Eurasian Business Review
https://doi.org/10.1007/s40821-024-00270-x

REGULAR ARTICLE

# So far, yet so close. Using networks of words to measure proximity and spillovers between firms

**Alessandro Marra**[1,2] · **Marco Cucculelli**[3,5] · **Alfredo Cartone**[1,4]

Received: 8 September 2023 / Revised: 12 February 2024 / Accepted: 7 May 2024


## Abstract

Textual data are the last frontier in the empirical literature on proximity between firms. While there are a growing number of studies using textual data, no robust methodology has yet emerged, nor has any attempt been made to compare the resulting findings with standard measures of proximity based on existing classification systems. The purpose of this paper is threefold. First, we propose a methodology that can be an effective and applicable tool for measuring proximity between companies. Second, we compare the resulting indicator of proximity, which we refer to as "business" proximity, with industrial and technological proximity scores based on activity codes and technology adoption, respectively. Third, we use business proximity to explain economic performance, assuming that knowledge sharing can occur between employees working in similar firms. Having established the soundness of the methodology, the empirical results confirm the substantial information content of the descriptive texts and provide evidence on the likelihood of spillover effects between firms that are close in the business and geographical dimension.

**Keywords** Industrial proximity · Technological proximity · Business proximity · Knowledge exchange · Firm performance

✉ Marco Cucculelli
m.cucculelli@univpm.it

[1] Dipartimento Di Economia, Università G. d'Annunzio Di Chieti E Pescara, Pescara, Italy

[2] Explo Academic Spinoff, Università G.D'Annunzio di Chieti e Pescara, Pescara, Italy

[3] Dipartimento Di Scienze Economiche E Sociali, Università Politecnica Delle Marche, Ancona, Italy

[4] Assistant Professor in Economic Statistics at "G. d'Annunzio" University of Chieti-Pescara funded by the program "Research contracts on innovation and green topics", FSE-REACT-EU Project by the European Commission-National Operational Program "PON Ricerca & Innovazione 2014-2020-DM 1062/2021 (D25F21001470007)", "G. d'Annunzio" University of Chieti-Pescara, Chieti-Pescara, Italy

[5] Fondazione Giorgio Fuà, Ancona, Italy

Published online: 26 June 2024

## 1 Introduction

Measuring proximity between companies is a task that is receiving increasing attention in the academic and policy worlds. From an operational perspective, it consists of two steps: choosing a classification system to define "what firms do" and estimating a score of proximity. Intuitively, two or more companies can be considered close to each other if, for example, they carry out similar industrial activities. This is a common measure of proximity in the literature, referred to as "industrial" relatedness or proximity. Many alternative measures can be proposed, changing the underlying classification system and purpose: firms can be connected based on alternative definitions of proximity (within a virtual network of firms or in an artificial multidimensional space) if, for example, they share common knowledge ("cognitive" proximity) or use the same technologies ("technological" proximity).

These two steps can be followed by a third step aimed at explaining the economic performance of firms by examining spillover effects between neighbouring firms: namely, "similar" firms are more likely to collaborate, exchange ideas, build 'new' knowledge, innovate and, hence, their growth may depend on the performance of their neighbours (Boschma, 2005; Nooteboom, 2003). The general proposition that greater proximity between firms can lead to more knowledge sharing and better performance has been revised: proximity can also lead to negative outcomes. This happens not only when proximity is "too low", but also when it is "too high" (Boschma, 2005; Nooteboom et al., 2007). Similarly, it is the perspective by which some absorptive capacity is needed to identify, interpret, and exploit new knowledge (Cohen & Levinthal, 1990).

Until a few years ago, the first step was accomplished by relying on existing classification systems. Thus, firms were usually classified either by industry codes (or activity codes), or by the product code assigned to the goods and services they offered, or the technological class to which their patents were related (Boschma & Gianelle, 2014). This approach was convenient for researchers and analysts but proved rather unsatisfactory for policy makers. To illustrate this point, we can consider industrial proximity based on activity codes: knowing that some firms belong to a particular sector (e.g. software) is not particularly useful for policy purposes if we cannot identify correctly their current specialisations. For example, it is possible that some software companies are currently developing software for CRM tools, while others are designing solutions for the Internet of Things, and still others involved in web design. Accordingly, it would be advisable to make a more differentiated classification that takes these specificities into account and enables the definition of clearly targeted actions. In other words, a correct and quick classification of companies can lead to more effective policies, as policy makers value timely and accurate information on the industrial composition of the economy.

The classification system based on activity codes is the most commonly used today. In the European Union, it takes the form of the Nomenclature of Economic Activities (Nace), created by Eurostat in 1970 and translated in Italy by the National Institute of Statistics (Istat) into the recently updated Ateco codes. Although they represent a well-established taxonomy, the activity codes have many limitations that are widely discussed in the literature (Nathan & Rosso, 2015; Papagiannidis et al.,

2017). Activity codes are often too strict and binding, lag far behind market developments and are selected on the basis of the "main economic activity," leading to the neglect of all other (secondary or ancillary) activities. Once companies are classified according to the economic activity codes, proximity can be measured.

Not all proximity scores are limited to the industrial sector of companies. Increasing attention is paid to firms' innovation capabilities and technology components. Therefore, companies can also be classified on the basis of their R&D activities by using the technology class in which their patent applications are filed. However, like activity codes, technology classes also have their limitations. For example, many companies do not file patents, which means that a measure of technological proximity based on patent classification cannot be determined for some companies. This is also the reason why in this paper we consider the adoption of new technologies (instead of patent classes) as the marker to capture the similarity of firms and then assess their technological proximity.

Recently, an unprecedented wave of technological change has had a profound impact on companies. On the one hand, due to technological progress and digitalization, standard classifications are less and less suitable for capturing the rapidly changing activities and products of companies; on the other hand, thanks to the emergence of big data, digital records can provide a wealth of information that can be collected and processed to classify companies. Several recent contributions take advantage of this data availability and propose original methods for classifying and analysing companies based on textual data. The latter allow researchers to capture crucial aspects of companies' actual operations, specialisations, products and markets that allow for more accurate comparisons between firms and more effective measures of proximity. When the data are in the form of descriptive text (e.g., descriptions of companies, products and services, and innovative research and development), the risk of simplification associated with standard classification systems is replaced by the complexity associated with developing a new and original system.

Although the use of textual data in empirical research is increasing, a sound methodology has not yet emerged; also, no attempts have been made to compare the results obtained from new methods with those obtained using standard measures of proximity based on existing classification systems. Moreover, many potential uses of the network approach are still unexplored. Given these gaps, the purpose of this paper is threefold. First, we propose a methodology that can be a valid, effective and immediately applicable tool for measuring inter-firm proximity. Second, we compare the resulting measure of proximity, which we refer to as "business" proximity (based on descriptive text on company websites), with industrial proximity (based on activity codes) and technological proximity (based on the adoption of new technologies by companies). Third, we propose to use business proximity in modelling economic performance to capture spillover effects due to knowledge sharing between similar firms.

The approach adopted consists of three stages. In the first stage, descriptive texts about the firms' activities are collected from the company websites, and keywords are created, organized into categories, and assigned to the companies. In the second stage, the keywords are used to create networks of words to measure proximity and compare with other scores. To build the networks for industrial and technological

proximity, we use data from local Chambers of Commerce and information from an ad hoc survey conducted for research purposes. In the last stage, we propose a model for firm performance that accounts for business proximity using spatial econometrics tools.

We apply our framework to a sample of 553 firms operating mainly in the manufacturing and service sectors and located in an Italian NUTS 2 region, the Marche region. This region is a relatively small but dynamic economy with a highly diversified industrial structure that, thanks to the variety of firm specialisations, technologies, competences, and knowledge, represents an ideal target for testing the effectiveness of the proposed methodology. The results highlight the advantages and disadvantages of this novel approach and provide evidence on the likelihood of spillover effects between neighbouring companies.

