant for some forms of innovation activity.

The descriptive statistics for *meta* features show that innovative firms have larger websites with respect to the number of subpages as well as with respect to the number of characters. The loading time is slightly faster for process innovators and innovators, but not for product innovators and firms with innovation expenditures. However, the differences are not statistically significant. The first occurrence on web.archive.org is significantly later for non-innovative firms. This indicates that their domain purchase year, i.e. website age, is slightly lower. Additionally, non-innovative firms have less often a version of their website for mobile end-user devices. Looking at *network* features, significant differences also exist for outgoing and incoming hyperlinks, as well as for hyperlinks to social media websites. Innovative firms have, on average, more hyperlinks. Moreover, the difference is greater

Table 2.3: **Descriptive statistics for selected variables.**

	Group-specific means											
	Product innovator			Process innovator			Innovator			Innovation expend.		
Feature (Variable name)	Yes	No	P-val.	Yes	No	P-val.	Yes	No	P-val.	Yes	No	P-val
Text-based features												
Emerging technology term (*emerging_tech*)	0.18	0.07	0.00	0.15	0.07	0.00	0.15	0.05	0.00	0.19	0.06	0.00
Percentage of English language (*english_language*)	0.16	0.10	0.00	0.14	0.10	0.00	0.14	0.09	0.00	0.17	0.08	0.00
Percentage of other languages (*other_lang*)	0.02	0.02	0.45	0.02	0.02	0.25	0.02	0.02	0.74	0.02	0.02	0.30
Topic popularity index (*pop_score*)	34.64	34.35	0.36	34.78	34.11	0.03	34.68	34.13	0.08	35.07	33.82	0.01
Share of numbers (*share_numbers*)	0.025	0.028	0.00	0.025	0.028	0.00	0.026	0.028	0.00	0.027	0.027	0.97
Flesch-reading-ease score (*flesch_score*)	40.09	41.22	0.01	40.54	41.03	0.26	40.47	41.26	0.09	39.28	41.28	0.01
Meta information features												
Website size: length (*text_length*)	75,269.35	56,746.84	0.00	71,629.95	55,685.73	0.00	71,193.63	52,859.37	0.00	75,334.75	52,462.63	0.00
Website size: nr. of pages (*nr_subpages*)	30.37	24.65	0.00	28.75	24.87	0.00	28.92	23.75	0.00	31.23	23.58	0.00
Loading time (*load_time*)	0.57	0.55	0.69	0.51	0.60	0.25	0.55	0.57	0.76	0.51	0.49	0.57
Mobile version (*mobile_version*)	0.76	0.70	0.00	0.76	0.68	0.00	0.75	0.67	0.00	0.73	0.69	0.06
Domain purchase year (*domain_purchase_year_proxy*)	2,004.22	2,004.98	0.00	2,004.42	2,004.96	0.00	2,004.37	2,005.17	0.00	2,004.38	2,005.01	0.01
Network features												
Outgoing hyperlinks (*outgoing_links*)	15.93	12.95	0.00	15.18	12.97	0.00	15.19	12.46	0.00	16.23	12.38	0.00
Incoming hyperlinks (*incoming_links*)	14.78	5.22	0.00	13.24	4.30	0.00	12.11	4.09	0.00	12.09	3.70	0.00
Use of social media (*social_media*)	1.62	1.02	0.00	1.51	0.98	0.00	1.47	0.92	0.00	1.62	0.91	0.00
PageRank centrality (*pagerank_index*)	$2 * 10^{-6}$	$1 * 10^{-6}$	0.00	$2 * 10^{-6}$	$1 * 10^{-6}$	0.00	$1 * 10^{-6}$	$1 * 10^{-6}$	0.00	$1 * 10^{-6}$	$1 * 10^{-6}$	0.01
Bridges (*bridge_index*)	0.43	0.26	0.01	0.38	0.28	0.05	0.37	0.27	0.04	0.31	0.27	0.35
Number of observations	4,487			4,484			4,487			1,893		

Notes: All variables were rounded to the second decimal place except the PageRank centrality, which was rounded to the sixth decimal place, and the share of numbers, which was rounded to the third decimal place (reason: some indicator values are very small).

for incoming than for outgoing or social media hyperlinks. Additionally, innovative firms are also significantly more important in firm networks when looking at the PageRank centrality. The statistical significance of the differences in the bridge index is, however, limited to certain forms of innovation activity. In summary, Table 2.3 confirms previous assumptions. Innovative firms seem more likely to apply emerging technologies, to have more technically advanced websites, and to be better connected with each other, according to most network indicators.

Figure 2.1 shows the average occurrence of different emerging technology terms on a firm's website with respect to its product innovator status. The emerging technology terms differ strongly in their probability of occurrence. The emerging technology term *Internet of Things* is most likely to occur. The term appears on more than 8% of all product innovator websites and only on less than 2% of all non-product innovator websites. Also, terms relating to different *machine learning* applications, *biometrics*, *blockchain* technology, and *mobile collaboration* appear relatively frequently. Moreover, nearly every emerging technology term is more likely to appear on a product innovator website than on a non-product innovator website. This result is the same for all innovation indicators.

Table 2.4 shows the ten most innovation-relevant LDA topics. The highest average Pearson correlation coefficients for all four innovation indicators and the document-topic probabilities are used to identify the most relevant LDA topics. The topics are sorted in descending order. LDA topic 98, which according to its keywords relates to research & development, has a positive and by far the strongest relationship to innovation. Also, LDA topic 35, which relates to ICT infrastructure, has a comparatively strong positive correlation with our innovation indicators. Among the top ten, the LDA topics 20 (tourism), 120 (consulting & customer support) and 23 (family business & craftsmanship) have the weakest correlation. Moreover, the correlation is negative.

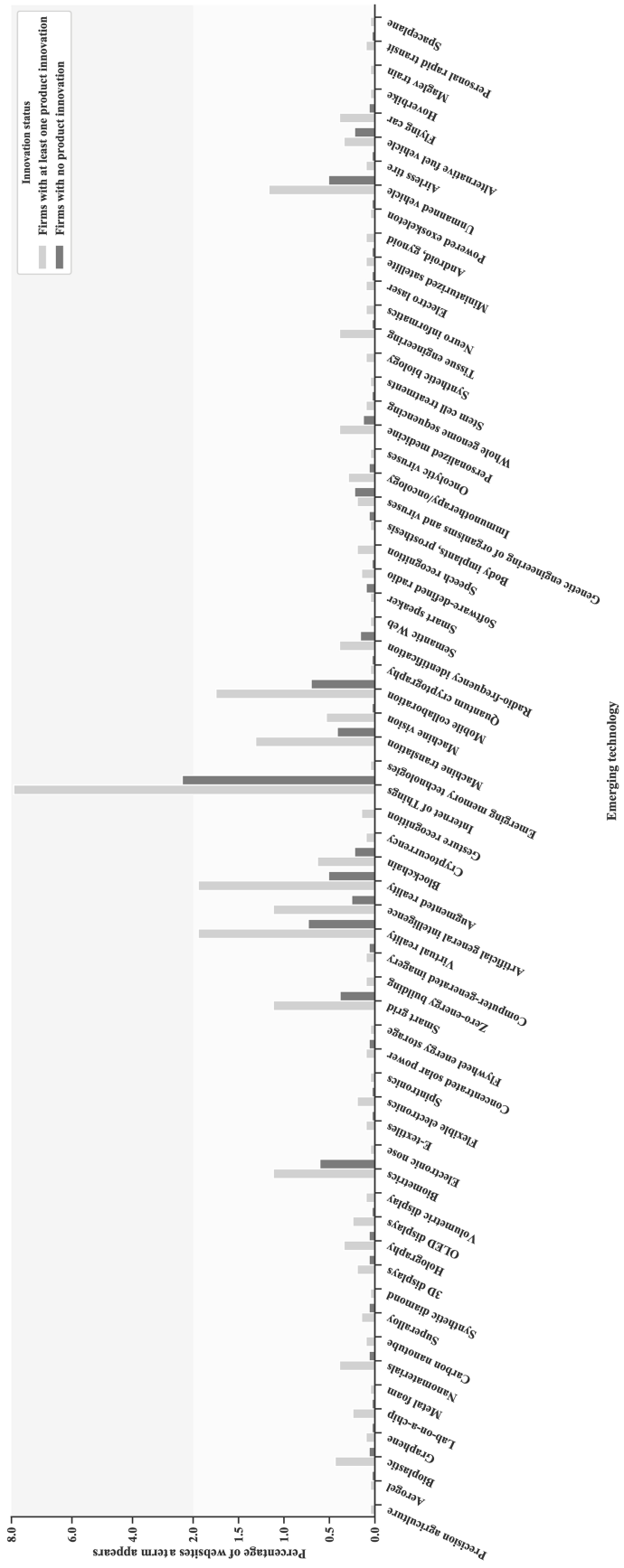


Figure 2.1: **Average occurrence of different emerging technology terms on firm websites with and without product innovations.** Emerging technology terms not appearing on firm websites are not illustrated. The y-axis has a scale break at 2%. Own illustration.

Table 2.4: **Content of the LDA topics with the strongest relationship to MIP-based innovation indicators.**

Topic number	Content	Translated	Top words	Correlation*
LDA topic 98	Research & development	yes	'company', 'customer', 'development', 'to develop', 'department', 'employee', 'partner', 'project', 'successful'	positive (0.15)
LDA topic 35	ICT infrastructure	yes	'system', 'software', 'data centres', 'server', 'version', 'support', 'date', 'windows', 'automatic', 'document'	positive (0.10)
LDA topic 65	Construction	yes	'to build', 'project', 'new building', 'architect', 'planning', 'renovation', 'reconstruction', 'construction', 'to plan', 'architecture'	negative (-0.09)
LDA topic 134	Business software	no	'array', 'value', 'news', 'office', 'paket', 'error', 'data', 'page', 'SAP', 'search'	positive (0.08)
LDA topic 7	Product experience	no	'centro', 'company', 'best', 'use', 'experience', 'world', 'please', 'product', 'may', 'find'	positive (0.08)
LDA topic 41	Common terms	yes	'and', 'far', 'to take place', 'to put', 'frame', 'that', 'information', 'total', 'receive', 'department'	negative (-0.07)
LDA topic 5	Carpentry	yes	'to tile', 'woods', 'to lay', 'laminate', 'tile', 'to put', 'material', 'stairs', 'floor', 'to glaze'	negative (-0.07)
LDA topic 20	Tourism	yes	'region', 'city', 'to be located', 'to offer', 'museum', 'old', 'historical', 'nature', 'tour', 'landscape'	negative (-0.06)
LDA topic 120	Consulting & customer support	yes	'pleased', 'to offer', 'customer', 'to advise', 'individual', 'consulting', 'available', 'question', 'competent', 'to find'	negative (-0.06)
LDA topic 23	Family business & craftsmanship	yes	'company', 'to operate', 'visit', 'to stand', 'roofing', 'Michael', 'son', 'specialize', 'work'	negative (-0.06)

*Measured by the average of all Pearson correlation coefficients between the average topic share per document and each innovation indicator.

Figure 2.2 also relates to the ten most innovation-relevant LDA topics. The histograms show the average share of the respective topic in a document for both innovative and non-innovative firms. The figure reflects the results presented in Table 2.4. The selected topics considerably differ between innovative and non-innovative firms. Also, relationships are constant; e.g. if a topic has a larger share on product innovator websites than on non-product innovator websites, it will also be relatively stronger represented on process innovator websites. Nonetheless, differences between innovation indicators exist. Average topic share differences diverge between indicators and are larger when considering firms' innovation expenditures than when taking product or process innovators into account.

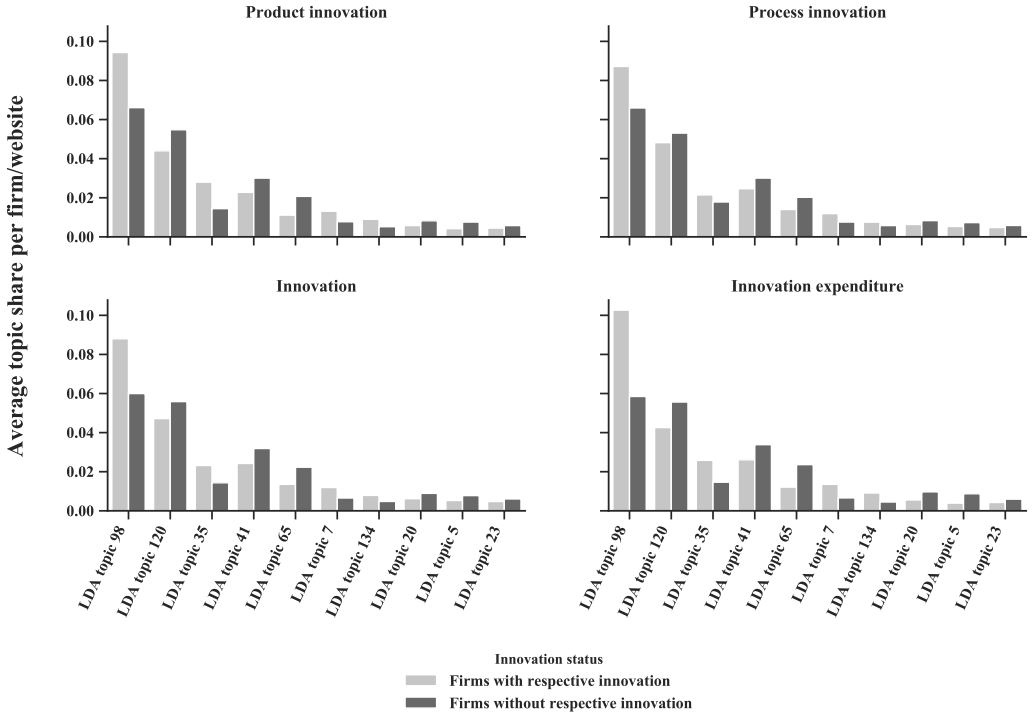


Figure 2.2: **Differences in the topic share of the top ten topics with the strongest average correlation with MIP-based innovation indicators.** For instance, LDA topic 98 has an average share of 10% in a document if a firm has innovation expenditures, compared to merely 6% if a firm does not have innovation expenditures. Own illustration.

2.5 Methodology

The objective of our work is the identification of website characteristics that allow the prediction of firm-level innovation activities. For this purpose, we integrate the described features as predictor variables in random forest classification models (Breiman 2001, Hastie et al. 2001). For each of our feature sets (*text*, *meta*, and *network* features), as well as for *all* features jointly, a separate random forest model is fitted. We use the Python package *scikit-learn* for the exercise. The random forest algorithm

is an ensemble method used for classification or regression tasks. Like any other machine learning algorithm, as defined in Mohri et al. (2018), it uses experience (in our case, survey data) to learn how to perform predictions. The random forest algorithm makes its decision based on the modus or mean of a multitude of decorrelated decision trees. Each tree is built based on bootstrapped samples of training data. By splitting the data at nodes into branches that are more *pure* with respect to the target variable, the algorithm learns to improve. We chose the random forest algorithm because it allows the calculation of feature importances while providing high predictive power and enabling the consideration of complex interactions.

For instance, feature importance can be derived by means of the MDI (Breiman et al. 1984), which is a measure based on a split criterion that is used to build single decision trees.[14] In our study, we use the *decrease in impurity* as a split criterion. A formal description of the *decrease in impurity* is given by Equation 2.1. $i(t)$ measures impurity at the node level, which is indicated by the Gini impurity index. t is a node within one tree, and s is a split at a certain value of a variable. N_x is the number of samples reaching node $x \in \{t, t_L, t_R\}$. Lastly, if t is the parent node, then t_L is the left child node, and t_R is the right child node for the split s at node t. The split s for node t that maximises $\Delta i(s,t)$ is iteratively chosen.

$$\Delta i(s,t) = i(t) - N_{t_R}/N_t * i(t_R) - N_{t_L}/N_t * i(t_L) \tag{2.1}$$

Feature importance is then derived from the sum of *decreases in impurity* of a single variable divided by the sum of *decreases in impurity* of all features used to build the tree. The value is additionally averaged over all trees in the forest and again normalised so that all values sum up to one. If multiple variables will lead to similar impurity decreases at one node, only one variable is selected for splitting. Hence, the (multi-)collinearity of features can bias feature importance. This issue can be illustrated by the following example: The same variable is included twice in a model. When choosing a variable for splitting, the model can randomly choose between the two, and the feature relevance is, thus, divided between both variables.

To evaluate the performance of the collected website characteristics, we use a random coin-toss model based on the sample distribution as a baseline model. A baseline model works as a benchmark to assess the performance of more complex solutions, i.e. the baseline model helps to analyse whether a trained model performs better than a random prediction. To estimate whether we achieve considerable improvements in comparison to baseline predictions, we perform a McNemar test (McNemar 1947). Assuming a chi-squared frequency distribution, the McNemar test measures if predictions from two machine learning models significantly disagree with each other, as illustrated in Equation 2.2. RF captures the number of

[14] For an overview of different split criterion measures, see Louppe et al. (2013).

observations misclassified by a fitted random forest model but not by the baseline model. BL captures the number of observations misclassified by the baseline model but not by a fitted random forest model.

$$\chi^2 = \frac{(RF - BL)^2}{(RF + BL)} \tag{2.2}$$

If a model that includes a distinct feature set significantly disagrees with baseline predictions according to the McNemar test and its evaluation metrics show superior values, we consider this feature set to be relevant for the prediction of firm-level innovation activity.

To further evaluate and compare models, we use the metrics *area under the curve* (AUC), *accuracy*, and *improvement of accuracy* in comparison to the baseline model. We also use *precision*, *recall*, and the *F1-score* for positive as well as negative observations (Fawcett 2006).

$$\text{False positive rate} = \frac{FP}{(FP + TN)} \tag{2.3}$$

$$\text{True positive rate (recall for the positive class)} = \frac{TP}{(TP + FN)} \tag{2.4}$$

The AUC can be explained as follows. The formulas listed in Equations 2.3 and 2.4 are based on the number of false positive predictions (FP), capturing non-innovative firms wrongly predicted as innovative; true positive predictions (TP), capturing innovative firms correctly predicted as innovative; false negative predictions (FN), capturing innovative firms wrongly predicted as non-innovative; and true negative predictions (TN), capturing non-innovative firms correctly predicted as non-innovative. The receiver operating characteristic (ROC) curve is a graphical illustration of the performance of a binary classifier. For different classification thresholds, the *false positive rate* is plotted against the *true positive rate*, and the AUC value is an approximation of the area below the ROC. Accordingly, the AUC value is the probability that a randomly chosen innovative firm is assigned a higher probability of being innovative than a randomly chosen non-innovative firm. Usually, AUC values above 0.7 are considered acceptable (Hosmer et al. 2013).

For the other metrics, a classification threshold has to be set. The classification threshold is also called the *cut-off* value and refers to the transformation of the regression output to a binary classification. Different cut-off values can be chosen if, for example, *false negatives* are considered more costly than *false positives* or if certain metrics need to be optimised. We select 0.5 as a cut-off value for all fitted models because this value is commonly used and we do not prefer one metric or class over the other.

$$\text{Precision for positive class} = \frac{TP}{(TP + FP)} \tag{2.5}$$

$$\text{Precision for negative class} = \frac{TN}{(TN + FN)} \tag{2.6}$$

$$\text{True negative rate (recall for the negative class)} = \frac{TN}{(FP + TN)} \tag{2.7}$$

Formal definitions of precision for innovative and non-innovative firms are illustrated in Equations 2.5 and 2.6. Recall for innovative and non-innovative firms is measured by the *true positive rate* or *true negative rate*, as illustrated in Equations 2.4 and 2.7. Precision measures, for instance, the share of correctly classified innovative firms in all firms classified as innovative, while recall measures the fraction of innovative firms that have been correctly identified as innovative.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.8}$$

$$\text{F1-score}_{P,N} = 2 * \left(\frac{(\text{Precision}_{P,N} * \text{Recall}_{P,N})}{(\text{Precision}_{P,N} + \text{Recall}_{P,N})}\right) \tag{2.9}$$

Accuracy and F1-score are presented in Equations 2.8 and 2.9. Accuracy measures the share of correct predictions in all predictions. The F1-score captures the harmonic mean between precision and recall for positive (P) and negative (N) observations, respectively. Baseline outcomes of accuracy, F1-scores, precision, and recall for our different innovation activity indicators are presented in Table 2.5 in Section 2.6. The random coin-toss model assumes a fixed chance of being innovative (based on the sample mean). Hence, results do not change when the threshold is varied, and therefore the AUC value is not displayed for baseline outcomes.

To control for overfitting, we analyse the model performance by using out-of-sample predictions. Accordingly, we do not evaluate the models' performance with the observations that are already used for learning. The data are split into a training sample (for fitting models) and a test sample (for evaluating models). To be more precise, the test sample is a *hold-out* sample and therefore never used for model training. The training sample consists of 75%, and the test sample consists of 25% of our observations. In the supervised learning context, this is a common partitioning method and constitutes a trade-off between the generalisation of the model and the validity of the evaluation. We also apply grid-search to tune the hyperparameters of all our models on our training sample (Hastie et al. 2001). We explore the hyperparameter space for the *number of trees* (100, 500, 1,000, and 1,500), *maximum tree depth* (50, 100, 150, and 200), and *minimum impurity decrease* (0.01, and 0.001). For all other hyperparameters, we use the default values provided by *scikit-learn* (Pedregosa et al. 2011). This leads to 32 different hyperparameter combinations

for every model. For each hyperparameter combination in our grid-search, a five-fold cross-validation is performed. The k-fold cross-validation belongs to the non-exhaustive cross-validation methods. The technique assesses the generalisability of machine learning models to new data and detects overfitting as well as potential sample biases. The data are split into k subsets, so that $100 - (100/k)$ % of the data are used for training the model and $100/k$ % for validation. In each of the k iterations, a different training and validation data set is used. Considering all models fitted in the cross-validated grid-search, we choose the model with the highest AUC value. The selected model is then evaluated on a test sample.[15]

2.6 Results

In this section, we present the predictions of MIP-based innovation indicators using a random forest classification approach. Table 2.5 shows evaluation metrics for all baseline as well as fitted models. We analyse four different innovation indicators (four target variables), which we predict based on three different subsets of features as well as their union (four different groups of features). Accordingly, we train 16 random forest models.

Looking at product innovators, the highest AUC score (73%) is achieved with *all* features. The baseline accuracy is 0.53. The largest increase can be observed for the *all-feature* model (17 percentage points). *Text*-based features alone, however, lead to an increase of 16 percentage points. Moreover, *network* and *meta* features have a relatively weak impact. They just lead to improvements of 13 and 11 percentage points, respectively. This indicates that a large share of predictive power results from website texts. The baseline F1-score for product innovators is 0.39, and for non-product innovators, it is 0.61. Hence, the sample is slightly imbalanced towards non-product innovators, and the chances of randomly predicting this class correctly are higher. Furthermore, the F1-scores show a similar result to the other metrics. Only the *text* and the *all-feature* models improve F1-scores notably. When solely applying *meta* or *network* features, F1-scores for innovative firms are even worse than the baseline performance. Precision values do not considerably differ between innovative and non-innovative firms and are always higher than the baseline prediction. Moreover, there is a comparatively large increase in precision for innovative firms. In contrast, there is a great difference between both classes with respect to recall values. For innovative firms, the recall values of fitted models are always worse than those of the baseline prediction. For non-innovative firms, the recall fluctuates between 88 and 95%.

[15]To ensure the reproducibility of our study, we fix the random seed when necessary. The random seed influences the model performance to some extent, e.g. observations are assigned to the train or test sample based on the random seed.

Table 2.5: **Results for random forest classification models using different feature sets and target variables.** Evaluation metrics are presented for the test sample.

Feature sets					Accuracy		F1-Score		Precision		Recall		McNemar	
Baseline	Text	Meta	Network	AUC	Value	Δ	Positive	Negative	Positive	Negative	Positive	Negative	P-values	Support
Product innovators														
x				-	0.53	-	0.39	0.61	0.39	0.61	0.39	0.61	-	1,122
	x			0.72	0.69	0.16	0.47	0.78	0.69	0.69	0.35	0.90	0.00	1,122
		x		0.66	0.64	0.11	0.37	0.75	0.59	0.66	0.27	0.88	0.00	1,122
			x	0.65	0.66	0.13	0.30	0.77	0.72	0.65	0.19	0.95	0.00	1,122
	x	x	x	0.73	0.70	0.17	0.49	0.79	0.71	0.70	0.37	0.90	0.00	1,122
Process innovators														
x				-	0.50	-	0.52	0.48	0.52	0.48	0.52	0.48	-	1,121
	x			0.62	0.59	0.09	0.63	0.54	0.59	0.59	0.67	0.50	0.00	1,121
		x		0.60	0.57	0.07	0.64	0.46	0.57	0.58	0.74	0.39	0.01	1,121
			x	0.59	0.57	0.07	0.62	0.52	0.58	0.56	0.66	0.48	0.01	1,121
	x	x	x	0.63	0.60	0.10	0.64	0.55	0.60	0.60	0.68	0.52	0.00	1,121
Innovators														
x				-	0.52	-	0.60	0.40	0.60	0.40	0.60	0.40	-	1,122
	x			0.67	0.63	0.11	0.75	0.30	0.63	0.59	0.91	0.20	0.00	1,122
		x		0.64	0.62	0.10	0.74	0.33	0.64	0.56	0.88	0.23	0.00	1,122
			x	0.62	0.60	0.08	0.75	0.00	0.60	0.00	1.00	0.00	0.00	1,122
	x	x	x	0.68	0.63	0.11	0.75	0.31	0.64	0.59	0.91	0.21	0.00	1,122
Innovation expenditures														
x				-	0.54	-	0.36	0.64	0.36	0.64	0.36	0.64	-	474
	x			0.74	0.73	0.19	0.55	0.80	0.68	0.74	0.47	0.88	0.00	474
		x		0.67	0.65	0.11	0.33	0.76	0.53	0.67	0.24	0.87	0.00	474
			x	0.65	0.67	0.13	0.25	0.79	0.68	0.67	0.16	0.96	0.00	474
	x	x	x	0.75	0.72	0.18	0.55	0.80	0.67	0.74	0.47	0.87	0.00	474

Notes: Numerical values are rounded. The baseline values are calculated assuming perfect knowledge about the test sample distribution, which means that the test sample mean is used for predictions. P-values relate to the significance level at which a model disagrees with its baseline model according to the McNemar test for 10,000 baseline prediction rounds. The significance levels are based on mean values.

Our evaluation metrics for models predicting process innovators have predominantly lower values than those predicting product innovators. Nonetheless, fitted models show for nearly all evaluation metrics better results than the process innovator baseline model, and the McNemar test also confirms a significant difference. Hence, website characteristics still improve predictions. The best performance, in terms of accuracy, is reached by our *all-feature* model, which leads to a performance increase of 10 percentage points. Moreover, *meta* and *network* features perform slightly worse than *text* features.

The performance of innovators is slightly better than that of process innovators in terms of AUC and accuracy. As the sample is slightly imbalanced towards innovators, this performance difference, however, is also partly related to different baseline values. Furthermore, similar to product innovators, we see remarkably higher AUC values for models including *text* features. However, considering all other evaluation metrics, *meta* features perform similarly to *text* features. Looking at F1-scores, predictions for the negative class always perform worse than the baseline model. In particular, the prediction solely based on *network* features leads to F1-scores of zero. This means the model predicts for every firm a probability that the firm is innovative greater than 0.5, which implies that the model always predicts the majority class. This is known as *zero-rule prediction*. For applying this rule, the information included in our baseline model is sufficient. In this regard, *network* features do not provide additional information for innovators. Looking at precision and recall (and not considering the *network* feature model), we find general improvements for innovative firms in comparison to the baseline model. For non-innovative firms, we only find improvements in precision. Recall values, however, are very low and worse than in the baseline model.

Even though the number of observations is the smallest, the predictive performance as well as the performance increase for firms with innovation expenditures are the highest in terms of AUC and accuracy. Looking at the *all-feature* model, firms with innovation expenditures can be predicted with an AUC value of 75% and an accuracy of 72%, which corresponds to an accuracy increase of 18 percentage points. The model solely based on *text* features performs even slightly better than the *all-feature* model considering accuracy. Besides, the values of all other evaluation metrics are always better than random for the *text* and *all-feature* models. Both models only using *network* or *meta* features show strict improvements in accuracy and precision, but F1-scores and recall are partly worse than the baseline model.

Furthermore, the McNemar test confirms that all fitted models significantly disagree with baseline predictions. The divergence is always highly significant (p-values are below 0.001), except for models that predict process innovators with either *meta* or *network* features, which are significant at the 0.01 level. This may be

because both feature sets and models predicting process innovators perform relatively poorly. Hence, the difference from baseline predictions is especially low when combining both. It is also noteworthy that even though the McNemar test is significant, it does not necessarily mean that the model is strictly better than the baseline model. Key evaluation metrics also have to show predominantly superior values. We want to highlight an example here: The random forest model that predicts innovators using *network* features has a large share of inferior values in comparison to baseline predictions. The model uses the zero-rule for its prediction. Accordingly, it significantly disagrees with the baseline model as it uses another decision rule. However, the fitted model is not strictly better. The model's predictions are solely based on the sample mean, and the fitted model is not learning from the provided features, as the evaluation metrics show.

Lastly, we want to note that we do not find a particular combination of hyperparameters across innovation indicators and feature sets that is always selected by the grid-search algorithm. However, the preferred *number of trees*, *maximum tree depth*, and *minimum impurity decrease* do exist across feature sets and target variables. For the *number of trees*, 1000 and 1500 are mostly chosen. The most dominant *maximum tree depth* values are 50 and 100. Moreover, a *minimum impurity decrease* of 0.001 is more frequently selected than 0.01. For more details, see Table A.5.

To analyse the robustness of the presented results, we re-estimate the *all-feature* model for each indicator using all possible combinations of splits between the training and test samples from 0.1/0.9 to 0.9/0.1 (in steps of 0.01). The change in respective AUC values with respect to an increasing training sample is displayed in Figure A.3. We find that AUC values for product innovators, process innovators, and innovators increase until a training sample size of 0.6 and then stay roughly constant at levels pointed out in Table 2.5. Hence, AUC values seem robust with respect to the sample split if a sufficiently large training sample size is reached. Besides, values fluctuate more strongly between 0.8 and 0.9, which is presumably related to a declining test sample size.

The performance of the model predicting innovation expenditures constantly increases until a training sample size of about 0.85. It has a comparatively large drop afterwards and generally tends to be more volatile. Both can be explained by a much smaller overall sample size for this indicator. For instance, a train/test split of 0.5 implies fewer observations are included in the training sample. Also, the test sample is always smaller, which makes the evaluation of the performance less robust. Furthermore, the increasing trend indicates that the model would have continued to improve if we had added more observations. AUC values based on training sample sizes between 75% and 85% fluctuate around the AUC value pointed out in Table 2.5.

In summary, it can be stated that the analysed website characteristics show a better performance in the prediction of product innovators and firms with innovation expenditures than of process innovators. Moreover, text-based features show greater relative relevance.

To compare the relevance of single features across feature sets, the ten most important predictor variables, measured by the MDI, are displayed in Figure 2.3 for each *all-feature* model, respectively.

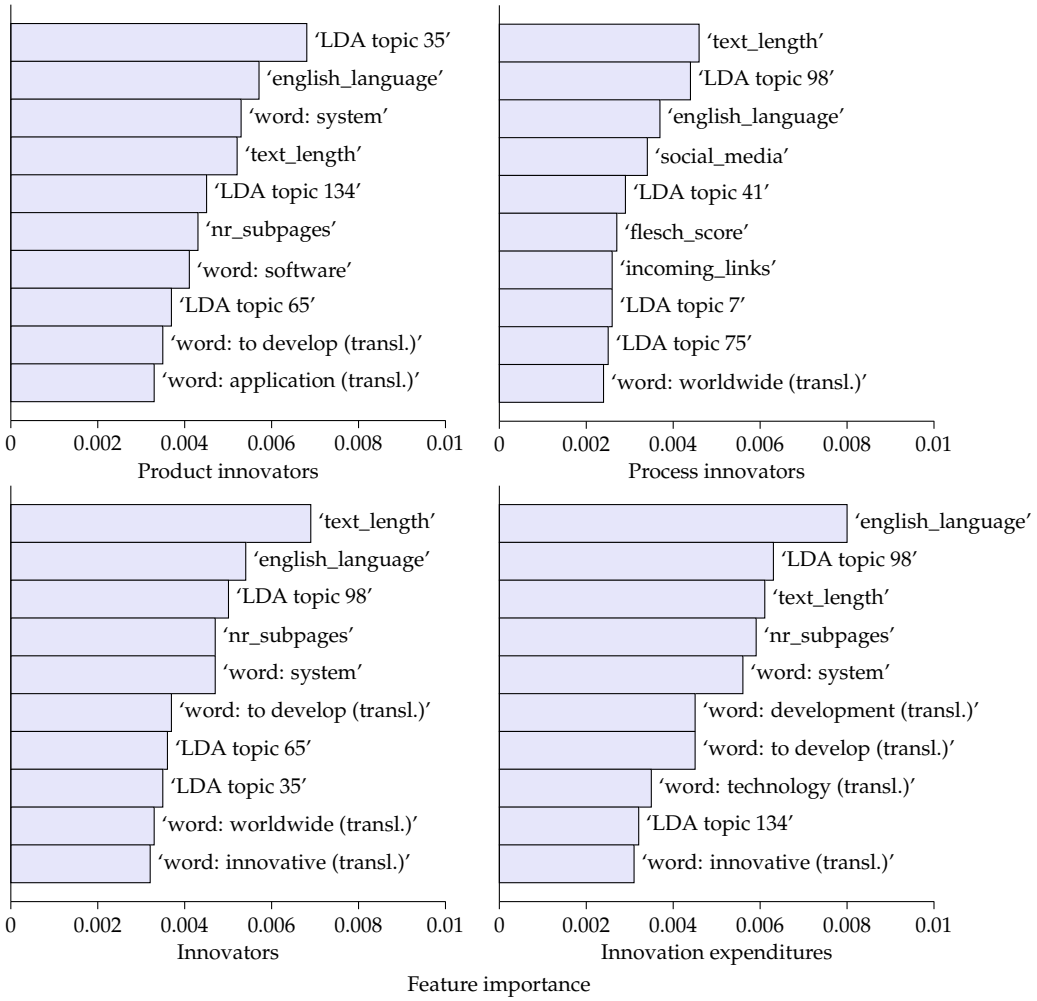


Figure 2.3: **Feature importance values for the *all-feature* models.** For instance, a value that is two times larger implies that the mean decrease in impurity of the related feature is twice as high. Target variables: Product innovators (top left), process innovators (top right), innovators (bottom left), and firms with innovation expenditures (bottom right). Own illustration.

Three features exist that nearly always appear among the ten most relevant: the total number of characters (*text_length*), the number of subpages (*nr_subpages*) (this feature only appears in the twelfth position for process innovators), and the share of the English language (*english_language*). A further investigation of the top 100 most

relevant features (see Appendix A.4) reveals that additional website characteristics exist with some general relevance. The words *worldwide*, *innovative*, *application*, *to develop*, *product*, *technology* (all translated), the word *system*, as well as certain LDA topics, and the topic popularity index (*pop_score*), incoming (*incoming_links*), outgoing (*outgoing_links*), as well as social media hyperlinks (*social_media*), the Flesch-reading-ease score (*flesch_score*), the loading time of a website (*load_time*), and the share of numbers (*share_numbers*) are among the 100 most relevant features for every indicator. This shows that particular website characteristics exist, which have some relevance across indicators. In contrast, it is also noteworthy that features exist that show a large difference in the descriptive statistics but seem less important when predicting the innovation status. For example, the *emerging technology term* dummy never appears among the top ten features for any indicator and is also not frequently observed among the top 100 features. Furthermore, some features are more relevant for certain innovation indicators than for others. For instance, IT-related features seem to be highly relevant for product innovators. The IT-related LDA topics 35 (*ICT infrastructure*) and 134 (*business software*), as well as the words *software* and *system*, are (only) among the top ten features for this indicator. Besides, LDA topic 7 with keywords linked to product experience and the word *application* appear among the top 15 features.

On the contrary, research & development-related LDA topic 98 is more important when estimating process innovators and firms with innovation expenditures. Besides, the LDA topic 65 occurs in Figure 2.3 for product innovators and innovators. The topic should have a negative relationship to innovation activity, as the descriptive statistics show that this LDA topic is more likely to appear on the websites of firms with no innovation activity. With respect to process innovators, only one single word can be found in the ten most important features, and it is the only indicator that has *network* features among its top ten. Furthermore, it is also interesting that the bottom left part of Figure 2.3, which relates to innovators, is, at least for most features, a combination of the most relevant features for product and process innovators. Last but not least, research & development-related words are highly important for predicting firms with innovation expenditures.

2.7 Discussion

Descriptive statistics as well as our fitted random forest models show that website characteristics are relevant predictors for firm-level innovation activity. We see a significant difference in means between innovative and non-innovative firms for most

of our features. For each innovation indicator, random forest models using all features jointly show almost always a considerably higher performance than the baseline prediction with respect to the presented evaluation metrics. Moreover, the McNemar test confirms a significant difference from baseline predictions for all models. Also, our results are in line with Kinne & Lenz (2021). Their statistical model has reached a similar accuracy for product innovators, only observed in one MIP wave.

Our exercise also reveals, especially when predicting product innovators and firms with innovation expenditures, that *text* features are more important than *meta* and *network* features. Besides, we see a pattern regarding the most important characteristics that is independent of different target variables: Across indicators, the total number of characters, the number of subpages, and the share of the English language belong always to the most relevant. It is also noteworthy that these features are more important than the word *innovative*. The findings suggest that website size and language should be considered for different types of website-based innovation indicators, which has not been done in previous studies. Meeting expectations, features that show insignificant differences in Table 2.3 rarely belong to the ten most relevant features in Figure 2.3. An exception is the *flesch_score* in the case of process innovators. Furthermore, considering the poor performance of the *meta* feature models and the result that *text* is the most relevant feature set, the relevance of website size is quite counter-intuitive. One has to consider, however, that the importance of features is considered separately. The relevance of, for example, the number of subpages is compared to the relevance of single words. If all words appearing in the term-document matrix were considered jointly instead, their aggregated relative relevance would lie between 74 and 77%, depending on the indicator. From this perspective, it becomes clear why *text* features and, in particular, textual content are important for an accurate prediction. Nonetheless, as explained before, relative MDI values should always be considered cautiously, as they are affected by multicollinearity. Moreover, other web-based features may exist that possess predictive power and have not been considered in our analysis. These features would most likely change the result.

Furthermore, it would also impact the relative MDI value if this study's website data were complemented with information from other sources, for example, non-web data from the MUP. In this case, innovation activity could potentially be predicted more accurately. However, we have deliberately decided against adding non-web data to our analysis since this study focuses on the comparison of website information that is up-to-date and freely accessible for everyone. Nonetheless, it would certainly be interesting to investigate in a further study the effect of adding additional non-web data. For potentially relevant features, see Gandin & Cozza (2019).

Another aspect that we want to emphasise is the fact that features that are highly important for one particular indicator usually relate to its form of innovation activity. We see this as a strong indication that the models use relevant information. Especially for firms with innovation expenditures, the selected word-based features appear particularly convincing. Terms like *to develop* (transl.) and *technology* (transl.) are highly ranked and have a very strong and direct connection to research & development expenditures. Another example is that the product experience-related LDA topic 7 (top 15 most important features) and the term *application* have high importance for product innovators. Additionally, the ten most relevant features of product innovators have a clear focus on information and communication technologies (ICT), which is in line with the innovation-spawning characteristic of ICT as well as with the result of Hall et al. (2013). They find that ICT investment intensity is positively associated with innovation and is more strongly linked to product than to process innovation.

Moreover, firms have a great incentive to present new products on their websites. Process innovators, however, have a smaller incentive to announce innovation activity because new processes are less relevant for most website visitors. This might explain why results show better predictive performance for product innovators than for process innovators and innovators. In addition, only a single word appears among the ten most relevant features of process innovators, and, even though this model differs on a higher significance level, *text* features alone do only lead to slightly better predictions than *meta* and *network* features. The results support the assumption that process innovations are often not mentioned explicitly on websites.

Regarding innovators, most of the top ten features either appear in the product or process innovator ranking, and the predictive performance of the *all-feature* model lies between both as well. The result meets our expectations as the innovator target variable is a combination of product and process innovators.

It is also interesting to note that, contrary to our expectations, some features are not so relevant. For instance, even though the descriptive statistics show a large difference between innovative and non-innovative firms, the emerging technology dummy does not seem to be very decisive for predictions. Looking at the Pearson correlation coefficients between this and all other features reveals that the emerging technology dummy has a comparatively strong relationship with other features. Hence, their relative MDI value is probably ranked lower due to multicollinearity. Besides, even though the descriptive statistics do not show a significant difference for every form of innovation activity, the Flesch-reading-ease score, the loading time of a website, and the share of numbers appear to be relevant for every indicator (according to the 100 most relevant features). These features, however, do not relate strongly to other features and might therefore provide some extra information.

Hence, they are comparatively relevant despite small differences.

Although we show a clear link between website characteristics and innovation status, the predictive performance of our models leaves room for improvement, as we, for example, still misclassify the existence of innovation expenditures for a considerable share of firms. Predictions might perform slightly better if neural networks were used. Our main criteria for choosing a random forest approach are the explainability of results and the fact that non-linear relationships can be learned. Unfortunately, neural networks do not offer the direct possibility of disclosing decision processes. Hence, there is a trade-off, which often occurs in practice, between performance and explainability. If explainability is not necessary, predictive performance can most likely be improved by neural networks.

Within our sample, there can, of course, also be innovative firms that do not mention their innovation activity (implicitly or explicitly) on their websites. In other words, some inaccuracy might relate to the nature of our data. In particular, product innovators, process innovators, and innovators might suffer from noise as they cover three years. Websites can change a lot during this period. Comparatively good results for firms with innovation expenditures may be explained by the fact that this information is observed on an annual basis. Solving this matching problem seems like a necessary step to improve predictions. Nonetheless, text data are always noisy, and models with perfect accuracy are rarely identified.

Furthermore, it could be criticised that website-based innovation indicators can only be applied to firms that have a website. Another point of criticism is that it could create noise if firms falsely claim on their website that they are innovative, e.g. for marketing purposes. The MIP contains self-reported data as well; however, firms do not have the incentive to make false declarations, as answers should not affect their public image. For this reason, we expect MIP data to reveal the actual innovation status. We consider the usage of MIP-based data as target variables as a solution to the problem of false declarations of innovation activity on firm websites. Besides, patent data could have also been used as an alternative target variable. However, patent-based indicators rather measure inventions than innovations.

2.8 Conclusion

In this research article, we contribute to the discussion on whether web-based innovation indicators are a feasible alternative to survey-based innovation indicators. We conduct our analysis with data on 4,487 German firms that reported different forms of innovation activity in a large-scale questionnaire-based survey (the MIP 2019). We extract website texts, additional website-related meta-information, as well as hyperlinks of these firms, and use the information to predict firm-level innovation activity reported in the MIP. The performance of our machine learning models shows that

website characteristics unambiguously relate to MIP-based innovation indicators. Furthermore, we find that website characteristics better predict product innovators and firms with innovation expenditures than process innovators. Hence, website characteristics appear to be suitable for measuring only certain aspects of innovation. Additionally, the importance of certain website characteristics varies between indicators. Accordingly, different features should be taken into account depending on the kind of innovation activity being analysed. Lastly, our work and related studies show that state-of-the-art web-based predictive modelling cannot fully replace traditional surveys as error rates remain quite high. However, our models provide information about innovation activities that can be quickly updated, are on a very granular level (firm level), and are less expensive than questionnaire-based surveys.

Chapter 3

Measuring Technological Change - A Novel Text Mining Approach

Joint work with Janna Axenbeck.

3.1 Introduction

General purpose technologies (GPTs) are important drivers of technological change. They are defined as "disruptive technologies [...] possessing a wider scope for continuous improvement and elaboration [...] and higher complementarity" (Petralia 2020, p. 1) and have the capacity to transform economies (Bresnahan & Trajtenberg 1995). The steam engine and electricity serve as prime examples of GPTs with profound impact. Therefore, it is of high importance to analyse the diffusion and impact of GPTs. A prerequisite for this is to collect reliable data and create measurement tools for their adoption. In this paper, we present a comprehensive now-casting tool for measuring GPT adoption, focusing on the example of *digital technologies*.

Digital technologies are GPTs due to their potential for technical improvements, their diffusion into all economic sectors, and the potential to enable their users to develop complementary innovations (Bresnahan & Trajtenberg 1995, Jovanovic & Rousseau 2005). E.g. the ongoing diffusion of information and communication technologies (ICT) affects firm-level production processes (Cardona et al. 2013, Vu et al. 2020), creates new markets, such as the app market, and reduces economic costs, such as search and transaction costs (Goldfarb & Tucker 2019). Due to these characteristics and their high potential to increase innovation and productivity growth, it is of particular importance to track their diffusion and economic effects.[16]

Measuring the digital transformation poses some fundamental challenges, as it does not relate to a single technology but to a group of already established or emerging digital technologies. Hence, either indicators for a single digital technology or

[16]The OECD provides a roadmap for measuring digital transformation in `https://doi.org/10.1787/9789264311992-en` [Last accessed: 24.02.2024].

composite indicators are utilised.[17] But both approaches entail disadvantages, as, on the one hand, single technologies cannot provide a complete picture, and, on the other hand, composite indicators require human assessment of how to measure and weigh different aspects related to digital technologies.[18] Measuring the adoption of digital technologies at the firm level also poses problems with respect to the collection of data. Traditionally, firm-level information is collected through questionnaire-based surveys, but these often come along with certain drawbacks. They are cost-intensive, lack timeliness and regional granularity, and require firm participation. In contrast, nowadays, a large share of firms have public websites that provide a wealth of firm data. Firm websites often include online shops, information about digital products, job postings, applied technologies, and links to social media websites.[19] Hence, information on a firm's website can relate to the firm's use of digital technologies in their products, processes, and distribution channels. Advances in computing power and natural language processing enable the collection and usage of large unstructured data sets. As a result, the texts on firm websites can be analysed quickly and on a large scale. Previous research has also illustrated that websites contain useful data for measuring economic outcomes, e.g. firm-level innovation activity (Axenbeck & Breithaupt 2021, Kinne & Lenz 2021).

One way of measuring digitalisation would be dictionary-based methods, such as keyword searches. A disadvantage of these methods is that words describing certain digital technologies are, to some extent, generic and do not always unambiguously relate to *digitalisation*, i.e. the words *apple* and *bit* can only be fully understood with contextual words. Both examples illustrate that measuring digital technologies on firm websites with a simple keyword search might not be sufficient. For this reason, we need a methodology that captures a more sophisticated and, at the same time, broad definition of firm digitalisation.

Measuring firm-level adoption of digital technologies faces challenges because no clearly defined ground truth exists, as digitalisation is an abstract construct. Hence, a suitable target variable for a supervised learning approach is not available.[20] To solve this problem, we propose to leverage firm websites in combination with classified *news articles* to create an indicator for firm-level adoption of digital technologies. We apply a *transfer learning* approach, in which a new task is solved through the transfer of knowledge from a related task (Torrey & Shavlik 2010, Lima et al. 2017). Accordingly, we use texts that are already labelled as either being about

[17] Examples: cloud capacity (single indicator) and DESI (composite indicator) of the European Commission (see: `https://ec.europa.eu/newsroom/dae/redirection/document/88764` [Last accessed: 24.02.2024]).

[18] Handbook on Constructing Composite Indicators (OECD): `https://www.oecd.org/sdd/42495745.pdf` [Last accessed: 24.02.2024].

[19] The information is based on a (presumably positive) self-representation and may be biased. Thus, our focus is rather on the relative value instead of the absolute indicator value.

[20] Existing firm-level indicators are typically not published.

or not being about digital technologies. For this purpose, we use newspaper articles, as they are often published within predefined sections or are labelled with keywords. The New York Times website, for example, has a section entitled *Tech*.[21] We use data from four different German newspaper outlets to ensure compatibility with our German firm data set. Two news outlets cover daily news, while the other two have a technical focus. As a result, we provide a broad definition of digital technologies. The labelled newspaper data build the fundamentals for an expert system on the subject of digital technologies. Using articles from these newspapers, we fit a multi-language *random forest regression* model that predicts the likelihood of a newspaper article addressing digital technologies (Breiman et al. 1984, Hastie et al. 2001). The model separates news articles on digital technologies from other topics with an accuracy of 97% on the test sample and is then applied to three annual data sets of German *firm website texts*.[22] The predictions are used as a continuous indicator for the firm-level degree of digitalisation and vary from zero to one.

Our results show that larger firms adopt digital technologies more often than smaller firms. Firms in the industries *"computer programming, consultancy, and related (service) activities"*, *"telecommunications"*, and *"publishing activities, video and television, music publishing, etc."* have a high degree of digitalisation, and firms in the industries *"accommodation and food and beverage service activities"*, *"beverages, food and tobacco"*, and *"construction"* are least digitalised. Moreover, firms in West Germany and in bigger cities adopt digital technologies more often than firms in East Germany and in rural areas. For plausibility checks, we use aggregated Eurostat data on ICT usage[23], regional digitalisation intensities by Prognos AG[24], and data from the Mannheim Innovation Panel (Rammer et al. 2021). Comparing our web-based indicator for firm digitalisation with established indicators shows similar and plausible patterns.

