PARIS LODRON UNIVERSITY OF SALZBURG

DISSERTATION

Web-based innovation indicators for microgeographic economic analysis

---

Author:
Jan Kinne

Supervisory Committee:
Prof. Dr. Bernd Resch
Prof. Dr. Thomas Blaschke
Dr. Christian Rammer

Submitted in fulfillment of the requirements for the degree of
DOCTOR RERUM NATURALIUM (Dr. rer. nat.)

in the

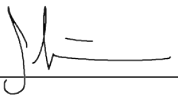
Faculty of Natural Sciences
Interfaculty Department of Geoinformatics – Z_GIS

July 6, 2020

Declaration of Authorship

I, Jan Kinne, declare that this thesis titled, "Web-based innovation indicators for microgeographic economic analysis" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: 

Date: July 6, 2020

"'We are living in the information age' is a popular saying; however, we are actually living in the data age. [...] An explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. Data mining has and will continue to make great strides in our journey from the data age toward the coming information age."

from "Data Mining: Concepts and Techniques", the first data mining book I read.

Abstract

The subject of this dissertation is the microgeographical analysis of firm locations with particular attention to the role of innovations. In the first two papers of this thesis (Chapter 2 and 3), detailed geodata are used for microgeographic mapping and econometric analysis of company locations and relevant location factors. The revealed drawbacks (lack of timeliness, coverage, granularity, and high data collection costs) of traditional innovation indicators based on surveys and patent data motivate the second part of this thesis, in which a novel approach for generating web-based innovation indicators is developed, tested and applied. The approach is based on a web mining framework, which relies on a purpose-built web scraping software (ARGUS) that is used to extract texts and hyperlinks from corporate websites (Chapter 4). These web data are then analyzed to find innovation-related information using data mining. The information thus obtained serves as the basis for a novel type of innovation indicators at the firm level, which can be collected on a large scale in a high-frequency, granular and cost-effective manner. Specifically, a "product innovator firm" prediction model based on deep learning and website texts is being developed, which uses firms from a traditional innovation survey as training data (Chapter 5). The proposed framework is also used for a web-based diffusion analysis, in which the dissemination of a information security standard (*ISO/IEC 27001*) is examined (Chapter 6). In the final paper (Chapter 7), the hyperlink networking of innovative and non-innovative companies on the Internet is researched and their relationship types are evaluated. This dissertation contributes to the understanding of microgeographic economic processes and furthermore develops a new methodological approach for measuring innovation. The latter in particular has a high societal relevance, since evidence-based policy-making depends on comprehensive and up-to-date indicators to successfully promote economic growth as well as to evaluate the effectiveness of economic policy measures.

Acknowledgements

My thanks go to the University of Salzburg and ZEW Mannheim as well as to the people they represent. In particular, I would like to mention my supervisors Bernd Resch and Thomas Blaschke from Salzburg. Bernd has been supervising me since my studies at the University of Heidelberg and I appreciate his openness towards all research directions and his ability to lead without being restrictive. I have generally appreciated this great openness and curiosity of the researchers in Salzburg. At the ZEW I would like to thank all my colleagues for their support and the friendly way they treat each other. I would like to especially emphasize Georg Licht and Christian Rammer, who both supported my admittedly quite daring research venture in the context of the TOBI project. I also received a lot of support from Knut Blind and Peter Winker. I would like to thank my co-authors Miriam, Janna and Mona for the great teamwork. Special thanks go to David Lenz, with whom I had more than one real "Eureka" moment and from whom I learnt a lot.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
CIS	**C**ommunity **I**nnovation **S**urvey
ESDA	**E**xploratory **S**patial **D**ata **A**nalysis
GIS	**G**eographical **I**nformation **S**ystem
GIScience	**G**eographical **I**nformation **Science**
ICT	**I**nformation and **C**ommunications **T**echnology
MAUP	**M**odifiable **A**real **U**nit **P**roblem
MUP	**M**annheimer **U**nternehmens **P**anel (Mannheim Enterprise Panel)
MIP	**M**annheim **I**nnovation **P**anel
NLP	**N**atural **L**anguage **P**rocessing
OSM	**O**pen **S**treet **M**ap
R&D	**R**esearch **&** **D**evelopment
STI	**S**cience **T**echnology **I**nnovation
tf-idf	**t**erm **f**requency - **i**nverse **d**ocument **f**requency

Für die Oma.

Chapter 1

Introduction

In this doctoral thesis I pursue the old question why companies locate in certain patterns and what influence the location of a company has on its evolution. In contrast to previous studies, I take a microgeographic perspective. This perspective has specific requirements on the data and methods used, but also promises previously unknown insights into the mechanisms behind the entrepreneurial choice of location and the interactions between companies and their surroundings. In this context, particular attention is paid to the role of innovations. This concerns both the different location choices of innovative and non-innovative firms and the role of location for the emergence of innovations in firms.

The contribution I am making with my co-authors regarding this dissertation is primarily of a methodological nature. Firstly, we use comprehensive microgeographic data for the first time to map company locations and analyze them with econometric and spatial-statistical methods. This aspect of my dissertation is described in *Part I: Microgeographic economic analysis*. The studies presented there revealed the shortcomings (lack of granularity, scope, timeliness and high costs) of traditional innovation indicators for microgeographical studies. In the second part of my dissertation *Part II: Web-based innovation indicators*, I and my co-authors work on the development and application of web-based innovation indicators. For this purpose, we propose a coherent framework based on web scraping of company websites and identify ways to generate and validate web-based innovation indicators at the company level.

This thesis consists of six mutually dependent scientific papers (the papers are included in the appendix) two of them in Part I and four in Part II, which are presented and summarized in the subsequent sections. Each paper is introduced by a short preface, which provides some theoretical background and ties the papers together. At the end of this thesis I give a comprehensive review of the results (*Part III: Synthesis*) and finish with a final conclusion.

Chapter 2

The Microgeography of Firm Locations

2.1 From Classical Location Theory to a Microgeographic Probability Grid

Location theory describes, explains and evaluates economic spatial systems, their geographic pattern (structure), interaction (relations), and dynamics (Schätzl, 2003). It is about the question of where, how and why certain kinds of economic activities are located. Given that "the location pattern of any industry is the product of a large number of individual decisions" (Smith, 1981), one of the main objectives is to investigate the location decisions of economic actors and the detection of determinants that trigger and influence these decisions (Haas and Neumair, 2008). These determinants are generally referred to as location factors. The scope of location theory can reach from the individual plant to industrial sectors to whole territorial production systems. Each of these different levels of analysis come along with unique requirements to the used data, the theoretical and methodological approach.

The work of Johann Heinrich von Thünen (Thünen, 1842) on the geographic variations of agricultural activity is widely regarded as the “first serious treatment of spatial economics” (Essletzbichler, 2011) and von Thünen has been called "the father of location theorists" (Isard, 1956). Von Thünen conceptualised a spatially varying location rent that causes agriculture activity to vary over geographic space. This location rent is the maximum rent that farmers are willing to pay for land, given constant costs of production, constant transportation cost per distance unit, constant price of the agricultural product and constant quantities produced per area unit of land. The profit of the same farming activity conducted at different locations is only determined by the costs to transport the produced goods to the market - the only location factor in this simple model. Hence, every location is associated with a certain location rent, which is a function of the location’s distance to the market. Given that the farmers act as rational homo economicus, they will produce each agricultural

product at its optimal location only. The optimal location for each product is given where the product's location rent is the highest of all alternative products. This results in concentric rings of differentiated agricultural land use around the model's hypothetical central city: the so-called *von Thünen Rings*.

Transportation costs remained the single most important location factor in "classical Germanic location theory" (Leyshon et al., 2011) with their focus on abstract and formalized models (Capello, 2014) in the nineteenth and early twentieth century. Weber (Weber, 1922), for example, used a *point of minimal transportation costs* between raw material markets and a final goods market to determine the optimal location for industrial production sites. However, Weber also considered other location factors such as spatially varying wages and dispersed raw materials. More importantly even, he incorporated *agglomeration economies*, which describe the (dis-) advantage that occur when economic activity (i.e. many firm locations) is concentrated geographically. Agglomeration economies have since become one of the main interests for location theorists, especially when investigating the interplay between firm location and innovation. Agglomeration economies have been subdivided into *localisation economies* and *urbanisation economies* (see for example Bathelt and Glückler, 2012). Localisation economies are believed to occur if firms of the same industrial sector cluster geographically, which causes knowledge spillovers and a larger pool of skilled labour and suppliers. Urbanisation economies, on the other hand, are the general advantages generated by the clustering of different industrial sectors, which results in inter-sectoral integration, high-quality infrastructure and a numerous, diverse workforce. The introduction of agglomeration economies in location theory adds a dynamic location factor where previous firm decisions can create a self-reinforcing momentum, which shapes the geographic pattern of firm locations and influences the future location decisions of other firms.

Very much since the early days of location theory, there has been a conflict between two opposing, yet sometimes complementing schools of thought that argue about the value or non-value of simplified and formalized location models. "A strong divergence of opinion between those who advocate rigorous formal models based on a series of limiting assumptions that seek to express regularities in mathematical terms, and those who deny that such models and technical tricks can explain why specific places change and develop in particular ways" (Leyshon et al., 2011). Smith (Smith, 1981), for example, states that "the main thrust of the critique may be summarized as follows. First, the preoccupation with optimally characteristics of conventional economic theory fails to recognize the evidently sub-optimal nature of much plant location practice, when judged against such criteria as cost minimization or the maximization of profits. Second and related too this, is the fact that theory [...] assuming as it does some idealized conception of omniscient 'economic man'

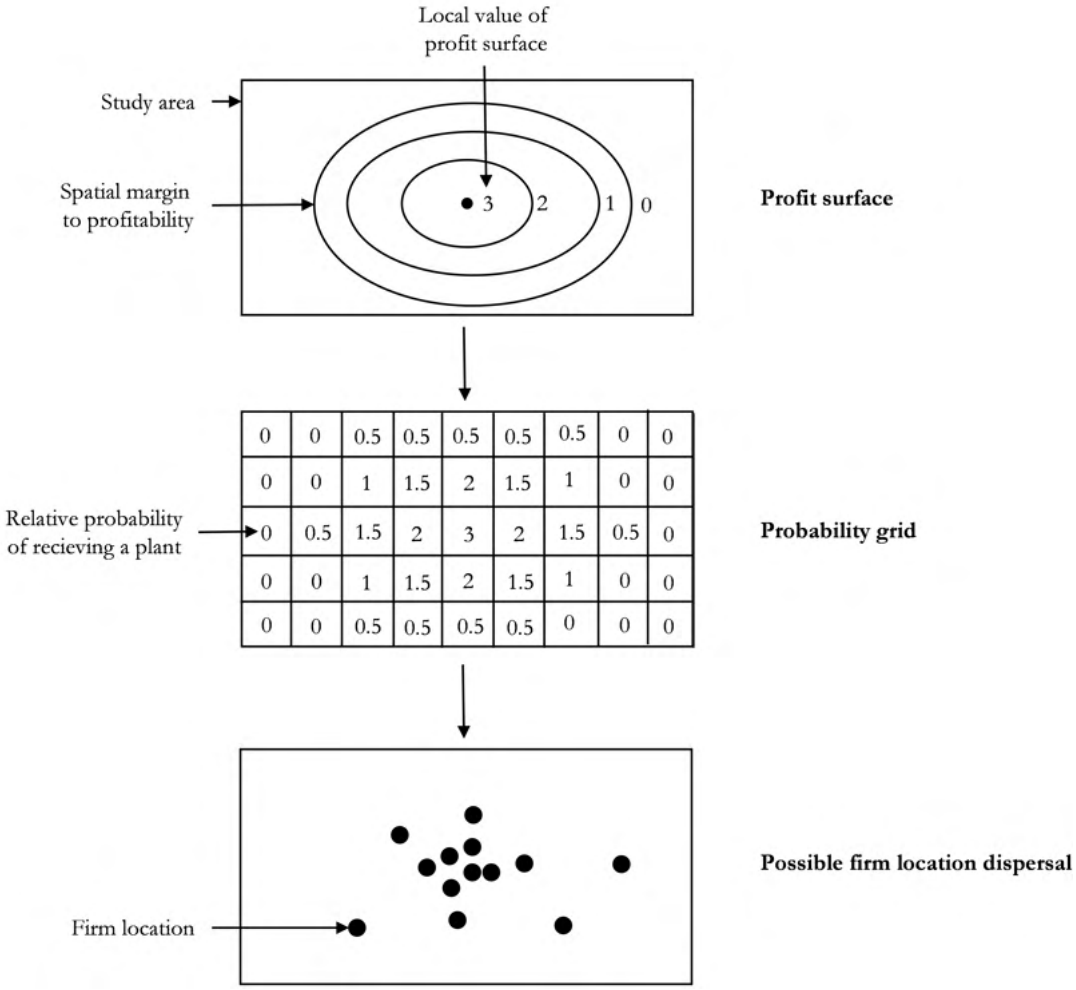
neglects to explore actual human behaviour". He recommends "a 'satisficing man' [who] could replace the all-knowing, perfectly able and rationale homo economicus of traditional theory". This satisficing man is characterised by satisficing behaviour and bounded rationality (see for example Pred, 1969). Such behaviour allows for location decisions that differ from the optimal location as it is predicted by deterministic location models.

Smith proposed the concept of a continuous spatial variable titled the *spatial margins of profitability* or *profit surface*, which determines the geographic space where a firm can (or cannot) make a profit. In this setup, firms may indeed choose a location outside the profitable areas, but they will not be able to make a profit nor be economically viable in the long run. Given that the location pattern of any industry is the product of a large number of individual decisions made by the firms of that industry and they "will tend to locate as near to the optimum location as knowledge and skill permit" (Smith, 1981), the overall firm location pattern will mostly trace the considered industry's spatial margins of profitability. This process of firm formation can be understood as a stochastic process taking place in the *spatial probability framework* of the industry which can be modelled as a *probability grid* where each cell is characterized by a certain probability to attract a firm. This probability can be derived from the location factors present in the respective cell. Figure 2.1 illustrates this concept.

The challenge of this approach is the allocation of the probability scores which requires knowledge of the profit surface. "Each cell accumulates scores according to the presence of conditions thought likely to attract a plant, perhaps according to their relative importance, and the aggregate becomes the probability value. [...] The drawback of this approach is the subjectivity involved in assigning scores to the attributes in a manner that supposedly reflects their relative attraction to industry" (Smith, 1981). In my first paper, which is summarized in the following section, I used microgeographic data and regression analysis to empirically estimate the profit surface of the Germany software industry in order to assess the relative importance of a range of location factors.

2.2 Micro-Location Patterns of Software Firms

In my first paper "Analyzing and Predicting Micro-Location Patterns of Software Firms", which I co-authored with Bernd Resch and which has been published in ISPRS International Journal of Geo-Information volume 7 issue 1, we combined open geodata, Volunteered Geographic Information (VGI), and a comprehensive firm dataset (the Mannheim Enterprise Panel (MUP)) containing about three million firm observations to empirically estimate a industry-specific probability grid in the sense of Smith (Smith, 1981, see previous chapter). We motivated our study



0	0	0.5	0.5	0.5	0.5	0.5	0	0
0	0	1	1.5	2	1.5	1	0	0
0	0.5	1.5	2	3	2	1.5	0.5	0
0	0	1	1.5	2	1.5	1	0	0
0	0	0.5	0.5	0.5	0.5	0	0	0

FIGURE 2.1: Schematic representation of a profit surface, a corresponding probability grid, and a possible firm location pattern. From Kinne, 2016.

with a general interest to the society because "a thorough understanding of the impact of location factors on firms' location decisions and firm performance can have important implications for stakeholders. Managers and entrepreneurs can integrate valuable information into the decision making process when choosing the location of a new venture. Policy makers at the regional, national, and multinational level want to promote economic growth by developing the right location factors to create a beneficial environment for firms" (Kinne and Resch, 2018). Even though we acknowledged the long-standing study of industrial location research, we argued that our study would, for the first time, study a nation-wide firm location pattern at the microgeographic level. Given that the location factors that we analyzed may

vary in their effect direction and strength, depending of the level of analysis, an effect known as the Modifiable Areal Unit Problem (MAUP, see for example Bluemke et al., 2017; Manley, 2014), our study was also intended as a robustness check to previous studies that used aggregated spatial units. We defined the following two research questions for our work:

1. Are the effects of location factors, as reported by previous studies using aggregated spatial units, robust at the microgeographic level?

2. How does a firm location prediction model perform at the microgeographic level and to what degree does it provide valuable new insights into the firm allocation process? What are the distinct requirements to the data and the statistical model?

At the methodological level, we first applied a Exploratory Spatial Data Analysis (ESDA) and then, in a consecutive step, fitted a Poisson regression model to a one kilometer grid where each grid cell contained the values of 24 different location factors. These location factors included agglomeration, infrastructure, socioeconomic, topographical, and amenity location factors which we derived from microgeographic data. The microgeographic data based mainly on OpenStreetMap (OSM) and official open data from statistical agencies. We focused on the software industry, which we argued to be rather unrestricted in its location decisions (Möller, 2018), inducing only little bias from unobservable location factors like local zoning restrictions. Figure 2.2 illustrates the analysis grid and the data we used.

In the initial ESDA, we found that Poisson regression is likely to be an appropriate method to model the pattern of software firms aggregated at a regular 1-km grid. Further, we found that software firms seem to be an urban phenomenon, as they are disproportionally frequent in and around urban areas and even form statistically significant hotspots in some city regions. We further concluded that the regional settlement structure (polycentric vs. monocentric) seems to have an impact on the location pattern of software firms. After fitting a Poisson regression model, we interpreted the estimated regression coefficients (see Table 2.1) to deduce the relationships between the location factors and software firm counts per cell. In a final analysis step, we investigated the model fit and adequacy using several goodness-of-fit measures and a spatial residual analysis. The latter was intended to unveil geographic areas of systematic overestimation or underestimation (i.e. significantly clustered regression residuals) of our regression model.

We concluded that the microgeographic level of analysis provided new insights into the firm site selection process, but also that most location factors are scale robust and that our findings with respect to location factor effects are in line with prior research using aggregated spatial units. We also concluded that our microgeographic

TABLE 2.1: Location factors and estimated coefficients with robust standard errors in parentheses. Number of software firms per cell as dependent variable. From Kinne and Resch, 2018.

Location Factor	Description	IRR
	Agglomeration Location Factors	
Firm density	Number of local firms (in 10)	1.028 *** (0.003)
Firm density^2	Squared number of local firms (in 10)	0.999 *** (0.000)
High-tech firms	Proportion of high-tech firms in local stock of firms (in %)	1.021 *** (0.000)
Major firms	Distance to next major firm in km	0.998 *** (0.000)
Commercial rent	Difference local rent to mean rent in neighborhood (in Euro)	1.127 *** (0.12)
Population	Population per cell (in 100)	1.081 *** (0.003)
Population^2	Squared population per cell (in 100)	0.999 *** (0.000)
Population centrality	Urban Centrality Index (in 0.1 UCI) high value ≙ monocentricity	1.079 *** (0.192)
	Infrastructure Location Factors	
Broadband Internet	Availability of ≥50 mb Internet (categories) high value ≙ low availability of Internet	0.764 *** (0.009)
Motorway	Distance to nearest motorway access (in km)	0.977 *** (0.001)
Railway	Distance to nearest main-line railway station (in km)	0.998 *** (0.000)
Airport	Distance to nearest main airport (in km)	0.998 *** (0.000)
Public transport	Weighted count of public transport stops	1.000 (0.001)
	Socio-economic Location Factors	
Wages	Median income of full time employee (in 100 Euro)	1.005 (0.003)
Universities	Distance to nearest university (in km)	0.980 *** (0.000)
Research institutes	Number of research institutes	1.004 (0.036)
Educated workforce	Proportion of graduate employees in %	1.063 ***(0.006)
Students	Proportion of students in local population in %	0.986 *** (0.003)
Business tax	Business tax factor (in 100) high values ≙ high taxes	0.925 ** (0.023)
	Quality of Life and Amenities Location Factor	
Life expectancy	Mean life expectancy of population	1.092 *** (0.012)
Crime	Violent and street crime incidents per 1000 inhabitants	1.021 (0.015)
Recreation	Number of recreational, community, and sports facilities	1.056 *** (0.008)
Culture	Number of cultural facilities	1.015 0.017
Leisure	Number of gastronomy, nightlife, and general leisure facilities	1.002 (0.002)
	Other	
Terrain	Difference in elevation to mean neighborhood elevation (in 100m) high values ≙ hillside location	0.919 *** (0.004)
Geocoding control variable	Geocoding match rate (in %) high value ≙ high completeness	1.018 *** (0.002)

** $p \geq 0.01$, *** $p \geq 0.001$.

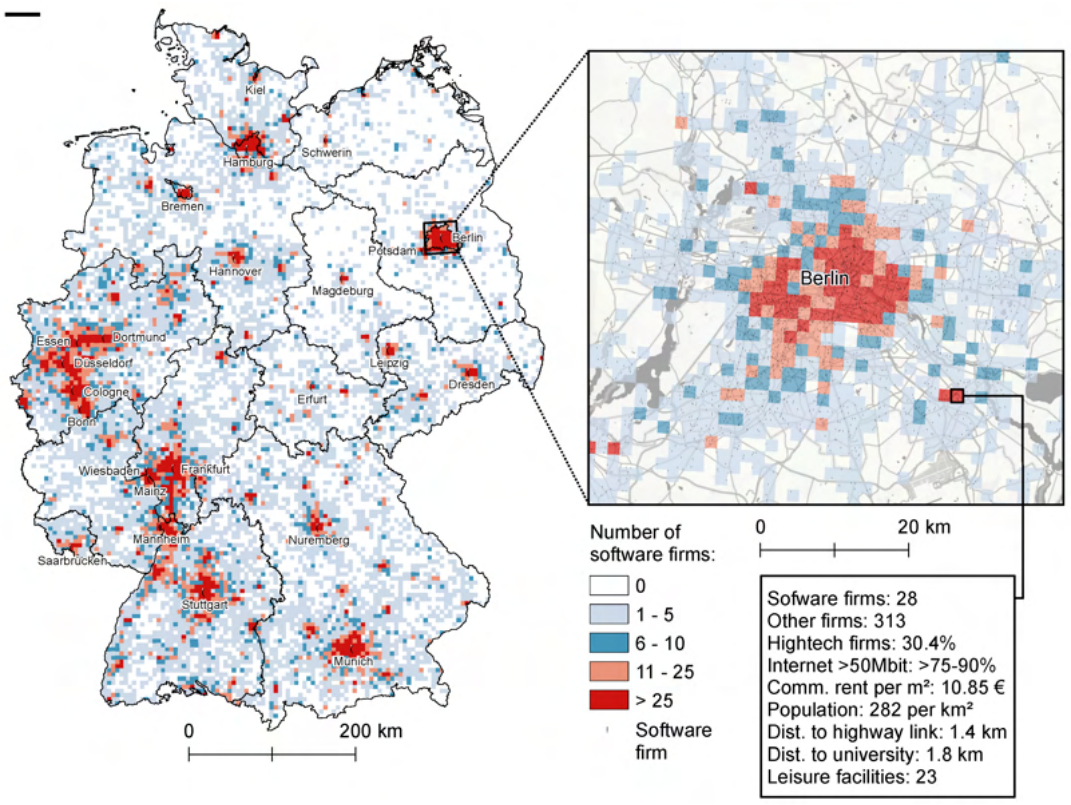


FIGURE 2.2: Overview (5 km scale) and zoom (1 km scale; with selection of location factors for exemplary cell) of the software firm location pattern. From Kinne and Resch, 2018.

prediction model is able to predict the location of software firms to a satisfying degree. However, we also pointed out the particular requirements to the statistical model and the data employed in a microgeographic location analysis, like the need for high resolution geodata, which was not available in all domains. We showed that this problem was most severe in cities, which often feature segregated populations and districts with very different socio-economic profiles. We discovered that this low resolution geodata lead to systematic prediction errors in cities. In the second paper of my dissertation, I therefore addressed specifically the question of company locations within cities, using completely disaggregated geodata.

Chapter 3

The Microgeography of Innovation

3.1 Urbanisation Economies and Knowledge Spillovers

Agglomeration economies are generally regarded as one of the most important location factors (see for example Rosenthal and Strange, 2004; Ahlfeldt and Pietrostefani, 2017; Smith and Florida, 1994) and also in my first paper (see previous chapter) agglomeration economies (operationalized using six different measures) showed a significant correlation with the location pattern of software companies. Especially in the context of innovation, which has long been recognized as a key element driving economic growth (Schumpeter, 1942), agglomeration economies and (urban) density in general are believed to play an important role. Close proximity to other firms and knowledge sources like scientific institutions are believed to provide firms with an convenient access to ideas, knowledge, and technologies. This concept of knowledge diffusion between nearby economic actors is commonly referred to as (knowledge) spillovers. Such spillovers may result from observing peers and competitors, the sharing of informal or tacit (i.e. non-codified) knowledge, personal and sometimes unintentional contacts, or from the fluctuation of workers and managers between firms (Audretsch, 1998; Glaeser, 1999). Last but not least, geographical proximity is also believed to increase trust and thus lower transaction costs between business partners through more frequent face-to-face communication, mutual understanding and a common background with shared cultural or institutional values (Florida, Adler, and Mellander, 2017; Porter, 1998).

Naturally, cities are the ideal location for this kind of close and frequent interaction between diverse economic actors and the concept of urbanisation economies, as it has been described in Chapter 2, refers to exactly this. Hereby, knowledge spillovers have long been identified as a key concept and one of the major drivers for the emergence, growth, and success of cities (Duranton and Puga, 2004; Marshall, 1890). There is also empirical evidence that spillovers increase the innovation performance of cities (Audretsch and Feldman, 2004; Simmie, 2002; Henderson, 2007) which may be a reason for the increasing importance of cities in a human society that is relying more and more on a knowledge-based economy (Bettencourt, 2013;

Helbing et al., 2007; Caragliu et al., 2015). A whole strain of literature is also dealing with the importance of amenities such as gastronomy, recreational facilities, and green spaces for firm location choice in a knowledge-based economy which relies heavily on *creative workers* who have a strong preference for a rich social and cultural life (Möller, 2018).

A key but yet little-explored aspect is the exact geographic scale at which proximity can stimulate knowledge spillovers and eventually innovation. Research on the benefits of personal interaction on the development of innovation within single institutions indicated that these benefits decay rapidly with distance and are of a truly microgeographic scope (Catalini, 2018; Kabo et al., 2014). In my second paper, my co-authors and I examined the extent to which proximity to urban knowledge sources (other companies, universities, etc.) is related to the innovation activity of companies. The results of this research are presented in the next section.

3.2 Knowledge Proximity and Firm Innovation

In my second paper "Knowledge Proximity and Firm Innovation: A Microgeographic Analysis for Berlin", which I co-authored with Christian Rammer and Knut Blind, and which has been published in Urban Studies volume 57 issue 5, we investigate the disaggregated location pattern of innovative and non-innovative firms in Berlin and their proximity to knowledge sources. The German capital of Berlin was chosen because of its insular geographical location in a otherwise rather sparsely populated German East, which makes it an ideal locally self-enclosed investigation area for a microgeographical study. As data, we used about 8,000 street-level geocoded Berlin-based firms from the Mannheim Enterprise Panel (MUP) that participated in the *Berlin Innovation Panel* survey. The Berlin Innovation Panel is conducted as part of the Mannheim Innovation Panel (MIP) which is the German contribution to the European-wide Community Innovation Survey (CIS). The CIS is a questionnaire-based survey which covers a sample of firms from manufacturing and business-oriented services sectors with at least five employees (see Peters and Rammer, 2013). In the survey, the firms are asked a range of questions concerning their innovation activities. The questions range from whether and what kind of research and development (R&D) they are conducting to whether they introduced new or significantly improved products or services. Thereby, the CIS follows a definition of innovation as it is laid out in the *Oslo Manual* (see OECD and Eurostat, 2018 for the most recent version). In addition to this data, from which we derived the locations of innovative and non-innovative firms in the years 2012-2016, we created a database on the location and size of university (campuses) and research institutes in Berlin. Figure 3.1 maps the locations of both innovative and non-innovative firms as well as university

campuses and research institutes that we used in this study. Figure 3.2 showcases the firm location data and the firms' corresponding innovation statuses derived from the CIS.

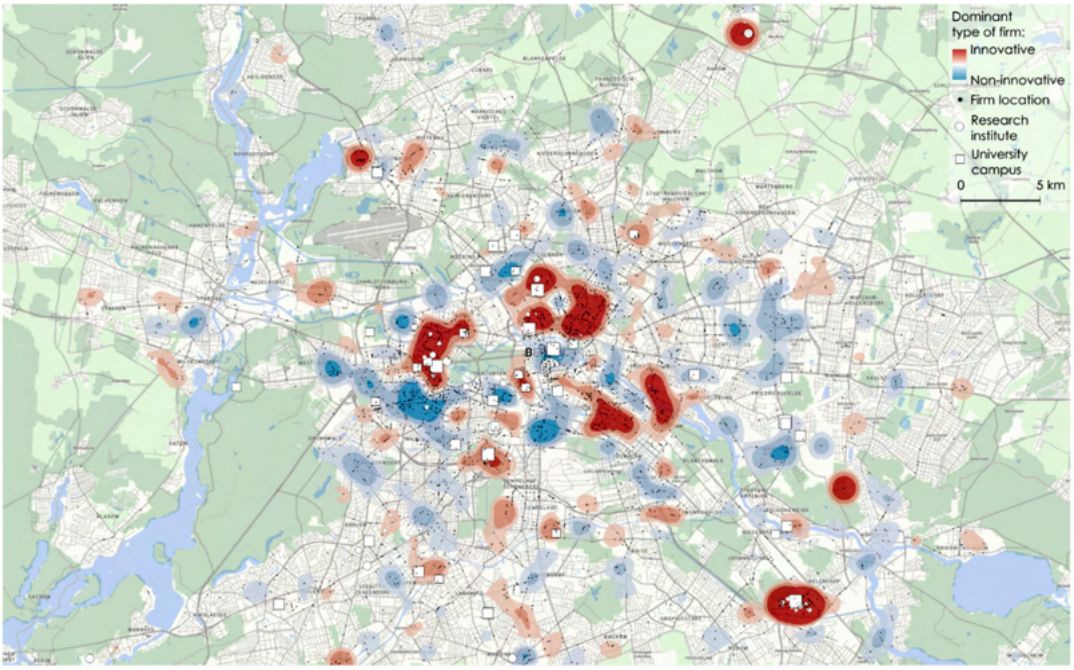


FIGURE 3.1: Innovative and non-innovative firms and knowledge sources in Berlin. From Rammer, Kinne, and Blind, 2020

We created four indicators to describe the local knowledge environment of a firm at a point in time *t*:

1. *Academic knowledge sources*: The number of students and researchers from research institutes in a firm's neighbourhood.

2. *Local buzz*: The number of new firms (both startups and firms that moved from other locations) in a firm's neighbourhood.

3. *Micro-clusters*: The number of neighbouring firms from the same industrial sector as the considered firm.

4. *Innovation dynamics*: The number of neighbouring firms that changed their innovation status from *t-1* (including new firms, closed firms, moving-in and moving-out firms).

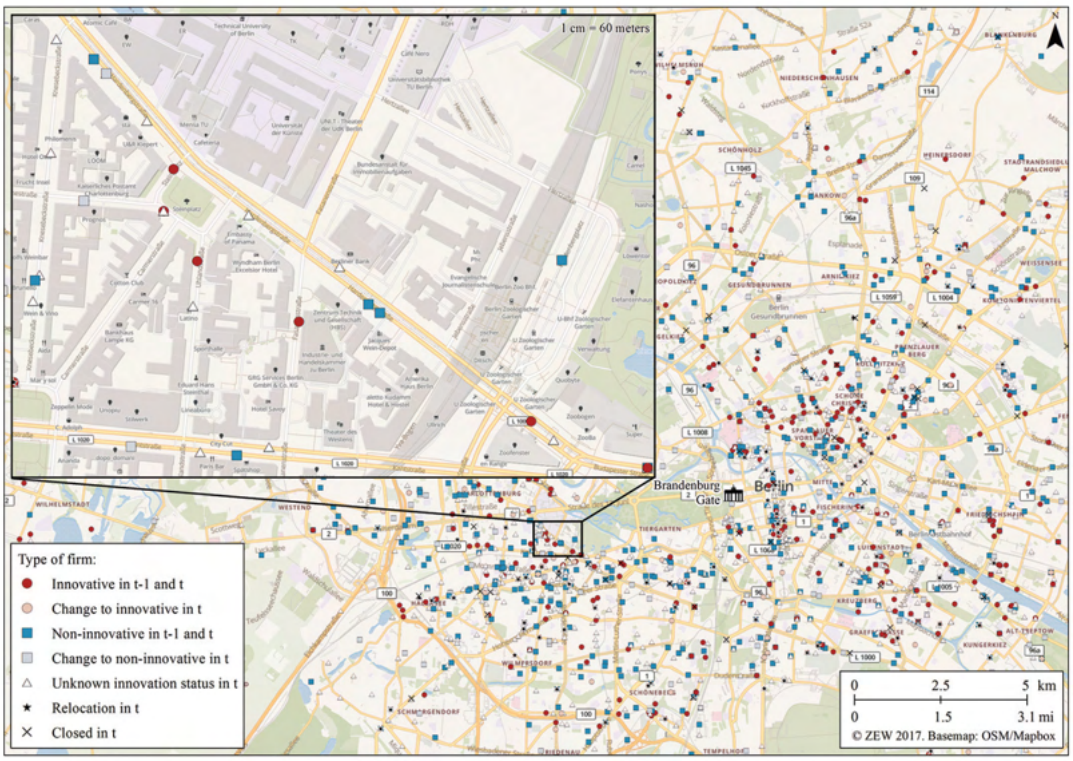


FIGURE 3.2: Example for the geographic distribution of firms in Berlin by innovation status. From Rammer, Kinne, and Blind, 2020

We then used six different distance thresholds (50, 100, 250, 500, 1000, and 2500 meters from the location of the firm under consideration) to define the neighbourhood of each company in our dataset. This approach allowed us to study the differences between innovative and non-innovative firms[1] in relation to their environment under different scale regimes. For this purpose, we used a matching approach (Heckman, Ichimura, and Todd, 1997) in order to compare innovative firms with non-innovative ones that share the similar basic characteristics (age, size, sector) so that differences in the local knowledge environment cannot be attributed to these characteristics. In addition to a descriptive statistical analysis, we also used a Probit regression model to analyse the relationship between a firm's innovation status (dependent variable in the regression analysis) and its knowledge environment (explanatory variables) as they were described above. In the regression analysis, we also controlled for the firm's size, age, and sector.

[1]Following the definition from the so-called Oslo Manual (OECD and Eurostat, 2018), a firm is considered a product (process) innovator if it introduced a new or significantly improved product (in-house process) within the past three years. Such innovations can be further divided into new-to-the-world, new-to-the-market, or new-to-the-firm novelties. Another information from the CIS that can be used as an innovation indicator is whether a firms conducts (continuous) in-house R&D. In our Berlin dataset, 39% of firms were classified as innovators, 5.4% introduced market novelties, and 21.2% conducted in-house R&D continuously.

We found that innovative, opposite to non-innovative firms, are located in places with a much higher *local buzz* and a higher density of *academic knowledge sources*. One of our main findings was that the geographic scope of these differences seems to be very confined. Concerning the proximity to research institutes, for example, the concentration of innovative firms already decreases beyond a 50 m radius. Beyond a distance of 1 km, we did not find a significant relation anymore. For firms with market novelties or continuous in-house R&D, close proximity (up to 250 m) to other firms from the same sector was a distinctive feature as well.

We concluded that our study could be seen a first attempt to zoom into the role of knowledge proximity for innovation in the city at a microgeographic level. However, we also added for consideration that we refrained from examining a more comprehensive model of the urban environment in which the firm operate, e.g. by including urban infrastructure, density or amenities in our analysis. We did also emphasize that, given our empirical setup, we could only observe correlations between innovation and local knowledge sources and no strictly identified causal relationships. The limitation of our study to the firm types that are covered in the CIS and the geographical restriction to a single city were other limitation that we discussed and eventually addressed in the second part of my dissertation.

Chapter 4

A New Generation of Innovation Indicators

4.1 Shortcomings of Traditional Innovation Indicators

In the first two papers of this dissertation we have shown that the microgeographic perspective offers new insights into entrepreneurial location selection and the associated location factors. We could also show that microgeographical analyses are accompanied by special demands on the used data and methods. For the modelling of "hard" location factors (i.e. infrastructure), we were able to rely in particular on high-resolution VGI data. In contrast, the availability of high-resolution "soft" location factors proved to be very limited in some cases. For example, we were still able to model agglomeration economies for the whole of Germany using the local density of company locations from the MUP company data set, whereas comprehensive information on the innovation activity of companies was only available for Berlin (derived from the singular Berlin Innovation Panel). An extension of our microgeographic study to other regions was not possible due to the complex and costly collection of innovation information via questionnaire-based surveys. Patents (applications, citations, and licensing) are often used as alternative innovation indicators at the firm-level. However, patents as innovation indicators also have some known shortcomings. This applies in particular to their very varying importance within some industries. For example, while companies in the pharmaceutical and mechanical engineering sectors often register patents in an attempt to protect their inventions, patents hardly play a role for companies in the information and communications technology (ICT) sector, as software is very difficult to patent in Europe.

In the second part of my dissertation I therefore focused on the development of a new generation of web-based innovation indicators, which are supposed to be available nationwide and thus suitable for comprehensive microgeographic analyses. These novel innovation indicators are based on the fact that companies, increasingly, leave digital traces that can be collected and analysed. An effective utilisation of these digital data has only become possible in recent years thanks to major

methodological advances in the field of data mining, especially Natural Language Processing (NLP). The adaptation of these novel methods for the purpose of extracting innovation-related information from web data, their transformation into innovation indicators at the firm-level, and their subsequent use in microgeographic analyses form the core of the second part of this dissertation. In the third paper of my dissertation, which will be presented in the next chapter, we develop and apply a coherent analysis framework for the generation of innovation indicators from company websites.

4.2 A Framework for Web-based Innovation Indicators

In my third paper "Web Mining for Innovation Ecosystem Mapping: A Framework and a Large-scale Pilot Study", which I co-authored with Janna Axenbeck and which is currently in minor revision status at Scientometrics, we develop and test a web mining framework for the generation of firm-level innovation indicators. We motivate our approach with the need to accurately measures innovation due to its overall importance as a key driver of economic growth. Measuring innovation activities to a sufficient degree of accuracy allows researchers and policy makers to analyze driving factors as well as the effectiveness of innovation policies. We invoke that there is evidence that traditional innovation indicators from questionnaire-based surveys and patent data struggle to provide a timely and sufficiently granular picture of the current state of innovation ecosystems (see for example Nagaoka, Motohashi, and Goto, 2010; Squicciarini, Dernis, and Criscuolo, 2013). The German MIP (which has been introduced in the previous chapters already), for example, covers about 10,000 firms every year, which corresponds to only 0.3% of the total number of firms in Germany. The total number of innovative firms remains unknown and can merely be estimated through statistical extrapolation. Furthermore, rare but potentially important innovation activities happening in unobserved sectors or technological fields may not be covered in the data. This also affects the analysis of geospatial innovation processes, some of which happen to operate on a fine (micro-)geographical scale (see previous chapter). Additionally, questionnaire-based surveys – especially when conducted on a large scale – are costly and time intensive. Indicators constructed from patents (patent applications, citations, licensing), on the other hand, cover only technological progress for which legal protection has been sought (Archibugi and Planta, 1996). Patent-based indicators also suffer from insufficient timeliness due to the time lag between patent priority date and the information becoming publicly available which usually is more than a year (OECD, 2009). The less popular literature-based innovation output indicators which are constructed by counting innovations reported in scientific, technical, or trade journals, on the other hand, do

not capture in-house innovations and are believed to under-represent innovations in smaller firms (Acs, Anselin, and Varga, 2002).

We summarized the shortcomings if traditional innovation indicators using the following four aspects:

1. *Coverage*: They cover only a fraction of the overall firm population.
2. *Granularity*: They suffer from insufficient sectoral, technological, and geographical granularity.
3. *Timeliness*: They capture the state of the STI (science, technology, and innovation) system as it was months or even years before.
4. *Cost*: They involve high data collection costs, especially when conducted on a large scale.

We then argue that web mining, the application of data mining techniques to uncover relevant data characteristics and relationships (e.g. data patterns, trends, correlations) from unstructured web data (see for example Askitas and Zimmermann, 2015), may be used to generate a new generation of innovation indicators from web data. We proposed firm websites as a particularly interesting data source for this purpose. Firms use their websites to present themselves, as well as their products and services. The information found on these websites can be used to assess firms' products, services, credibility, achievements, key personnel decisions, strategies and relationships with other firms (Gök, Waterworth, and Shapira, 2015). Even though surveying firms using their websites instead of conducting interviews or questionnaires or using other traditional methods offers some clear advantages (scale, cost, timeliness of the survey), it also comes with its own challenges concerning a challenging data collection, data harmonization, and data analysis. We argued that a consistent analysis framework for gathering, analysing, and validating web data for innovation studies is needed. Additionally, we claimed that the data source itself (i.e. the population of firm websites) has not been studied rigorously in terms of its qualitative and quantitative properties. Basic yet important data characteristics such as the structural properties of firm websites and their coverage of the overall firm population were unknown.

In the remainder of the paper, we first presented a coherent web mining framework based on ARGUS (Automated Robot for Generic Universal Scraping), a free-to-use web scraping tool which we developed for the purpose of large-scale *broad web scraping* (i.e. the scraping of several different websites, contrary to *focus* web scraping, where a single website is being scraped). We then applied ARGUS in two pilot studies. The first one was intended to assess firm websites as a data source by scraping and analysing a large number of firm websites. In the second study, we

used our proposed framework to map the innovation ecosystem "AI in Berlin" - a network of Berlin-based firms that engage in artificial intelligence. We framed our research using the following three research questions:

1. *URL coverage*: What subpopulation of firms can be surveyed using web mining of firm websites and is a systematic bias in terms of firm characteristics (age, size, sector, location etc.) to be expected?

2. *Website characteristics*: How do firm websites differ in terms of their size and content and how does that interfere with web mining studies?

3. *Innovation ecosystem mapping*: How can our proposed framework be used to map an innovation ecosystem?

Our proposed framework is show in Figure 4.1. It is based on a firm database which includes information on firm characteristics (e.g. sector, firm size) and, most importantly, the firms' website addresses (URLs). Ideally, the firm database has been matched to auxiliary databases containing established innovation indicators from questionnaire-based surveys, firm-level patenting data or literature data, such that traditional innovation indicators are available for a sub-sample of the firms in the main dataset. In a first step, the firms' web addresses are passed to a web scraper. The web scraper is then used to download website content (texts, hyperlinks etc.) from the firms' websites. In a third step, data mining techniques are applied to extract information on the firms' innovation activities from the downloaded website content. Based on this information, novel innovation indicators can be constructed. At this stage, additional metadata on the firm can be used to support the analysis (pre-classification, classification model selection based on firm characteristics, information from established innovation indicators etc.). In a final step, the new innovation indicators are merged back to the initial firm database. This last step also establishes a direct firm-level link between the novel innovation indicator and the established indicators available from the auxiliary databases. This link can then be used to evaluate the new indicators against traditional ones.

For the web scraping step shown in Figure 4.1, we developed the free-to-use, *Scrapy* Python based web scraping tool ARGUS (Kinne, 2018). ARGUS aims for a high degree of *adaptability* (ability to scrape a wide variety of web content from any website while maintaining the same structured output) and *scalability* (ability to scrape millions of webpages in a reasonable time to allow for frequent updating of the scraped web data). The graphical user interface of ARGUS is shown in Figure 4.2.

In the first part of the pilot study, we investigated corporate websites as a data source. First of all, we used the MUP to compare the subpopulation of companies that have their own website with those that do not. This basic analysis provides

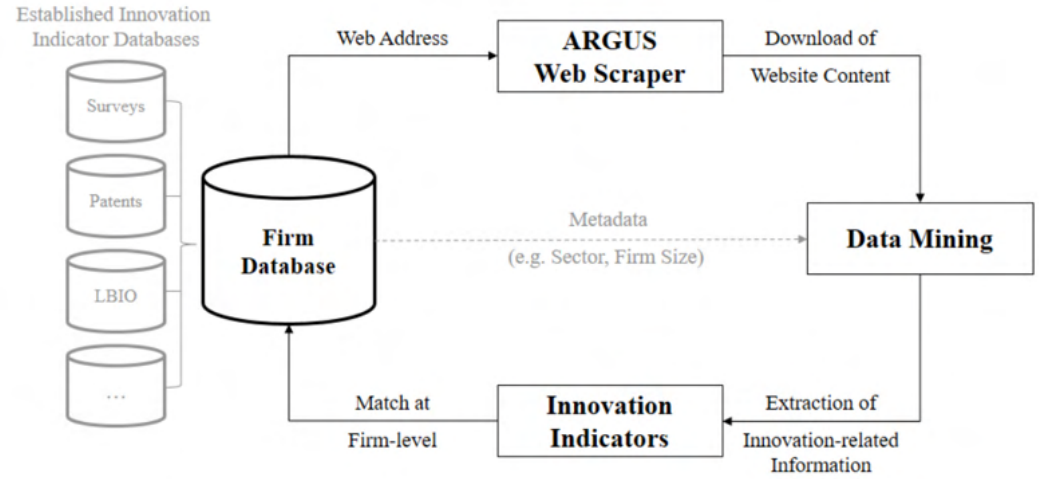


FIGURE 4.1: Analysis framework for generating web-based innovation indicators. From Kinne and Axenbeck, 2018

information about the type of companies that can be investigated with the methodology we proposed and the bias that can be expected in a web mining study with regard to the companies observed. It was found that especially larger and older companies from certain industries (for example mechanical engineering and ICT) operate their own websites. Very young and very small companies (younger than two years and less than 5 employees), on the other hand, are less likely to have their own websites. Moreover, regional broadband availability also seems to play a role in whether companies operate their own websites. We concluded that web mining studies are particularly suitable for the analysis of medium-sized to large companies and certain industries. In this subgroup a website coverage of more than 95% can be expected.

We then analysed the basic characteristics of corporate websites in terms of their structure and scope. For this purpose we used ARGUS and a comprehensive company sample from the MUP. We found that web mining studies have to deal with outliers. For example, 6% of corporate websites have a number of sub-pages that are more than four standard deviations above the population mean. Also with regard to hyperlinks the study showed that some websites are strong outliers and sometimes have hyperlinks to hundreds of thousands of other websites. Large companies in particular often have very large websites, with lots of text, many sub-webpages and many hyperlinks to other companies.

In the second part of the pilot study we used ARGUS to scrape the websites of all companies based in Berlin. We then used a keyword search to identify those companies that mention "artificial intelligence" on their websites. The goal was to map

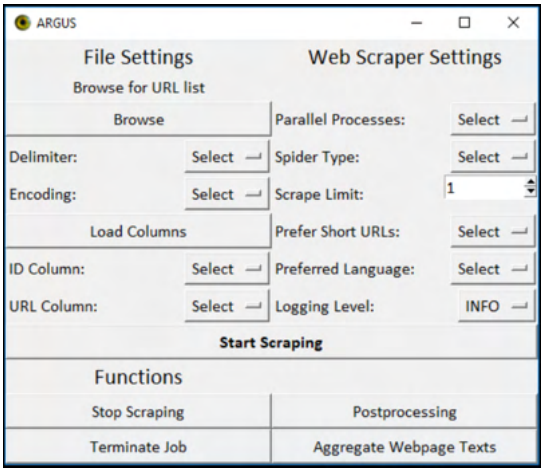


FIGURE 4.2: Graphical user interface of ARGUS.

the "Innovation Ecosystem AI Berlin", which consists of companies, interest groups and scientific institutions that are "AI engaged", i.e. are in some way involved with AI. Figure 4.3 shows the share of these companies in the local business population for the city of Berlin. In addition, we examined the share of "AI engaged companies" in different size and age categories, as well as different industries. To validate our results, we also compared our web-based results with projections from the MIP Innovation Survey, where companies were asked whether they use artificial intelligence. This revealed a similar distribution according to company size classes.

With this study we showed that our proposed multi-step web mining framework (web scraping, data mining, indicator generation and validation) is applicable for innovation ecosystem mapping and produces meaningful results. However, we also pointed out that a more sophisticated text mining approach would be necessary to distinguish the different actor groups (e.g. firms that offer AI-based products and services, universities that are engaged in basic research on AI, and interest groups that promote AI-centered agendas) that resulted from our simple keyword search. We addressed this issue in my fourth paper, which is presented in the following chapter.

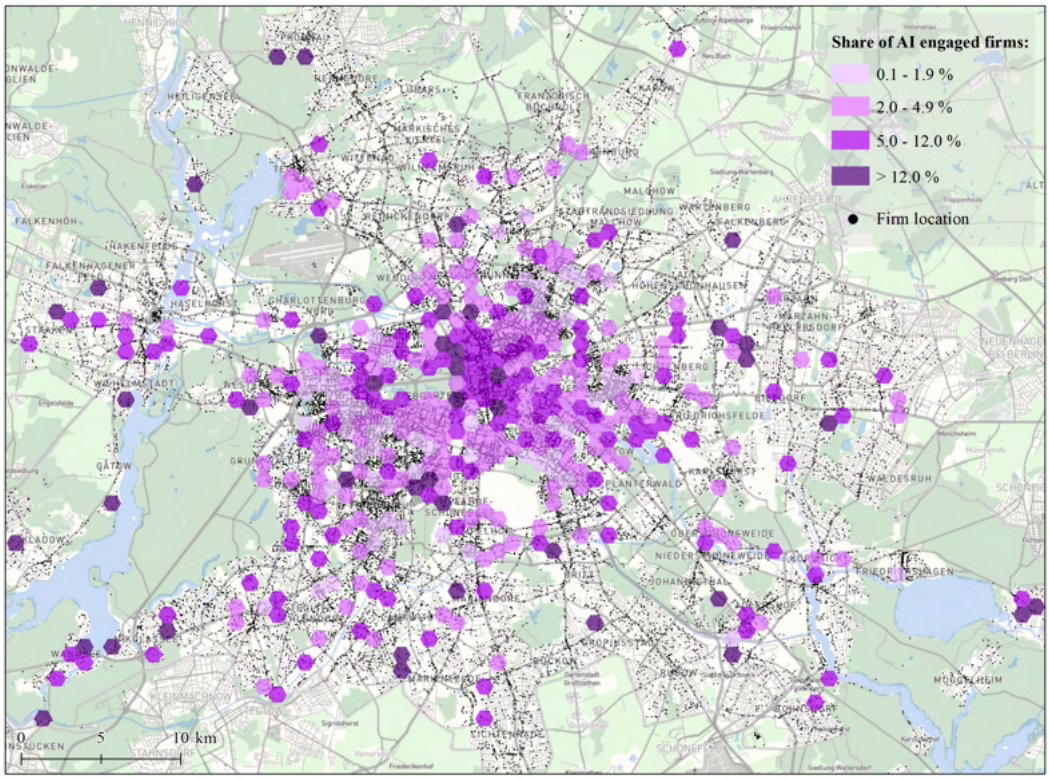


FIGURE 4.3: Share of Berlin-based firms that mention AI at least once on their websites. From Kinne and Axenbeck, 2018

Chapter 5

Deep Learning for Web Text Analysis

5.1 The Rise of Deep Learning in Natural Language Processing

In the previous chapter we proposed a first approach for the data mining step of our conceptualized web mining analysis framework, which was based on a simple keyword search. Our results already pointed in a promising direction, but we also found that a more sophisticated text analysis approach would be needed to extract valuable information from the downloaded web texts.

The field of Natural Language Processing (NLP) has made great progress in the analysis of natural language in recent years, especially through the use of Deep Learning based algorithms (see in particular Mikolov et al., 2011; Mikolov et al., 2013; Le and Mikolov, 2014; Pennington, Socher, and Manning, 2014) and, more recently, transfer learning based approaches (Devlin et al., 2018; Raffel et al., 2019; Liu et al., 2020). These methodological advances have been accompanied by an increasing availability of high-performing hardware (especially graphic processing units suitable for Deep Learning) and user-friendly software (e.g. the Python library Keras, Chollet et al., 2015). In my fourth paper, which is presented in the following section, I and my co-author David Lenz worked on the application of some of these novel NLP methods within the scope of the web mining framework we developed in the previous paper.

5.2 Predicting Innovative Firms

The goal of my fourth paper "Predicting Innovative Firms using Web Mining and Deep Learning", which I designed and wrote together with David Lenz, was to

adopt modern deep learning methods for the analysis of web texts and the identification of innovative companies (i.e. product innovators). We framed our study using the following two research questions:

1. Can deep neural networks be used to reliably identify product innovator firms solely based on their website texts?

2. Are the resulting firm-level, regional, and sectoral patterns from such a prediction model similar to the patterns observed from established innovation indicators when the model is applied to a large out-of-sample dataset of firm website texts?

Figure 5.1 outlines our approach. The websites of companies with unknown product innovator status are queried using ARGUS and all texts from these websites are downloaded. We then used tf-idf text vectorisation (see for example Manning, Raghavan, and Schutze, 2009) to represent the downloaded texts as numerical vectors. These vectors then serve as input to an artificial neural network (ANN), which was previously trained using web texts of companies for which traditional innovation indicators are available. For this purpose we used the companies surveyed in the MIP innovation survey. For each company respondent in this survey, the status of the company as product innovator (or non-innovator) is known and can be used as a "label" for the company's web texts (which were also downloaded using ARGUS). During the training the ANN "learns" which words or word combinations distinguish a product innovator from a non-innovator. After the training, the text of a company without a known innovation status can be entered into the model, which then makes a prediction regarding the probability of the company being a product innovator ("product innovator probability").

We limited the preprocessing of the web texts to a minimum and only standardized the web texts to a maximum of 5,000 characters per company, removed all characters that do not occur in the German alphabet and transferred all characters to lower case. For our classification model we tested different ANN architectures (e.g. convolutional neural networks) and also other, traditional models (e.g. logistic regression and decision trees). In this iterative process an "under-complete autoencoder-like" architecture (see e.g. Goodfellow, Bengio, and Courville, 2016) performed best in our classification task. Table 5.1 shows the performance of our final classification model in the so-called test set - a subset of the companies from the MIP that we did not use for the actual training, but held back for the subsequent performance evaluation of the model. The overall performance of the model, which can be gauged by the total f1-score of 0.80, can be considered very satisfactory. This is especially true since a model based only on traditional company information (company size, age, and industry) only achieves an f1-score of about 0.70. Put simply, our

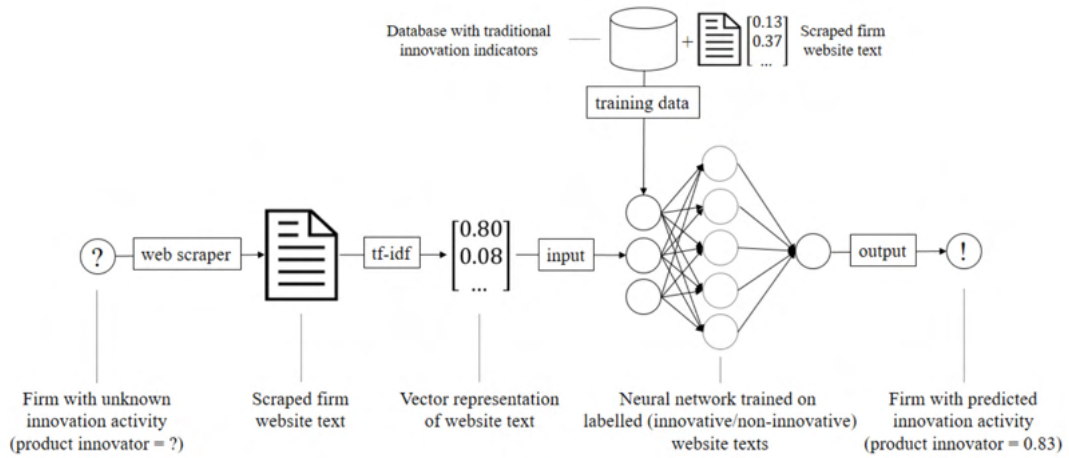


FIGURE 5.1: Framework to predict product innovator firms using their website texts. From Kinne and Lenz, 2019.

TABLE 5.1: Test set classification report. From Kinne and Lenz, 2019.

label	precision	recall	f1-score	support
non-innovator	0.81	0.91	0.86	429
product innovator	0.81	0.64	0.71	255
avg / total	0.81	0.81	0.80	684

model is able to assign a company a correct innovation status in 80% of the cases, and this only based on the company's website.

In the second part of the paper we used the trained model to calculate product innovator probabilities for about 700,000 German companies that have their own website. We then used these out-of-sample predictions for comprehensive robustness checks by comparing our new web-based innovation indicator with established indicators. We first used the MIP innovation survey to compare our results aggregated by size classes and industries with MIP extrapolations. This comparison showed that our results correspond very well to the extrapolations, but in some cases there are sector-specific divergences. As a second robustness check, we compared our new indicator with patent statistics at the firm level.This revealed similar correlations as those between the MIP survey and patent statistics (e.g. product innovators are more likely to hold patents and have more patents in their portfolio). As a third robustness check we compared our web-based results with regional innovation indicators from official statistics. For this purpose, we aggregated our results at the district level and then compared them, for example, with the proportion of the working population employed in R&D-related occupations. We found a high correlation between our new indicator and the established innovation indicators at the regional level. In the

fourth and final robustness check we compared the microgeoraphic pattern of our web-based indicator with the pattern of a special survey of the MIP in Berlin (the information from this special survey was not used during the training). This comparison is shown in Figure 5.2, where we can see a clear match between the two patterns.

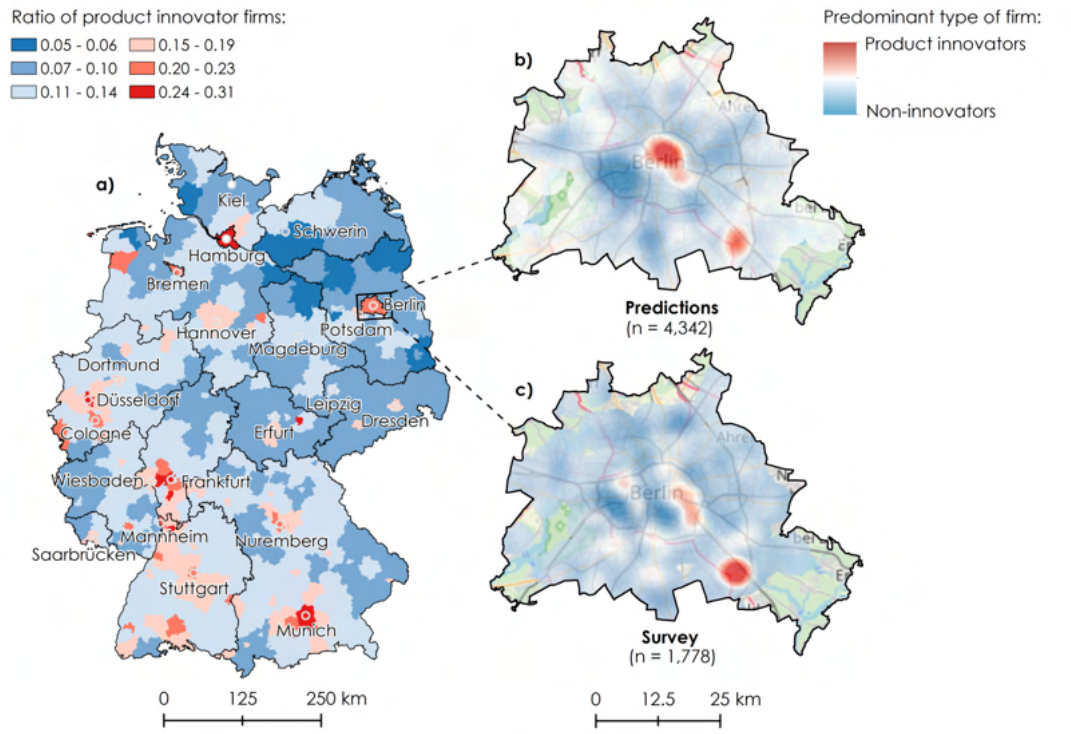


FIGURE 5.2: Map of product innovators. Left: Predicted share of product innovator firms by German districts. Right: Predicted (top) and surveyed (bottom) pattern of product innovator and non-innovator firms in Berlin. From Kinne and Lenz, 2019.