## 2 Background

It is well known that the most commonly used method to measure the similarity between companies is the one that evaluates "proximity" within the hierarchical structure of the classification of industrial activities. According to this method, two or more companies that share neighbouring codes are similar and therefore close within the network or observed space. Several studies have used this approach. Frenken et al. (2007) play a key role in this literature: using industry codes, the authors measure "related variety" as the average entropy of employment levels in five-digit sectors within each two-digit class and "unrelated variety" as the average entropy of employment levels in two-digit classes. Applying these measures to 40 NUTS3 areas in the Netherlands, the authors show that related variety positively affects employment growth, while unrelated variety is negatively correlated with unemployment growth. The analysis by Frenken and co-authors has stimulated a considerable strand of literature on the impact of related variety on economic development (Caragliu et al., 2016; Cortinovis & Van Oort, 2015; Falcioglu, 2011; Hartog et al., 2012; Quatraro, 2010; van Oort et al., 2015). However, the taxonomies used as a starting point have several limitations. First, the descriptions of the codes are too brief to capture the actual scope of activities of the observed companies. Second, although the codes are numerous, they are too limited in number to reflect the richness and diversity of existing business activities (Piore & Sabel, 1984). Third, such codes are often too strict and binding and lag far behind market developments (Feldman et al., 2005). Fourth, codes are selected on the basis of the "most important" economic activity, which leads to all other business activities being neglected (and thus information being lost) (Porter, 1998). Fifth, it should be noted that companies often opt for residual or "broader" (i.e. less informative) activity codes in order not to be limited to certain (too strict) areas of activity (OECD, 2013). Finally, organisations are constantly changing: the activity code chosen yesterday may only partially reflect what the organisation does today.

### 2.1 Measuring proximity between firms

To overcome these limitations, some researchers supplement the activity codes with information from additional sources, such as data on workflows (Boschma et al., 2009; Fitjar & Timmermans, 2017; Neffke & Henning, 2013; Timmermans & Boschma, 2014), input–output relationships (Boschma & Iammarino, 2009; Fan & Lang, 2000), or the co-occurrence of products in goods and services portfolios (Hidalgo et. al, 2007; Neffke & Henning, 2008; Neffke et al., 2011). In this way, they obtain more accurate measures of the proximity between firms.

Boschma et al. (2009) and Timmermans and Boschma (2014) use cross-industry work flows to measure proximity between groups of firms and assess the impact of hiring new employees. Fitjar and Timmermans (2017) compare the proximity measure developed by Frenken et al. (2007) with a new measure that uses labour mobility flows to determine the proximity between industries. The authors conduct an analysis of the Norwegian labour market and show that the two measures are highly correlated. Nevertheless, their index based on labour flows is better suited to identify proximity.

In addition, there are several papers that measure proximity between firms or groups of firms based on input–output linkages. Fan and Lang (2000) use data on commodity flows from the US input–output tables to construct a measure of proximity that captures the vertical correlation between industries. Their results show that the new input/output-based measure outperforms traditional measures based on activity codes. Other papers using this methodology include those by Saviotti and Frenken (2008), Boschma and Iammarino (2009), Cainelli and Iacobucci (2012) and Essleztibichler (2015).

In another method based on companies' products/services, proximity is measured by the co-occurrence of products in the portfolios of goods and services exported by countries (Hidalgo et al., 2007), regions (Neffke et al., 2011) and production sites (Neffke & Henning, 2008). Hidalgo et al., (2007) introduce the concept of product space, in which each product is close to another according to the frequency with which two products co-occur in countries' export portfolios. The method of co-occurrence of products is also used by Neffke and Henning (2008). In their paper, the authors measure the proximity between industries using a dataset of product portfolios manufactured in several Swedish factories: if two products are manufactured in the same factory, the industries to which the two products can be assigned can be considered close. A similar approach is used by Neffke et al. (2011) to analyse the proximity between industries.

Many studies focus instead on technological proximity. As expected, the technology class sharing methodology developed by Jaffe (1986) draws on the classification of firm patents to construct a measure of proximity in the technology space. This work was followed by several papers (Aldieri, 2013; Cantner & Meder, 2007; Petruzzelli, 2011; Schildt et al., 2005). Again, the existence of an extensive literature seems to confirm the soundness of the methodology, but the chosen taxonomy is characterised by several rigidities. First, some firms only hold patents that were granted a long time ago, which may contribute to a partial misrepresentation of the positioning of companies in the technology space. Second, the technology classes

are rigid categories based on definitions that are sometimes too general to capture the innovation content and scope of the underlying R&D activity. Finally, not all companies have patents, which means that not all firms are eligible for this analysis.

In addition to the technology classes, proximity can also be measured using patent citations. This methodology was introduced by Jaffe et al. (1993) and followed by several papers (e.g. Fischer et al., 2006; Fung, 2003; Mowery et al., 1998; Oikawa, 2017; Stuart & Podolny, 1996). In particular, Stuart and Podolny (1996) propose a network analysis approach to identify the development of the technological positioning of companies. This approach enables the mapping and clustering of firms based on similarities in innovation capacity and technology portfolio. Mowery et al. (1998) also use patent citations to measure the technological overlaps between companies and to evaluate the choice of technology partner, showing the influence of the technology portfolio on companies' decisions. Proxy measures based on patent citations are frequently used in the literature today. However, this methodology requires extensive citation flows between companies, which is computationally very expensive.

### 2.2 Using textual data

The main goal of the literature briefly described above is to propose coherent approaches to measuring corporate proximity. While most approaches are considerable, they are time-consuming and can only be implemented on a case-by-case basis (Boschma & Gianelle, 2014). Advances in text mining and text analytics have enabled the effective processing of information from corporate statutes, official documents and descriptive texts about companies' activities, products and services, patents and more (Gentzkow et al., 2019). This increasing availability of textual data is fuelling a promising strand of literature and opening up valuable research opportunities in regional economics, innovation economics and operational research.

Despite some technical difficulties, textual data allow researchers to classify firms based on more consistent and up-to-date information. To cite a few examples, Nathan and Rosso (2015) conduct a study of the information and communication technology (ICT) industry to determine which and how many companies operate in this sector in the UK. Using text mining, the authors create an industry-product map. The results show a more accurate and detailed mapping of companies operating in the ICT sector compared to a mapping based on industry classification codes, which can also suffer from lag as these codes are not updated in light of the continuous evolution of the most innovative sectors. Hoberg and Philips, (2016) apply text analysis to company product descriptions filed with the Securities and Exchange Commission (SEC). Based on the results of the text analysis, the authors construct an industry classification to capture the proximity between companies. Basically, proximity is measured by the similarity of the words used in the descriptions of the products and services offered. The advantages of this approach are that it overcomes the rigidity and limitations typical of traditional industry classifications and provides greater granularity in the resulting taxonomy. In addition, Shi et al. (2016), who develop a measure of dyadic proximity between companies, use the topic modelling technique to analyse unstructured, online textual data describing company activities.

The advantages of this approach are scalability with large data sets and accuracy in positioning companies in the product, market and technology domain.

With the aim of overcoming the aforementioned shortcomings of activity codes, Papagiannidis et al. (2017) apply data mining techniques and narrow down the business of each observed company by identifying multiple industry clusters in the geographic space. The identified industry clusters are the result of a more accurate proximity measure. Marra et al., (2017) perform text mining to analyze a sample of green tech firms in San Francisco, New York, and London: using metadata (i.e., information about firms' products, services, markets, and technologies) collected from Crunchbase, the authors classify firms' industrial activities and underlying specialisations, establish links for technological and market complementarities, and identify specific firm aggregations and emerging industrial clusters. Pavone and Russo (2017) apply text analysis techniques to the social objects of firms operating in the automotive supply chain in Italy. The authors calculate the proximity between companies based on the co-occurrence of keywords from company texts and create a classification of existing specialisations. Losurdo et al. (2019) measure the specialisations and competences of innovative digital companies based on the degree of digital technologies in the products and services offered. Their method makes it possible to overcome the limitations of defining industry specialisations in digital industries through SIC codes and to capture the specialisations of innovative companies at the metropolitan level. Marra et al. (2020) examine companies operating in the clean tech sector with the aim of providing methodological support for screening potential partners and helping companies to identify with whom they can collaborate and share knowledge. The authors apply a network analysis in which the proximity between two companies is determined not only by sharing one or more descriptive keywords of the companies' activities, but also on the basis of the presence of the keyword pairs that two companies have within the keyword vectors of each company in the observed population.