Lastly, we illustrate the indicator's potential for giving timely answers to pressing economic issues by analysing the link between digital technologies and *firm resilience* during the COVID-19 health crisis. For this purpose, we interpret changes in credit ratings as a proxy for firm resilience to the exogenous COVID-19 crisis. Our statistical analysis indicates the following: A high pre-crisis level of digital technology adoption in 2018 as well as an increase in the adoption of digital technologies between 2018 and 2020 are linked to improvements in credit scores between 2019 and 2021 (pre-crisis vs. crisis). The results are in line with the related economic literature on firm resilience.

[21] The New York Times: `https://www.nytimes.com/` [Last accessed: 24.02.2024].

[22] Number of firm-level observations per year: 663K for 2018, 894K for 2020, and 950K for 2022.

[23] isoc_e data: `https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm` [Last accessed: 24.02.2024].

[24] Prognos Digitalisation Compass 2018: `https://www.prognos.com/de/projekt/digitalisierungskompass-2018` [Last accessed: 24.02.2024].

Our contribution to the literature is an up-to-date, low-cost, and quality-tested methodology to measure digital technology adoption, covering all firm-size classes, regions, and industries in Germany.

The remainder of this paper is structured as follows: In Section 3.2, the related literature is summarised. Section 3.3 presents our data-driven framework, while Section 3.4 illustrates the plausibility of the results. The indicator is applied to a use case in Section 3.5. In Section 3.6, we discuss the results and our contribution. Section 3.7 contains a conclusion.

3.2 Literature Review

This section gives an overview of related text mining and statistical learning methods (Section 3.2.1), and literature on GPTs such as digital technologies (Section 3.2.2).

3.2.1 Text Mining and Statistical Learning

The newly developed indicator for firm digitalisation relies on web-based text data and predictions. The underlying methodology is, thus, related to the literature on economic indicators based on machine learning, text mining, and web data.

Previous studies already used web data to construct frequent real-time estimates, also known as nowcasting. Famous examples are Ginsberg et al. (2009) utilising Google search queries to detect influenza epidemics in the USA, and Choi & Varian (2012) showing that search engine data often correlate with economic activities such as automobile sales and unemployment claims. Some studies in the innovation literature illustrate that firm websites and other web data sources are suitable to generate firm-level indicators (Gök et al. 2015, Pukelis & Stanciauskas 2019, Axenbeck & Breithaupt 2021, Kinne & Lenz 2021). Some publications have a focus on topic modeling: Lenz & Winker (2020) apply a *Paragraph Vector Topic Model* for measuring the diffusion of new technologies (innovations). Larsen & Thorsrud (2019) decompose articles of business newspapers using a *Latent Dirichlet Allocation* topic model and show that content in newspaper articles entails predictive power for economic variables.[25] Newspaper articles were already successfully used to explain various economic outcomes and might therefore also be suitable for measuring firm digitalisation (e.g. Groseclose & Milyo 2005, Tetlock 2007, Engelberg & Parsons 2011).

Digitalisation is a broad and abstract concept without a well-defined ground truth that would enable precise supervised machine learning. In these situations, a transfer learning approach can be utilised that uses pre-trained models. Transfer learning has already been applied to image recognition tasks, e.g. Lima et al.

[25] For a further literature review, see Gentzkow et al. (2019).

(2017). As an example of an economic application, Xie et al. (2016) use a transfer learning approach to measure poverty rates with nighttime light intensity rates. We contribute to this literature by combining content from newspapers with texts from firm websites.

3.2.2 Measuring Digitalisation

Identifying and tracking general purpose technologies is associated with unique challenges. For example, related studies often have to rely on patent data (e.g. Hall & Trajtenberg 2006, Petralia 2020). For several technologies, there are studies dealing with their identification and tracking: machine learning (e.g. Goldfarb et al. 2023), nanotechnology (e.g. Youtie et al. 2008, Schultz & Joutz 2010, Graham & Iacopetta 2014), and blockchain (e.g. Kane 2017, Filippova 2019).

The focus of this study is on digitalisation which can be broadly defined as the "mass adoption of digital technologies that generate, process and transfer information" (Katz & Koutroumpis 2013, p.314).[26] It can be differentiated between internal and external digitalisation (see e.g. Büchel et al. 2020). Internal digitalisation refers, for example, to the digitalisation of firm-level processes and products. In contrast, external digitalisation includes, for example, the broadband availability in a region. Therefore, our web-based firm-level indicator focuses on internal digitalisation.

Governmental and affiliated institutions regularly capture data and provide a large variety of digitalisation indicators. We give a general overview of these below: Every year, the Federal Statistical Office of Germany (Destatis) collects data for Germany on the degree of firm digitalisation for Eurostat's *"ICT in enterprises"* survey.[27] For example, the data include information on enterprises with a website and enterprises using social media. The digitalisation indicator funded by the German *Federal Ministry for Economic Affairs and Climate Action* is measured at the firm level and describes the level of digitalisation with respect to industries, regions, and firm-size classes (Büchel et al. 2020). It captures data on firm-level properties such as digitalisation of processes and products and external factors like the availability of technical infrastructure. Further, questions about digital technologies used by firms are included in the Mannheim Innovation Panel survey (Rammer et al. 2021). The

[26] In contrast, digitisation describes "the process of converting something to digital form" (Source: `https://www.merriam-webster.com/dictionary/digitization` [Last accessed: 24.02.2024]).

[27] Description of the data: `https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Enterprises/ICT-Enterprises-ICT-Sector/_node.html` [Last accessed: 24.02.2024].
Access to the data on an aggregated level: `https://www-genesis.destatis.de/genesis/online?language=en&sequenz=statistikTabellen&selectionname=52911#abreadcrumb` [Last accessed: 24.02.2024].

OECD Going Digital index[28] consists of seven sub-indicators: access, use, innovation, jobs, society, trust, and market openness. The Digital Economy and Society Index (DESI) tracks the digital performance indicators of European countries and makes their progress comparable. Its dimensions are human capital, connectivity, integration of digital technology, and digital public services.[29] An overview of further country-level indicators is provided by Kotarba (2017). These composite indicators are predominantly used for comparisons between countries.[30] However, each of the presented indicators has its drawbacks. Especially, official statistics often cannot react to recent technological developments; the firm-level data are sometimes not accessible; the data exhibit significant temporal delays; the number of observations is usually small; and the underlying surveys are costly. For example, the present indicators do not cover cutting-edge technologies, such as large transformer models.[31]

Some approaches utilise big data and online data to measure the degree of firm digitalisation and related metrics. For example, Bertenrath et al. (2017) capture firm website metrics on applied technologies, mobile maturity, amount of traffic, search rankings, social media usage, keywords, and quality measures. Deviating from this, Ashouri et al. (2022) create a firm digitalisation score for products. The products are classified, e.g. by analysing the website content. Albeit, these indicators have some major drawbacks, as they only measure certain (technical) parts of firm digitalisation, sub-indicators need to be adjusted before aggregation, and firm-level indicators are, in some cases, not accessible.

Academic literature introduces a large set of proxies to measure firm digitalisation, e.g. for analyses of its economic effects: Hall et al. (2013) analyse, on the one hand, the relationship between research and development (R&D) expenditures and monetary ICT investments and, on the other hand, innovation and productivity at the firm level. Similarly, Dhyne et al. (2021) derive firm-level ICT capital stocks and find that high ICT capital is linked to an increase in the value added of a firm. Cardona et al. (2013) and Schweikl & Obermaier (2020) provide an overview of the empirical literature on ICT and productivity. Furthermore, the capital stock of computers or IT and survey-based spending data are used as indicators in a multitude of studies (Brynjolfsson & Hitt 1995, 1998, Brynjolfsson et al. 2002, Brynjolfsson & Hitt 2003). Other publications look at regional differentiation: Billon et al. (2010) analyse ICT adoption at the country level. They use, for example, the number of broadband subscribers and secure internet servers. Bloom et al. (2012) analyse the

[28] `https://goingdigital.oecd.org/en` [Last accessed: 24.02.2024] provides a view on the OECD toolkit and `https://doi.org/10.1787/9789264312012-en` [Last accessed: 24.02.2024] supplies the documentation.

[29] The methodological documentation of the DESI is provided by the European Commission at `https://ec.europa.eu/newsroom/dae/redirection/document/88557` [Last accessed: 24.02.2024].

[30] E.g. the EU 2020 innovation indicator captures technological progress (Janger et al. 2017). A broader overview of science and technology indicators is provided by Grupp & Schubert (2010).

[31] Transformer models are a prominent deep learning architecture (Vaswani et al. 2017).

effect of IT on productivity for (non-)multinational firms by using UK Census Bureau data on IT expenditures. Forman et al. (2012) analyse the effect of the diffusion of the internet on regional wage inequality. Similarly, the effects of mobile internet use on productivity are investigated (Bertschek & Niebel 2016) and the economic impacts of broadband internet are described in a literature overview by Bertschek et al. (2015). All of the presented studies have to deal with shortcomings in the measurement of digitalisation: Exact firm-level investment and capital stock data are often only available for a few firms. The proxies are based on specific technologies, such as broadband or cloud, and are subject to a time lag.

Therefore, our web-based approach makes a valuable contribution to this literature as it covers a broad definition of digital technologies, requires no (or only minor) human assessment of what digitalisation is, can be updated very quickly, and captures a large share of all German firms.

3.3 Framework

This section is about the framework for the measurement of digital technology adoption. First, we introduce a machine learning model (Section 3.3.1). Second, we describe the newspaper data, website data, and the data processing pipeline (Section 3.3.2). Third, we present the results from the model training (Section 3.3.3).

3.3.1 Approach

To create an indicator of firm-level digitalisation, we use a transfer learning approach. First, information from a large labelled data set is extracted, and a supervised machine learning model is trained. In the second step, the model is applied and optionally fine-tuned to a related problem.[32] The individual steps of our approach are illustrated and explained in Figure 3.1. In our case, the model is trained on news article texts because they are, for example, already labelled based on their overarching topic, such as politics, sports, or digitalisation.[33] Thus, by using news articles, we solve the problem of the missing definition of digitalisation. The random forest algorithm is employed for model training (Breiman et al. 1984, Hastie et al. 2001, Pedregosa et al. 2011). The multi-language machine learning model supports German and English texts, learns from a binary outcome (digitalisation vs. non-digitalisation)[34] and can predict the label of new or unseen news articles. Using this type of model, it is possible to create continuous regression forecasts between

[32] Usually, these machine learning models are called *pre-trained*.

[33] The labels of the news articles are extracted from the SEO keywords or their assignment to sections.

[34] Example (Digitalisation): News articles about the German government's Digital Summit are an example of positive data points. In contrast, articles about welfare reforms are labelled as non-digital.

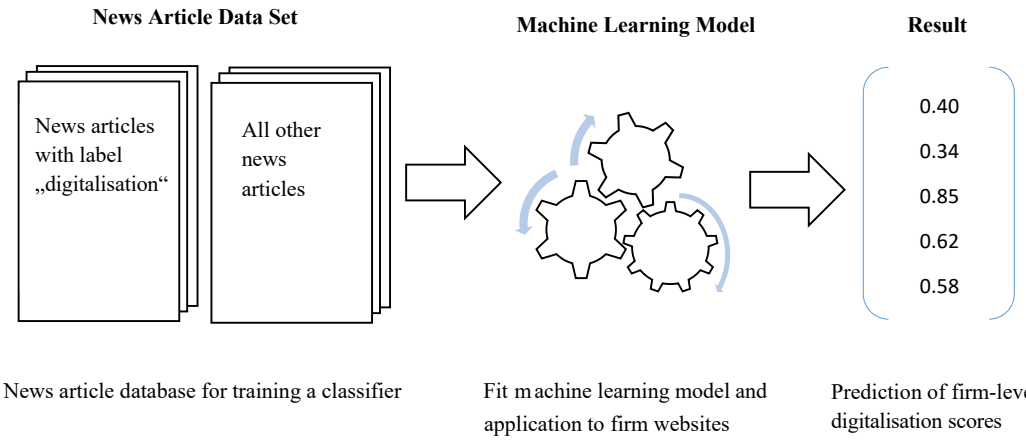


Figure 3.1: **Empirical approach for the development of a firm digitalisation indicator.** News article data with binary labels (left), machine learning model (middle), and continuous firm-level digitalisation scores based on scraped websites (right). Own illustration.

zero and one, indicating the probability of a text being about digitalisation.[35] The trained model is applied to firm website texts to obtain continuous predictions that we interpret as firm-level digitalisation scores.

3.3.2 Data

The following subsection describes the newspaper articles, firm websites, and their transformation into the vector space.

News Articles

We scrape newspaper data from four major German online news providers with the Python packages Selenium[36] and newspaper[37]. The news providers are not mentioned by their full names; instead we assign them the numbers (1), (2), (3), and (4) for referencing purposes. To cover a wide and diverse spectrum of content, we choose a mixture of technical and daily news.

In the first step, we capture the URLs on the news article websites and store them in a unified list. In the following step, each URL is called, and the HTML code is saved. Hereby, we scrape and process about 158K news article pages.[38] Errors occur as soon as the URL is not accessible or the HTML code cannot be parsed. Each article has the following data fields: the URL, an abstract if applicable, the text content,

[35] As long as it supports the output of regression forecasts, any machine learning model can be used.
[36] We choose the Selenium package because news websites are often dynamically loaded (`https://pypi.org/project/selenium/`) [Last accessed: 24.02.2024].
[37] Newspaper package: `https://pypi.org/project/newspaper/` [Last accessed: 24.02.2024].
[38] We use the abbreviations *K* for thousand and *M* for million, e.g. 80K instead of 80,000.

the newspaper section (optional), a publication date, and further meta information, e.g. search engine optimisation keywords. The publication date is explicitly considered, as the news articles reach many years into the past and might contain outdated information. Data fields are filled by analysing the HTML code. We remove an observation if a relevant data field cannot be extracted. The data extraction from news articles is exemplified in Figure 3.2.[39] We restrict the news article data set to recent articles as digital technologies are constantly evolving. All articles before 2017 are therefore removed. News articles after 2019 are removed because digital technolo-

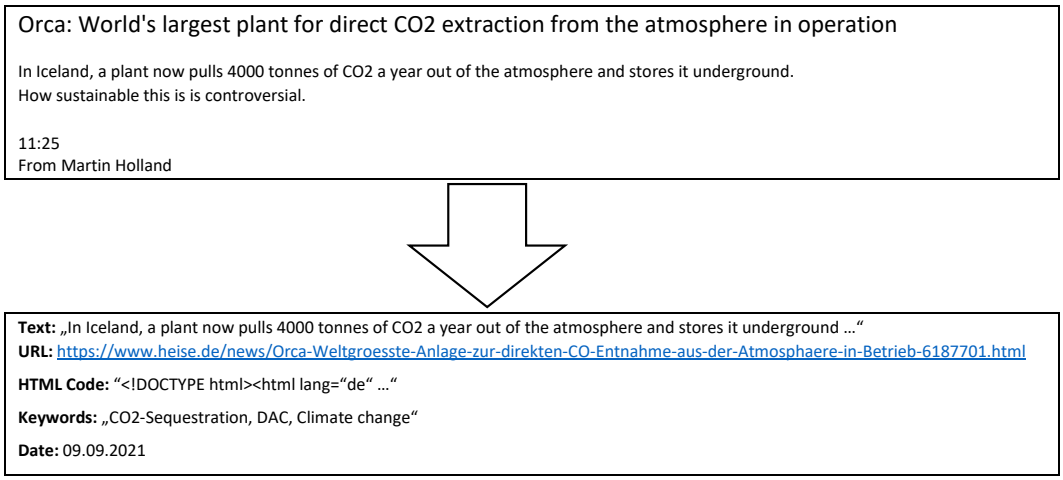


Figure 3.2: **Example of a news article (top) and an excerpt of the extracted data (bottom).** The presented news article was translated from German to English. Own illustration.

gies are often directly related to COVID-19. The data processing leads to a data set consisting of 70.2K news articles. We ensure that no news texts are exact duplicates by removing identical news articles that are published by different news providers. Furthermore, we only consider news articles with a minimum length of 1,000 characters (including spaces) for the machine learning model, which results in a slightly smaller data set consisting of 68.5K news articles without short and duplicated texts.

We label a subset of 25.3K newspaper articles with a binary variable. The variable is set to one if the article covers digitalisation and zero otherwise. In our example, for the daily news sources (1) and (2), the label is set to one if the search engine optimisation (SEO) keywords[40], text, or URL contain the word *digital*. For the two technical news sources (3) and (4), two research assistants independently labelled news articles depending on whether they cover the topic of digitalisation. We keep

[39]The illustrated news article is not in the data set of scraped articles. News article source: `https://www.heise.de/news/Orca-Weltgroesste-Anlage-zur-direkten-CO-Entnahme-aus-der-Atmosphaere-in-Betrieb-6187701.html` [Last accessed: 24.02.2024].

[40]The correlation between the appearance of the word digital in the text and the search engine optimisation keyword *digital* is high. We assume that some of the news providers automatically create the SEO keyword *digital* as soon as this word occurs in the text.

news articles if the research assistants agree on their labels.[41] Table 3.1 shows the summary statistics for the news providers. There are positive and negative training examples for each news provider, but the share of news articles on the topic of digitalisation is higher for the sources (3) and (4) because they report more frequently on technical topics.

Table 3.1: **Label statistics for German news articles.**

News source	Number of articles	Articles about topic digitalisation
(1)	9,048	1,014
(2)	15,325	1,243
(3)	651	451
(4)	297	245
Total	25,321	2,953

Notes: The articles are scraped from four different data sources. The translated English news articles have the same labels.

The model is trained on news articles and has to classify firm website texts written in German and English. Therefore, we machine translate the news articles into English using the Python package Deep-Translator.[42] Table B.2 shows an overview of the text body size of the news articles. For example, German news articles contain, on average, about 500 words and consist of 3,600 characters. The metrics for English texts differ for technical reasons.[43]

Website Data

We use the Mannheim Enterprise Panel (MUP) to retrieve website addresses. The MUP comprises almost all economically active firms in Germany. For example, by the end of 2013, it consisted of approximately 3.2M economically active German firms (Bersch et al. 2014). The MUP is fed with data from Creditreform[44], one of the largest credit rating agencies in Germany, and is updated every six months. From this data set, samples are taken for research and policy advice projects, e.g. for the Mannheim Innovation Panel (Rammer et al. 2021). For some of these firms, a web address for the firm's website is available.[45] The website URLs are the starting point for the web scraping approach. For our study, we use snapshots of the MUP data for the years 2018, 2020, and 2022. The ARGUS web-scraper (Kinne & Axenbeck 2020)

[41] The study can also be conducted without manual labelling, i.e. only using data sources (1) and (2).

[42] Deep-Translator software package: `https://pypi.org/project/deep-translator/` [Last accessed: 24.02.2024]. We use the *Google Translator* function of the package.

[43] The software package has a limit on the text length (5,000 characters) and the number of requests per day. As a result, less than 20% of texts are not completely translated. Optimisation potential: Separate long texts into smaller texts and translate them over a longer period.

[44] The Creditreform Group operates as a credit reporting agency and debt collection service provider. Further information is available at `https://www.creditreform.de/` [Last accessed: 24.02.2024].

[45] In 2018, around 1.15M economically active firms in Germany had a URL (Kinne & Axenbeck 2020).

is then used to capture website content in these years. The tool uses the Python package Scrapy[46] and has several settings. A firm website consists of at least one web page, i.e. a document that can be viewed in a web browser. In our case, we have limited the scraping to the 50 web pages with the shortest URL on a website. We assume that shorter URLs rather refer to general content, as the web page is usually fewer clicks away from the main page. Only internal web pages are considered, i.e. we exclude links to other sites such as Google or cooperation partners. Furthermore, we favour, based on a heuristic, German web page texts. For international firms with multiple versions of their website, this is an important decision. Therefore, there are mainly German and, to a lesser extent, English texts in our data set. In the next step, we remove irrelevant web pages with an approach similar to the machine learning-based *gold-bloat method* proposed by Kinne & Lenz (2021). For this, we train a model that predicts the probability of a text having relevant content. For the model training, we use web page texts that are divided into relevant and non-relevant classes. The model is then applied to the complete web page data set to remove login, document (file), contact, and legal content pages and, thereby, reduces noise and irrelevant text data. Lastly, the data are aggregated at the firm level by combining all web pages of a firm into one single text.

The preparation of the website data is presented in Figure 3.3. The number of scraped websites is 738K in 2018, 1.1M in 2020, and 1.1M in 2022. After merging and cleansing the data, 663K websites in 2018, 894K in 2020, and 950K in 2022 remain. We remove some observations, e.g. when more than one observational unit is likely to represent the same MUP firm. The number of firm websites differs due to the following reasons (no complete list): the available URL data set for 2020/2022 is larger than in 2018; websites are accessible for the first time in 2020/2022 or have been switched off since 2018; dealing with complex corporate structures; technical difficulties with the IT infrastructure; countermeasures against web scraping; and improvements in the web scraping software between the data collection time points.

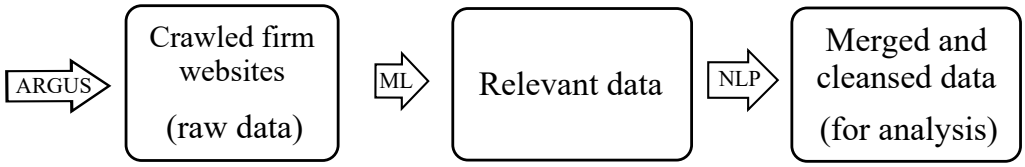


Figure 3.3: **Firm website data crawling and text processing pipeline.** Steps: [1] data crawling using the *ARGUS web-scraper*; [2] extraction of relevant data with the gold-bloat classifier; and [3] further cleansing and merging of the web page text data. Own illustration.

[46]Scrapy software package: `https://scrapy.org/` [Last accessed: 24.02.2024].

Furthermore, the scraped websites often contain errors. Reasons for this are, for example, dynamic websites that load parts of the texts only after user interaction.[47] The content of the website texts depends on the year in which the data were collected, as firms modify their websites over time. Websites can also change because of the characteristics of the visiting users, e.g. via cookies or the browser. The ARGUS web-scraper uses the software package Scrapy and, therefore, does not simulate a browser. Repeating the scraping process at a later point in time will likely yield different results, providing the opportunity to create a time-varying panel that tracks the websites.[48]

Transforming Texts to Matrices

Table B.1 describes the data preparation of the news and firm website texts. The steps are based on established methods from computational linguistics and data science (e.g. Manning & Schutze 1999, Khurana et al. 2023). The pipeline consists of data filtering, text tokenisation, stop word filtering, stemming of words, short word removal, unification of words, special character removal, and selection of words, e.g. based on a *tf-idf weighting*.[49] As a result, the text is decomposed into the most important standardised words. The news article and website text data processing need to be equal, i.e. the vocabularies must be comparable. For example, the vocabularies are not comparable if stemming is only performed for one of the two data sets. We assume that the uniform standardisation will mitigate the problem that the news article and firm website text types are different to some extent.

We cannot apply methods of statistical learning to text data directly because they usually need to be in a matrix structure (e.g. Hastie et al. 2001, Gentzkow et al. 2019). Therefore, news article and firm website texts are transferred to a term-document matrix M. The columns of M represent the documents D (firm websites or news articles). The rows represent the words (or terms) of the vocabulary, e.g. standardised words in a text corpus. Figure 3.4 highlights that the matrix entry at position $m_{i,j}$ corresponds to the frequency or weight of the word i in document j.

[47] The dynamic content on websites could also be scraped, e.g. with Selenium. However, this is not feasible, as we cannot implement tailored solutions for a huge number of websites. There may be a selection bias, as we expect especially larger and more digitalised firms to have dynamic websites.

[48] The website data sets are stored for replication purposes at ZEW's research data center. The newspaper data will be archived at ZEW's dark archive after the research project has been completed.

[49] The term frequency–inverse document frequency (tf-idf) reflects the importance of a word in a text relative to a corpus of texts (Salton & Buckley 1988).

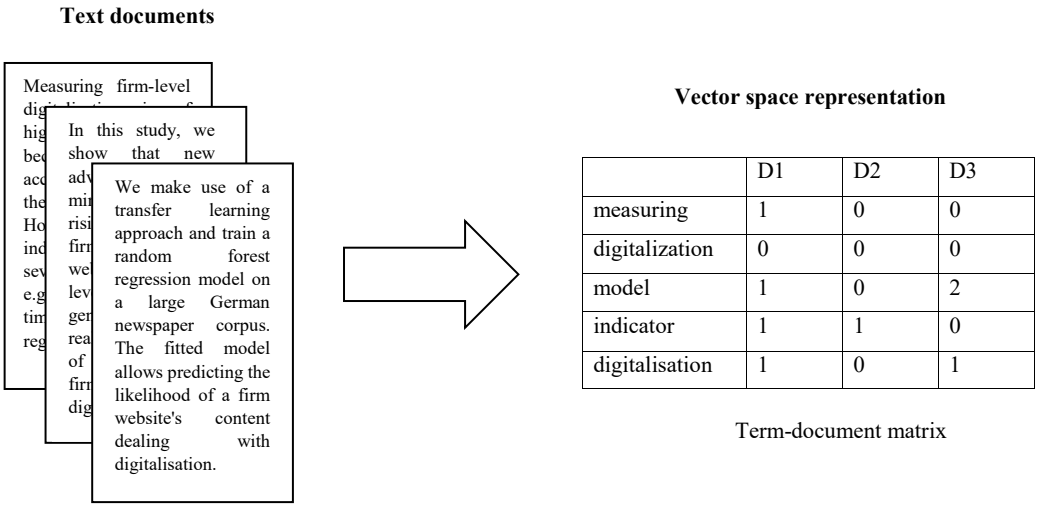


Figure 3.4: **Transformation of a text data set (left) to a term-document matrix M (right).** Matrix columns refer to text documents, and rows refer to words. Own illustration.

3.3.3 Model Training and Predictions

For the model training, we use 75% of the news articles.[50] Furthermore, grid-search optimisation is used for the selection of suitable hyperparameters[51] and a five-fold cross-validation is performed, inter alia, to identify overfitting. The term overfitting refers to a situation in which a model is too well fitted to the training data and then performs worse on the test data (Hastie et al. 2001).[52] The trained model is then applied to the texts of firm websites.

As a result, we calculate about 663K continuous digitalisation scores for the year 2018. For 2020, we have 894K scores and 950K for 2022. The number of firms contained in all three data sets amounts to 278K. All analyses hereafter refer to this subset unless stated otherwise. The predicted firm-level digitalisation score is between zero and one but can be re-scaled.

3.4 Plausibility of Digitalisation Scores

The next section deals with the plausibility of the derived digitalisation indicator. First, we present the performance metrics of the model (Section 3.4.1). Second, the validity is verified by comparing the novel web-based indicator with established digitalisation indicators (Section 3.4.2).

[50] The remaining 25% of the news articles are used as test data.

[51] Hyperparameters are set before training. We consider the following: number of features, minimum/maximum document frequency of words, stop words, number of estimators, and tree depth.

[52] We use the scikit-learn Python package (`https://scikit-learn.org/stable/` [Last accessed: 24.02.2024]) and fix the random seed to make the results reproducible.

3.4.1 Performance on Newspapers

We use the following metrics in our paper: *precision*, *recall*, *f1-measure*, *accuracy* (see formulas in Equation 3.1), *receiver operating characteristic* (ROC) curve, and the *area under curve* (AUC) value. A detailed discussion of these evaluation metrics is provided by Fawcett (2004) and Powers (2011).[53] The receiver operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) for a series of classification thresholds that are used to convert continuous predictions into class assignments. For example, class A is assigned if the prediction is smaller than 0.5; otherwise, class B is assigned. Lastly, the area under the curve (AUC) value is defined as the area under the ROC curve.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.1a}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3.1b}$$

$$\text{F-1 measure} = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{3.1c}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.1d}$$

$$\text{TPR} = \frac{TP}{TP + FN}, \tag{3.1e}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3.1f}$$

For a statistical model to be used in a transfer learning approach, good out-of-sample performance is a prerequisite. The evaluation metrics in Table 3.2 confirm that the model learns from news articles and makes highly accurate predictions about unseen observations. Model performance (precision and accuracy) is greater than 0.9 for both classes. However, the recall is slightly lower for the class *digitalisation*. The ROC curve is shown in Figure B.1, and the AUC value is 0.98. To evaluate the predictive power of our model, we use a baseline model that always predicts the majority class. The share of news articles about the topic of digitalisation is about 12%, giving a baseline accuracy of 0.88. Thus, our machine learning model performs better than the baseline model and learns from the data. In summary, our model provides good predictions on unseen data and is capable of generalising.

[53] The predictions fall into the classes *true positives* (TP), *false positives* (FP), *false negatives* (FN), and *true negatives* (TN). Furthermore, we use the *true positive rate* (TPR) and the *false positive rate* (FPR). The total number of predictions is denoted by N. Formulas for precision and recall are shown for the positive class. Please note that the recall for the positive class is equal to TPR.

Table 3.2: **Evaluation metrics for the news data test sample.**

Metrics on test sample					
classes	precision	recall	f-1 measure	support	accuracy
non-digitalisation	0.98	0.99	0.98	11,164	
digitalisation	0.93	0.83	0.87	1,498	
				12,662	0.97

Notes: The metrics precision, recall, f1-measure, and support are reported separately for the two classes, and accuracy is reported for both together. *Support* is the number of observations of each class.

To test the decision-making in the random forest, we look at the most important features. For this, the Mean Decrease in Impurity (MDI) feature importance is calculated (Breiman et al. 1984). The most important words and their feature importance values are presented in Table B.3. The list of words shows a clear link to the topic of digitalisation. However, words negatively correlated with digitalisation, such as *analog*, may also be included in this list. Some words are only relevant in combination with other words. For example, the word *apple* can refer to the firm or the fruit in English texts. Semantics only become clear when we look at the word in context or when we consider multiple decision layers of the trees in the random forest. To sum up, the random forest model sorts most of the unseen news article texts into the correct class and guarantees some interpretability of the decision rules.

3.4.2 Validity

Next, we present the predicted digitalisation scores and check the plausibility of the indicator. Our indicator is compared to several established digitalisation indicators at the industry, regional, and firm size level.[54] First, we examine the distributions of our web indicator with respect to firm characteristics. Second, we compare the novel indicator to survey-based data from the Mannheim Innovation Panel (Rammer et al. 2021) and the Eurostat *ICT usage in enterprises* data.[55]

Firstly, Figure 3.5 shows the predicted continuous firm digitalisation scores. The calculations are carried out for around 278K firms that were scraped in the years

[54]Data on NACE codes (<1%), employee counts (29%), and on the German federal state in which a firm is located (<1%) are not available for all MUP firms. The NACE codes are the *Statistical Classification of Economic Activities in the European Community* (`https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF` [Last accessed: 24.02.2024]). The relative shares of missing observations are listed in brackets for 2018. Therefore, we use subsets of the firms for which the data are available to validate the predictions.

[55]The *ICT usage in enterprises (isoc_e)* data are based on the annual surveys on *ICT usage and e-commerce in enterprises* by the National Statistical Institutes or Ministries in Europe. For more information, see `https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm` [Last accessed: 24.02.2024].

2018, 2020, and 2022. The distribution shows that many firms have a low digitalisation score and that there is a long tail on the right side of the distribution, representing highly digitalised firms. Furthermore, we find a mass point in the range of 0.7. The fact that we use a binary classifier for the news data, i.e. our model is designed to make discrete decisions, is one possible explanation for this mass point.[56] There are no strong outliers in the distribution, i.e. large jumps. The average firm

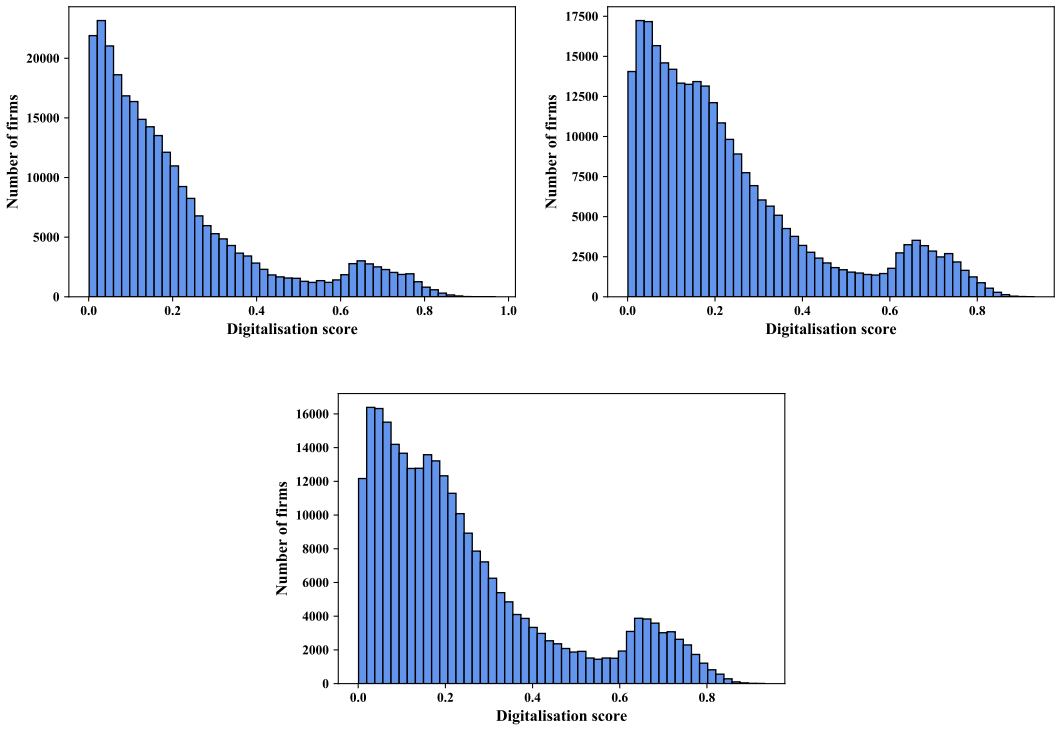


Figure 3.5: **Number of firms per digitalisation score interval.** The predictions are based on the websites in 2018 (top left), 2020 (top right), and 2022 (bottom). Own illustrations.

digitalisation is 0.205 in 2018, 0.237 in 2020, and 0.241 in 2022 (see Table 3.3). According to our indicator, firms have become more digitalised in recent years. In Figure B.2, it is confirmed that the digitalisation score has increased for the majority of the firms over the period observed. However, it is important to note that there are some firms for which we see a decline in digitalisation. The websites of these firms have often undergone significant changes, which affected their digitalisation score.

[56] A bimodal distribution could be a possible explanation for this finding (two groups: non-digital and digital firms). The predictions for the news article test data show a comparable distribution, i.e. most texts have a low predicted score, and the distribution has a small mass point at around 0.7.

Table 3.3: **Summary statistics for the digitalisation scores in 2018, 2020, and 2022.**

Year	∩	#Obs.	Mean	Median	Std.	Min.	Max.
2018	No	663K	0.2050	0.1412	0.1979	0.0011	0.9680
2020	No	894K	0.2376	0.1783	0.2049	0.0008	0.9740
2022	No	950K	0.2415	0.1807	0.2067	0.0008	0.9555
2018	Yes	278K	0.2073	0.1447	0.1978	0.0011	0.9680
2020	Yes	278K	0.2357	0.1764	0.2039	0.0008	0.9296
2022	Yes	278K	0.2456	0.1848	0.2079	0.0008	0.9320

Notes: Each prediction is based on the textual content of a firm website. The first three rows show all firms per year, and the last three rows present firms available in all three years.

Secondly, Figure 3.6 shows the average digitalisation scores by industry, i.e. the data are aggregated to 2-digit NACE codes. The *ICT* industry (NACE codes 26, 61–63) is highly digitalised, and the *food and beverage service activities* industry (NACE code 56) has a low level of digitalisation. For this analysis, it is essential to have a minimum number of observations for each industry to avoid sensitivity to outliers.[57]

For the next analysis, we use the industry definition from the Eurostat ICT survey, which is also based on aggregated 2-digit NACE codes. We use the data on *"buy cloud computing services used over the internet"* (2018), *"enterprises who have ERP software package to share information between different functional areas"* (2017), *"enterprises' total turnover from e-commerce"* (2018), *"enterprises analysing big data from any data source"* (2018), and *"use two or more social media (as of 2014)"* (2017) for the analyses. The data are available, aggregated by industry or firm size.[58] We select these variables because they are available at a fine granular level and are part of the *integration of digital technology* metrics in the DESI.[59] As several variables are not available from 2020 to 2022, we only provide a benchmark value for 2018. To evaluate our approach, we calculate the average of the five variables and use this as a composite indicator. Figure 3.7 shows the average web-based digitalisation scores for each year and industry, as well as the Eurostat composite indicator. There are some larger deviations between the Eurostat ICT survey data and the web indicator, e.g. for *"computer, electronic and optical products"*. Both indicators seem to be quite similar on the industry level, and our indicator does not produce counter-intuitive results.[60]

[57] In the context of this study, the observation count in each industry is at least 50 firms. Industries with coverage below 50 firms are removed from the visualisation to ensure the anonymity of the firms and to reduce the instabilities of the predictions due to low observation counts. Also, only a few 2-digit industries are dropped due to low numbers; other missing 2-digits are just not defined.

[58] For industries: without data from the financial industry; restricted to firms with at least 10 employees. For firm sizes: restricted to firms with at least 10 employees.

[59] The DESI is available at an aggregated level and, thus, not directly comparable with our results.

[60] For the industry data, we calculate the Pearson correlation coefficient, i.e. corr(web-based indicator 2018, Eurostat ICT survey) = 0.81 and corr(web-based indicator 2020, Eurostat ICT survey) = 0.81. The district data are not available for the *Digitalisierungskompass 2018*, and the number of firm-size classes is too small to reliably calculate a correlation coefficient.

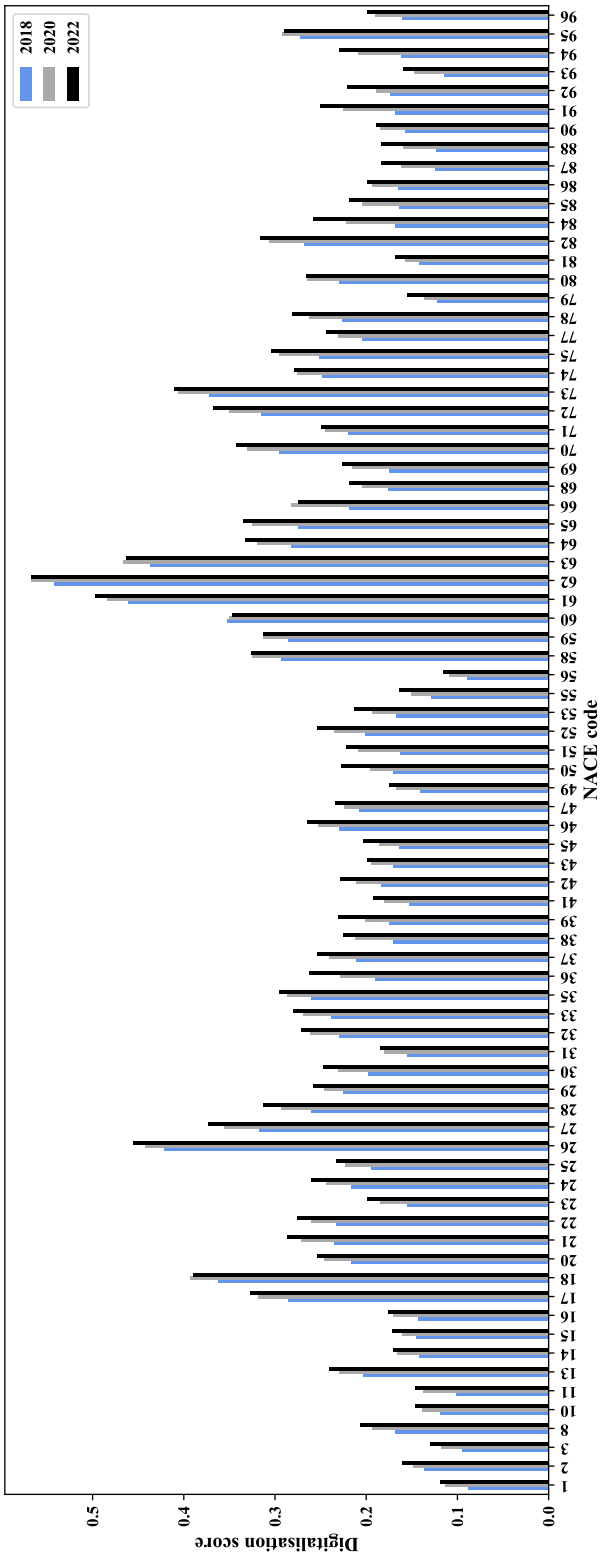


Figure 3.6: **Average digitalisation scores per industry in 2018 (blue colour), 2020 (grey colour), and 2022 (black colour).** The scores are based on the web data. The industries are defined by the 2-digit NACE codes. Groups with less than fifty observations are not shown. Own illustration.

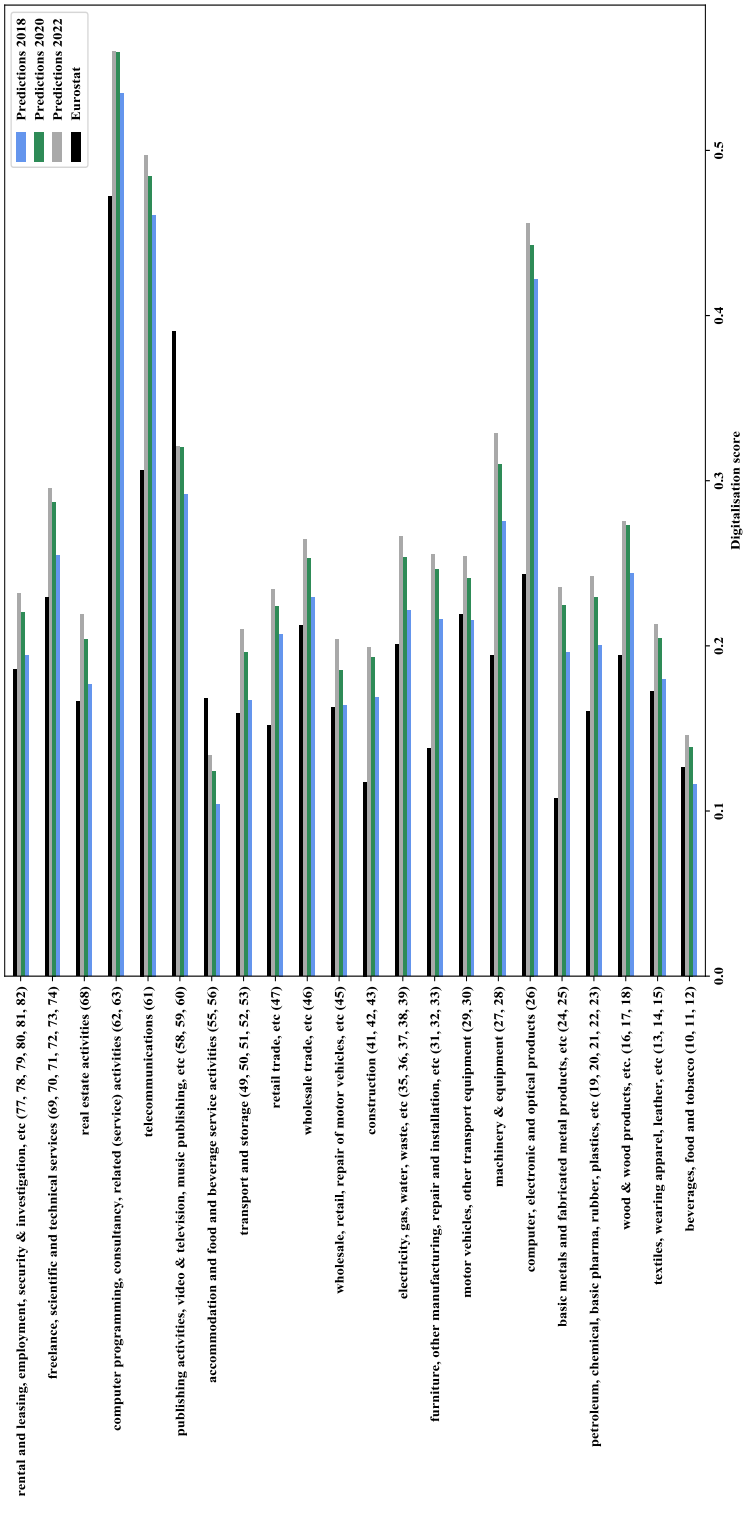


Figure 3.7: **Average digitalisation per industry in 2018, 2020, and 2022, respectively, based on MUP, Eurostat ICT survey (2017/2018), and web data.** Groups with at least fifty observations are shown. The industries are defined by the 2-digit NACE codes (see brackets). Own illustration.

For example, the industries *"computer programming, consultancy, and related (service) activities" (62-63)*, and *"telecommunication" (61)* have a high score, but *"beverages, food, and tobacco" (10-12)* has a low score.

Thirdly, Figure 3.8 shows the data for different firm-size classes (10-49 employees: *small*, 50-249 employees: *medium*, and 250+ employees: *large*). The figure shows that the larger the firm, the higher the level of digitalisation. The difference is evident across indicators, for example, when comparing large and small firms.[61] However, the differences between medium-sized and small firms in 2022 are minor.

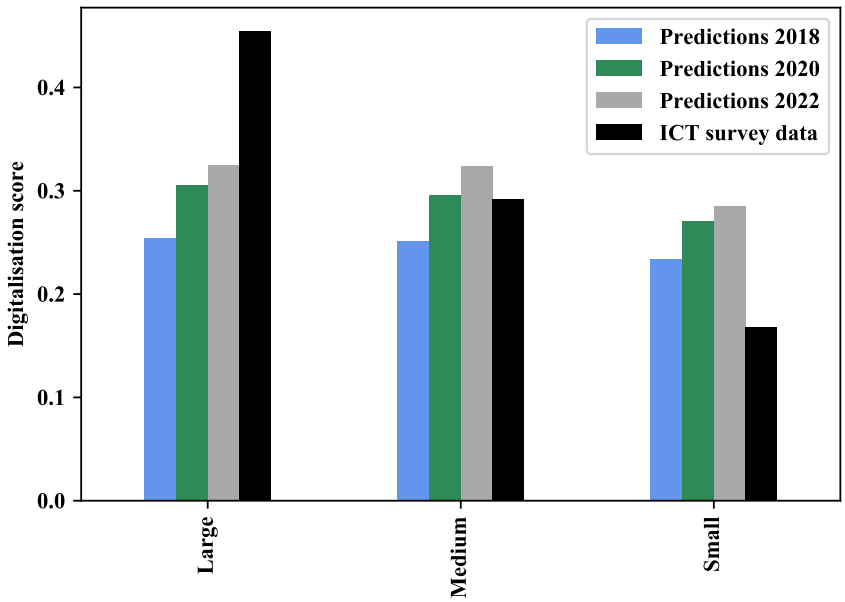


Figure 3.8: **Average digitalisation per firm size in 2018, 2020, and 2022, respectively, that is based on MUP, Eurostat ICT survey (2017/2018), and web data.** The data are grouped by the employee count (10-49 = Small, 50-249 = Medium, and 250+ = Large). Own illustration.

Our results suggest validity regarding regional differences: Figure 3.9 shows firm digitalisation for German districts. The first three figures show the web indicators for 2018, 2020, and 2022, respectively. The illustration on the bottom right is based on the Digitalisation Compass 2018 by Prognos AG[62] and uses a different scale. A dark red colour illustrates a high average level of digitalisation within a German district. Similar patterns are observed, e.g. the eastern part of Germany is less digitalised, and big cities, such as Berlin and Munich, are more digitalised than rural areas.

[61] Our results match findings from other studies, e.g. Büchel et al. (2020).

[62] Digitalisation Compass 2018: `https://www.prognos.com/de/projekt/digitalisierungskompass-2018` [Last accessed: 24.02.2024].