We concluded that our product innovator prediction model (later dubbed *Inno-Prob*) achieved a very good performance, both within the test set and compared to the established indicators. As a weakness we found the somewhat weak performance in relation to the recall of innovative companies (i.e. quite many innovative companies are not detected, while non-innovative companies are detected very reliably). Overall, we expressed confidence that we had created a valuable approach to the analysis of innovation in firms which enables researchers to "zoom in" to any region and perform microgeographic analysis on a comprehensive sample of companies. In the last two papers of my dissertation, which are presented in the next two chapters, we used our indicator for large-scale web-based innovation studies.

Chapter 6

Web Mining for Standards

6.1 Standards as Innovation Indicators

The revised version of the Oslo Manual 2018 (OECD and Eurostat, 2018), which is considered the international guideline for measuring innovation, listed standards as innovation indicators for the first time. Their important coordinating function in markets and their influence on product and process innovations were highlighted in particular. Standards or norms are usually developed on the basis of consensus and adopted by officially recognized organizations, such as the International Organization for Standardization ISO or its German member, the German Institute for Standardization DIN. Furthermore, the certification of important industry or market standards is considered an innovation benchmark. Standardization certificates are intended to demonstrate that product and process innovations meet the specifications defined in standards. In addition, the active participation of companies in standardisation bodies is also classified as an innovation activity. Finally, standards are also a source of information about innovation, which may be even more important than patents (Rammer et al., 2014). In addition to patents, standards can be seen as a new and complementary output indicator, which above all contributes to closing the monitoring gap regarding process innovations which are usually less patented. While patents can be used primarily to measure innovation output in manufacturing, standards are also implemented by service companies. Furthermore, standards also cover various aspects of sustainability and thus social innovations. These include not only the widely used environmental management standard ISO 14001, but also standards on corporate social responsibility, such as ISO 26001.

Standards thus represent an extension of the concept of innovation, and their creation and use can be classified as an innovation activity. However, activities in this area have so far hardly been covered by the usual innovation surveys and indicators. Up to now, standards have only been recorded and analysed on a highly aggregated level in the form of national records or numbers of certifications based on the above-mentioned management standards. In my fifth paper, which is presented in the next section, we designed a web mining based approach to measure

the diffusion of standards at the firm level.

6.2 A Web-based Diffusion Analysis

My fifth paper "Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis", which I co-authored with Mona Mirtsch and Knut Blind, has been published in IEEE Transactions of Engineering Management (issue pending). Within the scope of the paper we developed a web mining based approach to identify companies certified according to a certain standard. We validate our results by means of an elaborate manual review of the supposedly certified companies previously identified by web mining. The standard we chose for this study is the ISO/IEC 27001 standard, which is intended to help companies and organizations to develop and maintain "information security management systems". This standard has been described as "one of the most effective risk management tools for fighting off the billions of attacks that occur each year" (ISO, 2019). Such cyber attacks have become a global risk in recent years, with estimated values at risk of up to USD 5.2 trillion between 2019 and 2023 (Accenture and Ponemeon Institute, 2019). A certification according to ISO/IEC 27001 confirms the successful implementation of the measures defined in the standard. Studies on the current diffusion of the standard have so far been limited to small surveys and case studies. Our paper aims to close this knowledge gap by first investigating the adaptation of the ISO/IEC 27001 standard among German companies using Web Mining and then examining the drivers behind this adaptation.

Once again, the Mannheim Enterprise Panel (MUP) was used as the data basis, extended with web text information based on the web mining framework proposed and tested by us. In addition, the web-based innovation indicator we presented in the previous chapter is also available for each of the companies. In a first step, we identified (analogous to the keyword search in Kinne and Axenbeck, 2018) companies with at least one text reference to the management standards ISO 9001 (quality management), ISO 14001 (environmental management), ISO 50001 (energy management) and ISO/IEC 27001. 4.15% of the company websites contained at least one management standard reference. We then compared the number of companies identified in this way with projections from an ISO survey. It turned out that, with regard to the standards 9001, 14001 and 50001, we were able to capture between 33% and 67% of the number of companies extrapolated from the survey. Interestingly, however, we were able to cover twice as many companies with ISO/IEC 27001 reference as derived from the ISO survey.

In an elaborate manual process, we then categorized all companies identified via web mining that have an ISO/IEC 27001 reference on their website. The results of

this classification are shown in Figure 6.1 (here a company can be assigned to several categories, for example, if a consulting firm offers standardization-related services but is also certified). It can be seen that just under 30% of the identified companies are actually certified, but an equally large proportion only refer to certified partners. Other reasons for naming the standard are, for example, the employment of certified employees or the provision of services related to ISO/IEC 27001. Further examination showed that ISO/IEC 27001 certified companies are also certified to ISO 9001 in 42% of cases, and sometimes also to ISO 14001 (14%) or ISO 50001 (7%).

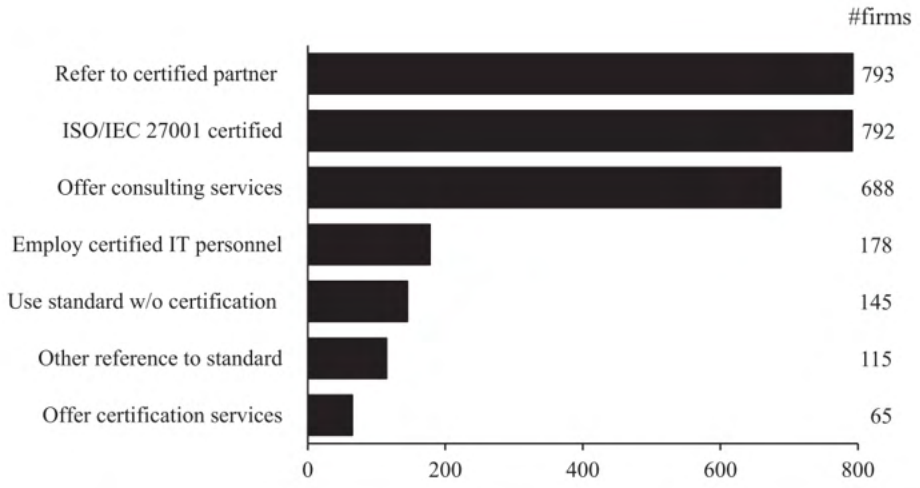


FIGURE 6.1: Firm categorization of 2,664 firms referring to ISO/IEC 27001 on their website. From Mirtsch, Kinne, and Blind, 2020.

Among the ISO/IEC 27001 certified companies identified by us, there is a strikingly large number of companies from the ICT sector (47% as opposed to the approximately 4% in the overall company sample). Table 6.1 shows a comparison of the means for different company characteristics, both for the ICT sector and for the entire company sample. Here it is apparent that certified companies are significantly larger than non-certified companies and also have a significantly higher innovation probability (based on the web-based innovation indicator developed by us) than non-certified companies. While certified companies are significantly younger than non-certified companies in the overall sample, certified ICT companies are significantly older than non-certified ICT companies. The results shown here have proven to be robust in a Probit regression.

As an additional robustness check we compared our results to a database of ISO/IEC certified companies maintained by TÜV Rheinland. This comparison showed a sector distribution that is very close to the distribution we generated using web mining. However, it was also found that 15% of the companies that are certified according to TÜV Rheinland do not publish their certificate on their website. A

TABLE 6.1: Firm characteristics of ISO/IEC 27001 certified firms versus non-certified firms. From Mirtsch, Kinne, and Blind, 2020.

All Sectors		
Mean	*Certified MUP firms*	*Non-certified MUP firms*
Firm Size***	76 (229.27) N=596	23 (483.61) N=458,933
Firm Age***	17 (14.03) N=768	24 (42.44) N=858,902
Innovation probability***	0.57 (0.20) N=774	0.25 (0.16) N=749,580
ICT Service Sector		
Mean	*Certified ICT service firms*	*Non-certified ICT service firms*
Firm Size***	61 (129.11) N=274	15 (78.82) N=16,896
Firm Age***	17 (12.53) N=333	14 (18.33) N=33,050
Innovation probability***	0.62 (0.18) N=331	0.49 (0.21) N=27,266

Notes: Standard deviation in parentheses. N = Number of observations. Significance from the t-test: * $p < 0.10$; ** $p < 0.05$; and *** $p < 0.01$.

small percentage (2%) publishes their certificate as a logo only and not as text, which makes it impossible to capture it via text mining. In addition, an extended manual analysis showed that larger companies with large websites tend to present their certification at a "deep" website level. With our limited web scraping (maximum 25 sub-webpages per company website in this study) we were not able to capture such references sometimes.

We concluded that our approach represents a first successful attempt to generate web mining based information for the diffusion of standards. In the discussion of our results we emphasized the importance of certified IT personnel and certified business partners, which we found in our analysis. We also have identified various implications for managers, policy makers and standardization authorities that can be derived from our study. For example, we discussed the possibility of a political push for the implementation of ISO/IEC 27001 measures in smaller companies without having them to actually obtain a certification, which promises significant cost advantages. In conclusion, our study has shown that web mining can be used to determine the current state of diffusion of standards in the firm population and to investigate the drivers of this diffusion. In contrast to traditional survey methods, our approach is particularly attractive because of the cost aspect, which allows for frequent and comprehensive inquiries. However, we also mention our limited web

scraping (maximum 25 sub-webpages per company website) as a weakness of our approach. Our study also showed that not all companies publish their certificates on their websites and that certificates are sometimes not presented in text form. As a third concern, we also mention the so far insufficient differentiation between companies that are actually certified and those where, for example, only a partner company is certified. Future web mining studies should address these shortcomings.

Chapter 7

Hyperlink Networks and Innovation

7.1 Networks and Proximity

In the web mining related papers presented so far, we have only touched on the relational aspect of our data, namely the existence of hyperlinks between company websites, but have not yet analysed it in detail. Companies link on their websites to, among others, partner companies, customer companies (for the purpose of reference marketing, for example) or associations they are members of. Hyperlinks of a purely technical nature are also possible, for example to retrieve external content and display it on the company's own website. Thus, an existing hyperlink between two company websites may have been created by the linking company for a variety of reasons, but it is always associated to an intentional decision to create this hyperlink and make it public. At the same time the linked company itself has no influence on who links to its own website. As such, the very existence of a hyperlink, its direction and reciprocity (mutuality) is a manifestation of corporate relationships that may be worth investigating.

Against the background of innovation studies, networks (between organisations, companies and people) are perhaps the central social construct, as they are needed for the exchange of information and knowledge. An entire scientific community has evolved around the study of economic networks in recent years (see for example Egeraat and Kogler, 2013; Ter Wal and Boschma, 2009; Hidalgo, 2015; Uzzi, 1996), based on the assumption that networks constitute the fundamental social morphology of our increasingly information-based society. A central assumption is that "the extent to which a network has access to technological know-how is at the roots of productivity and competitiveness" (Castells, 2000). The economy as a whole is understood as a social construct of interconnected companies and organisations (which themselves are networks of individuals) between which knowledge flows, learning is facilitated and innovations are created.

Hyperlink networks as a digital manifestation of real-world relationships between companies were already recognized almost twenty years ago (Park, 2003). So far, however, there have been hardly any studies in innovation research that have used hyperlink networks for comprehensive empirical studies. Especially for the exploration of multidimensional proximity, as envisioned by Boschma (Boschma, 2005), web-based relational data have advantages over traditional relational data from surveys and patents. Boschma and Frenken (Boschma and Frenken, 2010) described the different dimensions of proximity as follows: "In short, cognitive proximity indicates the extent to which two organizations share the same knowledge base; organizational proximity the extent to which two organizations are under common hierarchical control, social proximity the extent to which members of two organizations have friendly relationships, institutional proximity the extent to which two organizations operate under the same institutions, and geographical proximity the physical distance or travel time separating two organizations". When analyzing such proximities, one advantage of web data is that hyperlinks, unlike relationship information derived from patent data (e.g. via co-applications), can also map non-formal relationships. It can also be expected that web data will have advantages in terms of timeliness, granularity and coverage (see previous chapters).

In my final paper, which is summarized in the next chapter, my co-authors and I dealt with the development of web-based measures of cognitive and organizational proximity, and a subsequent comparison of the networking of innovative and non-innovative firms on the Internet.

7.2 How Innovative Firms Relate on the Web

Our paper "The Digital Layer: How Innovative Firms Relate on the Web", which I wrote together with Miriam Krüger, David Lenz, and Bernd Resch, is currently in working paper status. In the paper we develop an approach for mapping firm networks based on hyperlink mining, where in addition an evaluation of the hyperlinks is conducted (for example we distinguish between business and non-business relationships). We call the dataset we created the "Digital Layer", which is, so to speak, a digital model of the relationships between companies/organisations. In our study, the digital layer consists of about 500,000 companies in Germany, which are described by their website texts and are connected to each other by approximately seven million hyperlinks. In addition, for each company, we have company characteristics (size, age, location) from the MUP database and three innovation indicators at the company level: the product innovator status from the CIS (available for about 2,400 companies), the firms' patent holder statuses and our predicted product innovator probabilities (both available for all companies).

After we downloaded web texts and hyperlinks using ARGUS, we first had to decide for an approach to construct the hyperlink network. The most important decisions were whether or not to create a directed or undirected graph from our data and the question of a potential edge weighting using reciprocal (mutual) links. Lastly, a decision also had to be made regarding disconnected companies (companies that do not have a single hyperlink to another company in our data set). We decided to take a simple approach here, as this study constitutes a first attempt on large-scale hyperlink mining and focuses more on the methodological approach (web scraping and hyperlink classification). Thus, we constructed an undirected network, in which the relationships between companies with reciprocal hyperlinks, however, receive a higher weighting (see below). We also excluded the approximately 150,000 companies without hyperlinks in our data set (22%), which particularly affected smaller companies. For follow-up studies there are good arguments to choose a theoretically sound alternative network design, for example, one that only considers reciprocal relationships.

The Digital Layer that has been created in this way is shown in Figure 7.1. The upper panel shows the locations of the included companies and the average product innovator probabilities per grid cell. The middle panel shows an aggregated representation of the approximately seven million hyperlinks between the companies. The bottom panel shows the individual network of a single company (the Centre for European Economic Research in Mannheim) for all of Germany (left) and within its home region (right).

For each node (company) in our network, we first calculated three basic measures that describe its networking:

1. *Link count*: A simple count that measures the number of hyperlinks between this company and other companies (also known as *degree centrality*).

2. *Mean partner innovation*: This measure reflects the "innovativeness" of the company's partners. For this purpose, we average the predicted product innovator probabilities of all other firms that are directly hyperlinked to the firm of interest.

3. *Local firm density*: For this measure, we counted the number of other firms within a one kilometre radius of the (geographical) location of the company under consideration. This firm density measure is intended to be a control variable for localized spillovers that the company may experience from nearby sources. The search radius of one kilometre was chosen in correspondence to our results on rapidly diminishing knowledge spillovers (see Rammer, Kinne, and Blind, 2020).

In addition, we calculate for each company in the Digital Layer measures for the average geographic, cognitive and organizational proximity to their (hyperlinked) partners. We measure geographical proximity by calculating the geographical distance between the locations of linked companies. Our measure of cognitive proximity is intended to provide information on whether linked companies share a common "knowledge base". For this purpose, we calculate the similarities between the website texts of linked companies using the cosine similarity of tf-idf vectorized texts. We approximate organizational proximity by predicting the type of relationship (business vs. non-business) between two hyperlinked companies using a prediction model. Non-business relations are relations between firms that are not directly related to making business with each other and are of non-monetary nature. Business relation include all hyperlinks between firms that do or did business together. We used a logistic regression classifier as our prediction model, which we had trained using the texts of about 3,500 randomly sampled company pairs that are connected via a hyperlink. Previously, we had manually labelled these pairs to be either business or non-business partners by browsing the websites of both companies. The trained model achieved a very high accuracy of 0.92. We then used the trained model to predict the relationship type of all nearly seven million hyperlinks in our data set. The resulting prediction results range from 0.0 (high probability of a business relationship; small organizational proximity) to 1.0 (high probability of a non-business relationship; large organizational distance). These raw prediction results were directly used as a proxy measure for organizational proximity.

We then calculated mean proximity values for each company, as shown in Figure 7.2. Reciprocal relations between firms are assigned double the weighting (e.g. the relation between firms 1 and 3 in Figure 7.2) compared with one-way hyperlinks (Figure 7.2 shows such a relation for firms 1 and 2). As a result, each company in the Digital Layer is characterized by the following properties that describe its networking: Mean (geographical, cognitive, organisational) proximity to its partners, number of partners, mean product innovator probabilities of partners and local firm density.

In the main part of the paper, we examined the differences between innovative and non-innovative firms in terms of their networking characteristics. Central to this was a regression analysis (see Table 7.1), in which we used as dependent variables our web-based innovation indicator (both as a continuous probability and as a binary indicator), CIS product innovator status and patent holder status. We derived the following observations from the results as being either robust (significant and consistent for all specifications) or semi-robust (significant and consistent for most specifications):

1. Innovative firms have more partners. (robust)

TABLE 7.1: Regression results. From Krüger et al., 2020.

Variable	Web dataset (continuous y)	Web dataset (binary y)	Survey dataset (binary y)	Patent dataset (binary y)
		Constant		
Constant	0.2053***	-3.4465***	-2.6743	-5.5191***
		Firm-level controls		
Sector	Yes	Yes	Yes	Yes
Size	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes
Firm density (in 100)	0.0072***	0.0099***	0.0068	0.0021
Firm density (in 100) sq.	-0.0001***	-0.0002***	-0.0005**	-0.0002
		Hyperlink partners		
Link count (log)	0.0270***	0.0404***	0.0294***	0.0435***
Mean partner inno.	0.3036***	0.4602***	0.3803***	0.1745***
		Proximity		
Mean geo. distance	0.2404***	0.2688***	-0.1916	0.2455***
Mean geo. distance sq.	0.0260***	-0.0490**	-0.0966	-0.2399**
Mean cogn. distance	-0.1972***	-0.2045***	-0.0953	0.0284
Mean cogn. distance sq.	0.0733***	-0.0084	0.0398	0.0975*
Mean orga. distance	-0.4267***	-0.8022***	0.2706	0.0360
Mean orga. distance sq.	0.1151***	0.0994***	-0.5607	0.0652
		Proximity interactions		
Geo. dist. * orga. dist.	-0.0863***	-0.0377*	0.5660	0.0962
Geo. dist. * cogn. dist.	-0.2965***	-0.2566***	0.2122	-0.2845***
Cogn. dist. * orga. dist.	0.4326***	0.8583***	-0.1679	-0.1782
		Model statistics		
Model type	Robust OLS	Robust logit (average marginal effects)		
Observations	513,026	513,026	2,384	29,772
(Pseudo) R-squared	0.32	0.24	0.25	0.24
F-test/Wald chi2	3,187***	73,299***	379***	4,225***

2. Innovative firms have partners that are more innovative. (robust).

3. Innovative firms have transregional networks (geographically distant partners). (semi-robust)

4. Innovative firms are more likely to maintain business relationships. (semi-robust)

5. Innovative firms are more likely to connect to companies with similar knowledge bases. (semi-robust)

6. Innovative firms use geographic proximity to overcome cognitive distance to hyperlinked partners or use cognitive proximity to overcome geographic distance to their partners. (robust)

In conclusion, we find that our "Digital Layer" concept is an interesting approach for comprehensive empirical studies on the networking of companies, which also allows for the operationalization of proximity measures that have been difficult to

measure with traditional relational data. In particular, we emphasized the possibility of analyzing individual sectors or regions at any geographical level of analysis. For future studies, we particularly recommended the continuous recording of web data, which would also permit analysis of the Digital Layer over a longer period of time. Microgeographic firm-to-firm knowledge spillovers and the diffusion of technology between firms, industrial sectors, and regions would be particularly interesting research areas.

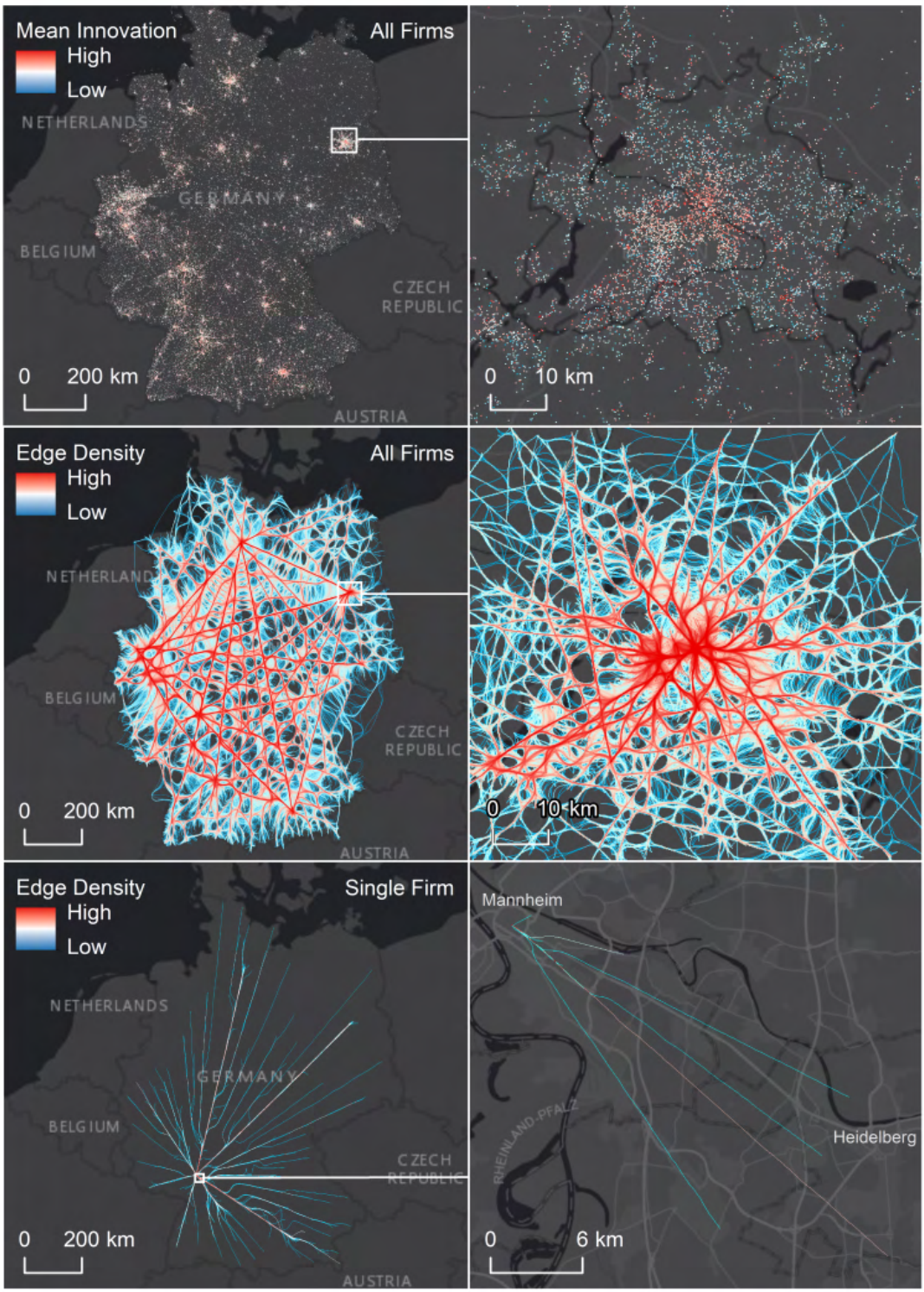
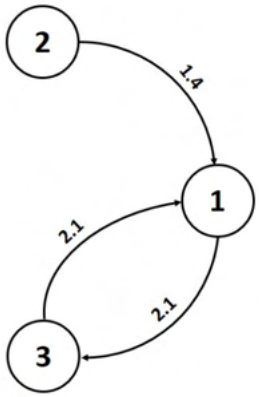


FIGURE 7.1: The Digital Layer of Germany. From Krüger et al., 2020.



Firm	Hyperlinks	Proximities	Mean Proximity
1	[2, 3, 3]	[1.4, 2.1, 2.1]	1.9
2	[1]	[1.4]	1.4
3	[1, 1]	[2.1, 2.1]	2.1

FIGURE 7.2: Schematic representation of a firm's hyperlink network. From Krüger et al., 2020.

Chapter 8

Synthesis

8.1 Summary of Research Results

The focus of my dissertation has changed with and through the progress of my research. Initially, the focus was on the use of microgeographical data for the investigation of location patterns and their determining location factors. It turned out that the new data sources (especially OpenStreetMap), which were not yet widely used in this context at that time, allowed for the microgeographic modelling of both "soft" (e.g. amenities) and "hard" (e.g. transport infrastructure) location factors. However, the econometric modelling of the relationships between these location factors and the location of software companies then also showed that unexplained variation remained particularly in cities and agglomerations (see Chapter 2), which is highlighted in Figure 8.1 for the urban area of Berlin.

As a direct result of this observation, the focus of my second paper was on a closer examination of firm locations within a single city. Inspired by research on the relevance of proximity in organizations for the development of innovations (Catalini, 2018; Kabo et al., 2014), my co-authors and I used the dataset of OSM data and georeferenced MUP company locations developed in my first paper to investigate the relationship between urban knowledge sources and the location of innovative companies. Berlin was an interesting area to study, as a comprehensive special survey of the MIP innovation survey is carried out there every year, which includes all Berlin based companies with at least five employees from the manufacturing and business-oriented services sectors. In contrast to my first paper, in which the proximity to external sources of knowledge could only be depicted by the proximity to universities and research institutes, the Berlin special survey enabled us to observe the proximity to innovative firms and to consider differences in the location decision of innovative and non-innovative firms. Our results confirmed the findings of previous studies that had an intra-organisational setting, which is that the effects of proximity in terms of innovation are of a truly microgeographic nature. Indeed, we were only able to find differences between the locations of innovative and non-innovative companies in terms of external knowledge sources on a very detailed

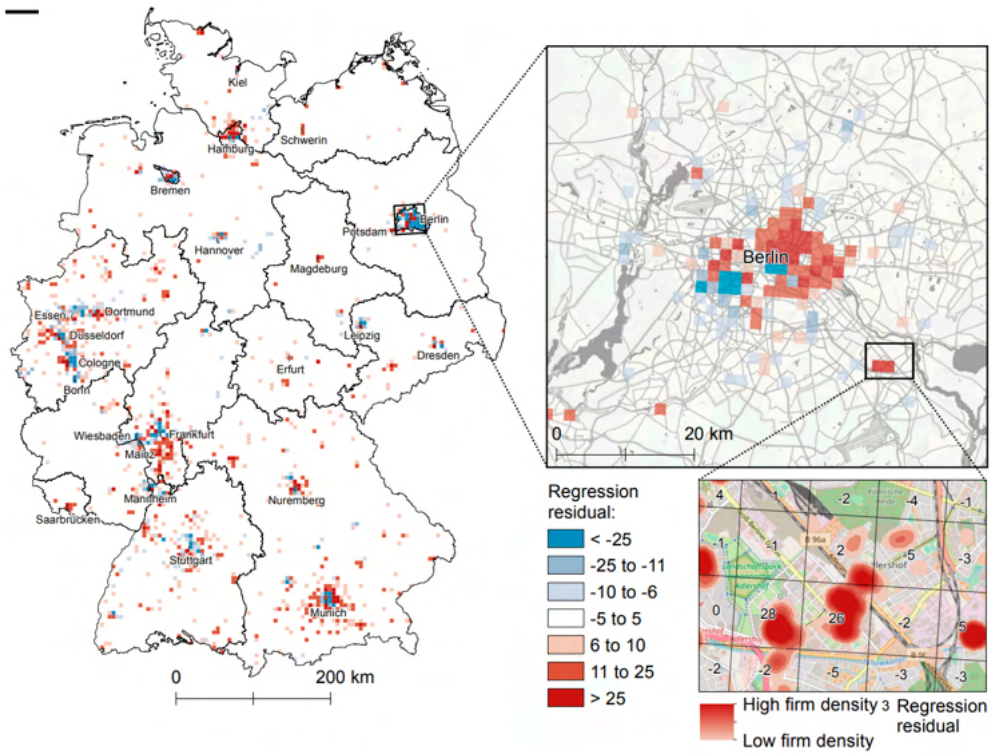


FIGURE 8.1: Regression residuals from Poisson location prediction model, Germany (left) and Berlin (right). From Kinne and Resch, 2018

.

geographical scope of up to one kilometre. However, it was not possible for us to verify our findings for other regions since the sample of the Germany-wide MIP survey covers only about 20,000 companies and the obtained geographical pattern is too sparse for microgeographical analysis. The popular alternative to survey-based innovation indicators, namely patent data, would have been available for all companies in Germany, yet it has the major disadvantage that patents are hardly relevant for a number of sectors. This is particularly true for the software industry, for example, since software is hardly patentable in Europe. Consequently, the patent-based measurement of innovation in the software industry may therefore only provide a distorted picture of the innovation activity in this industry.

In my search for alternative sources of information on the innovation activities of companies, I found that the Internet was hardly used as a source in innovation research at that time. While web data (especially user generated content from social media) had been used in the context of geographic research for at least ten years (compare Goodchild, 2007), for example for so-called event detection in the context of crime (Tonkin, Pfeiffer, and Tourte, 2012; Cheong, Ray, and Green, 2012a),

natural disasters (Collier, Son, and Nguyen, 2010; Gomide et al., 2011) and epidemiology (Sakaki, Okazaki, and Matsuo, 2010; Earle, Bowden, and Guy, 2011) or the classification of land use (Cheong, Ray, and Green, 2012b), there were only a few case studies (Gök, Waterworth, and Shapira, 2015; Arora et al., 2013; Youtie et al., 2012; Beaudry, Héroux-Vaillancourt, and Rietsch, 2016; Nathan and Rosso, 2017) in economic innovation research that used web data. In addition, there was no standardized approach to the generation of firm-level innovation indicators from web data, and the existing studies were one-time, hardly reproducible and transferable, often manually conducted investigations of small company samples. In 2017, I was then actively involved in the initiation and conception of a research project aiming at the development of a standardized framework for the generation of web-based innovation indicators from company websites. The project was eventually funded by the German Federal Ministry of Education and Research as a three-year joint project "Text Data Based Output Indicators as Base of a New Innovation Metric (TOBI)" by the ZEW Centre for European Economic Research and the University of Giessen.

In the context of the TOBI research project the four research papers constituting Part II of this dissertation were developed. The first TOBI-associated paper (Chapter 4) dealt with the motivation of our web mining approach (disadvantages of traditional innovation indicators in terms of timeliness, coverage, granularity, costs), the choice and fundamental investigation of our data basis (corporate websites, as well as their content and structural characteristics) as well as the conception of a consistent analysis framework and its application in a basic pilot study. In the same paper, we also introduced the ARGUS Web Scraper (Kinne, 2018), which we had developed specifically for the requirements of our web mining framework. In our second paper (Chapter 5), we developed an approach for generating a firm-level innovation indicator based on deep learning of website text. In our third paper (Chapter 6) we used our web mining framework and the new web-based innovation indicator to investigate the diffusion of a management standard. The fourth paper (Chapter 7) focused on the incorporation and classification of hyperlinks to explore the differences in the networking of innovative and non-innovative firms on the World Wide Web.

Our work was very well received by the scientific community as well as by stakeholders in official statistics and policy consulting. Thus, I have presented our approach and research results at more than 25 scientific workshops and conferences over the past two and a half years. The audience ranged from a primarily scientific spectrum (i.a. Swiss Federal Institute of Technology in Zurich, Laboratory for Innovation Science at Harvard) to official statistics (i.a. European Statistical Office, German Federal Statistical Office) and institutions with economic policy tasks (i.a. German central bank Bundesbank, OECD, German Federal Ministry of Education

and Research). In addition, several direct follow-up projects emerged from the original TOBI research, in which our web mining concept is either developed further or in which the web-based innovation indicator developed by us is used.

The application of our approach is not only limited to innovation research, but also offers itself for other research fields as an approach to generate research data. Our approach is also used, for example, to research gender equality at the firm level and to identify "green" business activities. This universal applicability for the analysis of companies via their websites is also the basis for the academic spin-off startup istari.ai, in which I am one of the co-founders. Istari.ai was founded in 2019 and specializes in the generation of up-to-date business information from corporate websites (e.g. on technologies used, corporate partners, and fields of activity).

One of the first projects in which istari.ai was involved is the web-based monitoring of the 2020 Coronavirus pandemic and the associated reactions of German companies (Kinne et al., 2020). In this project, the websites of more than one million German firms were analysed twice a week for three months in order to identify text references to the Coronavirus pandemic. Subsequently, the context (e.g. "problem" context or "adaption" context) of those identified references has been classified using a transfer learning based methodology. Compared to a questionnaire-based survey, which takes weeks to conduct and evaluate, our web mining approach, which is directly based on the research of my dissertation, provides up-to-date and comprehensive data necessary for evidence-based economic policy in the face of such a dynamic crisis. This example also shows that we might have created something of actual relevance through the research presented here.

8.2 Conclusion and Outlook

In the context of this dissertation an attempt was made to combine modern GI-Science methods and data with a topic in the field of innovation economics. To the best of my knowledge, this work represents a first attempt to provide a comprehensive, microgeographic perspective on the mechanisms behind the interactions between firms, innovation and location. This microgeographic perspective has only become possible through the emergence of widely available and detailed geodata, especially Volunteered Geographic Information. With regard to firms, on the other hand, the data situation turned out to be insufficient, since traditional innovation indicators are not suitable for this kind of analysis due to their insufficient coverage and granularity. This led my co-authors and I to develop a novel approach to measure innovation based on web data, since so far there was no generally applicable approach to generate innovation indicators from unstructured big web data. With our methodological work, we have opened up a field of research that has the

potential to make a relevant contribution to the generation of research data as well as indicators for evidence-based policy-making even outside of innovation research. Apart from a host of more specific research questions (e.g. mapping of geographical technology diffusion), a number of broader aspects concerning our approach and the data basis used remain open, some of which are summarized below:

1. *Corporate websites as an unbiased data source*: Additional research efforts should be made to investigate the effects of varying "digital and marketing capacities" of companies. For example, it is currently still unclear whether companies with better marketing (i.e. with a more favourable public profile) are systematically rated as more innovative by our model and whether this then also leads to a distortion in the aggregate (across all companies). The same applies to inadequately maintained company websites and their uneven impact on the timeliness of the information obtained from these websites. This was particularly evident in our web-based study on the impact of the 2020 Coronavirus pandemic (Kinne et al., 2020), which is not part of this dissertation. The absence of a textual corona reference in this case does not necessarily mean that the company is not affected by the pandemic at all. Just as well, the company may simply not have the digital capacity (e.g. a dedicated in-house online team) to keep its website up to date.

2. *Explainable AI*: In recent years there has been an increasing awareness (see for example Barredo Arrieta et al., 2020) that artificial intelligence systems in certain settings should not be black boxes, but must produce results that are interpretable, fair, transparent and accountable. Particularly against the background that political decisions will increasingly be based on the results of AI models (for example, assuming that our proposed approach is adopted in official statistics), these results must meet all of these criteria. For example, additional research is needed to understand the drivers of the output of our product innovator prediction model. It would be relevant to know which words and combinations of words have a particularly high weight when making the predictions. At the same time, there are also arguments in favour of a black box, considering that a black box is probably also more difficult to manipulate compared to an open and fully understood system.

3. *Expansion of the data basis and data fusion*: So far, we have only used HTML texts and hyperlinks as the data basis in our research. However, the texts published in PDF files as well as videos and images available on the company's website are a valuable source of information for a complete record of the company's activities. Yet the integration of images and videos poses quite another

methodological challenge for the analysis. In addition, the integration of further web data, such as social media profiles of companies or external news articles, promises to be an interesting extension of the data basis to be analysed and should be considered.

4. *Expanding the scope*: As already mentioned, the Web Mining approach presented here was developed specifically for the requirements of innovation research, but is also suitable for illuminating other aspects of a company. Companies also report on their websites about their employees and management, their skills and expertise, their plans, deployed technologies, partners and much more. Basically, a well maintained corporate website can be seen as a constantly updated document with extensive information about the general activity of the company, which can be easily accessed through our approach in order to generate pre-structured baseline data for further research.

5. *Time series analysis*: Innovation always means change. However, in our research to date we have only considered cross-sections and have omitted the temporal dimension. In fact, however, it is precisely the changes and updates on company websites that provide a valuable source of information on innovations in companies. Such studies will indeed soon be possible, as we have been regularly monitoring and storing data on at least all German companies for about two years now. In the aforementioned study on the 2020 Coronavirus pandemic, we have already conducted such a time series analysis, albeit only over a period of three months. It would also be worthwhile using Internet archives to set up a backward projection database with historical time series data from company websites.

6. *Internationalisation and multilingual models*: The research presented here focused exclusively on Germany, primarily due to the data basis, but also for language reasons. However, our approach can also be applied in other countries and for studies of international contexts, which seems appropriate against the background of a highly globalized world economy. Global databases containing information (including web addresses) for a large number of companies active worldwide are available and can be used for this purpose. Multilingual language models, which do not have to be adapted for each individual language, have particularly great potential for such analyses. Many of the novel NLP models based on transfer learning have this feature (Devlin et al., 2018; Raffel et al., 2019; Liu et al., 2020) and we have also successfully used them in international collaborations (König, Müller, and Wörter, 2020). Of course, this also poses special demands on the training data that is used. For example, "innovative companies" are not the same everywhere and a sector such as mechanical

engineering could be regarded as being often innovative in Germany, whereas in other countries it is perhaps regarded as being rather outdated. Here, "cultural awareness" must be developed and taken into account when constructing such models.

Bibliography

Accenture and Ponemeon Institute (2019). *The Cost of Cybercrime: Ninth Annual Cost of Cybercrime Study*. Tech. rep. Traverse City, Michigan: Ponemon Institute, p. 18.

Acs, Zoltan J., Luc Anselin, and Attila Varga (2002). "Patents and Innovation Counts as Measures of Regional Production of New Knowledge". In: *Research Policy* 31.7, pp. 1069–1085. DOI: 10.1016/S0048-7333(01)00184-6.

Ahlfeldt, Gabriel and Elisabetta Pietrostefani (2017). "The Economic Effects of Density: A Synthesis". London.

Archibugi, Daniele and Mario Planta (1996). "Measuring technological change through patents and innovation surveys". In: *Technovation* 16.9, pp. 451–468. DOI: 10.1016/0166-4972(96)00031-4.

Arora, Sanjay K. et al. (2013). "Entry strategies in an emerging technology: a pilot web-based study on graphene firms". In: *Scientometrics* 95.3, pp. 1189–1207.

Askitas, Nikolaos and Klaus F. Zimmermann (2015). "The Internet as a data source for advancement in social sciences". In: *International Journal of Manpower* 36.1, pp. 2–12. DOI: 10.1108/IJM-02-2015-0029.

Audretsch, B (1998). "Agglomeration and the location of innovative activity". In: *Oxford Review of Economic Policy* 14.2, pp. 18–29. DOI: 10.1093/oxrep/14.2.18.

Audretsch, David and Maryann Feldman (2004). "Knowledge Spillovers and the Geography of Innovation". In: vol. 4. DOI: 10.1016/S1574-0080(04)80018-X.

Barredo Arrieta, Alejandro et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Bathelt, Harald and Johannes Glückler (2012). *Wirtschaftsgeographie: ökonomische Beziehungen in räumlicher Perspektive*. ger.

Beaudry, Catherine, Mikaël Héroux-Vaillancourt, and Constant Rietsch (2016). "Validation of a web mining technique to measure innovation in high technology Canadian industries". In: *CARMA 2016–1st International Conference on Advanced Research Methods and Analytics*, pp. 1–25.

Bettencourt, Luís M a (2013). "The origins of scaling in cities". In: *Science* 340.6139, pp. 1438–1441. DOI: 10.1126/science.1235823.

Bluemke, M. et al. (2017). "Integrating Geographic Information into Survey Research: Current Applications, Challenges and Future Avenues". In: *Survey Research Methods* 11.3, pp. 307–327. DOI: http://dx.doi.org/10.18148/srm/2017.v11i3.6733.

Boschma, Ron (2005). "Proximity and Innovation: A Critical Assessment". In: *Regional Studies* 39.1, pp. 61–74. DOI: 10.1080/0034340052000320887.

Boschma, Ron and Koen Frenken (2010). "The Spatial Evolution of Innovation Networks: A Proximity Perspective". In: *The Handbook of Evolutionary Economic Geography*. Edward Elgar Publishing. Chap. 5.

Capello, Roberta (2014). "Classical Contributions to Location Theory". In: *Handbook of Regional Science*. Ed. by Manfred M. Fischer and Peter Nijkamp. Berlin, Heidelberg: Springer, pp. 507–526.

Caragliu, Andrea et al. (2015). "The winner takes it all: forward-looking cities and urban innovation". In: *Annals of Regional Science* 56.3, pp. 1–29. DOI: 10.1007/s00168-015-0734-5.

Castells, M. (2000). *The Rise of The Network Society: The Information Age: Economy, Society and Culture*. Information Age Series v. 1. Wiley.

Catalini, Christian (2018). "Microgeography and the Direction of Inventive Activity". In: *Management Science* 64.9, pp. 4348–4364. DOI: 10.1287/mnsc.2017.2798.

Cheong, Marc, Sid Ray, and David Green (2012a). "Interpreting the 2011 London riots from Twitter metadata". In: *International Conference on Intelligent Systems Design and Applications, ISDA*, pp. 915–920. DOI: 10.1109/ISDA.2012.6416660.

— (2012b). "Large-scale socio-demographic pattern discovery on microblog metadata". In: *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 909–914. DOI: 10.1109/ISDA.2012.6416659.

Chollet, François et al. (2015). *Keras*. https://keras.io.

Collier, Nigel, Nguyen Truong Son, and Ngoc Mai Nguyen (2010). "OMG U got flu? Analysis of shared health messages for bio-surveillance". In: *CEUR Workshop Proceedings* 714.Suppl 5, pp. 18–26. DOI: 10.1186/2041-1480-2-S5-S9.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". URL: http://arxiv.org/abs/1810.04805.

Duranton, Gilles and Diego Puga (2004). "Micro-foundations of urban agglomeration economies". In: *Handbook of Regional and Urban Economics*. Ed. by J. V. Henderson and J. F. Thisse. Vol. 4. Handbook of Regional and Urban Economics. Elsevier. Chap. 48, pp. 2063–2117.

Earle, Paul S., Daniel C. Bowden, and Michelle Guy (2011). "Twitter earthquake detection: Earthquake monitoring in a social world". In: *Annals of Geophysics* 54.6, pp. 708–715. DOI: 10.4401/ag-5364.

Egeraat, Chris van and Dieter F. Kogler (2013). "Global and Regional Dynamics in Knowledge Flows and Innovation Networks". In: *European Planning Studies* 21.9, pp. 1317–1322. DOI: 10.1080/09654313.2012.755827.

Essletzbichler, Jürgen (2011). "Locating Location Models". In: *The Sage Handbook of Economic Geography*. Ed. by Andrew Leyshon et al. London, Thousand Oaks, New Dehli, Singapore: SAGE. Chap. 1, p. 411.

Florida, Richard, Patrick Adler, and Charlotta Mellander (2017). "The City as Innovation Machine". In: *Regional Studies* 51.1, pp. 86–96. DOI: 10.1080/00343404.2016.1255324.

Glaeser, Edward (1999). "Learning in Cities". In: *Journal of Urban Economics* 46.2, pp. 254–277.

Gomide, J. et al. (2011). "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter". In: *Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany.*, pp. 1–8. DOI: 10.1145/2527031.2527049.

Goodchild, Michael F. (2007). "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69.4, pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, Massachusetts: MIT Press.

Gök, Abdullah, Alec Waterworth, and Philip Shapira (2015). "Use of web mining in studying innovation". In: *Scientometrics* 102.1, pp. 653–671. DOI: 10.1007/s11192-014-1434-0.

Haas, Hans-Dieter and Simon-Martin Neumair (2008). *Wirtschaftsgeographie*. Darmstadt: Wissenschaftliche Buchgesellschaft, p. 136.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". In: *The Review of Economic Studies* 64.4, pp. 605–654.

Helbing, Dirk et al. (2007). "Growth, innovation, scaling, and the pace of life in cities". In: *PNAS* 104.17, pp. 7301–7306.

Henderson, J. Vernon (2007). "Understanding knowledge spillovers". In: *Regional Science and Urban Economics* 37.4, pp. 497–508.

Hidalgo, C. (2015). *Why Information Grows: The Evolution of Order, from Atoms to Economies*. Penguin Books Limited.

Isard, Walter (1956). *Location and Space-Theory*. New York, London: MIT University Press, John Wiley & Sons, Chapman & Hall, p. 380. DOI: 10.1017/CBO9781107415324.004.

ISO (2019). "The Cyber Secrets". In: *ISOfocus*.

Kabo, Felichism W. et al. (2014). "Proximity effects on the dynamics and outcomes of scientific collaborations". In: *Research Policy* 43.9, pp. 1469 –1485. ISSN: 0048-7333. DOI: https://doi.org/10.1016/j.respol.2014.04.007.

Kinne, Jan (2016). "The Geographic Dispersal of the German Software Industry". Master Thesis. Ruprecht-Karls Universität Heidelberg, p. 120.

— (2018). *ARGUS - An Automated Robot for Generic Universal Scraping*. Mannheim. DOI: 10.1109/LPT.2009.2020494. URL: https://github.com/datawizard1337/ARGUS.

Kinne, Jan and Janna Axenbeck (2018). "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany". Mannheim.

Kinne, Jan and David Lenz (2019). "Predicting Innovative Firms using Web Mining and Deep Learning". Mannheim.

Kinne, Jan and Bernd Resch (2018). "Analyzing and predicting micro-location patterns of software firms". In: *ISPRS International Journal of Geo-Information* 7.1. DOI: 10.3390/ijgi7010001.

Kinne, Jan et al. (2020). *Coronavirus Pandemic Affects Companies Differently: A high-frequency website analysis of companies' reactions to the Coronavirus pandemic in Germany*. Tech. rep. Mannheim: ZEW Centre for European Economic Research, p. 15.

König, Michael, Oliver Müller, and Martin Wörter (2020). *Analyse der Webseiten Schweizer Unternehmen zur Reaktion auf die Corona-Pandemie – Methodenbericht.* Tech. rep. Zurich: KOF ETH Zurich, pp. 1–3.

Krüger, Miriam et al. (2020). "The Digital Layer: How innovative firms relate on the Web". Mannheim.

Le, Quoc V. and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32, p. 9. DOI: 10.1145/2740908.2742760.

Leyshon, Andrew et al. (2011). *The Sage Handbook of Economic Geography*. London, Thousand Oaks, New Dehli, Singapore: SAGE, p. 411.

Liu, Xiaodong et al. (2020). "Multi-task deep neural networks for natural language understanding". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 4487–4496.

Manley, David (2014). "Scale, Aggregation, and the Modifiable Areal Unit Problem". In: *Handbook of Regional Science*. Ed. by Manfred M. Fischer and Peter Nijkamp. Berlin, Heidelberg: Springer, pp. 1157–1171.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2009). *An Introduction to Information Retrieval*. Online edi. Cambridge University Press, p. 569. DOI: 10.1109/LPT.2009.2020494.

Marshall, A. (1890). *Principles of Economics*. Principles of Economics v. 1. Macmillan and Company.

Mikolov, Tomas et al. (2011). "Strategies for Training Large Scale Neural Network Language Models". In: *Proceedings of the Annual Conference of the International*

Speech Communication Association, INTERSPEECH. DOI: 10.1109/ASRU.2011.6163930.

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space".

Mirtsch, Mona, Jan Kinne, and Knut Blind (2020). "Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis". In: *IEEE Transactions on Engineering Management*, pp. 1–14. DOI: 10.1109/TEM.2020.2977815.

Möller, Kristoffer (2018). "Culturally clustered or in the cloud? How amenities drive firm location decision in Berlin". In: *Journal of Regional Science* 58, pp. 728–758. DOI: 10.1111/jors.12383.

Nagaoka, Sadao, Kazuyuki Motohashi, and Akira Goto (2010). "Patent Statistics as an Innovation Indicator". In: *Handbook of Economics of Innovation*. Ed. by Bronwyn H. Hall and Nathan Rosenberg. Vol. 2, pp. 1083–1127.

Nathan, Max and Anna Rosso (2017). "Innovative Events".

OECD (2009). *OECD Patent Statistics Manual*. OECD, p. 162. DOI: 10.1787/9789264056442-en.

OECD and Eurostat (2018). *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation*. 4th. Luxembourg, Paris: OECD and Eurostat, p. 258. DOI: 10.1787/9789264304604-en.

Park, Han Woo (2003). "Hyperlink network analysis: A new method for the study of social structure on the web". In: *Connections* 25, pp. 49–61.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Peters, Bettina and Christian Rammer (2013). "Innovation panel surveys in Germany". In: *Handbook of Innovation Indicators and Measurement*. Ed. by Fred Gault. Chapters. Edward Elgar Publishing. Chap. 6, pp. 135–177.

Porter, Michael E (1998). "Clusters and the New Economics of Competition". In: *Harvard Business Review* November-December.

Pred, A. (1969). "Behaviour and location: Foundations for a geographic and dynamic location theory." In: *Lund Series in Geography* 26.B.

Raffel, Colin et al. (2019). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". URL: http://arxiv.org/abs/1910.10683.

Rammer, C et al. (2014). *Innovationsverhalten der deutschen Wirtschaft 2013*. Tech. rep., pp. 1–20.

Rammer, Christian, Jan Kinne, and Knut Blind (2020). "Knowledge proximity and firm innovation: A microgeographic analysis for Berlin". In: *Urban Studies* 57.5, pp. 996–1014. DOI: 10.1177/0042098018820241.

Rosenthal, Stuart S. and William C. Strange (2004). "Evidence on the nature and sources of agglomeration economies". In: *Handbook of Regional and Urban Economics - Vol 4*. Ed. by J Vernon Henderson and Jacques-Francois Thisse. Vol. 4. Elsevier B.V. Chap. 49, pp. 2120–2167. DOI: 10.1016/S0169-7218(04)07049-2.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo (2010). "Earthquake shakes Twitter users: real-time event detection by social sensors". In: *WWW '10: Proceedings of the 19th international conference on World wide web*, p. 851. DOI: 10.1145/1772690.1772777.

Schätzl, Ludwig (2003). *Wirtschaftsgeographie 1 Theorie*. 9th ed. Paderborn: Verlag Ferdinand Schönigh, p. 280.

Schumpeter, J.A. (1942). *Capitalism, Socialism and Democracy*. Taylor & Francis.

Simmie, James (2002). "Knowledge Spillovers and Reasons for the Concentration of Innovative SMEs". In: *Urban Studies* 39.5-6, pp. 885–902. DOI: 10.1080/00420980220128363.

Smith, David M. (1981). *Industrial Location: An Economic Geographical Analysis*. 2nd. New York, Chichester, Brisbane, Toronto: John Wiley & Sons, p. 492.

Smith, Donald F. Jr. and Richard Florida (1994). "Agglomeration and Industrial Location: An Econometric Analysis of Japanese-Affiliated Manufacturing Establishments in Automotive-Related Industries". In: *Journal of Urban Economics* 36.1, pp. 23–41.

Squicciarini, Mariagrazia, Hélène Dernis, and Chiara Criscuolo (2013). "Measuring Patent Quality: Indicators of Technological and Economic Values". Paris.

Ter Wal, Anne and Ron Boschma (Mar. 2009). "Applying Social Network Analysis in Economic Geography: Framing Some Key Analytic Issues". In: *The Annals of Regional Science* 43, pp. 739–756. DOI: 10.1007/s00168-008-0258-3.

Thünen, Heinrich Johann von (1842). *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. 2013th ed. Paderborn: HWA, p. 679.

Tonkin, Emma, HD Pfeiffer, and Greg Tourte (2012). "Twitter, information sharing and the London riots?" In: *... Society for Information ...*, pp. 49–57.

Uzzi, Brian (1996). "The Sources and Consequences of Embeddedness for the Economic Performance of Organizations: The Network Effect". In: *American Sociological Review* 61.4, pp. 674–698.

Weber, Alfred (1922). *Über den Standort der Industrien: Reine Theorie des Standortes*. 2nd ed. Tübingen: J.C.B. Mohr, p. 264.

Youtie, Jan et al. (2012). "Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies". In: *Technology Analysis and Strategic Management* 24.10, pp. 981–995. DOI: 10.1080/09537325.2012.724163.

Appendix A

Paper 1: Analyzing and Predicting Micro-Location Patterns of Software Firms

International Journal of *Geo-Information*

MDPI

Article

Analyzing and Predicting Micro-Location Patterns of Software Firms

Jan Kinne [1,2,*] and Bernd Resch [2,3]

[1] Department of Economics of Innovation and Industrial Dynamics, Centre for European Economic Research, L7 1, 68161 Mannheim, Germany
[2] Department of Geoinformatics—Z_GIS, University of Salzburg, 5020 Salzburg, Austria; bernd.resch@sbg.ac.at
[3] Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA
* Correspondence: jan.kinne@zew.de; Tel.: +49-621-1235-297

Received: 20 November 2017; Accepted: 22 December 2017; Published: 24 December 2017

Abstract: While the effects of non-geographic aggregation on statistical inference are well studied in economics, research on the effects of geographic aggregation on regression analysis is rather scarce. This knowledge gap, together with the use of aggregated spatial units in previous firm location studies, results in a lack of understanding of firm location determinants at the microgeographic level. Suitable data for microgeographic location analysis has become available only recently through the emergence of Volunteered Geographic Information (VGI), especially the OpenStreetMap (OSM) project, and the increasing availability of official (open) geodata. In this paper, we use a comprehensive dataset of three million street-level geocoded firm observations to explore the location pattern of software firms in an Exploratory Spatial Data Analysis (ESDA). Based on the ESDA results, we develop a software firm location prediction model using Poisson regression and OSM data. Our findings offer novel insights into the mode of operation of the Modifiable Areal Unit Problem (MAUP) in the context of a microgeographic location analysis: We find that non-aggregated data can be used to detect information on location determinants, which are superimposed when aggregated spatial units are analyzed, and that some findings of previous firm location studies are not robust at the microgeographic level. However, we also conclude that the lack of high-resolution geodata on socio-economic population characteristics causes systematic prediction errors, especially in cities with diverse and segregated populations.

Keywords: firm location; location factors; software industry; microgeography; OpenStreetMap (OSM); prediction; Volunteered Geographic Information (VGI); Modifiable Areal Unit Problem (MAUP)

1. Introduction

The location pattern of any industry is the product of a large number of individual decisions. Industrial location analysis investigates these location decisions and seeks to detect location determinants that trigger and influence such decisions. These determinants are generally referred to as location factors. A thorough understanding of the impact of location factors on firms' location decisions and firm performance can have important implications for stakeholders. Managers and entrepreneurs can integrate valuable information into the decision making process when choosing the location of a new venture [1].

Policy makers at the regional, national, and multinational level want to promote economic growth by developing the right location factors to create a beneficial environment for firms. The long-standing study of industrial location research [2] has brought forward a wide range of location factors which can be studied at different levels of geographic aggregation, from the immediate firm neighborhood to highly aggregated spatial units. However, the analyzed location factors may vary in direction and

strength at different levels of analysis and findings from aggregated spatial vary depending on the spatial scale at which the analysis is conducted [3]. This issue is generally referred to as the Modifiable Areal Unit Problem (MAUP), which is defined through a location, a scale and a shape dimension [4–6]. The selection of the appropriate level of analysis is therefore crucial, especially in studies which evaluate public policies [7,8], and must be based on reasonable and transparent assumptions.

Such assumptions rely on a thorough understanding of geographic aggregation effects on statistical inference. While the effects of non-geographic aggregation on inference are well studied in economics [9,10], research on geographic aggregation is rather scarce. Amrhein [11] finds that scaling has strong effects on regression coefficients and correlation statistics. However, it is unclear how robust these results are in an empirical setting as simulated data was used in this study. Arauzo-Carod et al. [12] and Manjon-Antolin et al. [13] find only minimal zonation effects on regression results. Briant et al. [14] use administrative spatial units and gridding to assess both the scaling and shape dimension of the MAUP. They find that the use of different spatial units results in different regression coefficients. Overall, the understanding of the MAUP in industrial location analysis remains incomplete and Arauzo-Carod et al. conclude in their meta-study on industrial location research that "[. . .] the reported effects may not be robust to the use of alternative geographical units and the presence of spatial effects. In general, it is not clear what effects spatial aggregation and spatial dependence may have on the inference" [15]. Most previous studies analyzed firm location patterns aggregated at rather crude spatial scales, such as counties or metropolitan areas, and thus there is a lack of understanding of location determinants at the microgeographic level. The varying direction and strength of location factors at different levels of aggregation may lead to superimposed location factors which are missed when aggregated geographic units are analyses. Some location factor-firm relationships which are relevant at the macro level (aggregate) may not be so at the micro level (*ecological fallacy*).

Suitable data for such a microgeographic analysis has become available only recently through the emergence of Volunteered Geographic Information (VGI) [16] and the increasing availability of official (open) geodata [17–19]. The OpenStreetMap (OSM) project is of particular interest in the context of firm location analysis as it goes beyond mapping ordinary road networks: The informal OSM standard contains hundreds of tags in over 25 categories and includes map features such as amenities and public transport stations [20]. Up to now, only few studies have utilized the potential of OSM in firm location analysis and geographic economic analysis in general [21–23]. However, these studies did not use OSM in a large-scale spatial analysis but concentrated on single cities and a strongly limited set of location factors. Following the analysis of previous research efforts, the research questions for our work are defined as follows:

RQ1 Are the effects of location factors, as reported by previous studies using aggregated spatial units, robust at the microgeographic level?

RQ2 How does a firm location prediction model perform at the microgeographic level and to what degree does it provide valuable new insights into the firm allocation process? What are the distinct requirements to the data and the statistical model?

To answer the research questions above, we analyze firm location patterns at the microgeographic level using spatial firm-related data that are available in unseen detail compared to previous studies. We combine this unique data set of three million geocoded street-level firm observations in Germany with OSM data and other detailed geodata (population density, land cover, railway stations, education levels, life expectancy, and many others). We investigate whether findings from previous industrial location studies hold true at a small spatial scale, i.e., at fine spatial resolutions. In general, regular gridding reduces the bias induced by the use of predefined administrative units [24]. In our study, we focus on the software industry, which is rather unrestricted in its location decisions [22], inducing only little bias from unobservable location determinants.

First, we investigate the software firm location pattern in an Exploratory Spatial Data Analysis (ESDA). We find that Poisson regression is likely to be an appropriate method to model the pattern

of software firms aggregated at a regular 1 km grid, whereas negative binomial regression seems to be appropriate for higher levels of aggregation due to over-dispersion in the point pattern. Further, we find that software firms are an urban phenomenon, as they are disproportionally frequent in and around urban areas and even form statistically significant hotspots in some city regions. We further conclude that the regional settlement structure (polycentric vs. monocentric) seems to have an impact on the location pattern of software firms.

In a consecutive step, we construct a Poisson regression model to predict the number of software firms per 1 km grid cell using a large set of location factors. In the regression analysis, we include 24 different agglomerations, infrastructure, socio-economic, topographical, and amenity location factors. We interpret the estimated regression coefficients to deduce the relationships between the location factors and software firm counts. Due to identification limitations [25,26] in our model, we abstain from tagging causal relationships and rather concentrate on the predictive performance of our model. However, by comparing our estimates with estimates from previous studies, we are able to discuss differences in the location factor-firm count relationships at different levels of geographic aggregation. We find that our model's overall performance is good as it is able to redraw the software firm pattern to a high degree and yields reasonable coefficients, which are in line with prior research. Inter alia, we are able to show that regional population centrality (which we operationalize using the Urban Centrality Index [27]) is a significant predictor of local software firm numbers at the microgeographic level. However, we also find that our model has a weak performance in highly segregated cities with quarters characterized by populations with dissimilar socio-economic profiles. Due to data limitations, we are not able to capture this microgeographic heterogeneity in the population structure. When considered at the aggregate city level (25 km grid), this systematic prediction error is levelled and the model yields systematic (spatially autocorrelated) errors in areas which were identified as software industry hotspots in the ESDA. This indicates that our model specification misses some crucial location factors present in these areas or some of the model's assumption are violated (e.g., the independence between individual location choices).

2. Data

In this study, we utilize geographic data from three main sources: The OpenStreetMap project, official geodata from statistics agencies, and the geocoded Mannheim Enterprise Panel dataset.

2.1. OpenStreetMap Data

OpenStreetMap (OSM) is a collaborative mapping project, which allows users to create freely accessible geographic data. In addition to roads, OSM includes map features such as retail shops, public transport facilities, and a variety of natural features. Concerns about the quality of this kind of user-generated geographic information seem natural and emerged shortly after the launch of the project in 2004 [28]. An array of studies investigated OSM data and assessed the geometric, attributive and temporal accuracy, and completeness of the mapped features. Besides intrinsic approaches, most of these studies compare OSM data to established commercial or official geographic data on road networks [29–31], buildings [32], and land use data [33–35]. Their results show, first, that OSM data is only slightly inferior to official/commercial data in terms of accuracy. Second, OSM data completeness increases at a rapid rate and is assumed to have reached or exceeded the level of completeness of commercial data in the meantime. Third, the completeness of OSM is positively correlated to population density and can be considered to be particularly suitable for the spatial analysis of urban areas. In this study, we use motorway accesses, airport locations, public transport stops, and several types of amenities obtained from an unmodified OSM full copy [20]. We also use OSM geodata as base data for our address locator described below.