Kinne and Resch (2018) develop a novel approach to evaluate the innovation activity of companies based on textual data from company websites. They use automated web scrapers to collect text from websites, then extract semantic topics in a self-learning generative topic modelling approach and analyse these topics with a neural network method that evaluates the degree of innovation of each company. Similarly, Kinne and Lenz (2021) apply an artificial neural network model to the web texts of hundreds of thousands of companies in Germany by applying a previously trained model to web texts of companies identified as innovative and comparing the results with traditional indicators. According to the authors, this approach enables the creation of innovation indicators based on web data and is scalable. Bishop et al. (2022) conduct a bottom-up study with the aim of overcoming the limitations of industry classifications and examining the composition of the economy. By applying machine learning techniques and graph theory, the authors analyse business descriptions from corporate websites and create alternative taxonomies based on which they define industries as "communities within word networks". Marra and Baldassari (2022) use text mining and semantic algorithms to label innovative companies and provide an alternative perspective on classifying industrial activities. The results help the understanding of what companies do in a more effective and up-to-date way

than the standard industrial classification codes. In addition, the authors emphasise that by matching company keywords, the degree of closeness between companies can be examined in more detail.

Textual analysis is also widely used in the literature on technological proximity. There are numerous studies that use words and sentence structures in patent descriptions to create maps and classifications of companies. A widely used approach for patent textual analysis is the one based on subject-action-object (SAO) structures (Gerken & Moehrle, 2012; Park et al., 2013). Such structures form syntactically ordered sentences that can be automatically extracted from the processing of patent texts. For example, Gerken and Moehrle (2012) estimate technological proximity based on the number of semantically identical structures. Using patent data collected for the period between 1976 and 2006, the authors show that the method used is a valid solution for identifying highly novel inventions. Park et al. (2013) propose a framework that evaluates companies from a technological perspective to support target selection in merger and acquisition (M&A) decisions. The authors use patent textual analysis to create a map that can be used to identify companies suitable for strategic M&A and improve their technological capabilities. Finally, a similar analysis was conducted by Yoon and Kim (2012) and Yoon et al. (2013).

### 2.3 Measuring firm performance

The empirical literature on the relationship between firm proximity, knowledge spillover and firm performance is extensive (Raspe and Oort, 2008; Fallah et al., 2014; Ejdemo and Örtqvist, 2020; Aldieri et al., 2020). For example, performance can be estimated in different ways: scores of innovative activities, productivity and growth rates of companies in terms of number of employees or sales volume. Moreover, many contributions focus on a variety of available proximity measures. In this paper, we essentially aim to contribute by focusing on how business proximity may favour knowledge sharing and affect firm growth.

The chosen framework was inspired by several papers. Boschma et al. (2009) apply a methodology based on labor flows to measure the proximity between groups of companies and assess the impact of a new employee joining a firm. The authors find a positive impact on firm performance when the labour flow is between neighbouring firms in terms of skills. Furthermore, the effect is negative when the employee moves to a company that already possesses identical skills. Timmermans and Boschma (2014) also test the same framework in Denmark. The study shows a positive impact on the productivity growth of the firm when there is an inflow of skills closely related to those within the plant. Conversely, the inflow of skills that are similar to those already present in the plant has a negative effect. Aarstad et al. (2016) find that related variety is a positive driver of firm innovation, while unrelated variety is a negative driver of firm productivity. The results suggest that regions with high related variety and low unrelated variety are optimal for firm performance. Cainelli and Ganau (2019) find that related variety has a positive effect on employment growth. More specifically, the authors show that related variety posits knowledge spillovers between firms in different but related sectors, while firm

heterogeneity suggests spillovers between different firms in the same local sector. The results suggest a positive effect for both factors, with firm heterogeneity within a sector having a larger impact.

Thus, the underlying mechanism of our performance model is that similar firms (i.e., firms with similar specialisation, employing workers with comparable skills and knowledge, and using the same technologies) engage in processes of cooperation, competition or imitation and can therefore grow more (Wales et al., 2013; Choi and Williams, 2014). Therefore, the relationship between knowledge exchange and our measure of business proximity will assume an inverted U-shape when modelling performance (Fig. 1).

## 3 Data

Most studies on corporate behaviour focus on large, listed companies: for such companies, extensive data is available to assess the market's reaction to major changes in the company's business behavior or to understand how companies compete in the market. However, this information is not readily available for the majority of the firms operating in the small business sector: in this case, except for corporate accounts, publicly available data do not usually report the major factors affecting the business conduct, and most data can only be collected through direct interviews. With these considerations in mind, we have constructed a dataset by matching two complementary sources, namely a Bureau van Dijk (BvD) dataset (Aida) consisting of company financial statements and a cross-sectional survey dataset based on first-hand information collected directly from companies using questionnaire-based interviews.

Italian companies active in manufacturing and industry-related sectors according to the Ateco/NACE Rev. 2 classification and with between 10 and 1,000 employees in 2019 were selected for the survey. Using the Aida-BvD database, the number of Italian companies that met both criteria amounted to 55,124.

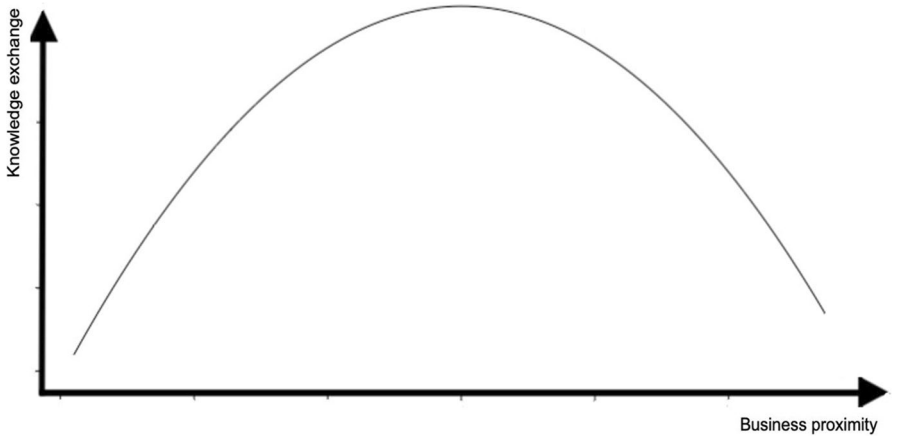


**Fig. 1** The U-inverted curve between business proximity and knowledge exchange

In order to obtain more accurate and reliable answers, we hired a team of postgraduate students with professional experience in the field who contacted the entrepreneurs or managers of the companies and assisted them in answering. The companies were contacted either by telephone (companies with 50 or more employees) or by email (companies with less than 50 employees). With the first method, the analysts were able to establish direct, effective and tailored communication, providing useful information and clearing up any doubts. In the second method, companies were sent an email with information about the questionnaire and the link to the online platform. To ensure a higher response rate, reminder emails were sent to the companies that had not completed the questionnaire within the set deadlines. At the end of the data collection (March 2020), we had 16,492 questionnaires (out of the original sample of 55,124 companies); the response rate was therefore 29.9%. We then checked the responses and discarded the questionnaires that were significantly incomplete (n. = 2,705) or contained potentially unreliable information, as well as duplicates and companies that have since exited the market (n. = 5,278), resulting in 8,509 usable questionnaires.