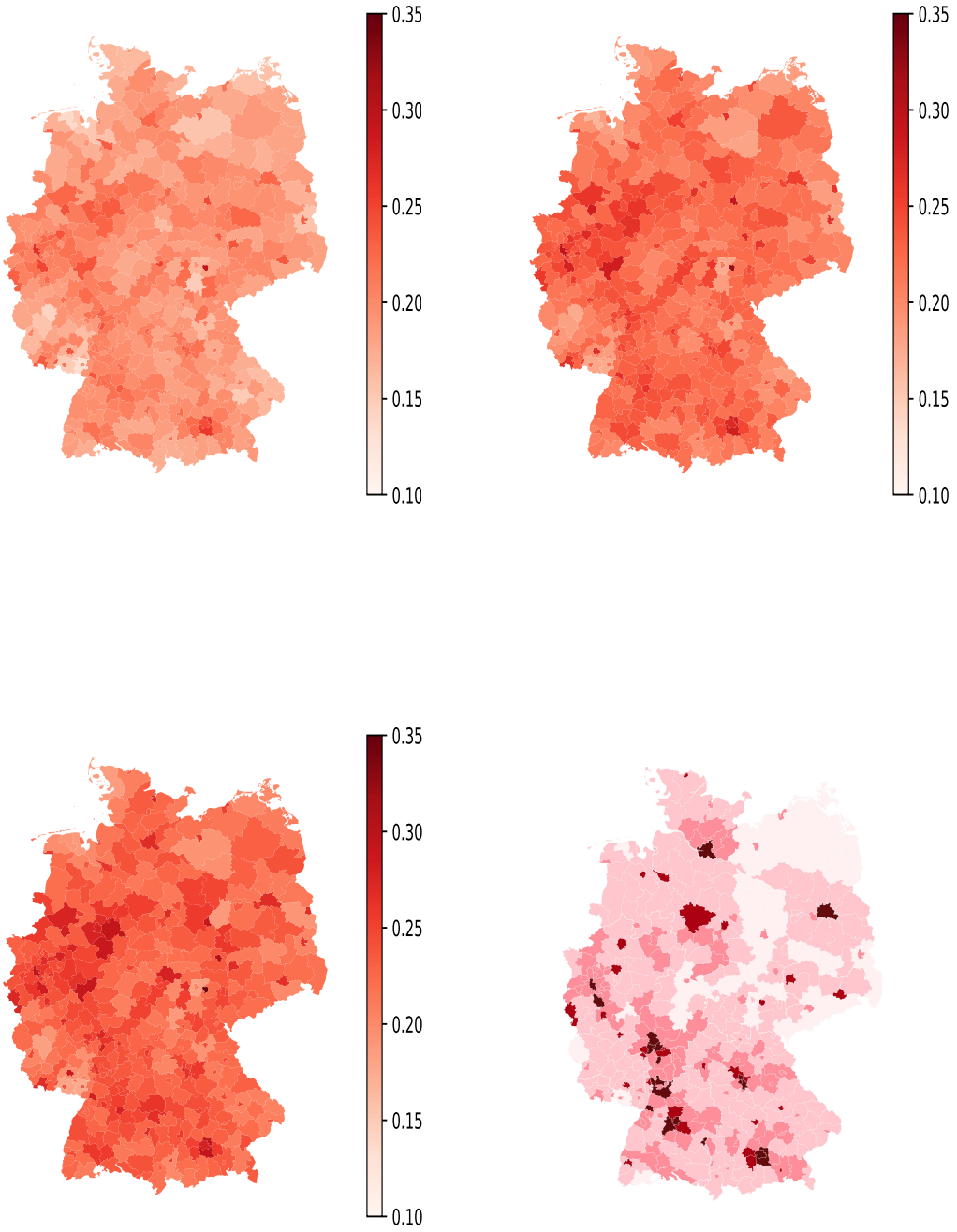


Figure 3.9: **Digitalisation at the district level. Average digitalisation scores per German district in 2018 (top left), 2020 (top right), and 2022 (bottom left) based on MUP and web data. Bottom right: Digitalisation Compass 2018 by the Prognos AG & index Gruppe.** The first three figures are our own illustrations. The source of the bottom right figure is `https://www.handelsblatt.com/politik/deutschland/digitalisierungskompass/` [Last accessed: 24.02.2024, the webpage is currently not available] and its description is available at `https://www.prognos.com/de/projekt/digitalisierungskompass-2018` [Last accessed: 24.02.2024].

Fourthly, we use the MIP 2020 survey to create a firm-level digitalisation indicator. Table B.4 provides a list of eight survey questions with a focus on digitalisation. Possible answers to the survey questions are *none*, *low*, *medium*, and *high*. The answer *none* is re-coded to 0, *low* to 1, *medium* to 2, and *high* to 3. The MIP indicator covers 894 firms and is defined as the unweighted average of the survey answers.[63] Figure 3.10 illustrates the MIP indicator and its link to the web-based digitalisation indicator. There is a positive link between both indicators, as indicated by the regression line. In summary, our indicator appears to be plausible with respect to size classes, districts, industries, and at the firm level.

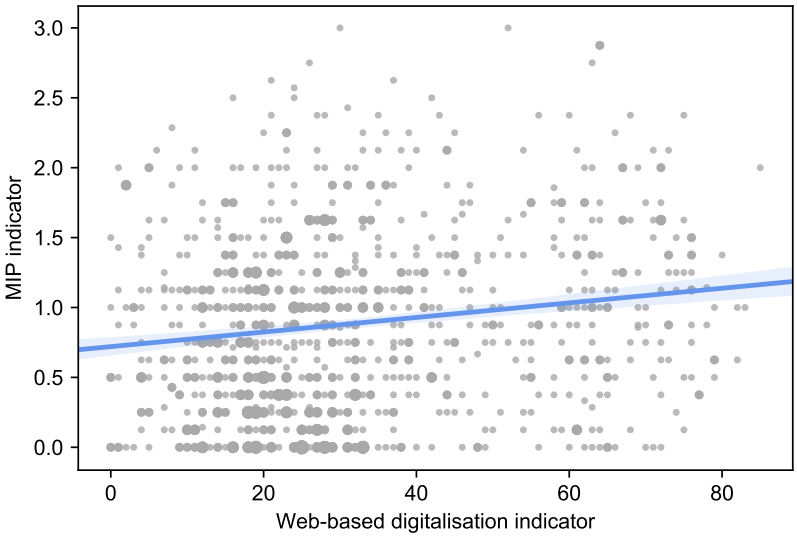


Figure 3.10: **Comparison of the web-based digitalisation indicator (2020) and the MIP digitalisation indicator (2020).** The grey dots represent the individual observations; the dot size scales with the observation count. The blue line is an estimate based on a linear regression; light blue illustrates the 90% confidence interval. Own illustration.

3.5 Use Case: Firm Resilience

Digitalisation is associated with a wide range of measures for firm performance. Economic literature provides evidence on the links between firm resilience and key firm characteristics such as digitalisation, e.g. Conz & Magnani (2020), Saad et al. (2021), Elgazzar et al. (2022). The relationship is found across different types of crises. Bertschek et al. (2019) show that the productivity of ICT-intensive firms was hit less hard during the economic crisis in 2008 and 2009 by analysing data from twelve countries and seven industries. More recently, Abidi et al. (2022) use data

[63]The number of observations is small, because not all firms answered the survey questions and were also recorded using web scraping.

from the Middle East and Central Asia to find that digitally-enabled firms had a smaller reduction in sales by about 4 percentage points during the COVID-19 health crisis. Thus, if our web-based indicator measures indeed digitalisation, it should also be positively linked to the economic resilience[64] of firms during the COVID-19 health crisis.

We use the MUP data (Bersch et al. 2014) and our digitalisation indicator to examine the relationship between digitalisation and the resilience of firms during the exogenous COVID-19 shock.[65] For this purpose, we assess the solvency of firms based on their credit ratings.[66] We expect that more resilient firms experienced a smaller decline in solvency during the COVID-19 crisis. Thus, we measure resilience as the change in firms' credit ratings before and after the shock. The credit rating ranges from 1 to 5, and a high number indicates a good rating.[67] Table B.5 shows the summary statistics of the estimation sample. The average credit rating worsened between 2019 and 2021 by 0.09 points. Average firm digitalisation in our sample increased by 0.03 between 2018 and 2020.

In the analyses, we control for firm characteristics at the regional (16 German federal states), industry (21 groups), founding period (4 groups), legal form (14 groups), and employee count level. The descriptive statistics are presented in Table B.6.

Equations 3.2 and 3.3 show the baseline model specifications for the firm resilience analysis. The main specification is presented in Equation 3.4. The variable Δrating_i refers to the change in the credit rating for firm i between 2019 and 2021. The main regressors $\Delta\text{digitalisation}_i$ and digitalisation_i are the change in firm digitalisation between 2018 and 2020 (before and after the crisis started) and the degree of firm digitalisation in 2018 (pre-crisis). Similarly, $\Delta\text{employees}_i$ is the change in

[64] Firm resilience is defined as "the ability of a firm to persist in the face of substantial changes in the business and economic environment and/or the ability to withstand disruptions and catastrophic events" (Acquaah et al. 2011, p. 5528) and is thus an attribute of a firm.

[65] Official firm exit data cannot be analysed during this period because there was no obligation to file for insolvency in Germany (`https://www.gesetze-im-internet.de/covinsag/` [Last accessed: 24.02.2024]).

[66] Some factors that influence the credit rating: credit assessments, annual financial statement data, industry risk, turnover, legal form, firm age, regional risk, order situation, capital, management experience, number of employees, turnover / employee, and capital / turnover. See also: `https://www.creditreform.de/aktuelles-wissen/praxisratgeber/wie-sie-ihren-bonitaetsindex-verbessern` [Last accessed: 24.02.2024].

[67] The MUP data are a panel and consist of semi-annual data points. We use the waves delivered in the middle of the years 2018, 2020, and 2021 for our analysis. The analysis is based on firms with websites scraped in 2018 and 2020. For data preparation, we delete observations as soon as a variable is missing, so that all regression analyses are conducted on the same data set. In the original definition, the credit rating is in the range of 100 and 600, where 100 is the best rating. Our credit rating data preparation is similar to the data processing in Dörr et al. (2021) and Dörr et al. (2022). Firms with a credit rating of less than 100 are deleted, as this indicates that too little information is available about a firm, e.g. a start-up. In addition, the credit rating of firms with values above 500 is truncated to 500, e.g. insolvent firms. For a simpler interpretation, we modify the credit rating using the function $(6 - (\text{credit rating}/100))$. We delete observations marked as duplicates or faulty, firms founded after January 2018, and firms with an exit year before 2018.

employee counts between 2018 and 2020. We expect a temporal delay between an increase in firm digitalisation and the change in firm resilience. For example, the use of a new technology may require a learning period after its introduction.

$$\Delta rating_i = \beta_1 \Delta digitalisation_i + u_i \tag{3.2}$$

$$\Delta rating_i = \beta_1 digitalisation_i + u_i \tag{3.3}$$

$$\Delta rating_i = \beta_1 \Delta digitalisation_i + \beta_2 digitalisation_i + ... + u_i \tag{3.4}$$

Table 3.4 shows the empirical results. Columns (1) - (2) show the baseline estimation results using a time lag and cross-sectional data, but without control variables. The results confirm that digitalised firms (pre-crisis) or firms that increased their degree of digitalisation between 2018 and 2020 are more resilient with respect to the COVID-19 crisis, as the respective coefficients are positive and significant. Column (3) shows estimation results, including control variables. Again, a highly significant positive link is found. If a firm increases its level of digitalisation from zero to one in the period between 2018 and 2020, then, on average, the firm has a credit rating increase of 0.0296 for the years 2019 to 2021. Similarly, the pre-crisis level of digitalisation is positively and significantly linked to firm resilience. However, the coefficients are small with respect to the credit rating range, and the R^2 value is

Table 3.4: **OLS regressions: Firm resilience.**

	(1)		(2)		(3)	
	Δrating 2021 - 2019		Δrating 2021 - 2019		Δrating 2021 - 2019	
Δdigitalisation 2020 - 2018	0.0202***	(0.000)			0.0296***	(0.000)
digitalisation 2018			0.0794***	(0.000)	0.0361***	(0.000)
Δemployees 2020 - 2018					0.00453	(0.551)
employees 2018					-0.000870	(0.662)
Constant	-0.0906***	(0.000)	-0.108***	(0.000)	-0.358***	(0.000)
Founding Period Dummies	No		No		Yes	
Legal Form Dummies	No		No		Yes	
Industry Dummies	No		No		Yes	
Location Dummies	No		No		Yes	
Observations	176,902		176,902		176,902	
R^2	0.000		0.002		0.035	

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The Columns (1) and (2) show the baseline specifications without control variables. Column (3) includes control variables. The number of employees is stated in thousands. The number of observations is reduced by the inclusion of firm attributes, i.e. in particular the number of employees.

only 0.035.[68] The results illustrate the indicator's potential for giving timely and plausible answers to pressing economic issues.

3.6 Discussion

To the best of our knowledge, our approach is the first purely web- and text-based method to measure firm-level digitalisation frequently and on a large scale. Unlike other approaches, we do not use survey data, which require extensive, costly, and time-consuming data collection and processing.

Websites are part of the public image of a firm that might be strategically altered. As a result, the degree of digitalisation may be biased. However, we do not consider this to be a severe issue, as our indicator is in line with other digitalisation indicators. Some specific firms might talk too little or too much about digital technologies on their websites, which may result in noise. Website data might also depend on firm characteristics. For instance, large firms in the chemical industry might use their websites to report about digitalisation efforts in a different way than a software development firm. Some firms might even use buzzwords on their websites to strengthen the public image of the firm or to improve ranks in search engines. Related to this, Vos (2009) discusses firm-level deceptions in climate protection efforts.

As general purpose technologies have an inherent potential for constant improvement, their characteristics can change over time. Hence, the concept of digitalisation may constantly evolve as new digital technologies appear. For instance, high-performance workstation computers were indicative of a digitalised firm a decade ago. Today, cloud solutions are preferable in many situations. The described problem can be tackled with an extension of the model, e.g. by ignoring older news articles or reducing their weight with respect to their age. The abstract definition of the term digitalisation is also a big challenge and illustrates the issue that we cannot evaluate against a clearly defined ground truth. Therefore, an important part of this work is to provide convincing arguments that the derived web-based indicator provides plausible results. The use of additional news article sources might provide an even broader and perhaps better definition of digitalisation. So far, articles from four news providers have been used, but they do not necessarily reflect all aspects of digital technologies. Further, it may be argued whether the content of the news articles is sufficient to recognise all digital technologies. If not, this may create a bias in the data and model. Subsequently, the question becomes evident whether news articles are a good choice for learning the definition of digitalisation.

[68] Adding the firm-level credit rating for 2018 to the model preserves the sign and significance levels of digitalisation 2018 and Δdigitalisation 2020 - 2018 (R^2 = 0.100). However, the sign of the constant becomes positive, and the credit rating in 2018 has a negative and significant coefficient.

Information like the firms' industry or size could also be used as input to the statistical model. For example, IT firms are expected to be more digital than the majority of non-IT firms. However, this requires information on the industry association for every firm in Germany. Other machine learning methods than the random forest model might also be more suitable for a transfer learning task, e.g. deep neural networks (e.g. LeCun et al. 2015). Further, the transfer of a model to a related task is associated with various challenges: For example, the news article and firm website texts can be quite different, so the vocabularies are not comparable. Lastly, natural language processing methods have not yet been exhausted, e.g. the use of word embeddings. Hence, our indicator still has room for improvement.

The web scraping of the data has some drawbacks. We cannot detect the correctness of the websites' content, and some dynamically loaded content may be missing. Website crawling can be further improved for dynamic websites, but a tailor-made solution for every firm website is not possible.[69] The content of firm websites varies in up-to-dateness because some firms update their websites much less frequently than others. The data processing is optimised for texts in German or English, but we cannot give a reliable estimate of how many firm website texts are neither German nor English. However, Axenbeck & Breithaupt (2021) show in a comparable study that only about 2% of web pages fall into this category.

Although our indicator has some weaknesses, which we are aware of, the results show that we measure firm digitalisation in a reliable, fast, and granular way. Therefore, we believe that this paper makes a valuable contribution to the literature on indicators.

3.7 Conclusion

In this paper, we introduce a web-based indicator for digital technology adoption and, thereby, tackle many problems with traditional indicators. For this purpose, we use about 25K news articles to train a random forest model that predicts whether a text is about digitalisation. Using a transfer learning approach, we apply the trained model to websites of German firms to create a large-scale indicator for digital technologies. We predict the scores of German firms for the years 2018, 2020, and 2022. Comparisons with established digitalisation indicators show that our approach provides plausible results at the firm, regional, and industry level and for different firm-size classes.

[69] Future studies could use other Python software packages, such as *Selenium* instead of *Scrapy*. Dynamically loaded content, e.g. after a click on a web element, can still not be captured on a large scale.

Our web-based indicator is a cost-effective way to measure firms' adoption of digital technologies as it can be updated quickly and covers many firms. The indicator covers all firm-size classes, regions, and industries in Germany. Thus, it does not have the shortcomings of traditional survey-based digitalisation indicators.

We illustrate the indicator's potential for giving timely answers to pressing economic issues by analysing the link between digitalisation and firm resilience during the COVID-19 shock. We find results that are consistent with the related literature, demonstrating the successful application of our web-based indicator. The analysis is only one example of a wide range of applications. Besides research, it is also applicable for economic policy advice and consulting, e.g. analysing the heterogeneity of digitalisation.

Chapter 4

Intangible Capital Indicators Based on Web Scraping of Social Media

Joint work with Reinhold Kesler, Thomas Niebel, and Christian Rammer.

4.1 Introduction

Knowledge-based capital is a key factor in productivity growth (Corrado et al. 2022). Over the past 15 years, it has been increasingly recognised that knowledge-based capital comprises much more than technological knowledge and that these other knowledge components are essential for understanding productivity developments and the competitiveness of both firms and economic aggregates (sectors, regions, and economies). In the tradition of the new growth theory (Romer 1986, Lucas 1988, Romer 1990), knowledge-based capital, also denoted as intangible capital, is often measured by the stock of technological knowledge and is approximated by accumulated R&D expenditures or the stock of patents.

Corrado et al. (2005, 2009) have proposed a classification of intangible capital goods that comprises three main components: (1) *innovative property*, (2) *computerised information,* and (3) *economic competencies*. While the first two components are already covered by different statistical surveys (R&D surveys on technical knowledge, innovation surveys on technical and non-technical innovation-related knowledge, and investment surveys capturing expenditures on computerised information such as software and databases), comprehensive statistical data on *economic competencies* are scarce. These competencies include, in particular, *firm-specific human capital*, *organisational capital*, and *brand equity*.

In this paper, we describe a new way of measuring investments in *economic competencies* that do not require firm surveys but are calculated based on publicly available data from online platforms. We focus on two types of economic competencies: investments in *brand equity* and investments in *firm-specific human capital*. For *brand equity*, we use the number of *likes* a firm has on Facebook as an indicator. Individual ratings (by employees) on the employer branding and review platform Kununu provide information for both the *firm image* (*brand equity*) and on-the-job training/career

development (*firm-specific human capital*). Both platforms are market leaders in their respective segments in Germany. Compared to survey-based data, publicly available platform data provide much broader coverage at substantially lower costs, a much higher timeliness, and a higher frequency.

However, the quality of platform data might be contested. To provide a first test of data validity, we compare the two newly developed indicators with survey-based expenditures on marketing (*brand equity*) and on-the-job training (*firm-specific human capital*), using data from the Mannheim Innovation Panel (MIP), which is the "German contribution to the European Commission's Community Innovation Surveys" (ZEW 2024*a*). The results show a positive and significant relationship between firm-level expenditures for marketing and on-the-job training and the respective information from the online platforms Facebook and Kununu. Based on this result, we explore the possibility of predicting *brand equity* and *firm-specific human capital* with machine learning methods. However, the additional explanatory power of the platform data is limited.

The rest of the paper is structured as follows: Section 4.2 provides an overview of the economic literature on intangible capital and the literature on the use of platform-based data for economic research. Section 4.3 introduces our data collection approach, the survey data used for comparison, and provides descriptive statistics of our estimation sample. The empirical approach and the results of our OLS regressions comparing the platform and the survey data are presented in Sections 4.4 and 4.5. Section 4.6 explores the possibility of predicting firm-level intangible capital expenditures with machine learning methods. Section 4.7 concludes.

4.2 Literature Review

Our research relates to the ongoing efforts to improve the measurement of knowledge-based capital. The terms knowledge-based capital and intangible capital are used as synonyms in this strand of the literature. Research related to intangible capital was largely initiated by the seminal papers of Corrado et al. (2005, 2009), which proposed a framework for the classification of intangible capital.

On the sectoral and total economy levels, a large number of studies have been released in the past ten years, trying to improve the measurement of intangible capital and, more importantly, analyse the economic impact of intangible capital. Notable contributions with respect to the economic implications of intangible capital are, among others, Corrado et al. (2013), Roth & Thum (2013), Chen et al. (2016), Niebel et al. (2017), Corrado et al. (2017), Chen (2018) and Adarov & Stehrer (2019). Roth (2019, 2022) offers a recent review of the literature, while Haskel & Westlake (2018) provide a more comprehensive overview of the topic.

Apart from measuring intangibles at the sectoral and total economy levels, several firm-level surveys with a special focus on intangibles were conducted (Awano et al. 2010*a*,*b*, Perani & Guerrazzi 2012, European Commission 2014). Furthermore, there are studies analysing the impact of knowledge-based capital on firm performance that are based on pre-existing general firm surveys (Crass et al. 2014, Di Ubaldo & Siedschlag 2020, Rammer et al. 2020, Roth et al. 2023).

The paper also relates to the growing literature on using web-scraped data for economic research. Claussen & Peukert (2019) show a strong increase in articles published in journals on the Financial Times 50 list between 2000 and 2018 that use data crawling for different use cases with data obtained from online platforms. Specifically, with the availability of many potential data sources on the internet and growing computing power, the possibility of viable web-based indicators has expanded in the last decades. For example, Ginsberg et al. (2009) use Google search query data to predict influenza-like disease activity in the United States. Similarly, Choi & Varian (2012) use search engine data to develop a set of economic indicators, e.g. for unemployment claims. Besides search query data being a viable predictor for a wide range of outcomes nowadays[70], other studies specifically leverage online platform data to approximate and predict economic outcomes. For instance, restaurant data from Yelp has been employed to measure local business activity, neighbourhoods' socioeconomic characteristics, and consumption patterns (Glaeser et al. 2018, Dong et al. 2019, Davis et al. 2019).

In recent years, there has also been a lot of research in the field of web-based innovation indicators. For example, Gök et al. (2015) develop an indicator for R&D activities based on website data. Similarly, Pukelis & Stanciauskas (2019) and Kinne & Lenz (2021) use texts on firm websites to create a statistical model to predict a firm's innovation status. Axenbeck & Breithaupt (2021) investigate the relationship between a wide variety of firm website characteristics and the firm's innovation status. In addition, Krüger et al. (2020) make use of texts and hyperlinks on firm websites, create an inter-firm network, and investigate its relationship with firm innovativeness.

Social media data have also been used to analyse the *brand equity* activities of firms, as social media has become a key channel for marketing and customer interaction (Bruhn et al. 2012). While many studies aim at deriving insights on firms' marketing performance (see Misirlis & Vlachopoulou 2018) or assessing the use of social media by firms (see Arora et al. 2014), fewer studies are linked to the subject of this paper, deriving a measure of *brand equity* at the firm level. For example, Coursaris et al. (2016) calculate an engagement score based on the number of likes, comments, and shares of firms' posts on Facebook and find that the engagement level

[70] For a brief overview on nowcasting, see Gentzkow et al. (2019).

has a positive effect on brand equity. Chung et al. (2015) use information (posts, comments, and likes) from Facebook pages of 100 large Korean firms and demonstrate that these indicators are positively related to market performance. Tirunillai & Tellis (2012) use indicators on chatter activities (product reviews on websites) for fifteen firms and find that the volume of chatter has a strong positive relationship with firms' returns. Luo et al. (2012) manually classify web blogs on nine large firms from the computer and software industry in terms of positive or negative sentiment along with blog volume data and find a strong leading effect of this brand indicator on firm equity value. All these studies focus on a relatively small number of large firms since they tend to be much more engaged on social media than small and medium-sized firms. In our study, we add to the literature by deriving social media-based indicators for a large number of firms across all industries and size classes.

Using data from professional networking platforms such as LinkedIn or employer review platforms (such as Glassdoor or RateMyEmployer) to assess firms' human capital is much less frequent. Most research papers that use this type of social media data focus on its role for employees (see Aguado et al. 2019), employee response to firm events (see Gortmaker et al. 2022), and recruiting (see Zide et al. 2014, Chiang & Suen 2015) rather than using it as a measure of employers' human capital. Ji et al. (2022) use data from the employer review platform Glassdoor to derive indicators of job satisfaction and find a positive association with lower financial reporting risk. Pisano et al. (2017) use LinkedIn data to analyse whether ownership concentration affects the disclosure of human capital information via social media platforms. Banerji & Reimer (2019) analyse the social connections of firm founders based on LinkedIn information to investigate the impact of connectedness (as a specific indicator of firm-specific human capital) on funds raised and find a strong positive relationship.

In this paper, we contribute to the literature in two ways. First, we describe a fairly generalisable method for matching and linking firm-level survey data and platform-based data. Second, using publicly available information from social media, we are able to derive new indicators of *firm-specific human capital* and *brand equity* that can complement firm surveys, thus improving the measurement of knowledge-based capital.

4.3 Data and Descriptive Statistics

4.3.1 Data Collection

Identifying Platform Profiles

Our data collection efforts aim to identify firm information related to *brand equity* and *firm-specific human capital* from digital platforms (Facebook and Kununu[71]) and compare this information with survey-based indicators on these two aspects of firms' economic competencies taken from the German Innovation Survey (the *Mannheim Innovation Panel* - MIP). The survey rests on a stratified random sample and is representative of the entire population of German enterprises (see Peters & Rammer 2013, ZEW 2024*a*).

Our data collection strategy is illustrated in Figure 4.1. For all firms that participated in the MIP survey conducted in 2017 (1), we create a specific search in Google based on the firm's website URL (2) to derive the platform URL of each firm for the platforms Facebook and Kununu. We download the page behind the URL and check whether it is the firm in the survey. If there is a match (3), we can analyse the platform profiles (4).

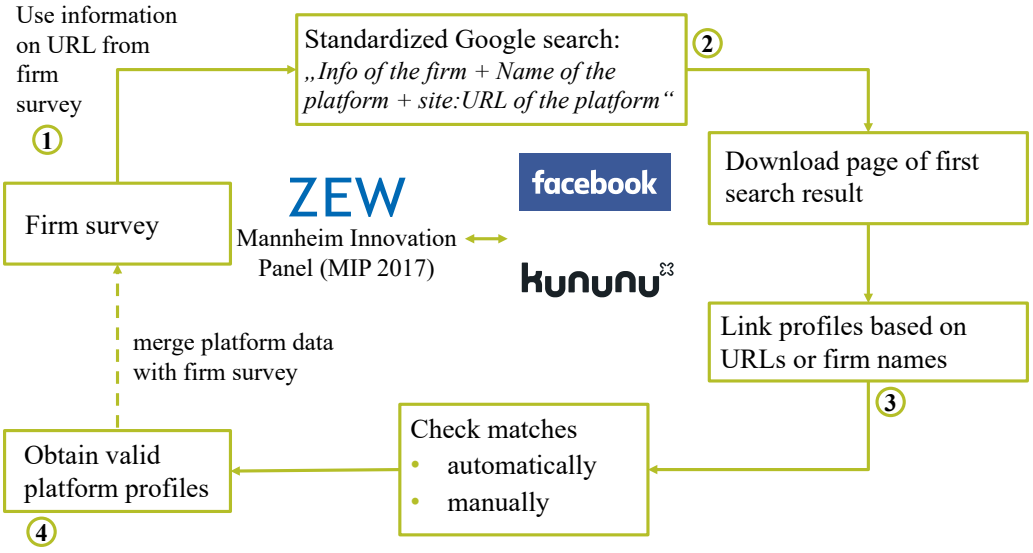


Figure 4.1: **Data collection approach.** The firm logos are taken from the websites of ZEW, Facebook, and Kununu. Own illustration.

The Google search included the firm's website URL, the platform name, and the search operator *'site:platform URL'*, which only returns results from the platform website. The approach only necessitates some identifying information from a firm

[71] We also scraped the MIP 2017 firm profiles on Twitter (the new name of Twitter is *X*). The number of firms on Twitter was rather low, so we decided not to use the Twitter data.

and allows it to be generally applied to different online platforms. We assume to be so specific that the first search result must be the platform profile of the searched firm, unless there is none. Subsequently, we take the received platform URL of the first search result and download the HTML code of the page, which is often the start page of the firm on the respective platform. Finally, we extract the initially selected firm's key information, which should ideally also be on the platform page, link it to the survey data, and verify the match.

The sample of our analysis includes 7,498 firms.[72] To verify whether our search, described in Figure 4.1, identified the right firm on the platform, we compare information from the platform profile with the information from the survey, such as the firm name or website URL. For Facebook, this is straightforward since firms are obliged to state their website URL on the start page. On the Kununu page of a firm that is not managed by the firm itself, there is no corresponding imprint. Therefore, we analyse the similarity of the firm name on Kununu and in the MIP. For this purpose, we carry out exact string matching. If no exact match is established, the Python package *fuzzywuzzy* is used to perform fuzzy string matching.[73]

The search was carried out at the end of 2017. We obtained 2,114 platform profiles of firms for Kununu and 1,539 for Facebook, representing 28.2% and 20.5%, respectively, of our sample (see Figure 4.2). In the case of Facebook, the share is comparable to other studies that also retrieve corporate profiles from online platforms (see Bertschek & Kesler 2022). For 598 firms, we found both a Facebook and Kununu page.

Obtaining Kununu Ratings for Training and Image

Kununu is an employer branding and review platform founded in 2007 that was acquired by XING in 2013.[74] Despite having a dedicated website for firms in the U.S., Kununu has a strong focus on the DACH region (Germany, Austria, and Switzerland). Besides an overall score/rating, employees can evaluate their firms within different categories. For our purposes, the individual ratings for *firm image* and *on-the-job training/career development* are the relevant categories (see Figure 4.3). *Firm image* is within the framework for intangible capital by Corrado et al. (2005, 2009) related to *brand equity* as it reflects the firm's public image, which is a major factor for a firm's marketing success. The rating for *on-the-job training/career development* is

[72] In total, 8,278 firms participated in the MIP 2017. For about 9% of the firms, no URL of the firm website was available. A first search result on Google produced Facebook URLs for 7,330 firms and Kununu URLs for 4,759 firms.

[73] For this purpose, we use the *fuzzywuzzy* functions *ratio*, *partial_ratio*, *token_sort_ratio*, and *token_set_ratio* and equally weight the results. In addition, a minimum threshold of 50% is defined. If multiple entries are above the threshold, we choose the MIP entry with the highest matching score. The package is available at `https://pypi.org/project/fuzzywuzzy/` [Last accessed: 24.02.2024].

[74] XING is a German competitor of LinkedIn.

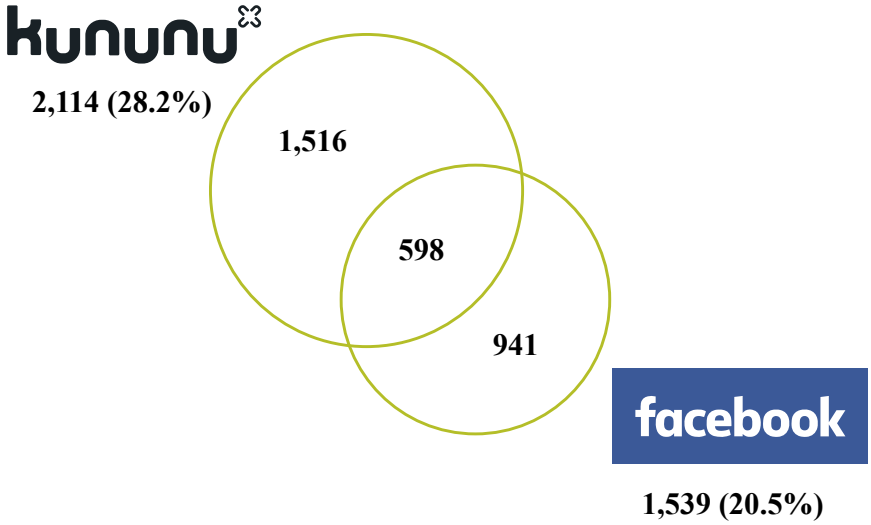


Figure 4.2: **Number of identified platform profiles.** Notes: In total, there are 7,498 firms with valid website URLs that participated in the MIP 2017 survey. Out of these 7,498 firms, we were able to identify 28.2% of the firms with a Kununu page and 20.5% of the firms with a Facebook page. Own illustration; the logos are taken from the websites of Facebook and Kununu.

directly linked to *firm-specific human capital* as it evaluates the relevance and effectiveness of a firm's human capital development efforts from the employees' point of view. Kununu data were collected in August 2018, with historical data dating back to 2010 (see Table C.1).

Obtaining Facebook Likes

Data from Facebook were collected in December 2017 and the first week of January 2018. The relevant data on the firms' Facebook pages are the number of *likes* and the URL of their firms' websites. The latter is needed to check whether our Google search indeed identified the right firm and to merge the number of *likes* on the Facebook page with the survey data on intangibles and other firm characteristics in the MIP 2017 survey. Within the framework of intangible capital by Corrado et al. (2005, 2009), the number of *likes* is related to *brand equity* as it reflects positive values associated with a firm in the general public. A high number of *likes* indicates that the firm's efforts to establish a favourable perception of its activities, products, and services have been, at least to some extent, successful. We scraped the start page of the firm profile on Facebook, which includes the number of *likes* and the URL of the firm website (see Figure 4.4). Obtaining historical Facebook data was not possible since Facebook has massively restricted API access (see Table C.1) as a result of the Facebook–Cambridge Analytica data scandal in early 2018.

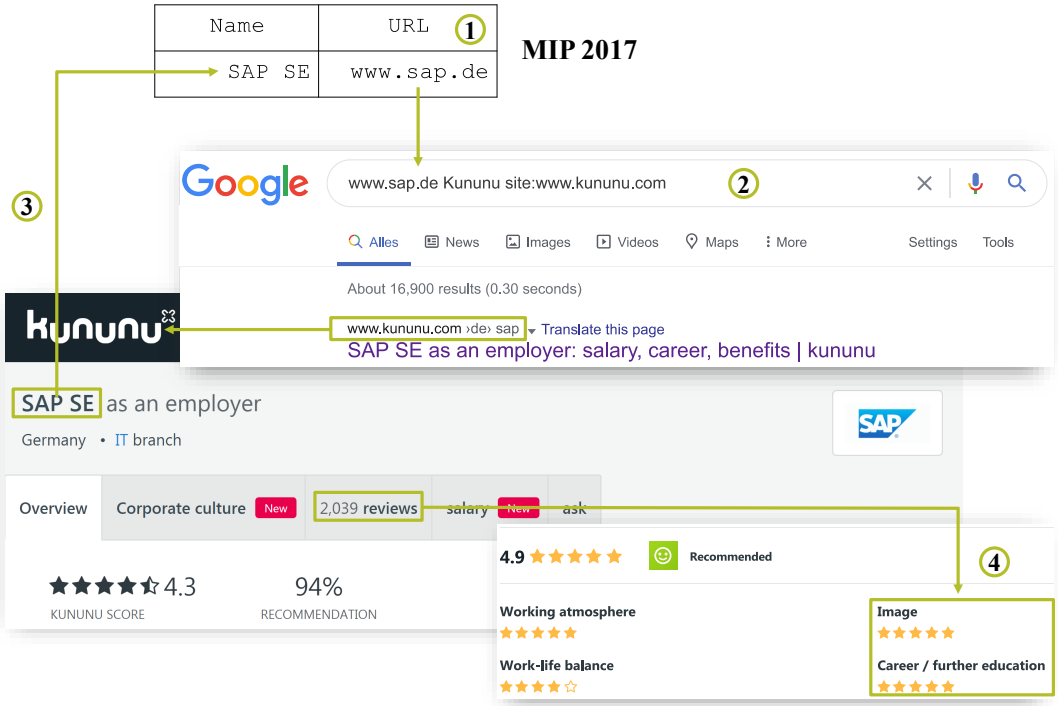


Figure 4.3: **Example of the methodology using the platform Kununu.** This is an example of a Kununu page for a firm. The presented firm is not necessarily in our sample. Screenshots of the Kununu page are in German language and were automatically translated by *Google Translate*. The numbers 1-4 refer to the numbers in Figure 4.1. Own illustration; some parts of the figure are taken from the websites of Google and Kununu.

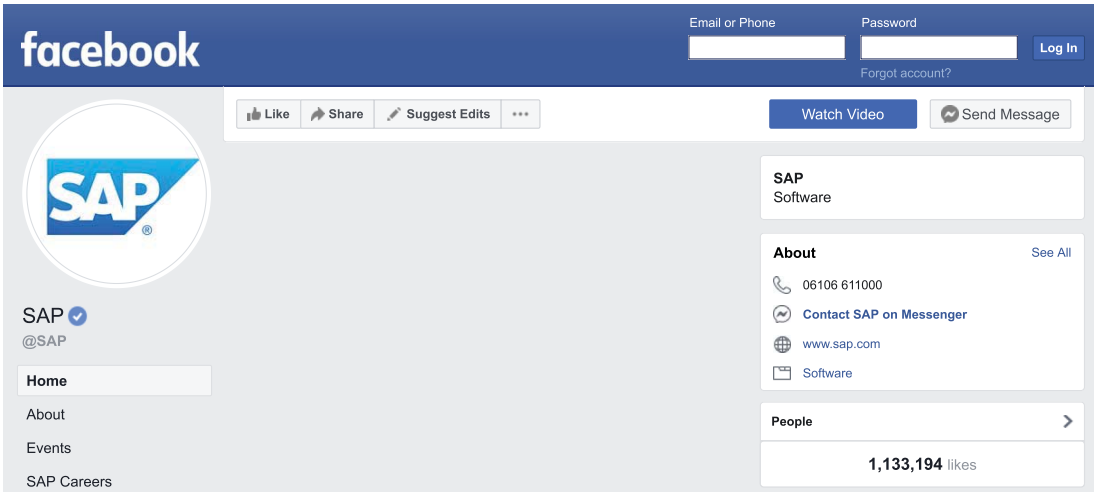


Figure 4.4: **Example of a firm profile on the platform Facebook.** This is an example of a Facebook page for a firm. The presented firm is not necessarily in our sample. Some elements of the Facebook profile of the firm are manually removed for clarity. Own illustration; the figure is based on the website of Facebook.

4.3.2 Survey Data: Mannheim Innovation Panel (MIP)

We use data from the Mannheim Innovation Panel (MIP) to test the relevance of our measures of *brand equity* and *firm-specific human capital* derived from information provided on online platforms. The MIP is the German contribution to the Community Innovation Survey (CIS) of the European Commission and follows the survey methodology of the CIS. The MIP is a stratified random sample of about 13% of the target population and includes firms with at least 5 employees from manufacturing and business-oriented services. The response rate of the 2017 survey was about 25%, resulting in 8,278 observations. For a more detailed description of the survey and data, see Peters & Rammer (2013), Behrens et al. (2017) and ZEW (2024*a*).

For our analysis, two variables from the MIP 2017 survey are used: the amount of expenditures for *marketing* in 2016 and the amount of expenditures for *employee training* in 2016. Marketing expenditures include all in-house and contracted-out expenditures for advertising and branding (incl. commercial marketing), reputation building, conceptual design of marketing strategies, market and customer research, and the installation of new distribution channels. Pure selling costs are not considered marketing expenditures. Employee training expenditures include all in-house and contracted-out expenditures for training and further education of employees, including payroll costs of employees for working time used to attend training. Expenditures for vocational education are not part of training expenditures. 6,339 firms provided data on their marketing expenditures, and 6,419 reported the amount of training expenditures.

4.3.3 Descriptive Statistics

Kununu Data

Table 4.1 and Table 4.2 show the summary statistics of our estimation sample for the analysis of the relationship between our knowledge-based capital indicators stemming from the employer branding and review platform Kununu and the survey-based expenditures for knowledge-based capital (MIP 2017).

We restrict our sample to firm profiles with at least four ratings between January 2017 and August 2018, as there is a trade-off between data quality and the number of data points (i.e. more ratings per firm on Kununu implies better data quality but fewer data points). In this way, we reduce the number of observations with Kununu data on *training* to 813 in the full sample (see Table C.2) and 519 in the estimation sample (see Table 4.1). The number of observations with Kununu data on *firm image* is reduced to 805 in the full sample (see Table C.2) and 492 in the estimation sample (see Table 4.2). These numbers are much lower than the total number of firms (8,278) participating in the MIP 2017 survey (see Table C.4), reflecting the fact that

only a smaller part of the entire firm population is represented on the platform Kununu. Overall, industry *J* (Information and Communication) is overrepresented in our estimation sample compared to the total MIP 2017 sample (see Table C.4). Furthermore, we do have far fewer firms with less than 10 employees in our sample compared to the full MIP 2017 sample (see Table C.6) because the ratings on Kununu come from employees of the firm. For firms with fewer than 10 employees, it is less likely to reach our minimum threshold of four ratings between January 2017 and August 2018.

Table 4.1 shows that the average rating for *on-the-job training* in our estimation sample is 3.31 (scale: 1-5, with 5 being the highest rating; for more details, see Figure C.1 and Figure C.4), while the expenditures for on-the-job training stemming from the MIP 2017 survey are, on average, 1.1 million euros. As there are very large firms with more than 122 thousand employees and almost 47 billion euros in turnover in our sample, the average number of employees is 1,238 and the average annual turnover is about 400 million euros. As expected, the respective median values are much lower. The average number of ratings for a firm on Kununu in our sample for *on-the-job training* is 22.9.

Table 4.1: **Summary statistics - Training: Kununu rating - Estimation sample.**

	N	Mean	Median	SD	Min	Max
Training: Kununu rating	519	3.31	3.38	0.79	1	5
Training expenditures (MEUR)	519	1.10	0.060	13.8	0.00097	300
Turnover (MEUR)	519	404.2	27.5	2,671.4	0.080	46,800
Number of employees	519	1,238.3	165	8,614.6	1	122,608
Number of Kununu ratings (Training)	519	22.9	9	74.3	4	1,174

Notes: These data contain only firms with at least four detailed ratings on Kununu between January 2017 and August 2018. The *Number of Kununu ratings (Training)* is not part of the regressions. Firms with less than four ratings between January 2017 and August 2018 are not part of the empirical analyses.

Table 4.2 displays the descriptive statistics of our estimation sample for the measure of *brand equity* at Kununu, which is the rating of the current and past employees for the *firm image*. Compared to the rating for on-the-job training, the average assessment of the employees for the image of their firm is noticeably larger (3.62 vs. 3.31). Figure C.2 and Figure C.5 provide more details about the distribution of the ratings. The average number of ratings for *firm image* is slightly lower than for *on-the-job-training* (21.7 vs. 22.9).

Table 4.2: **Summary statistics - Image: Kununu rating - Estimation sample.**

	N	Mean	Median	SD	Min	Max
Image: Kununu rating	492	3.62	3.75	0.80	1	5
Marketing expenditures (MEUR)	492	6.98	0.11	81.4	0.0010	1,480
Turnover (MEUR)	492	312.7	27.4	1,765.2	0.22	25,763
Number of employees	492	1,014.4	158	7,189.5	3	122,608
Number of Kununu ratings (Image)	492	21.7	8.50	72.7	4	1,171

Notes: These data contain firms with at least four detailed ratings on Kununu between January 2017 and August 2018. The *Number of Kununu ratings (Image)* is not part of the regressions. Firms with less than four ratings between January 2017 and August 2018 are not part of the empirical analyses.

Facebook Data

Table 4.3 displays the summary statistics of our estimation sample for the analysis of the relationship between the number of Facebook *likes* and the survey-based *marketing* expenditures (MIP 2017). In total, we have 944 firms with non-zero and non-missing data in our estimation sample. The average firm in our sample has 9,684 Facebook *likes* and 2.4 million euros in marketing expenditures.[75] Overall, the firms in our estimation sample are, on average, larger than in the entire MIP 2017 sample. Especially the size class of firms with 0 to 9 employees is underrepresented in our sample, as these firms are less likely to have a Facebook page (see Table C.7). Therefore, the average turnover in our estimation sample is 58.2 million euros, and each firm has, on average, close to 220 employees. However, these average numbers are generally driven by very large firms. All median values are smaller (see Table 4.3). For more details on the full sample, see Table C.3 and Table C.5.

Table 4.3: **Summary statistics - Image: Facebook likes - Estimation sample.**

	N	Mean	Median	SD	Min	Max
Image: Facebook likes	944	9,683.8	224	98,292.4	1	1,702,502
Marketing expenditures (MEUR)	944	2.40	0.032	48.8	0.00048	1,480
Turnover (MEUR)	944	58.2	5	434.9	0.027	11,630
Number of employees	944	218.4	40.5	1,090.8	1	25,247

Notes: These data contain only firms with at least one *like* on Facebook as of December 2017/January 2018 and non-zero marketing expenditures in 2016.

[75] Figure C.3 and Figure C.6 provide histograms for the number of *likes* and scatterplots for the relationship between the number of *likes* and the marketing expenditures.

4.4 Empirical Approach

We analyse the relationship between our newly developed platform-based indicators for *brand equity* and *firm-specific human capital* and the survey-based measures on marketing expenditures and on training expenditures (MIP 2017). These knowledge-based assets belong, within the framework of Corrado et al. (2005, 2009), to the group of *economic competencies* that are usually not measured within official statistics. For our cross-sectional data[76], we analyse this relationship with standard OLS regressions containing a set of firm-level control variables:

$$ln(Y_{image/likes,i}) = \beta_{exp} ln(expenditures\_marketing)_{2016,i} + X_i\gamma + e_i \tag{4.1}$$

$Y_{likes,i}$ denotes the number of *likes* on Facebook in December 2017/January 2018, and $Y_{image,i}$ is the average rating per firm i for the period from January 2017 to August 2018 for the item *image* on Kununu. The vector of control variables X_i includes turnover, the number of employees, and a set of industry dummies. The number of Facebook *likes* and the average score for the item *firm image* on Kununu are our indicators of *brand equity.*

Furthermore, the average score for the item *on-the-job training* on Kununu is used as an indicator for the intangible asset of *firm-specific human capital*. We study the explanatory power of our indicator via the following OLS regression:

$$ln(Y_{trainig,i}) = \beta_{exp} ln(expenditures\_training)_{2016,i} + X_i\gamma + e_i \tag{4.2}$$

$Y_{trainig,i}$ is the average rating for *on-the-job training* for the period from January 2017 to August 2018 on the employer *branding* and review platform Kununu. The vector of control variables X_i includes once more turnover, the number of employees, and a set of 23 industry dummies.

As an extension to OLS regressions, which are evaluating the relationship between the platform- and survey-based data, we employ a machine learning (ML) approach for predicting firm-level expenditures for marketing and on-the-job training based on our platform indicators. Details can be found in Section 4.6.

4.5 Estimation Results

4.5.1 Kununu

Table 4.4 shows OLS estimates for the relationship between the MIP 2017 survey-based measures for knowledge-based capital and our platform-based indicators. We

[76] We gathered historical Kununu data for the years prior to 2017. Thus, in principle, it would be possible to do fixed effects (FE) panel regressions with the full MIP panel and the historical Kununu platform data. However, due to the limited number of observations (ratings) in the early years of Kununu, this was not feasible.

estimate the models described in Equation 4.1 and Equation 4.2. In Column (1), we regress the Kununu rating for training on the *ln* of the survey-based expenditures for training and a set of control variables.[77] Column (3) displays the analogous results for the image of a firm. In Columns (2) and (4), we use *ln* transformations of the dependent variables, as the Kununu ratings are slightly skewed (see Figure C.1 and Figure C.2). As mentioned before, we restrict our Kununu sample to firm profiles with at least four ratings between January 2017 and August 2018. More ratings per firm on Kununu imply better data quality. But, on the other hand, the number of observations in our regressions is dramatically reduced. Given the data, the lower bound of at least four ratings per firm is sensible. We also provide robustness checks with varying thresholds.

Table 4.4: **OLS regressions: Kununu.**

Dependent variable	(1) Training: Kununu rating	(2) ln(Training: Kununu rating)	(3) Image: Kununu rating	(4) ln(Image: Kununu rating)
ln(Training expenditures)	0.0680*	0.0237*		
	(1.81)	(1.88)		
ln(Marketing expenditures)			0.0842***	0.0272***
			(3.13)	(3.19)
ln(Turnover)	0.0121	0.0103	-0.0407	-0.00946
	(0.28)	(0.68)	(-0.83)	(-0.61)
ln(Number of employees)	-0.0590	-0.0238	-0.0459	-0.0160
	(-1.06)	(-1.27)	(-0.78)	(-0.83)
Industry dummies	Yes	Yes	Yes	Yes
adj. R^2	0.139	0.139	0.128	0.132
Observations	519	519	492	492

Robust t statistics in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Robustness checks that use different minimum numbers of ratings (thresholds) can be found in Table C.8 and Table C.9. Requiring at least 3 or 5 ratings does not qualitatively change the results. With a minimum of 6 ratings, *on-the-job training* is insignificant due to the reduced number of observations. A minimum of 10 ratings leads to a further drop in the number of observations, resulting in insignificant results for both the image and training. For another robustness check, we removed the Kununu ratings of ex-employees. Especially if they were fired, the ratings could be biased. However, the removal of the ex-employees does not fundamentally change

[77] *ln* is the natural logarithm.

the results. Table 4.4 indicates a positive and significant relationship between the survey-based expenditures and our platform-based indicators. The main difference between the results for *training* (Columns 1 and 2) and *image* (Columns 3 and 4) is the higher significance level of the latter.

4.5.2 Facebook

Table 4.5 presents OLS estimates for the relationship between the MIP 2017 survey-based marketing expenditures and the number of Facebook *likes* in December 2017/January 2018. In Column (1), we regress the *ln* of the number of Facebook likes on the *ln* of the survey-based marketing expenditures of the firm. In Column (2), we add turnover and the number of employees as explanatory variables. In Column (3), we additionally include industry dummies. As before with the Kununu data, we observe a positive and highly significant relationship between the survey-based expenditures for marketing and our platform-based indicator (corresponding to the number of Facebook *likes*).

Table 4.5: **OLS regressions: Facebook.**

	(1)	(2)	(3)
	ln(Image:	ln(Image:	ln(Image:
Dependent variable	Facebook likes)	Facebook likes)	Facebook likes)
ln(Marketing expenditures)	0.522***	0.454***	0.455***
	(15.67)	(8.23)	(8.34)
ln(Turnover)		0.0589	0.106
		(0.72)	(1.11)
ln(Number of employees)		0.0550	0.0178
		(0.65)	(0.19)
Industry dummies	No	No	Yes
adj. R^2	0.261	0.265	0.364
Observations	944	944	944

Robust t statistics in parentheses
* p<0.10, ** p<0.05, *** p<0.01

4.6 Prediction of Expenditures on Knowledge-Based Capital based on Machine Learning

In this section, we discuss the use of machine learning methods to predict internal firm expenditures with the help of pre-existing firm and platform data. The underlying idea is that internal expenditures could be estimated using (semi-)public data.