2.2. Official Geodata

We use data issued by several German and European agencies, such as a downscaled population density grid issued by the European Environment Agency, which is available in 100 m resolution

and is based on communal census population data and land cover data (CORINE and LUCAS) [36]. Further, we use data on intercity railway stations and a 200 m resolution digital elevation model obtained from the German Federal Agency of Cartography and Geodesy. Socio-economic data on the level of education of the local workforce, wages, life expectancy, and number of resident students were obtained from the German Federal Institute for Research on Building, Urban Affairs and Spatial Development. Crime data was obtained from the German Federal Criminal Police Office. Due to the high data privacy awareness in Germany, the utilized socioeconomic data are only available at the municipality or district level. Local business tax rates were obtained from the German Federal Statistical Office. Local high speed broadband Internet availabilities are based on data from the German Federal Ministry of Transport and Digital Infrastructure. Locations of research institutes and universities were obtained from the German Federal Ministry of Education and Research. A 1 km resolution grid with the average commercial rent per square meter in 2016 was provided by the data company *empirica-systeme* GmbH, Berlin, Germany.

2.3. The Mannheim Enterprise Panel

The Mannheim Enterprise Panel (MUP) is a firm data base which covers the total stock of firms located in Germany. It contains about three million firm observations which are updated on a semi-annual basis. The data covers firm characteristics such as the branch of industry through NACE codes (a classification of economic activities in the European Union) and postal addresses [37]. Our definition of the software industry (the used NACE codes are: 62.01.0, 62.01.1, 62.01.9, 62.02.0, 62.03.0, 62.09.0, 63.11.0, 63.12.0) covers general programming activities, software development, web portals, data processing, and the development of web pages. In 2016, the MUP contained about 2.97 million active firms in Germany of which 70,009 are software firms (2.36%). We geocoded all MUP firm addresses using a self-made street type geocoding address locator based on an extended street network data model without house number interpolation. The geocoding results were assessed concerning their completeness and positional accuracy as proposed by Zandbergen [38].

The geocoding resulted in a completeness of 95.2% for the overall data set and 97.8% for the software firm subgroup in particular. The positional accuracy was verified by geocoding a random sample (n = 1000) of successfully geocoded addresses using a conventional geocoding service. The median positional offset between our geocoding results and the results obtained from the conventional service is 58 m (95% confidence interval: 53–69 m) and the mean is 252 m (95% confidence interval: 210–295 m), which is suitable for our level of analysis. A further analysis of the spatial distribution of the geocoding match rate aggregated at postal code areas revealed significant clustering (Moran's I = 0.13, *** $p < 0.001$) with few significant local clustering (Getis-Ord G_i*) of low match rates in rural areas. However, there is only a minor positive correlation (r_s = 0.006 ***) between the geocoding match rate and population density. Hence, known OSM data quality issues in rural areas (see above) do not seem to induce a systematic error in our geocoding results. We included an according control variable in the regression analysis (geocoding match rate at postal code area level) to cope with spatially varying geocoding completeness. We further used the MUP to identify the headquarter locations of the top 100 firms (by annual turnover) in Germany to include them as a location factor in the regression analysis.

3. Methods

Our analysis of the software firm location pattern is based on Exploratory Spatial Data Analysis (ESDA) and count data regression analysis.

3.1. Exploratory Spatial Data Analysis

Exploratory Spatial Data Analysis is a general term to describe the analysis of geospatial data in an explorative manner using a wide range of methods. It is similar to Geographic Knowledge Discovery [39], Spatiotemporal Data Mining [40], and GeoVisual Analytics [41,42]: Unexplored data is analyzed with the objective to uncover relevant and significant data characteristics or relationships

(e.g., data patterns, trends, correlations). Furthermore, the results should be summarized in an easily understandable way.

In this study, graphical techniques and geovisualization [43] are used to display and explore geographic data. Correlation analysis is used to measure the direction and strength of association between pairs of variables. We use the non-parametric Spearman's rank correlation coefficient r_s to measure the degree of monotonic relationship between variables. Quadrat analysis is used to evaluate the dispersion of point patterns by calculating their variance-to-mean ratio (VMR) using regular grids. The results of the quadrat analysis are used to assess whether the software firm location point pattern was produced by a random (homogenous Poisson) process [44,45]. We measure global spatial autocorrelation using Moran's Index I, which is arguably the most common measure to do so. We also utilize standardized Moran's I z-values, which allow us to compare I values between different levels of spatial aggregation. The generalized local G autocorrelation statistic G_i^* is used to evaluate local spatial association [46]. G_i^* was selected because we are mostly interested in detecting local pockets of positive spatial autocorrelation (e.g., "hotspots of the software industry"). Measures of spatial autocorrelation require us to hypothesize the spatial relationships in the study area [47]. We use the topological contiguity method with queen contiguity criterion (QNN) for our regular grids.

3.2. Count Data Regression Models

The most common way to model the relationship between location factors and the number of local firms per areal unit are count data regression models (CDM) [15]. The estimated coefficients from CDM provide evidence on how *ceteris paribus* variations in an explanatory variable affect the conditional mean of the number of local firm locations. However, it is not advisable to deduce causal relationships between the dependent variable and the explanatory variables without having a suitable identification strategy [26,48]. Relationships estimated in our regression analysis should be understood as correlations between our dependent variable (software firm counts) and a set of predictor variables (location factors).

We apply the most commonly used CDM: Poisson regression [26,49]. In a spatial setting, the data generating process can be understood as a spatial Poisson process. The standard (homogenous) spatial Poisson process generates points with complete spatial randomness (CSR) [44]. Spatial Poisson processes are used in many fields to model randomly distributed points [45,50]. An outcome Y is assumed to be Poisson distributed with a stationary density parameter λ. This density parameter defines both the mean and the variance of the distribution (equidispersion). A point pattern which features a spatially varying density parameter λ can be understood as a non-homogenous Poisson process. Here, the outcome Y depends on a location-dependent density parameter λ that varies systematically with a set of variables X (i.e., the location factors).

$$\begin{gathered} Y \sim \text{Poisson}(\lambda(X)) \\ E(Y) = \lambda(X)\ \text{Var} = \lambda(X) \end{gathered} \tag{1}$$

Hence, the local density parameter λ_i in cell i is conditional on the local values of $\mathbf{x}_i$:

$$\begin{gathered} y_i \mid \mathbf{x}_i \sim \text{Poisson}(\lambda_i) \\ E(y_i \mid \mathbf{x}_i) = \lambda_i\ \text{Var}_i = \lambda_i \end{gathered} \tag{2}$$

The effect of X on Y is defined by a set of unknown coefficients. These coefficients can be estimated in a Poisson regression, which is a generalized linear model with the natural logarithm as the link function. The parameter estimation is based on maximum likelihood. The expected count (i.e., the number of firms) in an area i of size A_i, given n location factors x, is then:

$$\hat{y}_i = \hat{\lambda_i} = e^{\ln(A_i)+\hat{\alpha}+\hat{\beta}_1 x_{1,i}+\ldots+\hat{\beta}_n x_{n,i}} \tag{3}$$

The coefficient $\exp(\hat{\alpha})$ is the offset, while $\exp(\hat{\beta})$ give the multiplicative effects of the location factors. The estimated coefficients can be reported as incidence-rate ratios (IRR) which make comparing rates easier. The IRR for a Δx_n change in x_n is $e^{\hat{\beta}_n \Delta x_n}$ (ceteris paribus). Cameron and Trivedi [26] recommend using robust standard errors for Poisson models.

We also use Negative Binomial regression (NBIN), which is a special case of Poisson regression [49]. In NBIN regression, it is assumed that an overdispersed Poisson process generated the point pattern under investigation. To cope with the additional variance, an additional shape parameter (over-dispersion parameter) is estimated, which allows for additional variance [26].

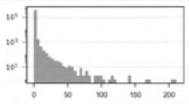
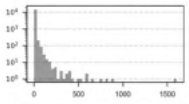
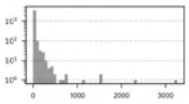
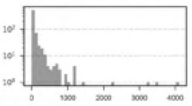
4. Results

In this section, we first present the results of our Exploratory Spatial Data Analysis (ESDA). Building on our findings from the ESDA, we construct a comprehensive set of location factors, which we use in a subsequent regression analysis. The results of the regression analysis are presented in the second part of this section. A detailed discussion of the results and their significance follows in Section 5.

4.1. Exploratory Spatial Data Analysis Results

Table 1 presents descriptive statistics of the software firm pattern aggregated at 1 km, 5 km, 10 km, and 25 km resolution grids. It can be seen that the variance-to-mean ratio (VMR) of the distribution strongly varies with the level of aggregation. At low levels of aggregation, the distribution is closer to equidispersion (indicating that the point generating process can be adequately modelled as a Poisson process). At higher levels of aggregation, the pattern appears to be increasingly clustered (over-dispersed). We conclude that Poisson regression is likely to be the appropriate regression model for low aggregation levels, while Negative Binomial regression, which can handle over-dispersed count data [49], seems to be more appropriate for higher levels of aggregation. These results show that the choice of level of aggregation highly influences the statistical characteristics of the spatial pattern under investigation and determines the choice of an appropriate statistical distribution.

Table 1. Descriptive statistics of the aggregated software firm location pattern.

Scale	Obs.	$\overline{X}$	$\tilde{X}$	SD	Min.	Max.	VMR	Histogram
1 km	361,453	0.19	0	1.64	0	211	14.12	
5 km	14,951	4.58	1	25.98	0	1604	147.39	
10 km	3860	17.74	4	87.07	0	3265	427.35	
25 km	671	102.06	27	301.74	0	4105	892.11	

Histogram: x = number of firms per cell; y = frequency.

Figure 1 maps the gridded distribution of software firms in Germany. An exemplary focus map of the German capital Berlin is shown to give an impression of the data's level of detail. It can be seen that the pattern largely redraws the population distribution: High numbers of software firms can be found in and around urban areas and low numbers in less densely populated areas. It is well known that the geographic pattern of economic activity is dominated by the influence of the population distribution: Humans tend to concentrate in specific areas, causing a high frequency of firm locations in those areas regardless of other factors. The population density can therefore be considered the reference pattern of the firm location distribution.

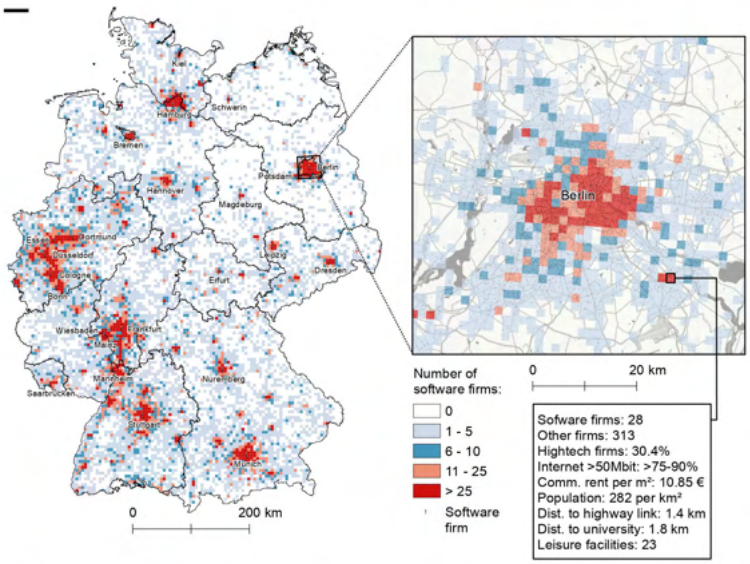


Figure 1. Overview (5 km scale) and zoom (1 km scale; with selection of location factors for exemplary cell) of the software firm location pattern.

However, Figure 2a indicates that software firms seem to have a location decision behavior different from the rest of the firm population. It can be seen that the share of software firms in the overall firm population is not distributed randomly over the study area (Moran's I = 0.36 ***; the standardized I values plotted in Figure 2b show that this applies to all scales). Instead, software firms are disproportionally frequent in and around urban areas and even form statistically significant ($p \leq 0.05$) hotspots (Getis-Ord G_i*) in the areas of Munich, Stuttgart and Rhine-Main (around Frankfurt). On the contrary, the absence of high software industry shares and hotspots in the very densely populated and large Ruhr area (around Essen) indicates that high population density alone does not necessarily imply large proportions of software firms in the local firm population.

Figure 3 helps to further investigate the relationship between firm numbers and population density by plotting Spearman's correlation coefficients for four levels of geographic aggregation. It can be seen that the positive monotonic relationship becomes stronger with the level of aggregation. Aggregated at 25 km, both software firms ($r_s = 0.94$ ***) and the total stock of firms ($r_s = 0.97$ ***) exhibit similarly strong monotonic relationships with population numbers. At the 1 km scale, software firm numbers show a distinctively lower correlation to local population numbers ($r_s = 0.38$ ***) than the rest

of the firm population (r_S = 0.65 ***). This indicates that population numbers alone do not predict the number of software firms very well at low levels of geographic aggregation.

Combining the findings from Figure 2 (large shares of software firms in densely populated areas) and Figure 3 (weaker correlation between software firm numbers and population numbers at the microgeographic level), which seem counterintuitive at first sight, leads us to the hypothesis that software firms do indeed locate in urban regions but prefer the less densely populated areas within cities (e.g., suburbs). Given that the overall firm population is largely dominated by firms from walk-in customer oriented sectors (retail, gastronomy, and personal services) it seems reasonable to assume that these firms seek to locate in the densest areas of cities (i.e., the city center/central business district). Software firms, on the other hand, are not dependent on walk-in customers and may locate disproportionally often in less dense areas, which are usually characterized by lower rents, but still offer most of the benefits of an urban environment. This location choice behavior, which we try to model in the upcoming sub-section, may lead to the observed location pattern of software firms.

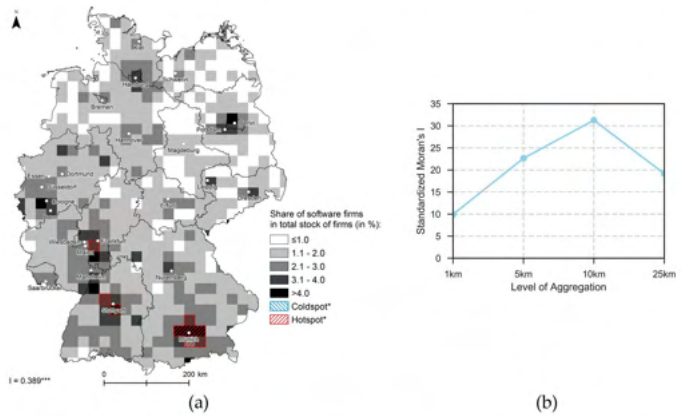


Figure 2. (**a**) Share of software firms in total stock of firms (25 km scale); (**b**) and standardized Moran's I by 1 km, 5 km, 10 km, and 25 km level of aggregation.

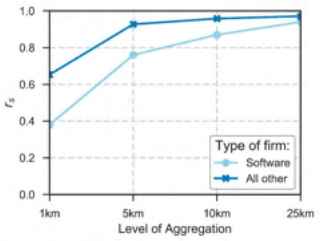


Figure 3. Correlation (r_S) between firm counts and population numbers by level of aggregation.

4.2. Regression Analysis Results

Based on the findings in the previous section, we specify a comprehensive model that correlates the number of software firms per 1 km grid cell to the values of 24 distinct location factors from five

groups: agglomeration, infrastructure, socio-economic, quality of life and amenities, and other location factors. Poisson regression was identified as the appropriate method to model the software location pattern at the 1 km level of aggregation. The location factors and the estimated coefficients yielded by the Poisson regression are given in Table 2. The regression coefficients are given as incidence-rate ratios (IRR) and can be read as follows: An increase in the population by 1 unit (equaling 100 inhabitants) is associated to an 1.081 (+8.1%) times larger number of local software firms and an increase in the distance to the next motorway access by 1 unit (1 km) is associated to an 0.977 (−2.3%) times smaller number of local software firms. The robust standard errors of the estimated coefficients are given in parentheses.

Table 2. Location factors and estimated coefficients with robust standard errors in parentheses.

Location Factor	Description	IRR
Agglomeration Location Factors		
Firm density	Number of local firms (in 10)	1.028 *** (0.003)
Firm density^2	Squared number of local firms (in 10)	0.999 *** (0.000)
High-tech firms	Proportion of high-tech firms in local stock of firms (in %)	1.021 *** (0.000)
Major firms	Distance to next major firm in km	0.998 *** (0.000)
Commercial rent	Difference local rent to mean rent in neighborhood (in Euro)	1.127 *** (0.12)
Population	Population per cell (in 100)	1.081 *** (0.003)
Population^2	Squared population per cell (in 100)	0.999 *** (0.000)
Population centrality	Urban Centrality Index (in 0.1 UCI) high value ≙ monocentricity	1.079 *** (0.192)
Infrastructure Location Factors		
Broadband Internet	Availability of ≥50 mb Internet (categories) high value ≙ low availability of Internet	0.764 *** (0.009)
Motorway	Distance to nearest motorway access (in km)	0.977 *** (0.001)
Railway	Distance to nearest main-line railway station (in km)	0.998 *** (0.000)
Airport	Distance to nearest main airport (in km)	0.998 *** (0.000)
Public transport	Weighted count of public transport stops	1.000 (0.001)
Socio-economic Location Factors		
Wages	Median income of full time employee (in 100 Euro)	1.005 (0.003)
Universities	Distance to nearest university (in km)	0.980 *** (0.000)
Research institutes	Number of research institutes	1.004 (0.036)
Educated workforce	Proportion of graduate employees in %	1.063 ***(0.006)
Students	Proportion of students in local population in %	0.986 *** (0.003)
Business tax	Business tax factor (in 100) high values ≙ high taxes	0.925 ** (0.023)
Quality of Life and Amenities Location Factor		
Life expectancy	Mean life expectancy of population	1.092 *** (0.012)
Crime	Violent and street crime incidents per 1000 inhabitants	1.021 (0.015)
Recreation	Number of recreational, community, and sports facilities	1.056 *** (0.008)
Culture	Number of cultural facilities	1.015 0.017
Leisure	Number of gastronomy, nightlife, and general leisure facilities	1.002 (0.002)
Other		
Terrain	Difference in elevation to mean neighborhood elevation (in 100m) high values ≙ hillside location	0.919 *** (0.004)
Geocoding control variable	Geocoding match rate (in %) high value ≙ high completeness	1.018 *** (0.002)

** $p \geq 0.01$, *** $p \geq 0.001$.

4.2.1. Interpretation of Regression Coefficients

We included the square of both the number of firms and the population to control for a nonlinear relationship with the number of software firms. The reason for taking this approach is because it is frequently stated that density may have an inverse u-shaped influence (an initially positive effect which, from a certain point on, turns into a negative effect, e.g., due to environmental pollution in very dense cities) on site attractiveness [21,51]. This seems to be confirmed by our estimation results. Both the number of firms and the population have a highly significant positive effect on the number of local software firms. The significant negative coefficients of their squared counterparts indicate the

assumed inverse u-shaped relationship. Population centrality is also estimated to have a significant effect. Increasing the monocentricity in the regional population distribution leads to an increase in the number of software firms. A high proportion of high-tech firms (classification according to [52]) in the local stock of firms is estimated to increase the number of software firms significantly as well. Increasing distance to major firms is associated to a significant decrease in the number of software firms. Higher commercial rents, expressed as the deviation from the mean rent in the immediate neighborhood (queen contiguity), are estimated to have a positive and significant influence. The model confirms that software firms locate in monocentric and dense areas, but avoid the densest areas. Geographic proximity to business customers (in the form of high-tech and major firms) matter as well. The strong positive effect of high (relative) commercial rents makes it a good predictor. However, there is severe endogeneity stemming from the simultaneity to the dependent variable (attractive locations causing high software firm numbers, which in turn cause high rents), an issue which is addressed in the Discussion section.

Increasing the distance to the motorway, railway, and aerospace network is associated with a significant decrease in the number of software firms. Access to public transport, on the other hand, has no significant effect. Decreasing the availability of broadband Internet is estimated to decrease the number of software firms significantly. These results indicate that software firms prefer locations with decent personal transport infrastructure and available broadband Internet. Local public transport does not seem to be of importance though.

The closeness to a university significantly increases the number of software firms. Counterintuitively, having a high proportion of students in the local population has a significant negative effect. The number of nearby research institutes and wages have no significant effect. Having a high share of graduate employees in the total stock of employees increases the number of software firms significantly, while high business taxes have a significant negative effect. These results indicate that software firms seek to locate close to universities and regions which offer an educated workforce and low business taxes. While this matches the image of the software industry as a knowledge intensive sector, the negative effect of students seems rather implausible. It shall be noted that wages, educated workforce, student population, and business tax levels are measured at a broad geographic scale (counties) and should therefore be understood as regional controls rather than microgeographic predictor variables.

High life expectancy is associated with a significant increase in the number of software firms. The same is true for the number of nearby recreational amenities. Crime rates, cultural amenities, and leisure amenities have no significant effect. These results indicate that a high quality of life does indeed increase the local attractiveness towards knowledge intensive software firms, which heavily rely on highly qualified and creative individuals who are assumed to have a strong preference for areas offering a high quality of living. The breakdown into different amenity types shows that only nearby recreational amenities seem to matter though. However, it should be kept in mind that other amenities could still play a role at different spatial scales: Having a cultural amenity in a city may increase the attractiveness of the city as a whole, but not necessarily the attractiveness of the immediate neighborhood around it.

We included a terrain variable, which captures the difference in elevation between focal cells and their neighborhood. This allows us to distinguish adjacent cells with almost identical location factors (e.g., distance to infrastructure) but different topographies (i.e., hillside location versus valley location). We assume that the identification of hillside location greatly improves the microgeographic predictive performance of our model. The large estimated negative and significant effect supports this assumption. The added geocoding control variable improves the predictive performance as well.

4.2.2. Model Fit and Spatial Residual Analysis

Model fit can be rather difficult to assess and there are a variety of measures of how adequately the model represents the data. We apply different goodness of fit measures (GoF) and spatial residual

analysis to assess the fit and adequacy of our model. Table 3 presents some GoF for the model based on the Poisson distribution assumption and the corresponding values from an estimation using Negative Binomial regression (NBIN). The pseudo-R^2 measures the badness of fit (deviance) of the model, i.e., how much worse the model is than a perfectly fitting model [49], and can only be interpreted against another model's pseudo-R^2. According to the root-mean-square error (RMSE) and pseudo-R^2 measure, the NBIN model's fit is inferior to the Poisson model. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are widely used measures to support model selection [26,49]. Both indicate that the NBIN model has the better fit (indicated by smaller values), contrary to the RMSE and pseudo-R^2.

Table 3. Poisson and Negative Binomial model goodness of fit.

GoF Measure	Poisson	Negative Binomial
Pseudo-R^2	0.58	0.33
RMSE	1.36	483,735
AIC	211,603	179,705
BIC	211,892	180,004

Figure 4 plots the frequencies of observed against predicted software firm counts (as proposed by [26]). It can be seen that the NBIN model yields severely overestimates firm counts. This is reflected by the RMSE but not the AIC and BIC, which are less sensitive towards severe over- and underestimation. In line with our prior assumptions, based on the descriptive statistics in Table 1, the Poisson model seems to be the better prediction model at this scale. However, it can also be seen that both models underestimate the number of zeros and low count cells. This indicates that an excess zero problem might be prevalent in our model. This can be the case if the study area includes areas (i.e., raster cells) that would never host any firms (e.g., water bodies). One way to deal with such structural zeros is to use Zero Inflated Poisson regression [48].

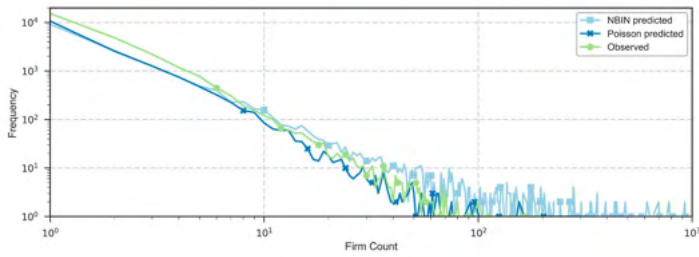


Figure 4. Frequencies of observed and predicted software firm counts.

Figure 5 maps the regression residual (prediction error) aggregated on a regular 5 km grid. Warm colors indicate cells which host more software firms than predicted by the model (underestimation), while cold colors indicate overestimated software firm counts. It can be seen that both under- and overestimation occur mostly in urban areas. Munich, which was identified as a software industry hotspot in the ESDA, has a notable contiguous "catchment area" where software firm numbers are uniformly underestimated, while firm numbers in the city center are overestimated. This pattern is reoccurring in and around other metropolitan areas as well. Due to the aggregation Berlin conveys a more "blue" impression in the Germany overview map, whereas the zoomed Berlin map (upper right hand side in Figure 5; original 1 km grid) shows largely red areas. The detailed map shows contiguous areas of severe overestimation (southwest) and underestimation (east and northeast) in different parts

of the city. Such positive autocorrelation in the residual pattern indicates that the prediction fails systematically in some areas. This may be due to one or several omitted explanatory variables or violations of the Poisson distribution assumption of independent events, which may be present if software firms themselves are a significant location factor, resulting in a self-enforcing process of accumulating firm locations. One possible explanation for the systematic prediction errors in northeast Berlin (around the district of *Prenzlauer Berg*) and southwest Berlin (around the district of *Wilmersdorf*) is unobserved heterogeneity in the sociodemographic composition of the local population. While Prenzlauer Berg is known for its young, alternative resident population and is often given as an example of ongoing gentrification, Wilmersdorf is a more middle-class residential area. The sociodemographic profile of Prenzlauer Berg could be considered a breeding ground for knowledge-intensive start-ups which rely on creative employees and entrepreneurs [53,54]. This location factor is not captured in our model but we propose solutions in the discussion section of this paper. Another case of a potentially omitted variable bias is highlighted in the detailed map on the lower right hand side of Figure 5. It highlights an area of isolated under-prediction in the district of *Adlershof* in the southeast of Berlin. For illustration reasons, the firm pattern with overlapping locations was transformed to a kernel density map (triweight kernel, uniform weights, 250 m bandwidth). The cause for this under-prediction is the presence of Germany's largest science park, which host several technology centers with office space dedicated to software firms [55].

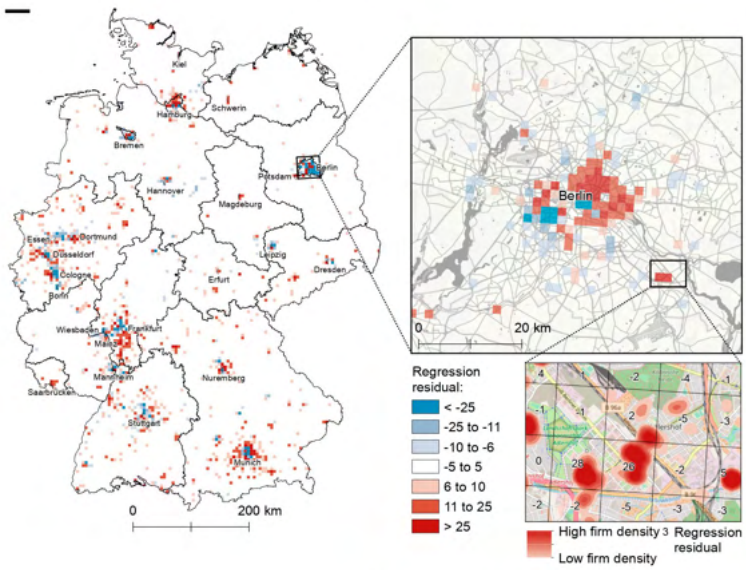


Figure 5. Regression residual aggregated at 5 km raster (**left**) and original 1 km grid (**right**).

Similar patterns as described above can be seen in other cities in Figure 5 too, resulting in significant spatial autocorrelation in the spatial distribution of the residual (Moran's I = 0.12 ***). However, with increasing aggregation, the spatial autocorrelation diminishes and becomes insignificant at the 25 km scale (see Figure 6a). At the 25 km scale (with single cities roughly aggregated into single cells) it seems that most local errors are levelled by the geographic aggregation. However, Figure 6b reveals that local pockets of spatial autocorrelation (G_i^*) still exist. The described prediction disparity

in Berlin is still present for example, because Berlin was, by chance, divided uniformly into four cells (cf. MAUP as mentioned above). This results in significant ($p < 0.05$) clustering of negative residuals (overestimation) in the south of Berlin (coldspot) and a hotspot of positive residuals (underestimation) in the north. Interestingly, other residual clusters occur mainly in areas which were identified as hotspots of the software industry (see Figure 6b).

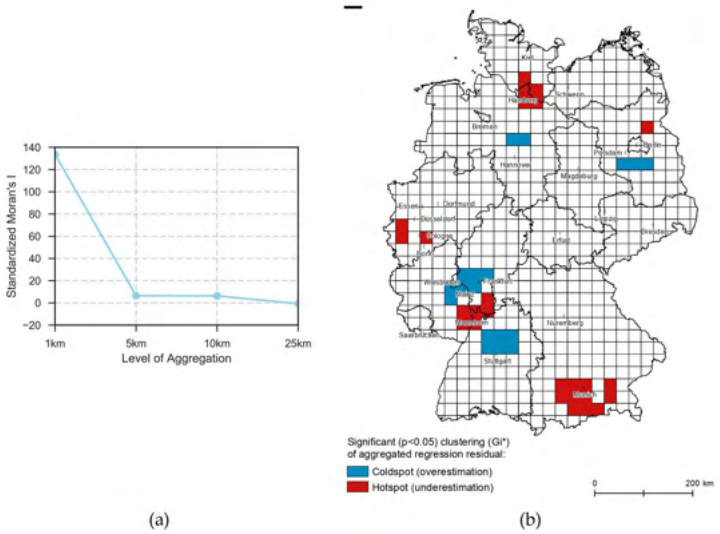


Figure 6. (**a**) Standardized Moran's I of regression residual aggregated at different levels of aggregation; (**b**) Significant clustering of regression residual aggregated at 25 km grid.

These results indicate that our prediction model produces good results at the microgeographic level, which can be used to generate even more decent software firm count predictions when aggregated at a larger scale. However, we find that our model shows weak performance in highly segregated cities with quarters characterized by populations with dissimilar socio-economic profiles. Due to data limitations, we are not able to capture this microgeographic heterogeneity in the population structure. At higher aggregation levels, the model fails to predict the correct firm numbers in areas with an extraordinary concentration of the software industry. This again may be seen as an indicator for unobserved location factors present in these areas, which go beyond the conventional set of location factors used in this study.

5. Discussion

In this section, we first discuss the coefficients resulting from the regression analysis results and interpret them in perspective of previous studies. We then discuss the model's fit and weaknesses, and the results of the spatial residual analysis. We also highlight opportunities for future research.

5.1. Discussion of Regression Coefficients

5.1.1. Agglomeration Location Factors

Agglomeration economies (and more generally density) are one of the earliest and most studied determinants of industrial location [56–58]. Our approach of modelling agglomeration economies as a function of density is a common empirical strategy [59]. Agglomeration economies manifest via dense customer-supplier linkages, labor pooling, knowledge spill overs, and high quality infrastructure. We included both the number of firms and the number of inhabitants as measures of density, even though these two are highly correlated, because they can differ at the microgeographic level as we showed in the ESDA. Empirical evidence for a positive effect of agglomeration on the location decision of firms, as we find it in our study, is confirmed in many studies [22,60–64]. There is a general agreement that the effect of density on location decisions is non-linear and follows an inverted U-shaped profile [15]. This means that, from a certain threshold, agglomeration diseconomies, i.e., negative economic effects caused by agglomeration, appear. We model this by including the squared number of firms and inhabitants. The estimated coefficient, which is negative and significant, confirms the assumed inverted U-shape effect of density on software firm location numbers.

We further included the Urban Centrality Index [27], which we calculated based on a 5 km grid, to measure the degree of centrality in the regional population distribution. The index ranges from 0.0 (absolute polycentricity) to 1.0 (absolute monocentricity). To our knowledge, population centrality has not been considered as a relevant location factor yet. Our analysis reveals that increasing population centralization is accompanied by an increase in software firm numbers. This indicates that firms (ceteris paribus) seek to locate in centrally located regions.

Software firms' products and services are demanded disproportionally intensely by high tech companies [65,66]. Hence, we included the proportion of high tech firms in the local firm population (excluding software firms). The large, positive and significant coefficient seems to confirm the importance of customer proximity for software firms. However, similar location choice behavior of software firms and high-tech firms could also cause this strong correlation.

Large firms may have a major impact on the location decision of software firms. We included the distance to the nearest headquarter of one of the 100 biggest (by turnover) firms in Germany to control for that. Our results suggest that software firms tend to locate nearby at least one of these major firms. Again, this correlation could also be caused by a similar location choice behavior and not by a causal positive influence of major firms on software firm numbers.

Commercial rent is a widely used proxy for the attractiveness of sites and measures the willingness-to-pay of firms for commercial property. Consequently, rents are often used as the dependent variable in empirical studies researching industrial location choice [21,63]. Rents are therefore highly endogenous when used as a location factor. Given that our considered industry only constitutes a minor fraction of the overall firm population (2.36%), rents may be considered as given (exogenous) to our software industry subset. Because rents exhibit severe regional disparities and a certain local rent level might be high at a nationwide perspective but comparatively low in the region, we included the difference to the mean commercial rent in the surrounding area (8 adjacent cells and the focal cell) as our commercial rent location factor. The estimated coefficient is large, positive, and significant, indicating commercial rents as a strong predictor of site attractiveness.

5.1.2. Infrastructure Location Factors

Transport infrastructures have been extensively studied in industrial location analysis and the positive effects of easily accessible transport infrastructure have been confirmed in many studies [62,67–69]. Unlike manufacturing, software firms are less dependent on moving inputs and outputs and rather rely on human capital. Thus, we included location factors which relate to the transportation of persons. In a highly developed and densely populated country such as Germany primary and secondary roads can be considered ubiquitous. Hence, we only included the distance

to the closest motorway link to measure accessibility to the road network. We further included the distance to the nearest long-distance railway station and major airport. A weighted count of local public transport facilities (bus stops, tram stops etc.) was also included. The weights are based on the transport capacities of the considered mean of transportation [70]. As software firms are highly dependent on the Internet, we also include the local availability of broadband Internet. Except for public transports, our analysis confirms the assumed positive relationship between advantageous infrastructure and software firm counts.

5.1.3. Socio-Economic Location Factors

Arguably the most researched socio-economic location factors are taxes, wages, and education of the local workforce. Most studies find a positive impact of workforce education [62,71], and proximity to universities and public research institutes [72,73] on firm numbers (especially for knowledge-intensive industries). High wages, on the other hand, are found to have a negative effect on firm numbers [61,74,75]. The same is true for high tax rates [61,68,75]. While our study can confirm the latter, wages have no significant effect on software firm numbers in our model. However, wages are strongly correlated ($r_s = 0.49$ ***) to the proportion of university graduated employees in the local workforce, which is found to have a strong positive effect on local software firm numbers (indicating multicollinearity). The software industry's need for highly educated employees is further emphasized by the strong positive effect of nearby universities. The number of local public research institutes has no significant effect though. It needs to be kept in mind that some socio-economic location factors are measured at a low spatial resolution (district and municipality level). While this is of no concern for tax levels, the share of graduate employees and wages can differ significantly within districts (ecological fallacy [4,76]). The lack of socio-economic location factors at the microgeographic level could in fact be a major issue of our model as we discuss further below.

5.1.4. Quality of Life and Amenities Location Factors

Qualified labor, the software industry's arguably most crucial input, is assumed to have a strong preference for a rich social and cultural life [53,77]. If software firms follow skilled labor [78] or locate at sites which attract skilled labor, the local quality of life becomes an important location factor. Quality of life is often measured through (exogenous) climate amenities [79] and the arguably more appropriate but endogenous urban consumption amenities [22,80]. We employed three different types of amenities in our study: Recreational, cultural, and leisure amenities. Recreational amenities encompass sports and natural spaces such as parks, playgrounds, and sports centers. Cultural amenities include features such as arts centers, cinemas, and museums. Leisure amenities cover all types of gastronomy (bars, pubs, and restaurants) as well as nightlife venues (e.g., nightclubs). To our knowledge, this is the first time a location study differentiates between these types of urban amenities.

Our results suggest that only recreational amenities are significant to software firm location choices. However, we suppose that measuring urban amenities at a different scale may yield different results. Having a theatre within the immediate neighborhood of a software firm may not be highly relevant, but having one in the same ward or city may be. The same is true for a vibrant night life, for example. Thus, future research could use location factors which operationalize urban amenities at different and maybe more appropriate scales.

We further included the local mean life expectancy, which was found to be the most import predictor for peoples' quality of life [81], and local levels of street and violent crime. While the estimated coefficient for life expectancy is large, positive, and significant, crime has no significant effect. Again, we assume that the spatial resolution of these two location factors (municipality level) are too low and unobserved within-city heterogeneity may compromise our results.

5.1.5. Other Location Factors

We also included a location factor that captures the terrain in the considered cell. We did so to be able to distinguish between neighboring and almost identical cells (e.g., considering their distance to the next motorway access) but different topographical properties (e.g., one is located at a steep hillside). Such a distinction becomes more important when small geographic units are analyzed and terrain roughness is not equalized by aggregating the smaller geographic units into larger ones. By including the difference between the mean elevation within the considered cell and the mean elevation in the surrounding area (8 adjacent cells plus the focal cell), we are able to identify hillsides and valleys. The estimated coefficient indicates that we created an important microgeographic predictor. Lastly, we included the local geocoding match rate to cope with unevenly distributed geocoding match rates.

5.2. Discussion of Model Adequacy

The prediction model based on Poisson regression, which is the most commonly used count data model (CDM) in firm location analysis [15], turned out to yield plausible results at the microgeographic level. The Poisson CDM generated better software firm count predictions than the Negative Binomial CDM, just as we assumed from the results of the Exploratory Spatial Data Analysis. We identified excess zeros as an issue in our prediction model. Excess zeros may arise if so called structural zeros are present in the dataset. Liviano and Arauzo-Carod [51] discuss the problem and interpretation of zero counts in count data models. They find that the zero excess problems may arise especially at very detailed geographical levels because most of the potential sites will never host any firms. They propose zero-inflated CDM to cope with that issue. In the first stage of such a two stage zero-inflated regression, the probability that each area with an observed count of zero is in one of two latent groups is estimated. The first group are those areas that would never host any firms (structural zeros) and the second group are those which might potentially host a firm in general [49]. For future research, we propose to use detailed land use data in a zero-inflated Poisson regression (ZIP) model to determine the membership of each grid cell to one of the two latent groups. Water bodies and forest, for example, could be identified as structural zero cells by doing so. In the second part of the ZIP the land use variable would be excluded. We expect that such an approach would yield better results than the pure Poisson regression approach chosen in our study.

Multicollinearity is likely to be present in our model. However, as multicollinearity is not a serious issue to the predictive performance of the model, it may cause the coefficient estimates to be unreliable [48] (i.e., the estimated coefficients may not coincide with the true influence of the explanatory location factor on the number of software firms). A possible solution for instable estimates in Poisson regression models due to multicollinearity is the application of a Poisson Ridge regression estimator [82].

Another deficit lies in the location factor operationalization. Indeed, we are able to show that OpenStreetMap data are suitable for microgeographic location analysis regarding their spatial accuracy, completeness, and type breakdown. The use of disaggregated amenity types suggests a promising approach towards more detailed firm location choice models. However, our analysis results indicate that the correct operationalization of location factors becomes even more difficult at the microgeographic level: Different location factors operate at different scales (*scale sensitivity*). A vibrant night life, for example, may have a positive impact on site attractiveness at the city level [53,54], but firms may still prefer calm neighborhoods (resulting in a negative influence of at a more detailed geographic scale). New scale-sensitive measures [83] or the use of spatially lagged variables [7] may help to solve this issue in future research.

The model's most serious issue is unobserved heterogeneity in the socio-economic characteristics of the population. This problem is most severe in cities, which often feature segregated populations and districts with very different sociodemographic profiles. The socio-demographic geodata used in our model does not have the appropriate geographic detail needed for a throughout consistent microgeographic firm count prediction. The imputation of macro-level socio-economic population

characteristics to the micro level causes the model to generate systematic (spatial autocorrelated) errors in some city districts. This became clear in the discussed Berlin districts of Prenzlauer Berg and Wilmersdorf: While the sociodemographic profile of Prenzlauer Berg can be considered a breeding ground for knowledge-intensive start-ups from the software industry, Wilmersdorf's more middle-class residential area is less so. Due to low resolution socio-economic geodata, both city districts have the same population profile, which causes our model to systematically overestimate the number of software firms in Wilmersdorf and to underestimate them in Prenzlauer Berg.

This issue may be tackled in two ways in future research. One solution may be the use of other regression models. Either a regional (city district) fixed effects regression model [26,48], which requires panel data where longitudinal observations are captured for the same geographic area. Given that appropriate longitudinal data is available, such a study layout would also allow for the analysis of the evolution of firm patterns over time. Spatial Error regression models, which can handle variables in the error term that are likely to be similar in adjacent regions, area another possible solution to spatial autocorrelated residuals [84,85]. Another straightforward option is the inclusion of geographically more detailed socio-economic geodata, which is not available in Germany though. In regions without such detailed geodata, future research may use alternative data sources and proxy data. New impulses for such data could come from the rich body of research concerned with the analysis of crowdsourced geodata and other Volunteered Geographic Information from social network sites (e.g. *Twitter*, San Francisco, CA, USA). Recent studies have shown that such data can be used to derive information on socio-demographics [86,87]. We also assume that OSM data has great potential in microgeographic location analysis, when appropriately deployed. The differences between the discussed Berlin districts of Prenzlauer Berg and Wilmersdorf also manifest in very different fertility rates. In 2016, Prenzlauer Berg had the highest fertility rate in Berlin, while Wilmersdorf had the second lowest out of 23 districts [88]. This condition could, for example, be measured by a proxy using OSM data on the number of day-care centers and pre-schools in the two districts.

6. Conclusions

In this paper, we presented a software firm location prediction model using Poisson regression and OSM data. We used a comprehensive dataset of three million street-level geocoded firm observations to explore the location pattern of software firms in an Exploratory Spatial Data Analysis (ESDA). Then, we used a variety of predictor variables to assess spatial factors that influence the location process of software firms. Our study shows that OSM can be used to construct location factors which are suitable for an encompassing microgeographic firm location analysis. Its coverage, completeness, and degree of detail makes OSM a promising yet underused data source in the context of firm location analysis and geographic economic analysis in general, also because the data are easy to obtain for many parts of the world. We also highlighted further application opportunities for OSM and other VGI data (e.g., geocoded data from social network sites) in this context. Our research questions defined in the introductory section can be answered as follows.

6.1. RS1: Scale-Robust Location Factors

We found that the microgeographic level of analysis provides new insights into the firm allocation process, but also that most location factors are scale robust. That is, our findings with respect to location factor effects are in line with prior research using aggregated spatial units. However, for a thorough understanding of MAUP scaling effects on location factor-firm correlations, our encompassing regression specification should be applied to different levels of geographic aggregation. Such an analysis could also investigate whether some location factors are more scale sensitive than others and whether the chosen operationalization approach alters the estimated effect of the location factors (e.g., "proximity to universities" could be measured by a binary variable, a count variable, or a continuous distance variable; recent research indicates that distance-based methods may be scale-robust [89–91]).

6.2. RS2: Microgeographic Location Prediction

We demonstrated that our microgeographic prediction model is able to predict the location of software firms to a satisfying degree, but it comes with particular requirements to the statistical model and the data employed in the analysis. The detailed level of geographic aggregation requires the researcher to employ a statistical model, which is adapted to the specific requirements of the level of analysis. In our specific case, statistical over-dispersion is less problematic, whereas excess zeros are a serious issue. At the same time, our analysis requires high resolution geodata, which may not be available in all domains. In our study, low resolution geodata on socio-economic population characteristics lead to unobserved microgeographic heterogeneity within cities, causing systematic prediction errors.

Acknowledgments: The authors thank the Centre for European Economic Research for providing the analyzed firm dataset. We also want to thank *empirica-systeme GmbH* for providing us with very helpful data on commercial rent. A special thank is due to Christian Rammer and René Westerholt who contributed valuable help and advice.

Author Contributions: Jan Kinne and Bernd Resch designed the study. Jan Kinne gathered, pre-processed, analyzed and visualized the data. Jan Kinne and Bernd Resch wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strotmann, H. Entrepreneurial survival. *Small Bus. Econ.* **2007**, *28*, 87–104. [CrossRef]
2. Capello, R. Classical contributions to location theory. In *Handbook of Regional Science*; Fischer, M.M., Nijkamp, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 507–526.
3. Clark, W.A.V.; Avery, K.L. The effects of data aggregation in statistical analysis. *Geogr. Anal.* **1976**, *8*, 428–438. [CrossRef]
4. Manley, D. Scale, aggregation, and the modifiable areal unit problem. In *Handbook of Regional Science*; Fischer, M.M., Nijkamp, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1157–1171.
5. Flowerdew, R. How serious is the modifiable areal unit problem for analysis of English census data? *Popul. Trends* **2011**, *145*, 106–118. [CrossRef] [PubMed]
6. Bluemke, M.; Resch, B.; Lechner, C.; Westerholt, R.; Kolb, J.-P. Integrating geographic information into survey research: Current applications, challenges and future avenues. *Surv. Res. Methods* **2017**, *11*, 307–327.
7. Arauzo-Carod, J.M.; Manjón-Antolín, M. (Optimal) spatial aggregation in the determinants of industrial location. *Small Bus. Econ.* **2012**, *39*, 645–658. [CrossRef]
8. Lee, Y. Geographic redistribution of US manufacturing and the role of state development policy. *J. Urban Econ.* **2008**, *64*, 436–450. [CrossRef]
9. Garrett, T.A. Aggregated versus disaggregated data in regression analysis: Implications for inference. *Econ. Lett.* **2003**, *81*, 61–65. [CrossRef]
10. Cherry, T.L.; List, J.A. Aggregation bias in the economic model of crime. *Econ. Lett.* **2002**, *75*, 81–86. [CrossRef]
11. Amrhein, C.G. Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environ. Plan. A* **1995**, *27*, 105–119. [CrossRef]
12. Arauzo-Carod, J.-M. Industrial location at a local level: Some comments about the territorial level of the analysis. *Tijdschr. Voor Econ. Soc. Geogr.* **2008**, *99*, 193–208. [CrossRef]
13. Manjon-Antolin, M.; Arauzo-Carod, J.M. Locations and relocations: Modelling, determinants, and interrelations. *Ann. Reg. Sci.* **2006**, *47*, 131–146. [CrossRef]
14. Briant, A.; Combes, P.P.; Lafourcade, M. Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *J. Urban Econ.* **2010**, *67*, 287–302. [CrossRef]
15. Arauzo-Carod, J.-M.; Liviano-Solis, D.; Manjon-Antolin, M. Empirical studies in industrial location: An assessment of their methods and results. *J. Reg. Sci.* **2010**, *50*, 685–711. [CrossRef]
16. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
17. Elwood, S.; Goodchild, M.F.; Sui, D.Z. Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Ann. Assoc. Am. Geogr.* **2012**, *102*, 571–590. [CrossRef]

18. Goodchild, M.F.; Longley, P.A. The practice of geographic information science. In *Handbook of Regional Science*; Fischer, M.M., Nijkamp, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1107–1122.
19. Sui, D.; Goodchild, M. The convergence of GIS and social media: Challenges for GIScience. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1737–1748. [CrossRef]
20. OpenStreetMap Foundation OpenStreetMap. Available online: http://www.openstreetmap.org (accessed on 1 November 2016).
21. Ahlfeldt, G.M. *Urbanity*; SERC Discussion Paper, 136; London School of Economics and Political Science: London, UK, 2013.
22. Möller, K. *Culturally Clustered or in the Cloud? Location of Internet Start-Ups in Berlin*; London School of Economics: London, UK, 2014; Volume 157.
23. Ahlfeldt, G.M.; Richter, F.J. *Urban Renewal after the Berlin Wall*; SERC Discussion Paper, 151; London School of Economics and Political Science: London, UK, 2013.
24. Grasland, C.; Madelin, M. *The Modifiable Areas Unit Problem*; ESPON: Luxembourg, 2006.
25. Wooldridge, J.M. *Econometric Analysis of Cross Section and Panel Data*; The MIT Press: Cambridge, MA, USA; London, UK, 2002.
26. Cameron, C.; Trivedi, P. *Microeconomics Using Stata*, Revised ed.; Stata Press: College Station, TX, USA, 2009.
27. Pereira, R.H.M.; Nadalin, V.; Monasterio, L.; Albuquerque, P.H.M. Urban centrality: A simple index. *Geogr. Anal.* **2013**, *45*, 77–89. [CrossRef]
28. Flanagin, A.J.; Metzger, M.J. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148. [CrossRef]
29. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. B Plan. Des.* **2010**, *37*, 682–703. [CrossRef]
30. Girres, J.F.; Touya, G. Quality assessment of the French OpenStreetMap dataset. *Trans. GIS* **2010**, *14*, 435–459. [CrossRef]
31. Neis, P.; Zielstra, D.; Zipf, A. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Futur. Internet* **2011**, *4*, 1–21. [CrossRef]
32. Hecht, R.; Kunze, C.; Hahmann, S. Measuring completeness of building footprints in OpenStreetMap over space and time. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 1066–1091. [CrossRef]
33. Arsanjani, J.J.; Mooney, P.; Zipf, A.; Schauss, A. Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In *OpenStreetMap in GIScience: Experiences, Research, and Applications*; Arsanjani, J.J., Zipf, A., Mooney, P., Helbich, M., Eds.; Springer: Heidelberg, Germany; New York, NY, USA; Dordrecht, The Netherlands; London, UK, 2015; p. 324.
34. Arsanjani, J.J.; Vaz, E. An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *35*, 329–337. [CrossRef]
35. Dorn, H.; Törnros, T.; Zipf, A. Geo-Information comparison with land use data in Southern Germany. *Int. J. Geo-Inf.* **2015**, *4*, 1657–1671. [CrossRef]
36. Gallego, F.J. A population density grid of the European Union. *Popul. Environ.* **2010**, *31*, 460–473. [CrossRef]
37. Bersch, J.; Gottschalk, S.; Müller, B.; Niefert, M. *The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany*; ZEW Discussion Paper, 14-104; Centre for European Economic Research: Mannheim, Germany, 2014.
38. Zandbergen, P.A. A comparison of address point, parcel and street geocoding techniques. *Comput. Environ. Urban Syst.* **2008**, *32*, 214–232. [CrossRef]
39. Miller, H.J.; Han, J. *Geographic Data Mining and Knowledge Discovery*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2009.
40. Cheng, T.; Haworth, J.; Anbaroglu, B.; Tanaksaranond, G.; Wang, J. Spatiotemporal data mining. In *Handbook of Regional Science*; Fischer, M.M., Nijkamp, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1173–1193.
41. Andrienko, N.; Andrienko, G. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*; Springer: Berlin/Heidelberg, Germany, 2005.
42. Andrienko, G.; Andrienko, N.; Bak, P.; Keim, D.; Wrobel, S. *Visual Analytics*; Springer: Heidelberg/Berlin, Germany, 2013.
43. Maciejewski, R. Geovisualization. In *Handbook of Regional Science*; Fischer, M.M., Nijkamp, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1137–1155.

44. Illian, J.; Penttinen, A.; Stoyan, H.; Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Patterns*; Senn, S., Scott, M., Barnett, V., Eds.; John Wiley & Sons: Chichester, UK, 2008.
45. Selvin, S. *Statistical Analysis of Epidemiologic Data*, 2nd ed.; Oxford University Press: New York, NY, USA; Oxford, UK, 1996.
46. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
47. Getis, A. Spatial weights matrices. *Geogr. Anal.* **2009**, *41*, 404–410. [CrossRef]
48. Greene, W.H. *Econometric Analysis*, 7th ed.; Pearson: Harlow, UK, 2014.
49. Coxe, S.; West, S.G.; Aiken, L.S. The analysis of count data: A gentle introduction to poisson regression and its alternatives. *J. Pers. Assess.* **2009**, *91*, 121–136. [CrossRef] [PubMed]
50. Lambert, D.M.; McNamara, K.T.; Garrett, M.I. An application of spatial poisson models to manufacturing investment location analysis. *J. Agric. Appl. Econ.* **2006**, *38*, 105–121. [CrossRef]
51. Liviano, D.; Arauzo-Carod, J.M. Industrial location and interpretation of zero counts. *Ann. Reg. Sci.* **2013**, *50*, 515–534. [CrossRef]
52. Gehrke, B.; Frietsch, R.; Neuhäusler, P.; Rammer, C. *Neuabgrenzung Forschungsintensiver Industrien und Güter*; EFI: Berlin, Germany, 2013.
53. Florida, R.; King, K. *Rise of the Urban Startup Neighborhood*; Martin Prosperity Institute Working Paper; Martin Prosperity Institute: Toronto, ON, Canada, 2016.
54. Florida, R.; Adler, P.; Mellander, C. The city as innovation machine. *Reg. Stud.* **2017**, *51*, 86–96. [CrossRef]
55. Projekt Adlershof Adlershof Science City. Available online: https://www.adlershof.de/en/sectors-of-technology/it-media/info/ (accessed on 1 October 2017).
56. Weber, A. *Über den Standort der Theorien: Reine Theorie des Standortes*, 2nd ed.; J.C.B. Mohr: Tübingen, Germany, 1922.
57. Marshall, A. *Principles of Economics*, 8th ed.; Macmillan Co.: London, UK, 1890.
58. Hoover, E.M. *Location Theory and the Shoe Leather Industries*; Harvard University Press: Cambridge, MA, USA, 1937.
59. Carlino, G.A.; Chatterjee, S.; Hunt, R.M. Urban density and the rate of invention. *J. Urban Econ.* **2007**, *61*, 389–419. [CrossRef]
60. Hansen, E.R. Industrial location choice in São Paulo, Brazil: A nested logit model. *Reg. Sci. Urban Econ.* **1987**, *17*, 89–108. [CrossRef]
61. Friedman, J.; Gerlowski, D.A.; Silberman, J. What attracts foreign multinational coproations? Evidence from branch plant location in the United States. *J. Reg. Sci.* **1992**, *32*, 403–418. [CrossRef]
62. Smith, D.F.J.; Florida, R. Agglomeration and industrial location: An econometric analysis of Japanese-Affiliated manufacturing establishments in automotive-related industries. *J. Urban Econ.* **1994**, *36*, 23–41. [CrossRef]
63. Ahlfeldt, G.; Pietrostefani, E. *The Economic Effects of Density: A Synthesis*; SERC Discussion Paper, 210; London School of Economics and Political Science: London, UK, 2017.
64. Rosenthal, S.S.; Strange, W.C. Evidence on the nature and sources of agglomeration economies. In *Handbook of Regional and Urban Economics*; Henderson, J.V., Thisse, J.-F., Eds.; Elsevier B.V.: Amsterdam, The Netherlands, 2004; Volume 4, pp. 2120–2167.
65. Eicher, T.S.; Strobel, T. *Information Technology and Productivity Growth*; Edward Elgar Publishing Ltd.: Cheltenham/Northampton, UK, 2009.
66. Jang, S.; Kim, J.; von Zedtwitz, M. The importance of spatial agglomeration in product innovation: A microgeography perspective. *J. Bus. Res.* **2017**, *78*, 143–154. [CrossRef]
67. List, J.A. US county-level determinants of inbound FDI: Evidence from a two-step modified count data model. *Int. J. Ind. Organ.* **2001**, *19*, 953–973. [CrossRef]
68. Coughlin, C.C.; Segev, E. Location determinants of new foreign-owned manufacturing plants. *J. Reg. Sci.* **2000**, *40*, 323–351. [CrossRef]
69. Arauzo-Carod, J.-M. Determinants of industrial location: An application for Catalan municipalitie. *Pap. Reg. Sci.* **2005**, *84*, 105–120. [CrossRef]
70. Peter, R. *Kapazitäten und Flächenbedarf Öffentlicher Verkehrssysteme in Schweizerischen Agglomerationen*; Term Paper; ETH Zürich: Zürich, Switzerland, 2005.
71. Coughlin, C.C.; Terza, J.V.; Arromdee, V. State characteristics and the location of foreign direct investment within the United States. *Rev. Econ. Stat.* **1991**, *73*, 675–683. [CrossRef]

72. Audretsch, D.B.; Lehmann, E.E. Does the knowledge spillover theory of entrepreneurship hold for regions? *Res. Policy* **2005**, *34*, 1191–1202. [CrossRef]
73. Rammer, C.; Kinne, J.; Blind, K. *Microgeography of Innovation in the City: Location Patterns of Innovative Firms in Berlin*; ZEW Discussion Paper; ZEW: Mannheim, Germany, 2016.
74. Basile, R. Acquisition versus greenfield investment: The location of foreign manufacturers in Italy. *Reg. Sci. Urban Econ.* **2004**, *34*, 3–25.
75. Barbosa, N.; Guimaraes, P.; Woodward, D. Foreign firm entry in an open economy: The case of Portugal. *Appl. Econ.* **2004**, *36*, 465–472. [CrossRef]
76. Goodchild, M.F. Scale in GIS: An overview. *Geomorphology* **2011**, *130*, 5–9. [CrossRef]
77. Cohendet, P.; Grandadam, D.; Simon, L. The anatomy of the creative city. *Ind. Innov.* **2010**, *17*, 91–111. [CrossRef]
78. Gottlieb, P.D. Residential amenities, firm location and economic development. *Urban Stud.* **1995**, *32*, 1413–1436. [CrossRef]
79. Glaeser, E.L.; Kerr, W.R.; Ponzetto, G.A.M. *Clusters of Entrepreneurship*; NBER Working Paper; NBER: Cambridge, MA, USA, 2009.
80. Ahlfeldt, G.M. Blessing or curse? Appreciation, amenities and resistance to urban renewal. *Reg. Sci. Urban Econ.* **2011**, *41*, 32–45. [CrossRef]
81. Eurostat. *Quality of Life: Facts and Views*; Mercy, J.-L., Litwinska, A., Dupré, D., Clarke, S., Ivan, G., Stewart, C., Eds.; Eurostat: Luxembourg, Luxembourg, 2015.
82. Månssona, K.; Shukur, G. A poisson ridge regression estimator. *Econ. Model.* **2011**, *28*, 1475–1481. [CrossRef]
83. Westerholt, R.; Resch, B.; Zipf, A. A local scale-sensitive indicator of spatial autocorrelation for assessing high- and low-value clusters in multiscale datasets. *Int. J. Geogr. Inf. Sci.* **2015**, 1–20. [CrossRef]
84. LeSage, J.; Pace, R.K. *Introduction to Spatial Econometrics*; Balakrishnan, N., Schucany, W.R., Eds.; Chapmann & Hall/CRC: Boca Raton, FL, USA, 2009.
85. Anselin, L. *Spatial Econometrics: Methods and Models*; Springer: Heidelberg/Berlin, Germany, 1988.
86. Sagl, G.; Loidl, M.; Beinat, E. A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. *ISPRS Int. J. Geo-Inf.* **2012**, *1*, 256–271. [CrossRef]
87. Miller, H.J.; Goodchild, M.F. Data-driven geography. *GeoJournal* **2015**, *80*, 449–461. [CrossRef]
88. Berlin-Brandenburg Bureau of Statistics Statistik Berlin-Brandenburg. Available online: https://www.statistik-berlin-brandenburg.de/ (accessed on 1 October 2017).
89. Carlino, G.A.; Carr, J.; Hunt, R.M.; Smith, T.E. The agglomeration of R&D labs. *J. Urban Econ.* **2017**, *101*, 14–26.
90. Scholl, T.; Brenner, T. Detecting spatial clustering using a firm-level cluster index. *Reg. Stud.* **2014**, *3404*, 1–15. [CrossRef]
91. Kukuliač, P.; Hor, J.R.I. W Function: A new distance-based measure of spatial distribution of economic activities. *Geogr. Anal.* **2016**, *49*, 1–16. [CrossRef]

Appendix B

Paper 2: Knowledge proximity and firm innovation: A microgeographic analysis for Berlin

Article

Urban Studies

Urban Studies
1–19

DOI: 10.1177/0042098018820241
journals.sagepub.com/home/usj
SAGE

Knowledge proximity and firm innovation: A microgeographic analysis for Berlin

Christian Rammer
Centre for European Economic Research (ZEW), Germany

Jan Kinne
Centre for European Economic Research (ZEW), Germany

Knut Blind
Technische Universitat Berlin, Germany

Abstract
We analyse the geographic proximity of innovative firms to different types of knowledge sources in an urban environment on a microgeographic scale. Based on a comprehensive panel data set of manufacturing and service firms in the German capital city Berlin, we investigate the characteristics of firms' knowledge environment while differentiating by the type of innovation. Geocoded firm locations at the level of individual addresses allows us to describe the knowledge environment of firms on a very fine microgeographic scale. We find that innovative firms are located in places with higher numbers of same-sector firms, more start-ups and a higher inflow of other firms. They also locate in closer proximity to universities and research institutes. These differences decay rapidly within a few metres (50–250 m), indicating a truly microgeographic scope of knowledge sources in urban environments.