All companies based in the Marche region were selected from this sample. The Marche region forms the eastern coast of Central Italy and borders both the Emilia-Romagna region (north) and the Abruzzo region (south). The region's GDP amounts to more than 42 billion euros, which corresponds to 2.4% of the national GDP. With a production model focused on manufacturing and characterised by the widespread presence of industrial districts, Marche has been hit hard by the economic recession over the last 15 years. Nevertheless, it has maintained a strong industrial and manufacturing character: 30% of value added is generated by industry, compared to the Italian average of 24%. The largest contribution to value added comes from the tertiary sector at 67%, while agriculture only accounts for 2%. The region's industrial and economic system has some peculiarities, such as an average small company size, a remarkable capacity for innovation, a moderate attraction to skilled labour and a remarkable ability to penetrate international markets (Banca d'Italia, 2021). For these reasons, Marche constitutes an interesting case study and a suitable target for our analysis.

We included in the sample all companies that have their headquarters in the region (n = 824). After some trimming and the exclusion of companies for which no reliable survey or financial data was available, 624 companies were included in the sample. The sample was representative in terms of provincial distribution (five provinces) and sector coverage. We checked the representativeness of the sample by comparing the firms' distribution with i) the distribution observed in our initial dataset collected from Aida BvD and ii) the distribution based on national data on Italian manufacturing firms compiled by the Italian National Statistical Office (ISTAT), which are supposed to convey the most exhaustive picture. As the analysis was restricted to companies with a company website, the final dataset comprises 553 companies, most of which belong to the manufacturing and production services sectors (Table 1).

Looking at the geographical distribution, we find that 34.4% of the companies are located in the province of Ancona, 20.4% in the province of Pesaro Urbino, 17.7% in the province of Macerata, 15.2% in the province of Ascoli Piceno and 12.3% in the province of Fermo (Fig. 2).

**Table 1** Frequency of firms by Ateco 2-digit class

| Ateco 2-digit class | Frequency |
|---|---|
| 46—Wholesale trade, except of motor vehicles and motorbikes | 8.7% |
| 15—Manufacture of leather and related products | 8.6% |
| 25—Manufacture of fabricated metal products, except machinery and equipment | 8.6% |
| 62—Computer programming, consultancy, and related activities | 6.1% |
| 43—Specialised construction activities | 5.2% |
| 28—Manufacture of machinery and equipment n.e.c | 4.2% |
| 31—Manufacture of furniture | 4.0% |
| 47—Retail trade, except of motor vehicles and motorbikes | 4.0% |
| 22—Manufacture of rubber and plastic products | 3.4% |
| 14—Manufacture of wearing apparel | 3.2% |
| 26—Manufacture of computer, electronic, and optical products | 3.2% |
| 74—Other professional, scientific, and technical activities | 2.5% |
| 71—Architectural and engineering activities; technical testing and analysis | 2.2% |
| 27—Manufacture of electrical equipment | 2.0% |
| 70—Activities of head offices; management consultancy activities | 2.0% |
| 33—Repair and installation of machinery and equipment | 1.5% |
| 41—Construction of buildings | 1.5% |
| 49—Land transport and transport via pipelines | 1.5% |
| 63—Information service activities | 1.5% |
| 10—Food industries | 1.3% |
| 55—Accommodation service activities | 1.3% |
| 56—Food and beverage service activities | 1.3% |
| 68—Real estate activities | 1.3% |
| 13—Textile industries | 1.2% |
| 16—Manufacture of wood and products of wood, except furniture, straw and plaiting | 1.2% |
| 17—Manufacture of paper and paper products | 1.2% |
| 23—Manufacture of other non-metallic mineral products | 1.2% |
| 42—Civil engineering | 1.2% |
| 73—Advertising and market research | 1.2% |
| 18—Printing and reproduction of recorded media | 1.0% |
| 11—Manufacture of beverages | 0.7% |
| 20—Manufacture of chemical products | 0.7% |
| 30—Manufacture of other transport equipment | 0.7% |
| 38—Waste collection, treatment, and disposal activities; materials recovery | 0.7% |
| 52—Warehousing and support activities for transportation | 0.7% |
| 81—Building and landscape service activities | 0.7% |
| 24—Metallurgical activities | 0.5% |
| 35—Electricity, gas, steam, and air conditioning supply activities | 0.5% |
| 64—Financial service activities (except insurance and pension funding) | 0.5% |
| 69—Legal and accounting activities | 0.5% |
| 72—Scientific research and development | 0.5% |
| 77—Rental and leasing activities | 0.5% |

**Table 1** (continued)

| Ateco 2-digit class | Frequency |
|---|---|
| 79—Travel agency, tour operator and other reservation service and related activities | 0.5% |
| 90—Creative, arts and entertainment activities | 0.5% |
| 96—Other personal service activities | 0.5% |
| 36—Water collection, treatment, and supply activities | 0.3% |
| 45—Wholesale and retail trade and repair of motor vehicles and motorbikes | 0.3% |
| 58—Publishing activities | 0.3% |
| 61—Telecommunications | 0.3% |
| 80—Investigation and security services | 0.3% |
| 85—Education | 0.3% |
| 88—Social work activities without accommodation | 0.3% |
| 93—Sporting, entertainment, and recreational activities | 0.3% |
| 95—Repair of computers and personal and household goods | 0.3% |
| 12—Tobacco industry | 0.2% |
| 32—Other manufacturing industries | 0.2% |
| 59—Motion picture, video, television programs, music and sound recording activities | 0.2% |
| 66—Activities auxiliary to financial services and insurance activities | 0.2% |
| 82—Office administrative, office support and other business support activities se | 0.2% |
| 86—Human health activities | 0.2% |

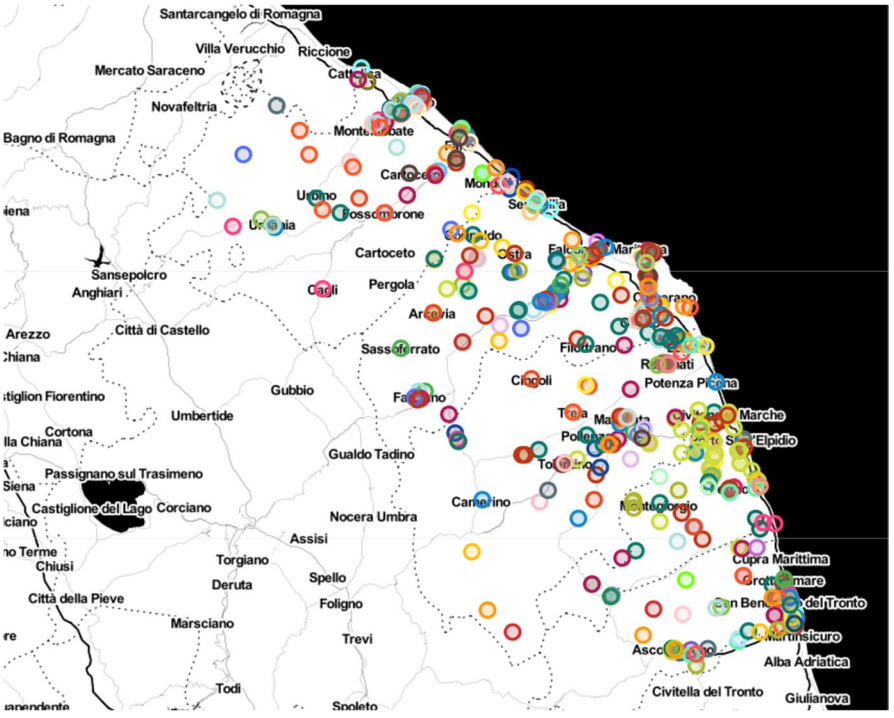


**Fig. 2** The localization of the firms in the Marche Region

The data in Aida BvD was accessed in 2021 and again in 2023 to collect up-to-date information. It should be emphasized that the data sources used are very different. The Aida BvD database provided us with financial and economic information on each company, while we collected descriptive texts from.html pages, which were used to generate keywords for classifying the companies and measuring their business proximity. We also conducted a survey to obtain detailed information about the companies' adoption of new technologies. Regarding this last source, the questionnaire allowed us to assess the adoption status of ten technologies, including big data and industry analytics, management systems, business intelligence and CRM, cloud computing, IT security, interconnected or modular production plants, collaborative robots, additive manufacturing systems (e.g., 3D printer), wearable devices, and augmented reality. Specifically, the survey included the following question, "Please, specify the digital technologies that are currently being developed or that the company is already using," followed by a list of ten different technologies that can be divided into two groups: "digital" and "smart manufacturing" technologies (Table 2).