The advantage is that the expenditures could be updated more regularly as the platform data are frequently updated. The firm data (number of employees, industry, turnover, and expenditures) are based on the Mannheim Innovation Panel (MIP).

Our analysis is based on the estimation samples described in Section 4.3.3. As feature variables of the machine learning models, we choose the turnover of the firm, the number of employees, the industry, and scraped data from Facebook and Kununu. The target variable is the amount of firm-specific expenditures on *employee training*, or *marketing* from the MIP survey. Based on each target variable, the observations above the 99th percentile (target variable) are removed, as they might all be randomly assigned to the test data. Additionally, the target variable is ln-transformed for machine learning, but the predictions are transformed back with the exponential function before the evaluation metrics are calculated. In doing so, we take into account that the target variable distributions are skewed. A standardisation of the feature variables *number of employees* and *turnover* is performed. As a result, the variables have a mean of zero and a standard deviation of one in the training data set. The transformation is carried out to improve the performance of neural networks. We use standardisation parameters of the training data to standardise the test data in order to avoid data leakage. As we use the standardisation parameters of the training data, the mean and standard deviation of the test data may not be zero or one. The categorical industry information is converted into dummy variables, as most machine learning methods only work with numerical data.

Several machine learning models are trained for our analysis: *neural networks (NN), random forests (RF), k-nearest-neighbour (KNN)*, and *support vector machines (SVM).* An overview of the methods is given in Hastie et al. (2001) and Kotsiantis (2007). For each model, we perform a 10-fold cross-validation to find suitable model parameters and robust models. The corresponding model parameters in the software packages are listed in italics and brackets. The neural network has three dense layers with sizes 16, 8, and 1, uses the *ReLu* and *linear* (last layer) activation functions, and the *mean absolute error* objective function. The optimisers *Adam*, *Adadelta* and *SGD (optimiser)* are taken into account. The random forest is trained with 1,000 and 5,000 trees (*n_estimators*), the *mean absolute error* (*criterion*), a maximum depth of 2, 5, 10, and 20 (*max_depth*) as well as *none* and *auto* as the maximum number of features (*max_features*). The k-nearest neighbours method considers the nearest 1, 2, 5, 10, 25, and 50 data points (*n_neighbours*) and the weighting schemes *uniform* and *distance* (*weights*). The support vector machine considers the kernel *linear*, *rbf*, *sigmoid*, and *poly* (*kernel*) to map the data into a high-dimensional vector space and the regularisation parameters 1, 2, and 5 (*C*). Non-specified parameters are set to the default options. The parameter spaces are expandable to retrieve improved results. For the training, 67% of the data was used, and 33% was reserved for testing. In addition,

we perform five random train-test splits to make a statement about the sensitivity of the splits.

For model training and evaluation, we use the Python packages *Keras* (Chollet et al. 2015) and *Scikit-learn* (Pedregosa et al. 2011). *Keras* is used as the neural network implementation, and the other machine learning models are based on the *Scikit-learn* software package. We use the Mean Absolute Error (MAE) metric as an evaluation measure because it is commonly used and can be interpreted easily (Hyndman & Koehler 2006). For example, a MAE of 0.1 corresponds to an average absolute difference of 100,000 euros between the actual expenditures and our prediction.

Table 4.6 shows the predictive performance for training expenditures based on Kununu and MIP data. Using Kununu data alone is as good as a baseline model with a MAE of 0.22, but the MIP data explain more of the data with a MAE of 0.14. Combining both feature sets does not improve the performance. In this case, the web-scraped data even worsen our predictions slightly.

Table 4.6: **Predictive power for training expenditures based on Kununu data.**

Features				Result		
Number of employees	Turnover	Industry	Training Kununu	N	Best model	MAE
			x	513	RF	0.22 (0.04)
x	x			513	RF	0.15 (0.03)
x	x	x		513	RF	0.14 (0.03)
x	x	x	x	513	RF	0.15 (0.03)

Notes: The results are based on the mean values of five random train-test splits. The standard errors for the MAE are mentioned in the brackets. A baseline model taking the non-transformed mean or median of the target variable in the train data set as prediction has a MAE of 0.31 and 0.22. Target variable: Training expenditures. Firms with zero marketing expenditures, turnover, or employees are dropped. Six outliers are dropped based on the 99th percentile of the target variable. The MAE reports the best *Mean Absolute Error* value for the specified set of models.

Table 4.7 shows the predictive performance for internal marketing expenditures based on Kununu and MIP data. The model based on web-scraped data is as good as the baseline model. The MIP data, on the other hand, can again explain more of the data with a MAE of 0.87. As expected, combining both feature sets (survey and platform data) only slightly affects the performance.

Table 4.7: **Predictive power for marketing expenditures based on Kununu data.**

Features					Result	
Number of employees	Turnover	Industry	Image Kununu	N	Best model	MAE
			x	487	KNN	1.04 (0.09)
x	x			487	SVM	0.88 (0.09)
x	x	x		487	RF	0.87 (0.08)
x	x	x	x	487	RF	0.86 (0.10)

Notes: The results are based on the mean values of five random train-test splits. The standard errors for the MAE are mentioned in the brackets. A baseline model taking the non-transformed mean or median of the target variable in the train data set as prediction has a MAE of 1.45 and 1.05. Target variable: Marketing expenditures. Firms with zero marketing expenditures, turnover, or employees are dropped. Five outliers are dropped based on the 99th percentile of the target variable. The MAE reports the best *Mean Absolute Error* value for the specified set of models.

Lastly, Table 4.8 shows the results for the prediction of the internal marketing expenditures based on the Facebook and MIP data. A model based on the platform data has a MAE of 0.26 but is outperformed by a model based on the MIP data. Combining both feature sets results in a model with a MAE of 0.21. Thereby, the result suggests that the web-scraped data have a positive effect on our predictions.

Table 4.8: **Predictive power for marketing expenditures based on Facebook data.**

Features					Result	
Number of employees	Turnover	Industry	Image Facebook	N	Best model	MAE
			x	934	KNN	0.26 (0.02)
x	x			934	NN	0.23 (0.02)
x	x	x		934	NN	0.22 (0.02)
x	x	x	x	934	RF	0.21 (0.01)

Notes: The results are based on the mean values of five random train-test splits. The standard errors for the MAE are mentioned in the brackets. A baseline model taking the non-transformed mean or median of the target variable in the train data set as prediction has a MAE of 0.42 and 0.27. Target variable: Marketing expenditures. Firms with zero marketing expenditures, turnover, or employees are dropped. Ten outliers are dropped based on the 99th percentile of the target variable. The MAE reports the best *Mean Absolute Error* value for the specified set of models.

The results are robust, in the sense that we can see the same pattern across different predictions. Platform data alone have relatively little or no predictive power, as the baseline models yield similar MAE values. MIP data explain a higher amount of the data and outperform the platform data. Combining platform data (Facebook or Kununu) with MIP data has at most a slight effect or no effect on the results.

The main problem in our analysis is the low number of observations, as we are limited to firms with data on firm-level training or marketing expenditures. Unfortunately, the coverage of MIP firms on Facebook and Kununu is relatively low, as illustrated in Figure 4.2. However, we expect a better performance with an increasing number of observations. Machine learning models based on small data sets are, to some extent, sensitive to sample splitting. For example, all large firms could fall into the test data set, leading to non-robust results. Therefore, our results are based on the mean values of five random train-test splits. The reported standard errors are, in some cases, relatively high (up to 0.10). Additionally, the data are not representative. For example, small firms are underrepresented, which might lead to problems with the generalisability of the machine learning models. The information about the *best models* should not be interpreted too extensively, as multiple models often perform only slightly differently. Random forests are known to have relatively good performance on tabular data. Therefore, it is not surprising that they are, in most cases, the best models. The performance could be further improved with modified random forest models, e.g. gradient-boosted trees (Hastie et al. 2001, Friedman 2002). Random forests, on the other hand, have the major disadvantage that the methodology is based on weighted averages. A random forest can never predict firm-level expenditures that lie outside the training set. In summary, the MIP and platform data can be used to a limited extent to estimate the internal expenditures of firms.

4.7 Conclusions and Future Research

This paper developed new indicators for the intangible capital of firms based on publicly available data from online platforms. These basic indicators for *brand equity* and *firm-specific human capital,* which are part of the intangibles framework developed by Corrado et al. (2005, 2009), were taken from the social media platform Facebook and the employer branding and review platform Kununu. We compare these indicators with firm-level survey data on marketing and training expenditures taken from the German part of the Community Innovation Survey. All OLS regressions show a positive and significant relationship between the firm-level expenditures for marketing and on-the-job training and the respective information stemming from the online platforms. Various robustness checks confirm the validity of the results.

However, there are also caveats with our current approach. Due to the limited presence of smaller firms on online platforms, we are currently predominantly capturing medium-sized and larger firms. Furthermore, although we do find a positive and significant relationship between our platform-based indicators and the survey-based expenditures in our OLS regressions, predicting expenditures based on an explorative machine learning approach shows that the platform data alone have little

or no predictive power.

Using data from online platforms can nevertheless provide a useful source for establishing firm-level indicators on intangible assets in the field of economic competencies, which are difficult to measure through surveys or from balance sheet data. But to better utilise this data source, more research is required. First, we need a better understanding of the dynamic relationship between activities on online platforms that are related to a firm's knowledge-based capital and the actual firm activities to build up and maintain such capital. Second, a comparative analysis of different platform data is needed to better assess the value of the information that can be derived from various platforms. Finally, analyses on the relationship between the newly derived indicators on firms' economic competencies, on the one hand, and firm performance (e.g. through productivity analysis), on the other, would provide additional insights into the validity of these indicators. For this purpose, time-series data on both platform-based indicators and firm performance measures are required.

Chapter 5

Mapping Employee Mobility and Employer Networks using Professional Network Data

Joint work with Hanna Hottenrott, Christian Rammer, and Konstantin Römer.

5.1 Introduction

The growing use of social media provides new sources of data for research purposes and the development of new economic indicators. Prominent examples of such sources are career-oriented social media platforms like LinkedIn or XING.[78] Career-oriented platforms are a natural candidate for the generation of large-scale employment indicators as these platforms, through network effects, attract many actors. There are incentives on both the users' and providers' sides to grow the platform through collecting and assembling data. Successful platforms are therefore a rich source of data on actors active on them. In the case of the career-oriented platform XING, they are, on the one hand, employees or job seekers and, on the other hand, employers such as firms, research institutes, or public administration.

The data available on such platforms allow for identifying employer-employee relationships over time and space, hence tracking individuals' mobility from one employment to another. Importantly, the employment types listed on such platforms are not limited to those with social security contributions but also capture unpaid, freelance, and entrepreneurial activities, typically unobserved in administrative employer-employee data. Further, the employment and employee mobility data are available immediately after adding them to the user profile.[79] Thus, there is no time lag in data provision, as is typically the case with administrative data.

[78] LinkedIn and XING are both employment-focused social media platforms. The former was launched in 2002 and is owned by Microsoft. XING is operated by New Work SE and was founded in August 2003. These platforms entail user and employer profile data managed by the users or representatives of the employer. XING is particularly popular in German-speaking countries.

[79] User profiles represent the online profiles of employees on the platform XING.

However, outdated information from the platform may also be part of the data. Furthermore, administrative data are often subject to stringent legal requirements and may, for example, not be linked to all types of third-party data. So far, it remains unclear whether the data extracted from career-oriented social networks are sufficiently representative for research purposes.

This study aims to assess the usefulness of career-oriented social media data for mapping and tracking employee mobility (here, also including non-conventional types of employment). Moreover, we explore the usefulness and plausibility of the employee flow data between employers by analysing the resulting networks. Measuring networks between employers through labour mobility is vital in innovation research (Görg & Strobl 2005, Balsvik 2011, Hottenrott & Lopes-Bento 2016). However, such networks are typically constructed from linked administrative employer-employee data (Maliranta et al. 2009, Collet & Hedström 2013, Kaiser et al. 2015), patent data, i.e. measuring inventor mobility (Somaya et al. 2008, Rahko 2017, van der Wouden & Rigby 2021), or data on scientific publications, i.e. capturing author mobility (Edler et al. 2011, Franzoni et al. 2014). In this study, we show that social network data are a valuable data source for exploring networks between employers. Exploring network data is valuable for many research applications, particularly for research on the performance of employers or regions (Schilling & Phelps 2007, Ozman 2009, Giuliani 2011). The approach has the potential to augment or replace data collected via surveys. While survey data are generally not well suited for mapping networks due to incomplete coverage and non-response, combining networks generated from a big-data source with survey data enriches the data portfolio and hence the scope of addressable research questions.

The data preparation consists of multiple consecutive steps: First, we disambiguate employers listed in publicly accessible employment data.[80] We link and classify employers using the names and addresses (including private employers, research institutes, public administration, (non-)governmental institutions, etc.) to identify these employers in the Mannheim Enterprise Panel (MUP) and other data sources such as the Mannheim Innovation Panel (MIP). The total number of available employments is about 46M.[81] We create a *Linked Employer-Employee* (LEE) data set by matching around 1.5M employers to 14M publicly accessible employments.[82] Second, we calculate employee flows based on the matched employments. We create an employee flow for each employee moving from one employer to another. As a result, we extract 9M employee flows between employers or into/out of employment, e.g. students entering the labour market or retirements out of the labour

[80] In simplified terms, employment is a tuple that consists of one user/employee and one employer.

[81] We use the abbreviations *K* for thousand and *M* for million, e.g. 2M instead of 2 million.

[82] About two-thirds of the employments are not considered because of a restriction on high-quality matches with the MUP and a limitation on data from publicly accessible user profiles.

market. Third, the flow data are used to create annual *employee flow networks* that cover the period from 2010 to 2020. For the series of networks, we calculate network measures, i.e. cliques, transitivity, reciprocity, and density. The resulting database contains high-quality employment, employer, employee flow, and annual network data.[83]

Next, we check the plausibility and representativeness of the data. For this purpose, we use MUP data, which covers almost all businesses in Germany[84] and a share of other organisations such as universities, research institutes, hospitals, and non-profit organisations. We compare the MUP employer data with the XING data. The results suggest that the employers on XING are sufficiently representative of all employers in Germany, regarding age, size, industry, legal form, and region. Moreover, the data capture a significant share of employers in Germany. However, some industries are underrepresented, e.g. utilities and trade.

We analyse the employments with regard to their experience, discipline, career stage, and type. The matched employers are investigated regarding their size, age, industry, and region. For both employments and employers, the distributions of these characteristics are plausible. Furthermore, we test the validity of the flow data by analysing the career level, employment discipline, employment type, employer size, employer industry, and employer region before and after the employment change. E.g. most employees switch employments within a discipline and typically move upwards on the career ladder. Lastly, we analyse the network data using different graph metrics and, as an additional check, visualise selected local networks.

Finally, since research shows a positive link between knowledge exchange through employee mobility and employer performance (Almeida & Kogut 1999, Godart et al. 2014, Wu et al. 2017, Abbasiharofteh et al. 2021), we also test the plausibility of selected network measures using employer data. The analyses yield promising results, e.g. changes in the degree centrality of employers are positively linked with changes in employment counts. In summary, with some limitations, the data represent a valuable novel research data source for studying the role of employee mobility and employer networks. The data go beyond paid employments, including internships, freelance work, and entrepreneurial activities. For network analyses, the coverage is sufficiently high, and network measures can be derived for employers that are neither active in patenting nor engaged in larger, visible alliances.

The remainder of this paper is structured as follows: In Section 5.2, we relate our

[83] For legal reasons, we decided against creating and analysing an in-detail employee data set.

[84] However, we mark some industries as missing/others according to their NACE classification. For example, agriculture (NACE code A), private households (NACE code T), and offshore organisations and bodies (NACE code U).

work to the literature, and Section 5.3 presents the data processing. Section 5.4 describes the resulting data sets, and Section 5.5 discusses the findings and concludes.

5.2 Literature Review

With the broader adoption of the internet, the conceptualisation and development of real-time economic indicators became increasingly popular. Choi & Varian (2012), for instance, use Google Trends to forecast unemployment claims as an indicator of economic activity. They demonstrate the viability of online search queries as indicators, leading the way to novel economic indicators. Allcott et al. (2019) analysed the magnitude of the fake news problem on Facebook. In contrast to Twitter, they find that the magnitude of the problem on Facebook had declined over time, but the spread of fake news was still predictable. Real-time indicators also provide new opportunities for innovation research, augmenting the portfolio of traditional innovation indicators. In traditional innovation studies, survey data or statistical data provided by the government are often used (e.g. Rammer et al. 2021). For example, innovation studies are often based on accounting data, firm survey data on expenditures for innovation, or patent (application) counts collected from patent office databases. While innovation surveys like the Community Innovation Survey by the European Union have substantially deepened and improved our understanding of innovation activities (Hong et al. 2012), key disadvantages of surveys are the cost of data collection, the time lag, and the problem of data availability for individuals or firms. The first two factors often limit reaching a sufficiently large sample, as asking thousands of firms directly costs time and money (Rammer & Es-Sadki 2023). In addition, the national statistical offices often limit access to raw data and only publish summary statistics or reports.

One approach to addressing these issues is to use available web data provided by employers or employees. Web data have been shown to offer value: Although scraping and processing web data takes time, and the quality of the data differs widely from employer to employer. For instance, Gök et al. (2015) construct a relatively accurate web-based R&D indicator. Kinne & Lenz (2021) extend the approach and illustrate that firm websites can be used to predict a firm-level innovation probability. In particular, the study uses survey data as a training sample to predict innovation activities for all German employers with a website. The authors illustrate the value of web mining and extend the previous approaches through deep learning, resulting in reliable innovation indicators available for employers that do not participate in innovation surveys or do not patent. In line with these ideas, Axenbeck & Breithaupt (2021) use the web-mining approach to identify website characteristics predicting firm-level product and process innovation activity, and Schwierzy et al. (2022) show that website data can also be used for the mapping of specific technologies such as

additive manufacturing. Axenbeck & Breithaupt (2022) demonstrate that employers' website data can be used to measure other corporate activities, such as those related to digitalisation. Other publications use, for example, employer-related data from the social media platforms Twitter, Kununu[85], and Facebook (e.g. Veltri 2013, Breithaupt et al. 2020).

First, this paper contributes to recent developments in web-based indicators by analysing a little-explored source of large-scale data on employee and employer activities. In particular, we test the use of data from a career-oriented social media platform for mapping the flows of individuals between employers. To do so, we build on graph theory. Graph theory is a branch of discrete mathematics and theoretical computer science (Bondy & Murty 1976, Diestel 2017). The underlying principles are linked to Leonhard Euler's work on the famous *Seven Bridges of Königsberg* problem. Graph theory is also related to *social network analysis* (SNA) and is frequently used in the innovation literature (e.g. Abbasiharofteh et al. 2021, Axenbeck & Breithaupt 2021). The theoretical foundations are described in Wasserman & Faust (1994) and Scott (2017). Our paper applies SNA methods in the context of employee flows and employer networks. Furthermore, we aim to account for characteristics of employee flows, such as the duration of employments and the career level of the employees.

Second, we contribute to the *Linked Employer-Employee* (LEE) literature, in which employees' individual data are linked with data on their employers. Our contribution is the creation of a LEE database from non-official data, i.e. the combination of web and proprietary data. Our LEE data have the advantage of using up-to-date and public data. Furthermore, it includes employees who immigrate from abroad, migrate abroad, or enter the labour market from education. However, in contrast to the official LEE data, we have only limited data on employees, such as wages, and on employers, such as financial indicators. Moreover, we do not have complete coverage of all employers and employees in Germany, and historical data might be sparse. Multiple LEE data sets have already been created from official data for Germany. For example, the SOEP-LEE extends the Socio-Economic Panel (SOEP) data by linking the employees' individual data with data on their employers (Weinhardt 2016, Weinhardt et al. 2016, 2017). The SOEP is a large, long-running multidisciplinary longitudinal study in Germany. Other German IAB data, such as WeLL-ADIAB and LIAB, provide, among others, historical *Linked Employer-Employee data* (Schmucker et al. 2014, Heining et al. 2016). Furthermore, the LEEP-B3 and linked ALLBUS data sets are also available in Germany (Gerhards et al. 2010, Abendroth et al. 2014). Lastly, there are LEE data sets for many more countries and regions, e.g. the *'US Worker Establishment Characteristics Database'*, the *'New Zealand's Linked Employer-Employee Database'*, the *'Norwegian Linked Employer-Employee Database'*, and

[85]Kununu: `https://www.kununu.com/de` [Last accessed: 24.02.2024].

a LEE data set for the European region (Jensen 2010).[86]

5.3 Data Processing

For our data analysis, we connect two data sources: First, we use data from the social and professional network XING. The data provide detailed information on users – mostly professionals – who create profiles on the platform primarily for professional networking. These profiles comprise personal, employment-related data, and data about the employer. Data access was granted in close cooperation with the platform provider New Work SE. Second, we use data from the Mannheim Enterprise Panel (MUP), which provides data about the population of registered businesses in Germany.[87] The data are maintained in collaboration with Creditreform, Germany's largest credit rating agency (Bersch et al. 2014). MUP data provide employer-level information that, besides others, contains addresses, employee counts, founding dates, and website URLs. We combine both data sources, i.e. we link XING data on employees and their mobility to the MUP.[88] The remainder of this section describes the data processing of employers and employments (Section 5.3.1), the derived employee flows (Section 5.3.2), and annual employer networks (Section 5.3.3).

5.3.1 Employers and Employments

Employer and employment data are exported from the data warehouse of XING. They contain about 1.9M employers, although not every one of them has to be a valid or active employer. These observations are not directly excluded, as we are also interested in historical data. XING users have deposited about 46M employment data points. About one-third of the employments are linked to the XING employer database.[89] The employment data are partly maintained by the users and include the employer name, employer URL, industry, employment type, career level, and field of activity (discipline). Some of the fields are optional, e.g. the employer URL. Some employment data points are of unclear quality. XING's employer data show specific features: First, the employer database is not standardised or linked to a uniform database like the MUP. Second, outdated, duplicate, or invalid/fake employers are contained. For example, insolvent employers whose XING profiles are

[86] *'European Structure of Earnings Survey'*: `https://ec.europa.eu/eurostat/web/microdata/structure-of-earnings-survey` [Last accessed: 24.02.2024].

[87] We use the term *employers* instead of *firms* because XING includes not only firms but also public institutions, universities, research institutes, and nonprofit organisations.

[88] In this paper, we are interested in the aggregated flows of employees between employers and not in the individual users. Fortunately, we have the legal permission to use the (aggregated) data.

[89] This is an internal database by XING that lists employers. Not all employers are linked to the database. These employers are only mentioned in the employment data.

later deleted.[90] Third, operating sites, subsidiaries, and employer groups can have their own XING profiles. Fourth, some employers are not located in Germany or the DACH region.[91]

Figure 5.1 (left) shows the XING profile of an employer. The employer size indicates the number of employees within the respective group (here: *51-200 employees*). Figure 5.1 (right) shows the work experience of a XING user. The user has had two employments. He has worked first as a Data Scientist and then as a researcher. His change of employment took place seamlessly in 2019 and is illustrated by the end date of the first employment and the start date of the second. The employer *'HMS Analytical Software GmbH'* maintains a XING profile. However, the employer *'ZEW – Leibniz Centre for European Economic Research'* does not. The missing logo is an indication, albeit not a sufficient one.

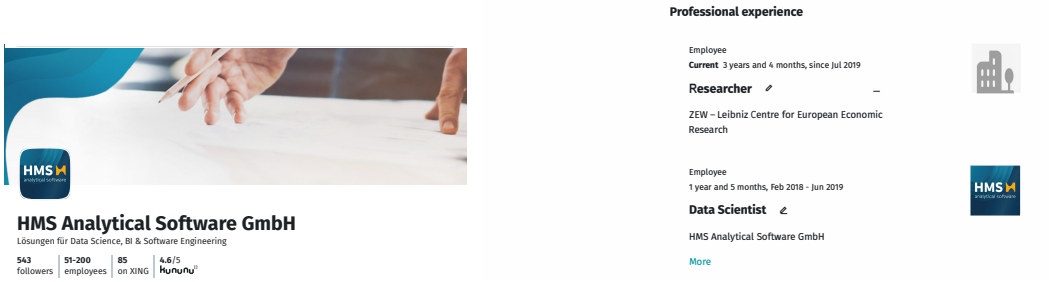


Figure 5.1: **XING employer profile (left) and a *professional experience* timeline for a platform user (right).** Source: The images were taken from the platform XING (www.xing.com).

The data preparation process consists of two consecutive steps. First, employers (1.9M) and employments (46M) are linked to the Mannheim Enterprise Panel (MUP) using the *SearchEngine tool* that implements a *string-matching algorithm*.[92] The linking relies on the text fields *employer name* or *employer URL*, if available. Employer profiles and employment data are separately linked with the tool, as their data quality is different. We assume that the data from the employer profiles have a higher quality than the employment data. As a result, we receive candidates from the Mannheim Enterprise Panel for the XING employer profiles and employers mentioned in employments on XING. Each candidate has a unique identifier called *crefo*, which is the identifier of employers in the MUP. There are 86M candidates for the employment data points based on employer name and 12M based on the URL.[93] Furthermore, about 187K employer profiles (XING) are linked to MUP employers. Second, the

[90] Here are some examples of the difference between relevant data and employers available on XING: Country data at the employer level are available for around 1.1M employers, of which 450K have indicated a country other than Germany. Furthermore, about 650K of the 1.9M employers have an invalid employer name, e.g. the names consist exclusively of dots and hyphens.

[91] DACH region: Germany, Austria, and Switzerland.

[92] GitHub project: `https://github.com/ThorstenDoherr/searchengine` [Last accessed: 24.02.2024] (Doherr 2023).

[93] A XING firm may have multiple candidates (potential matches) in the MUP database.

MUP candidates for the employments are enriched with additional employer-level data, like exit dates, if available. Some outliers are removed, e.g. implausible data such as employments before 1900 and after 2020 (after the data export), leading to a subset of 44M employments. To improve the data quality, we apply the *group-crefo* mapping, which combines affiliated employers within the MUP, e.g. subsidiaries. Lastly, binary indicators are created, which, for example, indicate if a *crefo* exists in the MARKUS database.[94]

We use a five-step heuristic to select the best candidate from the MUP for each employment data point (see Table 5.1). As a result, about 21M employments are matched to the MUP. Thus, the matching rate is about 47%. However, we deliberately extracted fewer matches than possible to ensure a higher quality of data, as there is a trade-off between the observation count and the data quality of the matches. For legal reasons, all employments of users without a public user profile must be removed, which reduces the number of matched employments to 14M.[95] We do not delete the 10M unmatched employments of users with public profiles but assign an artificial identifier for employers using the employer name. If users list the same employer name in their employment history and this employer has not been linked to the MUP, then we assign the same artificial identifier. For employers successfully linked to the MUP, we have additional data like the employee count, the location of the employer, and the year of foundation from this data preparation step on. These data are usually available as panels. For employers with an artificial identifier, the MUP characteristics are not available.[96]

The employer data consist of three types: employers only identified on XING, employers only identified in the MUP, and employers identified in both databases. In this paper, we are interested in employers listed in public employments linked to the MUP or non-matched employers (XING). For the following statistical analyses, we only use the matched employers. About 1.5 million unique employers have been successfully linked to the MUP. The number refers to the matched employers listed in XING employments because we do not count employers that are not mentioned in at least one employment. Furthermore, employers with an artificial identifier are not counted as well. In the MUP, some variables have missing values. The missing rates for the variables of the matched employers are: founding date (14%), district id

[94]The MARKUS database contains detailed and reliable data about employers: `https://www.creditreform.de/loesungen/marktanalyse-kundendaten/kundenbindung-akquise/markus` [Last accessed: 28.02.2024].

[95]We export a list of all publicly available user profiles (privacy settings) to ensure that we use only publicly accessible employments, as this is a legal requirement by New Work SE. Public employments can be viewed and saved by every visitor to XING. Each employment data point has a user reference, which is used to remove non-public data. Around 31M of 46M employment data points are public.

[96]The matched employment data have the following characteristics (missing rate in parentheses): employment type (0%), employment title (<1%), career level (47%), discipline (60%), start year (14%), and end year (35%). Some employment characteristics are re-coded for this project (see Table D.1).

(4%), legal form (<1%), two-digit NACE code (12%), and exit data (<1%). Employment counts are rather sparse. For example, 23% of matched employers have no employment data (MUP) between 2000 and 2021. However, many of these employers were not yet active or are no longer active. The imputation of missing numbers or carrying forward the last existing employment count are possible solutions to this problem. About 9 million employers are identified in the XING database without finding a link to the MUP, e.g. organisations outside the business enterprise sector or foreign employers. These employers have been issued an artificial identifier. The employer count is overestimated because unmatched employers like *'ZEW'* and *'ZEW Leibniz Centre'* do not receive the same identifier (hypothetical example).[97]

Table 5.1: **Data processing of employments.**

Step	Description	Data
1	Concatenate candidates (MUP) for employments; based on matching URLs or employer names.	Input: 86M + 12M = 98M candidates.
2	Drop duplicates based on a unique XING *employment identifier* and *crefo*; select unambiguous matches.	11M matches are selected and 83M candidates are left.
3	Use candidates with *'exist'* or *'missing'* exit status and select unambiguous matches.	6M matches are selected and 58M candidates are left.
4	Use candidates available in the *MARKUS* database and select unambiguous matches.	2M matches are selected and 38M candidates are left.
5	Use candidates with the highest fuzzy-matching score (employer name) and select unambiguous matches.	1M matches are selected and 28M candidates are left.
6	Concatenate results with disambiguated employer profile matches. Prefer employer profile matches.	21M employment matches.
7	Select matches for the subset of public employments.	Result: 14M employment matches.

Notes: Multi-stage selection of the best employment matches to the MUP. Includes the observation counts of the input and after processing in million employments (M).

5.3.2 Employee flows

Employee flows between employers are extracted from the employment data. Figure 5.2 shows a schematic representation of the data preparation steps. The steps are: data export, computation of match candidates, enrichment of candidates

[97] Lastly, we re-code employer characteristics of the MUP database (see Table D.2 and Table D.3).

with external data, selection of the best-matching candidate, deletion of non-public data, and computation of flows. For this, we select the employment data of users with a public profile and define a flow as the switch between two successive employments. Temporal breaks between employments, such as unemployment,

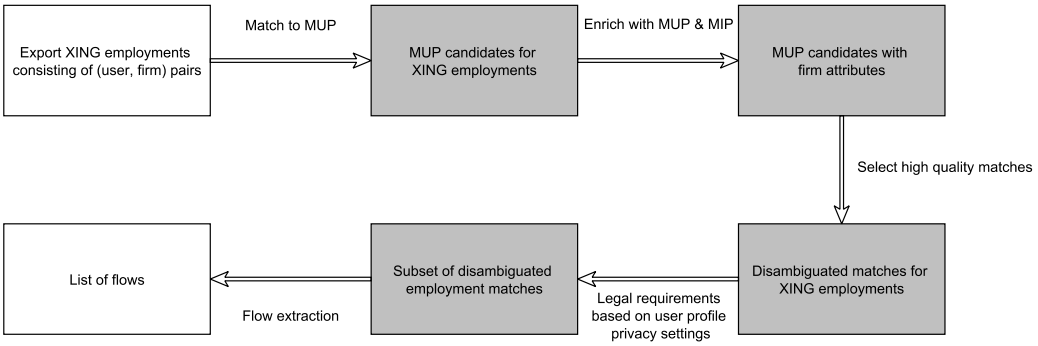


Figure 5.2: **Flow extraction procedure.** Steps: (1) Export data; (2) Get match candidates; (3) Enrich data; (4) Disambiguate match candidates; (5) Select subset following legal requirements; (6) Extract flow data. Own illustration (created with yEd - graph editor: `https://www.yworks.com` [Last accessed: 24.02.2024]).

are ignored. For example, if an individual was unemployed for three years between two employments, then a flow between the employment before and after unemployment is created. We create an employer identifier (*'missing'*) for the initial entry and final exit, e.g. for employment starters or retired individuals.

Table 5.2: **Extraction of employee flows.**

Step	Description	Data
1.	Start with all employments.	46M raw employments are exported.
2.	Keep the data of users with at least two employments.	34M employments are selected.
3.	Keep the employments of users with public profiles.	23M employments are kept: 13M are matched and 10M employments receive artificial identifiers.
4.	Extract matched flows.	9M flows between matched employers (incl. sink nodes). 7M flows model employment changes between two matched employers (excl. sink nodes).

Notes: Includes observation counts for each step in million employments (M). Input data: Employment and employer data. The sink nodes model the state before the first or after the final employment.

Furthermore, only user profiles with at least two employments are used. The steps lead to about 23M flows for matched and unmatched employments. Flows extracted from matched employments comprise about 9M observations, where 7M observations are flows between employers (see Table 5.2). Each flow entails the year

of the employment change and the characteristics of the old and new employment.

5.3.3 Networks

We create a temporally ordered set of graphs using the employee flow data.[98] We do not include the unmatched flow data as employer-level characteristics are missing. In addition, a higher number of employers is lengthening the subsequent calculations, and the data quality is lower. We create a series of graphs on an annual basis. For this purpose, we include a flow into an annual graph if the employment change has occurred within the respective year. Internships and student employments are not considered. We model the data as a weighted (each edge has a weight), directed (edges have a direction), and simple (no loops; at most, one edge per node pair) graph; see Bondy & Murty (1976) and Diestel (2017). The nodes of the graph are the employers, and the edges denote the employee flows between employers. Loops are deleted, e.g. a switch of employments or promotion within an employer, and multi-edges are aggregated to weighted edges. Furthermore, we model the direction of the employee flow in the graph (directed edge). If multiple employees move between the same two employers, we model this as a characteristic of the edge (*'weight'*).[99]

The annual graphs do not contain employers with zero flows in the respective year.[100] Each node contains employer characteristics like the employee count and location. The edges are associated with user characteristics, e.g. the year of the employment change. Table 5.3 shows the number of nodes and edges in the annual XING-based networks from 2010 until 2020. Employer- and flow-level characteristics are not shown for reasons of simplicity. Personalised data are no longer needed as the employment data have been mapped to the nodes and edges. Data for 2020 were not fully available at the time of the data export.[101] The series of networks has between 75K and, at most, 164K nodes. The edge count is at least 100K and at most 289K (see $Edges_1$). We find a slight decline in the number of nodes and edges for the year 2019. The employer and edge counts may be lower in the most recent years, as there is a time lag until newly established employers are available in the MUP data. The number of edges (including duplicates) is similar to the number of unique edges (see $Edges_2$ vs. $Edges_1$), as only a few employees move between the same employers in one year. Over time, the ratio between nodes and edges changes, i.e. the edge count per node increases. The edge count does not add up to 7M, as we ignore, for example, unpaid workers and do not consider all available years.

[98] The terms *graph* and *network* are often used interchangeably. We use the term graph to refer to the mathematical model. For the analytical applications, we use the term network.

[99] So-called *multigraphs* are an alternative, where multiple edges are allowed between node pairs.

[100] The isolated nodes can be easily added. However, we decided against it because the runtime and required memory for some graph algorithms scale quadratically with the node count of the graph.

[101] Date of data export: 26.10.2020.

Table 5.3: **Annual employee flow networks: Number of nodes and edges.**

Year	$\#Nodes$	$\#Edges_1$	$\#Edges_2$
2010	127,988	162,264	176,324
2011	137,723	186,636	203,013
2012	142,236	197,865	215,600
2013	147,168	207,711	227,067
2014	154,134	229,854	252,824
2015	159,902	253,068	281,432
2016	163,706	277,176	311,493
2017	163,364	288,805	324,304
2018	157,730	287,170	322,554
2019	135,470	241,574	271,399
2020*	74,367	99,602	110,319

Notes: Number of nodes, unique edges ($Edges_1$) and edges counting duplicates ($Edges_2$) for the XING networks. Networks are retrieved from employee flows. Source: TUM and ZEW based on XING data.
* The data for this year is only partially covered.

Next, we introduce the degree centrality measure (Nieminen 1973, Freeman 1978). The measure is calculated at the node level, i.e. at the level of employers. The definition depends on the network type and indicates how strongly or often a node is connected to other nodes. For directed networks, there is an in-degree and an out-degree centrality. For each node, the weights of the adjacent edges are summed up, i.e. total, incoming, or outgoing edges.[102] If the edge weight is not modelled, it is assumed that each edge has a weight of one. The centrality scores represent the number of employees moving between two employers within a year.[103] The number of centrality scores corresponds to the node count of the network. Consequently, we create a panel using the centrality scores of employers in the MUP. However, not every XING employer exists in all annual networks, e.g. if no employee left the employer or the employer ceased its business activities within a year.

Appendix D.2 gives an example of the data processing. The matching of employments, extraction of flows, and creation of annual networks are exemplified.

5.4 Data Description

Next, we describe the following data sets: employees and employments (Section 5.4.1), employers (Section 5.4.2), flows (Section 5.4.3), and networks (Section 5.4.4).

[102] Some definitions standardise the centrality score by dividing it with the node count of the network.
[103] In addition to degree centrality, further measures determine the relevance of nodes in a network, e.g. eigenvector and PageRank centrality (Page et al. 1999). However, this work does not use these due to more complex definitions and interpretations.

5.4.1 Employees and their Employments

Figure 5.3 shows selected characteristics of the platform users. The analyses intend to improve our understanding of the users active on the platform XING.[104] The most recent employment per user is utilised for the discipline, career stage, and employment type analyses. For the analysis of the user experience, the oldest employment per user is considered. Our findings suggest: First, most employees have between zero and nineteen years of professional experience. Few users have more than 39

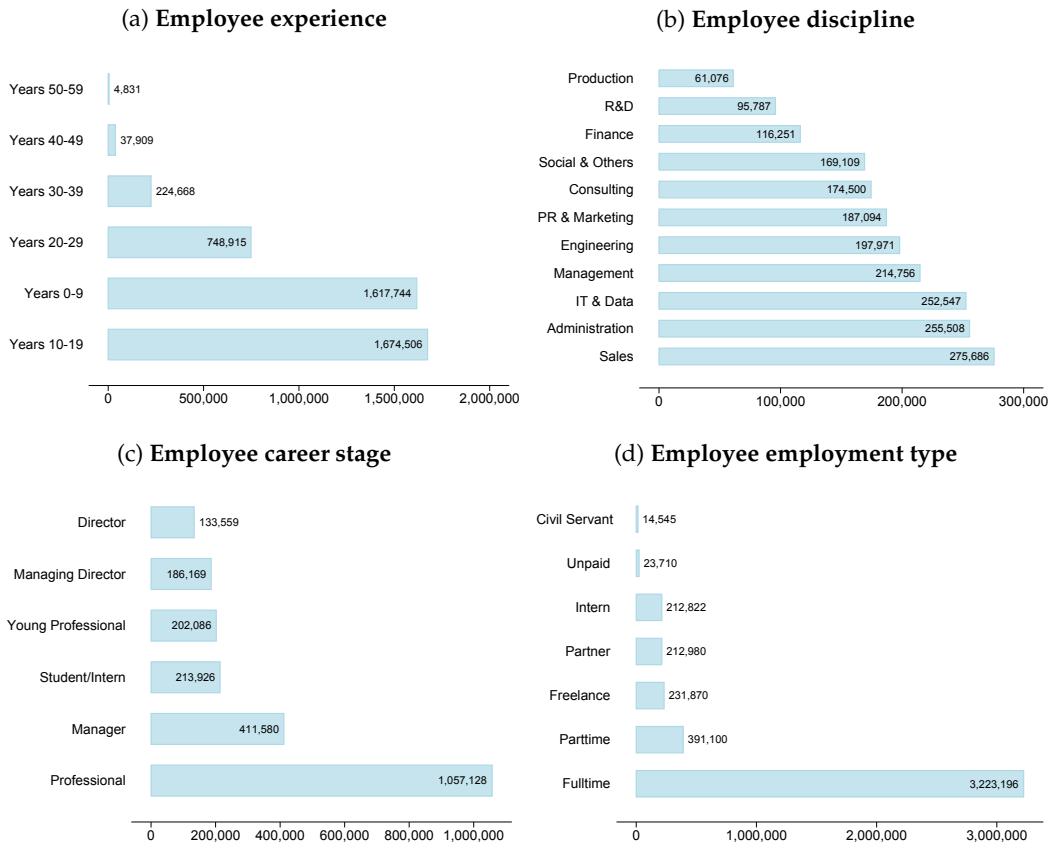


Figure 5.3: **XING users with respect to employment characteristics.** Note: (a): Per experience class; (b): Per discipline; (c): Per career stage; (d): Per employment type. For the analyses, only matched employments from publicly accessible user profiles are used. Source: TUM and ZEW based on XING data. Own illustration.

years of professional experience.[105] Second, the disciplines *'Sales'*, *'Administration'*, and *'IT & Data'* have the most employees. Surprisingly, few employees are working

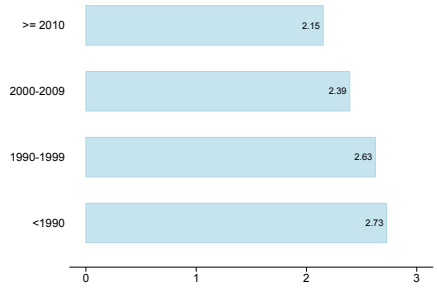
[104] We restrict the analysis to subsets of the data: employments have to be publicly available, matched to the MUP, and the respective characteristic, e.g. the career stage, needs to be available.

[105] German employees must work for 35 or 45 years to be considered *long-term* or *very long-term* insured to receive their full pensions. Primarily, highly educated employees are present on XING, who often start their careers later and, thus, work for fewer years overall.

in *'Production'*, *'R&D'*, and *'Finance'*. There may be a certain bias because the respective employees rarely use the platform (e.g. for *'Production'*). Third, most employees have professional experience but no managerial position. Managing directors and directors are found least often. Fourth, the majority of the employees work full-time.[106] Civil servants and unpaid workers are by far the least frequent. Overall, we can conclude that the XING data are not representative of German employees.

Figure 5.4 shows the employment length for employer and employment characteristics. First, employees stay longest at employers founded before 1990. For example, these employers are often larger and more established. However, we did not account for employments that were changed within an employer.[107] Second, the employment length is highest for partners and civil servants. By far, the shortest employment length is found for interns.[108] Figure 5.5 shows the number of employments by employer size, region, legal form, industry, and founding period.

(a) **Employment length by employer founding year class**



(b) **Employment length by employment type**

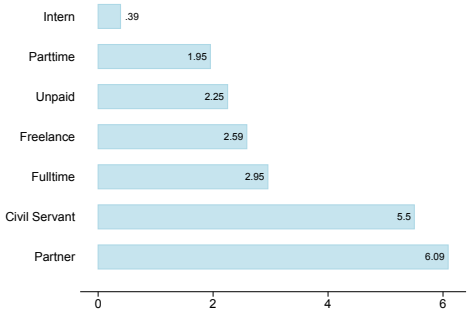


Figure 5.4: **Average length of employment (in years) for users by employer age (left) and employment type (right).** The start and end years of the employment data are used as a heuristic; based on public and matched employments. Source: TUM and ZEW based on data from XING and MUP. Own illustration.

[106] Students/Interns most often have the employment types *intern* and *part-time*.

[107] Larger employers often provide more opportunities to climb the career ladder in-house.

[108] The median length of employment relationships subject to social security contributions (in the data portfolio; excluding apprenticeships) is slightly longer than four years for 2017-2021. Source: `https://statistik.arbeitsagentur.de/Statistikdaten/Detail/202112/iiia6/beschaeftigung-sozbe-dauern/dauern-d-0-202112-xlsx.xlsx` [Last accessed: 24.02.2024].

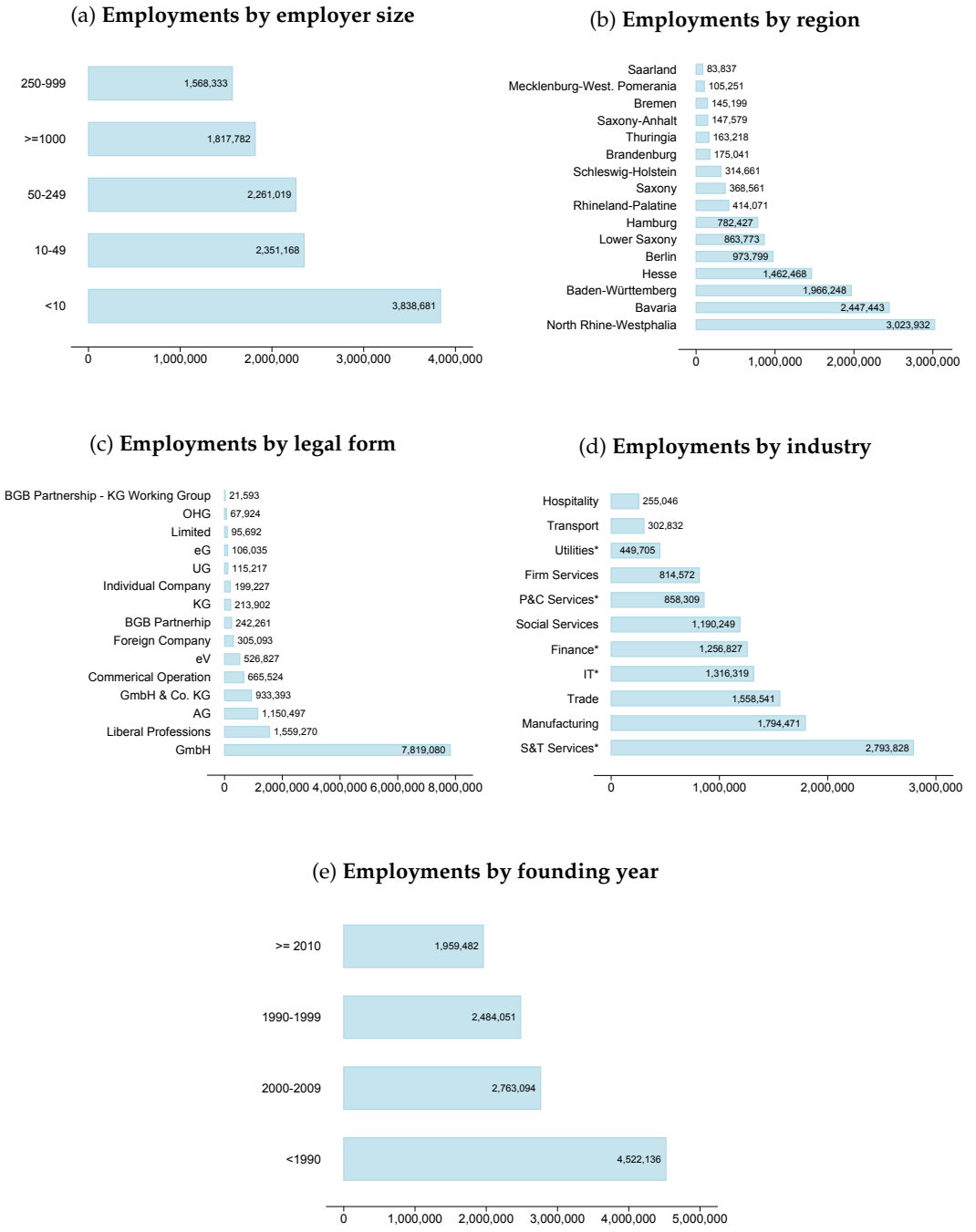


Figure 5.5: **Employment counts (XING) with respect to median employer size (top left), region (top right), legal form (middle left), employer industry (middle right), and employer founding year (bottom).** Only matched and public employments are used. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 and Table D.3 for a list of all labels). Source: TUM and ZEW based on data from XING and MUP. Own illustration.

First, small employers (<10 employees) are most often mentioned in employments. We expected this pattern as a large proportion of employees in Germany work for small and medium-sized enterprises (SMEs).[109] The smallest number of employments is found for employers with 250 to 999 employees. Second, most employments are linked to employers in North Rhine-Westphalia, Bavaria, and Baden-Württemberg. The fewest employments are found in Saarland, Mecklenburg-Western Pomerania, and Bremen. The results are plausible because these regions have the highest/lowest population counts. Third, most employments are linked to limited liability companies (*GmbH*) and *liberal professions*. Furthermore, employments within the *Freelance, scientific and technical services* (S&T Services) industry and for employers founded before 1990 are the most frequent. Fewer employments exist in the *hospitality* and *transport* industries and for employers founded since 2010.[110]

5.4.2 Employers

Figure 5.6 shows selected MUP characteristics for the matched XING employers.[111] The XING employer data set cannot be compared directly with the total stock of German employers in the MUP. For example, the XING data contain employers that are no longer economically active. Our findings suggest: First, most employers on XING have less than 10 employees, and only a few employers have at least 250 employees. Second, most employers are found in the industry *Freelance, Scientific and Technical Services (69-75)*; the fewest employers are in the industry *transport (49-53)*.[112] Third, most employers are founded in *'2000 - 2009'* or *'2010 & later'*; there is a small decrease in the number of founded employers *'before 2000'*. Fourth, most employers are located in the German regions of North Rhine-Westphalia, Bavaria, and Baden-Württemberg (in the former territory of West Germany). The fewest are located in Mecklenburg-Western Pomerania, Bremen, and Saarland. Many regions of the former German Democratic Republic (GDR) are in the lower half of the ranking, illustrating an east-west divide. The cities of Berlin and Hamburg are in the midfield of the ranking. In summary, the XING employer data seem plausible and, with some limitations, representative of Germany.[113]

[109] Legal entities and employees by employer size and industry: `https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/Unternehmensregister/Tabellen/unternehmen-beschaeftigtengroessenklassen-wz08.html` [Last accessed: 24.02.2024].