Keywords
agglomeration/urbanisation, economic processes, innovation, technology/smart cities

摘要
我们在微观地理尺度上分析创新企业与城市环境中不同类型知识来源的地理接近程度。基于德国首都柏林的制造和服务公司的综合面板数据集，我们在区分创新类型的基础上研究了企业知识环境的特征。单个地址层面的地理编码公司位置使我们能够以非常精细的微观地理尺度描述公司的知识环境。我们发现，创新型企业位于同行业公司数量较多、初创企业较多、其他企业流入较多的地方。它们也位于离大学和研究机构更近的地方。这些差异在数米（50-250米）内迅速衰减，表明了城市环境中知识来源地理辐射范围真实的微观特性。

关键词
集聚/城市化、经济过程、创新、技术/智慧城市

Received January 2018; accepted November 2018

Introduction

Knowledge spillovers are an important driver of innovation in firms. Using knowledge of others reduces the cost of innovation through input sharing and learning (Jaffe, 1986) and provides firms with ideas and technology they could not have developed based on their own capabilities. While knowledge spillovers may occur at any geographic scale, close proximity to external knowledge sources is likely to facilitate such spillovers (Audretsch and Feldman, 1996; Jaffe et al., 1993). Proximity eases exchange of workers among firms and hence the diffusion of knowledge embedded in people (Glaeser, 1999; Jovanovic and Nyarko, 1995; Jovanovic and Rob, 1989). Proximity also facilitates learning from observing, informal knowledge exchange through personal contacts of workers and managers, and knowledge exchange from joint business activities along the supply chain.

An urban environment provides an excellent ground for such knowledge flows (Audretsch, 1999; Feldman, 1999; Glaeser, 1999). Knowledge spillovers have hence been identified as one of the major sources for the emergence and growth of cities (Duranton and Puga, 2004; Marshall, 1890). There is strong empirical evidence that knowledge spillovers actually increase innovation performance in cities (Audretsch and Feldman, 2004; Feldman, 1999; Feldman and Audretsch, 1999; Henderson, 2007; Simmie, 2002). Less is known about the exact geographic scope of these spillovers, however. Many studies consider the entire stock of knowledge within a city as a source for spillovers (Audretsch and Feldman, 2004). When looking at proximity within a city, usually a rather large geographic scale is applied (Duranton and Overmans, 2005; Jang et al., 2017; Larsson, 2017; Murata et al., 2014; Rosenthal and Strange, 2008).

Recent research has paid more attention to the very local environment of a firm and how the configuration of a firm's direct neighbourhood and the local situation in which a firm operates affect knowledge exchange and innovation results (see Catalini, 2016; Kabo et al., 2014). These microgeographic studies, applying a very detailed scale of a few metres only, show that local environments provide special opportunities for meeting and contacting other innovative actors and can have a significant impact on innovation performance. Another recent strand of literature analysis the role of microgeographic configuration in a city (see Andersson et al., 2016, 2017). The concept of innovation districts within cities (Katz and Bradley, 2013; Katz and Wagner, 2014) is another recent approach that emphasises the role of microgeography for innovation in urban areas.

The present article links this microgeographic view with the literature on knowledge spillovers and innovation in an urban environment. By employing highly disaggregated geographic data on the location of knowledge sources and innovative firms in the German capital city Berlin, the article aims to contribute to the literature in two ways. First, we provide evidence on the role of microgeographic proximity within a city instead of treating a city as a single location. Second, we investigate how different knowledge sources (universities, research institutes, start-ups, other innovative firms) are geographically related to different types of innovation (product and process innovation, degree of novelty) which offers new insight into the role of knowledge diversity in a city for

Corresponding author:
Christian Rammer, Centre for European Economic Research (ZEW), Economics of Innovation and Industrial Dynamics, L7, 1, Mannheim 68161, Germany.
Email: rammer@zew.de

innovation. While this paper builds on the theoretical framework of knowledge spillovers, we do not empirically analyse knowledge spillovers as such.

Knowledge proximity and innovation in urban areas: A microgeographic perspective

A city is an ideal spot for accessing and exchanging knowledge that is critical to innovation. Innovative ideas of users, advanced technology from suppliers, new knowledge generated in science and research, innovations of other firms, or support from consultants and other service providers are often crucial inputs to innovation (Cassiman and Veugelers, 2006; Leiponen and Helfat, 2011; West and Bogers, 2013). In order to access this knowledge, firms do not only need absorptive capacities and adequate search strategies, they also need to interact with these external sources. Geographic proximity can certainly facilitate the exchange of knowledge (Figueiredo et al., 2015; Jaffe et al., 1993; Singh and Marx, 2013; Thompson, 2006). The more tacit knowledge is, the more important is face-to-face communication, mutual understanding, a common background and trust in order to learn from others (Audretsch and Feldman, 1996; Gertler, 2003; Howells, 2002). Combining specialised knowledge sources and diverse actors in a city helps to avoid lock-in that may emerge if close interaction among local actors restrict openness and searching beyond the local boundaries (Boschma, 2005).

Knowledge exchange can take place in different ways. One is when workers move between firms as much of the critical knowledge needed for innovation is embedded in workers (Combes and Duranton, 2006; Fosfuri and Rønde, 2004). Cities provide a particularly favourable environment for this type of knowledge exchange as workers can choose from many potential employers within a short commuting area from their home. Another way is to learn from observing and communicating with others (Duranton and Puga, 2004). Urban density clearly eases interaction, making urban areas a kind of school where entrepreneurs, managers and workers can continually add to their skills (Rosenthal and Strange, 2003). Urban density was also found to have a positive effect on wages, especially for university-educated workers, potentially capturing localised non-market interaction effects (Andersson et al., 2016). The concept of innovation districts introduced by Katz and Wagner (2014) also emphasises the role of local interaction between urban actors and infrastructures. They are becoming more important owing to the value of density and proximity in the evolution of a knowledge and technology driven economy and the emergence of open innovation approaches.

The link between local knowledge environments and innovation in firms is not a unidirectional one. Firms may not only seek for knowledge proximity, their innovation activities may also shape their local knowledge environment and affect innovation in other firms, owing to the public good property of innovation (Duranton and Puga, 2004). Close proximity between knowledge sources and innovative firms can hence form a local innovative milieu where actors mutually provide inputs for innovation (Gertler, 2003). The literature on innovative clusters has demonstrated how this process can form dynamic regional concentrations of innovative activities (see Audretsch, 2003; Feldman and Audretsch, 1999; Forman et al., 2016; Glaeser, 2000; Klepper, 2010; Porter, 1996; Saxenian, 1994).

A key but yet little-explored issue is the exact geographic scale at which the proximity to knowledge sources can stimulate innovation and knowledge spillovers. Personal interaction certainly facilitates this process

(Glaeser, 1999, 2000). The potential for personal interaction is assumed to decay rapidly with distance (Larsson, 2014). Close spatial proximity can also drive job-switching and associated knowledge flows (Larsson, 2017). Recent research has stressed the specific role of such microgeographic[1] configurations. Kabo et al. (2014) use path overlap within an academic research building as a measure of proximity and examine how physical space is shaping the formation and success of scientific collaborations. They find that when two researchers traverse paths with greater overlap, both their propensity to form new collaborations and to win grant funding for their joint work increase. Catalini (2016) shows that researchers' co-locations matter for the rate, quality and direction of scientific collaboration. Using data on research labs that were forced to move within a Paris university campus without being able to choose their new location, he found that collaboration between two labs increases significantly if the labs have moved to the same place, as long as the type of research done in both labs is sufficiently similar. Jang et al. (2017) demonstrate for the mobile gaming industry in Seoul that firms specialising in similar aspects of product innovation tend to locate in a single cluster within the city. Andersson et al. (2017) find that firms may benefit from both specialisation and diversification economies by locating in neighbourhood-level industry clusters within diversified cities.

While these microgeographic studies focus on collaboration and the performance of researchers and other individuals within the same organisation, there is little research on the role of the microgeographic configuration of the knowledge environment at the firm level. In this article, we analyse whether spatial proximity can play a decisive role for firms' innovation in urban areas. The direct neighbourhood to other firms and other knowledge providers may increase the chance of getting in contact with them and exchanging information, including coincidental contacts. Personal contacts can also facilitate the move of workers between firms. Direct neighbourhood may also increase the opportunity to observe activities of neighbours and stimulate learning.

We investigate the role of the local knowledge environment of innovative and non-innovative firms in urban areas, using Berlin as the place for our empirical analysis. Detailed firm address data allow for a microgeographic analysis at a scale of 50-m distances and below. At the same time, our data include information on innovation activities of a very large sample of firms in manufacturing and services in Berlin for a five-year period. This provides the opportunity to examine how innovation activities relate to changes in the innovation activities of surrounding firms, including relocations, start-ups and closures.

Our research is explorative in nature. The main aim is to describe the local knowledge environment of innovative and non-innovative firms and their spatial proximity to different knowledge sources on a microgeographic scale. We focus on three types of knowledge sources:

(1) The location of universities and research institutes as a major source of knowledge and talented people for innovative firms (Agrawal et al., 2014; Anselin et al., 1997; Feldman and Florida, 1994; Roper et al., 2017).
(2) Entrepreneurship clusters which provide both a source for innovation, and may challenge existing firms to respond to new market entries by increasing their own innovative efforts (Chatterji et al., 2014; Duvivier and Polèse, 2018; Glaeser et al., 2010).

(3) Micro-clusters of innovative firms that facilitate learning and specialisation in innovation (Boix et al., 2015; Jang et al., 2017).

Duranton and Puga (2001) stressed the importance to distinguish product and process innovation when analysing innovation in an urban environment. They showed that product innovation tends to be linked to diversified knowledge sources while cost-reducing process innovation are linked to more specialised places. We follow them and separate product from process innovation. In addition, we distinguish novel product innovations (new-to-the-market) from the imitation of innovative ideas and consider whether a firm conducts in-house R&D since different knowledge sources may be required for different degrees of novelty and types of innovative knowledge (Leiponen and Helfat, 2011; Rammer et al., 2009).

Methodology

When investigating the relation between a firm's local knowledge environment and its innovation activities, endogeneity problems emerge immediately. Since geography matters for innovation, firms will try to choose locations that fit best to their innovation activities and may locate in close proximity to other innovative firms and important knowledge sources (Leiponen and Helfat, 2011). This self-selection of innovative actors into certain urban neighbourhoods can be a main driver for the emergence of innovative districts within a city (Katz and Bradley, 2013; Katz and Wagner, 2014). Our data show that innovative firms indeed cluster in certain locations (see Figure 2) which suggests that a selection process may be at work. At the same time, locations may change their characteristics if they host innovative firms, e.g. if research institutes relocate towards existing innovative clusters.

We deal with this selection issue in two ways. For investigating differences in the local knowledge environment of innovative and non-innovative firms, it is important to consider firm characteristics which may affect both innovation decisions and the choice of location. We apply a matching approach (Heckman et al., 1998) to ensure that we compare innovative firms with non-innovative ones that share the same basic characteristics so that differences in the local knowledge environment cannot be attributed to these characteristics.

While matching is usually employed to identify treatment effects of policy intervention, the method is also useful for our purpose. We match each innovative firm i in our sample with a non-innovative firm j which shows the same basic characteristics. For this purpose, we estimate the propensity score for each innovative firm $P(x_i, \beta)$ and consider only innovative firms with common support, i.e. for which the probabilities do not exceed the maximum and do not fall below the minimum of the probabilities of non-innovative firms. x_i represents firm characteristics (size and age) and β is a parameter. In order to ensure that each pair of innovative and non-innovative firm comes from the same sector, we require exact matching for a firm's sector affiliation ($sec_i = sec_j$, sec being the 2-digit level of ISIC rev. 2). The difference δ between an innovative firm i and a non-innovative firm j (out of the entire group of non-innovative firms N^0) is given by:

$$\delta_{ij}^{k} = P^{k}\left(x_i', \hat{\beta}\right) - P^{k}\left(x_i', \hat{\beta}\right) \forall j = 1, \ldots, N^0 \tag{1}$$

Based on (1), we calculate the Malahanobis distance (MD)

$$MD_{ij} = \delta_{ij}{}^{k'} \Omega^{(-1)} \delta \forall j = 1, \ldots, N^0 for\, sec_i = sec_j \tag{2}$$

to find the nearest non-innovative firm j for each innovative firm i. Ω represents the covariance matrix based on non-innovative firms. Note that the matching is performed for each type k of innovation separately. Owing to the large data set we have at hand, the matching resulted in a high quality. The common support criteria was met for each innovative firm, and after matching size and age differences between innovative and non-innovative firms were completely insignificant. By ensuring that each innovative firm is matched with a non-innovative one from the same 2-digit sector, we take also into account the very different locational requirements of manufacturing and services firms.

The second approach to tackle endogeneity is to analyse whether changes in innovation activities of neighbouring firms in the past are correlated with a firm's current innovation activities in t. For this purpose, we consider five types of changes in innovation activities of firm j in period $t-1$ which is located in the neighbourhood n of firm i ($i \neq j$):

(1) transition of firm j from innovative to non-innovative between $t-2$ and $t-1$ (and vice versa) while firm j remains located in neighbourhood n in $t-2$, $t-1$ and t;
(2) moving in of an innovative or non-innovative firm j in $t-1$ into neighbourhood n (and staying in n in t);
(3) moving out of an innovative or non-innovative firm j in $t-1$ from neighbourhood n;
(4) foundation of a new firm j (innovative or non-innovative) in $t-1$ in neighbourhood n (which remains located in n in t);
(5) closure of a firm j (innovative or non-innovative) in $t-1$ in neighbourhood n.

We assume that past innovation dynamics in other firms are rather exogenous to a focus firm i's later innovation activities. At the same time, innovation dynamics in a firm i's local environment may alter the firm's opportunities for innovating. If neighbouring firms introduce innovations, innovative firms move into firm i's neighbourhood, or innovative start-ups open their business in firm i's neighbourhood, firm i may be stimulated by these activities and may learn for its own innovative efforts. Similarly, the loss of an innovative environment because of closures or moving out of innovative firms or stopping of innovative activities in neighbouring firms might discourage innovation in firm i. In order to test the relation between past innovation dynamics in the local environment on current innovation IN in firm i, we run the following regression model:

$$IN^k_{i,t} = \alpha + \sum_m \beta^k_m ID^k_{mj,t-1} + \sum_l \gamma_l CT_{li,t} + \varepsilon_{i,t} \; i \neq j; n_i = n_j \tag{3}$$

ID represents the different types m of innovation dynamics variables (transition of status, moving in and out, start-up and closure of firms) based on the number of firms reporting a respective dynamic in firm i's neighbourhood n. CT represents control variables l that may affect firm i's innovation decision in t, such as size, age and sector. α is a constant, β and γ are parameters to be estimated, and ε is the error term.

In both the matching and the regression model approach, we consider five types k of innovative firms:

(1) innovator (product and/or process)
(2) product innovator (goods and/or services)
(3) new-to-market innovator (i.e. novel product innovation)
(4) process innovator

(5) firms with continuous in-house R&D activity

Note that firms can have different types of innovations in the same period. All indicators refer to well-established definition and measures proposed in the Oslo Manual (OECD and Eurostat, 2005) and applied in the Community Innovation Surveys (CIS) of the European Commission and are measured in a binary (yes/no) way. Such indicators are well suited for measuring innovation in small firms since small firms usually only have one or few innovations within a certain period of time. The chosen indicators are also well suited for capturing innovation both in manufacturing and services. Using quantitative indicators such as R&D expenditure or sales with new products is often less useful as they can be subject to extreme values in small firms and may overrate differences in innovation performance (see Rammer et al., 2009). As about 80% of the firms in our empirical study are small firms with fewer than 50 employees, and about 70% are from service sectors, we believe that the choice of binary indicators is adequate for our study.

Data

The empirical analysis is based on a unique panel data set on innovation activities of Berlin-based firms, the 'Berlin Innovation Panel'. This panel survey has been initiated in 2012 by the Technical University of Berlin and has received funding since 2013 from the *Technologiestiftung* Berlin. The survey covers all legally independent enterprises with five or more employees in manufacturing and knowledge-intensive services[2] that are headquartered in Berlin. The survey is conducted as part of the German Innovation Survey,[3] which is the German contribution to the CIS. It shares all methodological features with the CIS, including questionnaire design, quality control and data processing routines (see Peters and Rammer, 2013, for more details). The Berlin Innovation Panel as well as the German Innovation Survey are conducted by the Centre for European Economic Research (ZEW) as a voluntary mail survey (including an online response option) on an annual base. This paper uses the first five waves of the Berlin Innovation Panel conducted in the years 2012 to 2016 which cover the reference years 2011 to 2015.

The gross sample of the survey includes basically all firms of the target population of the survey and has been refreshed in 2013 and 2015 to compensate for firm closures and firms moving out of Berlin. Panel mortality is substantially high in the Berlin panel, reflecting high dynamics in the urban firm sector. The response rate of the survey is around 20%, which is somewhat lower than the response rate in the German Innovation Survey (25–30%), reflecting a lower propensity of survey participation among smaller firms. Following the German Innovation Survey, the Berlin Innovation Panel includes a comprehensive non-response survey which collects information on the presence of product and process innovation as well as in-house R&D activities, applying the same definitions as the paper/online questionnaire. The non-response survey is based on a stratified sample of non-responding firms and is conducted by telephone. The number of firms covered by the non-response survey exceeds the number of responding firms, leading to a high share of firms from the gross sample for which innovation-related information has been collected (between 42% and 47%). Table 1 provides details on the sample size of the Berlin Innovation Panel.

The high dynamics in the Berlin firm sector reduces the panel nature of the data. In the first five survey years, a total of 7936

Table 1. Sample size of the Berlin Innovation Panel.

Survey year	Gross sample (#)	Not utili-sable[a] (#)	Responses (#)	Non-response (NR) survey (#)	Response rate (%)[b]	Response rate incl. NR (%)[b]
2012	4927	908	770	909	19.2	41.8
2013	5275	914	806	1101	18.5	43.7
2014	4886	782	752	997	18.3	42.6
2015	4810	918	791	1048	20.3	47.3
2016	4002	554	707	901	20.5	46.6

Notes: [a]Firm closure, moving out of Berlin, wrong address because of relocation, etc.
[b]As a percentage of gross sample net of not utilisable addresses.

different firms have been surveyed, but only 2092 firms were surveyed in all five years. As the unbalanced nature of the survey limits our analysis with respect to measuring innovation activities that take place in a firm's neighbourhood, we extended the data set towards a more balanced panel by interpolating and extrapolating observations. For this purpose, we exploit additional data from the Mannheim Enterprise Panel (MEP), which is the sampling pool for the survey,[4] on firm foundation and closure. Annual address, employment and sector data from the MEP are used to fill in geographic, employment and sector information. For innovation indicators, inter- and extrapolation is facilitated by the fact that each indicator refers to a three-year reference period, following the common practice of the CIS and the recommendations of the Oslo Manual (OECD and Eurostat, 2018). Details on the procedure are provided in the Supplementary Material (available online). In that way, we extended the total number of year-firm observations to 13,405.[5] The more balanced panel contains 3723 different firms. The average number of observations per firm is 3.6.

Address information for each firm in each year allows us to exactly geolocate firms and to calculate distances to other knowledge sources at a scale of 50 m and below. Details about geocoding of the address data are provided in the Supplementary Material (available online). Figure 1 illustrates the geographic detail of our firm-level innovation status data for the central area of Berlin.

Data on the five innovation indicators and on firm-level control variables are taken directly from the survey. While innovation is self-reported by firms, there are no incentives for firms to over- or under-report innovation. To clarify the concept of innovation, the survey includes examples for different types of innovations for the sector of the firm. 39.0% of the firms in our sample have introduced an innovation. 31.2% are classified as product innovators, 5.4% have introduced a market novelty, 24.4% are process innovators and 21.2% conduct in-house R&D continuously. The average size of firms is 86 employees (at full-time equivalents) and the average age of the firms is 21 years. 34.4% of the firms are from manufacturing sectors (including construction, energy and water supply, and waste treatment) and 65.6% from service sectors.

The geographic distribution of innovative and non-innovative firms is highly uneven. Figure 2 shows for product/process innovators a number of clusters with a high share of innovative firms as well as areas with predominantly non-innovative firms. Some

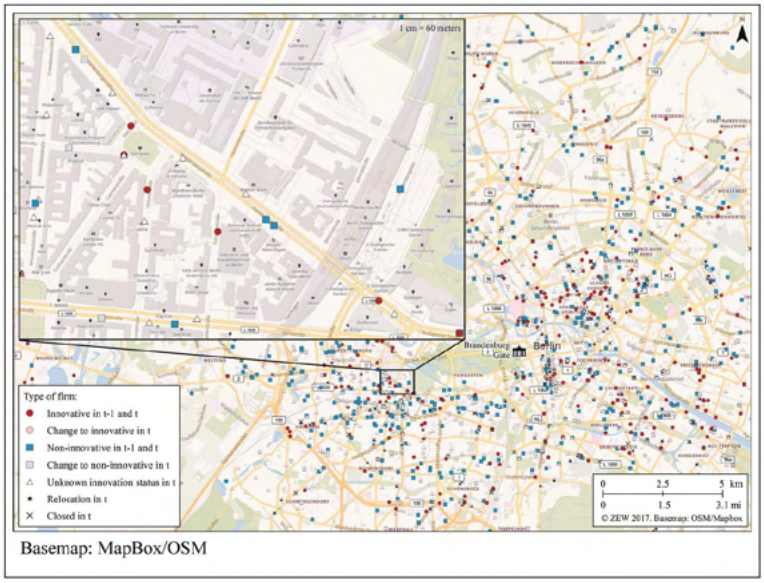


Figure 1. Example for the geographic distribution of firms in Berlin by innovation status.

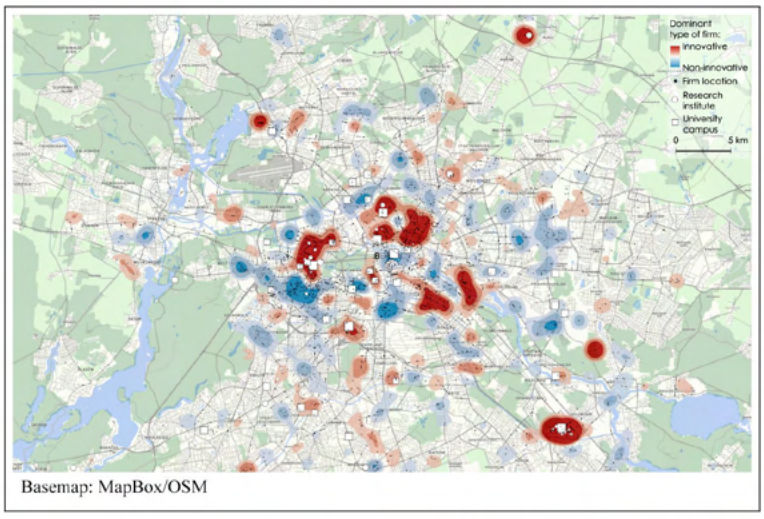


Figure 2. Clusters of innovative and non-innovative firms (product or process innovators) in Berlin.

Table 2. Location indicators.

Group	Indicator	Short name	Unit
Universities, research institutes	Universities	uni_[d]	No. of students
	Research institutes	ins_[d]	No. of researchers
Entrepreneurship	Firms moving in	fin_[d]	No. of firms
	Start-up activity	st_[d]	No. of firms
Micro-clusters	Stock of firms in same sector	fst_[d]	No. of firms
Innovation dynamics	Incoming innovators/non-innovators	in[k]1_[d]/in[k]0_[d]	No. of firms
	Outgoing innovators/non-innovators	ot[k]1_[d]/ot[k]0_[d]	No. of firms
	Transition into/out of innovator	ch[k]1_[d]/ch[k]0_[d]	No. of firms
	Innovative/non-innovative start-ups	nf[k]1_[d]/nf[k]0_[d]	No. of firms
	Closing innovators/non-innovators	cl[k]1_[d]/cl[k]0_[d]	No. of firms

Notes: [d]: alternative distance thresholds: 50 m, 100 m, 250 m 500 m, 1000 m, 2500 m; [k]: type *k* of innovation activity.

clusters relate to Katz and Wagner's (2014) 'anchor plus' model which describes innovation districts with a rich base of related firms, entrepreneurs and spin-off companies in downtown and mid-town areas of central cities centred around a major anchor institution (e.g. the Charlottenburg district around the Technical University) while others represent 're-imagined urban areas' (industrial or warehouse districts that underwent transformation, e.g. Friedrichshain). The Adlershof and Buch clusters in the southeast and the northeast correspond to Katz and Wagner's 'urbanised science park' type. More details on these clusters and their sector composition is provided in the Supplementary Material (available online).

The maps shows the difference in local firm densities between innovative and non-innovative firms The densities are based on kernel density estimation using triweight kernels with radius 1.5 km and identical weights (1.0) for all firms. In order to facilitate the visual presentation, the kernel density raster of innovative firms was multiplied by the relation of non-innovative to innovative firms.

Four groups of indicators describe the local knowledge environment of a firm (see Table 2).

(1) Universities and research institutes: number of students in universities and number of researchers in research institutes in firm *i*'s neighbourhood *n* in year *t*. The data have been collected based on our own inquiry and include 63 locations of universities and 83 locations of research institutes (see Figure 2).

(2) Entrepreneurship: number of firms newly started in firm *i*'s neighbourhood *n* in $t-1$, number of firms that moved into firm *i*'s neighbourhood in $t-1$ (either from Berlin or from elsewhere), considering only the sectors that are covered by the Berlin Innovation Panel. Data are taken from the MEP.

(3) Micro-clusters: number of firms active in the same 2-digit sector as firm *i* in firm *i*'s neighbourhood *n* in year *t*. Data are taken from the MEP.

(4) Innovation dynamics: number of firms in firm *i*'s neighbourhood *n* that changed their innovation status between years $t-1$ and *t*, including firms that have changed their innovation status, firms that moved in or out of the neighbourhood, and firms newly founded or closed. Data are taken from the Berlin Innovation Panel.

For defining a firm *i*'s neighbourhood *n*, we use six different distance thresholds: 50, 100, 250, 500, 1000 and 2500 m, measured as direct distance from the building in which a firm is located to the location of other firms, universities and research institutes. Descriptive statistics for all model variables, including data on innovation dynamics of firms by type of innovation, can be found in the Supplementary Material (available online).

Empirical results

Local knowledge environment of innovative firms

The result of the matching analysis reveals significant differences in the local knowledge environment of innovative and non-innovative firms in Berlin (see Table 3). For both product and process innovators we find statistically significant positive differences to non-innovators for the proximity to research institutes (*in_[d]*). An innovator on average has three researchers from public research institutes located within a 50 m radius, which is 79% more than for non-innovators. The difference is monotonously declining with increasing distance. The proximity to research institutes is significant up to a distance of about 1 km (except process innovators: 0.5 km).

For universities (*uni_[d]*), we find that within a distance of 0.5 to 1 km, the number of university students in a firm's neighbourhood is significantly higher for all product innovators and market novelty innovators while we do not find significant differences for smaller distances. This result may reflect the fact that most university buildings are rather huge and often located next to each-other on a campus, leaving little space for other firms to locate nearby, except for firms that provide direct services to the university or to students (e.g. printing shops, facility management). Consequently, other firms will have to choose locations that are rather distant from the university buildings.

Another significant difference relates to the proximity to other firms that recently have moved into a firm's neighbourhood (*fin_[d]*). A product innovator has on average 2.8 new neighbours in a 50 m radius, and 4.8 in a 100 m radius. This is 19% and 15% more than for firms without product innovation. Product innovators also show a larger number of recent start-ups (*st_[d]*) in their neighbourhood, exceeding the number of start-ups of firms without product innovation by 14% in a 100 m radius. We do not find higher local firm dynamics for process innovators. In addition, we find some indication of localisation economies for product innovators as the number of firms from the innovating firm's sector (*fst_[d]*) within a 250 m radius is significantly higher.

For market novelty innovators, the results largely correspond to those for product innovators. A main difference relates to localisation economies. Firms with market novelties are located in much closer proximity to other firms from their sector as compared with similar firms without market novelties. The number of same-sector firms (*fst_[d]*) located within a 50 m radius is 39% higher. The number of firms that have moved into the direct neighbourhood (*fin_[d]*) is 49% higher for a 50 m radius and 36% higher for a 100 m radius. The same holds for the number of start-ups (*st_[d]*). Significant positive differences can be observed up to 250 m for firms moving in (*fin_[d]*) and for start-ups (*st_[d]*), and up to 1 km for firms in the same sector (*fst_[d]*). For the proximity to research institutes and universities the results are similar to those for all product innovators.

For firms conducting in-house R&D on a continuous basis we find similar location patterns as for firms with market novelties. They show a higher number of firms in the same sector (*fst_[d]*) up to a 500 m radius. For the

Table 3. Differences in the proximity to knowledge sources by type of innovation: Results of matching analyses.

Variable	Innovator			Product innovator			Market novelty			Process innovator			Continuous R&D		
	mean	diff	*t* value	mean	diff	*t* value	mean	diff	*t* value	mean	diff	*t* value	mean	diff	*t* value
uni_50	11	−21	−0.28	11	22	0.38	6	0	0.00	17	52	1.09	11	41	0.61
uni_100	21	−1	−0.01	22	24	0.56	7	−9	−0.07	21	4	0.07	26	69	1.77
uni_250	147	16	0.75	164	19	0.93	157	32	0.69	159	16	0.70	188	56	2.73*
uni_500	982	25	2.99*	1068	27	3.22*	1517	48	3.28*	897	8	0.74	1356	52	6.85*
uni_1000	2760	14	2.51*	2928	17	3.03*	3657	43	4.47*	2622	4	0.55	3249	35	6.03*
uni_2500	12188	−2	−0.53	12437	2	0.48	12875	11	1.54	12086	−3	−0.66	12949	11	2.70*
ins_50	3.3	79	4.86*	3.7	76	4.36*	6.9	66	2.24*	3.1	70	3.65*	5	69	3.82*
ins_100	8.8	55	4.49*	9.3	52	3.76*	13	51	2.15*	7.8	30	1.78	12	51	3.81*
ins_250	48	40	4.35*	53	46	5.00*	76	51	2.85*	47	28	2.68*	76	63	7.58*
ins_500	117	31	4.73*	128	38	5.76*	194	51	4.21*	114	18	2.35*	161	56	9.10*
ins_1000	286	17	3.06*	305	22	4.00*	398	36	3.52*	280	8	1.22	342	34	6.18*
ins_2500	1192	0	0.10	1219	5	1.45	1275	9	1.16	1158	−7	−1.72	1244	7	1.62
fst_50	0.5	−1	−0.16	0.5	11	1.39	0.8	39	2.92*	0.4	−17	−1.71	1	26	3.14*
fst_100	0.7	1	0.12	0.7	12	1.77	0.9	30	2.40*	0.6	−11	−1.30	1	26	3.68*
fst_250	1.7	7	1.51	1.8	14	2.94*	2.5	34	3.71*	1.7	4	0.67	2	23	4.34*
fst_500	4.0	3	0.63	4.2	6	1.42	5.2	26	3.07*	3.9	3	0.50	5	13	2.71*
fst_1000	11	0	−0.07	11	3	0.64	13	19	2.19*	11	−2	−0.44	12	7	1.44
fst_2500	43	−8	−2.10*	44	−9	−2.08*	43	−2	−0.24	42	−8	−1.89	44	−7	−1.43
fin_50	2.6	17	1.81	2.8	19	2.00*	4	49	2.69*	2.5	−9	−0.77	3	25	2.75*
fin_100	4.6	13	1.96*	4.8	15	2.25*	6	36	2.51*	4.4	−5	−0.67	5	18	2.41*
fin_250	15	4	0.74	16	9	1.67	18	29	2.84*	15	0	0.02	16	6	1.01
fin_500	44	−7	−1.34	45	1	0.15	47	14	1.36	44	−6	−1.16	46	0	0.05
fin_1000	140	−8	−1.98*	144	0	−0.02	142	6	0.60	139	−9	−2.03*	146	1	0.16
fin_2500	646	−5	−1.56	657	−1	−0.19	650	5	0.71	644	−6	−1.89	674	2	0.58
st_50	0.2	12	1.28	0.2	13	1.35	0.3	51	3.14*	0.2	1	0.05	0	14	1.24
st_100	0.4	14	2.12*	0.5	14	1.98*	0.5	34	2.34*	0.4	4	0.53	0	11	1.26
st_250	1.7	5	0.89	1.7	9	1.62	1.9	24	2.02*	1.7	5	0.83	2	4	0.59
st_500	5.2	−1	−0.11	5.3	1	0.28	5.5	13	1.20	5.3	3	0.51	5	2	0.38

(continued)

Table 3. Continued

Variable	Innovator			Product innovator			Market novelty			Process innovator			Continuous R&D		
	mean	diff	*t* value	mean	diff	*t* value	mean	diff	*t* value	mean	diff	*t* value	mean	diff	*t* value
st_1000	17	0	−0.06	18	1	0.18	17	10	1.05	18	0	−0.09	18	4	0.72
st_2500	80	−2	−0.52	80	−1	−0.25	79	8	1.02	80	−1	−0.19	80	2	0.57
# obs.[a]	4119 / 5924			3305 / 6778			2558 / 7584			592 / 7158			318 / 7510		

Notes: *$p < 0.05$. diff. difference to control group in %. [a]Number of innovative/non-innovative firms.

number of ingoing firms (*fin_[d]*), significant effects are confined up to a 100 m radius. In addition, proximity to research institutes (*ins_[d]*) is much higher for continuously R&D performing firms. They also tend to be located closer to university campuses (*uni_[d]*) than any other type of innovation active firms. Proximity to start-ups (*st_[d]*) is not higher for firms with continuous R&D.

The results presented above refer to firms across all sectors. While the matching approach controls for firm sector affiliation, there might be a systematic difference between manufacturing and services owing to the different role of external knowledge sources in their innovation process. When matching separately for manufacturing and service firms (for results see the Supplementary Material, available online), we find that the results hold for manufacturing but not for services, except for continuous R&D where we also see that service firms with continuous R&D are located in closer proximity to universities and research institutes, and they have a larger number of other firms nearby.

Innovation dynamics in a firm's neighbourhood

The second part of our empirical analyses investigates the role of prior changes in other firms' innovation activity in the neighbourhood of a focal firm. These changes refer to firms already located in the area which start or stop to innovate (including closures) or move in or out (including start-ups). We assume that a firm's innovation of type k is particularly affected if changes in the same type of innovation occur nearby. In order to limit the variety of results, we define a firm's neighbourhood by a 250 m radius only since the analysis in the previous section has shown that the 250 m distance threshold is often a demarcation between significant and insignificant differences in

location patterns. We use regression models to analyse the link between a firm's innovation activities ('dependent variable') and innovation dynamics in its neighbourhood ('independent variable') while controlling for size, age and sector. Our results should be interpreted as correlations and not causal effects, as we cannot rule out that past innovation dynamics in a firm's local environment were influenced by the firm's current innovation activities, considering the steady-state nature of innovation.

We use three different dependent variables: (a) the probability of a firm i to report innovation of type k in year t (irrespective whether firm i has reported the same type of innovation in $t-1$), (b) the probability of a firm i to enter into innovation k in t (i.e. firm i hat no innovation of this type in $t-1$), and (c) the probability of a firm i to stop innovation k in t (while having reported this type in $t-1$). All independent variables are measured as change in the innovation status of neighbouring firms between $t-2$ and $t-1$. The estimation results are shown in Table 4.

We find that firms have a higher propensity to introduce a product innovation if other firms in their neighbourhood started or stopped product innovation activity in the year before. The estimated marginal effect is rather low: 1.2 percentage points for 'positive' innovation dynamics and 2.1 percentage points for 'negative' dynamics, compared to an average share of product innovators in the sample of 31.6%.

If other firms in the neighbourhood introduced a market novelty in the previous year, or if firms with market novelties moved away or ceased business, the probability to introduce a new-to-the-market innovation increases significantly. The estimated marginal effects are quite large: 3.8 percentage points if firms with a market novelty moved out and 6.6 percentage points if firms with market novelties closed. In case neighbouring firms have introduced such innovations in the previous period we find a much smaller value (0.9 percentage points, the average share of firms with market novelties is 5.9% in our sample).

The probability to introduce a process innovation increases if neighbouring firms have introduced process innovations in the previous year. The estimated marginal effect is 1.4 percentage points, while the average share of process innovators in the sample is 24.6%. This finding indicates a kind of local learning effect if firms can observe process innovation activities of their neighbours. For continuous R&D, we find similar results as for product innovation. If neighbouring firms started or stopped continuous R&D activities in the previous year, the probability to conduct continuous R&D increases. A high turbulence in local R&D activities seems to stimulate a firm's own R&D. In addition, the probability decreases if firms without continuous R&D moved into the neighbourhood.

We also run the regression analysis separately for manufacturing and service firms. Again, most of the above findings are confirmed for manufacturing firms but fewer for services, except for the findings on process innovation and continuous R&D which mainly apply to services. Owing to the lower number of observations in the split models, fewer statistically significant results are found.

Discussion and conclusion

This article made an attempt to explore the role of proximity to knowledge sources for innovation in firms in an urban environment. Using panel data on Berlin-based firms and exact address data, we investigated the firms' knowledge environments at a microgeographic scale, zooming into a firm's neighbourhood at the level of individual buildings.

When controlling for size, age and sector, we find that innovative firms, opposite to

Table 4. Estimation results of probit models on past innovation dynamics in a firm's neighbourhood and the firm's current innovation activities.

	Innovation		Product innovation		Market novelty		Process innovation		Continuous R&D	
	m.E.	std.err.	m.E.	std.err.	m.E.	std.err.	m.E.	std.err.	m.E.	std.err.
in1	−0.014	0.020	−0.021	0.022	0.013	0.026	0.008	0.022	−0.006	0.019
in0	0.025	0.020	0.008	0.017	−0.003	0.006	0.012	0.013	−0.027	0.013*
ot1	0.013	0.019	0.001	0.020	0.038	0.016*	−0.017	0.022	0.014	0.016
ot0	0.032	0.021	0.025	0.017	0.007	0.005	0.019	0.013	0.019	0.013
ch1	0.010	0.005	0.012	0.005*	0.009	0.005*	0.014	0.005*	0.023	0.011*
ch0	0.012	0.008	0.021	0.008*	−0.001	0.005	0.007	0.006	0.021	0.008*
cl1	−0.028	0.034	0.013	0.035	0.066	0.025*	−0.027	0.037	0.028	0.052
cl0	−0.024	0.035	−0.004	0.029	0.015	0.011	−0.019	0.025	0.000	0.022
nf1	0.133	0.143	−0.036	0.154	[a]		0.321	0.216	0.053	0.054
nf0	−0.032	0.040	0.014	0.034	0.004	0.013	−0.022	0.029	0.022	0.026
lna	−0.024	0.008*	−0.023	0.008*	0.009	0.002*	−0.016	0.007*	−0.024	0.007*
lnb	0.067	0.005*	0.053	0.004*	0.148	0.082*	0.059	0.004*	0.044	0.004*
LR Chi^2	1015.5		1103.5		372.9		467.5		1,209.8	
Pseudo R^2	0.112		0.131		0.168		0.062		0.179	
# observations	6736		6742		4915		6754		6311	

Notes: All models include sector dummies. [a]Variable omitted because of perfect correlation with dependent variable. *$p < 0.05$.

non-innovative firms, are located in places with a much higher number of other firms and start-ups in close vicinity (less than 250 m) and with a higher inflow of other firms. Close proximity to research institutes and universities is another distinctive feature of innovative firms. This finding is in line with a large number of studies that emphasise the role of local knowledge exchange between science and industry as a key factor of innovation (Anselin et al., 1997; Jaffe et al., 1993). What our study shows is that the geographic scope of this exchange seems to be very confined, at least in the context of urban spaces. Concerning the proximity to research institutes, the concentration of innovative firms already decreases beyond a 50 m radius. Beyond a distance of 1 km, we do not find a significant relation anymore. Product innovators and firms with continuous R&D in particular are very closely located to science and education institutions. These results point to the importance of opportunities for meeting each other in micro-spaces (Catalini, 2016; Kabo et al., 2014). The role of researcher mobility between universities, research institutes and firms in the innovation process may also enhance close proximity between them (Kaiser et al., 2015). In addition, some innovative firms may be spin-offs from research facilities or firms that choose a science park location in order to establish or ease interactions with science (see Löfsten and Lindelöf, 2002; Phan et al., 2005). However, the share of innovative firms located in science parks in all innovative firms in Berlin is at about 7%, implying that the overall results cannot be determined by this small group.

Another important finding of our analysis relates to the role of microgeographic industrial clusters. For firms with market novelties or continuous in-house R&D, close proximity (up to 250 m) to other firms from the same sector is a distinctive feature. In addition, a firm's probability to introduce a new-to-market innovation increases significantly if neighbouring firms have introduced such innovations in the previous period. These results hold across all industries, pointing to a general micro-scale localisation advantage in urban areas that can be related to learning from observing nearby competitors. Again, our contribution to the literature is that localisation economies in urban areas seem to operate on a very small geographic scale, which is in line with some sector-specific studies (see Arzaghi and Henderson, 2008; Jang et al., 2017).

The innovation dynamics in a firm's neighbourhood in the recent past (defined as changes in innovation activities in other firms located within a 250 m radius) do show some relation to current innovation in firms. As both 'positive' and 'negative' changes do play a role, it seems that local turbulence in innovation can motivate firms to consider innovation for themselves.

This research is a first attempt to zoom into the role of knowledge proximity for innovation in the city at a microgeographic level. While we believe that some of our results widen our understanding of the relation between localised knowledge and firm innovation, many important research questions remain open. First, we refrained from examining the exact urban environment in which a firm operates, e.g. urban infrastructure, density or amenities. In that way, our study is somewhat blind to the urban context that shapes innovation districts in cities (Katz and Bradley, 2013; Katz and Wagner, 2014). In future studies, one should compare the role of knowledge proximity for different innovation districts within a city.

Second, we can only observe correlations between innovation and local knowledge sources. We do not know to what extent firms actually interact with these knowledge sources and how knowledge is exchanged.

Third, though we do have panel data, the limited length of our time series (5 years) prevents a more in-depth analysis of how changes in the local knowledge environment are associated with changes in innovation. Finally, our results are based on an analysis across a large number of sectors of the urban economy. Analysis for a specific sector may add further insight as sectors may use different knowledge sources.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ORCID iD

Christian Rammer https://orcid.org/0000-0002-1173-9471

Notes

1. 'Microgeographic' refers to a micro-scale of spatial analysis that allows differentiation at a distance level of a few metres.
2. NACE rev. 2 divisions 10 to 39 (manufacturing) and 58 to 66, 70 to 74 (knowledge-intensive services).
3. Berlin-based firms surveyed in the German Innovation Survey are also part of the data set of the Berlin Innovation Panel.
4. The MEP (see Bersch et al., 2014, for details) is a kind of business register based on information collected by Germany's largest credit rating agency, *Creditreform*, and is maintained by ZEW. These data also serve as the base for the German firm data in the Bureau-van-Dijk company databases Amadeus and Orbis.
5. We also run the analyses based on the original data before inter- and extrapolation and found the same results, often at a higher level of statistical significance (see the Supplementary Material, available online).

References

Agrawal A, Cockburn I, Galasso A, et al. (2014) Why are some regions more innovative than others? The role of small firms in the presence of large labs. *Journal of Urban Economics* 81: 149–165.

Andersson M, Klaesson J and Larsson JP (2016) How local are spatial density externalities? Neighbourhood effects in agglomeration economies. *Regional Studies* 50(6): 1082–1095.

Andersson M, Larsson JP and Wernberg J (2017) *The economic microgeography of diversity and specialization*. IFN Working Paper No. 1167. Stockholm: Research Institute of Industrial Economics.

Anselin L, Varga A and Acs Z (1997) Local geographic spillovers between university research and high technology innovations. *Journal of Urban Economics* 42(3): 422–448.

Arzaghi M and Henderson JV (2008) Networking off Madison Avenue. *Review of Economic Studies* 75: 1011–1138.

Audretsch DB (1999) Agglomeration and the location of innovative activity. *Oxford Review of Economic Policy* 14(2): 18–29.

Audretsch DB (2003) Innovation and spatial externalities. *International Regional Science Review* 26(2): 167–174.

Audretsch DB and Feldman MP (1996) R&D spillovers and the geography of innovation and production. *American Economic Review* 86(3): 630–640.

Audretsch DB and Feldman MP (2004) R&D spillovers and the geography of innovation. In: Henderson JV and Thisse J-F (eds) *Handbook of Regional and Urban Economics, Volume 4, Cities and Geography*. Dordrecht: Elsevier, pp. 2713–2739.

Bersch J, Gottschalk S, Müller B, et al. (2014) *The Mannheim Enterprise Panel (MUP) and firm statistics for Germany*. ZEW Discussion Paper No. 14–104. Mannheim: Centre for European Economic Research (ZEW).

Boix R, Hervás-Oliver JL and Miguel-Molina BD (2015) Micro-geographies of creative industries clusters in Europe: From hot spots to assemblages. *Papers in Regional Science* 94: 753–772.

Boschma R (2005) Proximity and innovation. A critical assessment. *Regional Studies* 39: 61–74.

Cassiman B and Veugelers R (2006) In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition. *Management Science* 52(1): 68–82.

Catalini C (2016) *Microgeography and the Direction of Inventive Activity*. Rotman School of Management Working Paper No. 2126890, Cambridge, MA.

Chatterji A, Glaeser E and Kerr W (2014) Clusters of entrepreneurship and innovation. *Innovation Policy and the Economy* 14(1): 129–166.

Combes PP and Duranton G (2006) Labour pooling, labour poaching, and spatial clustering. *Regional Science and Urban Economics* 36(1): 1–28.

Duranton G and Overmans HG (2005) Testing for localisation using micro-geographic data. *Review of Economic Studies* 72(4): 1077–1106.

Duranton G and Puga D (2001) Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review* 91(5): 1454–1477.

Duranton G and Puga D (2004) Micro-foundations of urban agglomeration economies. In: Henderson JV and Thisse J-F (eds) *Handbook of Regional and Urban Economics, Volume 4: Cities and Geography*. Dordrecht: Elsevier, pp. 2063–2117.

Duvivier C and Polèse M (2018) The great urban techno shift: Are central neighbourhoods the next silicon valleys? Evidence from three Canadian metropolitan areas. *Papers in Regional Science* 97(4): 1083–1111.

Feldman MP (1999) The new economics of innovation, spillovers and agglomeration: A review of empirical studies. *Economics of Innovation and New Technology* 8: 5–25.

Feldman MP and Audretsch DB (1999) Innovation in cities: Science-based diversity, specialization and localized competition. *European Economic Review* 43: 409–429.

Feldman MP and Florida R (1994) The geographic sources of innovation: Technological infrastructure and product innovation in the United States. *Annals of the Association American Geographers* 84(2): 210–229.

Figueiredo O, Guimarães P and Woodward D (2015) Industry localization, distance decay, and knowledge spillovers: Following the patent paper trail. *Journal of Urban Economics* 89: 21–31.

Forman C, Goldfarb A and Greenstein S (2016) Agglomeration of invention in the Bay Area: Not just ICT. *American Economic Review* 106(5): 146–151.

Fosfuri A and Rønde T (2004) High-tech clusters, technology spillovers, and trade secret laws. *International Journal of Industrial Organization* 22(1): 45–65.

Gertler MS (2003) Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of Economic Geography* 3: 75–99.

Glaeser EL (1999) Learning in cities. *Journal of Urban Economics* 46(2): 254–277.

Glaeser EL (2000) The new economics of urban and regional growth. In: Clark G, Gertler M and Feldman M (eds) *The Oxford Handbook of Economic Geography*. Oxford: Oxford University Press, pp. 83–98.

Glaeser E, Kerr W and Ponzetto G (2010) Clusters of entrepreneurship. *Journal of Urban Economics* 67(1): 150–168.

Heckman J, Ichimura H and Todd P (1998) Matching as an econometric evaluation estimator. *Review of Economic Studies* 65(2): 261–294.

Henderson JV (2007) Understanding knowledge spillovers. *Regional Science and Urban Economics* 37(4): 497–508.

Howells JRL (2002) Tacit knowledge, innovation and economic geography. *Urban Studies* 39(5–6): 871–884.

Jaffe A (1986) Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value. *American Economic Review* 76: 984–1001.

Jaffe AB, Trajtenberg M and Henderson R (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 434: 578–598.

Jang S, Kim J and von Zedtwitz M (2017) The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research* 78: 143–154.

Jovanovic B and Nyarko Y (1995) The transfer of human capital. *Journal of Economic Dynamics and Control* 19(5–7): 1033–1064.

Jovanovic B and Rob R (1989) The growth and diffusion of knowledge. *Review of Economic Studies* 56(4): 569–582.

Kabo FW, Cotton-Nessler N, Hwang Y, et al. (2014) Proximity effects on the dynamics and outcomes of scientific collaborations. *Research Policy* 43: 1469–1485.

Kaiser U, Kongsted HC and Rønde T (2015) Does the mobility of R&D labor increase innovation? *Journal of Economic Behavior & Organization* 110: 91–105.

Katz B and Bradley J (2013) *The Metropolitan Revolution: How Cities and Metros are Fixing our Broken Politics and Fragile Economy*. Washington, DC: Brookings Institution Press.

Katz B and Wagner J (2014) *The rise of Innovation Districts: A New Geography of Innovation in America*. Washington, DC: Brookings Institution Press.

Klepper S (2010) The origin and growth of industry clusters: The making of Silicon Valley and Detroit. *Journal of Urban Economics* 67: 15–32.

Larsson JP (2014) The neighborhood or the region? Reassessing the density-wage relationship using geocoded data. *Annals of Regional Science* 52: 367–384.

Larsson JP (2017) Non-routine activities and the within-city geography of jobs. *Urban Studies* 54(8): 1808–1833.

Leiponen A and Helfat CE (2011) Location, decentralization, and knowledge sources for innovation. *Organization Science* 22(3): 641–658.

Löfsten H and Lindelöf P (2002) Science parks and the growth of new technology-based firms –Academic-industry links, innovation and markets. *Research Policy* 31(6): 859–876.

Marshall A (1890) *Principles of Economics*. London: Macmillan.

Murata Y, Nakajima R, Okamoto R, et al. (2014) Localized knowledge spillovers and patent citations: a distance-based approach. *Review of Economics and Statistics* 96(5): 967–985.

OECD and Eurostat (2005) *Oslo Manual. Guidelines for Collecting and Interpreting Innovation Data*. 3rd Edition. Paris: OECD Publishing.

Peters B and Rammer C (2013) Innovation panel surveys in Germany. In: Gault F (ed.) *Handbook of Innovation Indicators and Measurement*. Cheltenham: Edward Elgar, pp. 135–177.

Phan PH, Siegel DS and Wright M (2005) Science parks and incubators: Observations, synthesis and future research. *Journal of Business Venturing* 20(2): 165–182.

Porter M (1996) Competitive advantage, agglomeration economies, and regional policy. *International Regional Science Review* 19(1): 85–94.

Rammer C, Czarnitzki D and Spielkamp A (2009) Innovation success of non-R&D-performers: Substituting technology by management in SMEs. *Small Business Economics* 33: 35–58.

Roper S, Love JH and Bonner K (2017) Firms' knowledge search and local knowledge externalities in innovation performance. *Research Policy* 46: 43–56.

Rosenthal SS and Strange WC (2003) Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* 85(2): 377–393.

Rosenthal SS and Strange WC (2008) The attenuation of human capital spillovers. *Journal of Urban Economics* 64(2): 373–389.

Saxenian A (1994) *Regional Advantage: Culture and Competition in Silicon Valley and Route 12* 8. Cambridge, MA: Harvard University Press.

Simmie J (2002) Knowledge spillovers and reasons for the concentration of innovative SMEs. *Urban Studies* 39(5–6): 885–902.

Singh J and Marx M (2013) Geographic constraints on knowledge diffusion: Political borders vs. spatial proximity. *Management Science* 59: 2056–2078.

Thompson P (2006) Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citations. *Review of Economics and Statistics* 88: 383–389.

West J and Bogers M (2013) Leveraging external sources of innovation: A review of research on open innovation. *Journal of Product Innovation Management* 31(4): 814–831.

Appendix C

Paper 3: Web Mining for Innovation Ecosystem Mapping: A Framework and a Large-scale Pilot Study

Web Mining for Innovation Ecosystem Mapping: A Framework and a Large-scale Pilot Study

Jan Kinne[1,2,3]* and Janna Axenbeck[4 5]

[1] Department of Economics of Innovation and Industrial Dynamics, ZEW – Leibniz Centre for European Economic Research, Mannheim, Germany
[2] Z_GIS - Department of Geoinformatics, University of Salzburg, Austria
[3] Center for Geographic Analysis, Harvard University, Cambridge, MA, USA
[4] Department of Digital Economy, ZEW - Leibniz Centre for European Economic Research, Mannheim, Germany
[5] Justus-Liebig-University Giessen, Germany
* Corresponding author. Mail: jan.kinne@zew.de

Abstract: Existing approaches to model innovation ecosystems have been mostly restricted to qualitative and small-scale levels or, when relying on traditional innovation indicators such as patents and questionnaire-based survey, suffered from a lack of timeliness, granularity, and coverage. Websites of firms are a particularly interesting data source for innovation research, as they are used for publishing information about potentially innovative products, services, and cooperation with other firms. Analyzing the textual and relational content on these websites and extracting innovation-related information from them has the potential to provide researchers and policy-makers with a cost-effective way to survey millions of businesses and gain insights into their innovation activity, their cooperation, and applied technologies. For this purpose, we propose a web mining framework for consistent and reproducible mapping of innovation ecosystems. In a large-scale pilot study we use a database with 2.4 million German firms to test our framework and explore firm websites as a data source. Thereby we put particular emphasis on the investigation of a potential bias when surveying innovation systems through firm websites if only certain firm types can be surveyed using our proposed approach. We find that the availability of a websites and the characteristics of the website (number of subpages and hyperlinks, text volume, language used) differs according to firm size, age, location, and sector. We also find that patenting firms will be overrepresented in web mining studies. Web mining as a survey method also has to cope with extremely large and hyper-connected outlier websites and the fact that low broadband availability appears to prevent some firms from operating their own website and thus excludes them from web mining analysis. We then apply the proposed framework to map an exemplary innovation ecosystem of Berlin-based firms that are engaged in artificial intelligence. Finally,

we outline several approaches how to transfer firm website content into valuable innovation indicators.

Keywords: Web Mining; Web Scraping; Innovation;

JEL Classification: O30, C81, C88

Acknowledgments: The authors would like to thank the *German Federal Ministry of Education and Research* for providing funding for the research project (TOBI - Text Data Based Output Indicators as Base of a New Innovation Metric) of which this study is a part. Special thanks are due to Georg Licht who contributed valuable help and advice. We would also like to thank Sebastian Schmidt for his contribution to the development of ARGUS.

Author Contributions: Janna Axenbeck and Jan Kinne designed the study. Jan Kinne gathered, pre-processed, analyzed and visualized the data. Janna Axenbeck and Jan Kinne wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

1. Introduction

The disruptive force of radical innovation has the ability to reshape the economy and pave the way for new periods of long-term economic growth, while incremental innovation causes continuous change. It is therefore a matter of public interest to measure innovation activities within innovation ecosystems. Measuring these innovation activities to a sufficient degree of accuracy allows researchers to analyze a system's driving factors as well as the effectiveness of innovation policies. However, there is evidence that traditional indicators of innovation (e.g. questionnaire-based surveys and patent-based indicators) struggle to provide a timely and sufficiently granular picture of the current state of innovation ecosystems (Nagaoka, Motohashi, & Goto, 2010; OECD, 2009; Squicciarini & Criscuolo, 2013).

Firm-level innovation is often measured by means of indicators constructed using data from large-scale questionnaire-based surveys. Examples of such surveys include the Oslo Manual-based (OECD & Eurostat, 2018) biennial European Community Innovation Survey (CIS) and the annual Mannheim Innovation Panel (MIP), which also constitutes the German contribution to the CIS. Both surveys provide firm-level information about innovative and non-innovative enterprises as well as their R&D expenditures. Furthermore, they characterize an innovation by its degree of novelty (new to the firm, the market, the industry or the world) and the type of innovation (product, process, marketing, and organizational innovations). However, such indicators suffer from some major drawbacks. The German MIP, for example, covers 10,000 firms every year, which corresponds to only 0.3% of the total number of firms in Germany. Thus, the total number of innovative firms remains unknown and can merely be estimated through statistical extrapolation. Furthermore, rare but potentially important innovation activities happening in unobserved sectors or technological fields may not be covered in the data. This also affects the analysis of geospatial innovation processes, some of which happen to operate on a fine (micro-)geographical scale (Arzaghi & Henderson, 2008; Carlino & Kerr, 2015; Catalini, 2012; Jang, Kim, & von Zedtwitz, 2017; Kerr, Duranton, Glaeser, & Henderson, 2014). Consequently, established innovation indicators from questionnaire-based surveys lack sectoral, technological, and geographical granularity. Additionally, questionnaire-based surveys – especially on a large scale – are costly and time intensive. They also lack timeliness as it takes time to collect and process the data. Furthermore, surveys require firm participation as questionnaires have to be answered. As a result, voluntary surveys like the MIP suffer from uncompleted questionnaires and the desired information is not always accessible (Kleinknecht, Van Montfort, & Brouwer, 2002).

As an alternative to questionnaire-based surveys, innovation activity has been studied by analyzing patents (patent applications, citations, licensing). However, indicators constructed from patents cover only technological progress for which legal protection has been sought (Archibugi & Pianta, 1996). Moreover, most patents are never used (Shepherd & Shepherd, 2003); thus, they serve rather as indicators of inventions than of innovations. Another drawback of patent-based indictors, especially if they take a more selective approach, is that the dataset suffers from insufficient timeliness (Squicciarini & Criscuolo, 2013). The time lag between priority date and the information becoming available is usually more than a year (OECD, 2009).

Literature-based innovation output indicators (LBIO) are constructed by counting innovations in scientific, technical, or trade journals. This indicator type is usually used to measure the degree of radicalness of innovations. However, LBIOs do not capture in-house process innovations and the measure can be inflated for some technologies which might help firm profits to improve by signaling innovativeness (Coombs, 1996) or if other diverging incentives for firms to publish product innovations exist (Kleinknecht & Reijnen, 1993). In addition, Acs, Anselin, and Varga (2002) indicate that LBIOs under-represent innovations in smaller firms as their presence in the media is usually smaller.

We identified the following shortcomings which apply to a varying degree to the traditional innovation indicators described above:

- *Coverage*: They cover only a fraction of the overall firm population.
- *Granularity*: They suffer from insufficient sectoral, technological, and geographical granularity.
- *Timeliness*: They depict the state of the STI system as it was months or even years before.
- *Cost*: They involve high data collection costs, especially when conducted on a large scale.

The World Wide Web (Web) is a ubiquitous medium for communicating and disseminating information. Billions of private and commercial users worldwide (OECD, 2017) are producing increasing amounts of data. However, the sheer amount of data available, along with its mostly unstructured nature and its decentralized storage, imposes specific requirements on the collection, pre-processing, and analysis of the data. *Web mining*, the application of data mining techniques to uncover relevant data characteristics and relationships (e.g. data

patterns, trends, correlations) from unstructured web data, has been shown to be applicable in many fields of research (Askitas & Zimmermann, 2015; Raymond & Blockeel, 2000).

In economic research and ecosystem mapping, firm websites are a particularly interesting area of the Web. Firms use their websites to present themselves, as well as their products and services. The information found on these websites can be used to assess firms' products, services, credibility, achievements, key personnel decisions, strategies and relationships with other firms (Gök, Waterworth, & Shapira, 2015). Surveying firms using their websites instead of conducting interviews or questionnaires or using other traditional methods, offers some clear advantages (scale, cost, timeliness of the survey), but also comes with its own challenges (challenging data collection, data harmonization, and data analysis). However, no consistent approach for studying firm websites has been established yet. In addition, the data source itself (i.e. the population of firm websites) has not been studied rigorously in terms of its qualitative and quantitative properties. Basic yet important data characteristics such as the structural properties of firm websites and their coverage of the overall firm population are unknown.