Because the Ateco codes have the same limitations as any activity code-based classification system, we use an alternative approach to classify our companies that relies on descriptive text retrieved from company websites. An overview of the methodology is provided below.

## 4 Methodology

### 4.1 Textual analysis

In the proposed analysis, algorithms of text mining and semantics are used to process descriptive data. The aim of this first step is to profile and classify companies in a consistent and up-to-date manner based on the information collected from company websites (Marra & Baldassari, 2022). Since the descriptive texts can vary greatly from website to website, depending on how the company wants to present its

**Table 2** Frequency of firms by technology adoption

| No | Group | Technologies | Under development | Already in use |
|---|---|---|---|---|
| 1 | Digital | Big data & Industry Analytics | 8% | 3% |
| 2 | | Management systems | 6% | 19% |
| 3 | | Business Intelligence and CRM | 9% | 10% |
| 4 | | Cloud Computing | 5% | 10% |
| 5 | | IT security | 7% | 26% |
| 6 | Smart manufacturing | Interconnected or modular production plants | 54% | 6% |
| 7 | | Collaborative Robots | 2% | 2% |
| 8 | | Additive manufacturing systems (e.g., 3D printer) | 2% | 5% |
| 9 | | Wearable devices | 4% | 16% |
| 10 | | Augmented Reality | 3% | 1% |

features, we tried to capture four areas of interest, namely: specialisations, technologies, competences and knowledge of the companies. The keywords extracted from the descriptive texts were selected and grouped into four categories corresponding to the areas of interest mentioned above. The first category ("Specialisations") includes keywords that help us to describe the companies' activities and is similar to the classification by activity codes. The "Technologies" category includes a range of technologies that support the companies' industrial activities. The third category ("Competences") refers to the ability of company managers and employees to apply their knowledge to achieve results. The last category ("Knowledge") corresponds to the scientific disciplines held by the company's employees. It is thus possible to characterise companies according to their specialisations, technologies, competences and knowledge, which goes far beyond the usual industry codes.

We process the.html pages of the companies for each of the 553 companies included in the sample (by web crawling, if allowed). In doing so, we take the first step to obtain a multilabel classification. Specifically, we use a "general purpose" natural language recognition model based on machine learning algorithms trained on different knowledge bases. We perform pre-processing typical of text mining and use additional modules for spell checking and speech recognition. We use mixed models that draw on multiple taxonomies and use APIs (Access Programming Interfaces) to large libraries of software houses. We then switch to a more "specific" model to profile companies based on their activities. We perform labelling by assigning keywords to the observed companies, also using semantic understanding of the text. We perform manual check to assess the quality of the generated output and use automatic rules to reduce noise in the extraction phase. We then normalise the generated output and use predefined algorithms to obtain a multi-label classification and assign the keywords to the appropriate category.

Specifically, we use two families of algorithms: extraction algorithms, which identify the keywords that describe the business domains and classification algorithms, which assign the keywords to the categories. After extraction, the keywords are divided into categories following a thorough review and standardization. An important aspect is the selection of taxonomies: they must be as comprehensive and up-to-date as possible. We start with a set of taxonomies to classify specialisations and competences using our existing knowledge base and external sources. The latter consist of expert-driven and data-driven classifications (depending on whether the taxonomies are based on expert input or formulated using machine learning algorithms). We rely on taxonomies used by software houses specialized in text analytics as well as other useful taxonomies (National Research Council, 2010; European Commission, 2018; Federmanager, 2016; World Economic Forum, 2018; Assintel, 2020).

After performing these technical steps, we obtain 6,754 unique keywords that can be useful for company profiling. More specifically, we assigned the keywords to the 553 companies in the sample. Since each keyword can apply to more than one company, the companies were associated with vectors of keywords. To reduce the processing and storage costs required to create networks of words, we selected the unique keywords that appeared at least three times in the dataset (i.e., in at least three different companies or vectors of words), assigning 5,412 keywords to four different categories:

798 keywords in the "Specialisations" category, 1,451 in the "Technologies" category, 1,529 in the "Competences" category, and 1,634 in the "Knowledge" category.

### 4.2 Network analysis

After obtaining a set of keywords, we relied on graph theory to build networks of keywords. In this type of network, companies are the nodes, and an edge exists between two companies if they have at least one keyword in common; thus, links are based on the co-occurrence of words, and the larger the number of common keywords, the higher the weight of that edge.

It is appropriate to point out a few drawbacks associated with the use of words in network analysis. First, the collection of textual data requires careful filtering and cleaning to ensure the accuracy of the selected terms. Second, network analysis does not always allow reliable comparisons between different graphs, as different networks may have too different (and then non-comparable) structures. In this case, as we will see later, graph embedding may help.

We compare the matrix of adjacencies underlying the business proximity graph with two different matrices: one underlying the industrial proximity network and one underlying the technological proximity network.

The industrial proximity network is based on the hierarchy of Ateco 2007 activity codes: the lower the class that two companies share in the hierarchy, the more similar they are. According to this basic reasoning, firms in the same five-digit class are more closely related than firms that share only the same three-digit class. In the resulting network and in the underlying matrix of adjacencies, two firms are connected by a weight 5 link if they have the same five-digit class, by a weight 4 link if they have the same four-digit class, by a weight 3 link if they have the same three-digit class, and by a weight 2 link if they have the same two-digit class. This network is undirected and weighted, i.e., the higher the n-digit class, the stronger the edge between the nodes.

The technological proximity network is constructed based on the adoption of technologies. Like the industrial proximity network, two firms are connected by a link of weight $n$, being this the number of technologies used by the two firms. Since technologies are ten, $n$ goes from 0 to 10. When $n$ is 0, there is no connection between the two firm nodes; the higher $n$ is, the stronger the edge between the two firm nodes. The resulting network is undirected and weighted.

### 4.3 Statistical models on firm performance

Our goal is to model firm performance and test if spillovers have some influence in the network of business proximity. As a first stage, we define a regression model as:

$$y_i = \alpha + \boldsymbol{x}_i^t \boldsymbol{\beta} + u_i \tag{1}$$

where, $y_i$ is the sales growth for $1 \ldots N$ firms in the period between 2018 and 2020 (Daunfeldt et al., 2015; Lu et al., 2021) and $\boldsymbol{X}$ is a matrix of $i = 1...N$ units and $r = 1 \ldots P$ variables including logs of:

- the sales in 2018, representing the firms' initial scale, for which we expect firms with larger size to experience slower growth (Uhlaner et al., 2013);
- the average number of employees (each with her/his own knowledge) in the time interval 2018–2020, representing the overall company's knowledge base expected to affect firm performance (Jansen et al., 2005; Morris et al., 2020);
- the number of adopted technologies, or technology adoption, which is expected to contribute positively to sales growth (Autry et al., 2010; Büchi et al., 2020).

Further, $\boldsymbol{\beta}$ is the parameters vector of length $P$ related to each of the covariates in [1], $\alpha$ is the intercept, and $u$ is the error term. This model may have limitations due to the traditional assumption that observations are independent (Anselin, 2010; Postiglione et al., 2017).

To account for interdependencies between firms, we augment [1] using spatial econometrics methods as a second step. We expect firms to be significantly affected by business neighbours in the context under analysis and we start from a spatial lag model (SLM) specification in which "spatial" interdependencies are meant as business similarities. A SLM takes the form:

$$y_i = \alpha + \boldsymbol{x}_i^t \boldsymbol{\beta} + \rho \sum_{j=1}^{N} w_{ij}^b y_j + u_i with j \neq i \tag{2}$$

where ρ is the autocorrelation parameter for the lag of the dependent variable determined on the proximity matrix $\boldsymbol{W}^b$. Technically, $\boldsymbol{W}^b$. is an exogenous $NxN$ matrix based on our business network and the inverted U-shaped relationship presented in Sect. 2.3. If two firms have no common keywords, $w_{ij}^b$. is zero, while this value increases to one as the number of keywords grows towards the median of the common word count. It then drops back to zero when two firms share too many keywords. The values on the main diagonal of $\boldsymbol{W}^b$. are zero by definition.