[110] The presented figures pool historical XING data. Therefore, they are no representation of the current employments in Germany. For comparison, the employer characteristics for the stock of MUP employments (2002-2019) in Germany are described in Figure D.5. Larger deviations in the MUP and XING data are found for the founding period, employer size, and some industries.

[111] The employers are extracted from the XING employments that were matched to the MUP.

[112] The numbers in parentheses are the two-digit NACE codes. The NACE codes are the *Statistical Classification of Economic Activities in the European Community*. For a definition, see `https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF` [Last accessed: 24.02.2024].

[113] Figure D.4 shows the share of MUP employers (2002-2019) and XING employers in Germany with respect to regions, founding year periods, legal forms, employer size classes, and industries. The

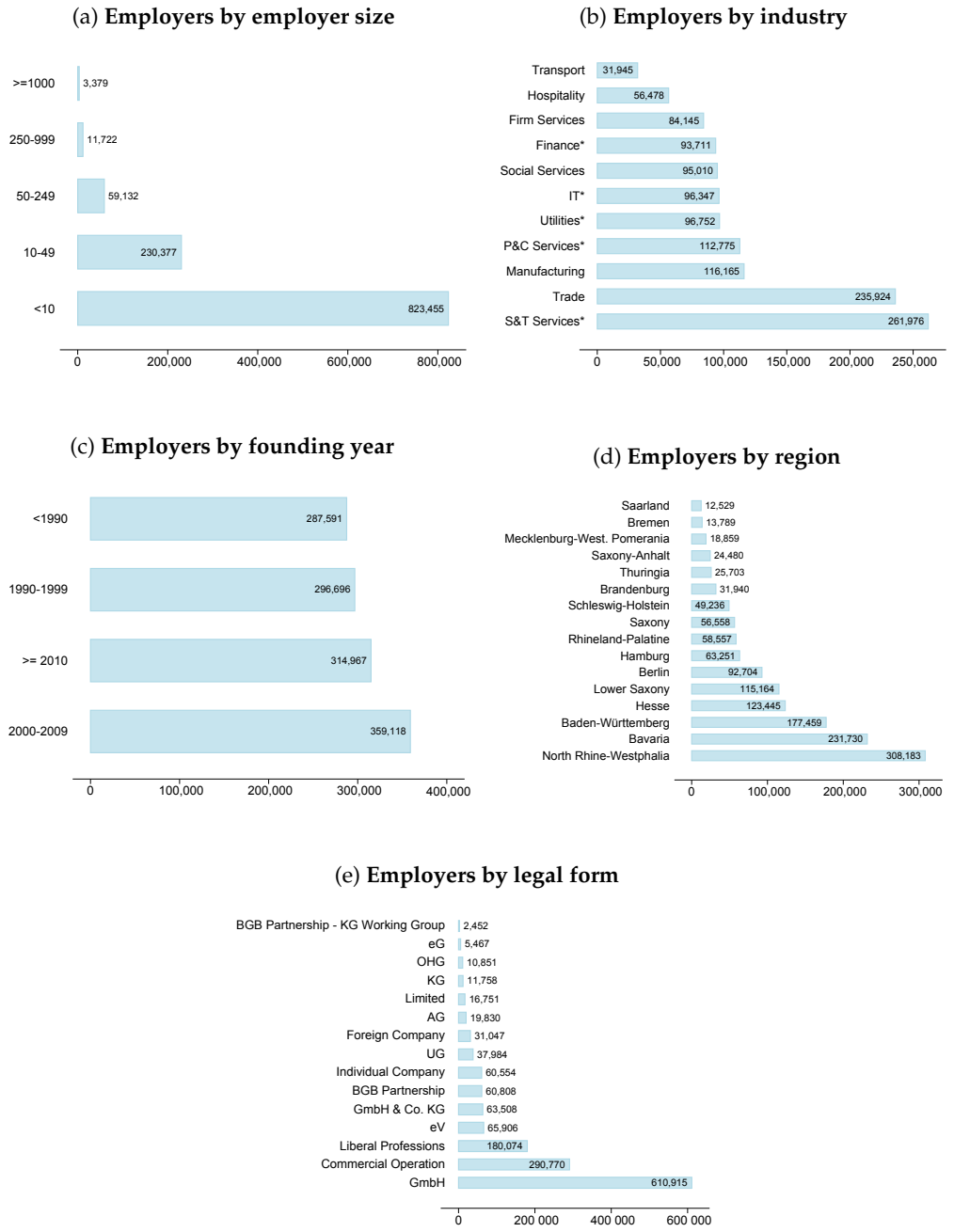


Figure 5.6: **Number of XING employers with respect to MUP characteristics; based on matched employers.** Note: (a): Per median employer size group; (b): Per industry; (c): Per founding year group; (d): Per German region; (e) Per legal form. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 and Table D.3 for full labels). Source: TUM and ZEW based on XING and MUP data. Own illustration.

distributions of the XING and MUP data appear to be roughly similar. However, some deviations exist, e.g. in terms of legal form and industry. For us, it seems plausible that, for example, employers in the utility industry are less covered as these employers and their workforce are less reliant on platforms.

5.4.3 Employee flows

Figure 5.7 shows the employee flows of XING users by their employment characteristics, i.e. career level, discipline, employment type, employer size, industry, and region. First, we analyse the flows between employers with respect to career levels. Most users transfer out of employments with professional experience. Users usually switch employments within an experience class or into a higher experience class, e.g. from *young professional* to *professional*. Employees rarely reduce their experience level. For example, if an employee is moving into a new industry. Second, we analyse the flows between employers for employment disciplines. Most users switch employments within their discipline.[114] However, we find flows for all discipline combinations. Third, we analyse the flows between employers by their employment type. Here, we find a deviating pattern: Many of the employment types show a flow into the *'Full-time'* class. Further validity checks show that civil servants move only occasionally into unpaid employments, and young professionals rarely move into (managing) director positions. Fourth, many flows occur within the same employer size class, and relatively few flows are found for employment changes to large-sized employers. Fifth, many employees are moving within industries. However, employers in the *Freelance, Scientific and Technical Services* (69-75) industry receive substantial inflows from all other industries. Sixth, the majority of employees change employers within their region. However, the regions of Bavaria, North Rhine-Westphalia, and Baden-Württemberg have substantial inflows from all other regions.

Figure D.2 shows the flows between the regions of West and East Germany.[115] Most employees change employers within West Germany. Inflows from West Germany do not offset outflows from East Germany. Lastly, we present the employee flows by district type (Figure D.2).[116] Most employees switch employers within big cities and the employee flow data do not show a strong rural exodus.

[114] Please note that the disciplines are quite coarsely defined.
[115] Berlin is not considered, as the assignment to West or East Germany is not clear.
[116] For definitions, see Table D.12.

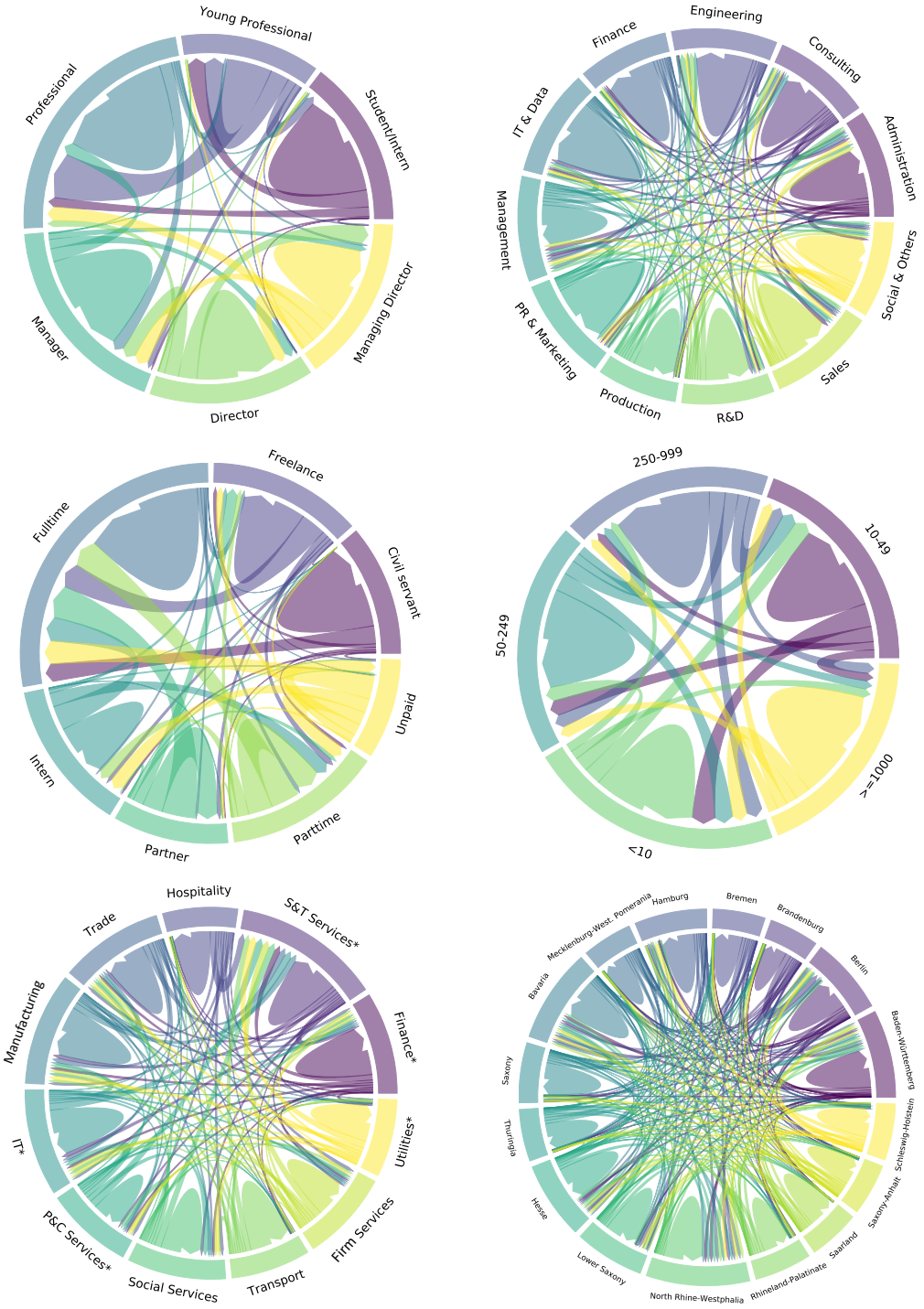


Figure 5.7: **Employee flows as chord diagrams.** Top left: Career levels. Top right: Employment discipline, Middle left: Employment type, Middle right: Employer size, Bottom left: Employer industry, Bottom right: Employer region. Flows are based on matched employments from public user profiles. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 for a list of all labels). Table D.8 and Table D.9 present the respective flow matrices. Source: TUM and ZEW based on XING and MUP data. Own illustration.

5.4.4 Networks

Table 5.4 provides the definitions of network metrics and references to their implementation.[117] Table 5.5 presents the metrics for the employee flow networks. The annual networks are available for the period 2010 to 2020. The nodes are the employers, while the edges are the employee flows.

Table 5.4: **List of network metrics.**

Metric	Definition	Function
#Nodes	Number of nodes in the network.	graph.vcount()
#Edges	Number of unique edges in the network.	graph.ecount()
#Cliques	Number of complete subgraphs, where an edge is existent between any two nodes (excl. loops).	len(list(graph.cliques()))
Max. clique size	The number of nodes in the largest clique.	graph.clique_number()
Transitivity	Measures the probability that two neighbours of a node are connected. Calculated for each node and then averaged. Vertices with less than two neighbours are ignored.	graph.transitivity_avg local_undirected()
Reciprocity	Reciprocity is defined as the probability that a directed edge's opposite counterpart (in the other direction) is also included in the network.	graph.reciprocity()
#Clusters	Number of strongly connected components in the network. A strongly connected component is a subgraph, where every node is reachable from every other node.	len(list(graph.clusters()))
Density	Ratio of the edge count by the maximum possible edge count.	graph.density()
Girth	Length of the shortest circle in the network. Circles consist of at least three nodes.	graph.girth()

Notes: Description of network metrics. The table provides the definitions of the metrics and the igraph function (Python: `https://igraph.org/python/` [Last accessed: 24.02.2024]).

The clique counts of the graphs vary between 181K and 649K. These cliques may be, for example, highly interconnected employers within the same industry or region. Large cliques consist of smaller cliques as subgraphs. Therefore, the number of cliques is quite high. The maximum clique size varies between 8 and 16 and is, on average, 12. The cluster counts lie between 64K and 133K, and the graphs' densities vary between $9.6\text{x}10^{-6}$ and $1.8\text{x}10^{-5}$. Thus, many pairs of employers have no employee flows and may also not be indirectly linked by employee flows. The transitivity lies between 0.10 and 0.12, and the reciprocity is between 0.05 and 0.07. Mutual, e.g. between direct competitors, and transitive employee flows between employers are, thus, not very common.

[117] Figure D.7 provides a directed example graph for the metrics from Table 5.5. Further details about the implementation of the metrics are available at `https://igraph.org/python/doc/api/igraph.Graph.html` [Last accessed: 24.02.2024].

Table 5.5: **Annual employee flow networks: Metrics.**

Metric	Min.	Max.	Mean
#Nodes	74,367	163,706	142,162
#Edges	99,602	288,805	221,065
#Cliques	181,363	649,221	456,504
Max. clique size	8	16	11.72
Transitivity	0.10	0.12	0.11
Reciprocity	0.05	0.07	0.06
#Clusters	64,501	133,471	118,768
Density	$9.6x10^{-06}$	$1.8x10^{-05}$	$1.1x10^{-05}$
Girth	3	3	3

Notes: Network metrics for the series of annual graphs (2010-2020). Some deviations are partly due to the lower number of employee flows in earlier years and the incomplete coverage of 2020.

Exemplary, Figure 5.8 presents the intra-city flows between employers that are located in Munich for the year 2019.[118] The network comprises 2,982 employers and 4,235 unique employee flows and is a subgraph of the complete employee flow data set. Large parts of the network are connected, and some central employers have many incident edges. Some employers have only one incident edge and are not or only sparsely connected to the rest of the network. Employers without flows are not represented in the network. Comparable analyses can be performed for arbitrary time periods and cities or regions; see Figure D.3 for Mannheim. For 2019, the network for a smaller city like Mannheim includes 236 employers and 221 unique employee flows. Therefore, the network is smaller than the network for Munich. Again, a large share of the employers and flows are found in the city center.

In contrast, Figure D.6 shows the relative flow counts for the cities of Berlin, Cologne, Hamburg, and Munich. We present the ten districts with the most employee flows to or from the four cities. Berlin, Hamburg, and Munich play an important role in employee mobility because there is a high employee exchange between larger cities. Furthermore, regional differences exist, and districts in the closer neighbourhood have a special role. For example, the city of Cologne has a high level of employee exchange with the nearby districts of *Bonn* and *Rhein-Sieg-Kreis*.

Figure 5.9 shows network-based measures for German regions. The figure on the left shows the log-scaled number of clusters per district. However, the number of clusters is not indicative of their size. Large cities and districts in West Germany have the most regional clusters. The figure on the right shows the network density scores per district. Large parts of East Germany have the highest scores. The maximum number of edges in a network scales quadratically with the node count. As a result,

[118] Table D.11 describes the matching of XING employers to geographical coordinates.

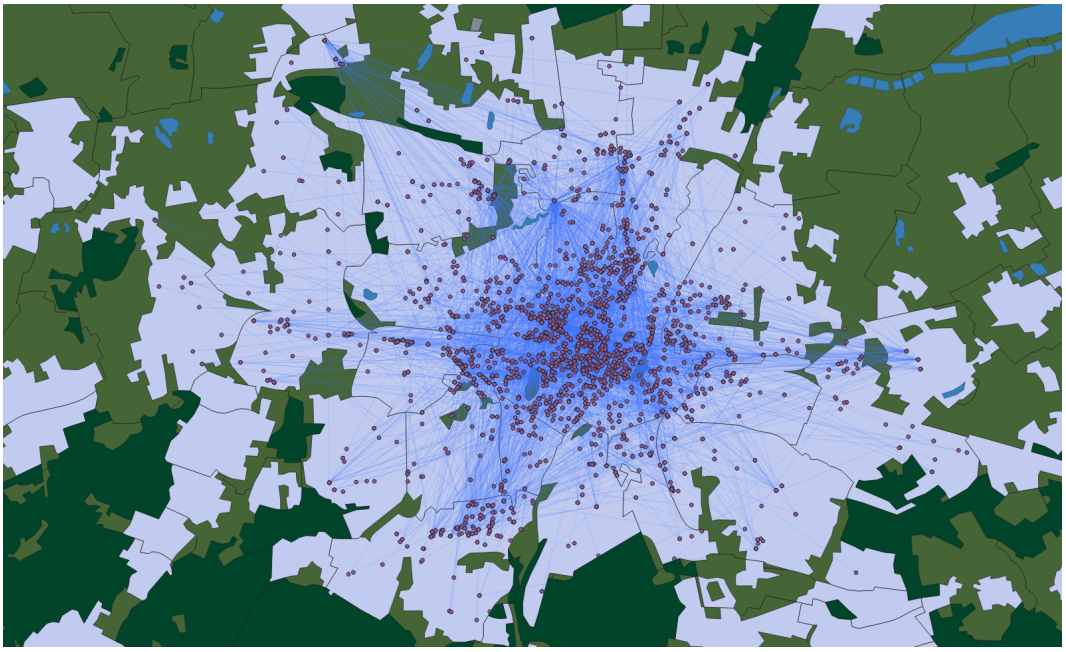


Figure 5.8: **Employee flow network for the city of Munich.** The employee flow data are restricted to flows within Munich in 2019. Source: TUM and ZEW based on XING data. Own illustration (created with QGIS: `https://www.qgis.org` [Last accessed: 24.02.2024]).

the density score is directly impacted in sparse networks. A possible explanation is that there are fewer employers in East Germany and, for example, in the region of Saarland, but those are better connected by employee flows. Figure 5.10 presents the degree centrality, i.e. the number of employees moving to or from an employer, by employer characteristics. First, we analyse the degree centrality by employer size. The employer size is the median number of employees per employer in the panel. Employers with larger employee counts have higher degree centrality scores. Second, employer age is positively linked with high degree centrality scores. We expect this link as employer age correlates with employer size. Third, employers listed as publicly traded (*'AG'*) have, by far, the highest degree centrality. We find the lowest centrality scores for entrepreneurial companies at limited liability (*'UG'*) and commercial operations.[119] Fourth, the industries *'Social Services'*, *'Finance, Insurance & Real Estate'* and *'Information and Communication'* have the highest centrality scores. The industries *'Hospitality'* and *'Utilities/Construction'* have the lowest scores. Fifth, big cities and urban districts have higher scores than rural districts. Sixth, the regions of Berlin and Hamburg have the highest scores. Brandenburg and Saarland have the lowest scores. The centrality scores show an east-west divide.

In the last step, we perform plausibility checks at the employer level to verify that the XING and MUP data are statistically related. Thereby, we demonstrate that

[119] Table D.3 presents the German and translated labels (English).

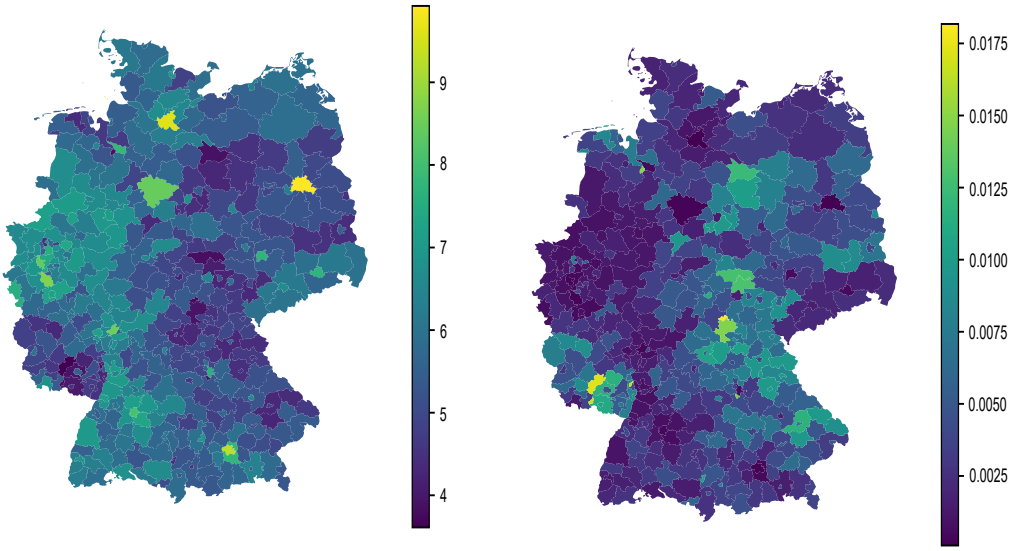


Figure 5.9: **Selected network measures per German district. Left: Number of clusters per district (log-scaled). Right: Network density per district.** Both measures are calculated for district-level networks without temporal restrictions. Data: Public and matched XING employments and flows. Source: TUM and ZEW based on XING and MUP data. Own illustration (created with geopandas: `https://geopandas.org/en/stable/` [Last accessed: 24.02.2024]).

the flow data are externally valid and model the actual inflows and outflows of German employers, although they are by no means complete. The absolute number of employees and flows cannot be observed over time on XING, as many employees and employers are not active on the platform.

We use MUP and XING data to analyse the link between the changes in the employee counts (see Table 5.6).[120] Changes in the employment counts (= in-degree minus out-degree of an employer within one year) are measured using employee flow data from the platform XING. The results show that the variables are positively and significantly related. However, the coefficients for ΔEmployees (MUP data) are smaller than one, indicating that we do not capture the entire inflow and outflow of employers. Reasons for this include inaccuracies in the MUP employment figures and the flow data. Furthermore, not all types of employees are well represented on XING. The link persists for the original data (Columns 1 and 3) and log-transformed variables (Columns 2 and 4). Unfortunately, the R^2 scores are relatively low.[121]

[120] We start with 1.6M observations from Table 5.3, i.e. the sum of network nodes over the period from 2010 to 2020. We delete observations if industry, region, founding date, or legal form data are missing, leaving 1.4M observations. Other observations are omitted due to missing employee count data.

[121] Robustness check: In each year, we consider all 964K employers with at least one employee flow between 2010 and 2020. Missing centrality scores are set to zero if no flows are found for the employers in the year in question. The results again show that the coefficients for the change in employee counts are positive and highly significant.

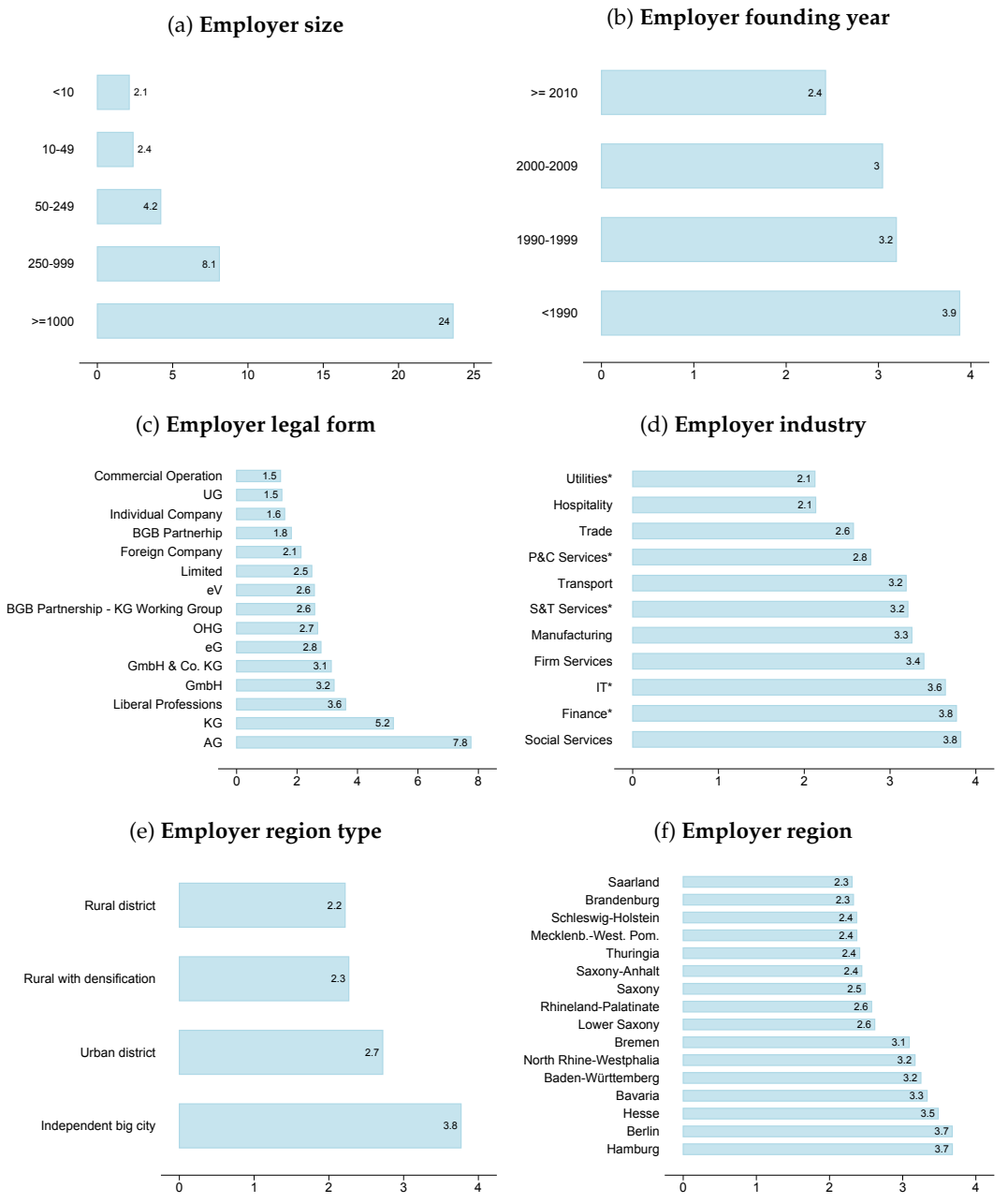


Figure 5.10: **Degree centrality by employer characteristics.** The employer data are modelled as a panel. Missing values are not shown. Employers without employee flows within a year are not considered in the respective annual network. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 and Table D.3 for a list of all labels). Source: TUM and ZEW based on XING and MUP data. Own illustrations.

Table 5.6: **Regressions: Degree centrality.**

	(1) ΔDegree	(2) Δlog(Degree+1)	(3) ΔDegree	(4) Δlog(Degree+1)
ΔEmployees	0.00266***		0.00267***	
	(<0.001)		(<0.001)	
Δlog(Employees+1)		0.176***		0.174***
		(0.00237)		(0.00237)
Year			0.00154	0.00105***
			(0.00118)	(<0.001)
Constant	-0.0341***	-0.0292***	-2.869	-2.104***
	(0.00335)	(<0.001)	(2.373)	(0.623)
Industry Dummies	No	No	Yes	Yes
Legal form Dummies	No	No	Yes	Yes
Region	No	No	Yes	Yes
Founding Dummies	No	No	Yes	Yes
N	767,652	767,652	767,652	767,652
R^2	0.004	0.009	0.005	0.011
adj. R^2	0.004	0.009	0.005	0.011

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Standard errors are in parentheses. We use degree centrality measures as the dependent variable (XING) and the employer characteristics as explanatory variables (MUP). Source: TUM and ZEW based on XING and MUP data. Definitions:
ΔDegree(t) = In-Degree(t) - Out-Degree(t)
Δlog(Degree+1) = log(In-Degree(t)+1) - log(Out-Degree(t)+1)
ΔEmployees(t) = Employees(t+1) - Employees(t-1).

5.5 Conclusion

The availability of social media data creates new opportunities for empirical economic research. This paper presents a detailed exploration of using data from the career-oriented social networking platform XING for measuring employee flows and employer networks. We obtain employment data from public user profiles and link them to the Mannheim Enterprise Panel (MUP). The novel link creates a unique platform-based LEE data set that allows tracking employee flows between employers. The data set comprises about 14M employments for 1.5M disambiguated employers. Furthermore, the matched employment data are used to extract 9 million employee flows and create annual flow networks for the years 2010 to 2020.

We check the plausibility of the data set and show that career-oriented social networking data contain meaningful and valuable data about employments, employers, employee flows, and employer networks. In doing so, we test the plausibility

of selected network measures against employer data and find that they show convincing patterns. For example, the difference in the employer-level degree centrality scores (in-degree minus out-degree) positively correlates with changes in MUP employment counts. Thus, the professional network data appear to be externally valid and can model the inflows and outflows of German employers.[122] Notably, the coverage goes beyond paid employment by including, for example, freelance work and some entrepreneurial activities. For the analysis of networks, the coverage is sufficiently high, and network measures can be derived for employers that are neither active in patenting nor engaged in larger, visible alliances. Nevertheless, using such data for research purposes should be done with care.

Platforms like XING and LinkedIn may have some biases concerning their users. For example, older employees, employees without training, and employees from specific industries, such as household services, are less frequently represented. Linked Employer-Employee data from official sources provide a better overview of these employment relationships, e.g. the IAB data from the Federal Employment Agency[123], even though this also comes with other restrictions as discussed at the beginning of the paper. Furthermore, self-employed workers may not be correctly matched in our data set because they could not be linked to a MUP employer. In principle, however, it is possible to recognise these employments and treat them automatically. Furthermore, we only use publicly available data from the platform XING to comply with the users' desire to keep their data non-public. As a result, we might create biased data as, for example, specific user groups are more concerned about their privacy, e.g. employees in the fields of law and IT. Unfortunately, investigating this hypothesis in more detail is not straightforward. Also, some users do not update their profiles frequently, which creates the impression that an employee has been with the employer longer than is the case. Another way this can happen is if a user changes platforms and instead maintains his LinkedIn profile. However, our platform-based approach has a multitude of advantages. The data are subject to fewer legal constraints, can be updated regularly, and are publicly accessible.

We want to highlight that employers on XING and MUP are linked using fuzzy string matching (Doherr 2023). However, our method is not error-free. The matching approach can be further improved, for example, by adjusting the parameters or using different matching methods. There is a discrepancy in the unit of observation: On XING, employer sites are often listed and linked to employments. The MUP, however, consists of employers that are legally independent entities. Therefore, a link of multiple employer sites to one employer is necessary. Furthermore, employers with few employees may have biased representation in the network, as there might not be any publicly accessible employments for them on XING. In addition,

[122] Another possibility for validation is a comparison with LEE data from official data sources.
[123] Institute for Employment Research: `https://iab.de/` [Last accessed: 24.02.2024].

the networks are constructed on an annual basis, so there are hard boundaries for the annual networks. For example, employment switches on New Year's Eve or New Year will thus end up in two different networks. Smooth transitions between the annual networks would improve the quality of the data, e.g. by including the previous year's flows with a smaller weight.[124] The extraction of employee flows does not consider many exceptional cases, e.g. concurrent employments are not modelled. Further, time gaps between multiple employments are, so far, not considered. In the future, we might model them as separate nodes representing unemployment.

In conclusion, despite these challenges, we are able to link a large share of employers due to careful disambiguation of employer names, URLs, and profiles. The plausibility checks suggest that the resulting data have no major shortcomings. Hence, the new database provides opportunities for being used in subsequent research on the role of employee mobility, networks, and local ecosystems for economic performance both at the employer and the regional level. The micro-nature of the data allows, for example, the calculation of indicators on the level of the network nodes. These include centrality measures for employers and aggregate measures for network characteristics at the regional level. Data availability over time further facilitates analyses of the network's development and the drivers of these changes.

[124]There are some dynamic time series methods to tackle this problem in future studies, e.g. the sliding window model (Datar et al. 2002).

Chapter 6

Concluding Remarks

In the following section, I conclude this dissertation. First, I summarise the four essays. Second, I describe the limitations and hurdles of the methodology. Third, I outline the potential of the indicators and give some directions for future work.

1. Summary

My co-authors and I used various methods to collect a wide range of public firm data, which have the potential to complement or replace traditional data for economic research. First, we scraped firm websites (see Chapter 2 and Chapter 3). Second, we used the scraped data from online platforms (see Chapter 4). Third, we gained access to firm data by cooperating with a platform provider (see Chapter 5).

These heterogeneous data sets were linked to a standardised firm database. First, Chapters 2 and 3 used Mannheim Web Panel (MWP), Mannheim Innovation Panel (MIP), and Mannheim Enterprise Panel (MUP) data. In this case, the linking between the databases was almost trivial, as the MUP is the starting point for the MIP and MWP data collection. Second, in Chapter 4, platform data from Facebook and Kununu were used. Although the initial search for firm profiles on the platforms Kununu, Facebook, and Google was based on data from the MUP, the results contained many false positives, i.e. firm profiles not referring to the input data. These had to be corrected using NLP and matching methods. Third, linking the data in Chapter 5 turned out to be the most complex, but a large proportion of the firm profiles on XING were linked to the MUP. The setup between the target database and the web data was highly relevant in terms of required effort and matching quality.

Two types of data were used in this dissertation: Unstructured text data from firm websites and (semi-)structured tabular data from platforms. First, text data were used in Chapters 2 and 3. The data consisted primarily of texts from firm websites and were, therefore, highly unstructured. The objective was to structure the data in such a way that a machine learning model could be trained on it. We used natural language processing methods for this data transformation. Second, Chapters 4 and 5 focused on tabular or (semi-)structured data. The main focus was

not on the conversion of data into a vector space but on the standardisation and cleansing of the data.

We used the web data to measure technological change at the firm level. In each essay, my co-authors and I calculated at least one indicator for technological change or a related measure of technology diffusion. These measures included firm-level indicators for innovation (Chapter 2), digitalisation (Chapter 3), marketing and on-the-job training (Chapter 4), as well as employee mobility (Chapter 5).

The indicators were then validated to confirm their quality. To do so, we used survey-based information from MIP and Eurostat, as well as the MUP database. In Chapter 3, comparisons with established digitalisation indicators showed that the approach provided plausible results at the firm, regional, size, and industry levels. The indicators in Chapters 2 and 4 were compared with firm-level MIP data. In Chapter 5, validation was carried out using firm-level MUP data. The web indicators in all essays showed a statistical link to the validation data sets.

The indicators are relevant for several reasons: Compared to traditional measures, web indicators have the advantage of being able to be collected quickly and cheaply updated. In addition, the scores are available for many firms, i.e. for around one million German firms which have a website. Furthermore, the web indicators are subject to fewer legal restrictions than data from official bodies. As a result, the indicators can be merged with other data more easily. Thus, the websites of firms and digital platforms provide a source for constructing novel indicators to answer research and policy questions in various fields. For example, we used the results of Chapter 3 to explore the link between digitalisation and the resilience of firms during the COVID-19 crisis.

2. Limitations and hurdles

A major limitation is that the publicly available data may not be entirely correct. Targeted misinformation, such as greenwashing or overstating digitalisation efforts, poses a problem to the data quality (e.g. Vos 2009). In our case, the content of the website data is also limited (up to 50 subpages per website), and the age of texts found on the web can often only be approximated. Interactive elements, meta data, image, audio, and video data on the websites are not recorded. Many small firms might not be covered by the data sets because they often manage their businesses on platforms such as Instagram, as opposed to a traditional website. This presents a challenge, as on many platforms, certain information cannot be retroactively collected or reconstructed. Web-based indicators also bring about limitations, as they are often not able to cover nuanced topics, e.g. monetary investments in certain technologies. Therefore, these web-based indicators are more suitable for broader subject areas. Further, each web indicator needs to be validated with independent

data to ensure its quality. A suitable technical infrastructure and a long-term project are needed to reliably calculate the indicators. Lastly, acceptance among researchers and statistical offices for the web-based indicators has yet to be established.

3. Potentials and outline for future work

In this section, I describe ways to expand web data, and I give directions for future research and related policy advice.

I propose scraping European and worldwide web data. There are more and more European regulations (Digital Markets Act[125], EU Supply Chain Law Initiative[126], etc.) as well as discussions about, for example, a worldwide minimum taxation. A successful evaluation of these regulations can only be implemented with web data if all (European) countries are covered by the data. To achieve this, additional data must be used, as the MUP primarily covers Germany. I propose using a worldwide database of firms, such as ORBIS[127], which contains data on around 450 million firms. Collecting and processing worldwide data presents new hurdles, such as the large number of languages used. However, the problem can be overcome with statistical machine translation. Further, web data should be scraped more frequently. In the event of an exogenous shock, such as COVID-19 or comparable future events, research and policy advice can be provided more quickly and accurately. Furthermore, the data benefits when a long-term time series is created. The MWP data has been collected since 2018[128], but the platforms are not scraped regularly.

There are various paths for future research. One promising area is the development of innovative methodologies for firm-level indicators. Statistical methods such as long short-term memory (LSTM) can be used to achieve better predictive performance (Hochreiter & Schmidhuber 1997). However, both accuracy and explainability should be considered, as they form the basis for improved acceptance in science and administrative offices. Further, not only the firms can be identified in the data but also places, people, and time points by using Named Entity Recognition (e.g. Nadeau & Sekine 2007, Li et al. 2020). Technological leaps, such as quantum computers, can also have a positive impact, as some types of calculations can be processed much faster, i.e. making certain algorithms applicable in the first place. New or supplementary data sets should also be used, e.g. public annual reports and job

[125]Digital Markets Act: `https://digital-markets-act.ec.europa.eu/` [Last accessed: 24.02.2024].
[126]EU Supply Chain Law Initiative: `https://www.csr-in-deutschland.de/EN/Business-Human-Rights/Europe/EU-supply-chain-law-initiative/eu-supply-chain-law-initiative.html` [Last accessed: 24.02.2024].
[127]ORBIS data: `https://www.bvdinfo.com/en-us/our-products/data/international/orbis` [Last accessed: 24.02.2024].
[128]MWP data: `https://kooperationen.zew.de/en/zew-fdz/provided-data/mannheim-webpanel` [Last accessed: 24.02.2024].

postings. The latter in particular provides very valuable insights into the firms, as the required technological experience of the applicants is listed in the job postings.[129]

Lastly, the indicators for firm-level technological change that are developed in this dissertation can be used for downstream economic research. The LEE data set created in Chapter 5 can be utilised to analyse the relationship between employee flows and the innovative capacity of firms. The digitalisation indicator from Essay 3 is already being used to explain changes in the mobility patterns of employees during the COVID-19 crisis (Axenbeck et al. 2023). There are also possible applications in neighbouring disciplines. For example, we use the MWP and MUP data sets to analyse the integrity of a scientific use file (SUF) in a data attack scenario, i.e. an attempt is made to de-anonymise the protected data in a monitored environment.

To summarise, the use of web-based data and machine learning techniques for measuring technological change promises many opportunities for economic research and policy advice, which are typically difficult to address with traditional data and methods.

[129] Data provider for job postings (example): `https://www.textkernel.com` [Last accessed: 24.02.2024].

Bibliography

Abbasiharofteh, M., Kinne, J. & Krüger, M. (2021), The strength of weak and strong ties in bridging geographic and cognitive distances, ZEW Discussion Paper No. 21-049, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *https://ftp.zew.de/pub/zew-docs/dp/dp21049.pdf*

Abendroth, A.-K., Melzer, S. M., Jacobebbinghaus, P. & Schlechter, F. (2014), Methodological Report Employee and Partner Surveys of the Linked Employer-Employee Panel (LEEP-B3) in Project B3 „Interactions Between Capabilities in Work and Private Life: A Study of Employees in Different Work Organizations“, SFB 882 Technical Report Series 12, University of Bielefeld, Bielefeld. [Online; last accessed 12.01.2024].
URL: *https://core.ac.uk/download/pdf/211841852.pdf*

Abidi, N., El Herradi, M. & Sakha, S. (2022), Digitalization and Resilience: Firm-level Evidence During the COVID-19 Pandemic, IMF Working Paper 2022/034, International Monetary Fund, Washington DC. [Preprint (Online); last accessed 12.01.2024].
URL: *https://www.imf.org/-/media/Files/Publications/WP/2022/English/wpiea2022034-print-pdf.ashx*

Ackland, R., Gibson, R., Lusoli, W. & Ward, S. (2010), 'Engaging With the Public? Assessing the Online Presence and Communication Practices of the Nanotechnology Industry', *Social Science Computer Review* **28**(4), 443–465.
URL: *https://doi.org/10.1177/0894393103627235*

Acquaah, M., Amoako-Gyampah, K. & Jayaram, J. (2011), 'Resilience in family and nonfamily firms: an examination of the relationships between manufacturing strategy, competitive strategy and firm performance', *International Journal of Production Research* **49**(18), 5527–5544.
URL: *https://doi.org/10.1080/00207543.2011.563834*

Adarov, A. & Stehrer, R. (2019), Tangible and Intangible Assets in the Growth Performance of the EU, Japan and the US, Research Report 442, Vienna Institute for International Economic Studies, Wien. [Preprint (Online); last accessed 12.01.2024].

URL: *https://wiiw.ac.at/tangible-and-intangible-assets-in-the-growth-performance-of-the-eu-japan-and-the-us-dlp-5058.pdf*

Aguado, D., Andrés, J. C., García Izquierdo, A. L. & Rodríguez, J. (2019), 'LinkedIn "Big Four": Job Performance Validation in the ICT Sector', *Journal Of Work And Organizational Psychology* **35**(2), 53–64.
URL: *https://doi.org/10.5093/jwop2019a7*

Allcott, H., Gentzkow, M. & Yu, C. (2019), 'Trends in the diffusion of misinformation on social media', *Research & Politics* **6**(2).
URL: *https://doi.org/10.1177/2053168019848554*

Almeida, P. & Kogut, B. (1999), 'Localization of Knowledge and the Mobility of Engineers in Regional Networks', *Management Science* **45**(7), 905–917.
URL: *https://doi.org/10.1287/mnsc.45.7.905*

Archibugi, D. & Planta, M. (1996), 'Measuring technological change through patents and innovation surveys', *Technovation* **16**(9), 451–468, 519.
URL: *https://doi.org/10.1016/0166-4972(96)00031-4*

Arora, A., Arora, A. S. & Palvia, S. (2014), 'Social Media Index Valuation: Impact of Technological, Social, Economic, and Ethical Dimension', *Journal of Promotion Management* **20**(3), 328–344.
URL: *https://doi.org/10.1080/10496491.2014.908803*

Arora, S. K., Youtie, J., Shapira, P., Gao, L. & Ma, T. (2013), 'Entry strategies in an emerging technology: A pilot web-based study of graphene firms', *Scientometrics* **95**(3), 1189–1207.
URL: *https://doi.org/10.1007/s11192-013-0950-7*

Arundel, A. & Kabla, I. (1998), 'What percentage of innovations are patented? empirical estimates for European firms', *Research Policy* **27**(2), 127–141.
URL: *https://doi.org/10.1016/S0048-7333(98)00033-X*

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C. & Cunningham, S. (2022), 'Indicators on firm level innovation activities from web scraped data', *Data in Brief* **42**, 108246.
URL: *https://doi.org/10.1016/j.dib.2022.108246*

Awano, G., Franklin, M., Haskel, J. & Kastrinaki, Z. (2010*a*), Investing in Innovation: Findings from the UK Investment in Intangible Asset Survey, Index Report July 2010, National Endowment for Science, Technology and the Arts (NESTA), London. [Online; last accessed 12.01.2024].

URL: *https://media.nesta.org.uk/documents/investing_in_innovation.pdf*

Awano, G., Franklin, M., Haskel, J. & Kastrinaki, Z. (2010*b*), 'Measuring investment in intangible assets in the UK: results from a new survey', *Economic & Labour Market Review* **4**(7), 66–71.
URL: *https://doi.org/10.1057/elmr.2010.98*

Axenbeck, J., Bertschek, I., Breithaupt, P. & Erdsiek, D. (2023), Firm Digitalisation and Mobility - Do Covid-19-Related Changes Persist?, ZEW Discussion Paper No. 23-011, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online), last accessed 12.01.2024].
URL: *https://ftp.zew.de/pub/zew-docs/dp/dp23011.pdf*

Axenbeck, J. & Breithaupt, P. (2021), 'Innovation indicators based on firm websites - Which website characteristics predict firm-level innovation activity?', *PLOS ONE* **16**(4), e0249583.
URL: *https://doi.org/10.1371/journal.pone.0249583*

Axenbeck, J. & Breithaupt, P. (2022), Measuring the Digitalisation of Firms – A Novel Text Mining Approach, ZEW Discussion Paper No. 22-065, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *http://ftp.zew.de/pub/zew-docs/dp/dp22065.pdf*

Baeza-Yates, R. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM press, ISBN: 978-0-201-39829-8, New York, NJ.
URL: *https://dl.acm.org/doi/10.5555/553876*

Balsvik, R. (2011), 'Is labor mobility a channel for spillovers from multinationals? Evidence from Norwegian manufacturing', *The Review of Economics and Statistics* **93**(1), 285–297.
URL: *https://doi.org/10.1162/REST_a_00061*

Banerji, D. & Reimer, T. (2019), 'Startup founders and their LinkedIn connections: Are well-connected entrepreneurs more successful?', *Computers in Human Behavior* **90**, 46–52.
URL: *https://doi.org/10.1016/j.chb.2018.08.033*

Becker, W. & Dietz, J. (2004), 'R&D cooperation and innovation activities of firms - evidence for the German manufacturing industry', *Research Policy* **33**(2), 209–223.
URL: *https://doi.org/10.1016/j.respol.2003.07.003*

Behrens, V., Berger, M., Hud, M., Hünermund, P., Iferd, Y., Peters, B., Rammer, C. & Schubert, T. (2017), Innovation Activities of Firms in Germany - Results of the German CIS 2012 and 2014: Background Report on the Surveys of the Mannheim Innovation Panel Conducted in the Years 2013 to 2016, Documentation 17-04, ZEW - Leibniz Centre for European Economic Research. [Online; last accessed 12.01.2024].
URL: *http://ftp.zew.de/pub/zew-docs/docus/dokumentation1704.pdf*

Belderbos, R., Carree, M. & Lokshin, B. (2004), 'Cooperative R&D and firm performance', *Research Policy* **33**(10), 1477–1492.
URL: *https://doi.org/10.1016/j.respol.2004.07.003*

Bellstam, G., Bhagat, S. & Cookson, J. A. (2020), 'A Text-Based Analysis of Corporate Innovation', *Management Science* **67**(7), 4004–4031.
URL: *https://doi.org/10.1287/mnsc.2020.3682*

Bersch, J., Gottschalk, S., Müller, B. & Niefert, M. (2014), The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany, ZEW Discussion Paper No. 14-104, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *https://ftp.zew.de/pub/zew-docs/dp/dp14104.pdf*

Bertenrath, R., Fritsch, M., Lichtblau, Karl & Schleiermacher, T. (2017), Digitale Wirtschaft Nordrhein-Westfalen, Studie im Auftrag der Initiative Digitale Wirtschaft NRW des Ministeriums für Wirtschaft, Energie, Industrie, Mittelstand und Handwerk des Landes Nordrhein-Westfalen, IW Consult, Köln. [Online; last accessed 12.01.2024].
URL: *https://www.iwkoeln.de/fileadmin/publikationen/2017/334156/IW-Gutachten_Digitale_Wirtschaft_NRW_Endbericht.pdf*

Bertschek, I., Briglauer, W., Hüschelrath, K., Kauf, B. & Niebel, T. (2015), 'The Economic Impacts of Broadband Internet: A Survey', *Review of Network Economics* **14**(4), 201–227.
URL: *https://doi.org/10.1515/rne-2016-0032*

Bertschek, I. & Kesler, R. (2022), 'Let the user speak: Is feedback on Facebook a source of firms' innovation?', *Information Economics and Policy* **60**, 100991.
URL: *https://doi.org/10.1016/j.infoecopol.2022.100991*

Bertschek, I. & Niebel, T. (2016), 'Mobile and more productive? Firm-level evidence on the productivity effects of mobile internet use', *Telecommunications Policy*

40(9), 888–898.
URL: *https://doi.org/10.1016/j.telpol.2016.05.007*

Bertschek, I., Polder, M. & Schulte, P. (2019), 'ICT and resilience in times of crisis: Evidence from cross-country micro moments data', *Economics of Innovation and New Technology* **28**(8), 759–774.
URL: *https://doi.org/10.1080/10438599.2018.1557417*

Billon, M., Lera-Lopez, F. & Marco, R. (2010), 'Differences in digitalization levels: a multivariate analysis studying the global digital divide', *Review of World Economics* **146**(1), 39–73.
URL: *https://doi.org/10.1007/s10290-009-0045-y*

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* **3**(Jan), 993–1022.
URL: *https://dl.acm.org/doi/10.5555/944919.944937*

Bloom, N., Sadun, R. & Van Reenen, J. (2012), 'Americans Do IT Better: US Multinationals and the Productivity Miracle', *American Economic Review* **102**(1), 167–201.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.102.1.167*

Bondy, J. A. & Murty, U. S. R. (1976), *Graph Theory With Applications*, Elsevier Science Ltd/North-Holland, ISBN: 0444194517.