In this paper, we develop and present a coherent web mining framework that is based on *ARGUS* (***A**utomated **R**obot for **G**eneric **U**niversal **S**craping*), an easy and free-to-use web scraping tool which allows for large-scale data retrieval from websites without requiring the user to have expert knowledge of web scraping technology. We then apply ARGUS in a pilot study using the entire firm population of Germany. The aim of this pilot study is to investigate and quantitatively assess firm websites as a data source for web-based innovation indicators and innovation ecosystem mapping, as well as to derive best practice guidelines for researchers who use ARGUS for large-scale web surveys. The following three research questions guideline our pilot study:

- **Research Question 1** *URL Coverage*: What subpopulation of firms can be surveyed using web mining of firm websites and is a systematic bias in terms of firm characteristics (age, size, sector, location etc.) to be expected?
- **Research Question 2** *Website Characteristics*: How do firm websites differ in terms of their size and content and how does that interfere with web mining studies?
- **Research Question 3** *Innovation Ecosystem Mapping*: How can our proposed framework be used to map an innovation ecosystem?

The remainder of this paper is organized as follows. First, we summarize the results of previous innovation research studies that used web mining. In the following Methods section,

we present our web mining framework and the ARGUS web scraping tool. In section 4, we present our data. The results of our pilot study are presented in section 5 and are discussed in section 6. Section 7 concludes and outlines future research.

2. Previous Research

There are only a few existing studies analyzing the usability of web-based innovation indicators and web mining for innovation ecosystem modelling. These studies either employ web content mining or web structure mining (Miner et al., 2012). The latter is the analysis of connections between entities (e.g. firms) via the hyperlink structure of websites. Katz and Cothey (2006) used this approach to develop a method that produces indicators for the web presence of innovation systems. In a case study on European and Canadian education institutions, they find that their method is suitable for measuring "the amount of recognition a nation or province's web presence receives from other nations and provinces in their innovation systems" (Katz & Cothey, 2006, p. 85). The authors emphasize the importance of reproducible and accurate indicators which are capable of dealing with the constantly changing properties of the Internet. Ackland et al. (2010) combine a web structure with a web content analysis. Other authors used such an approach in combination with visual network-based methods to identify business deals, funding relations, and alliances (Basole, Huhtamäki, Still, & Russell, 2016; Basole et al., 2015; Rubens, Still, Huhtamaki, & Russell, 2011).

In web content analysis, texts and other website content are analyzed. This approach is taken by the following studies: Youtie et al. (2012) use web scraping to explore the transitions from discovery to commercialization of 30 nanotechnology SMEs. Arora et al. (2013) use a similar approach to analyze entry strategies of SMEs commercializing emerging graphene technologies. Both study approaches are able to identify different innovation stages. Applying a keyword technique to explore the R&D activities of 296 UK-based enterprises, Gök, Waterworth, and Shapira (2015) find that web-based indicators offer additional insights when compared with patent and literature-based indicators. In addition, they emphasize that web mining as a research method has another advantage. The act of surveying a subject using web scraping does not cause certain problems such as altering the behavior of the study subject in response to being studied. The authors conclude "…that web mining is a significant and useful complement to current methods, as well as offering novel insights not easily obtained from other unobtrusive sources" (Gök, Waterworth and Shapira 2015, 653). However, they raise the criticism that obtaining information from website data is more difficult and care

needs to be taken when generating web-based indicators. The information on websites is generally more related to innovation output than input. In addition, websites are self-reported and firms are not publishing new information on their websites at equal rates. Beaudry, Héroux-Vaillancourt and Rietsch (2016) use a keyword technique to generate innovation indicators of Canadian aeronautic, space and defense, as well as nanotechnology-related firms based on the text on their websites. They find some significant correlation between their indicators and traditional ones. Nathan and Rosso (2017) combine UK administrative microdata, media and website content to develop experimental measures of firm innovation for SMEs. The authors use proprietary data gathered by a data firm which uses website and media content to model firms' lifecycle events such as new product and service launches. They are able to identify three times more product/service launches than patent applications from SMEs in 2014/2015. Nathan and Rosso (2017) conclude that web-based indicators are a useful complementary measure to existing metrics as they reveal additional information. Moreover, they find that past patent activities are related to a firm's current launch activities and that tech SMEs are substantially more launch-active than non-tech SMEs.

The study by Kim et al. (2012) is also worth mentioning here. They do not make use of firm websites but apply text mining methods to forecast technology developments. The use data from published papers and patents to detect emerging technologies and determine their stage of development. As patents tend to detect inventions rather than innovations, firm websites promise to provide additional insights for measuring technology developments with text mining tools.

Studies on web-based innovation indicators have thus confirmed that firm websites are an interesting and rich data source for examining the innovation activity of firms and innovation ecosystems in general. However, no consistent approach (like the one we presented in the previous section) on how to study firms' websites has yet been established. Moreover, the data source itself (i.e. the population of firm websites) has not been studied rigorously in terms of its qualitative and quantitative properties. A number of basic yet important data characteristics are still unknown:

- *Structure*: Structural properties (size/depth, type of information provided, technological framework, web technologies used, update frequencies, languages used) of firm websites are largely unknown.
- *Coverage*: Coverage and structure of firm websites may differ systematically depending on the sector, firm size, firm age or region.

3. Methods

Note on terminology: A *website* is the overall internet presence of a firm. A website consists of a number of *webpages* (e.g. "www.firm-name.com", "www.firm-name.com/products"). The highest level webpage is called the *homepage* or the *main page* (e.g. "www.firm-name.com"), while lower level webpages are called *subpages* (e.g. "www.firm-name.com/products"), if a distinction has to be made. The first webpage downloaded from a website (the webpage corresponding to a URL in the user given list of URLs; this is usually the website's homepage) is referred to as the *start page*.

3.1. A web mining framework for mapping innovation ecosystems

Nowadays, almost all (relevant) firms have their own websites which they use to publish information about their products and services. We assume that they also use this platform to highlight new and innovative features. In addition, firm websites provide additional information about firm credibility, achievements, key personnel decisions, strategies and relationships with other firms (Gök et al., 2015). These aspects can all be related to a firm's innovation activity. Therefore, firm websites may reveal directly or indirectly whether new products, technologies, and processes are being implemented. While this data is publicly available, it is unstructured and stored in a decentralized manner. Therefore, there is a need for a consistent methodology for gathering and harmonizing the data, as well as for extracting innovation-related information which can be used to generate innovation indicators.

In Figure 1, we outline such a methodology in the form of a general analysis framework for mapping innovation ecosystems and generating web-based firm-level innovation indicators. Similar to traditional innovation indicators, the base data is a firm database which includes information on firm characteristics (e.g. sector, firm size) and, most importantly, the firms' website addresses (URLs). Ideally, the firm database has been matched to auxiliary databases containing established innovation indicators from questionnaire-based surveys, firm-level patenting data or literature data (LBIO), such that traditional innovation indicators are available for a subsample of the firms in the main dataset. In a first step, the firms' web addresses are passed to a web scraper. The web scraper is then used to download website content (texts, hyperlinks etc.) from the firms' websites. In a third step, data mining techniques are applied to extract information on the firms' innovation activities from the down-

loaded website content. Based on this information, novel innovation indicators can be constructed. At this stage, additional metadata on the firm can be used to support the analysis (pre-classification, classification model selection based on firm characteristics, information from established innovation indicators etc.). In a final step, the new innovation indicators are merged back into the firm database. This last step also establishes a direct firm-level link between the novel innovation indicator and the established indicators available from the auxiliary databases. This link can later be used to evaluate the new indicators against the traditional ones.

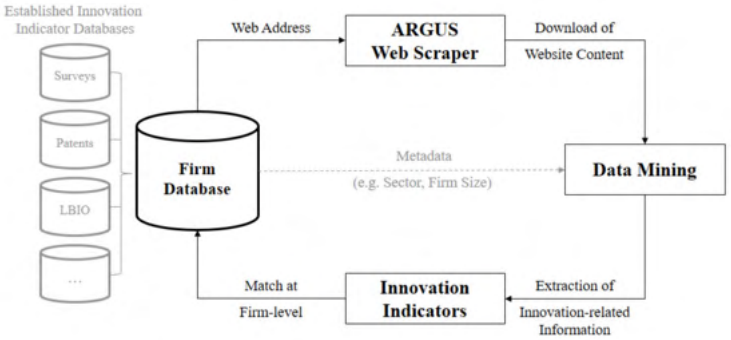


Figure 1. General analysis framework for mapping innovation ecosystems.

The proposed analysis framework allows for an automated, less costly mapping of entire firm populations that can be carried out faster and in shorter time intervals in comparison to traditional approaches. Also, this approach is easily expandable to map knowledge ecosystems (see e.g. Xu, Wu, Minshall, & Zhou, 2018) by scanning the websites of universities and research institutes. Furthermore, receiving firm information from websites does not require any effort on the part of the analyzed firms. As a result, web-based indicators created this way have the potential to outperform traditional indicators in terms of coverage, granularity, timeliness, and survey costs. The crucial point in our proposed framework is the identification and extraction of those pieces of information from the unstructured website content that reveal information about firms' innovation activities. Recent technological and methodological advances in analyzing unstructured data using machine learning (Grentzkow, Kelly, & Taddy, 2017; Mikolov, Deoras, Povey, Burget, & Cernocky, 2011; Steiger, Resch, & Zipf, 2016) may have that potential. Methods such as deep neural networks for natural language

processing and social network analysis are able to deal with the difficulties resulting from heterogeneous data sources and may be to extract interpretable and meaningful information on firms' innovation activities (see Conclusion and Future Research section).

3.2. ARGUS web scraper

ARGUS (**A**utomated **R**obot for **G**eneric **U**niversal **S**craping) is a web scraping software tool that was developed to meet the requirements that are determined by the web mining framework outlined in the previous section:

- **Adaptability**: The web scraper must be able to scrape a wide variety of web content from any website. At the same time, the web scraper's output must be in a structured and consistent format.
- **Scalability**: The web scraper must be able to scrape tens of millions of webpages from millions of firm websites in a reasonable time frame that allows for frequent iterations of the scraping process in order to build up a panel database of web data.
- **Easy-to-use**: The web scraper must be easy-to-use such that it can be used by researchers without profound knowledge in web scraping technology.
- **Free and Open Source**: In order to ensure a rapid dissemination as well as a sustainable further development of the web scraper, the program must be free-to-use and open source.

ARGUS is based on the Scrapy Python framework (Scrapy Community, 2008) and is available open source via Github (Kinne, 2018). The program features a graphical user interface (see Figure 2) that allows for a rather easy and command line free control.

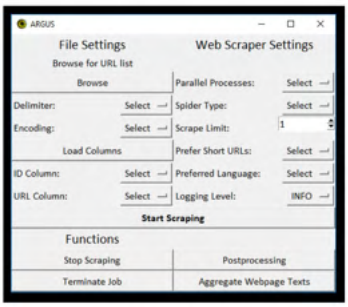


Figure 2. ARGUS graphical user interface.

4. Data

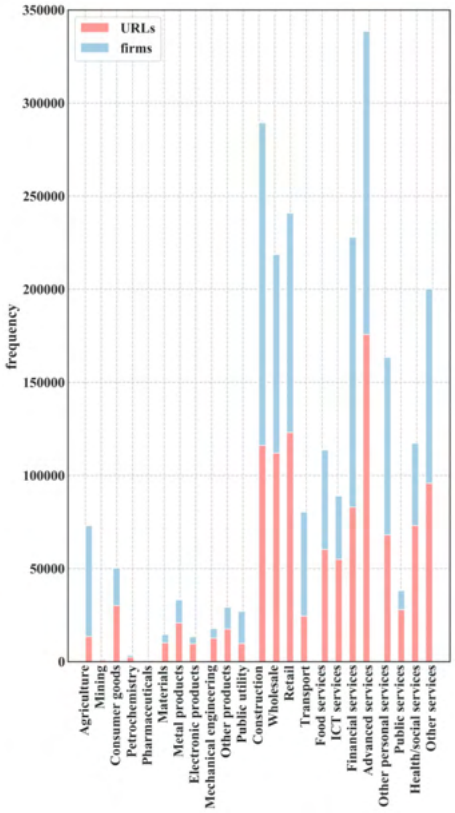
For the pilot study conducted in this paper, we use the *Mannheim Enterprise Panel* (MUP) as our base firm dataset. The MUP is a panel database that covers the total population of firms located in Germany. It contains about three million firm observations which are updated on a semi-annual basis. We restrict the dataset to firms that were definitely economically active in 2018 (2.52 million firms). The dataset also includes firm characteristics such as the industrial branch (NACE codes; a classification of economic activities in the European Union), postal addresses, number of employees, as well as the website address (URL) of the firm. For more information on the MUP see Bersch et al. (2014).

Patents are one of the most widely used and established innovation indicators (see e.g. Acs, Anselin, & Varga, 2002b; Archibugi & Pianta, 1996; Griliches, 1990; Nelson, 2009; OECD, 2009). We gathered patent data (patent stock end of 2017) from the European Patent Office and conducted a firm-patent match with our MUP firm database. Thereby, we restricted the patent dataset to patents that were filed after 2005 (10 years is the average lifetime of a patent in our database) to account for the decreasing economic and technological value of aging patents (Behrens, Hünermund, Leitner, Licht, & Peters, 2018).

5. Results

5.1. URL coverage

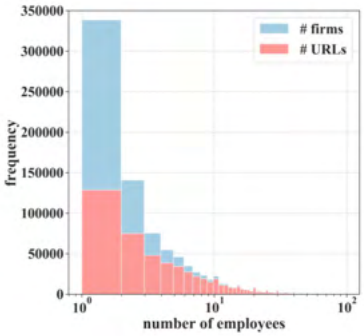
The overall URL coverage in our dataset is at 46% (1.15 million firms), but differs with firm size, sector, and location. Table 1 shows a breakdown of the firm population and URL coverage by sectors (a NACE code to sector mapping can be found in Table A1 in the appendix). Some sectors have a considerably higher URL coverage (≥70% coverage for materials, electronic products, mechanical engineering, and public services) than others (≤40% coverage for agriculture, public utility, construction, transport, financial services).



Sector	# firms	# URLs	:
Agriculture	73,111	13,507	0.18
Mining	1,528	795	0.52
Consumer goods	50,423	30,130	0.60
Petrochemistry	3,617	2,489	0.69
Pharmaceuticals	928	594	0.64
Materials	14,628	10,218	0.70
Metal products	33,267	20,934	0.63
Electronic products	13,432	9,675	0.72
Mechanical engin.	17,781	12,677	0.71
Other products	29,267	17,600	0.60
Public utility	27,038	9,718	0.36
Construction	289,399	116,137	0.40
Wholesale	218,664	112,070	0.51
Retail	240,857	123,141	0.51
Transport	80,373	24,585	0.31
Food services	113,688	60,258	0.53
ICT services	89,061	55,062	0.62
Financial services	227,927	83,064	0.36
Advanced services	338,519	175,774	0.52
Other personal serv.	163,391	68,133	0.42
Public services	38,217	27,918	0.73
Health/social serv.	117,383	73,084	0.62
Other services	200,293	95,896	0.48
MISSING sector	140,439	5,389	0.04
Total	**2,523,231**	**1,148,848**	**0.46**

Table 1. URL coverage by sector.

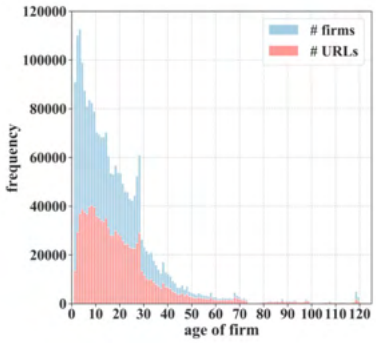
Table 2 shows firms' URL coverage by firm size groups (number of employees; variable available for 38% of firms). We can see that most firms are very small (micro-enterprises with less than 6 employees) and that coverage for this group is rather low (49%). For small firms (6-25 employees) coverage is decent (84%). Medium (26-250 employees) and large firms (>250 employees) are covered very well (94% and 97% respectively). These numbers are in line with official statistics, which cite the share of enterprises in Germany with websites at 87% for firms with 10 or more employees and 64% for firms with less than 10 employees (Eurostat, 2018). A two-sample t-test (see e.g. Krzywinski & Altman, 2013) indicated a highly significant difference in the number of employees between the overall firm population ($\bar{x}$=3.4) and the subpopulation covered by a URL ($\bar{x}$=19.6).



# employees	# firms	# URLs	:
1-5	655,617	324,393	0.49
6-25	229,995	193,648	0.84
26-250	71,778	67,132	0.94
>250	6,481	6,298	0.97
all	**963,871**	**591,471**	**0.61**

Table 2. URL coverage by firm size.

Table 3 shows firms' URL coverage by age (variable available for 91% of firms). Several historical events with an increased founding activity can be seen in the distribution (left panel): German Reunification (~28 years), constitution of the Federal Republic after the Second World War (~70 years), and the entrepreneurial boom of the *Gründerzeit* (~120 years). A trend of increasing URL coverage with firm age is visible: While very young firms (younger than two years) are poorly covered (18%), firms which are older than six years have better coverage (about 50%). It should be noted that firm age and firm size are positively correlated (Spearman's rho of 0.37; $p<0.001$). A two-sample t-test indicated a highly significant difference between the age the overall firm population ($\bar{x}$=16.7) and the URL covered subpopulation ($\bar{x}$=21.2).



Age	# firms	# URLs	:
0-1	138,922	25,005	0.18
2-5	395,725	146,670	0.37
6-25	1,210,762	616,722	0.51
26-100	521,821	273,435	0.52
>100	36,138	20,548	0.57
all	**2,303,368**	**1,082,380**	**0.47**

Table 3. URL coverage by firm age.

Figure 4 maps the ratio of firms with an available URL to the overall local firm population by district. Low and high ratios do not seem to be randomly scattered, but instead low coverage can be primarily found in the East of Germany, while the Western part seems to be well covered. This impression of non-randomness is confirmed by a high and significant *Moran's I* (see e.g. Fischer & Getis, 2010) value of 0.39 ($p<0.001$) indicating high positive spatial autocorrelation (clustering). We further identified several significant ($p<0.05$) local clusters of both high and low URL coverage using *Getis-Ord Gi** (Getis, 2009) measure of local autocorrelation. We also find that coverage is generally better in densely populated (urban) areas, indicated by a very high and significant correlation between population density and URL coverage at the level of districts (Spearman rho of 0.5; $p<0.001$).

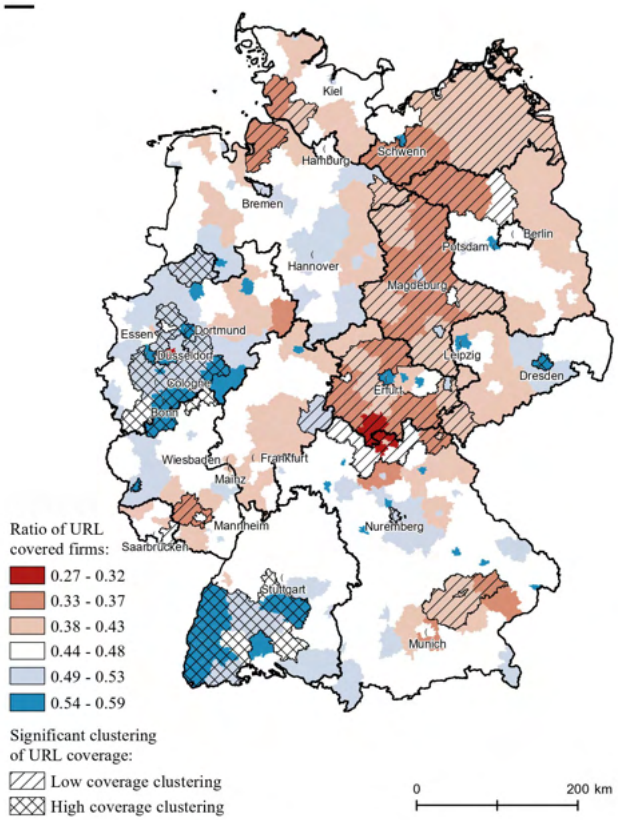


Figure 4. URL coverage by districts.

We investigate the relationships between the discussed firm characteristics and the availability of a URL in a probit regression analysis. The regression analysis results (as marginal effects) are shown in Table 4. *Broadband availability* is measured as the percentage of households in the firm's municipality that have potential access to broadband internet (≥50 Mbits download speed available; all technologies) (BKG, BMVI, & TÜV Rheinland, 2016). *Population density* controls for urban or rural firm locations and makes sure that broadband availability is not just a proxy for urban/rural firm location. *Employees*, *age*, and *sector* are defined as above.

Missing URLs in our data can result from either incomplete inquiry by our data provider or the fact that firms have actually no website. We investigate this issue by including two control variables in the regression analysis. Some legal forms do require a mandatory entry in official commercial registries – a procedure which makes surveying the firm a lot easier and, thus, likely increases the probability of a correctly entered URL in our data. We use information on the firms' *legal form* to control for this. The *search quality* variable controls for a possible bias in our data provider's search strategy too. We use the availability of a phone number in our data as an indicator for how well the firm was researched by the data provider.

The baseline firm in the regression is a mechanical engineering firm in a region with >95% broadband availability, 0 population density (rural area), >250 employees, >100 years of age, a legal form which requires an entry in the German commercial registry, and with an available phone number in our data. The pseudo-R^2 of the model is 0.19 and the mean variance inflation factor (VIF) is 9.36, which may indicate problematic multicollinearity in our model (the corresponding correlation Table A2 can be found in the appendix). While some authors emphasize a VIF of lower than 10 (Kutner, Nachtsheim, Neter, & Li, 2005), others suggest a significantly lower threshold of 3 (Tabachnick & Fidell, 2006).

Overall, the findings from the descriptive statistics are confirmed by the probit regression. Very young and very small firms do not have websites and the sector plays an important role. The regression also shows that firms in areas with low broadband availability are less likely to have a website. Our controls make us confident that this is not just a bias in the search strategy of our data provider. Instead, low broadband availability may detain firms from running their own website. According to our estimated effects, 30,000 firms in Germany (extrapolated to the total firm population) do not have an own websites because of their region's low high-speed Internet availability. This relates to 3.6% of firms in poor Internet regions, and to 1% of the total firm population in Germany respectively.

Table 4. Probit regression results. Dependent variable: Available firm website URL (yes/no).

Variable	Marginal effect	Robust Std. Error
Broadband (≥50Mbits) availability		
76-95%	-0.001	0.001
50-75%	-0.022***	0.001
10-50%	-0.044***	0.001
0-10%	-0.057***	0.002
Population density		
1,000 people/km²	0.008***	0.000
Employees		
MISSING	-0.484***	0.005
1-5	-0.373***	0.005
6-25	-0.134***	0.005
26-250	-0.041***	0.006
Age		
0-1	-0.242***	0.003
2-5	-0.093***	0.003
6-25	-0.061***	0.003
26-100	-0.072***	0.003
Sector		
Agriculture	-0.308***	0.004
Mining	-0.188***	0.013
Consumer goods	-0.052***	0.004
Petrochemistry	0.014	0.009
Pharmaceuticals	-0.027	0.016
Materials	-0.010	0.005
Metal products	-0.075***	0.005
Electronic products	0.030***	0.006
Other products	-0.041***	0.005
Public utility	-0.201***	0.005
Construction	-0.197***	0.004
Wholesale	-0.095***	0.004
Retail	-0.077***	0.004
Transport	-0.282***	0.004
Food services	-0.040***	0.004
ICT services	0.053***	0.004
Financial services	-0.176***	0.004
Advanced services	-0.060***	0.004
Other personal services	-0.129***	0.004
Public services	0.136***	0.004
Health/social services	0.021***	0.004
Other services	-0.030***	0.004
Legal form		
Registry entry not mandatory	-0.059***	0.001
Foreign legal form	0.358***	0.020
Search quality		
No other contact info	-0.362***	0.001

Baseline firm: Mechanical engineering firm in region with >95% broadband availability, >250 employees, >100 years old, has legal form which requires entry in commercial registry, and other contact info (phone) is available in data.

*p≤0.05, **p≤0.01, ***p≤0.001; n=2,108,104

Overall, 17,294 firms (0.6% of all firms) in our MUP dataset are patent holders and 71.47% of them are covered by a URL. Such a high URL coverage of patent holder firms was to be expected, given that mainly larger firms from sectors with a high URL coverage hold patents. As a result, patent holder firms will be overrepresented in web mining studies (1.3% patent holders after scraping compared to 0.6% in our base dataset). Figure 5 shows a breakdown of the share of patent holder firms by sector. While there is no eye-catching difference in the sector-level URL coverage of patent holder firms, the figures does highlight a well-known shortcoming of patents as innovation indicators. While patents play a crucial role to protect intellectual property in some sectors like mechanical engineering and pharmaceuticals other sectors where many firms may be considered as innovative patents do not fulfil this role. In the ICT services sector, for example, only 0.8% if firms hold patents, which is attributable to the fact that software is not patentable in Germany.

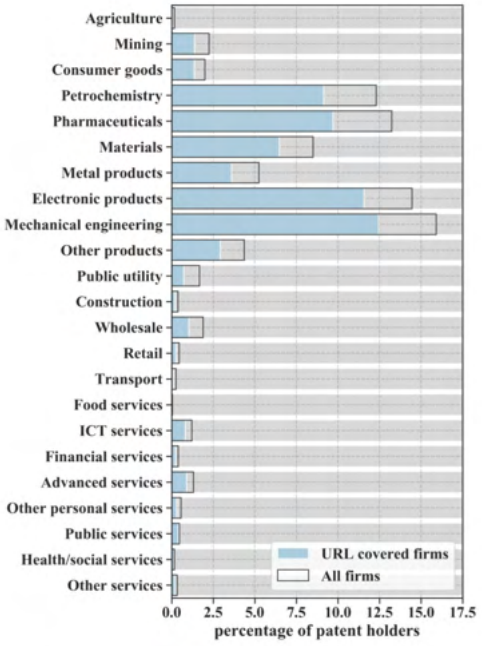


Figure 5. Share of patent holders by sector.

5.2. Website characteristics

For our further in-depth analysis of firm website characteristics, we randomly sampled 11,477 firms with a URL from our dataset and used ARGUS to scrape their websites. 84.2% of the websites could be scraped, while the he remaining 15.8% returned errors (DNS errors, timeouts, and HTTP errors) when requesting their start pages. T-tests between firms with successfully/not successfully requested websites showed no significant difference in firm size and age.

We then investigated the share of URLs for which initial requests are redirected. We only tag redirects if the redirect results in crawling a webpage from a different (second level) domain (e.g. "www.example.com" redirects to "www.sample.com"). Redirects between secure and standard HTTP (e.g. "http://www.example.com" to "https://www.example.com") and subdomain changes (e.g. "www.products.example.com" to "www.example.com") are not tagged as redirects. Redirects we tag can be both harmless (e.g. a firm registered a new domain and redirects there from its old domain) and severe (e.g. firm A was acquired by firm B and firm A's old URL now redirects to the website of its parent company B; small firms sometimes register domains but redirect to personal pages on social media like facebook.com). To be sure that the crawled website really belongs to the corresponding firm, redirected requests must either be checked thoroughly or excluded from the analysis. We opt for the latter and excluded 9.5% of the URLs that were successfully crawled but were also tagged as redirected. T-tests showed no significant difference in firms' age and size between redirecting and non-redirecting URLs. In sum, 23.8% of firms had to be excluded from further analysis due to redirect or request errors, reducing our sample to 8,744 firms.

For the remaining firms, the mean number of webpages per website is 218.8 (SD 604.7) and the median is 15, resulting in a highly skewed distribution, as it can be seen in Figure 6. A considerable share (5.86%) of the websites reached the *Scrape Limit* (see Methods section) of 2,500 subpages which we set for this analysis. Differences between sectors are stark as seen in Figure 7, where the mean number of webpages (indicated as red dots) vary considerably between sectors. Some of this variation is due to the positive correlation (Spearman's rho of 0.19; $p<0.001$) between firm size (which also varies systematically with the sector) and the number of webpages on a firm's website.

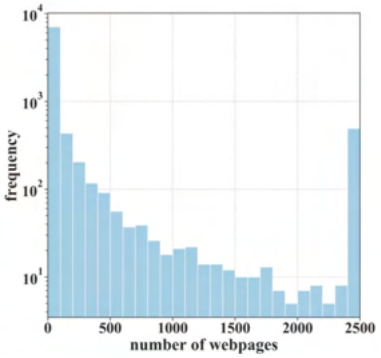


Figure 6. Number of webpages on a firm website.

Figure 7. Number of webpages on firm website by sectors.

On average, a webpage we have downloaded has 3295.86 characters (SD=9960.43) and half of them have 1970 characters or less (which equals about two thirds of a standard page of text), resulting in a highly skewed distribution as it seen in Figure 8. We did not find any statistically significant relationship between the mean text length per webpage and any firm characteristic.

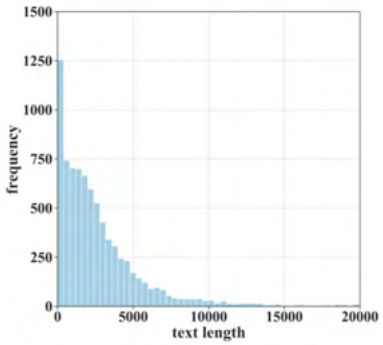


Figure 8. Mean text length per webpage.

We randomly sampled 911 websites and used Python's langdetect library (Danilak, 2015) to identify the languages used in each of their 193,504 sub-webpages. The algorithm was able to classify 91.9% of these webpages of which 88.2% were classified as being written in German. Most (60.8%) of the non-German language webpages were classified as written in English. Most of the firms have websites that are written almost completely in German (close to 100% of their webpages were classified as German), as it can be seen in Figure 9. Some firms only have non-German texts on their websites (share < 0.2; 4.5%). Figure 10 shows that the share of German language on a firm's website is related to the firm's sector (we do not show sectors with fewer than 10 observations). We do not find any other statistically significant relation to other firm characteristics.

It is important to keep in mind that sub-webpages were not selected uniformly or randomly from the firms' websites, as we used ARGUS' language selection heuristic set to German. Consequently, if a firm website was classified to be completely in German that does not automatically imply that the firm uses German exclusively on its website. Changing the preferred language from German to English decreases the share of German classified webpages from 88.2% to just 74.9% and increases the share of English webpages from 7.2% to 11.3%. This indicates that some firms have both German and English versions of their

website and ARGUS is indeed able to scrape a preferred language – a desirable feature as most natural language processing methods require text corpora in a single language.

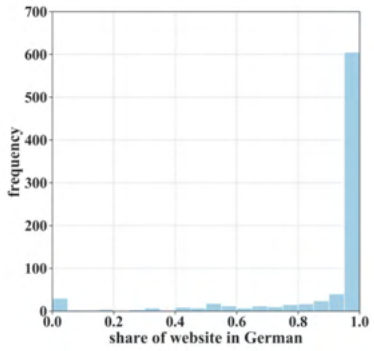


Figure 9. Share of website in German.

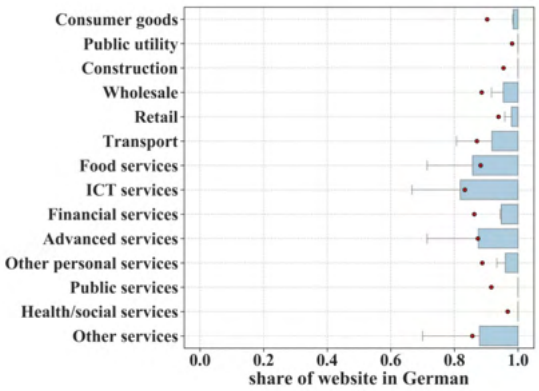


Figure 10. Share of website in German by sectors.

We also investigated the number of hyperlinks that connect a website to other websites in the World Wide Web by scraping our random sample of 11,477 firms using ARGUS' hyperlink scraping mode (*Scrape Limit* set to 100). We found that no website has less than 14 hyperlinks to other websites and some outlier websites have tens of thousands of such connections. The mean number of hyperlinks per website is 252.17 (SD 1779.69) and the median is 116. Unsurprisingly, the number of hyperlinks found on a firm's website is highly

correlated (Spearman's rho of 0.51; p<0.001) with the website's overall size (i.e. its number of sub-webpages). Looking at the mean number of hyperlinks per webpage, we see that, on average, a webpage contains 14.52 hyperlinks. The median number of hyperlinks per webpage is just 6, resulting in a highly skewed distribution as it can be seen in Figure 11. We did not find statistically significant relationships between the number of hyperlinks per webpage and any firm characteristics.

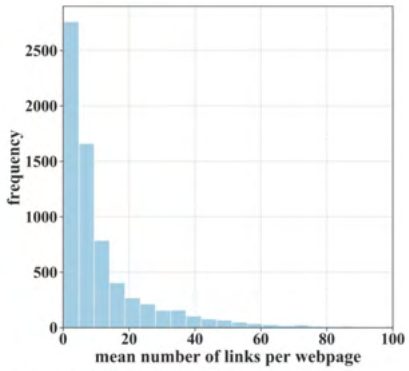


Figure 11. Mean number of hyperlinks per webpage.

5.2. Mapping an innovation ecosystem

In this section, we use our proposed framework (see Figure 1) and apply each outlined step (base dataset, web scraping, data mining, indicator creation, and evaluation/validation) to map an exemplary innovation ecosystem. We decided to investigate Berlin-based companies and scientific institutions that are engaged in artificial intelligence (AI). The German capital of Berlin is known for its thriving start-up tech scene. Its insular geographical location in the otherwise rather sparsely populated German East poses an ideal locally self-enclosed investigation area for a microgeographical study (see C. Rammer, Kinne, & Blind, 2020).

We used all entries in our MUP dataset with a postal address in Berlin and an available URL (n=74,202) as our **base dataset**. ARGUS was then used to **web scrape** the websites referenced by the URLs (*scrape limit* set to 50, *prefer short urls* activated, and language heuristic set go German). After excluding erroneous requests and redirects, 61,976 observations remained in our dataset.

For the **data mining** step, we decided to remain with a simple keyword search to identify firms and other institutions that are in some way engaged in AI. We defined a list of German and English keywords that comprise of different spellings and declensions of the word "artificial intelligence". We then tagged websites where at least one instance of any defined keyword is included. This simple **indicator** allows us to identify companies and institutions that report on their websites that they engage somehow in AI. One can argue that all Berlin-based companies that are part of this AI engaged community form an innovation ecosystem with actors that apply AI directly, use tools or have partners that incorporate AI, or at least have an AI related agenda. The latter especially applies to some of the many associations (industrial associations and other interest groups for example) that are located in the German capital. The overall share of firms and institutions that are part of this ecosystem is 2.49% when taking into account only those firms that mention AI on their websites and 7.86% when including also those firms with at least one AI engaged hyperlink partner.

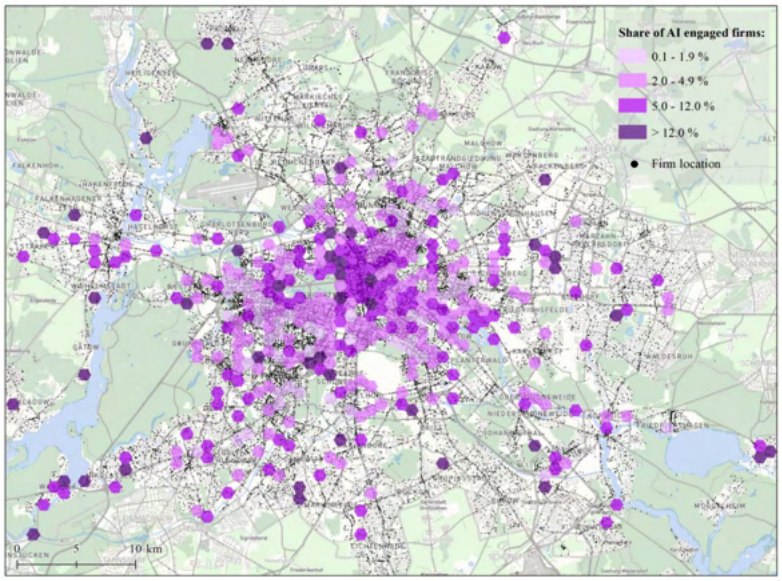


Figure 12. Share of Berlin-based firms that mention AI at least once on their websites. Basemap: Mapbox.

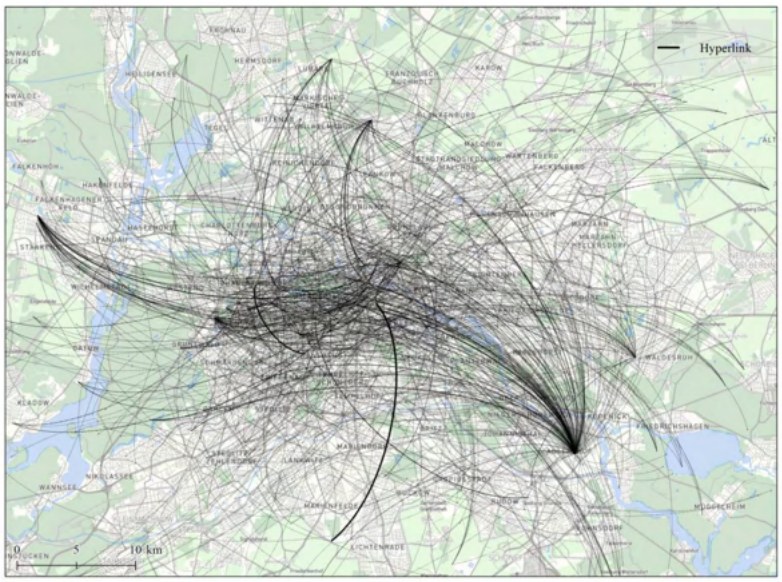


Figure 13. Incoming and outgoing hyperlinks to Berlin-based firms that mention AI at least once on their websites. Basemap: Mapbox.

Figure 12 maps the locations of all firms in our Berlin sample and the share of AI engaged firms per one kilometer hexagons (only hexagons with at least five firm observations shown). It can be seen that higher shares can be found all over the greater metropolitan area with some clustering in the city center, especially the Eastern part of the city center. Figure 13 maps incoming and outgoing hyperlinks to websites of firms that mention AI at least once on their website. For visualization purposes, we aggregated all firm locations using the same hexagons used in Figure 12. The edge weightings (displayed as edge thickness) results from the number of hyperlinks between individual hexagons (i.e. between the websites of firms located in each hexagon).

For **validation** purposes, we compare our results against survey data from the 2019 German Community Innovation Survey (CIS). The CIS is a European-wide, questionnaire-based innovation survey which is conducted annually using a stratified sample of about 20,000 firms from the MUP firm database (see Rammer et al., 2019). The CIS sample is restricted to firms with at least five employees and sectors from manufacturing and business-oriented services. In the 2019 CIS, firms were asked "Does your enterprise use artificial intelligence

methods?" with the possibility to tick either "yes" or "no". The survey answers were used to extrapolate numbers that are representative for the firm population covered in the CIS (i.e. manufacturing and business-oriented services firms with at least five employees). We use these extrapolations to compare our web-based results against the survey results. For Figure 13, we restricted our dataset to firms with available information on the number of employees that are from sectors which are in the CIS survey population (manufacturing and business-oriented services; n=5,785 in Berlin). In this subgroup, our web-based results indicate that 4.7% of firms are engaged in AI, ranging from 3.24% for firms with less than five employees to 24.53% for firms with at least 250 employees. This positive correlation between firm size and AI engagement can be observed from the survey data as well even though there are differences concerning the individual size groups.

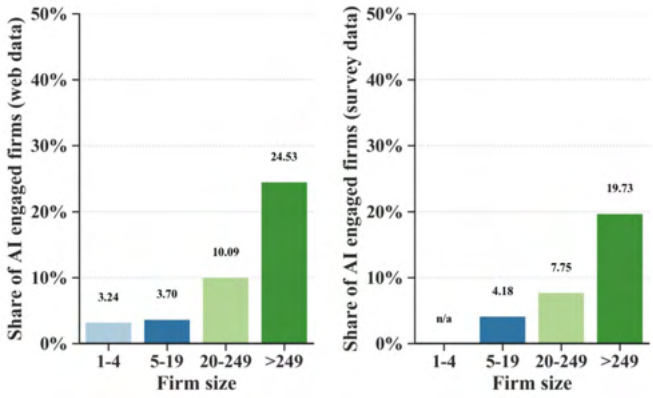


Figure 13. Share of Berlin-based firms that mention AI at least once on their websites (left panel) and share of Germany-based firms that state to use AI in the CIS survey data (right panel).

Figure 14 shows the sectoral breakdown of firms that are engaged in AI (i.e. name AI on their websites) and the share of firms that have at least one AI engaged partner (i.e. an existing hyperlink between the firm and another firm that is engaged in AI). It can be seen that associations seem to play a significant role in the ecosystem we are trying to map. This sector shows both the highest share of institutions that mention AI on their website (6.6%) and the highest share (17.7%) of institutions that are connected to at least one intuition that mentions AI on its website. Other sectors with a comparatively high share of AI engaged firms are

professional services (which also include software companies) and education (which in addition to schools also includes universities and research institutes). Low shares can be observed in sectors like construction, mining, the hospitality industry, and (rather surprisingly) in the chemical/pharmaceutical industry. Concerning the share of firms with at least one AI engaged partner, the healthcare sector shows a comparatively high share (9.7%) of institutions with at least one AI engaged firm, even though the sector itself shows a very low share of institutions that are engaged in AI themselves (0.2%). Our manual investigations reveals that this stems partly from the fact that websites of doctor's offices oftentimes hyperlink to medical organizations and societies that feature AI related agendas (for example on AI image recognition methods in radiology). Other sectors with a high share of firms with at least one AI engaged partner are professional services, education, and public utility. Low shares are exhibited by the construction and mining sector.

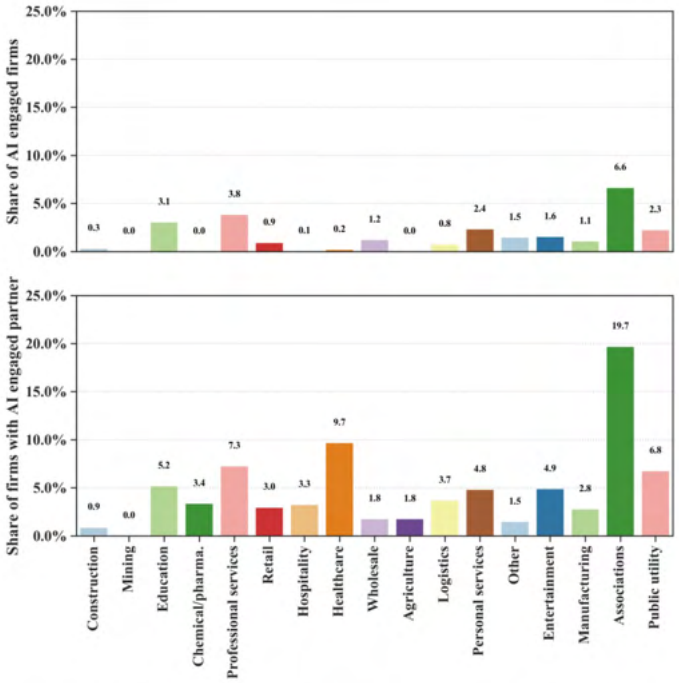


Figure 14. Share of firms that mention AI at least once on their websites (upper panel); share of firms with hyperlinked partner that mentions AI at least once (lower panel).

Figure 15 shows the age and size group breakdowns of firms that are engaged in AI or have hyperlink partners that are engaged in AI. Concerning firm size, the pattern seen in Figure 13 is repeated for this slightly altered size groupings (i.e. larger companies are more likely to be engaged in AI or to have at least one AI engaged partner). Interestingly, this pattern is reversed for the breakdown by firm age. Here, younger firms are more likely to be engaged in AI (4.90% of firms younger than one year, compared to 2.01% of firms older than 25 years). This is especially interesting, given that firm size and age are highly correlated (Spearman's correlation of 0.31). This result may indicate that young firms that engage in AI are also among the ones that grow the fastest in terms of their number of employees. Looking at the share of firms with at least one AI engaged hyperlink partner, we see fewer differences concerning the different age groups.

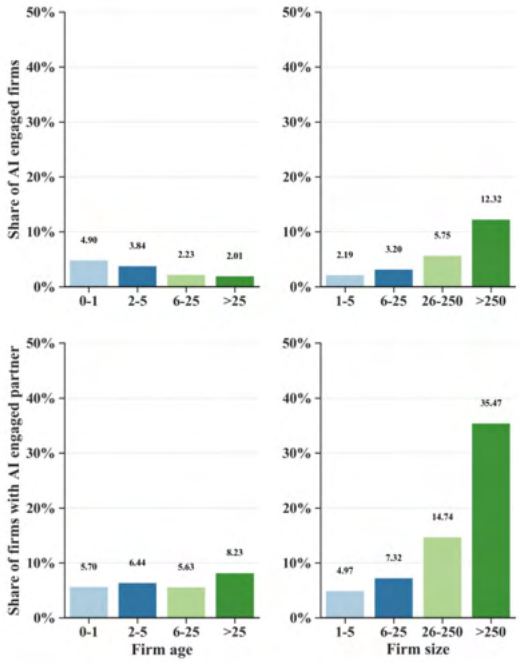


Figure 15. Firm age (left) and size (right) of firms that mention AI at least once on their websites (upper panel); firm age (left) and size (right) of firms with one or more hyperlinked partners that mentions AI at least once
(lower panel).

6. Discussion

In the first part of our study, we investigated what firms in the total population of firms actually have their own websites (URL coverage) which would allow researchers to survey them in a web-based study. Thereby, we put particular emphasis on firm characteristics and their statistical relations to the URL coverage in the overall firm population. For this purpose, we also tried to untangle the cause of missing URLs in our firm dataset and distinguish between *true* missing values (the firm has no website) and *false* missing values (the firm has a website, but it was not found by our data provider). Based on our case study results, regularities in URL coverage remain after controlling for a potential bias in the search strategy of our data provider. Researchers who conduct web mining to map innovation ecosystems - as we proposed it in our framework - will have difficulties observing very young and very small firms, especially those from certain sectors such as agriculture and those located in rural areas. In addition, low broadband availability seems to deter firms from setting up their own website and therefore systematically excludes them from any web-based studies. If one assumes that low broadband availability is associated to a generally lower use of the Internet (both private and commercial) in a region, this may actually indicate that firms with local target markets that are located in an area with a low broadband availability have no incentive to set up their own website in order to communicate with their customers. On the other hand, our results show that medium-sized and medium-aged, as well as large firms can be thoroughly surveyed using our proposed web mining framework. This is especially true in urban areas. Given that the vast majority of innovative activity in Germany is conducted by the latter firm type (Rammer, et al. , 2017), we can conclude that our web mining framework is suitable for analyzing the most important business-side parts of the German innovation ecosystem. This assumption is backed by our finding that patenting firms are overrepresented in web mining studies due to the higher URL coverage in patent-intensive firm subgroups.

We identified URL redirects as a potential issue when conducting web mining studies because outdated URLs can result in potentially harmful redirects. If conducting a large-scale web study based on a huge firm datasets, it is usually not possible to make sure that the available firm website addresses are all up-to-date. To minimize the share of erroneous scraped we content, we therefore recommend excluding firms such URL redirects. Given that less than 10% of successful URL requests were redirected and we did not find any systematic firm age or size bias, such an exclusion seems reasonable.

Our results showed that firm website size is highly correlated to firm size (number of employees) and sectors. Large firms have both more webpages on their websites and more

text on each of these webpages. In general, we find that outliers play an important role when conducting web mining studies. Some websites are extremely large in terms of the number of webpages and the amounts of text provided on them. This outlier issue also causes the mean number of webpages per website to vary quiet strongly between sectors. On the other hand, the median number of webpages per website is rather stable across sectors (about 15 webpages per website). To completely scrape two thirds of all firm websites, it is therefore sufficient to set the limit of downloaded webpages per website to 50. If this threshold is increased to 250, 90% of the websites can be scraped entirely. About 6% of firms can be seen as extreme outliers with 2,500 or more sub-webpages on their websites.

Based on these purely quantitative results, it is difficult to make any generally applicable best practice recommendation for an appropriate *Scrape Limit* for ARGUS. If researchers are interested in generating a more general textual description of the firms, they may select a rather low Scrape limit of 15 and would still scrape half of all firm websites entirely. If they are interested in highly specific information, that may be located on lower levels of the website, the need to set a rather high scrape limit around 250. In this sense, our results should provide researchers with a sound reference point when conducting their own web mining studies.

Unsurprisingly, our results showed that most websites of Germany-based firms are in German. However, a considerable share (about 5%) of the firms have mostly (≥80%) non-German texts on their websites. We were also able to show that the ARGUS simple language selection heuristic helps to restrict the downloaded texts downloaded to a certain language. Given that most natural language processing algorithms require text corpora to be in a single language, this is a significant result. We were also able to show that a considerable share of firms provide several versions of their website in different languages. The language selection heuristic of ARGUS is likely to be even more important when working with websites from multilingual countries (e.g. Switzerland, Belgium). Furthermore, we found significant sectoral differences in the use of language. Some sectors (e.g. agriculture, personal services, construction) mostly use German, while others (e.g. mechanical engineering, pharmaceuticals) use other languages as well. We assume that the sector's orientation towards either local/national or international markets may play an important role here.

The total number of hyperlinks that can be found on firm websites is, unsurprisingly, highly correlated to the number of webpages it has. The mean number of links per webpage, however, seems to be randomly distributed with no significant relationship to the firm size, age, or sector. If hyperlinks between firms are interpreted as some kind of relationship (e.g.

customer, cooperation), this would indicate that, on average, the connectedness of a firm grows with its size. A qualitative analysis of these connections could reveal whether certain types of firms (e.g. innovative ones) are connected differently (e.g. regional vs. transregional) compared to other firm types (e.g. non-innovative firms).

In the last part of this study, we used our proposed framework and applied the described workflow (using a firm base dataset to scrape firm websites, apply data mining, creation and validation of web-based indicators) for an exploratory analysis of the artificial intelligence (AI) ecosystem of Berlin-based institutions. The German capital has been chosen due to its thriving tech scene and its insular geographical location. We used a keyword-based approach to identify those institutions that mention AI at least once on their websites. Arguably, this approach does not necessarily inform about institutions that apply AI in their production process, offer products that incorporate AI features or conduct AI-related research and development. Nevertheless, it can be assumed that institutions that decide to mention AI on their websites at least somehow deal with this technology. This engagement may range from basic research to product development to a superficial marketing strategy. It can therefore be argued that these companies are in some way involved in the "Innovation Ecosystem AI Berlin". Although we assume that our simple approach is suitable to provide a first insight into this ecosystem, future research should definitely undertake a further distinction of the identified actors. Here we suggest, for example, that a sample of actors could be drawn, which could then be manually classified into a certain class (e.g. research, product development, marketing strategy, etc.). This manually labelled data set could then be used as training data for a text-based machine learning model, which would be trained for the classification of actors based on their web texts.

The comparison between our new web-based indicator on "AI engagement" and an indicator on the use of AI collected in a classical survey has shown that even our very simple, keyword-based approach seems to deliver meaningful results. At the same time this comparison also shows the potential of our approach. While the costly survey only provides information for a Germany-wide extrapolation, our web-based approach allowed us to collect information for about 60,000 companies in Berlin alone. Unlike the survey data, our data contains information on all industries and size classes. Overall, we are confident that the approach we have presented has the potential to provide valuable, comprehensive and cost efficient insights that compare well to traditional sources.

7. Conclusion and Future Research

7.1. Conlusion

In this paper, we proposed a web mining framework for the mapping of innovation ecosystems by generating innovation indicators from website contents. We argued that established innovation indicators have a number of shortcomings concerning their coverage, granularity, timeliness, and data collection costs and that web-based indicators have the potential to overcome some of these limitations. The proposed web mining framework is composed of four key parts: a firm database with firm-level metadata and the firms' web addresses, ARGUS web scraper which is used to download firm website content, a data mining part to extract innovation-related information from the downloaded web content, and the actual innovation indicators generated from the extracted information. In the remainder of the paper we conducted a large-scale pilot study to investigate firm websites as a potentially valuable data source for innovation ecosystem mapping and we used our proposed approach to study the "Innovation Ecosystem AI Berlin". Two research questions were the guideline for this pilot study.

- **URL coverage:** URL coverage (the availability of a website for a firm) differs systematically with firm characteristics. Certain types of firms can, thus, not be surveyed using our proposed web mining framework. Especially very young and very small firms, as well as firms from certain sectors and regions exhibit a very low URL coverage. Furthermore, we find that low local broadband availability can prevent firms from setting up their own internet presence. On the other hand, we find that almost all medium to large sized firms from sectors such as mechanical engineering and ICT services have websites. We also found that URL coverage is especially high among patenting firms. Given that the vast majority of innovative activity in Germany is conducted by these firm types, we can conclude that our web mining framework is suitable for analyzing the most important parts of the firm innovation systems.
- **Website characteristics:** We concluded that web mining studies have to deal with outlier issues. About 6% of firm websites have a number of sub-webpages four or more standard deviations above the population mean. Concerning the number of hyperlinks and the text volume found on these websites, this issue is even more evident. Large firms do not only operate larger websites, they also provide disproportionally more hyperlinks and text on them. We also found that

there are sectoral differences concerning the size of firm websites and the languages used on them. We were also able to show that the language selection heuristic of ARGUS effectively restricts text downloads to a certain language, which allows users to leverage the fact that many firms provide several versions of theirs websites in different languages. An important feature given that most natural language processing methods require texts in a single language.

- **Mapping an Innovation Ecosystem:** We showed that our proposed approach can be used to identify firms and other institutions that are engaged in a certain activity or technology and report on that on their websites. Using the example of AI-engaged institutions in the German capital of Berlin, we applied a simple keyword based approach and hyperlink mining to map an innovation ecosystem at the microgeographic level. Our results compared well to traditional survey data on the use of artificial intelligence in terms of firm size. However, we also pointed out that a more sophisticated text mining approach would be necessary to distinguish the different actor groups (e.g. firms that offer AI-based products and services, universities that are engaged in basic research on AI, and interest groups that promote AI-centered agendas) that resulted from our simple keyword search.

7.1. Future Research

In future research, the analysis of the downloaded web data and the inclusion of other subsystems of the innovation ecosystem (e.g. via the websites of universities and research institutes) should be in the focus. For the analysis of textual content, several approaches may be suitable. If researchers want to investigate a topic that can be adequately described using a set of keywords (e.g. specific technologies, standards, patent numbers, policy measures) a simple keyword search can be sufficient. In such a keyword search, firms can be identified that use these keywords on their websites. Smarter search strategies with additional filtering words and the like may be used to refine the results.

Recent developments in the field of natural language processing (NLP) (e.g. Mikolov *et al.*, 2011, 2013; Mikolov, Yih and Zweig, 2013), especially the ones involving artificial neural network language models, resulted an array of potentially valuable approaches to extract innovation related information from web scraped texts. A possible approach to predict a firm's innovation activity as outlined in Figure 16. A neural network is trained using texts scraped from websites of firms for which established innovation indicators are available.

Such indicators can be used to create a training dataset of labelled (innovative/non-innovative) website texts. After training the neural network, unlabeled website texts (i.e. texts from websites of firms with unknown innovation activity) can be examined by the network and given a probability of being scraped from an innovative firm's website. Given that such information is available, additional firm metadata (e.g. the sector of the firm) could be used to enhance the model.

Text mining methods based on neural networks and semantic topic models were also successfully applied in geographical information science (GIScience) to uncover social phenomena from geocoded unstructured text data. Resch, Usländer, & Havas (2018) for example, present an approach to assess the footprint of and the damage caused by natural disasters by combining machine learning techniques for semantic information extraction. They also showed that their approach can be used to identify relevant semantic topics without a priori knowledge. Their methodology may be applicable to detect and monitor the diffusion of technology, for example.

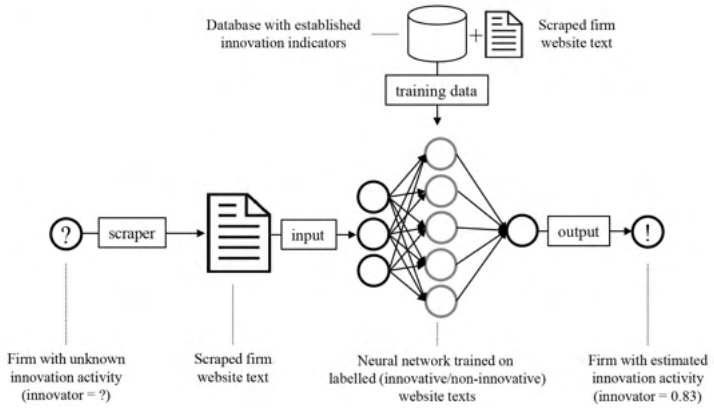


Figure 16. Proposed artificial neural network based innovation prediction model.

References

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.

Acs, Z. J., Anselin, L., & Varga, A. (2002a). Patents and innovation counts as measures of regional production of new knowledge. *Research Policy*, *31*(7), 1069–1085. https://doi.org/10.1016/S0048-7333(01)00184-6

Acs, Z. J., Anselin, L., & Varga, A. (2002b). Patents and Innovation Counts as Measures of Regional Production of New Knowledge. *Research Policy*, *31*(7), 1069–1085. https://doi.org/10.1016/S0048-7333(01)00184-6

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

Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: a pilot web-based study on graphene firms. *Scientometrics*, *95*(3), 1189–1207.

Arzaghi, M., & Henderson, J. V. (2008). Networking off Madison Avenue. *Review of Economic Studies*, *75*(4), 1011–1038. https://doi.org/10.1111/j.1467-937X.2008.00499.x

Askitas, N., & Zimmermann, K. F. (2015). The Internet as a data source for advancement in social sciences. *International Journal of Manpower*, *36*(1), 2–12. https://doi.org/10.1108/IJM-02-2015-0029

Basole, R. C., Huhtamäki, J., Still, K., & Russell, M. G. (2016). Visual decision support for business ecosystem analysis. *Expert Systems with Applications*, *65*(August), 271–282. https://doi.org/10.1016/j.eswa.2016.08.041

Basole, R. C., Russell, M. G., Huhtamäki, J., Rubens, N., Still, K., & Park, H. (2015). Understanding business ecosystem dynamics: A data-driven approach. *ACM Transactions on Management Information Systems*, *6*(2). https://doi.org/10.1145/2724730

Beaudry, C., Héroux-Vaillancourt, M., & Rietsch, C. (2016). Validation of a web mining technique to measure innovation in high technology Canadian industries. In *CARMA 2016–1st International Conference on Advanced Research Methods and Analytics* (pp. 1–25).

Behrens, V., Hünermund, P., Leitner, S. M., Licht, G., & Peters, B. (2018). *Investigating the Impact of the Innovation Union: State of Implementation and Direct Impact Assessment*. Maastricht.

Bersch, J., Gottschalk, S., Müller, B., & Niefert, M. (2014). *The Mannheim Enterprise Panel (MUP) and firm statistics for Germany. ZEW Discussion Paper*. https://doi.org/10.2139/ssrn.2548385

BKG, BMVI, & TÜV Rheinland. (2016). *Broadband Atlas*. Berlin. Retrieved from https://www.bmvi.de/DE/Themen/Digitales/Breitbandausbau/Breitbandatlas-Karte/start.html

Carlino, G., & Kerr, W. R. (2015). Agglomeration and Innovation. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of Regional and Urban Economics* (Vol. 5, pp. 349–404). Amsterdam: Elsevier North-Holland. https://doi.org/10.1016/B978-0-444-59517-1.00006-4

Catalini, C. (2012). *Microgeography and the Direction of Inventive Activity*. *Rotman School of Management Working Paper* (Vol. 2126890). https://doi.org/10.1287/mnsc.2017.2798

Coombs, R. (1996). Core competencies and the strategic management of R&D. *R&D Management*, *26*(4), 345–355. https://doi.org/10.1111/j.1467-9310.1996.tb00970.x

Danilak, M. (2015). langdetect. Retrieved from https://pypi.org/project/langdetect/

Eurostat. (2018). EUROSTAT. Retrieved July 18, 2018, from http://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-714389_QID_3257D732_UID_-3F171EB0&layout=TIME,C,X,0;SIZEN_R2,B,Y,0;GEO,B,Y,1;INDIC_IS,B,Z,0;UNIT,B,Z,1;INDICATORS,C,Z,2;&zSelection=DS-714389INDICATORS,OBS_FLAG;DS-714389UNIT,PC_ENT;DS-7143

Fischer, M. M., & Getis, A. (2010). *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Heidelberg, Berlin: Springer. https://doi.org/10.1017/CBO9781107415324.004

Getis, A. (2009). Spatial Weights Matrices. *Geographical Analysis*, *41*(4), 404–410.