Lastly, we take advantage of a Cliff-Ord type spatial model also known as SARAR (Kelejian & Piras, 2017) and expressed as:

$$y_i = \alpha + \boldsymbol{x}_i^t \boldsymbol{\beta} + \rho \sum_{j=1}^{N} w_{ij}^b y_j + u_i with u_i = \lambda \sum_{j=1}^{N} w_{ij}^g u_j + \varepsilon_i and j \neq i \tag{3}$$

where the error term $u$ is the sum of an autocorrelated component plus an independently and identically distributed term $\epsilon$. Here, $\lambda$ is a parameter expressing the magnitude of autocorrelation in $u$. Interestingly, the model in [3] allows us to use two different contiguity matrices. Namely, we use a business proximity matrix $\boldsymbol{W}^b$ to account for spatial dependence in the lag of $y$ and a $NxN$ matrix $\boldsymbol{W}^g$ based on the inverse of geographical distance to capture spatial autocorrelation in the errors.

Spatial models are used in the paper to get a measure of direct and indirect impacts. The interpretation of models [2] and [3] should not dwell on parameter estimates, but on impacts. Direct impacts are a measure of marginal effects merely attributable to a change of any covariates in the units $i$. Conversely, indirect impacts value the marginal effects due to a change in any of the independent

variable in neighbours. The total impacts are obtained as a sum of the two. Technically, direct and indirect impacts are computed by considering the closed form of models [2] and [3] and calculating the matrix of partial derivatives of $y$ on each $X^r$. The average on the diagonal elements of the matrix of partial derivatives $\partial y_i/\partial X_i^r$ will return us the direct impact of $X^r$, while the average of the off-diagonal matrix $\partial y_i/\partial X_j^r$ is to be meant as the indirect impact (LeSage & Pace, 2009).

Indirect impacts from equations [2] and [3] can be considered as a measure of business spillovers. The variance for the impacts (direct, indirect, and total) can be obtained by available MC routines and statistical significance of spillovers can be tested. As a results, the integration of our model into a spatial frame extends the understanding of the covariates and how these contribute (indirectly) through neighbours' spillovers. Neighbours with high sales volume are expected to impact firm performance via competition rather than coopetition (Bouncken & Fredrich, 2012; Bouncken et al., 2018), while neighbours with a higher number of employees (and a larger knowledge base) are expected to have a greater chance to profitably collaborate and interact with peers at close firms, exchange ideas and build 'new' knowledge (Dahl & Pedersen, 2004; Ter Wal & Boschma, 2011). Also, neighbours with more technologies are expected to enhance firms' capacity to exploit opportunities and grow (Aldieri et al., 2020; Guerrero et al., 2023).

Lastly, we emphasise that in this paper the purpose is to test spillovers in sales growth implied by business proximity. Thus, all direct and indirect impacts are based on the business matrix $\boldsymbol{W^b}$ considered to compute spatial lags of our dependent variable. Besides a spatial lag specification, we rely on the SARAR model to embed potential geographical factors through the disturbance components at least.

## 5 Results

### 5.1 Networks of proximity

Based on the adjacency matrices described above, three different graphs can be constructed (Table 3).

The first graph is that of industrial proximity, in which companies are linked on the basis of sharing the 2-, 3-, 4-, and 5-digit Ateco classes. This diagram has a peculiar configuration: the 553 firm nodes (connected by 12 459 edges) form a rather articulated network with strong fragmentation, in which several clusters are formed based on the proximity of the firms within the Ateco hierarchy.

The average degree is 45.1, which means that a node has about 45 connections on average. The average weighted degree–- the average sum of the weights of the edges of the nodes–- is 17.6. Both the network diameter (i.e., the maximum distance between any pair of nodes in the graph) and the average shortest path length (i.e., the average number of steps along the shortest paths for all possible pairs of nodes in the network) are equal to 1, indicating that the network has a fairly tight mesh structure. The graph density (which provides information about the degree of connectedness between nodes) is 0.1, i.e. 10% of all possible connections are realised. Caution should always be exercised when interpreting the value of the density, especially when comparisons

**Table 3** The metrics of the graphs underlying the different proximity measures

| Metrics | Industrial proximity | Technological proximity | Business proximity |
|---|---|---|---|
| Number of nodes | 553 | 553 | 553 |
| Number of edges | 12,459 | 118,721 | 101,641 |
| Average degree | 17.6 | 429.4 | 367.6 |
| Average weighted degree | 45.1 | 892.5 | 1408.7 |
| Graph diameter | 1 | 2 | 3 |
| graph density | 0.1 | 0.8 | 0.7 |
| average path lenght | 1 | 1.1 | 1.3 |
| modularity | 0.9 | 0.06 | 0.2 |
| average clustering coefficient | 1 | 0.9 | 1.3 |

with a network of similar size are not possible. Another popular network-level indicator (which is also referred to as a cluster-level indicator) is modularity, which is a measure of the extent to which a network is divided into groups or clusters. Networks with high modularity have dense connections between nodes within groups, but few connections between nodes in different clusters. Information about the tendency of nodes to form clusters is also provided by the average clustering coefficient, which captures the proportion of complete triangles formed by nodes. The modularity and average clustering coefficient are 0.9 and 1, respectively, indicating a strong tendency to cluster. In the network, clusters of companies belong to Ateco 46–- Wholesale trade, except of motor vehicles and motorcycles; Ateco 15–- Manufacture of leather products; Ateco 25–- Manufacture of fabricated metal products, except of machinery and equipment; Ateco 62–- Computer programming, consultancy and related activities; and Ateco 43–- Specialised construction activities.

The second graph consists of the technological proximity network, in which the companies are connected if they share at least one technology. The data used to create the diagram comes from the survey conducted among the companies in the sample. This graph shows a very different configuration from the previous one: the 553 firm nodes tend to cluster together into a single central cluster. The number of edges is significant: 118,721. The average degree is 429.4, and the average weighted degree is 892.5. The network diameter is 2 and the average path length is 1.1 (i.e., both values are higher than in the industry proximity network), while the graph density is 0.8, indicating a much higher connectivity compared to the first network. The modularity and the average clustering coefficient are 0.06 and 0.9, respectively. As there is no aggregation by Ateco sector in this case, it is more difficult to examine the network, which would allow to distinguish groups of companies sharing the same number of technologies. However, it should be borne in mind that the graphical representation of the connections between companies only has the function of visualising these connections from a bird's eye view; if the number of connections is very high, the visualisation of the graph is hardly informative. It should also be noted that some company nodes located at the edge of the graph have no relationship with other nodes because they do not use any of the ten technologies.

The third network is the graph comprising the links between the 553 company nodes based on textual data. The number of edges is high and amounts to 101,641. The average degree is 367.6 and the average weighted degree is 1408.7. The network diameter is 3, the graph density is 0.7 and the average path length is 1.3. The modularity and the average clustering coefficient are 0.2 and 1.3 respectively. Again, the network is too complex to allow an effective visual interpretation.

Even in this case, some company nodes remain unconnected. This limitation is mitigated by two elements, one related to the methodology initially proposed and the other to a simple operational decision in the exercise carried out. Indeed, if one intends to classify companies based on descriptive texts on company websites, some companies may be poorly described, which calls into question the soundness of the proposed classification. This drawback may be exacerbated if the textual description of the activities carried out by the company does not contribute to the classification system. For example, if the descriptive text is incomplete, useless or irrelevant: the result in this case can also be the so-called "empty" companies. However, in the present case, the limitation is exclusively due to the operational decision made at the stage of assigning keywords to companies. Namely, in order to reduce processing and storage costs, we have chosen to limit the number of keywords useful for classification to those that occur at least three times, thus limiting the number of keywords and the number of connections.