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.
URL: *https://doi.org/10.1023/A:1010933404324*

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees (1st Edition)*, Taylor & Francis, Routledge, New York, ISBN: 9781315139470.
URL: *https://doi.org/10.1201/9781315139470*

Breithaupt, P., Kesler, R., Niebel, T. & Rammer, C. (2020), Intangible Capital Indicators Based on Web Scraping of Social Media, ZEW Discussion Paper No. 20-046, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *http://ftp.zew.de/pub/zew-docs/dp/dp20046.pdf*

Bresnahan, T. F. & Trajtenberg, M. (1995), 'General purpose technologies 'Engines of growth'?', *Journal of Econometrics* **65**(1), 83–108.
URL: *https://doi.org/10.1016/0304-4076(94)01598-T*

Bruhn, M., Schoenmueller, V. & Schäfer, D. B. (2012), 'Are social media replacing traditional media in terms of brand equity creation?', *Management Research Review*

35(9), 770–790.
URL: *https://doi.org/10.1108/01409171211255948*

Brynjolfsson, E. & Hitt, L. (1995), 'Information technology as a factor of production: The role of differences among firms', *Economics of Innovation and New Technology* **3**(3-4), 183–200.
URL: *https://doi.org/10.1080/10438599500000002*

Brynjolfsson, E. & Hitt, L. M. (1998), 'Beyond the Productivity Paradox', *Communications of the Association for Computing Machinery (ACM)* **41**(8), 49–55.
URL: *https://doi.org/10.1145/280324.280332*

Brynjolfsson, E. & Hitt, L. M. (2003), 'Computing productivity: Firm-level evidence', *The Review of Economics and Statistics* **85**(4), 793–808.
URL: *https://doi.org/10.1162/003465303772815736*

Brynjolfsson, E., Hitt, L. M. & Yang, S. (2002), 'Intangible Assets: Computers and Organizational Capital', *Brookings Papers on Economic Activity* **2002**(1), 137–198.
URL: *https://www.brookings.edu/wp-content/uploads/2002/01/2002a_bpea_brynjolfsson.pdf*

Büchel, J., Demary, V., Goecke, H., Rusche, C., Burstedde, A., Engels, B., Kohlisch, E., Koppel, O., Mertens, A., Scheufen, M., Wendt, J., Ewald, J., Hünnemeyer, V., Kempermann, H., Lichtblau, K., Schmitz, E., Bertschek, I., Niebel, T., Rammer, C., Schuck, B., Birtel, F., Harland, T., Hicking, J. & Wenger, L. (2020), Digitalisierungsindex 2020 - Langfassung Ergebnispapier (Digitalisierung der Wirtschaft in Deutschland), Technical report, Bundesministerium für Wirtschaft und Energie (BMWi), Berlin. [Online; last accessed 12.01.2024].
URL: *https://www.de.digital/DIGITAL/Redaktion/DE/Digitalisierungsindex/Publikationen/publikation-download-Langfassung-digitalisierungsindex-2020.pdf*

Cardona, M., Kretschmer, T. & Strobel, T. (2013), 'ICT and productivity: conclusions from the empirical literature', *Information Economics and Policy* **25**(3), 109–125.
URL: *https://doi.org/10.1016/j.infoecopol.2012.12.002*

Cassiman, B. & Golovko, E. (2011), 'Innovation and internationalization through exports', *Journal of International Business Studies* **42**(1), 56–75.
URL: *https://doi.org/10.1057/jibs.2010.36*

Chen, W. (2018), 'Cross-Country Income Differences Revisited: Accounting for the Role of Intangible Capital', *Review of Income and Wealth* **64**(3), 626–648.
URL: *https://doi.org/10.1111/roiw.12305*

Chen, W., Niebel, T. & Saam, M. (2016), 'Are intangibles more productive in ICT-intensive industries? Evidence from EU countries', *Telecommunications Policy* **40**(5), 471–484.
URL: *https://doi.org/10.1016/j.telpol.2015.09.010*

Chiang, J. K.-H. & Suen, H.-Y. (2015), 'Self-presentation and hiring recommendations in online communities: Lessons from LinkedIn', *Computers in Human Behavior* **48**, 516–524.
URL: *https://doi.org/10.1016/j.chb.2015.02.017*

Choi, H. & Varian, H. (2012), 'Predicting the Present with Google Trends', *Economic Record* **88**(s1), 2–9.
URL: *https://doi.org/10.1111/j.1475-4932.2012.00809.x*

Chollet, F. et al. (2015), 'Keras'. GitHub code repository [Online; last accessed 12.01.2024].
URL: *https://github.com/fchollet/keras*

Chung, S., Animesh, A., Han, K. & Pinsonneault, A. (2015), The Business Value of Firms' Social Media Efforts: Evidence from Facebook, *in* 'Proceedings of the 17th International Conference on Electronic Commerce 2015', pp. 1–8. [Online; last accessed 12.01.2024].
URL: *https://doi.org/10.1145/2781562.2781604*

Claussen, J. & Peukert, C. (2019), Obtaining Data from the Internet: A Guide to Data Crawling in Management Research, Technical report. Available at SSRN [Preprint (Online); last accessed 12.01.2024].
URL: *https://ssrn.com/abstract=3403799*

Collet, F. & Hedström, P. (2013), 'Old friends and new acquaintances: Tie formation mechanisms in an interorganizational network generated by employee mobility', *Social Networks* **35**(3), 288–299.
URL: *https://doi.org/10.1016/j.socnet.2013.02.005*

Conz, E. & Magnani, G. (2020), 'A dynamic perspective on the resilience of firms: A systematic literature review and a framework for future research', *European Management Journal* **38**(3), 400–412.
URL: *https://doi.org/10.1016/j.emj.2019.12.004*

Corrado, C., Haskel, J. & Jona-Lasinio, C. (2017), 'Knowledge Spillovers, ICT and Productivity Growth', *Oxford Bulletin of Economics and Statistics* **79**(4), 592–618.
URL: *https://doi.org/10.1111/obes.12171*

Corrado, C., Haskel, J., Jona-Lasinio, C. & Iommi, M. (2013), 'Innovation and intangible investment in Europe, Japan, and the United States', *Oxford Review of Economic Policy* **29**(2), 261–286.
URL: *https://doi.org/10.1093/oxrep/grt017*

Corrado, C., Haskel, J., Jona-Lasinio, C. & Iommi, M. (2022), 'Intangible Capital and Modern Economies', *Journal of Economic Perspectives* **36**(3), 3–28.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.36.3.3*

Corrado, C., Hulten, C. & Sichel, D. (2005), Measuring Capital and Technology: An Expanded Framework, *in* C. Corrado, J. Haltiwanger & D. Sichel, eds, 'Measuring Capital in the New Economy', University of Chicago Press, pp. 11–46.
URL: *https://www.nber.org/system/files/chapters/c0202/c0202.pdf*

Corrado, C., Hulten, C. & Sichel, D. (2009), 'Intangible Capital and US Economic Growth', *Review of Income and Wealth* **55**(3), 661–685.
URL: *https://doi.org/10.1111/j.1475-4991.2009.00343.x*

Coursaris, C. K., van Osch, W. & Balogh, B. A. (2016), Do Facebook Likes Lead to Shares or Sales? Exploring the Empirical Links between Social Media Content, Brand Equity, Purchase Intention, and Engagement, *in* 'Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)', IEEE, pp. 3546–3555. GitHub code repository [Online; last accessed 12.01.2024].
URL: *http://dx.doi.org/10.1109/HICSS.2016.444*

Crass, D., Licht, G. & Peters, B. (2014), Intangible Assets and Investments at the Sector Level: Empirical Evidence for Germany, *in* A. Bounfour & T. Miyagawa, eds, 'Intangibles, Market Failure and Innovation Performance', Springer International Publishing Switzerland.
URL: *https://doi.org/10.1007/978-3-319-07533-4_4*

Crepon, B., Duguet, E. & Mairessec, J. (1998), 'Research, Innovation And Productivity: An Econometric Analysis at The Firm Level', *Economics of Innovation and New Technology* **7**(2), 115–158.
URL: *https://doi.org/10.1080/10438599800000031*

Datar, M., Gionis, A., Indyk, P. & Motwani, R. (2002), 'Maintaining Stream Statistics over Sliding Windows', *SIAM Journal on Computing* **31**(6), 1794–1813.
URL: *https://doi.org/10.1137/S0097539701398363*

Davis, D. R., Dingel, J. I., Monras, J. & Morales, E. (2019), 'How Segregated Is Urban Consumption?', *Journal of Political Economy* **127**(4), 1684–1738.
URL: *https://doi.org/10.1086/701680*

Dhyne, E., Konings, J., Van den bosch, J. & Vanormelingen, S. (2021), 'The Return on Information Technology: Who Benefits Most?', *Information Systems Research* **32**(1), 194–211.
URL: *https://doi.org/10.1287/isre.2020.0960*

Di Ubaldo, M. & Siedschlag, I. (2020), 'Investment in Knowledge-Based Capital and Productivity: Firm-Level Evidence from a Small Open Economy', *Review of Income and Wealth* **67**(2), 363–393.
URL: *https://doi.org/10.1111/roiw.12464*

Diestel, R. (2017), *Graph Theory (5th ed.)*, Electronic Edition, Springer Berlin, Heidelberg, ISBN: 978-3-662-53622-3.
URL: *https://doi.org/10.1007/978-3-662-53622-3*

Doherr, T. (2023), The SearchEngine: A Holistic Approach to Matching, ZEW Discussion Paper No. 23-001, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online), last accessed 12.01.2024].
URL: *http://ftp.zew.de/pub/zew-docs/dp/dp23001.pdf*

Dong, L., Ratti, C. & Zheng, S. (2019), 'Predicting neighborhoods' socioeconomic attributes using restaurant data', *Proceedings of the National Academy of Sciences* **116**(31), 15447–15452.
URL: *https://doi.org/10.1073/pnas.1903064116*

Dörr, J. O., Licht, G. & Murmann, S. (2021), 'Small firms and the COVID-19 insolvency gap', *Small Business Economics* **58**(2), 887–917.
URL: *https://doi.org/10.1007/s11187-021-00514-4*

Dörr, J. O., Kinne, J., Lenz, D., Licht, G. & Winker, P. (2022), 'An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers', *PLOS ONE* **17**(2), e0263898.
URL: *https://doi.org/10.1371/journal.pone.0263898*

Edler, J., Fier, H. & Grimpe, C. (2011), 'International scientist mobility and the locus of knowledge and technology transfer', *Research Policy* **40**(6), 791–805.
URL: *https://doi.org/10.1016/j.respol.2011.03.003*

Elgazzar, Y., El-Shahawy, R. & Senousy, Y. (2022), The Role of Digital Transformation in Enhancing Business Resilience with Pandemic of COVID-19, *in* 'In: Magdi, D.A., Helmy, Y.K., Mamdouh, M., Joshi, A. (eds) Digital Transformation Technology. Lecture Notes in Networks and Systems, volume 224', Springer, Singapore, pp. 323–333.
URL: *https://doi.org/10.1007/978-981-16-2275-5_20*

Engelberg, J. E. & Parsons, C. A. (2011), 'The Causal Impact of Media in Financial Markets', *The Journal of Finance* **66**(1), 67–97.
URL: *https://doi.org/10.1111/j.1540-6261.2010.01626.x*

European Commission (2014), Flash Eurobarometer 369 (Investing in Intangibles: Economic Assets and Innovation Drivers for Growth), Technical report, GESIS Datenarchiv, Köln. [Online; last accessed 12.01.2024].
URL: *https://doi.org/10.4232/1.11908*

Fawcett, T. (2004), 'ROC Graphs: Notes and Practical Considerations for Researchers', *Pattern Recognition Letters* **31**(8), 1–38.
URL: *https://www.researchgate.net/publication/228404321_ROC_Graphs_Notes_and_Practical_Considerations_for_Researchers*

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern Recognition Letters* **27**(8), 861–874.
URL: *https://doi.org/10.1016/j.patrec.2005.10.010*

Filippova, E. (2019), Empirical Evidence and Economic Implications of Blockchain as a General Purpose Technology, *in* '2019 IEEE Technology & Engineering Management Conference (TEMSCON)', IEEE, pp. 1–8. [Online; last accessed 12.01.2024].
URL: *https://doi.org/10.1109/TEMSCON.2019.8813748*

Flesch, R. (1948), 'A new readability yardstick.', *Journal of Applied Psychology* **32**(3), 221–233.
URL: *https://psycnet.apa.org/doi/10.1037/h0057532*

Forman, C., Goldfarb, A. & Greenstein, S. (2012), 'The Internet and Local Wages: A Puzzle', *American Economic Review* **102**(1), 556–75.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.102.1.556*

Franzoni, C., Scellato, G. & Stephan, P. (2014), 'The mover's advantage: The superior performance of migrant scientists', *Economics Letters* **122**(1), 89–93.
URL: *https://doi.org/10.1016/j.econlet.2013.10.040*

Freeman, L. C. (1978), 'Centrality in Social Networks Conceptual Clarification', *Social Networks* **1**(3), 215–239.
URL: *https://doi.org/10.1016/0378-8733(78)90021-7*

Frenz, M. & Ietto-Gillies, G. (2009), 'The impact on innovation performance of different sources of knowledge: Evidence from the UK Community Innovation Survey', *Research Policy* **38**(7), 1125–1135.
URL: *https://doi.org/10.1016/j.respol.2009.05.002*

Friedman, J. H. (2002), 'Stochastic gradient boosting', *Computational Statistics & Data Analysis* **38**(4), 367–378.
URL: *https://www.sciencedirect.com/science/article/pii/S0167947301000652*

Gandin, I. & Cozza, C. (2019), 'Can we predict firms' innovativeness? The identification of innovation performers in an Italian region through a supervised learning approach', *PLOS ONE* **14**(6), e0218175.
URL: *https://doi.org/10.1371/journal.pone.0218175*

Gentzkow, M., Kelly, B. & Taddy, M. (2019), 'Text as Data', *Journal of Economic Literature* **57**(3), 535–574.
URL: *https://www.aeaweb.org/articles?id=10.1257/jel.20181020*

Gerhards, C., Liebig, S. & Elsner, J. (2010), Datenhandbuch: Projekt "Verknüpfte Personen-Betriebsdaten im Anschluss an den ALLBUS 2008" - ALLBUS-Betriebsbefragung 2009, DSZ-BO Technical Report 1, University of Bielefeld, Bielefeld. [Preprint (Online); last accessed 12.01.2024].
URL: *https://portal.fdz-bo.diw.de/sites/default/files/DSZ-BO-Technical-Report_Nr01_0.pdf*

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009), 'Detecting influenza epidemics using search engine query data', *Nature* **457**(7232), 1012–1014.
URL: *https://doi.org/10.1038/nature07634*

Giuliani, E. (2011), Networks of innovation, *in* 'Cooke P., Asheim B., Boschma R., Martin R., Schwartz D., Tödtling F. (ed.), Handbook of Regional Innovation and Growth, chapter 12', Edward Elgar Publishing.
URL: *https://ideas.repec.org/h/elg/eechap/13482_12.html*

Glaeser, E. L., Kim, H. & Luca, M. (2018), 'Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change', *AEA Papers and Proceedings* **108**, 77–82. [Online; last accessed 12.01.2024].
URL: *https://www.aeaweb.org/articles?id=10.1257/pandp.20181034*

Godart, F. C., Shipilov, A. V. & Claes, K. (2014), 'Making the Most of the Revolving Door: The Impact of Outward Personnel Mobility Networks on Organizational Creativity', *Organization Science* **25**(2), 377–400.
URL: *https://doi.org/10.1287/orsc.2013.0839*

Gök, A., Waterworth, A. & Shapira, P. (2015), 'Use of web mining in studying innovation', *Scientometrics* **102**(1), 653–671.
URL: *https://doi.org/10.1007/s11192-014-1434-0*

Goldfarb, A., Taska, B. & Teodoridis, F. (2023), 'Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings', *Research Policy* **52**(1), 104653.
URL: *https://doi.org/10.1016/j.respol.2022.104653*

Goldfarb, A. & Tucker, C. (2019), 'Digital Economics', *Journal of Economic Literature* **57**(1), 3–43.
URL: *https://www.aeaweb.org/articles?id=10.1257/jel.20171452*

Görg, H. & Strobl, E. (2005), 'Spillovers from Foreign Firms through Worker Mobility: An Empirical Investigation', *The Scandinavian Journal of Economics* **107**(4), 693–709.
URL: *https://www.jstor.org/stable/34410121*

Gortmaker, J., Jeffers, J. & Lee, M. (2022), Labor Reactions to Credit Deterioration: Evidence from LinkedIn Activity, Technical report, Available at SSRN. [Preprint (Online); last accessed 12.01.2024].
URL: *https://dx.doi.org/10.2139/ssrn.3456285*

Graham, S. J. & Iacopetta, M. (2014), 'Nanotechnology and the Emergence of a General Purpose Technology', *Annals of Economics and Statistics/Annales d'Économie et de Statistique* **115/116**, 25–55.
URL: *https://doi.org/10.15609/annaeconstat2009.115-116.25*

Griffith, R., Huergo, E., Mairesse, J. & Peters, B. (2006), 'Innovation and Productivity Across Four European Countries', *Oxford Review of Economic Policy* **22**(4), 483–498.
URL: *https://doi.org/10.1093/oxrep/grj028*

Groseclose, T. & Milyo, J. (2005), 'A measure of media bias', *The Quarterly Journal of Economics* **120**(4), 1191–1237.
URL: *https://doi.org/10.1162/003355305775097542*

Grupp, H. & Schubert, T. (2010), 'Review and new evidence on composite innovation indicators for evaluating national performance', *Research Policy* **39**(1), 67–78.
URL: *https://doi.org/10.1016/j.respol.2009.10.002*

Hall, B. H., Jaffe, A. & Trajtenberg, M. (2005), 'Market Value and Patent Citations', *RAND Journal of Economics* **36**(1), 16–38.
URL: *https://www.jstor.org/stable/1593752*

Hall, B. H., Lotti, F. & Mairesse, J. (2013), 'Evidence on the impact of R&D and ICT investments on innovation and productivity in Italian firms', *Economics of Innovation and New Technology* **22**(3), 300–328.
URL: *https://doi.org/10.1080/10438599.2012.708134*

Hall, B. H. & Trajtenberg, M. (2006), Uncovering general purpose technologies with patent data, *in* 'in: Cristiano Antonelli & Dominique Foray & Bronwyn H. Hall & W. Edward Steinmueller (ed.), New Frontiers in the Economics of Innovation and New Technology, chapter 14', Edward Elgar Publishing, pp. 389–426.
URL: *https://ideas.repec.org/h/elg/eechap/3286_14.html*

Haskel, J. & Westlake, S. (2018), *Capitalism without Capital: The Rise of the Intangible Economy*, Princeton University Press, ISBN: 9780691175034.
URL: *https://press.princeton.edu/books/hardcover/9780691175034/capitalism-without-capital*

Hastie, T., Friedman, J. & Tibshirani, R. (2001), *The Elements of Statistical Learning (first edition)*, Springer Series in Statistics, New York, NJ, ISBN: 0387952845.

Heining, J., Klosterhuber, W., Lehnert, P. & Seth, S. (2016), Linked-Employer-Employee-Daten des IAB: LIAB-Längsschnittmodell 1993 – 2014 (LIAB LM 9314), FDZ Datenreport 10/2016, Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg. [Online; last accessed 12.01.2024].
URL: *https://doku.iab.de/fdz/reporte/2016/DR_10-16.pdf*

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.
URL: *http://dx.doi.org/10.1162/neco.1997.9.8.1735*

Hong, S., Oxley, L. & McCann, P. (2012), 'A survey of the innovation surveys', *Journal of Economic Surveys* **26**(3), 420–444.
URL: *https://doi.org/10.1111/j.1467-6419.2012.00724.x*

Hosmer, Jr., D. W., Lemeshow, S. & Sturdivant, R. X. (2013), *Applied Logistic Regression (third edition)*, Wiley Series in Probability and Statistics, John Wiley & Sons.
URL: *https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387*

Hottenrott, H. & Lopes-Bento, C. (2016), 'R&D Partnerships and Innovation Performance: Can There Be too Much of a Good Thing?', *Journal of Product Innovation Management* **33**(6), 773–794.
URL: *https://doi.org/10.1111/jpim.12311*

Hyndman, R. J. & Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**(4), 679–688.
URL: *https://www.sciencedirect.com/science/article/abs/pii/S0169207006002399*

Jaffe, A. B., Newell, R. G. & Stavins, R. N. (2002), 'Environmental policy and technological change', *Environmental and resource economics* **22**, 41–70.
URL: *https://link.springer.com/content/pdf/10.1023/A:1015519401088.pdf*

Janger, J., Schubert, T., Andries, P., Rammer, C. & Hoskens, M. (2017), 'The EU 2020 innovation indicator: A step forward in measuring innovation outputs and outcomes?', *Research Policy* **46**(1), 30–42.
URL: *https://doi.org/10.1016/j.respol.2016.10.001*

Jensen, P. H. (2010), 'Exploring the Uses of Matched Employer–Employee Datasets', *Australian Economic Review* **43**(2), 209–216.
URL: *https://doi.org/10.1111/j.1467-8462.2010.00594.x*

Ji, Y., Rozenbaum, O. & Welch, K. (2022), Corporate Culture and Financial Reporting Risk: Looking Through the Glassdoor, Available at SSRN. [Preprint (Online); last accessed 12.01.2024].
URL: *https://ssrn.com/abstract=2945745*

Johannessen, J.-A. (2008), 'Organisational innovation as part of knowledge management', *International Journal of Information Management* **28**(5), 403–412.
URL: *https://doi.org/10.1016/j.ijinfomgt.2008.04.007*

Jovanovic, B. & Rousseau, P. L. (2005), General purpose technologies, *in* 'Aghion P., Durlauf S. N., Handbook of Economic Growth', Vol. 1, Part B, Elsevier, pp. 1181–1224.
URL: *https://doi.org/10.1016/S1574-0684(05)01018-X*

Kaiser, U., Kongsted, H. C. & Rønde, T. (2015), 'Does the mobility of R&D labor increase innovation?', *Journal of Economic Behavior & Organization* **110**, 91–105.
URL: *https://doi.org/10.1016/j.jebo.2014.12.012*

Kane, E. (2017), Is Blockchain a General Purpose Technology?, SSRN paper 2932585. [Preprint (Online); last accessed 12.01.2024].
URL: *https://dx.doi.org/10.2139/ssrn.2932585*

Katz, J. S. & Cothey, V. (2006), 'Web indicators for complex innovation systems', *Research Evaluation* **15**(2), 85–95.
URL: *https://doi.org/10.3152/147154406781776922*

Katz, R. L. & Koutroumpis, P. (2013), 'Measuring digitization: A growth and welfare multiplier', *Technovation* **33**(10-11), 314–319.
URL: *https://doi.org/10.1016/j.technovation.2013.06.004*

Kelly, B., Papanikolaou, D., Seru, A. & Taddy, M. (2021), 'Measuring Technological Innovation over the Long Run', *American Economic Review: Insights* **3**(3), 303–320.
URL: *https://www.aeaweb.org/articles?id=10.1257/aeri.20190499*

Khurana, D., Koli, A., Khatter, K. & Singh, S. (2023), 'Natural language processing: state of the art, current trends and challenges', *Multimedia Tools and Applications* **82**(3), 3713–3744.
URL: *https://doi.org/10.1007/s11042-022-13428-4*

Kinne, J. (2018), 'ARGUS - An Automated Robot for Generic Universal Scraping'. GitHub code repository [Online; last accessed 12.01.2024].
URL: *https://github.com/datawizard1337/ARGUS*

Kinne, J. & Axenbeck, J. (2020), 'Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study', *Scientometrics* **125**, 2011–2041.
URL: *https://doi.org/10.1007/s11192-020-03726-9*

Kinne, J. & Lenz, D. (2021), 'Predicting innovative firms using web mining and deep learning', *PLOS ONE* **16**(4), e0249071.
URL: *https://doi.org/10.1371/journal.pone.0249071*

Kirbach, M. & Schmiedeberg, C. (2008), 'Innovation and export performance: Adjustment and remaining differences in East and West German manufacturing', *Economics of Innovation and New Technology* **17**(5), 435–457.
URL: *https://doi.org/10.1080/10438590701357189*

Klomp, L. & Van Leeuwen, G. (2001), 'Linking Innovation and Firm Performance: A New Approach', *International Journal of the Economics of Business* **8**(3), 343–364.
URL: *https://doi.org/10.1080/13571510110079612*

Kogan, L., Papanikolaou, D., Seru, A. & Stoffman, N. (2017), 'Technological Innovation, Resource Allocation, and Growth', *Quarterly Journal of Economics* **132**(2), 665–712.
URL: *https://doi.org/10.1093/qje/qjw040*

Kotarba, M. (2017), 'Measuring digitalization: Key metrics', *Foundations of Management* **9**(1), 123–138.
URL: *https://doi.org/10.1515/fman-2017-0010*

Kotsiantis, S. (2007), 'Supervised Machine Learning: A Review of Classification Techniques', *Informatica: An International Journal of Computing and Informatics* **31**(3), 249–268.
URL: *https://www.informatica.si/index.php/informatica/article/view/148/140*

Krüger, M., Kinne, J., Lenz, D. & Resch, B. (2020), The digital layer: How innovative firms relate on the web, ZEW Discussion Paper No. 20-003, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *https://www.econstor.eu/bitstream/10419/213105/1/1688207651.pdf*

Lachenmaier, S. & Wößmann, L. (2006), 'Does innovation cause exports? Evidence from exogenous innovation impulses and obstacles using German micro data', *Oxford Economic Papers* **58**(2), 317–350.
URL: *https://doi.org/10.1093/oep/gpi043*

Larsen, V. H. & Thorsrud, L. A. (2019), 'The value of news for economic developments', *Journal of Econometrics* **210**(1), 203–218.
URL: *https://doi.org/10.1016/j.jeconom.2018.11.013*

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.
URL: *https://doi.org/10.1038/nature14539*

Lenz, D. & Winker, P. (2020), 'Measuring the diffusion of innovations with paragraph vector topic models', *PLOS ONE* **15**(1), e0226685.
URL: *https://doi.org/10.1371/journal.pone.0226685*

Li, J., Sun, A., Han, J. & Li, C. (2020), 'A Survey on Deep Learning for Named Entity Recognition', *IEEE Transactions on Knowledge and Data Engineering* **34**(1), 50–70. [Online; last accessed 12.01.2024].
URL: *https://ieeexplore.ieee.org/abstract/document/9039685*

Lima, E., Sun, X., Dong, J., Wang, H., Yang, Y. & Liu, L. (2017), 'Learning and Transferring Convolutional Neural Network Knowledge to Ocean Front Recognition', *IEEE Geoscience and Remote Sensing Letters* **14**(3), 354–358.
URL: *https://doi.org/10.1109/LGRS.2016.2643000*

Loper, E. & Bird, S. (2002), NLTK: The Natural Language Toolkit, *in* 'ETMTNLP '02: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1', pp. 63–70. [Preprint (Online); last accessed 12.01.2024].
URL: *https://dl.acm.org/doi/10.3115/1118108.1118117*

Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. (2013), Understanding variable importances in forests of randomized trees, *in* 'Advances in Neural Information Processing Systems 26 (NIPS 2013)', pp. 431–439. [Online; last accessed 12.01.2024].
URL: *https://proceedings.neurips.cc/paper/2013/file/e3796ae838835da0b6f6ea37bcf8bcb7-Paper.pdf*

Lucas, R. (1988), 'On the Mechanics of Economic Development', *Journal of Monetary Economics* **22**(1), 3–42.
URL: *https://doi.org/10.1016/0304-3932(88)90168-7*

Luo, X., Zhang, J. & Duan, W. (2012), 'Social Media and Firm Equity Value', *Information Systems Research* **24**(1), 146–163.
URL: *https://doi.org/10.1287/isre.1120.0462*

Mairesse, J. & Mohnen, P. (2010), Using Innovation Surveys for Econometric Analysis, *in* B. H. Hall & N. Rosenberg, eds, 'Handbook of the Economics of Innovation', Vol. 2, North-Holland, Amsterdam, chapter 26, pp. 1129–1155.
URL: *https://doi.org/10.1016/S0169-7218(10)02010-1*

Maliranta, M., Mohnen, P. & Rouvinen, P. (2009), 'Is inter-firm labor mobility a channel of knowledge spillovers? Evidence from a linked employer–employee panel', *Industrial and Corporate Change* **18**(6), 1161–1191.
URL: *https://doi.org/10.1093/icc/dtp031*

Manning, C. & Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT press, Cambridge, MA, ISBN: 0262133601.

McNemar, Q. (1947), 'Note on the sampling error of the difference between correlated proportions or percentages', *Psychometrika* **12**(2), 153–157.
URL: *https://doi.org/10.1007/BF02295996*

Misirlis, N. & Vlachopoulou, M. (2018), 'Social media metrics and analytics in marketing – S3M: A mapping literature review', *International Journal of Information Management* **38**(1), 270–276.
URL: *https://doi.org/10.1016/j.ijinfomgt.2017.10.005*

Mohnen, P. (2019), R&D, Innovation and Productivity, *in* T. t. Raa & W. H. Greene, eds, 'The Palgrave Handbook of Economic Performance Analysis (first edition)', Palgrave Macmillan Cham, Amsterdam, pp. 97–122.
URL: *https://doi.org/10.1007/978-3-030-23727-1*

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018), *Foundations of Machine Learning (second edition)*, MIT Press, Cambridge, MA, ISBN: 9780262039406.

URL: *https://mitpress.mit.edu/9780262039406/foundations-of-machine-learning/*

Nadeau, D. & Sekine, S. (2007), 'A survey of named entity recognition and classification', *Lingvisticae Investigationes* **30**(1), 3–26.
URL: *https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad*

Nathan, M. & Rosso, A. (2022), 'Innovative events: product launches, innovation and firm performance', *Research Policy* **51**(1), 104373.
URL: *https://doi.org/10.1016/j.respol.2021.104373*

Niebel, T., O'Mahony, M. & Saam, M. (2017), 'The Contribution of Intangible Assets to Sectoral Productivity Growth in the EU', *Review of Income and Wealth* **63**, S49–S67.
URL: *https://doi.org/10.1111/roiw.12248*

Nieminen, U. J. (1973), 'On the Centrality in a Directed Graph', *Social Science Research* **2**(4), 371–378.
URL: *https://doi.org/10.1016/0049-089X(73)90010-0*

OECD/Eurostat (2018), Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation (4th Edition), The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg. [Online; last accessed 12.01.2024].
URL: *https://www.oecd-ilibrary.org/content/publication/9789264304604-en*

Ozman, M. (2009), 'Inter-firm networks and innovation: a survey of literature', *Economic of Innovation and New Technology* **18**(1), 39–67.
URL: *https://doi.org/10.1080/10438590701660095*

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999), PageRank: Bringing order to the web, Technical report, Stanford Digital Libraries Working Paper. [Preprint (Online); last accessed 12.01.2024 (webpage is currently not available)].
URL: *http://ilpubs.stanford.edu/422/1/1999-66.pdf*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. (2011), 'Scikit-Learn: Machine Learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.
URL: *https://arxiv.org/abs/1201.0490*

Perani, G. & Guerrazzi, M. (2012), The Statistical Measurement of Intangible Assets: Methodological Implications of the Results of the ISFOL 2011 Pilot Survey. Mimeo. Available upon request from the authors.

Peters, B. & Rammer, C. (2013), Innovation Panel Surveys in Germany, *in* F. Gault, ed., 'Handbook of Innovation Indicators and Measurement', Edward Elgar Publishing, pp. 135–177.
URL: *https://www.e-elgar.com/shop/gbp/handbook-of-innovation-indicators-and-measurement-9781782545170.html*

Petralia, S. (2020), 'Mapping general purpose technologies with patent data', *Research Policy* **49**(7), 104013.
URL: *https://doi.org/10.1016/j.respol.2020.104013*

Pisano, S., Lepore, L. & Lamboglia, R. (2017), 'Corporate disclosure of human capital via LinkedIn and ownership structure: An empirical analysis of European companies', *Journal of Intellectual Capital* **18**(1), 102–127.
URL: *https://doi.org/10.1108/JIC-01-2016-0016*

Powers, D. M. (2011), Evaluation: from precision, recall and F-measure to ROC, Informedness, Markedness and Correlation, Technical report. [Preprint (Online); last accessed 12.01.2024].
URL: *https://arxiv.org/ftp/arxiv/papers/2010/2010.16061.pdf*

Pukelis, L. & Stanciauskas, V. (2019), Using Internet Data to Compliment Traditional Innovation Indicators, *in* 'International Society of Scientometrics and Infometrics (ISSI) 2019'. [Preprint (Online); last accessed 12.01.2024].
URL: *https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf*

Rahal, C., Verhagen, M. & Kirk, D. (2022), 'The rise of machine learning in the academic social sciences', *AI & SOCIETY* pp. 1–3.
URL: *https://doi.org/10.1007/s00146-022-01540-w*

Rahko, J. (2017), 'Knowledge spillovers through inventor mobility: the effect on firm-level patenting', *The Journal of Technology Transfer* **42**(3), 585–614.
URL: *https://doi.org/10.1007/s10961-016-9494-3*

Rammer, C., Behrens, V., Doherr, T., Hud, M., Köhler, M., Krieger, B., Peters, B., Schubert, T., Trunschke, M. & von der Burg, J. (2019), Innovationen in der deutschen Wirtschaft: Indikatorenbericht zur Innovationserhebung 2018, Technical report, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *https://ftp.zew.de/pub/zew-docs/mip/18/mip_2018.pdf*

Rammer, C., Doherr, T., Krieger, B., Marks, H., Niggemann, H., Peters, B., Schubert, T., Trunschke, M. & von der Burg, J. (2021), Indikatorenbericht zur Innovationserhebung 2020, Technical Report, ZEW - Leibniz Centre for European Economic

Research, Mannheim. [Online; last accessed 12.01.2024].
URL: *https://ftp.zew.de/pub/zew-docs/mip/20/mip_2020.pdf*

Rammer, C. & Es-Sadki, N. (2023), 'Using big data for generating firm-level innovation indicators - a literature review', *Technological Forecasting and Social Change* **197**, 122874.
URL: *https://www.sciencedirect.com/science/article/pii/S0040162523005590*

Rammer, C., Roth, F. & Trunschke, M. (2020), Measuring Organisation Capital at the Firm Level: A Production Function Approach, ZEW Discussion Paper No. 20-021, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Preprint (Online); last accessed 12.01.2024].
URL: *http://ftp.zew.de/pub/zew-docs/dp/dp20021.pdf*

Rietsch, C., Beaudry, C. & Héroux-Vaillancourt, M. (2016), Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community, *in* 'CARMA 2016: 1st International Conference on Advanced Research Methods and Analytics', pp. 100–115.
URL: *http://dx.doi.org/10.4995/CARMA2016.2016.3140*

Romer, P. M. (1986), 'Increasing Returns and Long-Run Growth', *Journal of Political Economy* **94**(5), 1002–1037.
URL: *https://www.jstor.org/stable/1833190*

Romer, P. M. (1990), 'Endogenous Technological Change', *Journal of Political Economy* **98**(5, Part 2), 71–102.
URL: *https://www.jstor.org/stable/2937632*

Roth, F. (2019), Intangible Capital and Labor Productivity Growth: A Review of the Literature, Hamburg Discussion Papers in International Economics 4, University of Hamburg, Hamburg. [Preprint (Online); last accessed 12.01.2024].
URL: *https://www.econstor.eu/bitstream/10419/207163/1/hdpie-no04.pdf*

Roth, F. (2022), Intangible Capital and Labor Productivity Growth–Revisiting the Evidence: An Update, Hamburg Discussion Papers in International Economics 11, University of Hamburg. [Preprint (Online); last accessed 12.01.2024].
URL: *https://www.econstor.eu/handle/10419/253256*

Roth, F., Sen, A. & Rammer, C. (2023), 'The role of intangibles in firm-level productivity – evidence from Germany', *Industry and Innovation* **30**(2), 263–285.
URL: *https://doi.org/10.1080/13662716.2022.2138280*

Roth, F. & Thum, A.-E. (2013), 'Intangible Capital and Labor Productivity Growth: Panel Evidence for the EU from 1998-2005', *Review of Income and Wealth* **59**(3), 486–508.
URL: *https://doi.org/10.1111/roiw.12009*

Saad, M. H., Hagelaar, G., van der Velde, G. & Omta, S. W. F. (2021), 'Conceptualization of SMEs' business resilience: A systematic literature review', *Cogent Business & Management* **8**(1), 1938347.
URL: *https://doi.org/10.1080/23311975.2021.1938347*

Salton, G. & Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management* **24**(5), 513–523.
URL: *https://doi.org/10.1016/0306-4573(88)90021-0*

Schilling, M. A. & Phelps, C. C. (2007), 'Interfirm Collaboration Networks: The Impact of Large-Scale Network Structure on Firm Innovation', *Management Science* **53**(7), 1113–1126.
URL: *https://doi.org/10.1287/mnsc.1060.0624*

Schmucker, A., Seth, S. & Eberle, J. (2014), WeLL-Befragungsdaten verknüpft mit administrativen Daten des IAB, Datenreport 01/2014, FDZ der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- Berufsforschung, Nürberg. [Preprint (Online); last accessed 12.01.2024].
URL: *https://doku.iab.de/fdz/reporte/2014/DR_01-14.pdf*

Schultz, L. & Joutz, F. (2010), 'Methods for identifying emerging General Purpose Technologies: a case study of nanotechnologies', *Scientometrics* **85**(1), 155–170.
URL: *https://doi.org/10.1007/s11192-010-0244-2*

Schweikl, S. & Obermaier, R. (2020), 'Lessons from three decades of IT productivity research: towards a better understanding of IT-induced productivity effects', *Management Review Quarterly* **70**, 461–507.
URL: *https://doi.org/10.1007/s11301-019-00173-6*

Schwierzy, J., Dehghan, R., Schmidt, S., Rodepeter, E., Stoemmer, A., Uctum, K., Kinne, J., Lenz, D. & Hottenrott, H. (2022), Technology Mapping Using WebAI: The Case of 3D Printing, Technical report, Available at arXiv. [Preprint (Online); last accessed 12.01.2024].
URL: *https://arxiv.org/abs/2201.01125*

Scott, J. (2017), *Social Network Analysis (4th Edition)*, SAGE Publications Ltd. ISBN: 9781473952126.
URL: *https://doi.org/10.4135/9781529716597*

Solow, R. M. (1956), 'A contribution to the theory of economic growth', *The Quarterly Journal of Economics* **70**(1), 65–94.
URL: *https://doi.org/10.2307/1884513*

Solow, R. M. (1957), 'Technical change and the aggregate production function', *The Review of Economics and Statistics* **39**(3), 312–320.
URL: *https://doi.org/10.2307/1926047*

Somaya, D., Williamson, I. O. & Lorinkova, N. (2008), 'Gone but Not Lost: The Different Performance Impacts of Employee Mobility Between Cooperators Versus Competitors', *Academy of Management Journal* **51**(5), 936–953.
URL: *https://psycnet.apa.org/doi/10.5465/AMJ.2008.34789660*

Su, X., Huang, Z., Zhao, Y., Chen, Y., Dou, Y. & Pan, H. (2023), Recent Trends in Deep Learning Based Textual Emotion Cause Extraction, *in* 'IEEE/ACM Transactions on Audio, Speech, and Language Processing', Vol. 31, pp. 2765 – 2786.
URL: *https://doi.org/10.1109/TASLP.2023.3254166*

Tacchella, A., Napoletano, A. & Pietronero, L. (2020), 'The Language of Innovation', *PLOS ONE* **15**(4), e0230107.
URL: *https://doi.org/10.1371/journal.pone.0230107*

Tetlock, P. C. (2007), 'Giving Content to Investor Sentiment: The Role of Media in the Stock Market', *The Journal of Finance* **62**(3), 1139–1168.
URL: *https://doi.org/10.1111/j.1540-6261.2007.01232.x*

Tirunillai, S. & Tellis, G. J. (2012), 'Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance', *Marketing Science* **31**(2), 198–215.
URL: *https://doi.org/10.1287/mksc.1110.0682*

Torrey, L. & Shavlik, J. (2010), Transfer Learning, *in* E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito & A. J. S. López, eds, 'Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques', IGI Global, Hershey, pp. 242–264.
URL: *https://doi.org/10.4018/978-1-60566-766-9.ch011*

van der Wouden, F. & Rigby, D. L. (2021), 'Inventor mobility and productivity: A long-run perspective', *Industry and Innovation* **28**(6), 677–703.
URL: *https://doi.org/10.1080/13662716.2020.1789451*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), Attention Is All You Need, *in* '31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA'.
URL: *https://doi.org/10.48550/arXiv.1706.03762*

Veltri, G. A. (2013), 'Microblogging and nanotweets: Nanotechnology on Twitter', *Public Understanding of Science* **22**(7), 832–849.
URL: *https://doi.org/10.1177/0963662512463510*

Vos, J. (2009), 'Actions Speak Louder than Words: Greenwashing in Corporate America', *Notre Dame Journal of Law, Ethics & Public Policy* **23**(2), 673–697.
URL: *http://scholarship.law.nd.edu/ndjlepp/vol23/iss2/13*

Vu, K., Hanafizadeh, P. & Bohlin, E. (2020), 'ICT as a driver of economic growth: A survey of the literature and directions for future research', *Telecommunications Policy* **44**(2), 101922.
URL: *https://doi.org/10.1016/j.telpol.2020.101922*

Wasserman, S. & Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press, ISBN: 9780511815478.
URL: *https://doi.org/10.1017/CBO9780511815478*

Weinhardt, M. (2016), SOEP-LEE Betriebsbefragung - Datenhandbuch der Betriebsbefragung des Sozio-oekonomischen Panels, SOEP Survey Papers 306: Series D, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin. [Online; last accessed 12.01.2024].
URL: *https://www.econstor.eu/handle/10419/130173*

Weinhardt, M., Meyermann, A., Liebig, S. & Schupp, J. (2016), The Linked Employer–Employee Study of the Socio-Economic Panel (SOEP-LEE): Project Report, SOEPpapers on Multidisciplinary Panel Data Research 829-2016, Deutsches Institut für Wirtschaftsforschung, Berlin. [Preprint (Online); last accessed 12.01.2024].
URL: *https://www.diw.de/de/diw_01.c.530102.de/publikationen/soeppapers/2016_0829/the_linked_employer-employee_study_of_the_socio-economic_panel__soep-lee___project_report.html*

Weinhardt, M., Meyermann, A., Liebig, S. & Schupp, J. (2017), 'The Linked Employer–Employee Study of the Socio-Economic Panel (SOEP-LEE): Content, Design and Research Potential', *Jahrbücher für Nationalökonomie und Statistik* **237**(5), 457–467.
URL: *https://doi.org/10.1515/jbnst-2015-1044*

Wu, L., Jin, F. & Hitt, L. M. (2017), 'Are All Spillovers Created Equal? A Network Perspective on Information Technology Labor Movements', *Management Science* **64**(7), 3168–3186.
URL: *https://doi.org/10.1287/mnsc.2017.2778*

Xie, M., Jean, N., Burke, M., Lobell, D. & Ermon, S. (2016), Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping, *in* 'Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)', Elsevier. [Online; last accessed 12.01.2024].
URL: *https://ojs.aaai.org/index.php/AAAI/article/view/9906/9765*

Youtie, J., Iacopetta, M. & Graham, S. (2008), 'Assessing the nature of nanotechnology: can we uncover an emerging general purpose technology?', *The Journal of Technology Transfer* **33**, 315–329.
URL: *https://doi.org/10.1007/s10961-007-9030-6*

ZEW (2024*a*), Mannheim Innovation Panel - the Annual German Innovation Survey, Project website, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Online; last accessed 24.02.2024].
URL: *https://www.zew.de/WS109-1*

ZEW (2024*b*), Mannheim Innovation Panel: Innovation Activities of Enterprises in Germany , Project website, ZEW - Leibniz Centre for European Economic Research, Mannheim. [Online; last accessed 12.01.2024].
URL: *https://www.zew.de/en/research-at-zew/mannheim-innovation-panel-innovation-activities-of-enterprises-in-germany*

Zide, J., Elman, B. & Shahani-Denning, C. (2014), 'LinkedIn and recruitment: how profiles differ across occupations', *Employee Relations* **36**(5), 583–604.
URL: *https://doi.org/10.1108/ER-07-2013-0086*

Appendices

Appendix A

Innovation Indicators Based on Firm Websites – Which Website Characteristics Predict Firm-Level Innovation Activity?

A.1 Comparison of the Distributions Between the MIP and the Subsample

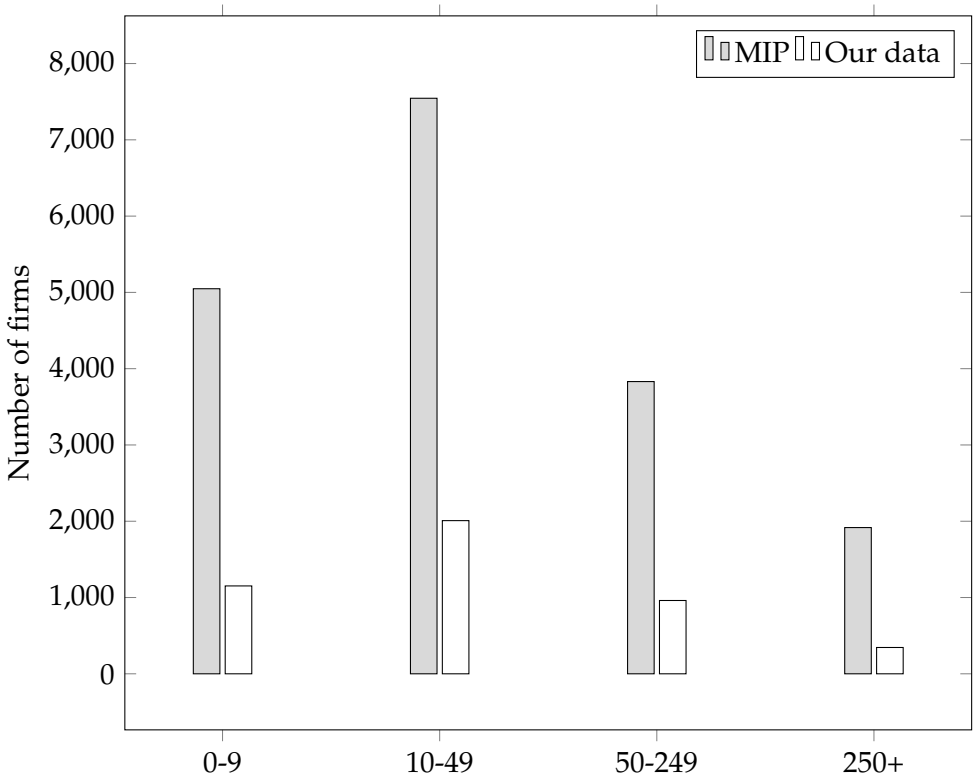


Figure A.1: **Firm distribution based on the number of employees.**
Own illustration.

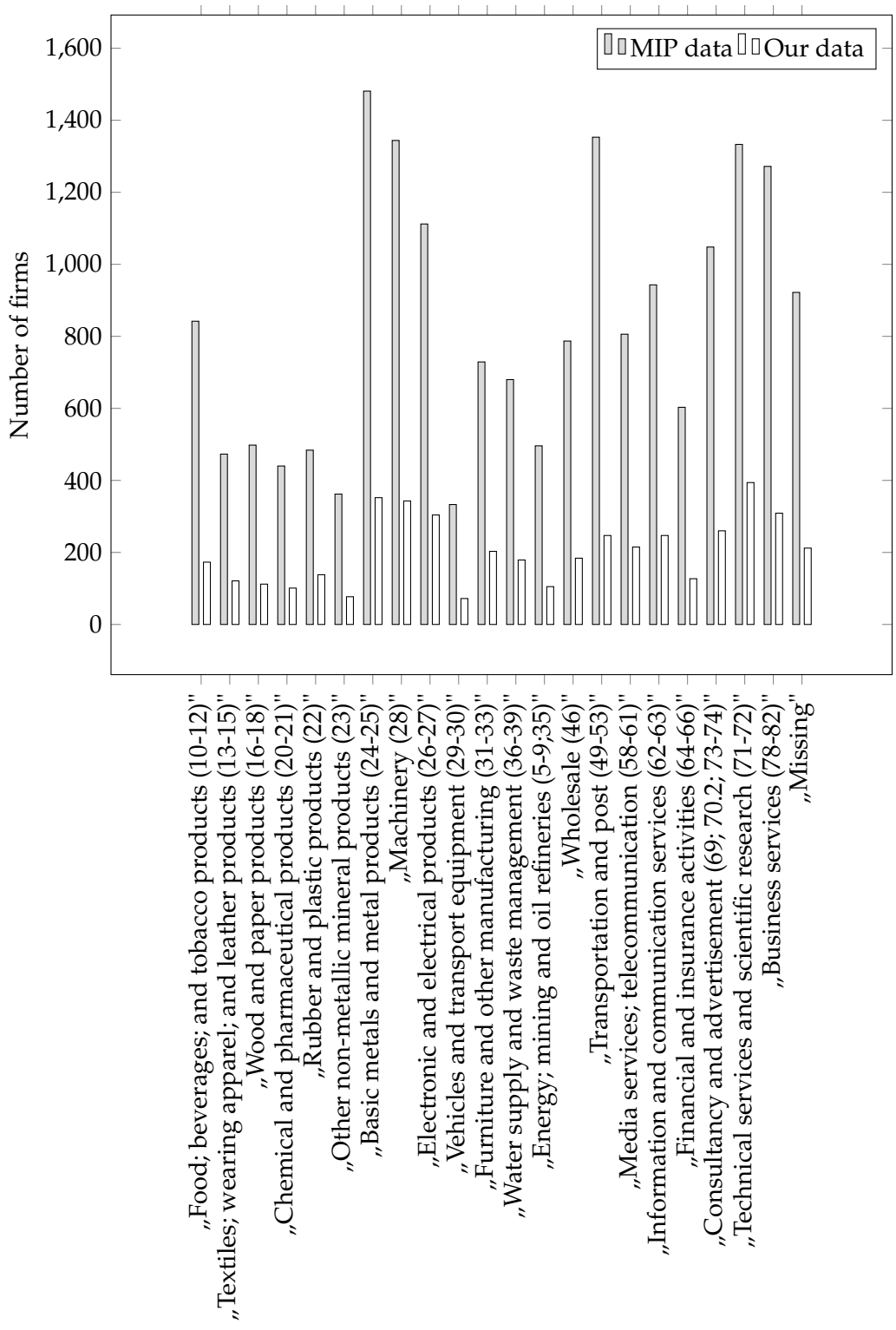


Figure A.2: **Firm distribution for industries based on 2-digit NACE codes.** Own illustration.