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

Grentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as Data* (NBER Working Paper Series No. 23276). Cambridge, Massachusetts.

Griliches, Z. (1990). *Patent statistics as economic indicators: A survey* (NBER working paper No. 3301). *NBER working paper*. Cambridge, Massachusetts.

Jang, S., Kim, J., & von Zedtwitz, M. (2017). The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research*, *78*(June), 143–154. https://doi.org/10.1016/j.jbusres.2017.05.017

Katz, J. S., & Cothey, V. (2006). Web Indicators for Compley Innovation Systems. *Research Evaluation*, *45*(5), 893–909. https://doi.org/10.1016/j.respol.2006.03.007

Kerr, W. R., Duranton, G., Glaeser, E., & Henderson, V. (2014). Agglomerative Forces and Cluster Shapes. *Review of Economics and Statistics*, *96*(3).

Kim, J., Hwang, M., Jeong, D.-H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. *Expert Systems with Applications*, *39*(16), 12618–12625. https://doi.org/https://doi.org/10.1016/j.eswa.2012.05.021

Kinne, J. (2018). ARGUS - An Automated Robot for Generic Universal Scraping. Mannheim: Centre for European Economic Research. https://doi.org/10.1109/LPT.2009.2020494

Kleinknecht, A., & Reijnen, J. O. N. (1993). Towards literature-based innovation output indicators. *Structural Change and Economic Dynamics*, *4*(1), 199–207. https://doi.org/10.1016/0954-349X(93)90012-9

Kleinknecht, A., Van Montfort, K., & Brouwer, E. (2002). The Non-Trivial Choice between Innovation Indicators. *Economics of Innovation and New Technology*, *11*(2), 109–121. https://doi.org/10.1080/10438590210899

Krzywinski, M., & Altman, N. (2013). Points of significance: Significance, P values and t-tests. *Nature Methods*, *10*(11), 1041–1042. https://doi.org/10.1038/nmeth.2698

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, NY: McGraw-Hill Irwin.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.1162/153244303322533223

Mikolov, T., Deoras, A., Povey, D., Burget, L., & Cernocky, J. (2011). Strategies for Training Large Scale Neural Network Language Models. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.1109/ASRU.2011.6163930

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL-HLT* (pp. 746–751). https://doi.org/10.3109/10826089109058901

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Cambridge, Massachusetts: Academic Press.

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent Statistics as an Innovation Indicator. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of Economics of Innovation* (Vol. 2, pp. 1083–1127).

Nathan, M., & Rosso, A. (2017). *Innovative Events* (Centro Studi Luca d'Agliano Development Studies Working Paper No. 429).

Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, *38*(6), 994–1005. https://doi.org/10.1016/j.respol.2009.01.023

OECD. (2009). *OECD Patent Statistics Manual*. Paris: OECD. https://doi.org/10.1787/9789264056442-en

OECD. (2017). *Broadband Portal*. Paris. Retrieved from www.oecd.org/sti/broadband/oecdbroadbandportal.htm

OECD, & Eurostat. (2018). *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation* (4th ed.). Luxembourg, Paris: OECD/eurostat. https://doi.org/10.1787/9789264304604-en

Rammer, C., Kinne, J., & Blind, K. (2020). Knowledge proximity and firm innovation: A microgeographic analysis for Berlin. *Urban Studies*, *57*(5), 996–1014. https://doi.org/10.1177/0042098018820241

Rammer, C, Aschhoff, B., Doherr, T., Peters, B., & Schmidt, T. (2017). *Innovationsverhalten der deutschen Wirtschaft. Indikatorenbericht zur Innovationserhebung 2016*. Mannheim.

Rammer, Christian, Behrens, V., Doherr, T., Hud, M., Köhler, M., Krieger, B., … von der Burg, J. (2019). *Innovationen in der deutschen Wirtschaft*. Mannheim.

Raymond, K., & Blockeel, H. (2000). Web Data Mining research: A survey. *SIGKDD Explorations*, *2*(1), 1–10. https://doi.org/10.1109/ICCIC.2010.5705856

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362–376. https://doi.org/10.1080/15230406.2017.1356242

Rubens, N., Still, K., Huhtamaki, J., & Russell, M. G. (2011). A network analysis of investment firms as resource routers in Chinese innovation ecosystem. *Journal of Software*, *6*(9), 1737–1745. https://doi.org/10.4304/jsw.6.9.1737-1745

Scrapy Community. (2008). Scrapy. Scrapinghub Ltd. Retrieved from https://github.com/scrapy/scrapy

Shepherd, W. G., & Shepherd, J. M. (2003). *The Economics of Industrial Organization*. Long Grove, IL: Waveland Press Inc.

Squicciarini, M., & Criscuolo, C. (2013). *Measuring Patent Quality* (OECD Science, Technologya dn Industry Working Papers No. 2013/03). Paris. https://doi.org/http://dx.doi.org/10.1787/5k4522wkw1r8-en

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographic Information Science*, *30*(9), 1694–1716.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). London: Pearson.

Xu, G., Wu, Y., Minshall, T., & Zhou, Y. (2018). Exploring innovation ecosystems across science, technology, and business: A case of 3D printing in China. *Technological Forecasting and Social Change*, *136*(June 2017), 208–221. https://doi.org/10.1016/j.techfore.2017.06.030

Youtie, J., Hicks, D., Shapira, P., & Horsley, T. (2012). Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. *Technology Analysis and Strategic Management*, *24*(10), 981–995. https://doi.org/10.1080/09537325.2012.724163

Appendix

NACE code range	Sector label	Level 1 codes
0-4999	Agriculture	A
5000-9999	Mining	B
10000-18999	Consumer goods	C
19000-20999	Petrochemistry	C
21000-21999	Pharmaceuticals	C
22000-24999	Materials	C
25000-25999	Metal products	C
26000-27999	Electronic products	C
28000-30999	Mechanical engineering	C
31000-34999	Other products	C
35000-40999	Public utility	D, E
41000-44999	Construction	F
45000-46999	Wholesale	G
47000-48999	Retail	G
49000-54999	Transport	H
55000-57999	Food services	I
58000-63999	ICT services	J
64000-68999	Financial services	K
69000-76999	Advanced services	M
77000-83999	Other personal services	M
84000-85999	Public services	O,P
86000-89999	Health/social services	Q
90000-99999	Other services	R

Table A1. Sectors' NACE code ranges.

Variable	Broadband	Population density	Employees	Age	Legal form	Search quality
Broadband	-					
Population density	-0.44	-				
Employees	-0.01	0.02	-			
Age	0.04	-0.08	0.04	-		
Legal form	-0.09	0.10	-0.10	-0.10	-	
Search quality	0.06	-0.10	0.15	0.37	-0.10	-

n=2,108,104; p≤0.001 for all correlations.

Table A2. Correlation (Spearman's rho) table.

Appendix D

Paper 4: Predicting Innovative Firms using Web Mining and Deep Learning

// NO.19-001 | 12/2019

DISCUSSION PAPER

// JAN KINNE AND DAVID LENZ

Predicting Innovative Firms using Web Mining and Deep Learning

ZEW

Predicting Innovative Firms using Web Mining and Deep Learning

Jan Kinne[a,b,c,1,2] **and David Lenz**[c,d,1]

[a]Department of Economics of Innovation and Industrial Dynamics, ZEW Centre for European Economic Research, Mannheim, Germany; [b]Department of Geoinformatics Z_GIS, University of Salzburg, Salzburg, Austria; [c]istari.ai, Mannheim, Germany; [d]Department of Econometrics and Statistics, Justus-Liebig-University, Gießen, Germany

Abstract

Evidence-based STI (science, technology, and innovation) policy making requires accurate indicators of innovation in order to promote economic growth. However, traditional indicators from patents and questionnaire-based surveys often lack coverage, granularity as well as timeliness and may involve high data collection costs, especially when conducted at a large scale. Consequently, they struggle to provide policy makers and scientists with the full picture of the current state of the innovation system. In this paper, we propose a first approach on generating web-based innovation indicators which may have the potential to overcome some of the shortcomings of traditional indicators. Specifically, we develop a method to identify product innovator firms at a large scale and very low costs. We use traditional firm-level indicators from a questionnaire-based innovation survey (German Community Innovation Survey) to train an artificial neural network classification model on labelled (product innovator/no product innovator) web texts of surveyed firms. Subsequently, we apply this classification model to the web texts of hundreds of thousands of firms in Germany to predict whether they are product innovators or not. We then compare these predictions to firm-level patent statistics, survey extrapolation benchmark data, and regional innovation indicators. The results show that our approach produces reliable predictions and has the potential to be a valuable and highly cost-efficient addition to the existing set of innovation indicators, especially due to its coverage and regional granularity.

Keywords: Web Mining | Innovation | Deep Learning | Natural Language Processing
JEL Classification: O30, C81, C83

First version: January 2019
This version: December 2019

Innovations can disrupt individual industries with game-changing technology and the most radical innovations can even reshape whole economies. Despite having a destructive element, innovation is widely considered to be a main driver of long-term economic growth. Such growth may be kick-started by radical innovations or driven forward by a constant stream of so called incremental innovations which cause continuous change. Measuring and promoting innovation is the main objective of STI (science, technology and innovation) policy, which requires an accurate and timely picture of the current state of the STI system to implement policy measures in an evidence-based manner. However, traditional innovation indicators from questionnaire-based surveys or patent data struggle to provide the full picture (1–3). (4) identified shortcomings of traditional innovation indicators concerning their coverage, granularity, timeliness, and cost. The authors proposed to use firm websites as a source of firm-level innovation indicators, leveraging the fact that almost all relevant firms have websites nowadays. Websites are used as platforms to provide information on a firm's products and services, achievements, strategies, and relationships. All these aspects may be related to innovations developed by the firm. Innovation, in this context, is defined as the introduction of a new or significantly improved product or process (5). Most of the information on websites is codified as text and extracting innovation-related information from these web texts and transferring this information into a reliable firm-level innovation indicator is the aim of this study.

Text mining algorithms can be used to extract knowledge from large document collections and turn them into valuable economic information (6–10). As a result, text mining became one of the most promising approaches in economic analysis to provide novel tools and insights to economists. At the methodological level, great progress has been made in natural language processing (NLP), driven by the rapid increase in computational power and the availability of large text corpora (11). Especially artificial neural networks have shown very promising results when used for the classification of text documents into certain categories (12, 13).

In this study, we use information from the Mannheim Innovation Panel (MIP), a questionnaire-based innovation survey of firms, to label the websites of surveyed firms as associated to either a product innovator firm or a non-innovator. This labelled data set is then used to train a deep neural network to predict the probability of firms to be product innovators based solely on their website text. Figure 1 outlines our proposed approach. The predicted product innovator probabilities can be interpreted as a continuous firm-level indicator of innovation.

We assess our proposed approach using the following two research questions:

- **Research Question 1**: Can deep neural networks be used to reliably identify product innovator firms solely based on their website texts?
- **Research Question 2**: Are the resulting firm-level, regional, and sectoral patterns from such a prediction model similar to the patterns observed from established innovation indicators when the model is applied to a large out-of-sample dataset of firm website texts?

The remainder of this paper is structured as follows. First, we present our data, followed by our methods. Our results

David Lenz and Jan Kinne designed the study, gathered, pre-processed, analyzed and visualized the data. Jan Kinne and David Lenz wrote the paper.

The authors declare no conflict of interest. The authors would like to thank the German Federal Ministry of Education and Research for providing funding for the research project (TOBI - Text Data Based Output Indicators as Base of a New Innovation Metric; funding ID: 16IFI001) of which this study is a part.

[1]Jan Kinne and David Lenz contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: jan.kinne@zew.de

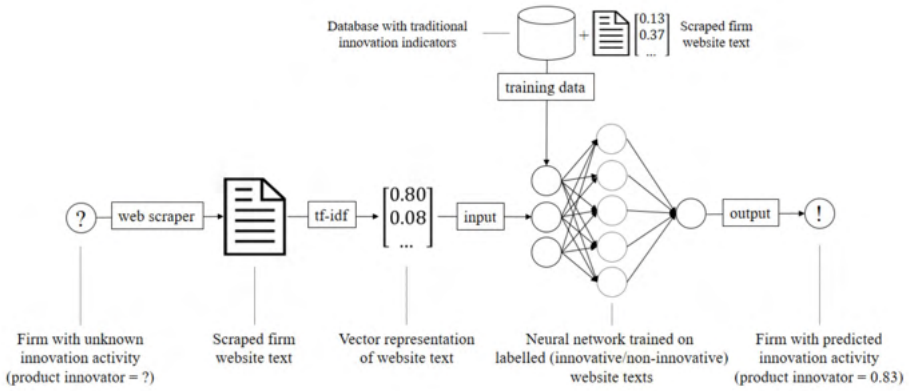


Fig. 1. Proposed product innovator prediction framework. Web scraped texts from firms with a known innovation status are used to train a artificial neural network. The trained model can then be used to predict the innovation status of out-of-sample (unlabelled) firms based on their website texts.

section is twofold. In the first part, we present the results of our artificial neural network classification model. In the second part, we apply our classification model to a large dataset of German firm websites and assess the resulting firm-level, sectoral, and regional patterns. The results are discussed and summarized in the last two sections.

Data

In the following section we present our base datasets, the Mannheim Enterprise Panel (MUP), a firm database containing all economically active firms in Germany, and its derivative, the Mannheim Innovation Panel (MIP), which is an annual innovation survey and the German contribution to the European-wide Community Innovation Survey (CIS). We also present how we obtained texts from the websites of firms in both the MUP and MIP using ARGUS web scraper (14).

MUP firm dataset. The Mannheim Enterprise Panel (MUP) is a panel database which covers the total population of firms located in Germany. It contains more than three million firms which are updated on a semi-annual basis. The data also includes firm characteristics such as the industry (NACE codes; a classification of economic activities in the European Union), postal addresses, number of employees, as well as the website address (URL) of the firms (see (15).

For our analysis, we use MUP data restricted to firms that were definitely economically active in early 2018. The resulting dataset contains 2.52 million firms and 1.15 million URLs (*URL coverage* of 46%). A prior analysis of this dataset by (4) showed that URL coverage differs systematically with firm characteristics. Only a fraction of very young (younger than two years) and very small firms (fewer than five employees) are covered by a URL after controlling for the search quality of the data provider. Sectoral and regional differences can be expected as well. Some regions (especially those with low broadband Internet availability and a low population density) and some sectors, like agriculture, exhibit lower URL coverage. However, given that most of the innovation activity is conducted by middle-sized and larger firms (16), which are well covered in the dataset, the dataset can be assumed to be suitable to analyze the German innovation system.

MIP innovation survey data. The Mannheim Innovation Panel (MIP) is an annual questionnaire-based innovation survey of firms sampled from the MUP database. The survey is designed as a panel survey, such that firms in the sample are surveyed every year. Firm closures and mergers are substituted by randomly sampled additional firms every two years. The MIP is the German contribution to the Community Innovation Survey (CIS), which is conducted every two years in the European Union. CIS data has been used as base data in an array of studies (17). The survey methodology and definition of innovation follows the "Oslo Manual" (5) and covers firms with five or more employees in manufacturing and business-oriented services. Each year, firms are asked whether they introduced new or significantly improved products ("product innovations") during the three years prior to the survey. The firms receive detailed and sector-specific descriptions with examples of product innovations to help them filling the survey. In our study, we use the firms' status as product innovators (yes/no) as the binary target variable.

Each MIP survey, following the CIS guidelines, covers a three-year reference period (the three years preceding the year the survey is conducted). The most recent survey available to us is 2017 which covers the years 2014, 2015, and 2016. Given that we use web texts scraped from the firms' websites in 2018 matched to the 2017 MIP survey, some firms may have changed their innovation status during this time lag. A relevant fraction of firms actually change their status as product innovators between years, such that they may be an innovator in 2015 but not in 2016 (18). We decided to cope with this issue by restricting our analysis to firms that had a "stable" innovation status in the surveys of 2015, 2016, and 2017 (covering the years 2012 to 2016). This means they replied to be product innovators in either all of these years

or none of these years. It can be assumed that such firms are less likely to have changed their innovation status between the survey of 2017 and our test day in 2018. This approach reduces our sample of MIP 2017 firms from 18,062 to 4,481.

The sampling procedure of the survey (oversampling of industries and size classes where innovation in more prevalent) results in an over-representation of innovative firms in the MIP. As a result 32% of the firms in our database are product innovators. Projected to the overall firm population in Germany, the share of product innovators can be expected to be in the range of 27% for the target population in the MIP (19)). This oversampling of innovative firms and a restriction to firms with at least five employees results in a dataset, in which firms are larger, in terms of their number of employees, than firms in the overall firm population in Germany. The mean number of employees in our final MIP sample is 277.5 and the median is 23. Very young firms cannot be included in our sample, as firms have to have taken part in the survey at least three times. Nevertheless, we are able to show that the restriction to firms with a "stable" innovation status results in a higher quality of our training data (see Appendix).

ARGUS website texts. We used ARGUS (4), a free web scraping tool based on the Scrapy Python framework, to scrape texts from the websites of all firms in both our MUP and MIP datasets. We used ARGUS simple language selection heuristic (set to German), which was shown to help limiting the downloaded texts to a certain language (4). According to (4), about 90% of the webpages downloaded this way can be expected to be in German, with some sectors, like the pharmaceuticals and mechanical engineering sector, exhibiting higher shares of non-German webpages (most of them in English).

Webpages are not downloaded randomly from the firms' websites. Instead, ARGUS starts at a firm's main page ("homepage") and then continues downloading those sub-webpages with the shortest URL. The rationale is that more general information on the firm is available at the top level of its website (e.g. "firm-name.com/products", "firm-name.com/team"), which should be given priority over more specific information (e.g. "firm-name.com/news/2017/august/most_read"). This top-level approach is intended to capture texts that represent firm-level business activity profiles instead of specific product-level descriptions which are found on low-level webpages. Even though the latter may inform about individual product innovations, we assume that a top-level business profile description may allow our prediction model to learn combinations of more general signal words that reliably predict product innovator firms, regardless of what exact product innovations they implemented. Such a generalization would be desirable, especially as our training data consists of firms from both ends (consistently innovative or consistently non-innovative) of the overall firm distribution.

The number of downloaded webpages per firm website is defined by a limit parameter in ARGUS, which was set to 25 for this analysis. Hereby, we follow the recommendations of (4) who found that 50% of all firm websites can be scraped completely when this limit parameter is set to 15. Reaching 90% would require raising this threshold to 250, highlighting that web-based studies have to deal with outlier websites, as some firms (especially large ones) have massive websites with ten-thousands of webpages. Following the suggested practice by (4), we excluded websites which redirect to a different domain when requesting their first webpage (i.e. homepage). A practice which should ensure that crawled websites belong to the corresponding firms. This was also shown to not result in any sectoral or firm age selection bias. Table 1 presents our data after excluding such initial redirects and download errors caused by non-existing websites.

During web scraping, all texts found on a firm's website are downloaded, regardless of their content and relevance to the study. To filter out unwanted (*bloat*) sub-webpages, we applied an intermediate filtering step which is described in the Appendix.

Methodology

In this section, we present how the website texts were preprocessed and transferred to term frequency–inverse document frequency (tf-idf) vectors. Tf-idf represents documents as a fixed size vector by counting words in each document (term frequency) and weighting each frequency by the inverse of the term's overall document frequency. We then describe the architecture of our deep neural network model for binary text classification, and how we evaluate the model's classification performance.

Web text preprocessing. We reduced the preprocessing of our texts to a minimum. The scraped web texts were standardized to lowercase and all characters not in the German alphabet were removed (keeping *Umlaut* special characters, whitespaces, and ampersands). Tests with word stemming procedures, which reduce words to their stem (e.g. "innovation" and "innovator" to "innovat"), did not increase our classification performance and we refrained from using it.

Web texts as numerical tf-idf vectors. We used the term frequency–inverse document frequency (tf-idf) scheme to represent the website texts as sparse vectors (see e.g. (20)). The tf-idf algorithm transfers each document to a fixed size sparse vector of size V, where V is the size of a dictionary composed of all words found in the overall text corpus. We restricted our dictionary to words with a minimum document frequency of 1.5% and a maximum document frequency of 65% (*popularity based filtering*), resulting in a dictionary size V of 6,144 words. Each entry in the tf-idf vector of a document corresponds to one word in the dictionary, representing the relative importance of this word in the document (i.e. the website). Words that do not appear in a given document are represented by a 0 value. Specifically, in a first step (the tf step) the number of appearances per word in a single document are counted. In a second step, the inverse document frequency (idf) is used as a weighting scheme to adjust the tf counts. Conceptually, the idf weights determine how much information is provided by a specific word by means of how frequently a word appears in the overall document collection. The intuition is that very frequent words that appear in a lot of documents, should be given less weight compared to less frequent words, as infrequent words are more useful as a distinguishing feature.

Web text classification with a deep neural network. Deep neural networks showed remarkable success when applied as text classification models (12, 13). Different deep neural network architectures were proposed and showed varying performance in different NLP tasks. We tried several neural network architectures (convolutional neural networks, recurrent neural

Table 1. Firms in datasets after filtering steps.

dataset	base data	unstable innovation	no URL	errors/redirects	final data
MUP	2,523,231	N/A	-1,374,383	-463,791	685,057
MIP	18,062	-13,587	-456	-893	3,126

networks, both with long short-term memory and gated recurrent units) and also compared their performance in our specific classification task with more traditional models (naive Bayes classifier, logistic regression, decision trees). In this iterative process, an architecture that could be described as a *undercomplete autoencoder-like neural network* turned out to be the model with the best classification performance. Autoencoder-style neural networks (see e.g. (21)) impose a "bottleneck" (hidden layers with very few neurons) in the network architecture which are intended to force the learning of a highly compressed representation of the network's input. While the output of a standard autoencoder network has the same dimensionality as the network's input, the output of an undercomplete autoencoder network has a smaller dimension than the network's input.

Our final network consists of four hidden layers with intermediary dropout layers, which are intended to improve the network's generalization by ignoring (*dropping*) neurons during the training phase (22). The network's first hidden layer consists of 250 neurons, the following two hidden layers consist of only five neurons each (the "bottleneck"), while the forth and last hidden layer contains 125 neurons. We used *scaled exponential linear units* (SELU, (23)) as activation functions in the hidden layers. The network's output layer consists of a single neuron with a *sigmoid* activation function, a common approach to obtain an output between 0 and 1 from a neural network in binary classification tasks (see e.g. (20)). We used the common Adam optimizing algorithm (24) for the stochastic optimization of the network weights.

Results

In this section, we use the dataset of MIP firms with a surveyed innovation status to train our product innovator classification model and test the model's performance using a retained part of the training data (the test set). In the second part of this section, we apply the innovation prediction model to about 700,000 firms from the MUP to predict whether they are product innovators and then examine the resulting firm-level, sectoral, and regional patterns.

Product innovator prediction model performance. After filtering bloat webpages from our MIP dataset (see Appendix), we aggregated all remaining webpages to the firm level, keeping only the first 5,000 words per firm. As a result, each firm is represented by a single document with a maximum length of 5,000 words from non-bloat webpages only. We randomly selected 75% of the data as training set and retained 25% as test set. Table 2 details precision, recall, f1-score, and support for the resulting classifier applied to the test set (classification threshold for the probabilities of 0.5). If the model classifies a firm as a product innovator, it is correct in roughly 4 out of 5 cases, as it can be seen by the 81% precision for the product innovator class. The model retrieves 64% of all product innovator firms and 91% of all non-innovator firms in the test dataset (recall). The overall f1-score of the model is 80%.

Table 2. Product innovator classification report for test set.

label	precision	recall	f1-score	support
non-innovator	0.81	0.91	0.86	429
product innovator	0.81	0.64	0.71	255
avg / total	0.81	0.81	0.80	684

We also tested the classification model using two alternative configurations: a first model that takes a vector with only the firms' (normalized) age, number of employees, and (*one-hot* encoded) sectors as input and a second model for which we concatenated these firm characteristics vectors and the corresponding tf-idf text vectors. For both alternative models we did not alter the original model's architecture, except for changing the number of neurons in the input layer in order to allow for the input of the new vectors. It turned out that the original text-only model outperforms the firm characteristics-only model by about ten percentage points in overall precision, recall, and f1-score. Most notably, the text-only model showed a twenty percentage points higher recall in the already difficult to detect product innovator class. Furthermore, we found that the alternative text plus firm characteristics model does not show an improved performance over the text-only model.

Out-of-sample product innovator predictions. We used the trained model to predict product innovator probabilities for 685,057 MUP firms. The resulting distribution of product innovator probabilities is shown in Figure 2. The mean probability is 0.253 and the median is 0.203. The lowest predicted probability is 0.029 and the highest is 0.944.

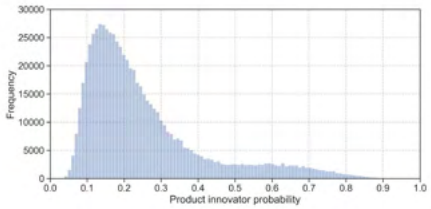


Fig. 2. Product innovator probabilities distribution. Histogram of predicted product innovator probabilities for 685,057 firms.

Comparison to firm-level patent statistics. We obtained firm-level patent statistics (a long-time established indicator for innovation in firms, see e.g. (25, 26)) for 2017 from the European Patent Office. Because patents vary greatly in their importance from sector to sector, we limited our comparison to sectors where at least five percent of the companies are patent holders (mechanical engineering, electronic products, petrochemistry, pharmaceuticals, metal product, other products, and materials). We also excluded patents that were filed

prior to 2006 (ten years is the average lifetime of a patent in our database) to account for the decreasing economic and technological value of aging patents.

The significant Spearman's correlation between our predicted product innovator probabilities and the firms' statuses as patent holders is 0.29. The correlation with the firms' number of patents is 0.25. We also ran two regression analyses to control for sector and size effects in these correlations. The results presented in Table 3 indicate that our predicted product innovator probabilities are strongly related to the probability of a company to hold a patent and having a high patent count. We need to emphasize that this regression analysis is meant to provide a framework for a controlled correlation analysis and the results should not be interpreted as indicating causality.

Comparison to survey extrapolation statistics. To sort the firms into one of the two classes of product innovators and non-innovators (to make them comparable to existing sector and size level survey benchmark data), we are required to set a rather arbitrary product innovator probability classification threshold. Setting this threshold to 0.5 (the same that was used during the model's training) results in 10.31% of the MUP firms being classified as product innovators. Validating this number is difficult as there is no available reference value for the share of product innovator firms in the overall firm population of Germany because MIP survey data can only be used to extrapolate representative shares for that part of the firm population which is covered by the survey (firms with five or more employees in manufacturing and business-oriented sectors) (19). As a result, we can benchmark our prediction results against extrapolated MIP data only in these sectors and size groups, which corresponds to 89,372 of the MUP firms.

For this subgroup, the predicted share of product innovators is 21% (classification threshold of 0.5) while the MIP survey extrapolations indicate a higher share of 27%. This underprediction of product innovators can be related to our model's rather low recall of the product innovator class (see previous section). To adjust for this discrepancy, we decided to calibrated our classification threshold to a value that produces the same number of innovative firms anticipated by the survey extrapolation benchmark of 27%. This calibrated classification threshold of 0.401 was subsequently used to label all 685,057 MUP firms as either product innovators or non-innovators. Naturally, this resulted in an increased share of product innovators in the overall firm population from 10.31% to 15.12%.

In the Appendix, breakdowns by sectors and size classes can be found. Concerning sectors, they show that even though the overall trend and the proportions between sectors are similar to the survey benchmark, underprediction can be seen in all sectors except for wholesale, consulting, and especially ICT firms. Concerning size classes, it can be seen that our predictions match the survey benchmark very well, except for very large firms with more than 1,000 employees where we underestimate the share of product innovators.

Geographic patterns. Map *a* in Figure 3 shows the predicted ratio of product innovator firms to all firms for 432 German districts. It can be seen that city districts exhibit higher shares of product innovator firms. This fact is confirmed by a high and significant Spearman's correlation coefficient of 0.61 between district population density (a proxy for urbanity) and the ratio of product innovator firms. It can also be seen that the vicinities of some major agglomeration areas in the South-West of Germany (Munich, Nuremberg, Stuttgart, Mannheim, and Frankfurt) exhibit high shares of product innovator firms as well. We also find positive and significant correlations between the local share of predicted product innovator firms and the share of the local population working as research and development staff (0.72) as well as with the number of local high-tech patent application per one million inhabitants (0.67). Both these statistics were obtained from the European Statistical Office at the level of NUTS-2 regions for the most recent years available.

The detailed address information in our MUP firm database allows us to disaggregate the geographic pattern as it is shown in map *b* of Figure 3 and to analyze individual regions at a microgeographic level. For the German capital of Berlin, a special survey of the MIP is conducted every year (27), covering a high proportion of firms from manufacturing and business-oriented services in Berlin (the data from this survey was not used for the training of our prediction model). This comprehensive dataset allows us to map the density of our predicted product innovators in Berlin against the observed density from the MIP special survey (map *c* in Figure 3). For map *b* which shows our prediction results, we selected firms that are from sectors and size classes covered in the

Table 3. Firm-level patent statistics regression results.

Dependent variable	Patent holder	Patent count
	Variable of interest	
Product innovator probability	0.265***	35.441***
	Controls and constant	
Sector	Yes	Yes
Employees (log)	0.054***	2.274***
Constant	-5.221***	0.003***
	Model statistics	
Regression model type	Robust logit (average marginal effects)	Robust Poisson (incident rate ratios)
Observations	35,291	35,291
Pseudo R-squared	0.24	0.52
Wald chi2/F-test	4,653.30***	911.10***

survey (4,342 of 35,998 firms in Berlin). Map *c* shows the same pattern for 1,778 firms that answered the innovation survey questionnaire. Both firm location patterns were used to calculate kernel density maps using the same set of parameters. It can be seen that the two densities resemble each other in their overall appearance, with major hotspots in the eastern city center of Berlin (city districts of Mitte, Prenzlauer Berg, and Friedrichshain-Kreuzberg) as well as in the area around Adlershof (a major science and technology park) in the South-East.

Discussion

In this paper, we introduced a deep learning approach to predict product innovator firms based on their website texts. With our deep learning classification model we achieved an overall f1-score of 0.80 in the test dataset, which indicates a very good but by no means perfect prediction performance. While we found that precision is balanced for both classes, the model performed less good concerning the recall of the "product innovator" class. Further performance improvements regarding the latter are therefore desirable and could be achieved with larger training data or an improved model architecture, for example by relying on recent developments concerning pre-trained language models (28). Given our rather small training (2,531 observations) and test sets (684 observations), we are satisfied with the performance of the model. We are especially impressed by the model's ability to decode that much information from a given firm website text, such that adding explicit size, age, and sector information does not further improve the model's prediction performance. We also found that our text-only model outperforms an alternative firm characteristics-only model by more than ten percentage points in overall precision, recall, and f1-score. This is especially distinct for the difficult-to-detect product innovator class, where our text-based model achieved a more than twenty percentage points higher recall.

The classification of 685,057 out-of-sample MUP firms resulted in a product innovator probability distribution where most firms have a probability between 10% and 40%. 10.31% of all firms would be classified as product innovators when using the training step classification threshold of 0.5. As there is no reference value for the overall population of firms in Germany, we compared our prediction results to the MIP survey sampling population for which extrapolated population reference numbers are available (19). In total, 89,372 MUP firms fall into this group of manufacturing and business-oriented service firms with five or more employees. Within this group, a classification threshold of 0.5 would result in a share of 21% predicted product innovators, just short of the surveyed 27%. For the further analysis, we calibrated the classification threshold to 0.401 which results in a share of product innovators that matches the survey benchmark. We also used this threshold to classify MUP firms from sectors and size classes that are not covered in the MIP, given the lack of a better reference values. In the MUP dataset, this shifts the share of predicted product innovators from from 10.31% (0.5 threshold) to 15.12% (0.401 threshold). We continued comparing our prediction results against established innovation indicators at the firm-level (using patent statistics), the regional level (using patent statistics and statistics on R&D personnel), as well as size and sector-levels (using MIP survey extrapolation numbers).

The breakdown by sector showed that the overall sectoral pattern (i.e. the proportions between sectors) is similar to the MIP benchmark. However, our model underestimates the share of product innovator firms in most sectors for which a reference value is available. We attribute this to our model's low recall of the "product innovator" class. Wholesale and ICT services, however, are exceptions, as our model overestimates the share of product innovators in these sectors. Concerning the wholesale sector, we assume that our model is not be able to distinguish between products produced or just sold by a firm. This may lead to a high share of predicted product innovators in the wholesale sector because some assumed product innovator firms are actually just presenting and selling products of other firms. One possible explanation for the overpredicted share of product innovators in the ICT service sector is that the "tech" sector is nowadays widely considered the sector with the most innovative and future-oriented technologies (with buzz words like Digitalization, Industry 4.0, Internet of Things, Artificial Intelligence and the like). Firms with an innovative agenda or self-concept may use these technologies (or at least the associated buzz words) and mention them on their websites. This could result in a bias in our classification model such that the artificial neural network learns to over-relate these words to innovativeness. ICT firms, which naturally use such "tech" vocabulary on their websites, may then be classified as product innovators too often. A preliminary analysis concerning the importance of word features using SHAP (29) points in exactly this direction and suggest that tech sector affiliated words (e.g. "software", "data", "cloud") may indeed play an important role during the classification. The share of product innovator firms in sectors for which no survey benchmark is available can be considered to be reasonable and indicate that our model may be used for predictions in sectors for which no training data is available. This assumption is supported by our findings on the regional and firm-level (see below). However, we assume that the retail sector, may suffer from overestimated product innovator shares for the same reasons as the wholesale sector.

In conclusion, we suggest to use the raw (continuous) product innovator probabilities for future studies. If a binary indicator is needed, sector-level classification thresholds should be used to cope with the bias we found to be present in our predictions. Given that survey data is available for some sectors, researchers may want to select different classification threshold for each sector such that the predicted share of product innovators matches the shares from the extrapolated survey. Another approach would be to train separate models for each sector.

The breakdown by firm size revealed an interesting, non-linear relationship between firm size and the predicted product innovator probabilities. Over all sectors, the product innovator probability peaks at 500 to 1,000 employees before decreasing until about 2,000 employees. This effect is even more distinct for the exemplary sectors of ICT services and mechanical engineering. The well-known German *Mittelstand* (the bulk of mid-sized and highly innovative German firms) may be visible here. Compared to the MIP survey benchmark, our model almost perfectly predicts the share of product innovators for all size classes, except for very large firms with 1,000 or more employees, where the predictions are clearly below the MIP benchmark. This issue has to be examined in follow-up studies.

At the level of individual firms, we used patent statistics to

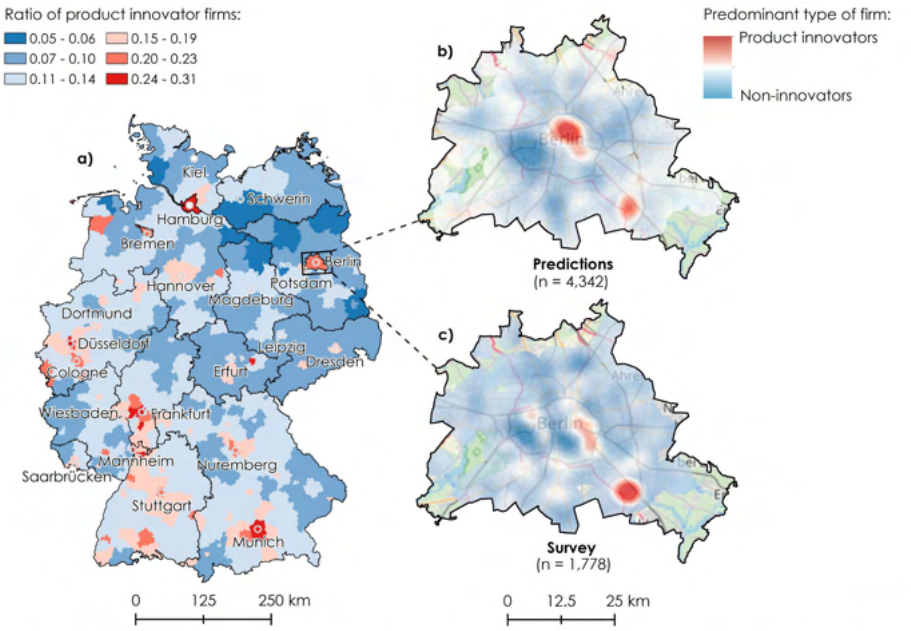


Fig. 3. Geographic pattern of product innovator firm predictions. Predicted share of product innovator firms in local firm population by German districts (a). Microgeographic pattern of product innovator firms in Berlin, predicted (b) and surveyed (c).

evaluate our out-of-sample predictions. Considering the raw correlations, we observed a significant Spearman's correlation between our predicted product innovator probabilities and a firm's status as a patent holder (0.29) as well as the number of patents a firm holds (0.25). These correlations remained even after controlling for potential size and sector effects in a regression analysis. They also compare very well to the correlations between a firm's product innovator status from the MIP survey and its status as a patent holder (0.29) and the number of patents the firm holds (0.30). Consequently, we are confident that our predicted innovation indicator is robust even in an out-of-sample setting. However, it has to be noted that patents and the concept of product innovations capture rather different aspects of the innovation system. Patents do not only tend to detect inventions rather than innovation, they are also used as legal protection for technological progress, for example.

The regional patterns of our predictions, with distinct East-West, North-South, and urban-rural trends, are in line with what was reported in the respective literature (30). We also found very high and significant positive correlations between our predictions and two established innovation indicators available at the regional level (high-tech patent applications and R&D personnel). The microgeographic innovation density maps for Berlin highlighted two further aspects. First, the results of our prediction model compare very well to the benchmark data from the MIP special survey of Berlin. We identified similar product innovator hotspots in both patterns (city center East and the technology park of Adlershof in the South-East). Again, we assume that the model's bias towards the ICT sector may be the cause for a more pronounced innovation hotspot in the eastern city center, an area with exceptionally high shares of firms from this sector (31). Second, our predicted innovation indicator can be used to conduct large-scale analysis of regions in any desired geographical resolution, from individual firm locations to aggregated geographical units. The latter can be considered an important contribution because it allows scientist to analyze innovation policies with unprecedented regional and sectoral granularity.

Conclusion

In this paper, we presented a novel approach on how to predict a highly granular firm-level innovation indicator using deep learning of website texts. We motivated our approach with the need to provide innovation policy making with an innovation indicator that overcomes some of the limitations of traditional indicators from questionnaire-based surveys and patents. Using the website texts of firms included in a traditional innovation survey as training data, we developed an

artificial neural network classification model to predict the product innovator probabilities of firms using only their website texts as input. Our choice of training data, as well as our web text selection and preprocessing procedure were intended to allow the neural network to learn from firms' business activity profiles. Eventually, we did not intend to identify distinct product innovations, but firms with business activity profiles that make product innovations very likely. This likelihood (i.e. probability) can be interpreted as a continuous firm-level innovation indicator. The following two research questions were intended to answer the question on the credibility of this new web-based innovation indicator.

Prediction performance. We concluded that our product innovator classification model achieves a good performance within the test set of firms with a surveyed innovation status. However, we found that the model tends to underpredict the share of product innovators firms, which is reflected in a rather low recall concerning the "product innovator" class. An alternative model that uses only firm characteristics (size, age, and sector) as input, was outperformed by our text-only model by more than ten percentage points f1-score and twenty percentage points recall concerning the difficult-to-detect "product innovator" class. We also found that a combination of web texts and firm characteristics did not improve over a text-only model.

Patterns from out-of-sample prediction. We predicted product innovator probabilities for 685,057 out-of-sample firms and examined the resulting sectoral, size, regional, and firm-level patterns using MIP survey extrapolations, regional innovation indicators, and firm-level patent statistics.

Compared to the survey extrapolations, we found that our model underestimates the overall share of product innovator firms, if a classification threshold of 0.5 is used (we attribute this to our model's low recall of innovative firms). We subsequently adjusted the classification threshold to a value that produces the same number of innovative firms anticipated by the survey extrapolation benchmark.

The resulting sectoral patterns showed two distinct features. First, the overall trend and proportions between sectors followed the trend anticipated from the survey. Second, we identified a positive bias towards ICT firms, resulting in an overestimated share of product innovators in this sector which we discussed thoroughly in the Discussion section. Aggregated to size classes, our predictions almost perfectly matched the surveyed benchmarks. Looking at the relation between firm size and the raw product innovator probabilities, we identified a non-linear relationship that may reflect innovative and mid-sized German *Mittelstand* firms.

We also found high and positive correlations between our predictions and firm-level patent statistics (patent counts and patent holder status) that are also robust in a regression setting which controls for sector and size effects. These correlations also matched the correlations we could observe when comparing patent statistics and the original MIP innovation survey results.

The geographical patterns yielded by our model showed very high correlations to regional high-tech patent applications and regional R&D personnel. Lastly, we showed that the microgeographic prediction patterns match external survey data very well. Overall, we are confident that we created a valuable tool for scientist to analyze innovation at any geographical scale and sectoral level.

Future research. Future research should concentrate on both the methodological development and the application of our approach. Methodologically, it would be interesting to further investigate which words and word combinations have a significant impact on the neural network's prediction outcome. Additional development of the network's architecture and additional training data, as well as a different preprocessing of the training data, could lead to a better prediction performance. Our proposed approach could also be applied to other target variables from surveys in economics (e.g. process innovators and patent holders) or other fields of social science. Empirical follow-up studies could apply our approach to a wide array of research questions, from innovation policy evaluations to the analysis of knowledge spillovers and technology diffusion. Frequent crawling of firm websites may also allow us to build up a panel database of web-based innovation indicators suitable for time-series analysis.

Bibliography

1. S Nagaoka, K Motohashi, A Goto, Patent Statistics as an Innovation Indicator in *Handbook of Economics of Innovation*, eds. BH Hall, N Rosenberg. Vol. 2 edition, pp. 1083–1127 (2010).
2. M Squicciarini, C Criscuolo, Measuring Patent Quality. (2013).
3. OECD, *OECD Patent Statistics Manual*. (OECD, Paris), p. 162 (2009).
4. J Kinne, J Axenbeck, Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany. (2018).
5. OECD, Eurostat, *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation*. (OECD/eurostat, Luxembourg, Paris), 4th edition, p. 258 (2018).
6. A Levenberg, S Pulman, K Moilanen, E Simpson, S Roberts, Predicting Economic Indicators from Web Text Using Sentiment Composition. *Int. J. Comput. Commun. Eng.* **3**, 109–115 (2014).
7. VH Larsen, LA Thorsrud, The Value of News, (Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School), Technical Report 6 (2015).
8. J Lüdering, P Winker, *Forward or Backward Looking ? The Economic Discourse and the Observed Reality*. (MAKGS Joint Discussion Paper Series in Economics), (2016).
9. M Grentzkow, BT Kelly, M Taddy, Text as Data. (2017).
10. S Rönnqvist, P Sarlin, Bank distress in the news: Describing events through deep learning. *Neurocomputing* **264**, 57–70 (2017).
11. J Schmidhuber, Deep learning – An overview. *Neural Networks* **61**, 85–117 (2015).
12. Y Kim, Convolutional Neural Networks for Sentence Classification in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (Association for Computational Linguistics, Doha, Qatar), p. 1746–1751 (2014).
13. Z Yang, et al., Hierarchical Attention Networks for Document Classification. (2016).
14. J Kinne, ARGUS - An Automated Robot for Generic Universal Scraping (2018).
15. J Bersch, S Gottschalk, B Müller, M Niefert, The Mannheim Enterprise Panel (MUP) and firm statistics for Germany. (2014).
16. C Rammer, B Aschhoff, T Doherr, B Peters, T Schmidt, Innovationsverhalten der deutschen Wirtschaft, (Centre for European Economic Research (ZEW), Mannheim), Technical report (2017).
17. F Gault, E Aho, M Alkio, A Arundel, C Bloch, *Handbook of Innovation Indicators and Measurement* ed. F Gault. (Edward Elgar Publishing Ltd, Glos, UK), p. 486 (2013).
18. B Peters, Persistence of innovation: Stylised facts and panel data evidence. *J. Technol. Transf.* **34**, 226–243 (2009).
19. C Rammer, et al., Innovationen in der deutschen Wirtschaft, (ZEW Centre for European Economic Research, Mannheim), Technical report (2019).
20. CD Manning, P Raghavan, H Schutze, *An Introduction to Information Retrieval*. (Cambridge University Press, Cambridge, England), Online edi edition, p. 569 (2009).
21. I Goodfellow, Y Bengio, A Courville, *Deep Learning*. (MIT Press, Cambridge, Massachusetts), (2016).
22. N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
23. G Klambauer, T Unterthiner, A Mayr, S Hochreiter, Self-Normalizing Neural Networks. *Adv. Neural Inf. Process. Syst.* **30**, 99–112 (2017).
24. DP Kingma, JL Ba, ADAM: A Method for Stochastic Optimization in *ICLR Conference Paper*. p. 15 (2015).
25. Z Griliches, Patent statistics as economic indicators: A survey. (1990).
26. ZJ Acs, L Anselin, A Varga, Patents and innovation counts as measures of regional production of new knowledge. *Res. Policy* **31**, 1069–1085 (2002).
27. D Feser, Innovationserhebung Berlin 2017, (Technologiestiftung Berlin, Berlin), Technical report (2018).
28. J Devlin, MW Chang, K Lee, K Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
29. S Lundberg, SI Lee, A Unified Approach to Interpreting Model Predictions. **16**, 426–430 (2017).

30. European Commission, Regional Innovation Scoreboard 2017, (Bruxelles), Technical report (2017).
31. C Rammer, J Kinne, K Blind, Knowledge Proximity and Firm Innovation: A Microgeographic Analysis for Berlin. *Urban Stud.* **forthcoming** (2019).
32. M Danilak, langdetect (2015).

Appendix

Bloat webpage filtering

Training dataset of bloat and non-bloat webpages. During web scraping, all texts found on a firm's website are downloaded, regardless of their content and relevance to the study. Alongside valuable web texts describing the firm itself, products, employees, employed technologies, and the like, web texts from imprints, legal information, and HTTP cookie pop-ups are downloaded as well. We also face the problem that our textual data is highly ambiguous in the sense that many websites share common features, e.g. login pages or contact and legal sections. To filter webpages which contain text of mostly unwanted nature (*bloat webpages*), we created a dataset of webpages labeled as either no-bloat (containing unwanted information) or gold (containing relevant information) which we used to train a bloat/no-bloat classification model.

For this purpose, we sampled 10,000 firms from our MUP base dataset and used ARGUS to scrape their websites with a limit of 100 webpages per website and German as the preferred language. We then kept only non-empty webpages written in German (as classified by Python's langdetect library; (32)). From this sample, we drew 10,000 webpages of which 8,080 could be unambiguously labeled as either gold or bloat by hand.

Bloat webpage filtering results. Training our classification model with the bloat webpage training data and testing it with a retained part of the bloat webpage data (test set), resulted in a precision, recall, f1-score and support indicated in Table 4. The *precision* score of 0.81 indicates that the trained model is correct in 81% of cases if the predicted label is "bloat" (i.e. in 19 % of cases the prediction is bloat even though the webpage is *no bloat*). Out of all bloat webpages, we identify 48% of webpages correctly (*recall* of 0.48) as being bloat, but fail to detect 52% of bloat webpages. Combining precision and recall by applying a harmonic mean, results in a *f1-score* of 87%. *Support* indicates the respective number of cases. Thus, while having high precision, the recall of the bloat class leaves room for improvements. However, in our case we think it is reasonable to prefer a high precision over high recall, as we primarily want to dismiss webpages that are certainly not relevant.

Table 4. Bloat classification report for test set.

label	precision	recall	f1-score	support
bloat	0.81	0.48	0.61	368
no bloat	0.89	0.97	0.93	1652
avg / total	0.88	0.89	0.87	2020

Based on these findings, we decided to set the threshold of the classification model to 0.9, i.e. we only kept a webpage if the model was highly certain that the webpage is no bloat (probability(no bloat) > 0.9). This filtering step resulted in the exclusion of 309 MIP firms because their websites consisted of bloat webpages only.

Alternative training data

To assess the adequacy of our training data selection approach, we reran the entire procedure of web scraping, text preprocessing, model training and testing using three alternative datasets. First, we used only the product innovator variable from the 2017 survey (the same survey we used to create our main training dataset) instead of creating our "stable innovator" training dataset. Using this significantly larger dataset (11,506 firm websites) resulted in a f1-score of just 68% in the corresponding test set (see Table 5).

Table 5. Product innovator classification report for alternative data test set A.

label	precision	recall	f1-score	support
non-innovative	0.72	0.83	0.77	1,784
innovative	0.62	0.47	0.53	1,093
avg / total	0.68	0.69	0.68	2,877

Second, we used the product innovator variable of the more recent MIP of 2018. The results in Table 6 show that this convergence in time of survey data and web data results in a better f1-score, compared to the results using the same survey variable with a one year longer time lag (see Table 5).

Table 6. Product innovator classification report for alternative data test set B.

label	precision	recall	f1-score	support
non-innovative	0.75	0.88	0.81	1,264
innovative	0.61	0.38	0.47	601
avg / total	0.70	0.72	0.70	1,865

Third, we used an alternative product innovator variable of the MIP 2018 which relates directly to the year of the survey. Instead of asking about product innovations introduced by the firm in the three consecutive years prior to the survey, product innovations in 2018 are surveyed. This survey data, which covers about the same time as the web data scraped in 2018, increases the predictive performance of the model in the corresponding test set even more (f1-score of 0.74; Table 7).

Table 7. Product innovator classification report for alternative data test set C.

label	precision	recall	f1-score	support
non-innovative	0.79	0.95	0.86	751
innovative	0.63	0.26	0.37	258
avg / total	0.75	0.77	0.74	1,009

Share of predicted product innovator firms by sectors and size classes

Figure 4 presents the share of product innovator firms by sector after applying the calibrated classification threshold of 0.401. We also indicate the share of product innovator firms by sector as they are calculated from the MIP questionnaire-based survey (transparent bars) if available for the respective sector. Even though the overall trend and the proportions between sectors are similar to the survey benchmark, underprediction can be seen in all sectors except for wholesale, consulting, and especially ICT firms. For sectors without a survey benchmark,

assessing the results is difficult, but overall these predictions look decent. Very low shares of product innovators in construction and agriculture, for example, and higher shares in management services is what was to be expected.

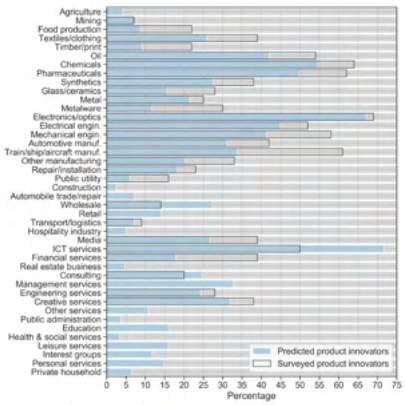


Fig. 4. Predicted product innovator firms by sector. Share of product innovator firms with five or more employees by sector. Blue bars indicate the predicted shares. Transparent bars indicate extrapolated shares from the MIP innovation survey.

Figure 5 shows a breakdown of our predictions by firm size (number of employees). In the left panel, the number of employees is plotted against the predicted product innovator probabilities for all sectors (blue). ICT service (purple) and mechanical engineering firms (green) are plotted as exemplary sectors. It can be seen that the fitted polynomial regression lines of third order indicate a positive non-linear relationship between the size of firms and their product innovator probabilities. The right panel of Figure 5 shows the share of product innovator firms by size groups for sectors covered in the MIP survey (blue) and the corresponding survey extrapolation benchmarks (transparent bars). It can be seen that our predictions match the survey benchmark very well, except for very large firms with more than 1,000 employees where we underestimate the share of product innovators by about 20 percentage points.

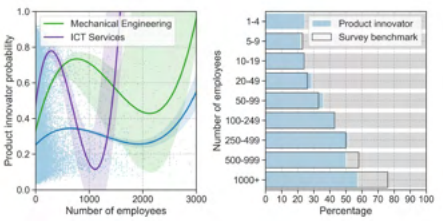


Fig. 5. Predicted product innovator firms by size. Left panel: Product innovator probabilities by firms size (number of employees) with fitted third order polynomial regression line with 95% confidence interval; all sectors (blue), ICT services (purple), mechanical engineering (green). Right panel: Share of product innovator firms by size classes (predictions in blue; survey extrapolations transparent).

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

//
IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de

Appendix E

Paper 5: Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis

Mona Mirtsch, Jan Kinne, and Knut Blind

***Abstract*—In the light of digitalization and recent EU policy initiatives, information is an important asset that organizations of all sizes and from all sectors should secure. However, in order to provide common requirements for the implementation of an information security management system, the internationally well-accepted ISO/IEC 27001 standard has not shown the expected growth rate since its publication more than a decade ago. In this article, we apply web mining to explore the adoption of ISO/IEC 27001 through a series of 2664 out of more than 900 000 German firms from the Mannheim Enterprise Panel dataset that refers to this standard on their websites. As a result, we present a "landscape" of ISO/IEC 27001 in Germany, which shows that firms not only seek certifications themselves but often refer on their websites to partners who are certified instead. Consequently, we estimate a probit model and find that larger and more innovative firms are more likely to be certified to ISO/IEC 27001 and that almost half of all certified firms belong to the information and communications technology (ICT) service sector. Based on our findings, we derive implications for policy makers and management and critically assess the suitability of web mining to explore the adoption of management system standards.**

***Index Terms*—Adoption, information security, management system standards, standards, web mining.**

I. Introduction

IN ADDITION to the advantages of digitalization, the growing connectivity also entails risk with regard to information security [1]–[3]. Security breaches have, therefore, become a global concern, with a value at risk arising from direct and indirect attacks of USD 5.2 trillion between 2019 and 2023 [4]. To achieve information security and reduce the risk of security breaches, organizations must take appropriate measures to protect their information assets and ensure business continuity [5]. The international management system standard ISO/IEC 27001 assists organizations in developing and maintaining an information security management system (ISMS) on the organizational level [6] and "remains one of the most effective risk management tools for fighting off the billions of attacks that occur each year" [1].

After implementing this management system, firms can additionally seek certification to ISO/IEC 27001 to provide confidence to stakeholders that risks are adequately managed [7]. Certification against (preferably international) standards, such as ISO/IEC 27001, is increasingly moving into the focus of policy makers in the light of recent European initiatives. While the Directive on security of network and information systems (NIS-Directive EU 2016/1148) targets operators of essential services in critical infrastructures and digital service providers, the Regulation on information and communications technology (ICT) cybersecurity certification (EU 2019/881 - Cybersecurity Act) sets up a European cybersecurity certification framework for ICT products, ICT services, and ICT processes.

However, apart from the number of valid certificates, which are published in the context of the annual ISO Survey (2018), surprisingly little is known about the adoption of ISO/IEC 27001. According to Castka and Corbett [8], research is often neglected in the early stages of management system standards, probably due to the limited data available. While initial studies often focus on the motives and impacts of adoption, usually based on firm-level data and interviews or surveys, later studies on diffusion often determine diffusion patterns based on macrolevel data [8]. According to Rogers [9], adoption is the decision of an adopting unit (such as firms) "to make full use of an innovation as the best course of action available." Diffusion, on the other hand, being the aggregation of individual (in our case firm) decisions, involves a time aspect and is defined as "the process in which an innovation is communicated through certain channels over time among the members of a social system" [9].

Existing studies on ISO/IEC 27001 analyze the adoption mainly from a theoretical perspective [10]–[12], based on

Manuscript received September 2, 2019; revised December 20, 2019; accepted January 29, 2020. This work was supported in part by the European Commission under Grant Agreement 778420—EURITO and in part by the German Federal Ministry of Education and Research project TOBI under Grant 16IFI001. Review of this manuscript was arranged by Department Editor E. Viardot. *(Corresponding author: Mona Mirtsch.)*

Mona Mirtsch is with the Bundesanstalt für Materialforschung und -prüfung (Federal Institute for Materials Research and-Testing— BAM), 12489 Berlin, Germany, and also with the Technische Universität Berlin, 10587 Berlin, Germany (e-mail: mona.mirtsch@bam.de).

Jan Kinne is with the ZEW—Leibniz Centre for European Economic Research, 68161 Mannheim, Germany, and with the istari.ai UG (haftungsbeschränkt), 68199 Mannheim, Germany, and also with the Department of Geoinformatics—Z_GIS, University of Salzburg, 5020 Salzburg, Austria (e-mail: jan.kinne@zew.de).

Knut Blind is with the Fraunhofer Institute of Systems and Innovation Research, 76139 Karlsruhe, Germany, and also with the Chair of Innovation Economics, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: knut.blind@tu-berlin.de).

Digital Object Identifier 10.1109/TEM.2020.2977815

surveys with the pitfalls of low response rates [13]–[17] or based on case studies [18]. To the best of our knowledge, no studies have empirically investigated the adoption of ISO/IEC 27001 at the national level.

To help fill this gap, the aim of our article is twofold. First, to explore the adoption of ISO/IEC 27001 in Germany, not only by taking into account firms certified to ISO/IEC 27001, but also adopting this standard in other ways. Second, to identify drivers for the certification to ISO/IEC 27001 in Germany. Therefore, we introduce a new method to analyze the adoption of management system standards using web scraping and web mining. Web mining describes the application of data mining techniques to uncover relevant data characteristics and relationships (e.g., data patterns, trends, and correlations) from previously web scraped unstructured web data [19]. We do so by using data from the Mannheim Enterprise Panel (MUP) as the firm database, and then categorize web scraped firms using their website texts and conduct multivariate analyses based on firm characteristics and a deep-learning-based product innovator probability indicator [20].

The remainder of this article is structured as follows. Section II discusses the literature on ISO 9001 and ISO 14001 as well as existing studies on ISO/IEC 27001. Based on the assumption that management system standards are organizational innovations [21]–[23], we present the Technology-Organization-Environment (TOE) framework as an applicable innovation adoption model [24] for firms adopting ISO/IEC 27001. Section III describes the research methodology starting with web mining as a data collection process. Section IV presents the results of the manual categorization of firms that refer to ISO/IEC 27001 on their websites. Using a probit model, we estimate determinants of firm-specific characteristics (firm size, age, innovativeness, and sector affiliation) for the certification to ISO/IEC 27001. In Section V, we discuss our findings and derive a number of managerial implications and recommendations for standards development organizations and policy makers. In our conclusion, we summarize our findings, outline the limitations of our article, and discuss the suitability of web mining to explore the adoption of ISO/IEC 27001 and management system standards in general, including the need for further research.

II. Literature Background

A. Literature Review on the Adoption of Management System Standards

Management system standards, also referred to as meta standards [25], "help organizations improve their performance by specifying repeatable steps that organizations consciously implement to achieve their goals and objectives [...]" [26]. Thereby, organizations can decide whether to implement a management system standard or additionally seek certification through the attestation by an independent third party, also sometimes referred to as registration [27].

Certificates can help organizations signal attributes [28], [29], and hence decrease information asymmetries, one aspect of market failures according to Akerlof [30]. As shown by Terlaak and King [31], the certification to management system standards, such as ISO 9001, is particularly beneficial when there is a high information asymmetry between producers and buyers.

As highlighted by Castka and Corbett [8], in their review of the adoption and diffusion of management system standards (focusing on ISO 9001 and ISO 14001), many studies emphasize on who adopts a standard, why, how and when. The decision to adopt a management system standard is driven by internal or external reasons [8]. The benefits of certification include regulatory compliance [32], meeting customer requirements [33], internal improvements [34], [35], access to markets [36], and innovation performance [37]. Although the motives for seeking certification to ISO 9001 and ISO 14001 are quite similar, the adoption of the latter is often determined by the regulatory environment [38].

DiMaggio and Powell [39] argued that firms are driven by coercive, mimetic, and normative isomorphism, which make organizations similar over time. The desire to improve performance drives the first movers, whereas the second movers are more driven to improve their image [40]. Therefore, according to Naveh *et al.* [40], first movers benefit more from implementing a managerial practice, such as ISO 9001, from their own experience, whereas second movers can benefit by learning from the experiences of others. In this context, the later adoption can be explained by the "bandwagon effect," where previous adopters either reveal information about the value of the adoption or increase the value of the adoption and thereby set off bandwagons [41].

In the case of ISO 14001, Delmas and Montes-Sancho [42] noted that mandatory forces (e.g., derived from regulation) dominate in the early adoption phase, whereas normative pressures and trade-related aspects are more prevalent in the later phase. This effect is evidenced by Arimura *et al.* [43] in relation to ISO 14001, who also recommended government assistance programs to encourage the adoption of ISO 14001 for addressing public objectives.