### 5.2 Comparing networks of proximity

Given the complexity of graphs and the resulting difficulties in studying them, we propose to measure the similarity between adjacency matrices. Intuitively, we intend to overlay the different adjacency matrices and measure some kind of correlation between them. Specifically, we use the graph embedding technique to measure the distance between each of the three matrices and the other two. The resulting output preserves as much structural information and attributes of the graph as possible by transforming the data into a low-dimensional space (Goyal & Ferrara, 2018). To compare the three networks, we use the Network Laplacian Spectral Descriptor (NetLSD) framework. NetLSD is a graph embedding method that generates a vector representation for each graph. Since the method allows us to map the different graphs in a Euclidean space, we consider the distance d as a measure of dissimilarity between the vectors associated with each of the networks, which ranges from $d=0$ when the networks are the same to $\infty$ when the networks are different (Han et al., 2011; Frankl et al., 2020).

The results are summarised in a matrix in which we scaled the Euclidean pairwise distance matrix $\boldsymbol{D}$ by dividing all elements $d_{ij}$ by their maximum value (corresponding to 2.6). In this way, we obtain the scaled $d_{ij}^{*}$ elements of $\boldsymbol{D}^{*}$ (Table 4).

Surprisingly, the maximum distance is the one between industrial proximity and technological proximity. If we give this distance a value of 1, it represents the benchmark against which we can compare all other distances. On the other hand, the business proximity is very close to technological proximity, with 0.07. Finally, the distance between industrial proximity and business proximity is large and equal to 0.93.

**Table 4** Distance matrix between the vectors associated to the different proximity graphs

| | Industrial proximity | Technological proximity | Business proximity |
|---|---|---|---|
| Industrial proximity | 0.0 | 1.0 | 0.93 |
| Technological proximity | | 0.0 | 0.07 |
| Business proximity | | | 0.0 |

The results give rise to a number of considerations. First, the method based on textual data from company websites allowed for a new and original classification that is both more informative than industrial proximity based on activity codes and less time-consuming than technological proximity based on survey data. As mentioned above, this second measure could only be estimated based on the effort required to create a questionnaire, its telematic transmission and completion by entrepreneurs and managers, as well as the final data processing. The advantages of the business proximity measure are also reflected in the relatively low response rate, which makes it difficult to process the lists of companies. Secondly, we will show how the business proximity measure allows us to examine the information underlying this measure in detail. Companies can be assimilated based on the different dimensions of interest: Specialisations, Technologies, Competences, and Knowledge. These dimensions can be considered in isolation or all together.

### 5.3 Using business proximity to explain firm performance

Estimations obtained by the three specifications proposed are shown in Table 5. Both SLM and SARAR are estimated using maximum likelihood.

In both SLM and SARAR, the lag of the dependent variable is calculated using the business proximity matrix. In SARAR, the autocorrelation in the error term is accounted for by the inverse geographic distance between each *i* and the nearest 80 units. Accordingly, the diffusion of local shocks can be taken into account in the firm growth model.

Values of ρ confirm the role of business proximity in explaining sales growth. Hence, we will consider direct and indirect impacts for these models and obtain spillovers due to business proximity. On the other hand, the estimated λ confirms that geographical autocorrelation affects the residuals. This feature together with AIC figures justify the choice of the SARAR model.

In Table 6, we focus on the impacts for the interpretation of the results (LeSage & Pace, 2009). It can be seen that the impact of initial size is negative, as smaller firms tend to grow more and vice versa. This is in line with the expected outcome and an extensive empirical literature that uses firm size as a moderating variable for sales growth (Uhlaner et al., 2013). The interpretation regarding indirect effects should be emphasized, whereby if the business neighbours are larger, the growth rate is further reduced, while smaller neighbours could have a positive effect on sales growth. Such a result is interesting when considering previous research that tends to argue that coopetition rather than competition can improve firm performance (Bouncken & Fredrich, 2012; Bouncken et al., 2018).

**Table 5** Parameter estimates obtained by linear regression, spatial lag SLM, and SARAR. Estimated standard errors are in brackets, while the p-values levels are (*) = 0.10, (**) = 0.05, (***) = 0.01

| | OLS (1) | SLM (2) | SARAR (3) |
|---|---|---|---|
| (Intercept) | 1.6279*** (0.1039) | 1.4551*** (0.1037) | 1.4518*** (0.1048) |
| Sales at year 2018 | -0.2140*** (0.0194) | -0.1996*** (0.0181) | -0.1994*** (0.0179) |
| Average number of employees | 0.1278*** (0.0219) | 0.1200*** (0.0201) | 0.1176*** (0.0200) |
| Number of technologies adopted | -0.0174** (0.0074) | -0.0149** (0.0068) | -0.0138** (0.0200) |
| ρ (business proximity) | | 0.3020*** (0.0733) | 0.3011*** (0.0067) |
| λ (geographical proximity) | | | 0.4077* (0.2348) |
| AIC | 178.72 | 165.31 | 164.68 |

**Table 6** Direct, indirect, and total impacts calculated for the SARAR specification. Estimated standard errors are in brackets, while the p-values levels are (*) = 0.10, (**) = 0.05, (***) = 0.01

| | Direct impacts | Indirect Impacts | Total Impacts |
|---|---|---|---|
| Sales at year 2018 | -0.2128*** (0.0193) | -0.0726*** (0.0273) | -0.2853*** (0.0386) |
| Average number of employees | 0.1255*** (0.0217) | 0.0428** (0.0175) | 0.1683*** (0.0340) |
| Number of technologies adopted | -0.0147* (0.0065) | -0.0050 (0.0030) | -0.0198** (0.0090) |

As far as the average number of employees is concerned, both the direct and indirect effects are positive (Jansen et al., 2005; Uhlaner et al., 2013). Directly, a company with many employees grows faster, as its employees can contribute unique knowledge that promotes company growth. Indirectly, as the number of employees increases, so does the knowledge base, which increases the potential for interorganizational connections and collaboration, thereby increasing overall performance (Dahl & Pedersen, 2004; Ter Wal & Boschma, 2011). This important finding highlights the role of companies with a large workforce in promoting knowledge sharing and enhancing the performance of their business neighbours. The result highlights the potential insight offered by spillovers thanks to the combined use of a spatial model and a business proximity matrix.

Regarding the last variable considered, we find a slightly negative direct effect of technology adoption: firms do not benefit from more technologies, in contrast to the existing literature (Giotopoulos et al., 2017). This result requires further analysis and investigation, considering the existing debate between technological breadth and depth (Autry et al., 2010; Büchi et al., 2020). As we have no useful evidence to assess the dimension of technological depth (low breadth does not imply higher depth), we cannot derive an interpretation to emphasise the benefits of greater technological specialisation. The indirect effect is not statistically significant, so firms

would not benefit from the widening of technology adoption by business neighbours (Guerrero et al., 2023).

## 6 Conclusion

Textual data represent the last frontier in the empirical literature on firm proximity. Although a growing number of studies have used such data to create more consistent classification systems, make more accurate comparisons between companies and measure firm proximity, no reference methodology has yet emerged, and no attempts have been made to compare the results with standard measures of proximity.

With this in mind, the aim of this paper was to propose a methodology for measuring similarity between firms and to compare the resulting measure with industrial and technological proximity. In addition, an attempt was made to explain sales growth, assuming that greater knowledge sharing can take place between similar companies. The results provided initial indications of the likelihood of knowledge exchange between firms that are close to each other in the business proximity network. In addition, the study made it possible to highlight the advantages and disadvantages of the proposed methodology.

The proposed approach based on textual data has two main advantages: first, it makes it possible to obtain a result that is simultaneously more clearly articulated than industrial proximity based on activity codes and less time-consuming than technological proximity based on survey data. Second, this methodology allows for a more detailed examination of the information.