A.2 List of Emerging Technology Terms Used in the Keyword Search

English terms: agricultural robot, closed ecological systems, cultured meat, precision agriculture, vertical farming, micro air vehicle, neural-sensing headset, four-dimensional printing, arcology, aerogel, bioplastic, conductive polymers, cryogenic treatment, fullerene, graphene, lab-on-a-chip, magnetorheological fluid, metamaterials, metal foam, multi-function structures, nanomaterials, carbon nanotube, quantum dots, superalloy, synthetic diamond, translucent concrete, 3D displays, ferroelectric liquid crystal display, holography, interferometric modulator display, laser video displays, OLED displays, micro LED displays, telescopic pixel display, time-multiplexed optical shutter, volumetric display, biometrics, digital scent technology, electronic nose, e-textiles, flexible electronics, memristor, molecular electronics, nano electro mechanical systems, spintronics, thermal copper pillar bump, three-dimensional integrated circuit, concentrated solar power, electric double-layer capacitor, flywheel energy storage, grid energy storage, home fuel cell, lithium iron phosphor battery, lithium-sulfur battery, magnesium battery, nanowire battery, ocean thermal energy conversion, smart grid, vortex engine, wireless energy transfer, zero-energy building, computer-generated imagery, virtual reality, ultra-high-definition television, 5G cellular communications, artificial general intelligence, augmented reality, blockchain, carbon nanotube field-effect transistor, civic technology, cryptocurrency, exascale computing, gesture recognition, internet of things, emerging memory technologies, emerging magnetic data storage technologies, fourth generation optical discs, holographic data storage, general purpose computing on graphics processing units, exocortex, machine translation, machine vision, mobile collaboration, nano radio, optical computing, quantum computing, quantum cryptography, radio-frequency identification, semantic web, smart speaker, software-defined radio, speech recognition, subvocal recognition, hybrid forensics, body implants, prosthesis, cryonics, de-extinction, genetic engineering of organisms and viruses, suspended animation, artificial hibernation, immunotherapy/oncology, nano medicines, nano sensors, oncolytic viruses, personalized medicine, whole genome sequencing, robotic surgery, stem cell treatments, synthetic biology, synthetic genomics, tissue engineering, tricorder, brain-computer interface, neuro informatics, electro encephalography, neuro prosthetics, caseless ammunition, directed energy weapon, electro laser, electromagnetic weapons, electrothermal-chemical technology, green bullet, laser weapon, particle beam weapon, sonic weapon, stealth technology, vortex ring gun, wireless long-range electric shock weapon, artificial gravity, stasis chamber, inflatable space habitat, miniaturized satellite, android, gynoid, nanorobotics, powered exoskeleton,

self-reconfiguring modular robot, unmanned vehicle, airless tire, alternative fuel vehicle, electro hydrodynamic propulsion, flying car, fusion rocket, hoverbike, jetpack, backpack helicopter, maglev train, vactrain, magnetic levitation, mass driver, personal rapid transit, physical internet, scooter-sharing system, propellant depot, reusable launch system, space elevator, spaceplane, supersonic transport, vehicular communication systems.

German terms: Agrarroboter, geschlossenes ökologisches System, Zuchtfleisch, Präzisionslandwirtschaft, vertikale Landwirtschaft, Mikro-Luftfahrzeug, neuronales Headset, vierdimensionales Drucken, Arkologie, Aerogel, Bio-Kunststoff, leitfähige Polymere, kryogene Behandlung, Fulleren, Graphen, Labor auf einem Chip, magnetorheologische Flüssigkeit, Metamaterialien, Metallschaum, Multifunktionsstrukturen, Nanomaterialien, Kohlenstoffnanoröhre, Quantenpunkte, Superlegierung, synthetischer Diamant, durchsichtiger Beton, 3D-Display, ferroelektrische Flüssigkristallanzeige, Holographie, interferometrische Modulatoranzeige, Laser-Video-Display, OLED Display, Mikro-LED Display, Teleskop-Pixelanzeige, zeitgemultiplexter optischer Verschluss, volumetrische Anzeige, Biometrie, digitale Dufttechnologie, elektronische Nase, E-Textil, flexible Elektronik, Memoristor, molekulare Elektronik, nanoelektromechanisches System, Spintronik, Thermo-Kupfer-Säulen-Stoß, dreidimensionale integrierte Schaltung, konzentrierte Solarenergie, elektrischer Doppelschicht-Kondensator, Schwungradspeicherung, Speicherung von Netzenergie, Heim-Brennstoffzelle, Lithium-Eisen-Phosphor-Batterie, Lithium-Schwefel-Batterie, Magnesium-Batterie, Nanodraht-Batterie, Ozean-Thermische Energieumwandlung, intelligentes Netz, Vortex-Motor, drahtlose Energie-Übertragung, Nullenergiehaus, computergeneriertes Bild, virtuelle Realität, hochauflösendes Fernsehen, 5G zellulare Kommunikation, künstliche Intelligenz, erweiterte Realität, Blockchain, Kohlenstoffnanoröhren-Feldeffekttransistor, zivile Technik, Kryptowährung, Exascale-Computing, Gestenerkennung, Internet der Dinge, neue Speichertechnologie, neue magnetische Speichertechnologie, optische Platten der vierten Generation, holografischer Speicher, allgemeines Rechnen auf Grafikprozessoren, Exokortex, maschinelle Übersetzung, maschinelles Sehen, mobile Zusammenarbeit, Nano-Funk, optische Datenverarbeitung, Quantencomputer, Quantenkryptographie, Radiofrequenz-Identifikation, semantisches Web, intelligenter Lautsprecher, Software-definiertes Radio, Spracherkennung, subvokale Erkennung, Hybrid-Forensik, Körperimplantat, Kryonik, Wiederbelebung ausgestorbener Tierarten, Gentechnik, verzögerte Reanimation, künstlicher Winterschlaf, Immuntherapie/-onkologie, Nanomedizin, Nanosensoren, onkolytische Viren, individualisierte Medizin, whole

genome sequencing, Roboterchirurgie, Stammzellentherapie, synthetische Biologie, synthetische Genomik, Gewebezüchtung, Tricorder, Gehirn-Computer-Schnittstelle, Neuroinformatik, Elektroenzephalographie, Neuroprothetik, hülsenlose Munition, gerichtete Energiewaffe, Elektro-Laser, elektromagnetische Waffen, elektrothermisch-chemische Technologie, grünes Geschoss, Laser-Waffe, Strahlenwaffe, Schallwaffe, Tarntechnologie, Wirbelringkanone, Elektroschockwaffe, künstliche Schwerkraft, Stasiskammer, aufblasbares Weltraum-Habitat, Miniatursatellit, Android, Nanorobotik, Exoskelett, selbstkonfigurierender Roboter, unbemanntes Fahrzeug, luftlose Reifen, Fahrzeug mit alternativen Kraftstoffen, Elektrohydrodynamischer Antrieb, Fluidik, Fusionsrakete, Schwebefahrrad, Jetpack, Rucksackhelikopter, Magnetschwebebahn, Vactrain, magnetische Schwebetechnik, Massenantrieb, Personal Rapid Transit, physisches Internet, Roller-Sharing-System, fliegendes Treibstofflager, wiederverwendbares Startsystem, Raumaufzug, Raumflugzeug, Überschalltransport, Fahrzeugkommunikationssystem.

A.3 Detailed Information on the Calculation of Text-Based, Meta-Information and Network Features

Text-Based Features

1) **Texts** – To identify the most relevant terms when predicting a firm's innovation status, we transform the scraped texts into a format that allows us to do mathematical operations: We convert the website texts into a term-document matrix, e.g. see Baeza-Yates & Ribeiro-Neto (1999) and Blei et al. (2003), which is a matrix that counts the frequency of terms that occur in a collection of documents (websites in this particular case). Every column represents a document, and a row represents a word from a predefined vocabulary space. Accordingly, every cell counts how often a particular word appears in a particular document. We define our vocabulary space as the 5,000 most frequent words in our entire training text corpus. Before we calculate the term-document matrix, we conduct the following preprocessing steps: We merge all scraped subpages related to a single firm and delete irrelevant subpages (imprints, information about cookies, or texts that are prescribed by law) by using the gold-bloat method based on a supervised machine learning regression model (see Kinne & Lenz 2021). Also, every word is converted into lower case and lemmatised using the Python package *spacy*. We exclude punctuation as well as English and German stop words (word lists are derived from the Python package *nltk*). Additionally, we manipulate the term-frequency counts with the TF-IDF scheme (Baeza-Yates & Ribeiro-Neto 1999), as it usually improves predictions. Therefore, each document is tokenised, and the modified term-document frequency is calculated using the *Tfidf-Vectorizer* algorithm from *scikit-learn*.

2) **Emerging technology terms** – To capture firms that mention emerging technologies, we conduct a keyword search in which we calculate whether a technology from Wikipedia's list of emerging technologies appears on a firm's website using all subpages and the entire vocabulary as well as the Python package *regex*.[130] We only search for a selection of technologies that are in a research, development, diffusion, or commercialisation stage, as it is a criterion for an innovation to be brought into use. A detailed list of all used keywords is provided in Appendix A.2. The feature *emerging_tech* is a dummy variable that captures whether an emerging technology term appears on a firm website.

3) **Latent patterns** – Latent patterns on a website, which might reveal a firm's innovation status, are captured by the latent dirichlet allocation model (LDA) (see Blei et al. 2003). The LDA algorithm assumes that a document consists of a set of

[130] See `https://en.wikipedia.org/wiki/List_of_emerging_technologies` [Last accessed: 24.02.2024].

topics, while every topic is a distribution of words. By linking each word in a document to a topic and iteratively improving assignments, the algorithm learns the distribution of topics in the text corpus as well as the distribution of words related to each topic. Moreover, after applying the LDA algorithm, the topic-document matrix shows how much every topic contributes to a document (website). We do not want our topic model to be exclusively valid for our sample. Hence, we calibrate our topics on a separate sample, which consists of 32,276 websites of firms observed in the MUP 2019 but not in the MIP 2019. We apply the same text preprocessing to it as to our MIP sample, but with two differences. First, we use a larger vocabulary space (15,000 most frequent words). Second, we do not manipulate word counts using the TF-IDF formula but generate a TF-IDF stop word dictionary, excluding words with a lower sum of TF-IDF scores than three within the LDA corpus. The latter is applied to ensure that words that are characteristic of particular websites are considered. Also, to improve our model performance, we delete all words that appear less than 50 times and in more than 90% of all documents in the LDA corpus. We use the *TfidfVectorizer* to calculate the stop word dictionary. This dictionary as well as the *CountVectorizer* from *scikit-learn* are applied to generate a term-document matrix for our LDA sample. A term-document matrix for the MIP sample is calculated in the same manner. The Python package *scikit-learn* is used to train the LDA model. In the standard LDA approach, the number of topics needs to be defined. To solve this issue, we apply the grid-search technique to optimise the number of topics. For this, we use the *GridSearchCV* algorithm from *scikit-learn*. We evaluate which model parameter combination leads to the best result according to the log likelihood. We conduct a grid-search over different values for the 'number of topics'-parameter as well as the document-topic prior. We try 150, 180, 200, and 250 topics and values of 0.05, 0.1 for the document-topic prior. The optimal number of topics is 150; the highest log-likelihood is achieved with a document-topic prior of 0.1. After fitting the LDA model with the separate sample, the topic distribution for each website in our MIP sample is predicted (*LDA topic*) and used in our random forest models, i.e. the predicted topic share in a document for each topic is used as a feature.

4) **Topic popularity index** – The topic popularity index is the sum of document-topic probabilities weighted by the relative frequency with which each topic appears in the entire text corpus (*pop_score*). A topic is considered to appear in a document if the document-topic probability is greater than 2%.

5) **Language classification** – The export orientation of a website might provide information about a firm's innovation status. English is worldwide the most widely spoken language by the total number of speakers. Therefore, it is quite likely that firms with international customers describe their products in English. We measure the share of subpages in the English language, as well as all other languages

except German, to approximate the export orientation of a firm (*english_language*, *other_lang*). For the language classification of subpages, we apply the Python package *langdetect*.[131]

6) **Share of numbers** – We test whether the share of numbers in the total text length per document relates to the innovation status. The share is calculated by the ratio of digits within a string (document). For example, the text 'This book costs 500 dollars.' has a ratio of 3/28, i.e. 10.7%. The corresponding variable is named *share_numbers*.

7) **Flesch-reading-ease score** – The Flesch-Reading-Ease score is a metric used to assess the complexity of texts. The main idea for the index is that short words and short sentences are easier for readers to understand. The Python package *ReadabilityCalculator* was used to calculate the score.[132] The full definition can be found in Flesch (1948) and the corresponding variable is named *flesch_score*.

Meta-Information Features

8) **Website size** – Approximating the firm's size might help to predict a firm's innovation status. For example, Kinne & Axenbeck (2020) show that the number of subpages correlates with firm size, and larger firms tend to be more likely to implement an innovation. Hence, we use the number of subpages as a feature to predict a firm's innovation status (*nr_pages*). One problem related to this feature is that it is truncated at 50 subpages due to the scraping limit of the web-scraper. However, as only 1.5% of our observations exceed the scraping limit, we do not see a severe problem here. Moreover, we use a random forest model that selects cut-off points for splitting. Hence, it can cope with truncated features. We additionally analyse to what extent the number of characters per website (*text_length*), which might also relate to firm size, informs about the firm's innovation status.

9) **Loading time** – This feature serves as a proxy for a firm's hardware infrastructure. A website's loading time (*load_time*) is determined by a http or https request. The time from sending the request until the arrival of the response is measured. Servers that are far away or that only process the requests slowly (e.g. due to bad hardware or an overload) have a higher loading time (in milliseconds). However, it should be noted that the IT infrastructure can also be outsourced to professional hosting firms. We retrieved the loading time using the Python packages *requests* and *time*.[133] The latter is a standard Python library.

[131] *Language detection* package: `https://pypi.org/project/langdetect/` [Last accessed: 24.02.2024].
[132] *Readability Calculator* package: `https://pypi.org/project/ReadabilityCalculator/` [Last accessed 04.01.2024].
[133] *Requests* package: `https://pypi.org/project/requests/` [Last accessed: 24.02.2024].

10) **Mobile version** – For each website, it is retrieved whether a version for mobile end-user devices exists. A Google API ("PageSpeed Insights") is used to extract this information from the websites. The data are delivered as JSON objects. Within the delivered data, the binary variable *score* within the data structure *usability* is used (*mobile_version*). It indicates Google's mobile version passing score. The Python packages *json*, *mechanize*, *socket* and *urllib* are used for this exercise.

11) **Website age** – To determine the website age, we use web.archive.org. The website includes an internet archive that allows you to look at websites at earlier stages. We wrote a small programme that automatically goes to web.archive.org and searches for the first entry of a particular website. This characteristic serves as a proxy for the digital age of a firm (*domain_purchase_year_proxy*). Our programme uses the Python package *urllib*.

Network Features

12) **Centrality** – Relationships with other firms might also link to a firm's innovation status. If a firm is related to another firm, the firm will likely refer to it on its website. Hence, to capture relationships with other firms, the sum of outgoing (*outgoing_links*) and incoming (*incoming_links*) hyperlinks to/from other firms is observed. Outgoing hyperlinks are measured by the number of external links on a firm's website. We measure incoming hyperlinks by counting how often firms that are listed in the entire MUP refer to a particular firm. Additionally, a directed graph is constructed. Here, a vertex represents a firm, and an edge is a hyperlink from one firm to another. The Pagerank centrality measure is calculated with the Python package *igraph*[134] and the function *pagerank*. The default parameters are used, and the resulting variable is called *pagerank_index*.

13) **Social media** – The use of social media could also be correlated with the firm's innovation status. Therefore, the sum of hyperlinks to the websites Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub, Flickr, and Vimeo is counted and used as another feature (*social_media*). This is calculated by means of *regex* again.

14) **Bridges** – We construct an undirected graph as well. A bridge is an edge of a graph whose removal increases the number of connected components. For each vertex, we count the number of times it is part of a bridge. The Python package *networkx* and the function *bridges* are used to calculate the bridges and the described measure.[135] The resulting variable is named *bridge_index*.

[134] *Igraph* package: `https://igraph.org` [Last accessed 04.01.2024].
[135] *Networkx* package: `https://networkx.github.io` [Last accessed 04.01.2024].

A.4 Most Relevant Features for Each *All-Feature* Model[136]

Table A.1: **Most relevant features for product innovators.**

Model	Top 100 most relevant features
Product innovators	'LDA topic 35', 'english_language', 'word: system', 'text_length', 'LDA topic 134', 'nr_subpages', 'word: software', 'LDA topic 65', 'word: to develop (transl.)', 'word: application (transl.)', 'LDA topic 105', 'word: test', 'LDA topic 7', 'word: product (transl.)', 'incoming_links', 'word: worldwide (transl.)', 'word: innovative (transl.)', 'LDA topic 98', 'domain_purchase_year_proxy', 'word: version', 'word: innovative', 'LDA topic 41', 'LDA topic 20', 'word: technology (transl.)', 'share_numbers', 'word: sensor', 'LDA topic 127', 'social_media', 'flesch_score', 'word: development (transl.)', 'emerging_tech', 'LDA topic 34', 'word: technology', 'LDA topic 38', 'LDA topic 96', 'LDA topic 75', 'LDA topic 46', 'pop_score', 'LDA topic 39', 'word: automatic (transl.)', 'LDA topic 101', 'LDA topic 70', 'LDA topic 78', 'LDA topic 84', 'LDA topic 128', 'outgoing_links', 'LDA topic 148', 'LDA topic 97', 'word: to optimize (transl.)', 'word: software development (transl.)', 'word: application (transl.)', 'LDA topic 119', 'LDA topic 36', 'word: component (transl.)', 'LDA topic 69', 'load_time', 'LDA topic 52', 'LDA topic 56', 'LDA topic 60', 'LDA topic 143', 'word: digital', 'LDA topic 8', 'LDA topic 113', 'LDA topic 120', 'word: complex (transl.)', 'LDA topic 53', 'LDA topic 138', 'LDA topic 144', 'LDA topic 51', 'LDA topic 15', 'LDA topic 19', 'word: support', 'LDA topic 103', 'LDA topic 106', 'word: user (transl.)', 'LDA topic 57', 'LDA topic 107', 'LDA topic 49', 'LDA topic 104', 'word: deployment (transl.)', 'LDA topic 5', 'LDA topic 111', 'word: interfaces (transl.)', 'LDA topic 85', 'LDA topic 61', 'LDA topic 114', 'LDA topic 43', 'LDA topic 45', 'LDA topic 26', 'LDA topic 132', 'LDA topic 16', 'word: production (transl.)', 'LDA topic 125', 'LDA topic 146', 'word: year (transl.)', 'LDA topic 140', 'LDA topic 91', 'word: integrate (transl.)', 'LDA topic 79', 'word: special (transl.)'

Table A.2: **Most relevant features for process innovators.**

Model	Top 100 most relevant features
Process innovators	'text_length', 'LDA topic 98', 'english_language', 'social_media', 'LDA topic 41', 'flesch_score', 'incoming_links', 'LDA topic 7', 'LDA topic 75', 'word: worldwide (transl.)', 'outgoing_links', 'nr_subpages', 'LDA topic 84', 'word: product (transl.)', 'word: system', 'LDA topic 65', 'LDA topic 20', 'LDA topic 57', 'LDA topic 53', 'share_numbers', 'LDA topic 106', 'LDA topic 148', 'LDA topic 104', 'load_time', 'LDA topic 99', 'LDA topic 122', 'LDA topic 140', 'word: technology (transl.)', 'pop_score', 'word: to develop (transl.)', 'LDA topic 35', 'LDA topic 31', 'LDA topic 127', 'LDA topic 12', 'word: ISO', 'LDA topic 39', 'LDA topic 121', 'LDA topic 32', 'LDA topic 36', 'word: innovative (transl.)', 'LDA topic 2', 'LDA topic 100', 'LDA topic 6', 'LDA topic 13', 'LDA topic 120', 'word: standard', 'word: successful (transl.)', 'LDA topic 43', 'LDA topic 103', 'LDA topic 60', 'LDA topic 64', 'LDA topic 96', 'LDA topic 23', 'LDA topic 133', 'LDA topic 93', 'LDA topic 78', 'LDA topic 40', 'LDA topic 146', 'LDA topic 74', 'LDA topic 101', 'LDA topic 97', 'word: to start (transl.)', 'word: international', 'LDA topic 147', 'LDA topic 86', 'LDA topic 73', 'LDA topic 144', 'LDA topic 14', 'LDA topic 46', 'word: partner', 'LDA topic 19', 'LDA topic 68', 'word: team', 'LDA topic 30', 'LDA topic 141', 'LDA topic 123', 'LDA topic 111', 'LDA topic 34', 'LDA topic 134', 'word: application (transl.)', 'LDA topic 22', 'word: as well as (transl.)', 'LDA topic 0', 'LDA topic 24', 'LDA topic 113', 'LDA topic 88', 'LDA topic 105', 'LDA topic 8', 'LDA topic 94', 'LDA topic 44', 'LDA topic 79', 'LDA topic 114', 'LDA topic 5', 'LDA topic 126', 'LDA topic 83', 'LDA topic 45', 'LDA topic 129', 'LDA topic 56', 'LDA topic 117', 'LDA topic 145'

[136]Transl.: translated from German to English language.

Table A.3: **Most relevant features for innovators.**

Model	Top 100 most relevant features
Innovators	'text_length', 'english_language', 'LDA topic 98', 'nr_subpages', 'word: system', 'word: to develop (transl.)', 'LDA topic 65', 'LDA topic 35', 'word: worldwide (transl.)', 'word: innovative (transl.)', 'LDA topic 84', 'LDA topic 134', 'LDA topic 41', 'LDA topic 20', 'word: product(transl.)', 'LDA topic 7', 'LDA topic 31', 'social_media', 'flesch_score', 'domain_purchase_year_proxy', 'word: development (transl.)', 'word: application (transl.)', 'incoming_links', 'LDA topic 78', 'outgoing_links', 'LDA topic 96', 'LDA topic 75', 'word: successful (transl.)', 'LDA topic 103', 'word: complex (transl.)', 'LDA topic 101', 'LDA topic 100', 'LDA topic 140', 'share_numbers', 'LDA topic 5', 'LDA topic 105', 'LDA topic 122', 'LDA topic 0', 'LDA topic 56', 'LDA topic 114', 'load_time', 'LDA topic 127', 'LDA topic 50', 'LDA topic 6', 'LDA topic 53', 'LDA topic 69', 'LDA topic 94', 'LDA topic 51', 'LDA topic 46', 'LDA topic 120', 'pop_score', 'LDA topic 102', 'LDA topic 90', 'LDA topic 113', 'word: to offer (transl.)', 'LDA topic 121', 'LDA topic 36', 'LDA topic 52', 'LDA topic 32', 'LDA topic 19', 'LDA topic 89', 'word: experience (transl.)', 'LDA topic 2', 'LDA topic 60', 'LDA topic 142', 'word: innovative', 'LDA topic 43', 'LDA topic 23', 'LDA topic 87', 'LDA topic 28', 'LDA topic 39', 'LDA topic 148', 'LDA topic 133', 'LDA topic 106', 'LDA topic 11', 'LDA topic 34', 'LDA topic 82', 'LDA topic 37', 'LDA topic 13', 'LDA topic 86', 'word: as well as (transl.)', 'LDA topic 61', 'LDA topic 33', 'LDA topic 12', 'LDA topic 126', 'word: high (transl.)', 'LDA topic 22', 'LDA topic 71', 'LDA topic 85', 'LDA topic 138', 'LDA topic 144', 'LDA topic 117', 'LDA topic 83', 'LDA topic 16', 'word: deployment (transl.)', 'LDA topic 136', 'LDA topic 147', 'LDA topic 123', 'LDA topic 64', 'LDA topic 68'

Table A.4: **Most relevant features for innovation expenditures.**

Model	Top 100 most relevant features
Innovation expend.	'english_language', 'LDA topic 98', 'text_length', 'nr_subpages', 'word: system', 'word: development (transl.)', 'word: to develop (transl.)', 'word: technology', 'LDA topic 134', 'word: innovative (transl.)', 'word: innovation', 'incoming_links', 'word: international', 'LDA topic 148', 'LDA topic 105', 'word: product (transl.)', 'word: application (transl.)', 'word: research (transl.)', 'word: worldwide (transl.)', 'LDA topic 84', 'LDA topic 7', 'domain_purchase_year_proxy', 'LDA topic 36', 'LDA topic 106', 'outgoing_links', 'LDA topic 35', 'flesch_score', 'LDA topic 28', 'LDA topic 5', 'LDA topic 20', 'LDA topic 65', 'load_time', 'LDA topic 100', 'word: innovative', 'LDA topic 39', 'LDA topic 125', 'share_numbers', 'LDA topic 41', 'LDA topic 120', 'LDA topic 73', 'LDA topic 1', 'integration', 'pop_score', 'LDA topic 82', 'LDA topic 13', 'social_media', 'emerging_tech', 'LDA topic 104', 'LDA topic 57', 'LDA topic 6', 'LDA topic 53', 'LDA topic 109', 'LDA topic 26', 'LDA topic 75', 'word: high', 'LDA topic 34', 'LDA topic 32', 'LDA topic 89', 'LDA topic 49', 'LDA topic 140', 'LDA topic 81', 'word: workshop', 'LDA topic 83', 'LDA topic 113', 'word: management', 'LDA topic 22', 'LDA topic 59', 'LDA topic 56', 'LDA topic 31', 'LDA topic 67', 'LDA topic 24', 'LDA topic 0', 'LDA topic 79', 'LDA topic 68', 'LDA topic 102', 'LDA topic 61', 'LDA topic 3', 'LDA topic 138', 'LDA topic 44', 'LDA topic 40', 'LDA topic 128', 'LDA topic 146', 'LDA topic 141', 'word: to optimize', 'LDA topic 70', 'LDA topic 78', 'LDA topic 132', 'LDA topic 95', 'word: process (transl.)', 'LDA topic 80', 'LDA topic 127', 'LDA topic 60', 'LDA topic 93', 'LDA topic 133', 'LDA topic 114', 'LDA topic 46', 'word: high', 'word: as well as', 'LDA topic 96', 'LDA topic 8'

A.5 Learned Hyperparameters for Random Forest Models Using Different Feature Sets and Target Variables

Table A.5: **Learned hyperparameters for random forest models using different feature sets and target variables.**

Feature sets			Number of trees	Max. depth	Min. impurity decrease
Text	Meta	Network			
Product innovators					
x			1000	50	0.001
	x		1000	50	0.001
		x	1500	50	0.001
x	x	x	1500	100	0.001
Process innovators					
x			1000	50	0.001
	x		1500	50	0.010
		x	1000	50	0.001
x	x	x	1500	50	0.001
Innovators					
x			1500	50	0.001
	x		1000	50	0.001
		x	500	50	0.001
x	x	x	1000	50	0.001
Innovation expenditures					
x			1500	100	0.001
	x		1000	50	0.010
		x	1000	50	0.010
x	x	x	1000	50	0.001

A.6 AUC Values for Different Splits Between the Training and Test Sample

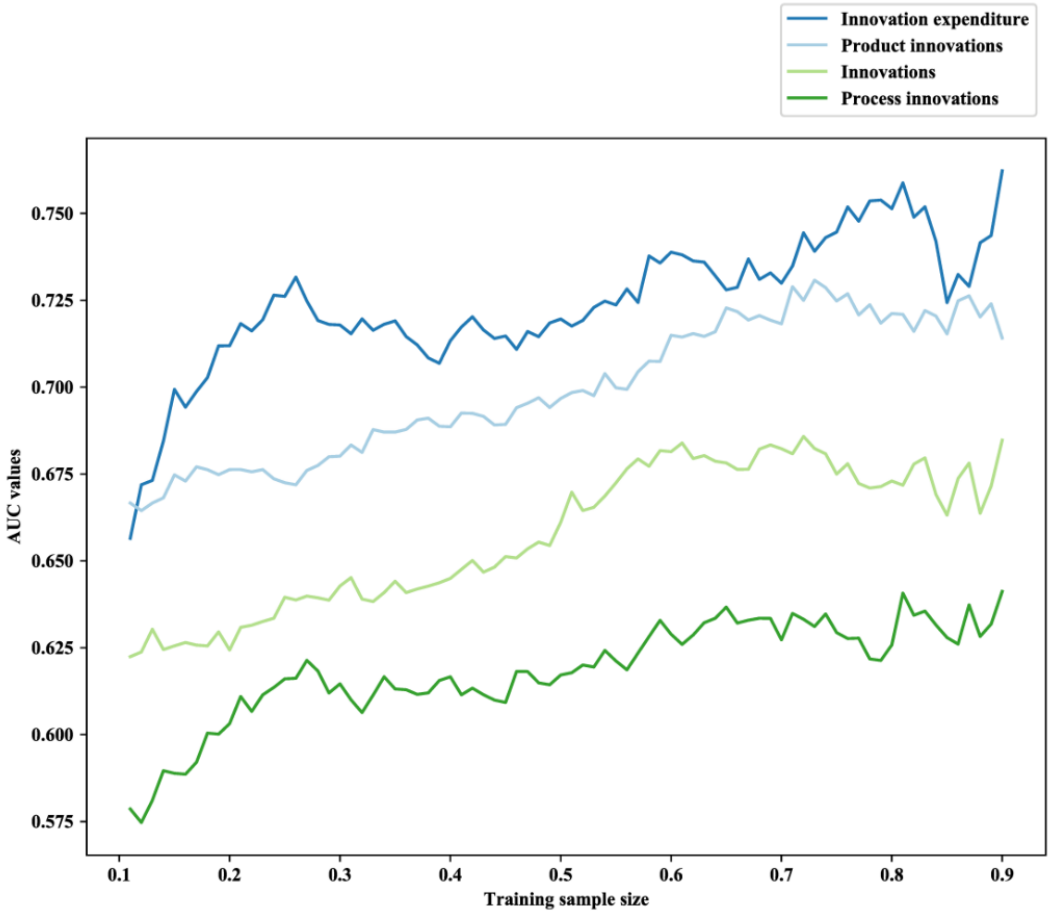


Figure A.3: **AUC values for different splits between the training and test samples.** Line plot that illustrates for each indicator how AUC values of the *all-feature* models change if the train/test split changes from (0.1/0.9) to (0.9/0.1) in steps of 0.01. Own illustration.

Appendix B

Measuring Technological Change - A Novel Text Mining Approach

B.1 Description of the Text Processing Pipeline

Table B.1: **Text data processing for news article and firm website data.**

Steps	Description
1. Data filtering	Delete data points without text and text duplicates.
2. Tokenisation	The texts are split into individual words using the Python Natural Language Toolkit (NLTK) software package (Loper & Bird 2002).
3. Stop word filter	Delete words using multiple stop word lists. The deleted words are usually not relevant for classification. An example of a stop word is *'and'*.
4. Stemming	Different variants of a word are reduced to their base form. E.g.: The words *'tree'* and *'trees'* become *'tree'*. The Snowball Stemmer (NLTK package) is used.
5. Short word removal	Words with a character length of one or two are deleted. These are usually punctuation marks or special characters.
6. Unification of words	All capital letters are converted to lower case to reduce the vocabulary size.
7. Special character removal	Special characters are removed.
8. Word selection	The 10,000 words with the highest *term frequency – inverse document frequency* (tf-idf) score are extracted. The remaining words are deleted to reduce the dimension and *noise* of the data.

Notes: The pipeline filters irrelevant data, extracts words from the text, and standardises.

B.2 Descriptive Statistics of the Newspaper Data

Table B.2: **Descriptive statistics for 25K news articles (per language).**

Metric	Language	Mean	Median	Std. dev.	Min.	Max.
Number of characters	German	3,652	3,450	1,820	1,000	33,004
Number of words	German	515	485	261	115	4,563
Number of characters	English	3,125	3,188	1,180	795	5,196
Number of words	English	525	535	200	130	899

Notes: The statistics (mean, median, standard deviation, minimum, and maximum) are reported before the natural language processing (NLP) pipeline is applied to the newspaper data.

B.3 Performance of the Machine Learning Model

Table B.3: **Most important words in the random forest model that was trained on news articles.** The feature importance is the *Mean Decrease in Impurity* (MDI) measure.

Word	Importance	Word	Importance
digital	0.13423	onlin	0.00578
digitalis (*translated*)	0.01936	internet	0.00549
digitization	0.01674	app	0.00529
googl	0.01244	apps	0.00452
smartphon	0.01144	pixel	0.00436
softwar	0.01003	gmbh	0.00434
appl	0.00846	android	0.00421
user	0.00818	analog	0.00403
comput	0.00675	samsung	0.00401
facebook	0.00662	algorithm	0.00401
job market (*translated*)	0.00656	data	0.00391
bitcoin	0.00592	iphon	0.00374
use (*translated*)	0.00590	virtual	0.00364

Notes: Some of the listed words are manually translated from German to English.

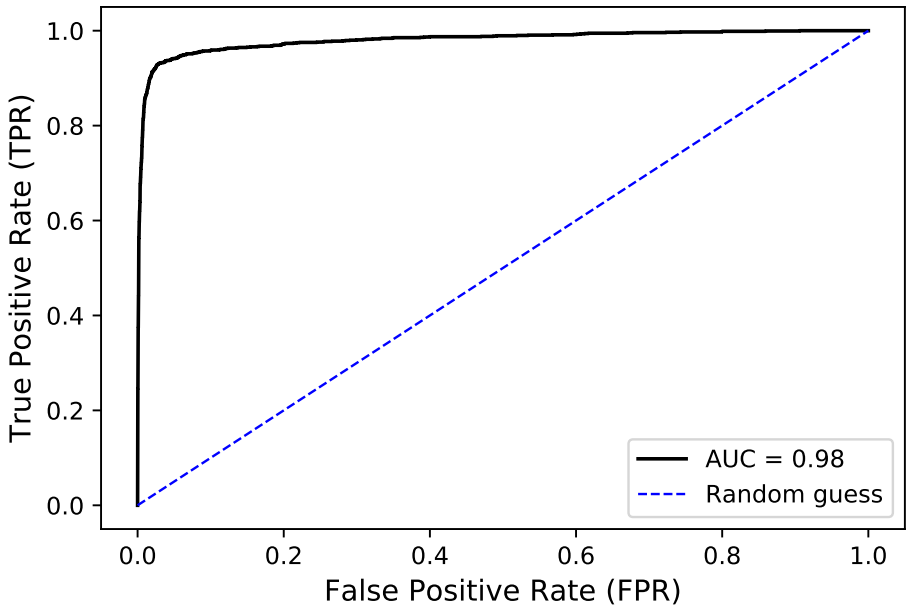


Figure B.1: **ROC plot and AUC value (bottom right) for the regression model trained on the news articles.** The baseline values are shown on the diagonal. Own illustration.

B.4 Digitalisation over Time

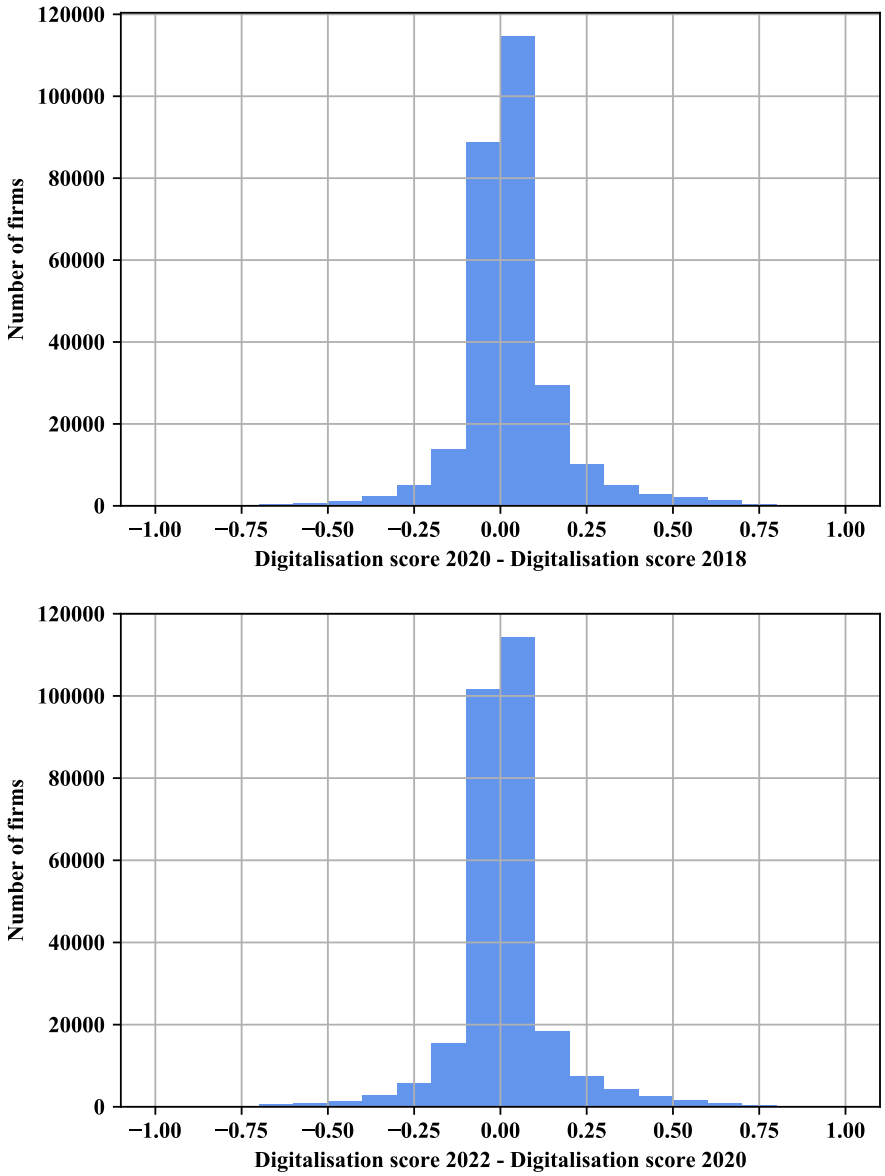


Figure B.2: **Temporal change in firm digitalisation scores.** The histograms are based on the predictions for the years 2018/2020 (top) and 2020/2022 (bottom). Own illustration.

B.5 Mannheim Innovation Panel Data: Questions about Digitalisation

Table B.4: **List of questions about digitalisation in the Mannheim Innovation Panel 2020 (MIP 2020) with the response options (answers).**

Number	Question	Answers
1	Use of digital platforms for delivering products or services	None, Low, Medium, High
2	Use of social networks to contact customers and obtain new customers	None, Low, Medium, High
3	Customisation of products through digital channels	None, Low, Medium, High
4	Methods of digital price differentiation	None, Low, Medium, High
5	Use of digital sources to collect data	None, Low, Medium, High
6	Digital integration of suppliers, business and other cooperation partners	None, Low, Medium, High
7	Use of digital media/tools for crowdsourcing of innovative ideas	None, Low, Medium, High
8	Use of machine learning or artificial intelligence	None, Low, Medium, High

Notes: The question block has the following title: “How important are the following digital elements for the current business model of your enterprise?".

B.6 Firm Resilience Use Case: Summary Statistics

Table B.5: **Summary statistics of the *firm resilience* estimation sample that is based on MUP and web data.**

	Mean	Std. dev.	Min.	Max.	Count
digitalisation 2020	0.258363	0.2091660	0.001176	0.9266	176,902
digitalisation 2018	0.227544	0.2041340	0.001122	0.9165	176,902
Δdigitalisation 2020-2018	0.030818	0.1455610	-0.817273	0.8684	176,902
rating 2021	3.350102	0.5029563	1	5	176,902
rating 2020	3.379925	0.4906057	1	5	176,902
rating 2019	3.440050	0.4767542	1	5	176,902
rating 2018	3.458093	0.4701994	1	5	176,902
Δrating 2021-2019	-0.089947	0.3549380	-3.29	3	176,902
employees 2020	0.033506	0.3395655	0.001	64.5	176,902
employees 2018	0.032793	0.3173355	0.001	64.5	176,902
Δemployees 2020-2018	0.000713	0.1276257	-36.77	29.3	176,902

Notes: The number of employees per firm is stated in thousands. The statistics for the categorical data are shown in a separate table.

Table B.6: **Firm counts for the categorical MUP characteristics location, industry, legal form, and founding period (for the *firm resilience* estimation sample).**

Location (German State)	Count
Baden-Württemberg	27,556
Berlin	10,292
Brandenburg	3,951
Bremen	2,250
Hamburg	6,531
Bavaria	31,544
Saxony	4,059
Thuringia	1,641
Hesse	14,982
Mecklenburg–Western Pomerania	1,268
Lower Saxony	19,034
North Rhine-Westphalia	35,117
Rhineland-Palatinate	8,059
Saarland	2,130
Saxony-Anhalt	1,269
Schleswig-Holstein	7,219

Industry	Count
Agriculture, forestry, and fishing (1-3)	1,647
Mining and quarrying (5-9)	200
Manufacturing industry (10-33)	20,579
Energy supply (35-35)	939
Water supply, sewage and waste disposal, and pollution clean-up (36-39)	1,040
Construction (41-43)	19,597
Wholesale and retail trade, and repair of motor vehicles (45-47)	35,602
Transport and storage (49-53)	4,108
Accommodation and food service activities (55-56)	6,910
Information and communication (58-63)	8,743
Provision of financial and insurance services (64-66)	6,133
Real estate and housing (68-68)	6,849
Provision of professional, scientific, and technical services (69-75)	26,745
Administrative and support service activities (77-82)	11,440
Public administration, defense, and social security (84-84)	824
Education and teaching (85-85)	3,460
Health and social services (86-88)	11,197
Art, entertainment, and recreation (90-93)	3,210
Provision of other services (94-96)	7,658

Legal form	Count
Self-employed profession (*Freie Berufe*)	9,715
Commercial operation (*Gewerbebetrieb*)	38,634
BGB partnership (*BGB-Gesellschaft*)	9,592
Single firm (*Einzelfirma*)	8,542
GmbH & Co. KG	10,231
OHG	1,388
KG	1,187
GmbH	89,512
AG	2,081
eG	907
eV	4,697
UG	415

Founding period	Count
$<$ 1990	49,571
1990 - 1999	44,932
2000 - 2009	49,723
$\geq$ 2010	32,676

Notes: Groups with less than fifty observations are not shown in the table. The legal form labels have been translated from German into English where possible (see brackets). The definition of industries is based on the 2-digit NACE codes and is shown in the brackets.

Appendix C

Intangible Capital Indicators Based on Web Scraping of Social Media

C.1 Overview of Dimensions on the Platforms Facebook and Kununu

Table C.1: **Overview of dimensions on platforms.**

	Kununu	Facebook
Information on the start page	Grade, scale, recommendations, hits, benefits.	Number of "likes" obtained without the Facebook Graph API.
Historic information	Individual/detailed ratings (including "firm image" and "on-the-job training/career development") are available with a time stamp, so the data can be reconstructed historically.	Get firm and user contributions, including comments and "likes", via the Facebook Graph API. You only get the current number of fans.
Remarks	Problem of deletion/change.	The API access was massively restricted in the wake of the Facebook–Cambridge Analytica data scandal in early 2018.

C.2 Summary Statistics for Facebook and Kununu Data

Table C.2: **Summary statistics - Kununu - Full sample.**

	N	Mean	Median	SD	Min	Max
Training: Kununu rating	813	3.29	3.35	0.77	1	5
Image: Kununu rating	805	3.56	3.67	0.79	1	5
Training expenditures (MEUR)	1,568	0.42	0.019	7.96	0	300
Marketing expenditures (MEUR)	1,548	2.50	0.040	46.0	0	1,480
Turnover (MEUR)	1,961	252.3	10.3	2,222.4	0	57,550
Number of employees	2,051	774.8	70	6,643.5	0	156,487
Number of Kununu ratings (Training)	2,114	9.91	2	43.3	0	1,174
Number of Kununu ratings (Image)	2,114	9.83	2	43.1	0	1,171
ln(Training: Kununu rating)	813	1.16	1.21	0.26	0	1.61
ln(Image: Kununu rating)	805	1.24	1.30	0.25	0	1.61
ln(Training expenditures)	1,356	-3.49	-3.69	1.84	-7.71	5.70
ln(Marketing expenditures)	1,321	-2.63	-2.81	2.19	-8.11	7.30
ln(Turnover)	1,957	2.47	2.33	2.15	-5.30	11.0
ln(Number of employees)	2,045	4.35	4.25	1.74	0	12.0

Notes: *Full sample* indicates the merge of the MIP 2017 survey with the identified Kununu profiles (see Figure 4.2). The "Number of Kununu ratings" shows the descriptive statistics for all 2,114 MIP 2017 firms with a Kununu profile. The number of observations for "Kununu rating" is lower as we restrict the data to firms with at least 4 ratings between January 2017 and August 2018.

Table C.3: **Summary statistics - Facebook - Full sample.**

	N	Mean	Median	SD	Min	Max
Image: Facebook likes	1,498	15,905.7	215.5	142,384.2	1	3,687,320
Marketing expenditures (MEUR)	1,165	1.95	0.020	44.0	0	1,480
Turnover (MEUR)	1,425	134.5	4.23	1,891.5	0	57,550
Number of employees	1,498	465.2	35	5,436.4	0	156,487
ln(Image: Facebook likes)	1,498	5.71	5.37	2.30	0	15.1
ln(Marketing expenditures)	1,012	-3.24	-3.51	2.11	-8.11	7.30
ln(Turnover)	1,421	1.69	1.46	2.09	-5.30	11.0
ln(Number of employees)	1,492	3.77	3.56	1.69	0	12.0

Notes: *Full sample* indicates the merge of the MIP 2017 survey with the identified Facebook profiles (see Figure 4.2).

Table C.4: **Industry coverage: Kununu.**

	MIP 2017		**Full Sample**		**Estimation Sample (Training)**		**Estimation Sample (Image)**	
	N	Percent	N	Percent	N	Percent	N	Percent
A - Agriculture, forestry, and fishing	13	0.16						
B - Mining and quarrying	98	1.18	5	0.24	1	0.19	2	0.41
C - Manufacturing	3,664	44.26	996	47.11	241	46.44	220	44.72
D - Electricity, gas, steam, air conditioning supply	135	1.63	56	2.65	11	2.12	9	1.83
E - Water supply, sewerage, waste management, remediation	386	4.66	44	2.08	11	2.12	8	1.63
F - Construction	204	2.46	25	1.18	4	0.77	5	1.02
G - Wholesale, retail trade, repair of motor vehicles	435	5.25	106	5.01	26	5.01	27	5.49
H - Transportation and storage	543	6.56	109	5.16	20	3.85	16	3.25
I - Accommodation and food service activities	15	0.18	1	0.05				
J - Information and communication	615	7.43	260	12.30	78	15.03	79	16.06
K - Financial and insurance activities	255	3.08	95	4.49	23	4.43	21	4.27
L - Real estate activities	53	0.64	7	0.33				
M - Professional, scientific, and technical activities	1,329	16.05	282	13.34	75	14.45	75	15.24
N - Administrative and support service activities	502	6.06	119	5.63	28	5.39	28	5.69
O - Public administration and defence, compulsory social security	2	0.02						
P - Education	10	0.12	4	0.19				
Q - Human health and social work activities	2	0.02						
R - Arts, entertainment, and recreation	7	0.08	3	0.14	1	0.19	1	0.20
S - Other service activities	10	0.12	2	0.09			1	0.20
Total	8,278	100.00	2,114	100.00	519	100.00	492	100.00

Notes: Empty fields indicate that there are no observations.

Table C.5: **Industry coverage: Facebook.**

	MIP 2017		Full Sample		Estimation Sample	
	N	Percent	N	Percent	N	Percent
A - Agriculture, forestry, and fishing	13	0.16	1	0.06		
B - Mining and quarrying	98	1.18	9	0.58	3	0.32
C - Manufacturing	3,664	44.26	686	44.57	407	43.11
D - Electricity, gas, steam, air conditioning supply	135	1.63	28	1.82	19	2.01
E - Water supply, sewerage, waste management, remediation	386	4.66	37	2.40	24	2.54
F - Construction	204	2.46	27	1.75	19	2.01
G - Wholesale, retail trade, repair of motor vehicles	435	5.25	97	6.30	66	6.99
H - Transportation and storage	543	6.56	104	6.76	66	6.99
I - Accommodation and food service activities	15	0.18	5	0.32	4	0.42
J - Information and communication	615	7.43	173	11.24	112	11.86
K - Financial and insurance activities	255	3.08	57	3.70	32	3.39
L - Real estate activities	53	0.64	7	0.45	5	0.53
M - Professional, scientific, and technical activities	1,329	16.05	191	12.41	116	12.29
N - Administrative and support service activities	502	6.06	109	7.08	67	7.10
O - Public administration and defence, compulsory social security	2	0.02				
P - Education	10	0.12	2	0.13	1	0.11
Q - Human health and social work activities	2	0.02	1	0.06	1	0.11
R - Arts, entertainment, and recreation	7	0.08	3	0.19	2	0.21
S - Other service activities	10	0.12	2	0.13		
Total	8,278	100.00	1539	100.00	944	100.00

Notes: Empty fields indicate that there are no observations.