The motivation to seek certification may also depend on the sector in which the firm operates. Singh *et al.* [44] found that manufacturers are more likely to focus on developing export potential and reducing costs, whereas service providers tend to meet external expectations, such as from customers or government agencies. In addition, internationally active firms are more likely to adopt standards and be certified [45], especially when export markets are affected by EU regulations [33].

However, the adoption of a management system standard and particularly seeking certification is time consuming and costly, especially regarding the costs for external auditors [46]. These costs involve the setting up of a management system, the involvement of consultants, and, in the case of additional certification, the cost of external auditing [47]. These costs vary by firm size and sector, ranging from $10 000 to $200 000 for ISO 14001 [48]. In terms of time invested, the average duration of certification to ISO 9001 is 12 months [44]. Since these investments could outweigh the benefits [49], firms might adopt a management system standard but not seek a third-party attestation (certification).

Once firms have already invested in the adoption of a standard, this can also change their decision-making process when adopting an additional standard [50]. Therefore, a firm's experience

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

in implementing a management system standard could spur the implementation of another management system standard [32], [51], [52]. However, the implementation of a previous management system standard could also hinder the adoption of another management system standard, if it is not fully complementary to the previously adopted standard [32]. Tuczek *et al.* [52], who also referred to Castka and Corbett [8], pointed out that this "coupling effect" is not sufficiently investigated in the context of the adoption of standards.

Firms are increasingly making use of integrated management systems that cover the aspects of quality (ISO 9001), environment (ISO 14001), energy (ISO 50001), occupational health and safety (OHSAS 18001 or ISO 45001), and, also, information security (ISO/IEC 27001) [53]. The aim of integrating compatible management system standards is to reduce administrative burden [54] and costs, e.g., when combined audits and multiple certifications can be obtained. Furthermore, organizations can use the meta structuring of standards similar to the structuring of technologies as a way to deal with the multiplicity of standards, as Gey and Fried [55] showed in the case of a software company.

Previous studies have investigated the adoption of international standards, e.g., by counting valid certificates. However, little attention has been paid to the various forms of adoption (i.e., implementation versus certification) [56] and to the actors and activities to promote the diffusion of organizational standards, which Stamm [57] has recently termed as *diffusion work*. By introducing four modes of standard diffusion along the dimensions direct/indirect and explicit/implicit, namely concrete diffusion (I), broad diffusion (II), selective diffusion (III), and ideational diffusion (IV), Stamm [57] emphasized on the role of consultants to connect activities of standards developing organizations, governments, business associations, and academics. The analysis of this *diffusion work* is particularly suitable for earlier stages, in which the mimetic behavior is not largely evident [57], and from the perspective of the policy stage, since the adoption of the standard does not necessarily immediately follow the creation of the standard.

B. Literature Review on ISO/IEC 27001

Spurred by the success of ISO 9001 and ISO 14001, ISO/IEC 27001 was initially published at the end of 2005 by the International Organization of Standardization (ISO) together with the International Electrotechnical Commission (IEC) and technically revised with the second edition of ISO/IEC 27001:2013. This standard was reviewed and confirmed in 2019, and hence this version remains current.

The underlying ISO/IEC 27000 series is based on the British Code of Practice BS 7799 (see Disterer [6] for the development of this standard), which currently comprises over 40 international standards, including information security controls (ISO/IEC 27002), cloud security (ISO/IEC 27017 and ISO/IEC 27018), and investigation of incidents (ISO/IEC 27043) (ISO, 2019). As the best-known standard within this family, ISO/IEC 27001 [1] "provide[s] requirements for establishing, implementing, maintaining, and continually improving an information security management system" [7]. Within the ISO/IEC 27000 series, information security is

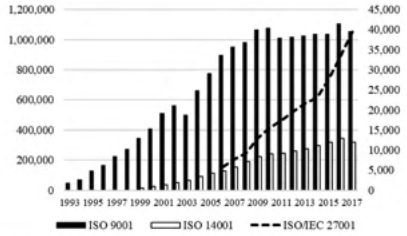


Fig. 1. Evolution of ISO 9001, ISO 14001, and ISO/IEC 27001 over time in terms of valid certificates worldwide. Source: [60].

defined as "preservation of confidentiality [...], integrity [...] and availability [...] of information" [58].

Information security, therefore, differs from concepts such as ICT security (limited to information stored or transmitted using ICT) and cybersecurity (extending information security by including noninformation-based assets), although these terms are often used interchangeably (though indeed overlap—see [59] for details).

Fig. 1 shows the diffusion of the three common management system standards with ISO 9001 and ISO 14001 (bars with the left y-axis) and ISO/IEC 27001 (dashed lines with the right y-axis) from the year in which they became certifiable or corresponding data from the ISO survey [60] are available.

Looking at the number of valid certificates according to the annual ISO survey, ISO/IEC 27001 has shown high growth rates in recent years (e.g., +19% in 2017), but still remains on a comparatively low absolute level (with less than 40 000 valid certificates at the end of 2017), especially compared to other common management system standards, such as ISO 9001 with more than one million valid certificates and ISO 14001 with roughly 360 000 valid certificates in 2017 [60]. This also applies to these management system standards in the early years, when more than 660 000 certificates for ISO 9001 and almost 240 000 certificates for ISO 14001 were valid a decade after their publication [61]. Furthermore, digitalization has been expected to spur the adoption of ISO/IEC 27001. Since firms increasingly store their information based on ICT and governments and suppliers more and more require firms to ensure information security, it has been expected that ISO/IEC 27001 would also be adopted apart from the IT sector [10]. These aspects led to expectations for a higher adoption rate of ISO/IEC 27001 globally [11].

Therefore, previous studies on ISO/IEC 27001 often focused on the reasons for the (low) adoption of ISO/IEC 27001 by firms, alongside the impact of this management system standard as well as the means to increase adoption [10], [11]. Based on case studies in the U.K. and in the Netherlands, Van Wessel and de Vries [18] found that firms adopt ISO/IEC 27001 and ISO/IEC 27002 both for internal reasons (quality enhancement, cost reductions, and increasing the company's risk profile) and for external reasons (meeting legal or customer requirements and improving image). However, firms, especially small and medium-sized enterprises (SMEs) [12], often do not implement information security standards due to high costs and the lack of evidence that the benefits outweigh the costs [62].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Existing studies show that the adoption of ISO/IEC 27001 or other ISMS standards neither leads to less frequent or severe security breaches nor to positive economic impacts through certification against ISO/IEC 27001 [11], [63]–[65]. Therefore, the motives for adopting this standard differ significantly from those for adopting other management system standards such as ISO 9001, the positive economic impact of which has been demonstrated in several studies [8]. However, Barlette and Fomin [11] point out that it is difficult to quantify the benefits of the adoption since ISO/IEC 27001 can be considered as a means to avoid potential losses rather than gaining immediate profits. As a specific positive economic effect, the implementation of ISO/IEC 27001 might result in lower insurance premiums [5].

Other possible reasons for the low adoption include the consideration of competing ISMS standards [11] and the fact that firms outsource their "information-related business" to other countries, e.g., the Far East [10]. However, Fomin *et al.* [10] found no statistical evidence for the latter, as the number of valid certificates in India, for example, was no higher than in the U.K., which is still the case [60]. Fomin *et al.* [10] also concluded (inter alia) that it is worth investigating the need perceived by firms to seek certification instead of just adopting this standard.

Benslimane *et al.* [66] examined the role of certification of IT personnel and ISMS standards, such as ISO/IEC 27001. Looking at online job postings, they found that organizations value work experience and personnel certifications related to IT security more than knowledge of IT security standards. These findings indicate that firms can implement ISMS requirements [66] without fully complying with or being certified to the management system standard.

A limited number of studies conducted surveys investigating motives, obstacles, and impact of ISO/IEC 27001 [14]–[17]. However, the number of respondents were comparably low ranging from 4 and 20 firms per survey also due to the limited number of valid certificates in countries such as Finland, Saudi Arabia, and Bosnia and Herzegovina, where the surveys were conducted. A recent study among Portuguese firms (with 25 participating companies) showed that more than half of these certified firms belong to the IT sector [67]. As regards the implementation and certification process, it took between 6 and 12 months for the firms to obtain ISO/IEC 27001 certification, which in most cases cost more than € 50 000 (including costs for personnel, technical equipment, and external consultancy [67]).

In order to increase the adoption of ISO/IEC 27001, most scholars place focus on the legal environment [10], [68]. From an institutional perspective, governmental intervention may be necessary, as a standard requires a certain adoption rate that triggers further adoption across other organizations, i.e., the bandwagon effect, which is not (yet) evident for ISO/IEC 27001 [68].

C. Theoretical Framework to Analyze Drivers for Certification to ISO/IEC 27001

The Schumpeterian definition of innovation [69] already goes beyond the narrow focus on technical innovations. One type of innovation is organizational innovation such as the implementation of management system standards as intraorganizational procedural innovation according to Armbruster *et al.* [22]. This approach is supported by Hashem and Tann [23] who stated that the introduction of ISO 9001 is an innovation and applied the TOE framework of Tornatzky *et al.* [24] to investigate key determinants of the adoption of the ISO 9000 standard series of Egyptian manufacturers [23].

Fig. 2. Conceptual model based on [23] and [24].

The TOE framework describes how the adoption of innovations is influenced by three aspects in the context of firms. It comprises the following.

1) The Technological context, which includes both internal and external technologies relevant to the firm.
2) The Organizational context, which features firm-specific factors, such as scope, size, and the managerial structure.
3) The Environmental context, which comprises surrounding factors, such as industry, competitors, and governmental influence.

According to Oliveira and Martins [70], the TOE framework has already been used to empirically validate factors that influence the adoption, such as electronic data interchange (EDI) [71], radio frequency identification (RFID) [72], and enterprise resource planning (ERP) systems [73].

For our article, we therefore examine the influence of selected factors on the adoption of ISO/IEC 27001 on firm level, as shown in our conceptual model in Fig. 2 based on the TOE model. As the depth or quality of implementation of management system standards may vary [8], [74], we focus on firms that have implemented this ISO/IEC 27001 standard and additionally received a certificate. We consider this as an indicator of making full use of ISO/IEC 27001.

We have *chosen firm size, firm age,* and *firm innovativeness* as organizational factors, as these factors were identified in previous studies as relevant factors for the analysis of the certification to management system standards [8], [23], [33], [43], [75], [76] or IS innovation adoption on firm level in general [70], [77].

In the technological context, "current practices" can determine the adoption of innovations [70], especially in terms of their compatibility with the new practice [77]. We, therefore, consider *certified to other management system standards* a "current practice" since certification to one management system standard is often linked to the certification to other management system standards [32], [51].

Taking into account that ISO/IEC 27001 is strongly associated with the IT sector [60], [61], we selected the *sector* as an external environmental factor for our study.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

III. Methodology

A. Web Mining for Innovation Indicators

Web mining based on previously web scraped websites has proven itself to be applicable in many research areas [78], [79]. In economic research, firm websites are a particularly interesting area of the World Wide Web. Firms use their websites to present themselves as well as their products and services. The information found on these websites can be used to assess firms' products, services, credibility, achievements, key personnel decisions, strategies, and relationships with other firms [80]. Surveying firms through their websites, rather than conducting interviews, questionnaires, or using other traditional methods, offer clear advantages (coverage, granularity, cost, and timeliness), but it is also associated with its own challenges (data collection, harmonization, and data quality) [19].

There are only a few existing studies that analyze the usability of web-based innovation indicators. These studies either use web content mining or web structure mining [81]. The latter is the analysis of connections between entities (e.g., firms) via the hyperlink structure of websites. Katz and Cothey [82] used this approach in a case study on European and Canadian education institutions. They find that their method is suitable for measuring the degree of recognition of a nation's or province's web presence they receive from other nations and provinces. The authors emphasize the importance of reproducible and accurate indicators capable of dealing with the constantly changing properties of the Internet.

In web content analyses, texts and other website contents are analyzed. This approach is taken by the following studies: Youtie *et al.* [83] used web mining to explore the transitions from discovery to commercialization of 30 nanotechnology SMEs. Arora *et al.* [84] used a similar approach to analyze entry strategies of SMEs commercializing emerging graphene technologies. Both study approaches are capable of identifying different innovation stages. Applying a keyword technique to explore the R&D activities of 296 UK-based enterprises, Gök *et al.* [80] found that web-based indicators provide additional insights compared to patent and literature-based innovation indicators. In addition, they emphasize that web mining has another advantage as a research method. The act of surveying a subject using web scraping and web mining does not cause particular problems, such as altering the behavior of the study object in response to being studied. The authors conclude "[...] that web mining is a significant and useful complement to current methods, as well as offering novel insights not easily obtained from other unobtrusive sources" [80]. However, they raise the criticism that obtaining information from website data is more difficult and that caution is required when generating web-based indicators. Information on websites is generally more related to innovation output than to input. In addition, websites are self-reported, and firms do not publish any new information on their websites at equal frequencies. Beaudry *et al.* [85] used a keyword technique to generate innovation indicators of Canadian aeronautic, space, and defense as well as nanotechnology-related firms based on the text on their websites. They found a significant correlation between their web-based and traditional innovation indicators.

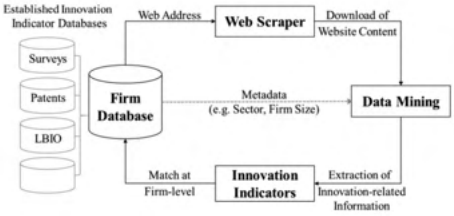


Fig. 3. General analysis framework for generating web-based innovation indicators. Source: [19].

Nathan and Rosso [86] combined the UK administrative microdata, media, and website content to develop experimental measures for innovation in SMEs. The authors used proprietary data gathered by a data firm that uses website and media content to model lifecycle events of firms such as new product and service launches. They were able to identify three times more product/service launches than patent applications from SMEs. Nathan and Rosso [86] concluded that web-based indicators are a useful complementary measure to existing metrics as they reveal additional information. Moreover, they found that previous patent activities are related to a firm's current launch activities and that tech SMEs are much more likely to launch new products or services than nontech SMEs. Studies on web-based innovation indicators have thus confirmed that firm websites are an interesting and rich data source for examining the innovation activity of firms and science, technology, and innovation systems in general.

B. Data Collection and Sample

Kinne and Axenbeck [19] proposed a generally applicable framework for studying firm websites based on established firm databases (see Fig. 3). Starting from the firms' website addresses, a web scraper queries the websites and downloads their content (e.g., texts). In a subsequent data mining step, which can be enriched with available firm metadata (e.g., for data mining model preselection), the so-called innovation-related information is extracted and transferred to firm-level innovation indicators. In the final step, these new innovation indicators are matched back to the firm database at the firm level. This last step also established a link between the new indicators and the traditional ones (e.g., patents) that can be used for validation.

In this article, we apply the web mining approach as described in Fig. 3 to identify and analyze German companies that mention the ISO/IEC 27001 standard on their websites.

Therefore, we use the Mannheim Enterprise Panel (Mannheimer Unternehmenspanel—MUP) from 2019 as a basic dataset. The MUP is based on a firm data pool of Germany's largest credit rating agency (Creditreform e.V.) and, as a panel firm database, comprises all economically active firms located in Germany and the associated metadata (e.g., sector, firm size, and location) [87].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

In the beginning of 2019, the MUP comprised 2 497 412 firms that were definitely economically active at that time and 1 155 867 corresponding website addresses (URLs). With these 1 155 867 URLs, we were able to successfully scrape texts from 912 850 firm websites using the open-source ARGUS web scraping tool [19]. Referring to the findings of Kinne and Axenbeck [19], we downloaded a maximum of 25 webpages per website (the median number of webpages per firm website in Germany is 15). We also used ARGUS' options to download preferably German language webpages and those with shorter URLs. The latter follows the idea that the most general information about a firm can be found on its top-level webpages (e.g., "firm-name.com/about-us"). Based on the results of a comprehensive study performed by Kinne and Axenbeck [19], it can be expected that the coverage of our sample of scraped website texts will differ systematically between sectors and firm types; only a small fraction of very young and very small firms (smaller than five employees and younger than two years) will be included. Sparsely populated regions and certain sectors, such as agriculture, are also less well covered. Medium-sized and larger firms are expected to be almost fully covered, especially in technology-intensive sectors, such as mechanical engineering [19].

The web scraping process described above resulted in approximately 47 GB of raw text data for the 912 850 firms. To identify firms that mention ISO/IEC 27001 on their websites, we used a simple keyword search. Taking into account the possible writing options for the individual management system standard, we have included all combinations of DIN (the German Institute for Standardization), ISO and IEC with 27000 and 27001 and tagged all firm websites with at least one occurrence of at least one of the search string combinations.

C. Methodology to Analyze the Adoption of ISO/IEC 27001 in Germany

The first step of the analysis focused on the number of firms that refer to ISO/IEC 27001 on their websites. In a subsequent step, we categorized the firms according to the reason why they refer to ISO/IEC 27001 on their website, assuming that not all firms are certified, but refer to this management system standard for other reasons. To ensure a correct manual categorization of the firms in this sample, the webpages of these firms were analyzed in detail per firm using predefined codes (e.g., firm is certified, adopts a standard without certification, offers consulting or certification services, and any other reference) and two additional codes derived during the coding process (firms employing certified IT specialists and firms that are not certified themselves but refer to certified business partners). This coding was conducted by three persons and all certified firms were independently validated by another person to ensure consistent results.

D. Methodology to Analyze Driving Factors for ISO/IEC 27001 Certification in Germany

For our following statistical analysis, we use the variables as described in Table I. We rely on the firm data in the MUP, which are available to 50% in terms of firm size, to 94% in terms of firm age, and to 99% in terms of affiliation to the sector of all web scraped firms. Furthermore, a firm-level product innovator probability is available for 82% of all web scraped firms.

TABLE I
DESCRIPTION OF VARIABLES

Model variable	Description
Dependent variable	
Certification to ISO/IEC 27001	1 if a firm obtained a certificate for ISO/IEC 27001, 0 otherwise. Derived from web mining and subsequent manual firm categorization
Independent variables	
Firm Size	Logarithm of number of the firms' employees. Derived from MUP firm data base
Firm Age	Logarithm of years since founding date. Derived from MUP firm data base
Innovation Probability	Probability (0.0 to 1.0) that the firm is a product innovator. Derived from a deep learning model
Sector	Sector affiliation using NACE classification Derived from MUP firm data base [20]

This prediction is based on the firm's website text and a deep learning model trained on the websites of firms surveyed in the German Community Innovation Survey (CIS) (see [20] for more details). In particular, traditional firm-level indicators from a questionnaire-based innovation survey (German CIS) were used to train an artificial neural network classification model on labeled (product innovator/no product innovator) web texts of surveyed firms. Subsequently, this classification model was applied to the web texts of hundreds of thousands of firms in Germany to predict whether they are product innovators or not. The authors compared these predictions to firm-level patent statistics, survey extrapolation benchmark data, and regional innovation indicators. The results showed that this approach produces reliable predictions and has the potential to be a valuable and highly cost-efficient addition to the existing set of innovation indicators, especially due to its coverage and regional granularity [20].

IV. RESULTS

A. Results of the Adoption Analysis of ISO/IEC 27001 in Germany

Out of the 1.15 million web scraped firms, a total of 47 919 firms refer to one of the management system standards, which corresponds to about 4.15% of all scraped firms. Most firms refer to ISO 9001, followed by ISO 14001, ISO 50001, and ISO/IEC 27001. This also corresponds to the ranking of valid ISO certificates published in Germany in 2017 as part of the ISO survey (see Table II).

As a first finding, only in the case of ISO/IEC 27001, the number of firms referring to this standard on their website is larger than the number of valid certificates according to the ISO survey [60]. Since firms can obtain more than one certificate

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
COMPARING CERTIFIED FIRMS OF MUP SAMPLE WITH VALID CERTIFICATES IN GERMANY

	# firms (MUP)	*# certificates (Germany)*	*Relation:*
ISO 9001	35,706	64,658	0.55
ISO 14001	6,789	10,176	0.67
ISO 50001	2,760	8,314	0.33
ISO/IEC 27001	2,664	1,339	2.00

Source: [60].

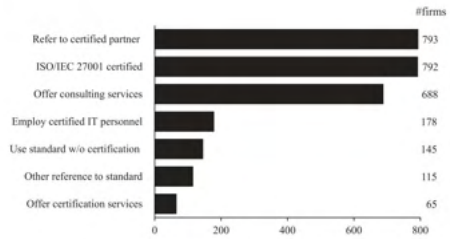


Fig. 4. Firm categorization of 2664 firms referring to ISO/IEC 27001 on their websites.

per management system standard (e.g., for different branches or organizational units within one firm), our comparison can, however, only serve as a rough proxy. Furthermore, firms can refer to the management system standards on their websites for other reasons than being certified.

Fig. 4 shows the results of manually categorizing the reasons why firms refer to ISO/IEC 27001 on their websites. In general, it should be noted that firms can belong to several categories, e.g., a consulting firm offering services in connection with ISO/IEC 27001 can also be certified to ISO/IEC 27001.

In total, 29.7% of the firms refer to ISO/IEC 27001 on their websites because they are ISO/IEC 27001 certified. A relatively small proportion (5.4%) stated that they have adopted a standard, but are not officially certified, although they often claim on their websites to seek certification in the future. Total 6.7% of firms employ certified IT personnel without having obtained a certificate for the firm's ISMS. However, the highest proportion of 29.8% of firms was not certified themselves but referred to a certified partner. Many firms referring to ISO/IEC 27001 offer consultancy (25.8%) or certification services (2.4%) related to ISO/IEC 27001. Overall 4.3% of all firms have referred to ISO/IEC 27001 for other reasons, e.g., to provide news about this management system standard.

For the companies certified to ISO/IEC 27001, we have also investigated the likelihood that firms will be certified to other international management system standards as technological context factor (see Fig. 2). Therefore, we have manually visited their websites and have searched for a different management system certificate. As a finding, a large proportion of firms

TABLE III
OBSERVED CO-OCCURRENCES OF REFERENCES TO MANAGEMENT SYSTEM STANDARDS IN ABSOLUTE AND RELATIVE TERMS

	ISO 9001	*ISO 14001*	*ISO 50001*
ISO/IEC 27001	339 (42%)	111 (14%)	52 (7%)

TABLE IV
SECTOR AFFILIATION OF ISO/IEC 27001 CERTIFIED FIRMS VERSUS NONCERTIFIED MUP FIRMS

	Certified MUP firms		*Non-certified MUP firms*		*Difference*
	Obs.	%	Obs.	%	in % points
ICT services	338	43.00	34,051	3.75	+39.25
Other services	58	7.38	54,990	6.06	+1.32
Consulting	53	6.74	49,634	5.47	+1.27
Financial services	52	6.62	26,352	2.90	+3.72
Management services	44	5.60	21,005	2.31	+3.29
Retail	32	4.07	93,971	10.36	-6.29
Engineering services	31	3.94	37,315	4.11	-0.17
Public utility	28	3.56	7,902	0.87	+2.69
Personal services	24	3.05	34,660	3.82	-0.77
Wholesale	24	3.05	55,917	6.16	-3.11
Transport/logistics	19	2.42	19,828	2.19	+0.23
Creative services	17	2.16	29,485	3.25	-1.09
Leisure services	11	1.40	22,313	2.46	-1.06
Construction	9	1.15	99,233	10.94	-9.79
Electronics/optics	6	0.76	4,654	0.51	+0.25
All other	40	5.10	316,148	34.84	-29.74
Sum	786	100%	907,458	100%	

certified to ISO/IEC 27001 is also certified to ISO 9001, followed by ISO 14001 and ISO 50001, as shown in Table III.

Out of the 792 ISO/IEC 27001 certified firms, 30% are certified to one additional standard, 9% against two further standards, and 5% against all three other management system standards.

B. Results on the Analysis of Driving Factors for ISO/IEC 27001 Certification in Germany

1) Descriptive Statistics: In terms of sector affiliation, almost half (43%) of all ISO/IEC 27001 certified firms offer ICT services, which is significantly higher than approximately 4% of all firms in the MUP data sample offering ICT services (see Table IV). ISO/IEC 27001 certified firms providing consultancy and financial services are also overrepresented as well as public utilities compared to noncertified firms in the MUP database. The results also show that ISO/IEC 27001 certification is not very common in "traditional" sectors, such as construction, retail, or manufacturing.

To differentiate between firms providing ICT services and other firms, we present the following descriptive statistics for all firms (all sectors), and in a second step, we focus just on

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
FIRM CHARACTERISTICS OF ISO/IEC 27001 CERTIFIED FIRMS VERSUS NONCERTIFIED MUP FIRMS

All Sectors		
Mean	*Certified MUP firms*	*Non-certified MUP firms*
Firm Size***	76 (229.27) N=596	23 (483.61) N=458,933
Firm Age***	17 (14.03) N=768	24 (42.44) N=858,902
Innovation probability***	0.57 (0.20) N=774	0.25 (0.16) N=749,580
ICT Service Sector		
Mean	*Certified ICT service firms*	*Non-certified ICT service firms*
Firm Size***	61 (129.11) N=274	15 (78.82) N=16,896
Firm Age***	17 (12.53) N=333	14 (18.33) N=33,050
Innovation probability***	0.62 (0.18) N=331	0.49 (0.21) N=27,266

Notes: Standard deviation in parentheses. N = Number of observations. Significance from the t-test: $^{*}\ p < 0.10$; $^{**}\ p < 0.05$; and $^{***}\ p < 0.01$.

TABLE VI
PROBIT ESTIMATION RESULTS

Independent variables	*Certification to ISO/IEC 27001 MUP total*	*Certification to ISO/IEC 27001 ICT Service Sector*
Firm Size (in logs)	0.179*** (0.00078)	0.238*** (0.00986)
Firm Age (in logs)	-0.057** (-0.00025)	0.011 (0.00046)
Innovation Probability	1.564*** (0.00679)	0.624*** (0.02591)
Sector Dummies	Yes (base: ICT service sector)	
Constant	-3.249***	-2.992***
Observations	345,607	14,333
Model Chi-square	2063.20***	252.43***

Notes: The table displays the coefficients of all observations in the MUP and ICT service sectors and the marginal effects of each in brackets. A correlation matrix of the variables is provided in Table VIII and the probit estimation results for the sector dummies in Table X $^{*}p < 0.10$.$^{**}\ p < 0.05$.$^{***}p < 0.01$.

TABLE VII
SECTOR AFFILIATION OF TÜV RHEINLAND ISO/IEC 27001 CERTIFIED FIRMS

	Observations	Percent
ICT services	123	47.13
Other services	23	8.81
Public utility	17	6.51
Consulting	13	4.98
Management services	12	4.60
Financial services	10	3.83
Health & social services	9	3.45
Electronics/optics	8	3.07
Engineering services	7	2.68
Personal services	6	2.30
All other	33	12.64
Sum	**261**	**100**

the companies that are attributed to ICT services, as they are responsible for almost half of all certifications. In both cases, the results of the descriptive statistics on firm size, firm age, and innovation probability presented in Table V reveal significant differences between the firms certified to ISO/IEC 27001 and noncertified firms.

Taking into account firms of all sectors, first, the certified firms with 76 employees are more than three times as large as the average noncertified firm in the MUP. Second, and in contrast, certified firms aged 17 years are on average seven years younger than the average of noncertified firms. Third, the innovation probability of 57% is twice as high as the average innovation probability of noncertified firms.

Surprisingly, when focusing on firms attributed to ICT services, the average age is the same as for all ISO/IEC 27001 certified companies. Certified ICT service firms are still larger than noncertified ICT service firms with 61 employees compared to 15 employees. Aged 17 years, however, they are also older than noncertified firms in the ICT service sector aged 14 years. After all, firms in the ICT service sector have a product innovation probability of almost 50%, i.e., almost twice the probability of all noncertified firms. However, certified firms in the ICT sector have an even higher product innovation probability with 62%.

Summarizing the findings from the analysis of the descriptive statistics, we can see a positive relationship between firm size and the probability of certification. A positive correlation with firm age can only be observed within the ICT service sector. Furthermore, innovativeness increases the likelihood of certification, while the high proportion of certified firms belonging to the ICT service sector (see Table V) indicates that this sector is strongly linked to certification against ISO/IEC 27001.

2) Probit Model: Finally, we run a probit model. Our probit models test the probability of the event (= certification to ISO/IEC 27001) as a dependent variable and the independent variables as shown in Table I.

The results of our two probit models are shown in Table VI. In the general model, which covers all MUP firms, significant results are shown for all explanatory variables. First, the likelihood to be certified to ISO/IEC 27001 increases significantly with firm size. Second, older firms are significantly less likely to be certified to ISO/IEC 27001. Third, firms with a higher innovation probability are more likely to be certified to ISO/IEC 27001. Finally, firms operating in the ICT service sector are more likely to be ISO/IEC 27001 certified than firms operating in any other sector as shown in Tables VII and X.

Consequently, we run a second probit regression model just for the firms active in the ICT service sector. Here, too, the firm size is significantly positively associated with the likelihood of being certified to ISO/IEC 27001. However, the age of firms in this sector does not significantly explain the likelihood of

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

certification. Finally, firms in the ICT service sector with a higher innovation probability are more likely to be certified to ISO/IEC 27001. In addition, this relationship is stronger than in the sample of all firms based on the marginal effects shown in brackets.

Since only a very small proportion of the firms in the MUP sample are ISO/IEC 27001 certified (less than 0.1%), we encounter the problem of a small sample bias. In our search for rare events, we, therefore, apply the method proposed by King and Zeng [89] and run a corrected logit estimate for our independent variables firm size, firm age, and innovation probability. The corrected logit estimates provided in Table IX confirm the results of our probit models.

C. Validation

To validate our findings and to avoid a single source bias, we relied on another independent dataset. Therefore, we have manually analyzed the ISO/IEC 27001 certified firms of the German certification body TÜV Rheinland, which publishes their valid certification[1]. In this certification database, we have identified 358 valid certificates of 261 German firms that are certified to ISO/IEC 27001.

First, we examined which sector these firms belong to. Second, we analyzed whether these firms publish their certificates on their websites, and if not, whether they publish a logo instead. Third, we analyzed how many certified firms would have been identified using our web scraping.

We found a similar sector breakdown (see Table VII) as our web mining results (see Table IV), which confirms that most ISO/IEC 27001 certified firms offer ICT services, followed by other services. Firms belonging to the public utility sector (e.g., energy providers) rank higher in this sample compared to our web mining sample, but this could also indicate a certain affiliation of this sector to this particular certification body.

Out of the 261 ISO/IEC 27001 certified firms, 39 firms (equaling 15%) did not publish a written reference to an ISO/IEC certification on their websites, one-third of them offering ICT services. Out of these 39 firms, 5 firms displayed a logo instead, representing less than 2% of the 261 firms.

Since our web scraper only searched for the top 25 webpages per firm, our web scraper would have identified 44% of these certified firms that are included in the MUP. This finding shows that the remaining ISO/IEC 27001 certified firms would have only be identified with a higher scraping effort, i.e., more webpages per company. Our manual analysis, furthermore, revealed that especially larger firms do not display their certificates on the top 25 webpages, but at lower level webpages—e.g., on the webpages of specific products or news pages.

V. Discussion

A. Discussion on the Adoption of ISO/IEC 27001 in Germany

The initial finding of our web mining revealed that double the number of firms refer to ISO/IEC 27001 on their websites as valid certificates according to ISO (2018) are available in Germany. Our manual categorization, however, showed that out of the 2664 firms identified, only 792 firms are certified to ISO/IEC 27001, which now represents roughly 60% of all valid certificates. This finding shows that many firms refer to this management system standard in relation to ISO/IEC 27001 for reasons other than being certified. Therefore, the manual categorization of all firm websites in our ISO/IEC 27001 analysis has helped to create a "landscape" of the adoption of ISO/IEC 27001 (see Fig. 5) including a demand side and a supply side to gain a better understanding of the ISO/IEC 27001 adoption in Germany.

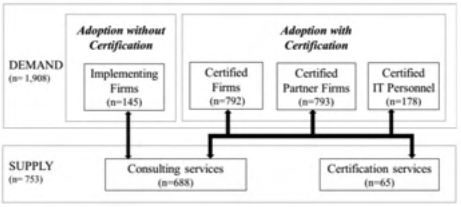


Fig. 5. ISO/IEC 27001 "landscape" of German firms.

On the demand side, the landscape does not only include certified firms, which is often the case with previous studies about management system standards using ISO survey data. Firms can also adopt this management system standard without seeking certification for themselves, which we refer to as implementing firms. The results show a comparatively small number of firms that have not (yet) received a certificate but have only adopted the standard. Referring to a study by Irish managers, which stated that 12% of firms use standards, such as ISO/IEC 27001, but only 2% are certified [90], it could have been expected that more firms had implemented the standard instead of being additionally certified. However, it may not be worthwhile to communicate on the website, if firms have implemented a standard without a formal attestation.

The landscape also shows the important role of IT personnel, as discussed above by Benslimane *et al.* [66], as it also implements security practices in firms according to the ISO/IEC 27001 standard, which can also serve as a signal to stakeholders. For example, IT personnel may have obtained certificates such as Information Security Officer or Auditor according to ISO/IEC 27001 (e.g., [91] as an example).

A key finding of our explorative research is the possibility to refer to partners (such as cloud computing providers or data centers) that are certified. This option shows the main difference between ISO/IEC 27001 and the other management system standards, as it is possible to outsource information security to some extent, which is unlikely for quality, environmental, and energy management. It is, therefore, possible that outsourcing will not take place in the Far East, for example, as discussed by Fomin *et al.* [10], but to IT service providers within Germany or Europe. This could also be spurred by the General Data Protection Regulation (GDPR), which entered into force in May

[1] [Online]. Available: www.certipedia.com

2018. Although ISO/IEC 27001 certification should not be seen as a tool to signal GDPR compliance, ISO/IEC 27001 can help to comply with the GDPR [92]. In order to elaborate this effect of "indirect certification" theoretically, one can apply theories about networking and, in particular, brand leveraging and co-branding, concepts that traditionally originate from marketing and, in particular, consumer research [93]. In our case, firms can be "embedded" in a network and gain reputation and trust by claiming an alliance with a partner who is certified, as shown by Hu *et al.* [94], in the case of technical standard alliances.

The "landscape" (see Fig. 5) also includes the supply side of ISO/IEC 27001, by involving certification bodies and consultants as important actors in the diffusion work [57] of this management system standard. The large number of consultants active in the field of ISO/IEC 27001 and providing knowledge of this standard indicates, first, a need for firms to use consulting firms for the implementation of ISO/IEC 27001. Second, it indicates that firms may implement this standard with the help of consultants rather than to be officially certified. This can also help explain the low adoption of ISO/IEC 27001 in Germany given the low number of valid certificates [60], although on average almost 30% of all German companies claim to have a formally defined ICT security policy that takes into account the confidentiality, integrity, and availability of their data and ICT systems [95].

B. Discussion on Driving Factors for ISO/IEC 27001 Certification in Germany

Our regression analysis revealed that larger and more innovative firms, most of them belonging to the ICT service sector, are more prone to ISO/IEC 27001 certification.

The significant size effect supports the findings of previous studies on other management system standards [33], both for all firms and for ICT service providers. Obviously, certification costs present a problem for smaller companies that may not be compensated by the benefits of achieving certification to ISO/IEC 27001 [12]. Since the firm size is often correlated with firm age [96], we expected a positive effect that is only the case for ICT service firms (see Table V), though not significant (see Table VI). Therefore, different organizational factors and IT skills may lead to differences in the perception of firms in terms of information security and related investments, apart from size, age, and innovativeness, which should be subject to future research.

Our findings have several implications for managers, policy makers, and standard development organizations. From a managerial perspective, it shows that firms can make use of ISO/IEC 27001 either in terms of implementation versus certification (1), the use of certified IT personnel (2), and the reference to a certified partner (indirect certification) (3) without having to bear the time and cost for certification. Therefore, depending on their individual objectives, firms should critically examine whether it is worthwhile to seek certification (e.g., as a competitive advantage or because stakeholders require an independent attestation) or not. In some cases, the implementation of ISO/IEC 27001 might be a good start to increase the overall level of information security, including employee awareness, without bearing the immediate costs for certification.

From a policy perspective, our findings have an impact when policy makers decide to make use of ISO/IEC 27001 to increase the overall level of information security in firms. First, the significant firm size effect may require action. Policy makers could, for example, spur the diffusion of ISO/IEC 27001 among SMEs by providing incentives to firms that seek services, e.g., from consultants, to implement an ISMS according to ISO/IEC 27001. Second, the benefits for smaller firms implementing an ISMS according to ISO/IEC 27001 may not be sufficiently known or measurable for smaller companies. Therefore, standards development organizations could publish practical guidance documents, in particular, to help SMEs apply the ISO/IEC 27000 series, as proposed by the European Commission in its recent rolling plan for ICT standardization [98]. Third, it is worth investigating whether independent third-party certification is required or whether a self-declaration of conformity might be useful to achieve the respective goal. Finally, looking closely at the ISO/IEC 27001 certified firms, they most often belong to the ICT service sector. Hence, the question arises as to whether the concentration of certifications among ICT service firms is sufficient for an overall adequate level of information security because they provide services to companies throughout the entire economy, or whether we have a significant gap here. This might be true, in particular, for manufacturing firms, particularly in view of the increasing connectivity related to Industry 4.0, which may require further actions from policy makers.

VI. Conclusion

For the first time, we used web mining as a data source and method to examine German firms in the MUP database with a website with reference to ISO/IEC 27001 in this article.

A manual categorization of all firms with ISO/IEC 27001 reference on their websites enabled the development of an ISO/IEC 27001 "landscape", as outlined in Fig. 5, covering both the demand side (firms making use of this management system standard) and the supply side of this management system standard (firms providing services related to ISO/IEC 27001).

The implications of our findings can lead to a better understanding of the reasons for the (low) adoption of ISO/IEC 27001. First, the small number of valid certificates reported in the ISO survey is not necessarily due to the low adoption rate of the standard. Firms can also benefit from either implementing the management system standard without seeking certification or by using certified IT personnel. Second, firms make use of certified partners to which they refer on their websites, a phenomenon that we term "indirect certification." These partners (mostly cloud suppliers and data centers), therefore, have a multiplier effect by providing information security to a larger number of firms.

Our web mining based analysis of firms that refer to ISO/IEC 27001 on their websites showed that this method can be used in combination with a manual firm-by-firm evaluation to gain a better understanding of the drivers for certification to ISO/IEC 27001. We have shown that firm size, innovativeness, and affiliation to the ICT service sectors are potential drivers

for ISO/IEC 27001 certification. In particular, smaller firms seek less certification than larger firms, which may call for the need for supporting SMEs in implementing ISO/IEC 27001 and seeking certification.

From a legal perspective, certification against ISO/IEC 27001 is voluntary for firms per se. However, this could change in the near future not only in the light of the NIS-Directive but also of the latest EU Cybersecurity Act. In addition, firms can adopt ISO/IEC 27001 to demonstrate compliance with the principles of technical and organizational measures to protect information for the purpose of the GDPR [99]. Thereby, the results of this article can help to derive more substantial recommendations for the application of this management system standard, e.g., if a mandatory certification for firms in specific sectors or alternative measures to increase the adoption of ISO/IEC 27001 are discussed.

From a methodological perspective, web mining of firm websites supplements the traditional methods of standard adoption research, which are often based on surveys and are qualitative in nature, or in the case of diffusion research based on national macrodata.

However, web mining and this article are not without limitations. As far as the applicability of the method is concerned, our web scraping first covered only the top 25 webpages per website. A previous study showed that the median number of subweb pages per website of German firms is 15, but this number of webpages is also strongly correlated with the size of the firm [19]. This suggests that our rather low per-website scraping limit can induce a bias against larger firms, which we also found in our validation, indicating that German ISO/IEC 27001 certified firms may be even larger than our empirical results suggest. For future web mining studies, we therefore suggest either using a higher scraping limit for all firms or adjusting the scraping limit according to the available firm size information.

Second, our analysis assumes that all firms certified to ISO/IEC 27001 would announce this on their websites. However, firms are not obliged to do so, and some sectors, such as ICT services or electronics, may be more prone to the presentation of their certificates on their websites than other sectors [67]. Therefore, firms active in the health or tourism sector may see a lower value for their goal of publishing their certificates and hence there may be a distortion in certain sectors.

Third, our web mining (by keywords only) cannot distinguish whether firms are certified or otherwise refer to this management system standard. Therefore, only a combination of web mining and manual analysis allowed a suitable categorization. In order to make use of this method to a greater extent, further automation would be needed using a web scraper. This could include the recognition of images to identify certificates, or the use of neural networks to predict whether a firm is certified to a particular management system standard.

Finally, the positive relationship of firm drivers for ISO/IEC 27001 certification does not necessarily imply causality. Further research is needed to examine the drivers and barriers to the adoption of ISO/IEC 27001. As a first step, our categorized firms that are certified to ISO/IEC 27001 or have adopted this standard (without certification) can be used to analyze the context in which firms refer to the use of ISO/IEC 27001 on their website as a motive for adoption and further sector segmentation. This analysis could also be extended to firms that refer to certified partner firms to examine the drivers for this type of "indirect certification". Additional methodological approaches, such as interviews and surveys, are needed to theoretically support these correlations and to identify further drivers and barriers in connection with ISO/IEC 27001 certification. Our identified firms can therefore serve as a sample.

Our approach of defining certifications based on management system standards as organizational innovation itself opens up a new research field to investigate the relationship between product innovation and certifications in the context of international management system standards as organizational innovations [22]. This raises the question of timing, i.e., whether product innovations trigger certification to management system standards as organizational innovations [97] or vice versa. However, this question cannot be answered by the available cross-sectional data but requires time-series data.

APPENDIX

See Tables VIII–Table X

TABLE VIII
CORRELATION MATRIX OF THE VARIABLES

	Firm Size	*Firm Age*	*Innovation probability*	*ISO/IEC 27001 certified*
Firm Size	1.000 (1.000)			
Firm Age	0.337* (0.329*)	1.000 (1.000)		
Innovation probability	-0.042* (0.266*)	-0.146* (0.033*)	1.000 (1.000)	
ISO/IEC 27001 certified	0.026* (0.117*)	-0.007* (0.044*)	0.050* (0.069*)	1.000 (1.000)

Notes: The table shows the pairwise correlation coefficients of all observations in the MUP. ICT sector service coefficients are in brackets.
* $p < 0.01$.

TABLE IX
CORRECTED LOGIT ESTIMATES

Independent variables	*Certification to ISO/IEC 27001 MUP total*	*Certification to ISO/IEC 27001 ICT Service Sector*
Firm Size (in logs)	0.398***	0.553***
Firm Age (in logs)	-0.287***	0.062
Innovation Probability	5.804***	1.415***
Constant	-8.923***	-6.188 ***
Observations	374,145	14,333

Notes: The table displays the coefficients of all observations in the MUP and ICT service sectors applying rare event logistic regression.
* $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE X
PROBIT ESTIMATION RESULTS FOR SECTOR DUMMIES

Sector	*Coefficient*	*Marginal effect*
Automobile trade/repair	-1.232***	-0.00653
Automotive manuf.	-1.076***	-0.00641
Chemicals	-1.152***	-0.00648
Construction	-0.946***	-0.00625
Consulting	-0.460***	-0.00474
Creative services	-0.601***	-0.00540
Education	-1.274***	-0.00656
Electrical engineering	-0.995***	-0.00632
Electronics/optics	-0.975***	-0.00629
Engineering services	-0.627***	-0.00549
Financial services	-0.383***	-0.00427
Health & social services	-1.487***	-0.00664
Interest groups	-1.090***	-0.00642
Leisure services	-0.613***	-0.00544
Management services	-0.158**	-0.00224
Mechanical engineering	-1.569***	-0.00666
Media	-0.646***	-0.00556
Metalware	-1.203***	-0.00651
Mining	-0.427	-0.00455
Other manufacturing	-1.041***	-0.00637
Other services	-0.444***	-0.00465
Personal services	-0.540***	-0.00514
Public utility	0.002	0.00004
Real estate business	-0.748***	-0.00586
Repair/installation	-0.731***	-0.00582

Notes: The table displays the coefficients and marginal effects based on the ICT service sector. * $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$.

ACKNOWLEDGMENT

M. Mirtsch would like to thank G. Dudek for valuable insights, S. Mareschow and G. Miklis for assisting in categorizing web scraped firms, M. Franke for IT support, and S. Stobbe for the language editing and proofreading. Finally, the authors gratefully acknowledge the valuable suggestions of three anonymous reviewers.

REFERENCES

[1] ISOfocus, "The cyber secrets," Jan./Feb. 2019. [Online]. Available: https://www.iso.org/files/live/sites/isoorg/files/news/magazine/ISOfocus%20(2013-NOW)/en/2019/ISOfocus_132/ISOfocus_132_en.pdf
[2] S.-Y. Peng, "'Private' cybersecurity standards? Cyberspace governance, multistakeholderism, and the (Ir) relevance of the TBT regime," *Cornell Int. Law J.*, vol. 51, no. 2, pp. 445–469, 2018.
[3] S. Shackelford and S. O. Bradner, "Have you updated your toaster? Transatlantic approaches to governing the internet of everything," Kelley School Bus. Res. Paper No. 18-60, pp. 1–31. 2018. [Online]. Available: https://ssrn.com/abstract = 3208018
[4] Accenture, "The cost of cybercrime," Ninth Annual Cost of Cybercrime Study, Independently Conducted by Ponemon Institute LLC and Jointly Developed by Accenture, 2019. [Online] Available: https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf
[5] R. Saint-Germain, "Information security management best practice based on ISO/IEC 17799," *Inf. Manage. J.*, vol. 39, no. 4, pp. 60–66, 2005.
[6] G. Disterer, "ISO/IEC 27000, 27001 and 27002 for information security management," *J. Inf. Secur.*, vol. 4, no. 2, pp. 92–100, 2013.
[7] *Information Security Management Systems*, ISO/IEC 27001:2013 (EN), 2013.
[8] P. Castka and C. J. Corbett, "Management systems standards: Diffusion, impact and governance of ISO 9000, ISO 14000, and other management standards," *Foundations Trends Technol. Inf. Oper. Manage.*, vol. 7, no. 3/4, pp. 161–379, 2013.
[9] E. M. Rogers, *Diffusion of Innovations*, 5th ed. New York, USA: Free Press, 2003.
[10] V. Fomin, H. Vries, and Y. Barlette, "ISO/IEC 27001 information systems security management standard: Exploring the reasons for low adoption," in *Proc. 3rd Eur. Conf. Manage. Technol.*, 2008, pp. 1–13.
[11] Y. Barlette and V. Fomin, "The adoption of information security management standards: A literature review," in *Proc. Inf. Resour. Manage.: Concepts, Methodologies, Tools Appl.*, 2010, pp. 69–90.
[12] Y. Barlette and V. V. Fomin, "Exploring the suitability of IS security management standards for SMEs," in *Proc. 41st Annu. Hawaii Int. Conf. Syst. Sci.*, 2008, pp. 308–317.
[13] Z. Abu Bakar, N. A. Yaacob, Z. M. Udin, J. R. Hanaysha, and L. K. Loon, "The adoption of business continuity management best practices among malaysian organizations," *Adv. Sci. Lett.*, vol. 23, no. 9, pp. 8484–8491, Sep. 2017.
[14] A. Skopak and S. Sakanovic, "Adoption of standard for information security ISO/IEC 27001 in Bosnia and Herzegovina," in *Proc. Int. Conf. Econ. Social Stud. Sarajevo*, 2016, pp. 35–42.
[15] C. Candiwan, "Analysis of ISO27001 implementation for enterprises and SMEs in Indonesia," in *Proc. Int. Conf. Cyber-Crime Investigation Cyber Secur.*, 2014, pp. 50–58.
[16] K. I. Alshitri and A. N. Abanumy, "Exploring the reasons behind the low ISO 27001 adoption in public organizations in Saudi Arabia," in *Proc. Int. Conf. Inf. Sci. Appl.*, 2014, pp. 1–4.
[17] B. AbuSaad, F. A. Saeed, K. Alghathbar, and B. Khan, "Implementation of ISO 27001 in Saudi Arabia—Obstacles, motivations, outcomes, and lessons learned," in *Proc. Australian Inf. Secur. Manage. Conf.*, 2011, pp. 1–9.
[18] R. Van Wessel and H. J. de Vries, "Business impact of international standards for information security management. Lessons from case companies," *J. Inf. Commun. Technol. Standardization*, vol. 1, pp. 25–40, 2013.
[19] J. Kinne and J. Axenbeck, "Web mining of firm websites: A framework for Web scraping and a pilot study for Germany," Leibniz Assoc., Berlin, Germany, ZEW Discussion Paper 18-033, 2019.
[20] J. Kinne and D. Lenz, "Predicting innovative firms using web mining and deep learning," Leibniz Assoc., Berlin, Germany, ZEW Discussion Paper 19-001, 2019.
[21] K. Blind, "Certifications based on international management system standards as innovation indicators: An explorative feasibility analysis," in *Proc. 24th EURAS Annu. Standardisation Conf., Standards, Bio-Based Econ.*, 2019, pp. 51–69.
[22] H. Armbruster, A. Bikfalvi, S. Kinkel, and G. Lay, "Organizational innovation: The challenge of measuring non-technical innovation in large-scale surveys," *Technovation*, vol. 28, no. 10, pp. 644–657, 2008.
[23] G. Hashem and J. Tann, "The adoption of ISO 9000 standards within the Egyptian context: A diffusion of innovation approach," *Total Qual. Manage. Bus. Excellence*, vol. 18, no. 6, pp. 631–652, 2007.
[24] L. G. Tornatzky, M. Fleischer, and A. Chakrabarti, *The Processes of Technological Innovation (Issues in Organization and Management Series)*. Lexington, MA, USA: Lexington Books, 1990.
[25] M. V. Uzumeri, "ISO 9000 and other metastandards: principles for management practice?" *Acad. Manage. Perspectives*, vol. 11, no. 1, pp. 21–36, 1997.
[26] ISO, "Management system standards." Accessed on: Mar. 1, 2019. [Online]. Available: https://www.iso.org/management-system-standards.html
[27] *Conformity Assessment—Vocabulary and General Principles*, EN ISO/IEC 17000:2004, 2004.
[28] M. Spence, "Job market signaling," *Quart. J. Econ.*, vol. 87, no. 3, pp. 355–374, 1973.
[29] W. K. Viscusi, "A note on "lemons" markets with quality certification," *Bell J. Econ.*, vol. 9, no. 1, pp. 277–279, 1978.
[30] G. A. Akerlof, "The market for "lemons": Quality uncertainty and the market mechanism," *Quart. J. Econ.*, vol. 84, no. 3, pp. 488–500, 1970.
[31] A. Terlaak and A. A. King, "The effect of certification with the ISO 9000 quality management standard: A signaling approach," *J. Econ. Behav. Org.*, vol. 60, no. 4, pp. 579–602, 2006.

[32] M. Delmas and I. Montiel, "The diffusion of voluntary international management standards: Responsible Care, ISO 9000, and ISO 14001 in the chemical industry," *Policy Stud. J.*, vol. 36, no. 1, pp. 65–93, 2008.

[33] S. W. Anderson, J. D. Daly, and M. F. Johnson, "Why firms seek ISO 9000 certification: regulatory compliance or competitive advantage?" *Prod. Oper. Manage.*, vol. 8, no. 1, pp. 28–43, 1999.

[34] K. D. Gotzamani and G. D. Tsiotras, "An empirical study of the ISO 9000 standards' contribution towards total quality management," *Int. J. Oper. Prod. Manage.*, vol. 21, no. 10, pp. 1326–1342, 2001.

[35] M. Terziovski, D. Power, and A. S. Sohal, "The longitudinal effects of the ISO 9000 certification process on business performance," *Eur. J. Oper. Res.*, vol. 146, no. 3, pp. 580–595, 2003.

[36] M. Potoski and A. Prakash, "Information asymmetries as trade barriers: ISO 9000 increases international commerce," *J. Policy Anal. Manage.* vol. 28, no. 2, pp. 221–238, 2009.

[37] B. Manders, H. J. de Vries, and K. Blind, "ISO 9001 and product innovation: A literature review and research framework," *Technovation*, vol. 48, pp. 41–55, 2016.

[38] H. A. Quazi, Y.-K. Khoo, C.-M. Tan, and P.-S. Wong, "Motivation for ISO 14000 certification: development of a predictive model," *Omega*, vol. 29, no. 6, pp. 525–542, 2001.

[39] P. DiMaggio and W. W. Powell, "The iron cage revisited: Collective rationality and institutional isomorphism in organizational fields," *Amer. Sociol. Rev.*, vol. 48, no. 2, pp. 147–160, 1983.

[40] E. Naveh, A. Marcus, and H. Koo Moon, "Implementing ISO 9000: Performance improvement by first or second movers," *Int. J. Prod. Res.*, vol. 42, no. 9, pp. 1843–1863, May 2004.

[41] A. Terlaak and A. A. King, "Follow the small? Information-revealing adoption bandwagons when observers expect larger firms to benefit more from adoption," *Strategic Manage. J.*, vol. 28, no. 12, pp. 1167–1185, Dec. 2007.

[42] M. A. Delmas and M. Montes-Sancho, "An institutional perspective on the diffusion of international management system standards: The case of the environmental management standard ISO 14001," *Bus. Ethics Quart.*, vol. 21, no. 1, pp. 103–132, 2011.

[43] T. H. Arimura, N. Darnall, and H. Katayama, "Is ISO 14001 a gateway to more advanced voluntary action? The case of green supply chain management," *J. Environ. Econ. Manage.*, vol. 61, no. 2, pp. 170–182, 2011.

[44] P. J. Singh, M. Feng, and A. Smith, "ISO 9000 series of standards: comparison of manufacturing and service organisations," *Int. J. Qual. Rel. Manage.*, vol. 23, no. 2, pp. 122–142, 2006.

[45] G. M. P. Swann, "The economics of standardization: An update," Innov. Econ. Limited, London, U.K., Rep. U.K. Dept. Bus., Innov. Skills, 2010.

[46] X. Cao and A. Prakash, "Growing exports by signaling product quality: Trade competition and the cross-national diffusion of ISO 9000 quality standards," *J. Policy Anal. Manage.*, vol. 30, no. 1, pp. 111–135, 2011.

[47] B. Manders, "Implementation and impact of ISO 9001," Ph.D. dissertation, Erasmus Res. Inst. Manage. Rotterdam, The Netherlands, 2015.

[48] P. Bansal and W. C. Bogner, "Deciding on ISO 14001: Economics, institutions, and context," *Long Range Planning*, vol. 35, no. 3, pp. 269–290, 2002.

[49] K. Blind and A. Mangelsdorf, "Zertifizierung in deutschen Unternehmen–zwischen Wettbewerbsvorteil und Kostenfaktor," in *Zertifizierung als Erfolgsfaktor*. Berlin, Germany: Springer, 2016, pp. 23–32.

[50] M. L. Katz and C. Shapiro, "Network externalities, competition, and compatibility," *Amer. Econ. Rev.*, vol. 75, no. 3, pp. 424–440, 1985.

[51] C. J. Corbett and D. A. Kirsch, "International diffusion of ISO 14000 certification," *Prod. Oper. Manage.*, vol. 10, no. 3, pp. 327–342, 2001.

[52] F. Tuczek, P. Castka, and T. Wakolbinger, "A review of management theories in the context of quality, environmental and social responsibility voluntary standards," *J. Cleaner Prod.*, vol. 176, pp. 399–416, 2018.

[53] D. Maier, A. M. Vadastreanu, T. Keppler, T. Eidenmuller, and A. Maier, "Innovation as a part of an existing integrated management system," *Procedia Econ. Finance*, vol. 26, pp. 1060–1067, 2015.

[54] T. H. Jørgensen, A. Remmen, and M. D. Mellado, "Integrated management systems—Three different levels of integration," *J. Cleaner Prod.*, vol. 14, no. 8, pp. 713–722, 2006.

[55] R. Gey and A. Fried, "Metastructuring for standards: How organizations respond to the multiplicity of standards," in *Corporate and Global Standardization Initiatives in Contemporary Society*. Hershey, PA, USA: IGI Global, 2018, pp. 252–276.

[56] H. J. de Vries and F. El Osrouti, "Impact studies on standards and standardisation - Looking back and moving forward," in *Proc. 24th EURAS Annu. Standardisation Conf., Standards, Bio-Based Econ.*, 2019, pp. 131–142.

[57] C. B. Stamm, "ISO 26000 gets taken around: Diffusion work as crucial link between standard creation and adoption," in *Corporate Social Responsibility and Corporate Change*. Berlin, Germany: Springer, 2019, pp. 135–158.

[58] *Information Technology—Security Techniques—Information Security Management Systems—Overview and Vocabulary*, ISO/IEC 27000:2018 (en), 2018.

[59] R. Von Solms and J. Van Niekerk, "From information security to cyber security," *Comput. Secur.*, vol. 38, pp. 97–102, 2013.

[60] ISO, "The ISO survey of management system standard certifications 2017," 2018. [Online]. Available: https://www.iso.org/the-iso-survey.html, Accessed on: Feb. 2, 2019.

[61] D. Tunçalp, "Diffusion and adoption of information security management standards across countries and industries," *J. Global Inf. Technol. Manage.*, vol. 17, no. 4, pp. 221–227, 2014.

[62] T. Neubauer, A. Ekelhart, and S. Fenz, *Interactive Selection of ISO 27001 Controls Under Multiple Objectives*. Boston, MA, USA: Springer, 2008, pp. 477–492.

[63] N. F. Doherty and H. Fulford, "Do information security policies reduce the incidence of security breaches: an exploratory analysis," *Inf. Resour. Manage. J.*, vol. 18, no. 4, pp. 21–39, 2005.

[64] C. Hsu, T. Wang and A. Lu, "The impact of ISO 27001 certification on firm performance," in *Proc. 49th Hawaii Int. Conf. Syst. Sci.*, 2016, pp. 4842–4848.

[65] G. P. Tejay and B. Shoraka, "Reducing cyber harassment through de jure standards: A study on the lack of the information security management standard adoption in the USA," *Int. J. Manage. Decis. Making*, vol. 11, no. 5/6, pp. 324–343, 2011.

[66] Y. Benslimane, Z. Yang, and B. Bahli, "Information security between standards, certifications and technologies: An empirical study," in *Proc. Int. Conf. Inf. Sci. Secur.*, 2016, pp. 1–5.

[67] A. Longras, T. Pereira, P. Cameiro, and P. Pinto, "On the track of ISO/IEC 27001: 2013 implementation difficulties in portuguese organizations," in *Proc. Int. Conf. Intell. Syst.*, 2018, pp. 886–890.

[68] S. Uwizeyemungu and P. Poba-Nzaou, "Understanding information technology security standards diffusion: An institutional perspective," in *Proc. Int. Conf. Inf. Syst. Secur. Privacy*, 2015, pp. 5–16.

[69] J. A. Schumpeter, *Theorie der wirtschaftlichen entwicklung. Leipzig: Duncker & Humblot. English Translation Published in 1934 As the Theory of Economic Development*. Cambridge, MA, USA: Harvard Univ. Press, 1912.

[70] T. Oliveira and M. F. Martins, "Literature review of information technology adoption models at firm level," *Electron. J. Inf. Syst. Eval.*, vol. 14, no. 1, pp. 110–121, 2011.

[71] K. K. Kuan and P. Y. Chau, "A perception-based model for EDI adoption in small businesses using a technology–organization–environment framework," *Inf. Manage.*, vol. 38, no. 8, pp. 507–521, 2001.

[72] Y.-M. Wang, Y.-S. Wang, and Y.-F. Yang, "Understanding the determinants of RFID adoption in the manufacturing industry," *Technol. Forecasting Social Change*, vol. 77, no. 5, pp. 803–815, 2010.

[73] M.-J. Pan and W.-Y. Jang, "Determinants of the adoption of enterprise resource planning within the technology-organization-environment framework: Taiwan's communications industry," *J. Comput. Inf. Syst.*, vol. 48, no. 3, pp. 94–102, 2008.

[74] I. Heras-Saizarbitoria and O. Boiral, "ISO 9001 and ISO 14001: towards a research agenda on management system standards," *Int. J. Manage. Rev.*, vol. 15, no. 1, pp. 47–65, 2013.

[75] J. Llach, R. D. Castro, A. Bikfalvi, and F. Marimon, "The relationship between environmental management systems and organizational innovations," *Hum. Factors Ergonom. Manuf. Serv. Ind.*, vol. 22, no. 4, pp. 307–316, 2012.

[76] G. Mangiarotti and C. A. F. Riillo, "Determinants of ISO9000:2000 certification in services and manufacturing: An empirical analysis for luxembourg," in *Proc. 4ème Colloque Luxembourgeois sur l'économie de la Connaissance Dans une Perspective Européenne*, 2010, pp. 7–9.

[77] E. Hoti, "The technological, organizational and environmental framework of IS innovation adaption in small and medium enterprises. Evidence from research over the last 10 years," *Int. J. Bus. Manage.*, vol. 3, no. 4, pp. 1–14, 2015.

[78] N. Askitas and K. F. Zimmermann, "The internet as a data source for advancement in social sciences," *Int. J. Manpower*, vol. 36, no. 1, pp. 2–12, 2015.

[79] R. Kosala and H. Blockeel, "Web mining research: A survey," *ACM SIGKDD Explorations Newslett.*, vol. 2, no. 1, pp. 1–15, 2000.

[80] A. Gök, A. Waterworth, and P. Shapira, "Use of web mining in studying innovation," *Scientometrics*, vol. 102, no. 1, pp. 653–671, 2015.
[81] G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, and D. Delen, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Cambridge, MA, USA: Academic, 2012.
[82] J. S. Katz and V. Cothey, "Web indicators for complex innovation systems," *Res. Eval.*, vol. 15, no. 2, pp. 85–95, 2006.
[83] J. Youtie, D. Hicks, P. Shapira, and T. Horsley, "Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies," *Technol. Anal. Strategic Manage.*, vol. 24, no. 10, pp. 981–995, 2012.
[84] S. K. Arora, J. Youtie, P. Shapira, L. Gao, and T. Ma, "Entry strategies in an emerging technology: a pilot web-based study of graphene firms," *Scientometrics*, vol. 95, no. 3, pp. 1189–1207, 2013.
[85] C. Beaudry, M. Héroux-Vaillancourt, and C. Rietsch, "Validation of a web mining technique to measure innovation in high technology Canadian industries," in *Proc. 1st Int. Conf. Adv. Res. Methods Anal.*, 2016, pp. 1–25.
[86] M. Nathan and A. Rosso, "Innovative events," Centro Studi Luca d'Agliano, Torino, Italy, Develop. Stud. Work. Paper 429, 2017.
[87] J. Bersch, S. Gottschalk, B. Müller, and M. Niefert, "The Mannheim Enterprise Panel (MUP) and firm statistics for Germany," Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, Germany, ZEW Discussion Paper 14-104, 2014.
[88] Eurostat, "Statistical classification of economic activities in the european community," NACE Rev. 2, 2008. Accessed on: Feb. 2, 2019. [Online]. Available: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl = LST_NOM_DTL&StrNom = NACE_REV2&StrLanguageCode = EN&IntPcKey = &StrLayoutCode = HIERARCHIC
[89] G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no. 2, pp. 137–163, 2001.
[90] O. Hogan, R. Jayasuriya, and C. Sheehy, "Economic Contribution of Standards in Ireland: A report for the National Standards Authority of Ireland," Centre for Econ. Bus. Res. (CEBR), London, U.K., Dec. 2015.
[91] DEKRA, "Informationssicherheit." [Online]. Available: https://www.dekra-akademie.de/de/iso2700x-schulung/, Accessed on: March 7, 2019.
[92] I. M. Lopes, T. Guarda, and P. Oliveira, "How ISO 27001 can help achieve GDPR compliance," in *Proc. 14th Iberian Conf. Inf. Syst. Technol.*, 2019, pp. 1–6.
[93] K. L. Keller, "Brand synthesis: The multidimensionality of brand knowledge," *J. Consum. Res.*, vol. 29, no. 4, pp. 595–600, 2003.
[94] J. Hu, Y. Zhang, and X. Fang, "Research on partner selection mechanism of technological standard alliance: From the perspective of network embeddedness," in *Proc. Portland Int. Conf. Manage. Eng. Technol.*, 2015, pp. 585–595.
[95] Eurostat. "ICT security in enterprises," 2015. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php/ICT_security_in_enterprises. Accessed on: Nov. 19, 2018.
[96] H. Mintzberg, S. Ghoshal, J. Lampel, and J. B. Quinn, *The Strategy Process: Concepts, Contexts, Cases*. Harlow, UK: Pearson Educ., 2003.
[97] J. M. Utterback and W. J. Abernathy, "A dynamic model of process and product innovation," *Omega*, vol. 3, no. 6, pp. 639–656, 1975.
[98] European Commission, "2019 Rolling plan for ICT standardisation," DG Internal Market, Ind., Entrepreneurship SMEs, Eur. Commission, Brussels, Belgium, 2019.
[99] C. Tankard, "What the GDPR means for businesses," *Netw. Secur.*, vol. 2016, no. 6, pp. 5–8, 2016.

Mona Mirtsch received the M.Sc. degree in business administration from the San Diego State University, San Diego, CA, USA, in 2004, and the Diploma in business administration from the European University Viadrina Frankfurt (Oder), Frankfurt (Oder), Germany, in 2006. She is currently working toward the Ph.D. degree in innovation economics with the Technische Universität Berlin, Berlin, Germany, in the field of cybersecurity and conformity assessment.

From 2006 to 2010, she was a Trainee and a Brand Manager for a multinational fast-moving consumer goods corporation in Hamburg, Germany. From 2010 to 2017, she was a Sales Manager also responsible for quality management for a family-owned metal forming company in Berlin, Germany. Since 2017, she has been working with the Department for Accreditation and Conformity Assessment at the Bundesanstalt für Materialforschung und -prüfung (Federal Institute for Materials Research and Testing—BAM), Berlin, Germany, dealing with questions of quality infrastructure issues.

Jan Kinne received the master's degree in geography from the Heidelberg University, Heidelberg, Germany, in 2016. He is currently working toward the Ph.D. degree in applied geoinformatics at the University of Salzburg, Salzburg, Austria in the field of microgeograhic innovation research using web data.

He was a Visiting Fellow with the Institute for Quantitative Social Sciences, Harvard University in 2019. Since 2016, he has been working as a Researcher with the Economics of Innovation Department, ZEW Centre for European Economic Research. Based on his Ph.D. research, he co-founded istari.ai (istari artificial intelligence), a startup company for AI-driven web analysis of company websites. His main areas of study were geoinformatics and spatial analysis (GIScience).

Knut Blind received the Bachelor's degree of Arts from Brock University, St. Catharines, ON, Canada, in 1990 and the Diploma in economics and the Doctoral degree in economics from Freiburg University, Freiburg, Germany in 1995. He studied economics, political science, and psychology at Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

In April 2006, he was appointed as a Professor of Innovation Economics with the Faculty of Economics and Management, Technische Universität Berlin. Between 2008 and 2016, he also held the endowed Chair of Standardisation at the Rotterdam School of Management, Erasmus University. Since 1996, he has been with the Fraunhofer Society (currently the Fraunhofer Institute of Systems and Innovation Research).

Appendix F

Paper 6: The Digital Layer: How innovative firms relate on the Web

// NO.20-003 | 01/2020

DISCUSSION PAPER

// MIRIAM KRÜGER, JAN KINNE,
DAVID LENZ, AND BERND RESCH

The Digital Layer: How Innovative Firms Relate on the Web

Leibniz Association

ZEW

The Digital Layer: How innovative firms relate on the Web

Miriam Krüger[a,1], **Jan Kinne**[b,c,d,e], **David Lenz**[b,f], **and Bernd Resch**[d,e]

[a]Technical University of Berlin, Berlin, Germany; [b]istari.ai, Mannheim, Germany; [c]Department of Economics of Innovation and Industrial Dynamics, ZEW Centre for European Economic Research, Mannheim, Germany; [d]Department of Geoinformatics - Z_GIS, University of Salzburg, Salzburg, Austria; [e]Center for Geographic Analysis, Harvard University, Cambridge, Massachusetts, USA; [f]Department of Econometrics and Statistics, Justus-Liebig-University, Gießen, Germany

Abstract

In this paper, we introduce the concept of a Digital Layer to empirically investigate inter-firm relations at any geographical scale of analysis. The Digital Layer is created from large-scale, structured web scraping of firm websites, their textual content and the hyperlinks among them. Using text-based machine learning models, we show that this Digital Layer can be used to derive meaningful characteristics for the over seven million firm-to-firm relations, which we analyze in this case study of 500,000 firms based in Germany. Among others, we explore three dimensions of relational proximity: (1) Cognitive proximity is measured by the similarity between firms' website texts. (2) Organizational proximity is measured by classifying the nature of the firms' relationships (business vs. non-business) using a text-based machine learning classification model. (3) Geographical proximity is calculated using the exact geographic location of the firms. Finally, we use these variables to explore the differences between innovative and non-innovative firms with regard to their location and relations within the Digital Layer. The firm-level innovation indicators in this study come from traditional sources (survey and patent data) and from a novel deep learning-based approach that harnesses firm website texts. We find that, after controlling for a range of firm-level characteristics, innovative firms compared to non-innovative firms maintain more numerous relationships and that their partners are more innovative than partners of non-innovative firms. Innovative firms are located in dense areas and still maintain relationships that are geographically farther away. Their partners share a common knowledge base and their relationships are business-focused. We conclude that the Digital Layer is a suitable and highly cost-efficient method to conduct large-scale analyses of firm networks that are not constrained to specific sectors, regions, or a particular geographical level of analysis. As such, our approach complements other relational datasets like patents or survey data nicely.

Keywords: Web Mining | Innovation | Proximity | Network | Natural Language Processing
JEL Classification: O30, R10, C80

1. Introduction

Since Schumpeter (1) innovation has been recognized as the key element driving economic growth (2). As a consequence, for decades both researchers and policy makers have focused on understanding innovation dynamics in networks of firms and the drivers behind them. One of the well researched aspects thereby is the impact of proximity on learning, knowledge creation and innovation. Boschma (3) conceptualized five dimensions of proximity that are related to the innovativeness of a firm in a network of firms: cognitive, geographical, organizational, institutional and social proximity. The theoretical approach of (3) found wide adaption in economic geography but has proven to be difficult to operationalize in large-scale empirical studies (see Literature review section). In this paper, we introduce a novel approach based on web mining to map firm networks and to analyze the characteristics of innovative firms in them. For that, we create a so-called Digital Layer of the network of firms located in Germany from large scale web scraping of firm websites, their textual content and the hyperlinks among them. This allows us to analyze firm-to-firm relations and firm characteristics at a larger scale and higher granularity compared to studies using traditional data based on questionnaire-based surveys or patents.

This way, we are able to investigate the characteristics of over half a million firms located in Germany and over seven million relations among them. Using text-based classification and text similarity models from machine learning, we create quantitative measures that describe the position and relationships of each firm in the Digital Layer. We demonstrate that these measures offer meaningful insights on firm-level innovativeness. These measures include the number of partners that a firm has in the network, the innovativeness of its partners, as well as several proximity measures describing the relation to the link partners of each firm.

We then relate these measures (and several firm-level control variables) to the innovativeness of firms in a regression analysis. In this regression analysis, we use two different firm-level innovation indicators as the dependent variable. First, we use a traditional indicator from the questionnaire-based German Community Innovation Survey (CIS) which includes information for about 2,500 firms in our dataset. Second, we use a web-based firm-level innovation indicator developed by (4) which is based on an artificial neural network classification model trained on website texts of firms surveyed in the CIS. The latter indicator is available for all 513,026 firms in our dataset.

With this study we aim to answer the following research

Miriam Krüger and Jan Kinne designed the study and wrote the paper. Jan Kinne gathered the data. Jan Kinne, Miriam Krüger and David Lenz analyzed the data. Bernd Resch supervised the study, discussed the results and proof-read the paper.

The authors declare no conflict of interest. The authors would like to thank the German Federal Ministry of Education and Research for providing funding for the research project (TOBI - Text Data Based Output Indicators as Base of a New Innovation Metric and NETINU - Networks of Innovative Firms; funding ids 16IFI001 and 16IFI106) of which this study is a part.

[1]To whom correspondence should be addressed. E-mail: jan.kinne@zew.de

questions:

1. **Research Question 1**: Is our approach to create a *Digital Layer* of interrelated and textually described firms suitable for a large scale web-based analysis of firm networks?

2. **Research Question 2**: How do innovative and non-innovative firms differ concerning their relationships in the Digital Layer and are the observed statistical relations between the different dimensions of proximity and firm innovation in line with the established theory?

The remainder of this paper is structured as follows: First, we give an overview of the literature related to this study. We then present the datasets used to create the Digital Layer and to assess firm-level innovation. In the following methodology section, we outline how we developed measures of firm-to-firm proximity and firm-level embeddedness. We then present our results and discuss them in the following two sections. We finalize this paper with our conclusions and an outlook to potential future research.

2. Literature review

Firm networks, proximity and innovation. More than two decades ago, (5) pointed out that technological change has brought into existence a new type of economy where "information is the key ingredient of social organization and flows of messages and images between networks constitute the basic thread of social structure." According to his reasoning, it is now networks that form the social morphology of our societies and "the extent to which a network has access to technological know-how is at the roots of productivity and competitiveness". In his book "Why information grows" (6) further builds upon this concept of our economy as a social construct of connected firms. Firms again are regarded as networks of individuals and the degree to which firms and networks of firms are capable of producing and crystalizing information lies at the core of why some places are economically successful and others are not. This paradigm differs from the previous view on innovative places and competitive firms as summarized by (7):

> "For a long time, a fundamental debate existed in economic geography about the question whether places are more relevant for the competitiveness of firms, or whether networks matter more (Castells 1996). While the concept "space of places" expresses the idea that the location matters for learning and innovation (being in the right place is what counts), the concept of "space of flows" focuses more on the idea that networks are important vehicles of knowledge transfer and diffusion (meaning that being part of a network is crucial). In a nutshell, the cluster literature claimed that regions are drivers of innovation and economic development: firms in clusters benefit almost automatically from knowledge externalities that are "in the air", as Marshall once put it. [...] This is not to say that the cluster literature overlooked the importance of networks. The problem was, however, that the cluster literature suggested that the space of place and the space of flows showed a great deal of overlap (Boschma and Ter Wal 2007). [...] Knowledge networks are not territorial, [though], but social constructs that may cross the boundaries of regions. Knowledge diffuses through social networks which may be dense between local agents, but may also span across the world."

And it is not only geography that matters for effective knowledge flows, learning and innovation. (3) conceptualized five dimensions of proximity that play a crucial role for inter-organizational interaction and innovation: cognitive, institutional, social, organizational and geographical proximity. (8) wrote:

> "In short, cognitive proximity indicates the extent to which two organizations share the same knowledge base; organizational proximity the extent to which two organizations are under common hierarchical control, social proximity the extent to which members of two organizations have friendly relationships, institutional proximity the extent to which two organizations operate under the same institutions, and geographical proximity the physical distance or travel time separating two organizations."

Traditional relational data for innovation networks. To empirically assess these different dimensions of proximity and their relation to innovation in firms, relational data is needed. So far relational data has been obtained from either primary survey data or secondary data sources such as patent data. Even though other sources of secondary network data exist, e.g. strategic alliance databases or co-publications, patent data is the most widely used. (7) review and assess the advantages and drawbacks of primary survey data and secondary patent data as relational datasets:

Primary survey data is obtained through interviews and/or questionnaires either by means of the roster-recall methodology or the snowball method (for more information on these methods see (7)). As this is very costly and time-intensive, primary survey data generally fails to capture an entire firm population and is thus regionally or sectorally bounded. Moreover, the quality of the obtained data is very dependent on the response rate of firms. Most datasets represent a static network at one point in time, as the conduction of longitudinal surveys for a potential dynamic analysis of firm networks is even more costly and time-intensive. They therefore conclude that "network analysis on the basis of primary data is most appropriate for small clusters of firms or relatively small sectors within a region." An advantage of survey data is that it can record different dimensions of relationships across the same set of actors. An example for that is (9) study, in which a network of business relations and a network of knowledge-based relationships is identified.

Secondary patent data provides relational links based on the information about the patent applicant or the inventors. The node in the network is hence either the firm or the inventor. A link between firms or inventors exists in case of co-patenting or multi-applicant inventorship. Patent data therefore only reveals relatively formal cooperative links that resulted in a filed patent. Many other forms of inter-firm cooperation and more informal inter-firm interaction are not captured. Moreover, there are only some sectors that strongly rely on patents to protect their innovations, such as the pharmaceutical or the semiconductor industries. Other sectors, such as software industries and services, protect their innovations more likely

via secrecy or trademarks. Network studies based on patent data are thus more appropriate for certain sectors than for others. An advantage is, however, that one can construct and analyze networks back in time, as patent data is available for a long time-series. This allows for dynamic analyses of inter-firm networks.

The Digital Layer as a new generation of web-based relational data. In this study, we introduce the concept of a Digital Layer created from large-scale web scraping of geolocated firm websites. The relations among firms in the Digital Layer are constructed from the hyperlinks between their websites, enriched with quantitative measures based on the websites' textual content. (10) identified hyperlinks as the "basic structural element of the internet". He points to hyperlinks as a new social or communication channel and as a means for organizations to exchange information and sustain cooperative relationships. According to him, a hyperlink system is comprised of organizations that are linked together around a common background, interest, or project. In this sense, we expect the Digital Layer to reveal relationships among firms which are of cooperative rather than competitive nature. We explore how the position and relationships of each firm in the Digital Layer relate to firm innovation based on quantitative measures, which operationalize the cognitive, organizational, and geographical proximity to link partners. This way, our dataset allows us to empirically investigate the characteristics of inter-firm interaction and innovation at a larger scale and higher granularity than with previous datasets available. Our dataset does not constrain us to specific sectors (see data section) and bears great potential for a dynamic network analysis of inter-firm relationships (see future work section).

Innovative and non-innovative firms in the Digital Layer. Based on the findings of previous studies using patent and survey data (11–13), we expect that innovative and non-innovative firms differ with regard to their position and relationships in the Digital Layer. (7), for example, reference the study of (12) which found "empirical evidence that firms with cutting-edge technology are usually positioned in the core of inter-firm collaboration networks." Moreover, (13) and (11) found a positive relationship between network centrality of firms and their innovative performance. We hence expect innovative firms to have a higher degree centrality, meaning more hyperlinks, than non-innovative firms in the Digital Layer. Based on the concept of *homophily* (14, 15), meaning that actors link to actors that are similar to them, we also expect that innovative firms especially link to firms of their own sector and to other innovative firms.

Proximity and innovation in the Digital Layer. (8) claim that "it depends on the optimal level of proximity between agents whether their connection will lead to a higher level of innovative performance or not". This means that both too much and too little proximity to partners can hamper interactive learning and innovation. Hence, we expect that close relationships between firms in terms of their cognitive, geographical, and organizational proximity (social and institutional proximity are not assessed in this study), are not necessarily related to higher innovativeness.

Concerning geographical proximity, it is argued that remotely located firms with merely distant partners will not be able to catch the *local buzz* and knowledge spillovers that firms in densely urban locations can grasp from more frequent and sometimes serendipitous face-to-face interactions with other economic actors. On the other hand, local over-embeddedness without any *global pipelines* might lead to missing the next crucial development from another place (16). Some trans-regional linkages are considered crucial to protect from so-called *technological lock-ins* (7, 17). In this sense, we expect innovative firms to have a mixture of local and trans-regional links.

In the case of cognitive proximity, (8) argue that a firm's cognitive base needs to be "close enough to new knowledge in order to communicate, understand and process it successfully". If the cognitive distance between actors becomes too large, learning and knowledge flows are hampered. (18) found that "firms innovate in areas close to their current cognitive capabilities along well-defined technological trajectories". (19) showed that cognitive proximity may enable RD alliances and (20) identified cognitive proximity of actors via patent citations. We thus expect innovative firms in the Digital Layer to be linked to firms that are in close cognitive proximity.

In the case of organizational proximity, "a continuum is assumed ranging from one extreme of 'on the spot' market, to informal relations between firms [...] to the other extreme of a hierarchically organized firm" (8). (11) found a positive relationship between firm survival and a mixture of embedded trust-based ties and arm's length market based ties of a firm. In this sense, we expect innovative firms to be linked with both organizationally close and distant firms.

3. Data

In this section, we first present the base firm dataset of this study. We then outline how web scraping was used to transfer the base dataset into the Digital Layer - a network of hyperlinked firms with associated web texts. Lastly, we present two innovation datasets (the German Community Innovation Survey and a large scale dataset of web-based innovation indicators) that are used in this study.

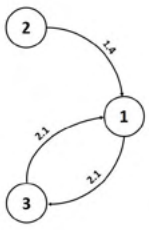
Firm base data. We use the Mannheim Enterprise Panel (MUP) of 2019 as our base dataset. The MUP is a firm panel database that covers the entire population of firms in Germany. It is updated on a semi-annual basis (21). In addition to firm-level characteristics, such as firm size, age, and location, the MUP also includes the web addresses (URL) for 1,155,867 of the 2,497,412 firms in early 2019 (*URL coverage* of 46%). A prior analysis of this dataset (22) showed that URL coverage differs systematically by sectors, regions, firm size and age groups. Very small and young firms (smaller than five employees and younger than two years), especially from sectors such as agriculture, are not covered as well as medium sized and larger firms from sectors like manufacturing and ICT (information and communication technology) services. The MUP, nonetheless, represents a comprehensive dataset with a very high URL coverage in those firm groups that are the most relevant for the development of innovation (22, 23). We removed firms without address information from our dataset and geocoded the remaining firms using street-level geocoding (without house numbers; see e.g. (24)).

The geocoded firms were also used to calculate a firm-level location control variable by counting the number of other firms within one kilometer around each individual firm.

The resulting local firm densities are used as a control for potential local spillovers. The search radius of one kilometer was selected according to (25) who showed that spillovers from local knowledge sources decay within a few hundred meters.

Constructing the Digital Layer. For the web scraping of the firm websites, we used ARGUS (26), an open source web scraping tool based on Python's Scrapy scraping framework. ARGUS was used to scrape texts from the websites of all MUP firms as well as the hyperlink connections among the firms. After the web scraping, we excluded erroneous downloads and potentially misleading redirects (see (22)) from the data. After this step, 684,873 firms remained in the dataset.

We then created a network of firms where the edges are constructed from the extracted hyperlinks between firms (see Figure 1 for an schematic representation). At this, edges are given either weight 1.0, if the hyperlink connection between a pair of firms is unidirectional, or weight 2.0, if the firms are mutually linked (i.e. both firms have a hyperlink connection to the other firm on their respective websites). As an example, in Figure 1, *firm 3* appears two times in the hyperlink vector of *firm 1* because the firms are mutually linked. As a result, the corresponding exemplary proximity value (say, the geographical distance between *firm 1* and *firm 3*) is weighted by 2.0 when calculating the mean proximity of *firm 1*.



Firm	Hyperlinks	Proximities	Mean Proximity
1	[2, 3, 3]	[1.4, 2.1, 2.1]	1.9
2	[1]	[1.4]	1.4
3	[1, 1]	[2.1, 2.1]	2.1

Fig. 1. Schematic representation of a firm hyperlink network. Network of three firms with hyperlink connections and a corresponding exemplary proximity measure.

After constructing the network, we excluded 150,116 firms (21.9%) without any hyperlink connections to other firms. Firms without any links have considerably fewer employees (11.9 vs. 27.7) than firms with hyperlinks and are younger (23.0 vs. 24.8 years) as well. Both values are different at a highly significant level according to a t-test. Both firms with and without hyperlinks were used to calculate a local firm density control variable though (see below). Overall, there are 7,076,560 hyperlink connections in our dataset.

Firm-level innovation data. We use two datasets with firm-level innovation indicators: The Mannheim Innovation Panel (MIP), a traditional questionnaire-based innovation survey of firms sampled from the MUP, and a web-based innovation indicator developed by (4).

The MIP survey is the German contribution to the Community Innovation Survey (CIS), which is conducted every two years in the European Union, and has been used in an array of innovation studies (27). The survey methodology and the definition of innovation follows the Oslo Manual (28) and covers firms with five or more employees from the sectors of manufacturing and business-oriented services. In the survey, firms are asked whether they introduced new or significantly improved products or services (*product innovations*) during the three years prior to the survey, as well as whether they will introduce such products or services in the current year. In this study, we use the latter indicator from the MIP survey of 2018 which relates to the same year and is available for 2,463 firms.

Our second innovation dataset consists of predicted firm-level product innovator probabilities based on a deep learning model and website texts. For this web-based indicator, an artificial neural network (ANN) was trained on the website texts of firms surveyed in the MIP. After training on this dataset of labelled (product innovator/no product innovator) firm website texts, the ANN is able to predict the product innovator probability of any out-of-sample firm with a website. (4) have shown that this approach can be used to generate reliable firm-level innovation indicators even in industrial sectors and size groups that are not covered in the training data (i.e. in the MIP survey). This web-based indicator is available for all 534,757 firms in our dataset.

Table 1 presents key descriptive statistics for both innovation datasets (i.e. the MIP *survey dataset* and the deep learning based *web dataset*). Due to the sampling scheme of the MIP, the survey dataset includes larger and older firms on average and certain sectors are over-represented (for more information see (23)). Even though the web dataset is closer to the overall German firm population, the results of (22) showed that it is not unbiased. Larger and older firms from certain sectors are more likely to have a website and thus are over-represented in the web dataset. Firms in the survey dataset are located in more densely populated areas on average. All these differences are statistically significant according to a t-test. The number of hyperlinks per firms, on the other hand, are not significantly different, but the distribution is extremly skewed especially for the *web dataset*. As a consequence, we use logs of this variable for the further analysis.

We report both the original continuous (C in Table 1) web-based innovation indicator and a binary (B) recast to make it comparable to the binary MIP survey indicator. The mean product innovator probability in the web dataset is 25%. Casted to a binary variable using a classification threshold of 0.4 (see (4)) results in only 16% predicted product innovators compared to 25% in the survey dataset. Given that the latter dataset intentionally over-samples innovative firm types (due to the sampling procedures outlined in (28)) while the web dataset is closer to the overall firm population, these values are credible (see also (4) for details).

4. Methodology

In this section, we outline how we operationalize the network position of each individual firm. Geographical, cognitive, and organizational proximity to each firm's link partners reflect the distances between firms with values of 0.0 indicating closest proximity and values of 1.0 indicating farthest distance. We

Table 1. Firm characteristics.

Variable	Mean	Median	Min	Max	Filled
	Survey dataset (n=2,463)				
Link count	11.36	5	1	992	1.00
Employees	81.97	39	1	5,060	0.80
Age	42.85	28.99	2.95	908	0.99
Firm density	879.50	79	0	3,879	1.00
Surveyed inno.	0.25	0	0	1	1.00
Pred. inno. (C)	0.30	0.23	0.37	0.93	1.00
Pred. inno. (B)	0.24	0	0	1	1.00
	Web dataset (n=543,825)				
Link count	13.01	4	1	168,961	1.00
Employees	27.65	6	1	244,038	0.52
Age	24.79	17.03	0.91	1019	0.94
Firm density	176.80	53	0	3,930	1.00
Surveyed inno.	-	-	-	-	-
Pred. inno. (C)	0.25	0.20	0.03	0.93	1.00
Pred. inno. (B)	0.16	0	0	1	1.00

also create firm-level measures that grasp the innovativeness of hyperlinked partners and the overall number of partners a firm is hyperlinked to. For all these measures we calculate the mean as it was outlined in Figure 1. In an earlier version of this paper, we also calculated standard deviations to capture the heterogeneity of each individual firm's network but found that a simple hyperlink count per firm sufficiently predicts for network heterogeneity.

Link count and mean partner innovation. *Link count* is a simple count of all the hyperlinks a firm maintains to other firms. In Figure 1, *firm 1* has a link count of 3 and *firm 3* has a link count of 2, for example. As such, the link count variable is analogous to the *degree* measure in social network analysis.

The *mean partner innovation* is a simple measure that reflects the innovativeness of the hyperlinked partners that a firm has in the Digital Layer. It is calculated by taking the mean of the firm-level web-based innovation indicator (see Data section) of the hyperlinked partners of a firm.

Geographical proximity. We measure geographical proximity by calculating the euclidean distance between firms that are hyperlinked. For each firm, we then calculate the mean euclidean distance to its partners. We normalize the resulting distances to values between 0.0 (0.0 meters) and 1.0 (840,858 meters, the maximum value in our dataset) to make it easier to compare geographical proximity with the other two dimensions of proximity, which naturally range from 0.0 to 1.0.

Cognitive proximity. The cognitive proximity between hyperlinked firms is operationalized by calculating the similarity between their website texts. We know that firms use their websites to present themselves, their products and services. These information are usually codified as text and can be extracted and analyzed to assess a firms' products, services, credibility, achievements, key personnel decisions, and strategies (29). In its entirety, website texts are a description of a firm's knowledge base and we use it to calculate the cognitive proximities between the firm and its hyperlinked partners.

We represent the firms' website texts in a high-dimensional vector space by transferring them using a term frequency-inverse document frequency (tf-idf) scheme (see e.g. (30)). The tf-idf algorithm transfers each document to a fixed size sparse vector of size V, where V is the size of a dictionary composed of all words found in the overall text corpus. We restricted our dictionary to words with a minimum document frequency of 1.5% and a maximum document frequency of 65% (*popularity based filtering*). Each entry in the tf-idf vector of a document corresponds to one word in the dictionary, representing the relative importance of this word in the document. Words that do not appear in a given document are represented by a 0 value.

Specifically, in a first step (the tf step) the number of appearances per word in a single document are counted. In a second step, the inverse document frequency (idf) is used as a weighting scheme to adjust the tf counts. Conceptually, the idf weights determine how much information is provided by a specific word by means of how frequently a word appears in the overall document collection. The intuition is that very frequent words that appear in a lot of documents, should be given less weight compared to less frequent words, as infrequent words are more useful as a distinguishing feature.

We then use the tf-idf vector of a firm to calculate its similarity to the website texts of other firms, which have a hyperlink to the firm under consideration. We quantify the similarity between the two website texts by computing the cosine similarity of their vector representations (see e.g. (30)), an approach widely adopted in natural language processing studies (see e.g. (31–33)). For the sake of consistency, we transform the calculated cosine similarities to cosine distances, which range from 0.0 (identical texts) to 1.0 (maximal dissimilar texts). Again, we then calculate the mean of the cognitive distances between a firm and its hyperlinked partners.

Organizational proximity. We operationalize organizational proximity as a binary variable by classifying the nature of each relation between hyperlinked firms as one of the following two classes:

- **Non-business relation**: Non-business relations are relations between firms that are not directly related to making business with each other and are of non-monetary nature. Such relations primarily include the membership in (industrial) associations or chambers of commerce, and references to regulatory or legal bodies (e.g. commercial courts, commercial registries). Hyperlinks to purely informative web contents are also part of this class. Such references may include, for example, hyperlinks from a pharmacy to an external website that informs about healthy diets or a hyperlink from a firm to the website of a local news outlet that reports about the firm's latest achievements.

- **Business relation**: This class includes all hyperlinks between firms that do or did business together. Oftentimes, firms include hyperlinks to the websites of other companies to present them as testimonials or because they have an ongoing business relation (e.g. web hosting, web design, web mail providers, certification services). If a firm hyperlinks to its own social media profiles, the company that operates the social media platform is a business partner of that firm as well (because they provide the platform and make money from it). Hyperlinks between entities of the same corporate group or between personal

websites of employees and their employer (e.g. professor to university) are also part of this class.

In terms of the degree of organizational proximity, the business relation is closer than the non-business relation as the ties represented by it are usually more formal and reoccurring. In that sense, we quantify the nature of each hyperlink connection between two firms as either value 0.0 (weak non-business relation) or 1.0 (strong business relation) that can be predicted in a binary machine learning classification task. For this classification, we again use the firms' website texts and relate them in the tf-idf vector space (see cognitive proximity section above).

First, we created a training dataset for that classification task by sampling 5,000 random pairs of hyperlinked firms from our dataset. Subsequently we labelled each hyperlink as representing either a business or non-business relation. We were able to label 3,632 hyperlink connections unambiguously. Figure 2 shows that more than two thirds of the hyperlinks were labelled as *business relations* with only few of them being hyperlinks between firm of the same corporate group. *Non-business relations* on the other hand are of information only and legal/regulatory nature to about equal shares.

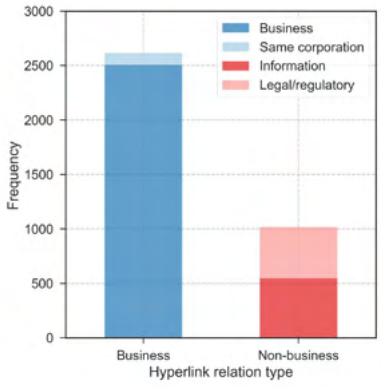


Fig. 2. Organizational proximity classes in training dataset. Manually labelled training dataset of hyperlinked firm pairs.

We then created numerical vectors for each hyperlinked firm pair by concatenating their respective tf-idf vectors. The resulting vectors have two times the dimension of our initial dictionary and effectively encode the texts of both firms. We tested several binary classifiers with these vectors and their corresponding labels from the training data and decided for a basic logistic regression classifier with balance class weights. For our classification task, the performance of the logistic regression classifier was overall superior in terms of accuracy and more balanced compared to more sophisticated binary classifiers which we tested (e.g. artificial neural networks and random forest). We trained the logistic regression classifier on two thirds of the labelled dataset and used one third (952 firms) as a test set to evaluate the performance of the model.

Table 2. Classification report for organizational proximity type prediction in the test set.

Label	Precision	Recall	f1-score	Support
Non-business	0.86	0.88	0.87	271
Business	0.95	0.94	0.95	681
Macro average	0.90	0.91	0.91	952
Weighted average	0.92	0.92	0.92	952
	Accuracy			
Overall	0.92			

Table 2 reports precision, recall, f1-score and accuracy of the trained model in the test set. The overall accuracy of 0.92 and an f1-score of 0.92 indicate a very good performance.

We used the trained model to predict the type of each of the 7,076,560 hyperlink connections in our dataset. The predictions range from 0.0 (high probability of business relation; small organizational distance) to 1.0 (high probability of non-business relation; large organizational distance). We summarized each firm's network by calculating the mean organizational distance over all its hyperlink connections.

5. Results

Figure 3 maps the Digital Layer of Germany which we created according to the procedure described in the previous section. The top panel of Figure 3 shows the distribution of product innovator firms in Germany (left) and Berlin (right) where the coloring of each cell gives the mean innovation probability for the companies contained in the respective cell. The middle panel shows the distribution of hyperlink connections in Germany (left) and Berlin (right). The lower panel shows the *ego* network of an exemplary firm (the Centre for European Economic Research) both for overall Germany (left) and for the Rhine-Neckar region (right) where the firm is located. The networks shown in Figure 3 were created using a graph bundling method based on kernel density estimation (34). Unsurprisingly, the density of hyperlink connections between any two areas seems to be highly dependent on population. However, Figure 3 is not intended to be of high analytical value but rather to give an overview of the dataset and its granularity.

Figure 4 shows kernel density estimations for all three types of firm-level proximity as well as for link count, mean partner innovation, and local firm density. The distribution of the normalized mean geographic proximity has a mean and a median of 0.28 (235 km) and follows a normal distribution with an over-proportional accumulation of observations at mean distance 0.0 (i.e. companies that maintain hyperlinks to other companies located in the same street). Mean cognitive distance and organizational distance follow a similar normal-like distribution with higher means (0.74 and 0.75) and medians (0.75 and 0.75). Considering mean cognitive distance, an over-proportional frequency of 0.0 observations can be seen (i.e. firms that share identical texts with their hyperlink partners). In the case of mean organizational distance, on the other hand, a high frequency of 1.0 values can be seen (i.e. a lot of of firms have partner networks that consist of only non-business relations). In table 1 we already saw that the distribution of link count is highly skewed. The mean link count is 13.01 and the median is 4, while the maximum link count in our dataset is 168,961 (the German branch of a major tech com-

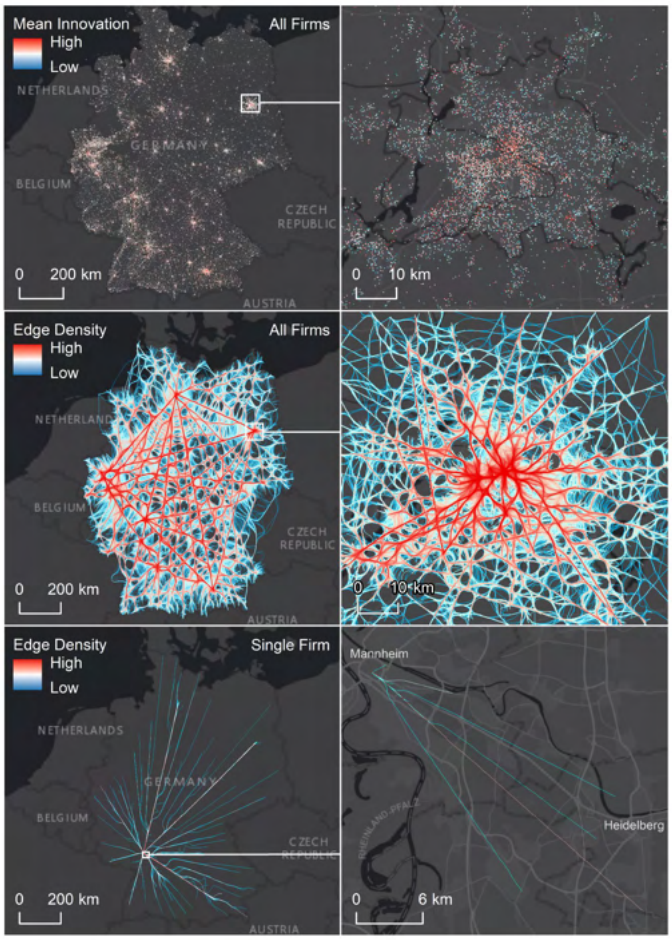


Fig. 3. The Digital Layer of Germany. Top row: Mean product innovator probability for Germany (left) and Berlin (right). Middle row: Hyperlink connections between firms in Germany (left) and Berlin (right). Bottom row: Hyperlink connections of a single firm observation in Germany (left) and the Rhine-Neckar region (right).

pany from the Silicon Valley). Mean partner innovation is again somewhat normal distributed with a mean of 0.36 and a median of 0.34. The distribution of the local firm density variable is very skewed again. On average, firms in our dataset have 176.8 other firms within one kilometer of their geographic location. The median is at 53 and the maximum value is 3,930 (downtown Hamburg).

Figure 5 shows the correlation table for all variables except for the *sector* variable which is categorical. The high correlation between the size of a company (*employees*) and its *age* is well known. However, there is also a strong positive correlation between firm size and the number of hyperlinked partners (*link count*) that a firm has. The *innovation* of firms shows a strong positive correlation to their hyperlinked partners' innovation (*mean partner innovation*). Having many partners (*link count*) is strongly negative correlated to *mean cognitive distance* (i.e. firms with many partners usually have similar partners). We also see a strong positive correlation between *mean geographic*

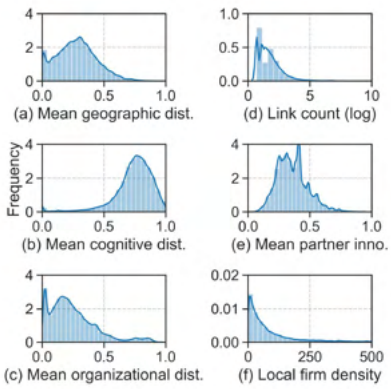


Fig. 4. Kernel density estimations for variables of interest. Geographical (a), cognitive (b), and organizational (c) distances. Link count (d), mean partner innovation (e), local firm density (f).

distance and *mean partner innovation* (i.e. firms with innovative partners maintain long-distance relationships). The strong negative correlation between *mean partner innovation* and *mean organizational distance* indicates that firms with innovative partners usually maintain stronger organizational ties.

Fig. 5. Correlation table. Correlation table with Spearman's correlation coefficients.

Figure 6 shows scatterplots and fitted regression lines of second order between innovation and our main variables of interest. We also tested regressions of third order which yielded only slightly different results. Both the number of partners of a firm (*link count*) and the mean innovation probability of these partners (*mean partner innovation*) show a strong positive and linear relation to the firm's own innovation probability. The relations between a firm's innovation probability and the mean cognitive and organizational distance to its hyperlink partners are both negative but less distinct. The mean geographical distance to a firm's partners as well as the local firm density show inverse-U shaped relationships to the firm's innovation probability.

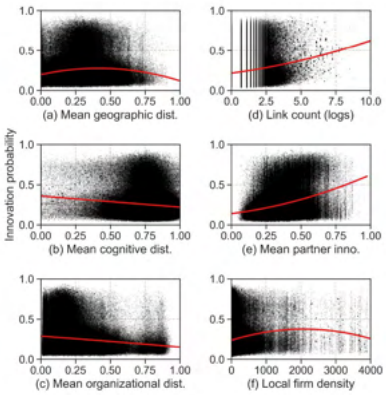


Fig. 6. Scatter plots for firm-level predicted innovation probability and variables of interest. Scatter plots and fitted regression lines of second order for geographical (a), cognitive (b), and organizational (c) distances, link count (d), mean partner innovation (e), and local firm density (f).

In Table 3, we present the results of an ordinary least square (OLS) regression with firm-level innovation as the dependent variable, control variables, and our variables of interest. The results are presented for both the innovation indicators from our *web dataset* (513,026 observations) and the *survey dataset* (2,405 observations) as a robustness check. Additionally, we also report the results of a fourth regression where we used the firms' statuses as patent-holders (1) or non-patent-holders (0). Concerning the web dataset, we run the regression for both the original continuous *product innovator probability* indicator and a binary recast (classification threshold 0.4). We use dummy variables to control for firms' sectors (*mechanical engineering* as baseline sector), size (number of employees; *missing* category as baseline), and age (*missing* category as baseline). We further control for the local firm density at the location of each firm.

6. Discussion

Our analysis revealed that innovative and non-innovative firms differ in terms of their network positions and hyperlink relations in the Digital Layer. Thereby our results are consistent as raw correlations (see Figure 5) and in our regression setting (see Table 3), which additionally controls for several firm characteristics. We find that innovative firms compared to

Table 3. Regression results

Variable	Web dataset (continuous y)	Web dataset (binary y)	Survey dataset (binary y)	Patent dataset (binary y)
		Constant		
Constant	0.2053***	-3.4465***	-2.6743	-5.5191***
		Firm-level controls		
Sector	Yes	Yes	Yes	Yes
Size	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes
Firm density (in 100)	0.0072***	0.0099***	0.0068	0.0021
Firm density (in 100) sq.	-0.0001***	-0.0002***	-0.0005**	-0.0002
		Hyperlink partners		
Link count (log)	0.0270***	0.0404***	0.0294***	0.0435***
Mean partner inno.	0.3036***	0.4602***	0.3803***	0.1745***
		Proximity		
Mean geo. distance	0.2404***	0.2688***	-0.1916	0.2455***
Mean geo. distance sq.	0.0260***	-0.0490**	-0.0966	-0.2399**
Mean cogn. distance	-0.1972***	-0.2045***	-0.0953	0.0284
Mean cogn. distance sq.	0.0733***	-0.0084	0.0398	0.0975*
Mean orga. distance	-0.4267***	-0.8022***	0.2706	0.0360
Mean orga. distance sq.	0.1151***	0.0994***	-0.5607	0.0652
		Proximity interactions		
Geo. dist. * orga. dist.	-0.0863***	-0.0377*	0.5660	0.0962
Geo. dist. * cogn. dist.	-0.2965***	-0.2566***	0.2122	-0.2845***
Cogn. dist. * orga. dist.	0.4326***	0.8583***	-0.1679	-0.1782
		Model statistics		
Model type	Robust OLS	Robust logit (average marginal effects)		
Observations	513,026	513,026	2,384	29,772
(Pseudo) R-squared	0.32	0.24	0.25	0.24
F-test/Wald chi2	3,187***	73,299***	379***	4,225***

non-innovative firms:

1. Have more hyperlinked partners.
2. Have partners that are more innovative.
3. Use geographic proximity to overcome cognitive distance to hyperlinked partners or use cognitive proximity to overcome geographic distance to their partners.

These findings are consistent for all of our used innovation datasets that we included as a robustness check. Finding *3* is consistent for the web dataset and the patent dataset but not for the survey dataset. Due to the comparatively small number of observations in the survey dataset, we were not able to identify statistically significant coefficients for our proximity measures in this dataset.

Link count. Previous studies like (13) and (11) found a positive relationship between the network centrality of firms and their innovation performance. Network centrality is equivalent to the number of hyperlink relations of each firm (*degree centrality*) in our study setup. In the scatter plots in Figure 6, we identified a strong positive and linear relationship between a firm's innovation probability and the number of hyperlinked partners. This positive relationship holds true when controlling for firm characteristics and other explanatory variables (see Table 3) and is consistent for all four datasets.

Mean partner innovation. All our results reveal a strong and positive relationship between a firm's innovation status and the mean product innovator probabilities of its hyperlinked partners, indicating that innovative firms are linked to other innovative firms. This is very much in line with the concept of *homophily* (14, 15), meaning that actors connect to other actors that are similar to them.

Firm density. (11) suggested that locally embedded firms have higher survival chances, but that the positive effect of embeddedness can reach a turning point, after which it reverses into a negative effect. If we assume that local firm density is a valid proxy for local firm embeddedness, our results in Figure 6 confirm the findings of (11) on a much larger scale. The existence of an optimal level of firm density is also revealed in the regression results for our web datasets. As the survey dataset is governed by a different sampling procedure and has very different descriptive statistics in terms of firm density (see Data section), we found no significant relationship between firm density and innovation for the survey dataset. On the basis of (16, 17)'s concept of "local buzz and global pipelines", we expected that successful firms are able to catch local knowledge flows (high local firm density) but maintain global pipelines (high mean geographical distance) to other innovative firms.

Mean geographical distance. Looking at the scatter plots in 6, we find that the mean geographical distance to hyperlinked partners has an optimum in its relation to a firm's innovation probability. As (35) explain, an optimum does not indicate that there is an optimal geographical distance but rather that a balanced level of local and non-local linkages to other companies generates an average distance that is most conducive to innovation. Concerning the regression results, we can confirm this for the patent dataset only, while we find a monotonically positive relationship between mean geographical distance and firm innovation in the web dataset.

Mean organizational distance. The raw correlations in Table 5 show a negative relation between mean organizational distance and innovation (i.e. innovative firms tend to form business instead of non-business relations). This negative relationship is also revealed in our regression results for both web datasets, where an increase in a firm's innovation probability is associated with a decline of its mean organizational distance in the variable range from 0.0 to 1.0. We assume business relationships to be closer than non-business relationships, because they are generally more formal and reoccurring. In this sense, it appears reasonable that knowledge flows and learning are more effective among organizationally close firms and go along with a higher innovativeness in the focal firm.

Mean cognitive distance. Both the raw correlations (see 6) and the regression results for our two web datasets reveal a negative relationship between cognitive distance and innovation within the value range of our dependent variable (0.0 to 1.0). This indicates that innovative companies connect to other companies that have a similar knowledge base (i.e. small *cognitive distance*). These findings are in in line with theory of (18) who argued that firms innovate in areas close to their own knowledge base. However, our measure for cognitive proximity has to be understood as a one-dimensional mapping of a high-dimensional process. There may be companies with quite different backgrounds (e.g. a software and a mechanical engineering company) that both participate in the same market (e.g. internet-of-things) and consequently share a similar knowledge base according to our measure for cognitive proximity. So our results might indicate that innovative firms and their partners share similar target markets rather that they are from the same sector.

We also found a negative relationship between innovation and an interaction term of cognitive and geographical distance for both our web datasets and the patent dataset. This may indicate that cooperation with cognitively distant companies can be successful (i.e. relate positively to firm innovation) when such partners are geographically close. It seems reasonable that geographic proximity helps to bridge knowledge gaps between dissimilar companies, for example by allowing for frequent face-to-face contact and the communication of tacit knowledge. Similarly, large geographical distances may not be hampering knowledge flows between partners if they share a common knowledge base which eases mutual understanding.

7. Conclusion

The Digital Layer. The aim of this study was to introduce a new approach to generate a web-based dataset of interrelated and textually described firms, the so-called *Digital Layer*. We constructed this Digital Layer by web mining the content of over half a million websites of German firms, resulting in a geolocated network with over seven million hyperlink relations. Making use of text-based machine learning models, we were able to operationalize proximity concepts that were difficult to analyze in large-scale empirical studies using other data sources. In a second step, we were able to empirically assess the relationship of these proximity measures and innovation

in firms. For this, we used three different firm-level innovation indicators: a traditional indicator from the questionnaire-based German Community Innovation Survey (CIS), a novel indicator generated from deep learning of website texts (4), and firm-level patent statistics. Our results showed that the Digital Layer is suitable for conducting large-scale analyses of firm networks that are not constrained to specific sectors, regions, or geographical levels of analysis.

Proximity and innovation. Our case study revealed that innovative firms are differently connected within the Digital Layer compared to non-innovative firms. We were able to confirm the results of previous studies, showing that innovative firms have more (hyperlinked) partners and that their partners are on average more innovative compared to the partners of non-innovative firms. Analogous to the theory of "local buzz and global pipelines" (16, 17), we found that innovative firms are located in high density areas and still maintain relations to firms that are geographically farther away. We were able to operationalize meaningful and convenient measures of geographical, organizational, and cognitive proximity from the Digital Layer. Our results indicate that close relationships are not necessarily related to higher firm innovativeness but that it rather depends "on the optimal level of proximity between agents" (8). We also found that the relation between innovation and proximity may be indeed rather complex and that different dimensions of proximity interact with each other.

Future research. We believe that the Digital Layer approach bears great potential for the empirical analysis of firm networks. As of now, we only have gathered data for one year, but we plan to reconstruct the hyperlink networks of previous years on the basis of web archive data and also to collect data in future years by continuing to gather web data using our presented approach. Such time series data would allow researchers to investigate innovation dynamics such as firm-to-firm knowledge spillovers and the diffusion of technology between firms, industrial sectors, and regions. The high level of granularity of the Digital Layer also allows for further analyses of microgeographical intraurban firm networks as well as the analysis of networks of cities. We also expect that the depth of the Digital Layer allows for many more studies in other economic or social science settings. Moreover, the Digital Layer can add meaningful insights to the research on the multilayered structure of corporate networks (36).

1. Schumpeter J (1942) *Capitalism, Socialism and Democracy.* (Harper & Brothers).
2. van Egeraat C, Kogler DF (2013) Global and regional dynamics in knowledge flows and innovation networks. *European Planning Studies* 21(9):1317–1322.
3. Boschma RA (2005) Proximity and innovation: A critical assessment. *Regional Studies* 39(1):61–74.
4. Kinne J, Lenz D (2019) Predicting Innovative Firms Using Web Mining and Deep Learning.
5. Castells M (1996) *The rise of the network society.* (Blackwell Publishers Cambridge, MA).
6. Hidalgo C (2015) *Why information grows: The evolution of order, from atoms to economies.* (Basic Books).
7. Ter Wal AL, Boschma RA (2009) Applying social network analysis in economic geography: Framing some key analytic issues. *The Annals of Regional Science* 43(3):739–756.
8. Boschma R, Frenken K (2009) The spatial evolution of innovation networks: A proximity perspective.
9. Giuliani E (2005) The structure of cluster knowledge networks: uneven and selective, not pervasive and collective in *DRUID Tenth Anniversary Summer Conference*. pp. 27–29.
10. Park HW (2003) Hyperlink network analysis: A new method for the study of social structure on the web. *Connections* 25:49–61.
11. Uzzi B (1996) The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review* pp. 674–698.
12. Gay B, Dousset B (2005) Innovation and network structural dynamics: Study of the alliance network of a major sector of the biotechnology industry. *Research policy* 34(10):1457–1475.
13. Giuliani E, Bell M (2005) The micro-determinants of meso-level learning and innovation: Evidence from a chilean wine cluster. *Research policy* 34(1):47–68.
14. Skvoretz J (1991) Theoretical and methodological models of networks and relations. *Social networks* 13(3):275–300.
15. Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American journal of sociology* 110(4):1132–1205.
16. Bathelt H, Malmberg A, Maskell P (2004) Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in human geography* 28(1):31–56.
17. Asheim BT, Isaksen A (2002) Regional innovation systems: the integration of local 'sticky' and global 'ubiquitous' knowledge. *The Journal of Technology Transfer* 27(1):77–86.
18. Winter SG, Nelson RR (1982) An evolutionary theory of economic change. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.*
19. Nooteboom B, Van Haverbeke W, Duysters G, Gilsing V, Van den Oord A (2007) Optimal cognitive distance and absorptive capacity. *Research policy* 36(7):1016–1034.
20. Breschi S, Lissoni F (2006) Mobility of inventors and the geography of knowledge spillovers: new evidence on us data. *KITeS Working Papers* (184).
21. Bersch J, Gottschalk S, Müller B, Niefert M (2014) The Mannheim Enterprise Panel (MUP) and firm statistics for Germany.
22. Kinne J, Axenbeck J (2018) Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany.
23. Rammer C, et al. (2019) Innovationen in der deutschen Wirtschaft, (ZEW Centre for European Economic Research, Mannheim), Technical report.
24. Zandbergen PA (2008) A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32(3):214–232.
25. Rammer C, Kinne J, Blind K (2019) Knowledge Proximity and Firm Innovation: A Microgeographic Analysis for Berlin. *Urban Studies* forthcomin.
26. Kinne J (2018) ARGUS - An Automated Robot for Generic Universal Scraping.
27. Gault F, Aho E, Alkio M, Arundel A, Bloch C (2013) *Handbook of Innovation Indicators and Measurement* ed. Gault F. (Edward Elgar Publishing Ltd, Glos, UK), p. 486.
28. OECD, Eurostat (2018) *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation.* (OECD/eurostat, Luxembourg, Paris), 4th edition, p. 258.
29. Gök A, Waterworth A, Shapira P (2015) Use of web mining in studying innovation. *Scientometrics* 102(1):653–671.
30. Manning CD, Raghavan P, Schutze H (2009) *An Introduction to Information Retrieval.* (Cambridge University Press, Cambridge, England), Online edi edition, p. 569.
31. Rahimi S, Mottahedi S, Liu X (2018) The Geography of Taste: Using Yelp to Study Urban Culture. *ISPRS International Journal of Geo-Information* 7(9):376.
32. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space.
33. Grentzkow M, Kelly BT, Taddy M (2017) Text as Data.
34. Hurter C, Ersoy O, Telea A (2012) Graph Bundling by Kernel Density Estimation. *Computer Graphics Forum* 31(3pt1):865–874.
35. Boschma R, Frenken K (2010) The spatial evolution of innovation networks: A proximity perspective in *The Handbook of Evolutionary Economic Geography.* (Edward Elgar Publishing).
36. de Jeude JvL, Aste T, Caldarelli G (2019) The multilayer structure of corporate networks. *New Journal of Physics* 21(2):025002.

in firms. For this, we used three different firm-level innovation indicators: a traditional indicator from the questionnaire-based German Community Innovation Survey (CIS), a novel indicator generated from deep learning of website texts (4), and firm-level patent statistics. Our results showed that the Digital Layer is suitable for conducting large-scale analyses of firm networks that are not constrained to specific sectors, regions, or geographical levels of analysis.

Proximity and innovation. Our case study revealed that innovative firms are differently connected within the Digital Layer compared to non-innovative firms. We were able to confirm the results of previous studies, showing that innovative firms have more (hyperlinked) partners and that their partners are on average more innovative compared to the partners of non-innovative firms. Analogous to the theory of "local buzz and global pipelines" (16, 17), we found that innovative firms are located in high density areas and still maintain relations to firms that are geographically farther away. We were able to operationalize meaningful and convenient measures of geographical, organizational, and cognitive proximity from the Digital Layer. Our results indicate that close relationships are not necessarily related to higher firm innovativeness but that it rather depends "on the optimal level of proximity between agents" (8). We also found that the relation between innovation and proximity may be indeed rather complex and that different dimensions of proximity interact with each other.

Future research. We believe that the Digital Layer approach bears great potential for the empirical analysis of firm networks. As of now, we only have gathered data for one year, but we plan to reconstruct the hyperlink networks of previous years on the basis of web archive data and also to collect data in future years by continuing to gather web data using our presented approach. Such time series data would allow researchers to investigate innovation dynamics such as firm-to-firm knowledge spillovers and the diffusion of technology between firms, industrial sectors, and regions. The high level of granularity of the Digital Layer also allows for further analyses of microgeographical intraurban firm networks as well as the analysis of networks of cities. We also expect that the depth of the Digital Layer allows for many more studies in other economic or social science settings. Moreover, the Digital Layer can add meaningful insights to the research on the multilayered structure of corporate networks (36).

1. Schumpeter J (1942) *Capitalism, Socialism and Democracy.* (Harper & Brothers).
2. van Egeraat C, Kogler DF (2013) Global and regional dynamics in knowledge flows and innovation networks. *European Planning Studies* 21(9):1317–1322.
3. Boschma RA (2005) Proximity and innovation: A critical assessment. *Regional Studies* 39(1):61–74.
4. Kinne J, Lenz D (2019) Predicting Innovative Firms Using Web Mining and Deep Learning.
5. Castells M (1996) *The rise of the network society.* (Blackwell Publishers Cambridge, MA).
6. Hidalgo C (2015) *Why information grows: The evolution of order, from atoms to economies.* (Basic Books).
7. Ter Wal AL, Boschma RA (2009) Applying social network analysis in economic geography: Framing some key analytic issues. *The Annals of Regional Science* 43(3):739–756.
8. Boschma R, Frenken K (2009) The spatial evolution of innovation networks: A proximity perspective.
9. Giuliani E (2005) The structure of cluster knowledge networks: uneven and selective, not pervasive and collective in *DRUID Tenth Anniversary Summer Conference*. pp. 27–29.
10. Park HW (2003) Hyperlink network analysis: A new method for the study of social structure on the web. *Connections* 25:49–61.
11. Uzzi B (1996) The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review* pp. 674–698.
12. Gay B, Dousset B (2005) Innovation and network structural dynamics: Study of the alliance network of a major sector of the biotechnology industry. *Research policy* 34(10):1457–1475.
13. Giuliani E, Bell M (2005) The micro-determinants of meso-level learning and innovation: Evidence from a chilean wine cluster. *Research policy* 34(1):47–68.
14. Skvoretz J (1991) Theoretical and methodological models of networks and relations. *Social networks* 13(3):275–300.
15. Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American journal of sociology* 110(4):1132–1205.
16. Bathelt H, Malmberg A, Maskell P (2004) Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in human geography* 28(1):31–56.
17. Asheim BT, Isaksen A (2002) Regional innovation systems: the integration of local 'sticky' and global 'ubiquitous' knowledge. *The Journal of Technology Transfer* 27(1):77–86.
18. Winter SG, Nelson RR (1982) An evolutionary theory of economic change. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.*
19. Nooteboom B, Van Haverbeke W, Duysters G, Gilsing V, Van den Oord A (2007) Optimal cognitive distance and absorptive capacity. *Research policy* 36(7):1016–1034.
20. Breschi S, Lissoni F (2006) Mobility of inventors and the geography of knowledge spillovers: new evidence on us data. *KITeS Working Papers* (184).
21. Bersch J, Gottschalk S, Müller B, Niefert M (2014) The Mannheim Enterprise Panel (MUP) and firm statistics for Germany.
22. Kinne J, Axenbeck J (2018) Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany.
23. Rammer C, et al. (2019) Innovationen in der deutschen Wirtschaft, (ZEW Centre for European Economic Research, Mannheim), Technical report.
24. Zandbergen PA (2008) A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32(3):214–232.
25. Rammer C, Kinne J, Blind K (2019) Knowledge Proximity and Firm Innovation: A Microgeographic Analysis for Berlin. *Urban Studies* forthcomin.
26. Kinne J (2018) ARGUS - An Automated Robot for Generic Universal Scraping.
27. Gault F, Aho E, Alkio M, Arundel A, Bloch C (2013) *Handbook of Innovation Indicators and Measurement* ed. Gault F. (Edward Elgar Publishing Ltd, Glos, UK), p. 486.
28. OECD, Eurostat (2018) *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation.* (OECD/eurostat, Luxembourg, Paris), 4th edition, p. 258.
29. Gök A, Waterworth A, Shapira P (2015) Use of web mining in studying innovation. *Scientometrics* 102(1):653–671.
30. Manning CD, Raghavan P, Schutze H (2009) *An Introduction to Information Retrieval.* (Cambridge University Press, Cambridge, England), Online edi edition, p. 569.
31. Rahimi S, Mottahedi S, Liu X (2018) The Geography of Taste: Using Yelp to Study Urban Culture. *ISPRS International Journal of Geo-Information* 7(9):376.
32. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space.
33. Grentzkow M, Kelly BT, Taddy M (2017) Text as Data.
34. Hurter C, Ersoy O, Telea A (2012) Graph Bundling by Kernel Density Estimation. *Computer Graphics Forum* 31(3pt1):865–874.
35. Boschma R, Frenken K (2010) The spatial evolution of innovation networks: A proximity perspective in *The Handbook of Evolutionary Economic Geography.* (Edward Elgar Publishing).
36. de Jeude JvL, Aste T, Caldarelli G (2019) The multilayer structure of corporate networks. *New Journal of Physics* 21(2):025002.