Table C.6: **Firm size coverage: Kununu.**

# of	**MIP 2017**		**Full Sample**		**Estimation Sample (Training)**		**Estimation Sample (Image)**	
employees	N	Percent	N	Percent	N	Percent	N	Percent
0-9	2,326	28.69	224	10.92	21	4.05	24	4.88
10-49	3,274	40.38	633	30.86	108	20.81	108	21.95
50-249	1,752	21.61	708	34.52	190	36.61	176	35.77
250+	755	9.31	486	23.70	200	38.54	184	37.40
Total	8,107	100.00	2,051	100.00	519	100.00	492	100.00

Notes: For 171 firms in the MIP 2017 sample, the number of employees is not available.

Table C.7: **Firm size coverage: Facebook.**

# of	**MIP 2017**		**Full Sample**		**Estimation Sample**	
employees	N	Percent	N	Percent	N	Percent
0-9	2,326	28.69	277	18.49	132	13.98
10-49	3,274	40.38	573	38.25	379	40.15
50-249	1,752	21.61	430	28.70	292	30.93
250+	755	9.31	218	14.55	141	14.94
Total	8,107	100.00	1,498	100.00	944	100.00

Notes: For 171 firms in the MIP 2017 sample, the number of employees is not available.

C.3 Robustness Checks for Regression Analyses

Table C.8: **Robustness check: OLS regressions Kununu - At least 3 ratings.**

	(1)	(2)	(3)	(4)
	Training:	ln(Training:	Image:	ln(Image:
Dependent variable	Kununu rating	Kununu rating)	Kununu rating	Kununu rating)
ln(Training expenditures)	0.0945***	0.0341***		
	(2.70)	(2.87)		
ln(Marketing expenditures)			0.0886***	0.0274***
			(3.37)	(3.17)
ln(Turnover)	0.0199	0.0119	-0.0403	-0.00998
	(0.47)	(0.81)	(-0.88)	(-0.68)
ln(Number of employees)	-0.0916	-0.0331*	-0.0544	-0.0160
	(-1.64)	(-1.71)	(-0.99)	(-0.88)
Industry dummies	Yes	Yes	Yes	Yes
adj. R^2	0.126	0.129	0.136	0.137
Observations	613	613	582	582

Robust t statistics in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table C.9: **Robustness check: OLS regressions Kununu - At least 5 ratings.**

	(1)	(2)	(3)	(4)
	Training:	ln(Training:	Image:	ln(Image:
Dependent variable	Kununu rating	Kununu rating)	Kununu rating	Kununu rating)
ln(Training expenditures)	0.0820**	0.0284**		
	(2.00)	(2.10)		
ln(Marketing expenditures)			0.0907***	0.0275***
			(3.11)	(3.05)
ln(Turnover)	0.00118	0.00684	-0.0714	-0.0196
	(0.03)	(0.46)	(-1.39)	(-1.20)
ln(Number of employees)	-0.0659	-0.0266	-0.0211	-0.00571
	(-1.15)	(-1.38)	(-0.34)	(-0.28)
Industry dummies	Yes	Yes	Yes	Yes
adj. R^2	0.131	0.131	0.105	0.104
Observations	433	433	413	413

Robust t statistics in parentheses
* p<0.10, ** p<0.05, *** p<0.01

C.4 Additional Graphs: Histograms and Scatter Plots

Histograms

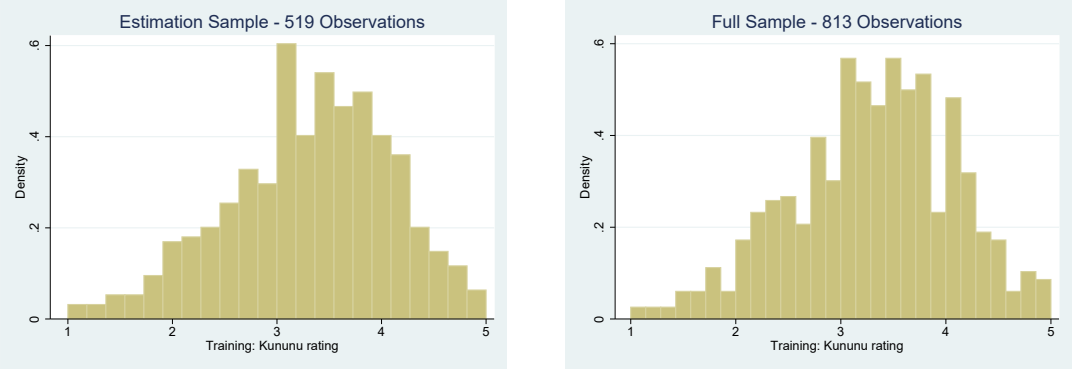


Figure C.1: **Histograms. Training: Kununu rating.** Own illustrations.

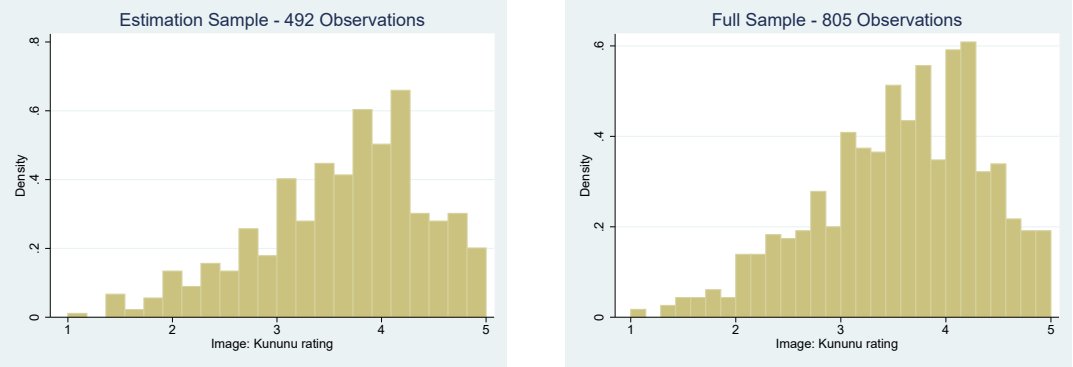


Figure C.2: **Histograms. Image: Kununu rating.** Own illustrations.

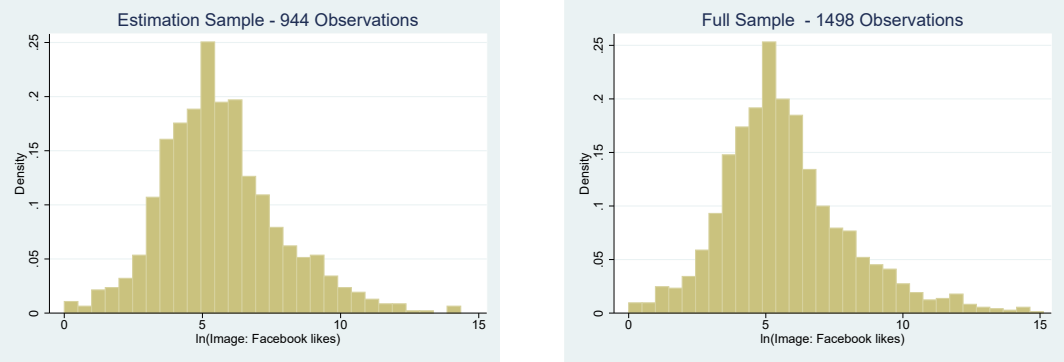


Figure C.3: **Histograms. Ln(Image: Facebook likes).** Own illustrations.

Scatter plots

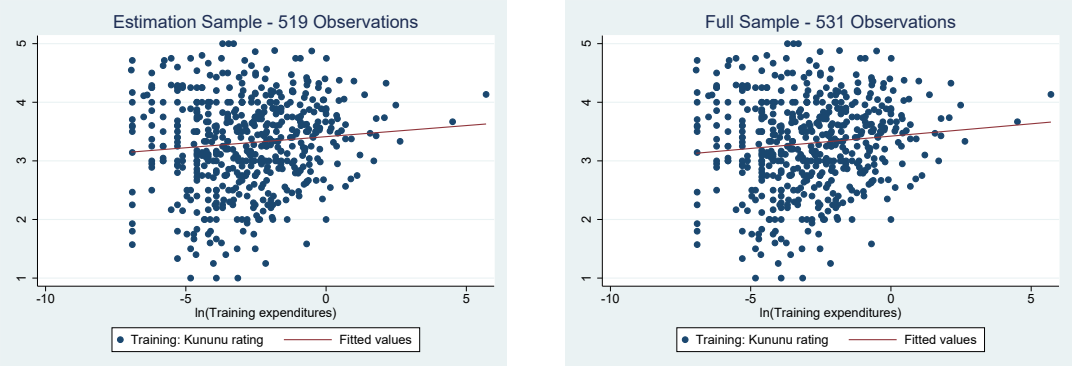


Figure C.4: **Scatterplots. Training: Kununu rating vs MIP Ln(Training expenditures).** Own illustrations.

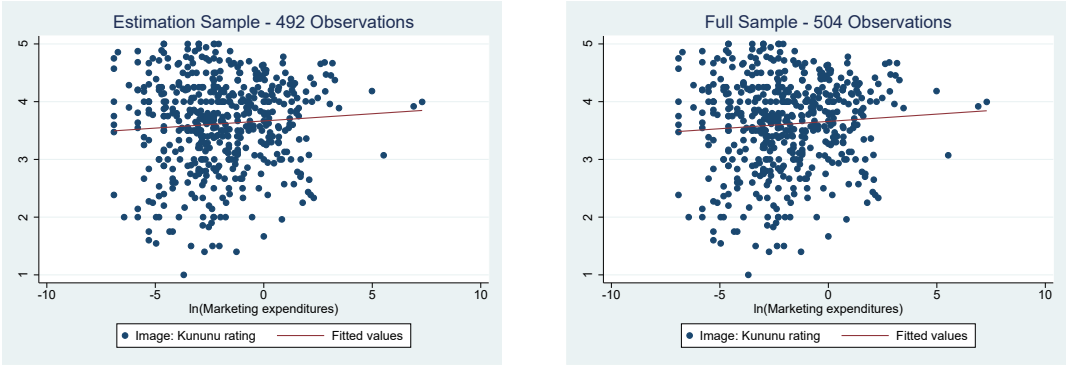


Figure C.5: **Scatterplots. Image: Kununu rating vs MIP Ln(Marketing expenditures).** Own illustrations.

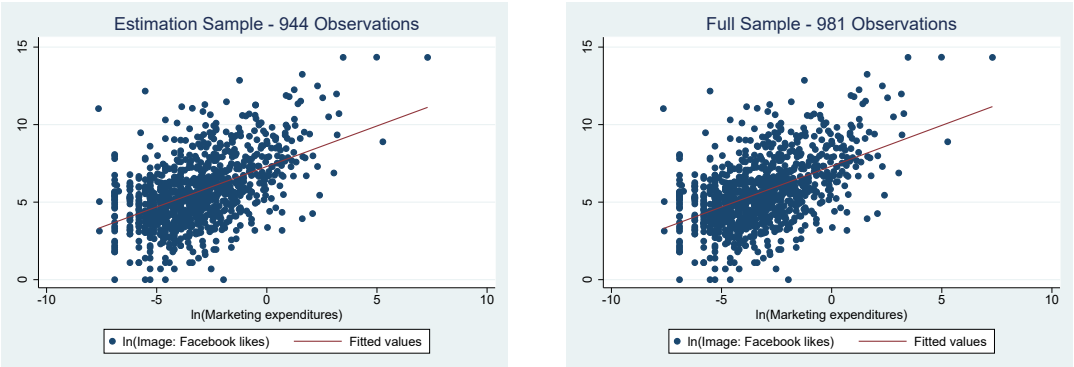


Figure C.6: **Scatterplots. Ln(Facebook likes) vs MIP Ln(Marketing expenditures).** Own illustrations.

Appendix D

Mapping Employee Mobility and Employer Networks using Professional Network Data

D.1 Processing of Employee and Employer Data from the Platform XING

Data Processing of Employment Characteristics

Table D.1: **Definitions of employment characteristics.**

	Original definition	New definition
Discipline	Administration, Clerking	Administration
	Purchasing, Materials Management, Logistics	Administration
	Human Resources	Administration
	Management and Corporate Development	Management
	Product Management	Management
	Project Management	Management
	Analysis and Statistics	IT & Data
	IT and Software Development	IT & Data
	Consulting	Consulting
	Law	Consulting
	Controlling and Planning	Finance
	Finance, Accounting and Controlling	Finance
	Customer Service and Support	Sales
	Distribution and Trade	Sales
	Research, Teaching and Development	R&D
	Health, Medicine and Social Affairs	Social & Others
	Other Fields of Activity	Social & Others
	Graphics, Design and Architecture	PR & Marketing
	PR, Public Relations and Journalism	PR & Marketing
	Marketing and Advertising	PR & Marketing
	Engineering and Technical Professions	Engineering
	Process Planning and Quality Assurance	Engineering
	Production and Craft	Production
Career level	Intern/Student	Intern/Student
	Job starter	Young Professional
	With Job Experience	Professional
	Managers (With/Without Pers. Responsibility)	Manager
	Director (Division Manager, VP, SVP etc)	Director
	Managing Director (CEO etc)	Managing Director
Employment type	Civil Servant	Civil Servant
	Honorary	Unpaid
	Freelancer	Freelancer
	Freelancing	Freelancer
	Self Employed	Freelancer
	Recruiter	Freelancer
	Shareholder	Partner
	Shareholder/Partner	Partner
	Owner	Partner
	Partner	Partner
	Intern	Intern
	Parttime	Parttime
	Fulltime	Fulltime
	Member of the Board	Fulltime

Notes: Processing of employment characteristics that are based on data from the platform XING. The original categories are translated from German into English (left column). The discipline, career level, and employment type categories are mapped to more coarse-granular groups (right column). Further and non-existing values are re-coded to *'missing/others'*.

Data Processing of Employer Characteristics

Table D.2: **Data processing of employer characteristics.**

	Original data	New data
Industry	$05 \leq$ NACE code ≤ 33	Manufacturing
	$35 \leq$ NACE code ≤ 43	Utilities / Construction (*Utilities*)
	$45 \leq$ NACE code ≤ 47	Trade
	$49 \leq$ NACE code ≤ 53	Transport
	$55 \leq$ NACE code ≤ 56	Hospitality
	$58 \leq$ NACE code ≤ 63	Information and Communication (*IT*)
	$64 \leq$ NACE code ≤ 68	Finance/ Insurance/ Real Estate/ Property and Housing (*Finance*)
	$69 \leq$ NACE code ≤ 75	Freelance, Scientific and Technical Services (*S&T Services*)
	$77 \leq$ NACE code ≤ 82	Firm Services
	$84 \leq$ NACE code ≤ 88	Social Services
	$90 \leq$ NACE code ≤ 96	Personal / Cultural Services (*P&C Services*)
Founding	founding year < 1990	< 1990
	$1990 \leq$ founding year ≤ 1999	1990 - 1999
	$2000 \leq$ founding year ≤ 2009	2000 - 2009
	founding year ≥ 2010	≥ 2010
Employer size	employer size < 10	< 10
	$10 \leq$ employer size ≤ 49	10 - 49
	$50 \leq$ employer size ≤ 249	50 - 249
	$250 \leq$ employer size ≤ 999	250 - 999
	employer size ≥ 1000	≥ 1000

Notes: Processing of employer characteristics that are based on data from the MUP. The two-digit NACE code, founding year, and employer size data (left column) are mapped to categorical variables (right column). The abbreviations of long class labels are shown in parentheses (right column). Further and non-existing values are re-coded to *'missing/others'*.

Data Processing of Legal Forms of Employers

Table D.3: **Legal form of employers.**

German label	English label
Aktiengesellschaft (AG)	Joint-stock Company
Kommanditgesellschaft (KG)	Limited Partnership
Gesellschaft mit beschränkter Haftung (GmbH)	Limited Liability Company
Freie Berufe	Liberal Professions
Gmbh & Co. KG	Gmbh & Co. KG
Eingetragene Genossenschaft (eG)	Registered Cooperative
Limited	Limited
BGB-Gesellschaft - Arbeitsgemeinschaft KG	BGB Partnership - KG Working Group
Offene Handelsgesellschaft (OHG)	General Partnership
Eingetragener Verein (eV)	Registered Association
Firma (Ausland)	Foreign Company
BGB-Gesellschaft	BGB Partnership
Einzelfirma	Individual Company
Unternehmergesellschaft (UG)	Entrepreneurial Company at Limited Liability
Gewerbebetrieb	Commercial Operation
Einzelperson	Single Person
Privatperson (Ausland)	Private Person (Foreign)

Notes: Legal forms of employers in the Mannheim Enterprise Panel (MUP). Some legal forms may not be available in the data analyses, e.g. *private person (foreign)*.
Left column: Original labels (German). Right column: Translated labels (English).
Some translations are based on `https://www.ihk.de/stuttgart/english/services2/business-support/legal-forms-of-doing-business-in-germany-3977966` [Last accessed: 24.02.2024].

D.2 Example: Data Processing

In this section, we give an example of the data processing in Section 5.3. Each row (see Table D.4) corresponds to an employment and consists of a user identifier, a start and end year, and an employer identifier, i.e. a reference to the MUP (crefo) or one that was created artificially. We assume that the employers *'HMS'* and *'IW'* could not be matched to the MUP. The employers named *'ZEW Gmbh'* and *'ZEW - Mannheim'* were linked to the same employer identifier. Table D.5 shows a simplified list of employee flows retrieved from the previous example. Each row corresponds to a flow between employers and contains a user identifier, the year of the employment switch, and a reference to the old and new employer. The employees who work for the first time (e.g. they were previously in school) or do not have follow-up employment (e.g. retired or pensioned) link to the *missing* node. The start year of the new employment is used as a timestamp for the employee flow.[137] After this transformation, the user reference is deleted to meet the privacy requirements. Table D.6 presents an example of filtered employee flows. We are interested in the employee flows within 2015, which are based neither on the employment of students nor on internal employment changes (loops). The flows that fulfill all requirements are marked. The employee flows are modelled as a graph consisting of five nodes and four edges (see Figure D.1). Table D.7 shows the degree centrality scores for the network in Figure D.1.

Table D.4: **Example: List of (un)matched employments.**

User	Employer-Name	Start-Year	End-Year	Matched	Employer identifier
0	zew mannheim	2010	2015	Yes	crefo: 2344
0	DIW	2015	Missing	Yes	crefo: 9988
1	ZEW Gmbh	2010	2012	Yes	crefo: 2344
1	IW	2012	Missing	No	artificial: 1
2	ZEW - Mannheim	2009	2015	Yes	crefo: 2344
2	IFO	2015	2020	Yes	crefo: 2885
2	DICE	2020	2021	Yes	crefo: 2367
3	IW Koeln	1980	2010	Yes	crefo: 7781
3	IAB	2010	2015	Yes	crefo: 4887
3	DIW	2015	Missing	Yes	crefo: 9988
4	zew mannheim	2015	2018	Yes	crefo: 2344
4	HMS	2018	Missing	No	artificial: 2

Notes: Example list of employments that are matched to the MUP or received an artificial identifier. Employer characteristics are not shown. The identifiers of the employers (crefo) consist of dummy data and do not match the MUP data.

[137] Exception: For the switch to pension/retirement, we use the end year of the last employment.

Table D.5: **Example: List of extracted flows.**

User	Old Employer	New Employer	Year of Switch
0	Missing	crefo: 2344	2010
0	crefo: 2344	crefo: 9988	2015
1	Missing	crefo: 2344	2010
1	crefo: 2344	artificial: 1	2012
2	Missing	crefo: 2344	2009
2	crefo: 2344	crefo: 2885	2015
2	crefo: 2885	crefo: 2367	2020
2	crefo: 2367	Missing	2021
3	Missing	crefo: 7781	1980
3	crefo: 7781	crefo: 4887	2010
3	crefo: 4887	crefo: 9988	2015
4	Missing	crefo: 2344	2015
4	crefo: 2344	artificial: 2	2018

Notes: The list of flows are retrieved from the (un)matched employments that are presented in Table D.4. Employer characteristics are not shown. The identifiers of the employers (crefo) consist of dummy data and do not match the MUP data.

Table D.6: **Example: Filtered employee flows.**

User	Old Employer	New Employer	Intern / Student	Year of switch	Keep
0	Missing	crefo: 2344	No	2010	No
0	crefo: 2344	crefo: 9988	No	2015	Yes
1	Missing	crefo: 2344	No	2010	No
1	crefo: 2344	artificial: 1	No	2012	No
2	Missing	crefo: 2344	Yes	2009	No
2	crefo: 2344	crefo: 2885	No	2015	Yes
2	crefo: 2885	crefo: 2367	No	2020	No
2	crefo: 2367	Missing	No	2021	No
3	Missing	crefo: 7781	No	1980	No
3	crefo: 7781	crefo: 4887	No	2010	No
3	crefo: 4887	crefo: 9988	No	2015	Yes
4	Missing	crefo: 2344	No	2015	Yes
4	crefo: 2344	artificial: 2	No	2018	No

Notes: The filtered list of flows is based on Table D.5. The edges that refer to 2015 and fulfill the additional requirements (no students, no loops) are marked in the last column. The identifiers of the employers (crefo) consist of dummy data and do not match the MUP data.

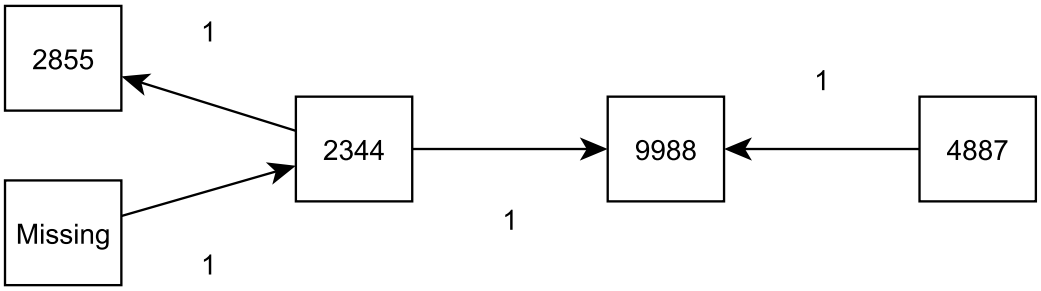


Figure D.1: **Example: A weighted, directed, and simple graph for the year 2015.** The nodes are the employers, and the edges model the employee flows. The edge weight is the number of employees moving between two employers. The figure is based on the flows from Table D.6. The identifiers of the employers (crefo) consist of dummy data and do not match the MUP data. Own illustration (created with yEd - graph editor: `https://www.yworks.com` [Last accessed: 24.02.2024]).

Table D.7: **Example: Degree centrality.**

Node Identifier	Degree	In-Degree	Out-Degree
Missing	1	0	1
crefo: 2855	1	1	0
crefo: 2344	3	1	2
crefo: 9988	2	1	1
crefo: 4887	1	0	1

Notes: Degree centrality scores for the network in Figure D.1. The metrics are calculated at the node level. The identifiers of the employers (crefo) consist of dummy data and do not match the MUP data.

D.3 Analysis of Employer, Employee, and Flow Data

Employee Flows by Employer Characteristics

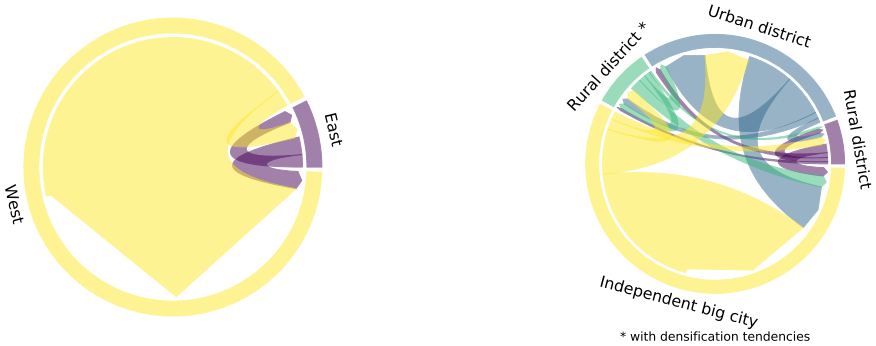


Figure D.2: **Employee flows by regions.** Left: Flows between East and West Germany. Data from Berlin are excluded because of the unclear assignment. Right: Flows between the different district types. Table D.10 presents the employee flow data as matrices. Source: TUM and ZEW based on XING and MUP data. Own illustration.

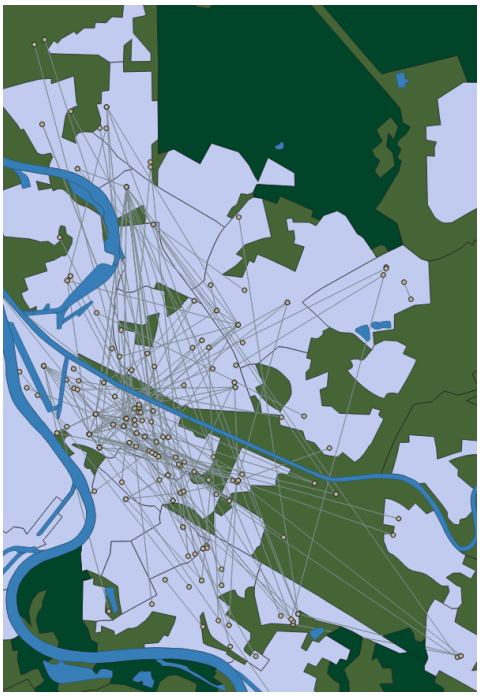
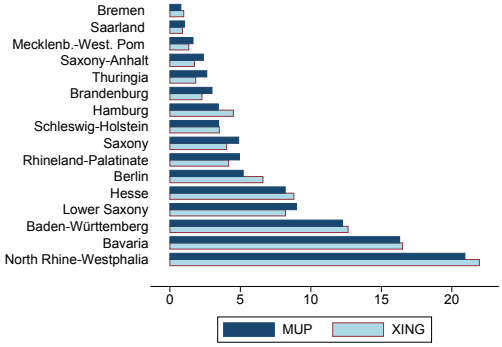


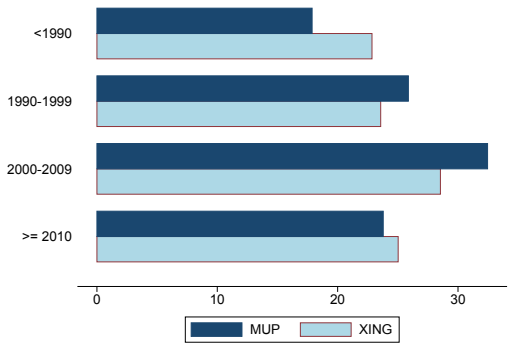
Figure D.3: **Employee flow network for the city of Mannheim.** The data are restricted to flows within Mannheim in 2019. Source: TUM and ZEW based on XING data. Own illustration (created with QGIS: `https://www.qgis.org` [Last accessed: 24.02.2024]).

Employers by Characteristics

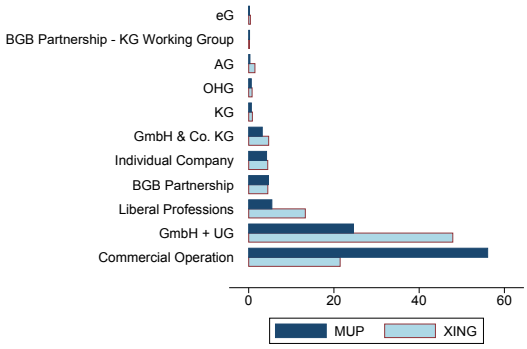
(a) **Employers by region**



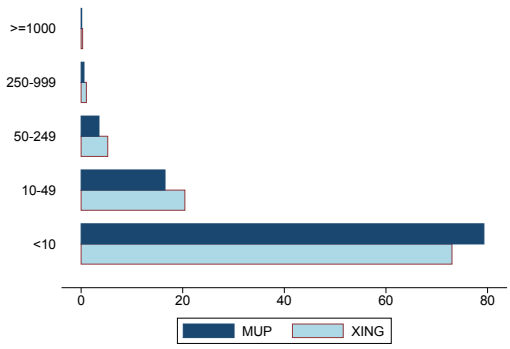
(b) **Employers by founding year**



(c) **Employers by legal form**



(d) **Employers by size**



(e) **Employers by industry**

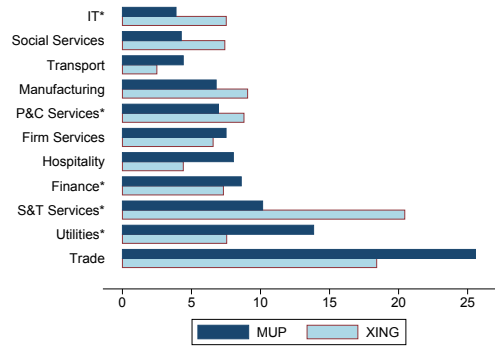


Figure D.4: **Share of MUP and XING employers with respect to employer characteristics (region, founding year, legal form, size, and industry).** The numbers sum up to one hundred percent for each data source. The MUP data are restricted to unique German employers that are listed from 2002 to 2019 (we use the latest entry). Missing values are not shown. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 and Table D.3 for full labels). Source: TUM and ZEW based on XING and MUP data. Own illustrations.

Employments by Characteristics

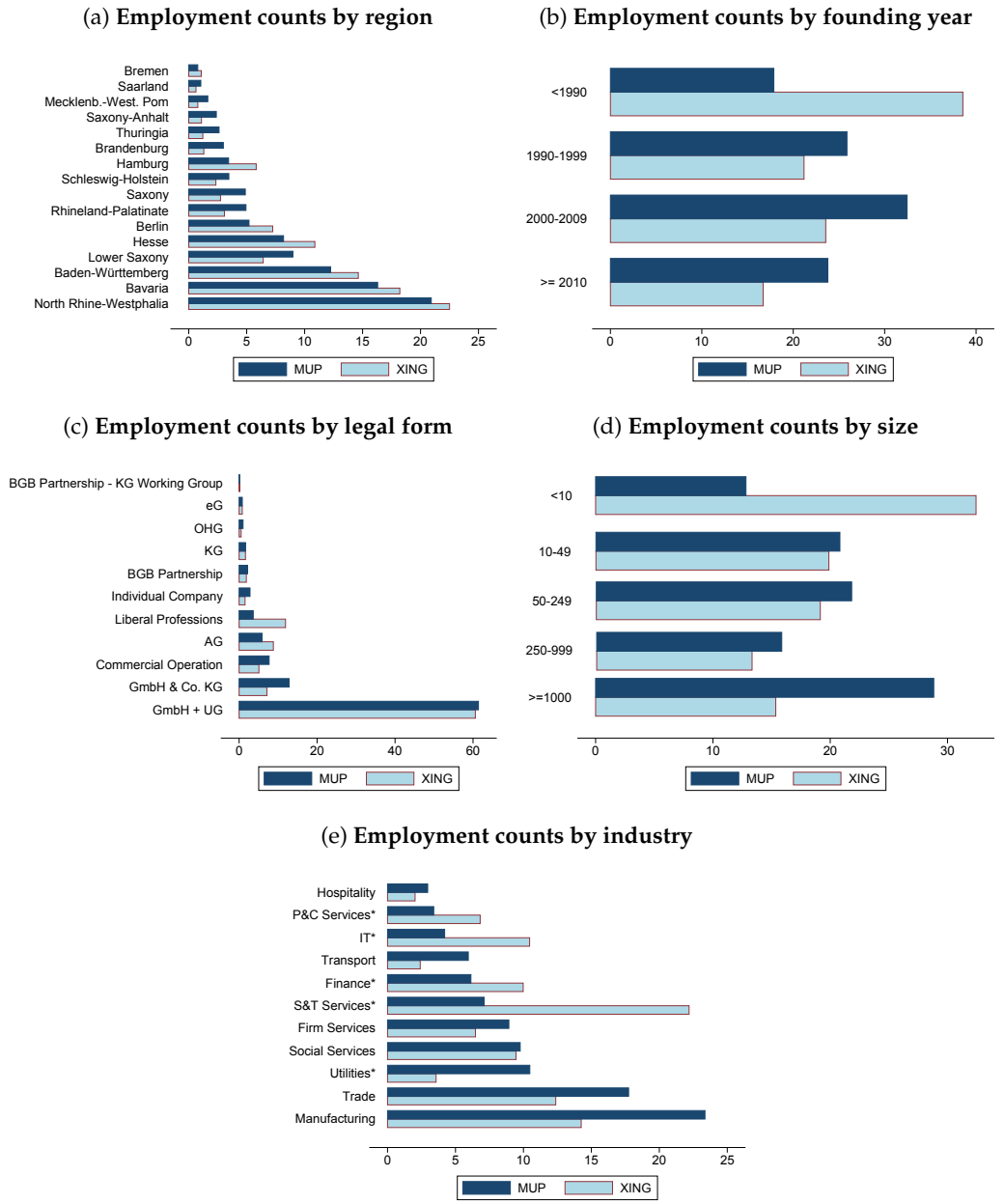


Figure D.5: **Share of MUP and XING employees with respect to employer characteristics (region, founding year, legal form, size, and industry).** The numbers sum up to one hundred percent for each data source. The MUP data are restricted to unique German employers listed from 2002 to 2019 (we use the latest entry for the employment counts). The data are not extrapolated and therefore include only a subset of all employments. Missing values are not shown. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 and Table D.3 for list of all labels). Source: TUM and ZEW based on MUP and XING data. Own illustrations.

Employee Flows by Characteristics

Table D.8: **Employee flow matrices: Discipline, employment type, and career level.**

Discipline		New employment										
		Administration	Consulting	Engineering	Finance	IT & Data	Management	PR & Marketing	Production	R&D	Sales	Social & Others
Old employment	Administration	**57.79**	4.71	2.12	4.26	3.40	6.01	4.20	1.41	2.09	8.69	5.32
	Consulting	6.59	**54.01**	2.30	4.52	7.16	9.00	3.91	0.57	2.68	5.40	3.85
	Engineering	2.64	2.20	**65.71**	1.05	4.10	7.44	1.11	2.71	7.19	3.00	2.85
	Finance	8.33	6.83	1.89	**59.65**	3.30	6.20	2.40	0.80	1.62	5.45	3.54
	IT & Data	2.57	4.87	3.18	1.45	**73.13**	4.55	1.98	0.53	3.23	2.47	2.05
	Management	6.34	8.10	6.18	2.98	5.53	**45.89**	7.38	1.70	3.21	8.23	4.46
	PR & Marketing	4.37	3.30	0.96	1.13	2.65	6.93	**68.04**	0.69	2.22	5.51	4.20
	Production	6.29	1.76	12.46	1.65	2.58	6.86	2.59	**47.16**	5.84	7.04	5.77
	R&D	3.66	4.27	12.09	1.25	6.49	6.48	3.45	1.85	**52.23**	2.63	5.60
	Sales	9.56	4.19	2.06	2.80	3.21	7.14	5.31	1.40	1.35	**57.73**	5.25
	Social & Others	8.88	3.89	3.86	2.56	3.29	5.26	5.84	2.48	4.84	8.42	**50.68**

Type		New employment						
		Civil Servant	Freelance	Fulltime	Intern	Partner	Parttime	Unpaid
Old employment	Civil Servant	**61.27**	4.66	25.46	2.54	1.04	4.03	0.97
	Freelance	0.10	**44.15**	35.68	4.91	5.57	8.34	1.21
	Fulltime	0.10	4.47	**84.64**	2.53	2.57	5.17	0.50
	Intern	0.03	5.14	34.46	**38.28**	0.68	19.49	1.88
	Partner	0.06	15.29	47.47	1.87	**28.79**	5.22	1.27
	Parttime	0.11	6.31	40.64	15.59	1.56	**34.35**	1.41
	Unpaid	0.20	9.42	33.30	20.14	2.38	18.20	**16.31**

Career level		New employment					
		Student/Intern	Young Professional	Professional	Manager	Director	Managing Director
Old employment	Student/Intern	**64.22**	21.27	11.83	1.73	0.22	0.70
	Young Professional	10.60	**23.65**	56.94	6.76	0.69	1.34
	Professional	2.71	3.42	**74.46**	15.16	1.57	2.65
	Manager	0.85	0.91	16.26	**63.40**	11.48	7.08
	Director	0.56	0.44	7.57	17.76	**50.72**	22.91
	Managing Director	1.91	1.70	17.59	18.11	14.66	**46.00**

Notes: Top: Discipline. Middle: Type. Bottom: Career level. The flows are based on matched and public employments. Each row has been normalised to sum up to 100 percentage points. Missing values are not shown. Source: TUM and ZEW based on XING data.

Table D.9: **Employee flow matrices: Industries, employees, and regions.**

Industry		New employment										
		Finance*	S&T Services*	Hospitality	Trade	Manufacturing	IT*	P&C Services*	Social Services	Transport	Firm Services	Utilities*
Old employment	Finance*	**56.62**	12.79	0.74	5.10	5.84	5.48	3.14	3.79	1.17	3.72	1.63
	S&T Services*	5.65	**52.86**	0.76	6.47	8.78	8.57	4.19	5.38	1.14	4.27	1.93
	Hospitality	5.53	13.28	**44.12**	6.48	4.87	4.27	5.73	6.41	1.69	6.30	1.32
	Trade	4.83	13.64	0.76	**48.50**	10.24	7.26	3.45	3.94	1.36	4.20	1.82
	Manufacturing	3.98	13.82	0.43	7.57	**57.20**	4.38	2.71	4.13	0.93	3.09	1.76
	IT*	4.62	15.65	0.41	6.08	4.84	**53.60**	4.74	3.99	0.84	4.16	1.07
	P&C Services*	5.15	16.61	1.27	6.13	6.85	9.94	**34.34**	11.59	1.29	5.14	1.68
	Social Services	4.31	15.64	0.92	4.65	7.28	6.31	7.84	**46.52**	1.01	4.02	1.50
	Transport	5.43	12.01	0.95	6.70	6.66	4.81	3.39	4.05	**48.82**	5.35	1.83
	Firm Services	5.84	15.83	1.26	6.85	7.27	8.44	4.84	5.59	1.91	**40.19**	1.97
	Utilities*	5.72	16.15	0.79	6.88	9.75	4.51	3.65	5.16	1.37	4.51	**41.52**

Employer size		New employment			
		<10	10-49	50-249	250-999	≥1000
Old empl.	<10	**50.97**	17.47	14.35	8.44	8.74
	10-49	23.11	**45.66**	15.25	8.15	7.81
	50-249	18.28	14.71	**49.62**	8.98	8.38
	250-999	15.83	11.53	13.14	**49.27**	10.20
	≥1000	14.81	10.11	11.13	9.24	**54.68**

Region		New employment														
		Baden -Württemberg	Berlin	Brandenburg	Bremen	Hamburg	Bavaria	Saxony	Thuringia	Hesse	Meckl.-West. Pomerania	Lower Saxony	North Rhine -Westphalia	Rhineland -Palatinate	Saarland	Saxony -Anhalt	Schleswig -Holstein
Old employment	Baden-Württemberg	**65.40**	2.89	0.42	0.37	1.98	9.35	0.97	0.81	4.81	0.21	2.05	7.58	1.81	0.33	0.37	0.65
	Berlin	5.49	**55.05**	2.57	0.44	4.26	8.87	1.51	0.54	5.32	0.53	2.54	9.90	1.22	0.19	0.62	0.96
	Brandenburg	5.76	18.35	**36.93**	0.52	3.27	7.82	3.03	0.81	4.57	1.00	3.17	10.67	1.33	0.25	1.14	1.37
	Bremen	5.47	3.43	0.48	**47.39**	6.84	6.78	1.04	0.42	4.24	0.54	10.77	9.11	1.03	0.16	0.47	1.83
	Hamburg	4.61	5.13	0.53	1.09	**56.13**	7.13	0.82	0.36	4.55	0.77	4.13	9.18	0.89	0.15	0.38	4.13
	Bavaria	7.38	3.72	0.46	0.36	2.52	**65.93**	1.22	0.57	4.83	0.29	2.15	8.20	1.07	0.19	0.43	0.68
	Saxony	6.07	4.84	1.16	0.44	2.17	8.92	**55.00**	1.56	4.06	0.46	2.72	8.38	1.13	0.26	2.04	0.81
	Thuringia	11.07	4.26	0.80	0.35	2.17	9.85	3.78	**45.60**	5.37	0.40	3.37	8.90	1.54	0.25	1.38	0.92
	Hesse	6.95	4.05	0.47	0.45	3.07	8.86	0.99	0.57	**56.92**	0.28	2.49	10.93	2.37	0.32	0.42	0.86
	Meck.-West.Pomerania	5.00	6.53	1.49	0.92	7.47	7.90	1.81	0.67	4.33	**43.51**	4.44	9.61	1.12	0.22	0.88	4.09
	Lower Saxony	5.49	3.62	0.60	1.98	4.90	6.87	1.17	0.64	4.47	0.51	**54.13**	11.69	1.15	0.19	0.89	1.69
	North Rhine-Westphalia	5.35	3.65	0.53	0.45	2.83	7.11	1.03	0.47	5.24	0.30	3.23	**66.75**	1.50	0.24	0.44	0.91
	Rhineland-Palatine	10.03	3.41	0.53	0.40	2.21	7.24	1.01	0.57	9.60	0.27	2.44	11.98	**47.93**	1.12	0.44	0.84
	Saarland	9.46	3.04	0.46	0.34	2.01	6.92	1.27	0.56	6.16	0.26	2.12	9.74	5.74	**50.65**	0.41	0.86
	Saxony-Anhalt	5.95	5.42	1.32	0.49	2.75	8.19	5.65	1.76	4.41	0.62	5.71	9.84	1.29	0.23	**45.16**	1.20
	Schleswig-Holstein	4.94	3.62	0.71	0.91	13.18	6.25	1.10	0.43	3.96	1.26	4.68	9.06	1.06	0.21	0.49	**48.14**

Notes: Top: Industry. Middle: Employer size. Bottom: Region. The flows are based on matched and public employments. Each row has been normalised to sum up to 100 percentage points. Missing values are not shown. Labels marked with a star (*) are listed in abbreviated form (see Table D.2 for the list of all labels). Source: TUM and ZEW based on XING data.

Table D.10: **Employee flow matrices: East/West Germany, and district type.**

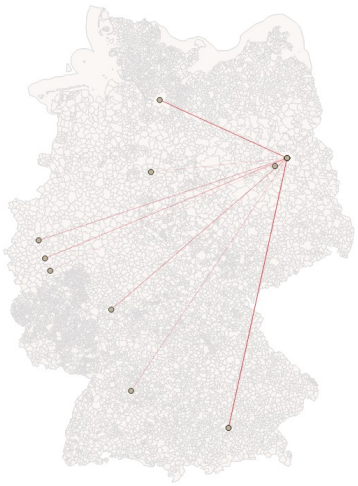
East-West		New employment	
		East	West
Old empl.	East	**41.10**	58.89
	West	4.48	**95.51**

District type		New employment			
		Rural district	Rural district*	Urban district	Independent big city
Old empl.	Rural district	**21.71**	10.76	23.93	43.58
	Rural district*	7.99	**21.73**	25.74	44.52
	Urban district	4.43	6.38	**42.62**	46.55
	Independent big city	3.94	5.33	22.73	**67.99**

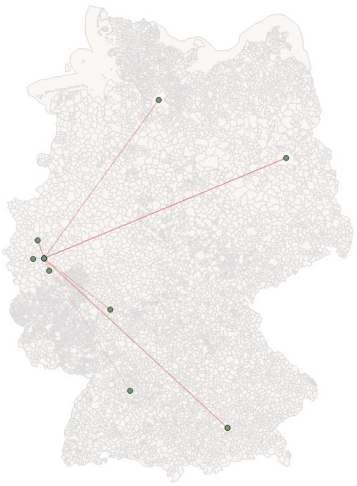
Notes: Top: East-West. Bottom: District type. The flows are based on matched and public employments. Each row has been normalised to sum up to 100 percentage points. The class *'Rural district*'* is the abbreviation for *'Rural district with densification tendencies'*. Missing values are not shown. Source: TUM and ZEW based on XING data.

D.4 Illustration of the Ten Districts with the Most Flows to or from Berlin, Cologne, Hamburg, and Munich

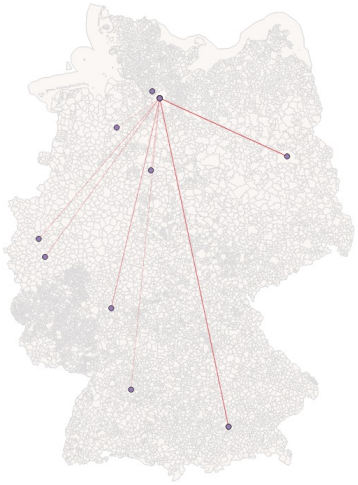
(a) **Flows for Berlin**



(b) **Flows for Cologne**



(c) **Flows for Hamburg**



(d) **Flows for Munich**

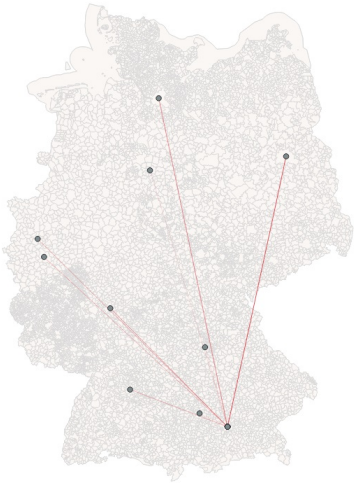


Figure D.6: **The top 10 districts with the most flows to or from the German cities of Berlin, Cologne, Hamburg, and Munich.** The edge thickness illustrates the relative flow count in the employee flow data set. The number of flows is normalised so that it adds up to one for the top 10 districts. Source: TUM and ZEW based on XING and MUP data. Own illustration (created with QGIS: `https://www.qgis.org` [Last accessed: 24.02.2024]).
Note: Some districts and cities are related, e.g. the city and district of Munich.

D.5 Mapping Employers from the Mannheim Enterprise Panel to Geographical Coordinates

Table D.11: **Example: Linking employee flows with coordinates.**

Steps	Observations	Description
1	870K	Load geo-referenced address data (longitude and latitude) for Germany from external data sources. Jan Kinne (colleague) kindly provided the data.
2	58K	Load XING employer data that are matched to the MUP. We restrict the data to employers located in Mannheim or Munich. The employers are selected based on district numbers 8222 and 9162.
3	57K	Load address data of matched XING employers. We remove all observations if no postal code or address data are available in the MUP.
4	57K	The MUP data contain postal codes and addresses. We split the addresses into street names and numbers (using a heuristic).
5	57K / 870K	Standardise the address text data. For example, transform texts into lowercase and treat special characters (ä, ü, ö, ß).
6	48K	Link 57K employers to 870K geo-referenced address candidates. High-quality matches: Check if the postal code, street name, and number are substrings of candidate addresses. Medium-quality matches: Check if the postal code and street name are substrings of candidate addresses. Merge both match data sets, but prefer high-quality matches over medium-quality matches. If there are several matches within one quality class, select the first one (heuristic). Result: Mapping of geo-coordinates to XING employers and employee flows.

Notes: Matching a subset of XING employers to geographical coordinates. The match is based on a heuristic and does not claim to be error-free. Number of employers in thousand (K).

D.6 Definition and Explanation of District Types

Table D.12: **Definition of district types.**

Type	Description
Rural district	Population share in large and medium-sized cities below 50% and population density excluding large and medium-sized cities below 100 inhabitants/km².
Rural district with densification tendencies	Population share in large and medium-sized cities of at least 50%, but a population density of fewer than 150 inhabitants/km²; a population share in large and medium-sized cities of less than 50% with a population density without large and medium-sized cities of at least 100 inhabitants/km².
Urban district	Population share in large and medium-sized cities of at least 50% and a population density of at least 150 inhabitants/km²; a population density without large and medium-sized cities of at least 150 inhabitants/km².
Independent big city	Independent cities with at least 100K inhabitants.

Notes: `https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Raumabgrenzungen/deutschland/kreise/siedlungsstrukturelle-kreistypen/kreistypen.html` (Source) [Last accessed: 24.02.2024].

D.7 Example: Calculation of Network Metrics

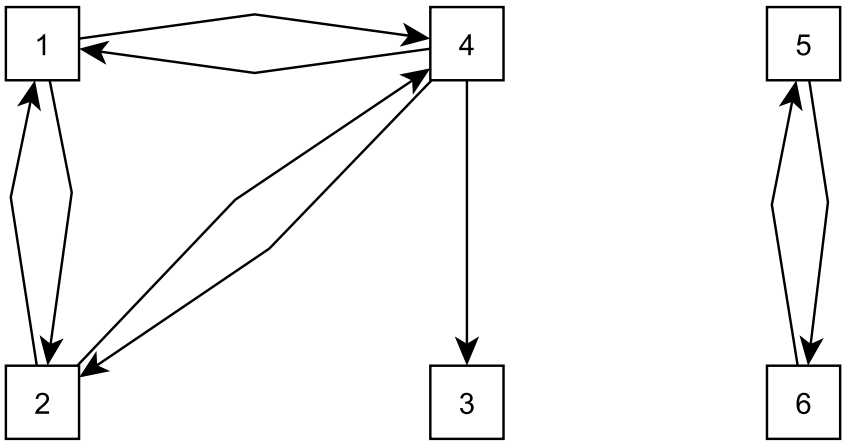


Figure D.7: **Example network and calculation of metrics:** #Nodes = 6, #Edges = 9, #Cliques = 12, Maximum clique size = 3, Transitivity = 0.777, Reciprocity = 0.888, #Clusters = 3, Density = 0.3, Girth = 3. Some *igraph* functions transform the directed edges into undirected edges, e.g. the calculation of cliques. Own illustration (created with yEd - graph editor: `https://www.yworks.com` [Last accessed: 24.02.2024]).