Companies can be aggregated on the basis of the various dimensions of interest: Specialisations, Technologies, Competences and Knowledge. Using textual data to classify companies can help policy makers answer precise questions and narrow down firms by areas of interest. Suppose we want to narrow down the clusters of companies with a “green",” “Industry 4.0” or “digital” footprint: this would not be possible if we only had access to the activity code. Instead, classification based on textual data could make this task feasible. Text analysis would also allow us to identify the “green” companies (in our sample, at least 42 companies refer to an environmentally friendly approach on their website), the companies that belong to industry 4.0 (based on the keywords identified, there are at least 170 companies that use smart manufacturing technologies in their production processes), or the so-called digital companies (185 companies in the sample confirm that they use digital technologies that enable data collection and processing). In this way, text analysis helps policy makers to better identify the objective of a given policy measure. More specifically, text analytics allows adding informative content (via keywords) to the description of the companies we want to study, such as green, Industry 4.0 or digital companies.

However, some limitations may arise from the data source, in our case the companies’ websites. If companies do not provide information about their business, the technologies they use, their products and services, it is not possible to classify them coherently. It should also be noted that in rare cases the opposite phenomenon may occur. Some companies, especially the larger ones, may be too diversified: this

circumstance can complicate the step of classifying companies, as the individual company is associated with too many keywords. Although this phenomenon may cause further complications in the interpretation of the analysis results, it is important to realize that the same phenomenon would make classification based on activity codes meaningless and inappropriate. There is also the possibility that companies may decide to display misleading information on their corporate website.

It is important to develop new methods to classify companies with increasing accuracy and to enable timely and effective policy action. To this end, it is crucial to provide researchers and analysts with a complete and up-to-date database so that they can focus on activities that enable coherent classification and measurement of similarity. In this direction, advanced analyses using textual data are on the rise, especially in the fields of regional economics, innovation economics and operational research. The texts used refer to the products and services offered, to the activities of the firms as found on the company websites, to the mission statement as expressed in the corporate purpose, to the R&D activities as found in the description of patents, and so on. However, it must be acknowledged that the novelty is offset by the challenges associated with the massive collection and systematisation of vast amounts of unstructured textual data.

A central concern of this work was to combine the stages of collecting and processing textual data, classifying companies based on the co-occurrence of keywords and measuring their proximity, with proposing statistical models that use this information to explain sales growth. The underlying assumption, as outlined earlier, is that two companies that are similar in business have a greater chance of sharing knowledge. The future steps in our research program are several. First, it is desirable to test the assumed model on larger samples of companies and compare the results across different geographical areas. Second, the initial phase on textual data will be conducted using company descriptions available on multiple databases, to check against the processing carried out on the companies' websites. Third, the estimation of proximity between companies will be conducted assuming additional functional relationships beyond the inverted U-curve used above to test further theoretical hypotheses on knowledge sharing. Fourth, we intend to test the results regarding technology adoption in order to propose robust and alternative interpretations should the results of a negative relationship with the dependent variable be confirmed. Finally, it is desirable to refine the statistical models by using different variables such as research and development expenditure, the number of patents filed or the number of new technologies adopted, and by using different contiguity matrices, alternating the business proximity matrix with the technological and industrial matrix, in order to maximize model performance.

**Funding** Open access funding provided by Università Politecnica delle Marche within the CRUI-CARE Agreement. The paper was funded within the PRIN project "Firms within networks of words: Using text data (instead of SIC codes) to measure proximity between firms and firm performance" - PRIN – Bando2022 - Prot. 20224EATBN.

**Data Availability** Data used in the analysis are available from Authors upon request.

## Declarations

**Competing interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.



## References

Aarstad, J., Kvitastein, O. A., & Jakobsen, S. E. (2016). Related and unrelated variety as regional drivers of enterprise productivity and innovation: A multilevel study. *Research Policy, 45*(4), 844–856.

Aldieri, L. (2013). Knowledge technological proximity: Evidence from US and European patents. *Economics of Innovation and New Technology, 22*(8), 807–819.

Aldieri, L., Bruno, B., Senatore, L., Vinci, C. P. (2020). The future of pharmaceuticals industry within the Triad: The role of knowledge spillovers in innovation process. *Futures, 122*, 102600. https://doi.org/10.1016/j.futures.2020.102600

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science, 89*, 3–25.

Assintel. (2020). Il mercato ICT e l'evoluzione digitale in Italia [The ICT Market and Digital Evolution in Italy]. Available at: https://www.assintel.it/osservatori-2/

Autry, C. W., Grawe, S. J., Daugherty, P. J., & Richey, R. G. (2010). The effects of technological turbulence and breadth on supply chain technology acceptance and adoption. *Journal of Operations Management, 28*(6), 522–536.

Banca d'Italia. (2021). L'economia delle Marche - Aggiornamento congiunturale, [The Economy of the Marche Region - Conjunctural Update, Managerial Skills] november 2021. Available at: https://www.bancaditalia.it/media/notizia/l-economia-delle-marche-aggiornamento-congiunturale-novembre-2021/?dotcache=refresh

Bishop A, Mateos-Garcia J., & Richardson, G. (2022). *Using text data to improve industrial statistics in the UK*. Economic Statistics Centre of Excellence (ESCoE) Discussion Papers ESCoE DP-2022-01.

Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies, 39*(1), 61–74. https://doi.org/10.1080/0034340052000320887

Boschma, R., & Gianelle, C. (2014). Regional Branching and Smart Specialisation Policy. S3 Policy Brief Series No. 06/2014. EUR 26521 EN. Luxembourg (Luxembourg): Publications Office of the European Union; 2014. JRC88242, DOI: 10.2791/039340, 10.2791/65062 (online); https://publications.jrc.ec.europa.eu/repository/handle/JRC88242

Boschma, R., & Iammarino, S. (2009). Related variety, trade linkages, and regional growth in Italy. *Economic Geography, 85*(3), 289–311.

Boschma, R., Eriksson, R., & Lindgren, U. (2009). How does labour mobility affect the performance of plants? The importance of relatedness and geographical proximity. *Journal of Economic Geography, 9*(2), 169–190.

Bouncken, R., & Fredrich, V. (2012). Coopetition: Performance implications and management antecedents. *International Journal of Innovation Management, 16,* 1250028.

Bouncken, R. B., Fredrich, V., Ritala, P., & Kraus, S. (2018). Coopetition in new product development alliances: Advantages and tensions for incremental and radical innovation. *British Journal of Management, 29*(3), 391–410.

Büchi, G., Cugno, M., & Castagnoli, R. (2020). Smart factory performance and Industry 4.0. *Technol Forecast Soc Change [Internet], 150*(November 2019), 119790. https://doi.org/10.1016/j.techfore.2019.119790

Cainelli, G., & Ganau, R. (2019). Related variety and firm heterogeneity. What really matters for short-run firm growth? *Entrepreneurship & Regional Development, 31*(9–10), 768–784.

Cainelli, G., & Iacobucci, D. (2012). Agglomeration, related variety and vertical integration. *Economic Geography, 88*(3), 255–277.

Cantner, U., & Meder, A. (2007). Technological proximity and the choice of cooperation partner. *Journal of Economic Interaction and Coordination, 2*, 45–65.

Caragliu, A., de Dominicis, L., & de Groot, H. L. F. (2016). Both Marshall and Jacobs were right! *Economic Geography, 92*(1), 87–111.

Choi, S. B., & Williams, C. (2014). The impact of innovation intensity, scope, and spillovers on sales growth in Chinese firms. *Asia Pacific J Manag [Internet], 31*(1), 25–46. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84894625454&doi=10.1007%2Fs10490-012-9329-1&partnerID=40&md5=11af59b03da24cd163a9e66b08785288.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152. https://doi.org/10.2307/2393553

Cortinovis, N., & Van Oort, F. (2015). Variety, economic growth and knowledge-intensity of European regions: A spatial panel analysis. *Regional Studies, 41*(5), 685–697.

Dahl, M. S., & Pedersen, C. Ø. R. (2004). Knowledge flows through informal contacts in industrial clusters: Myth or reality? *Research Policy, 33*(10), 1673–1686.

Daunfeldt, S., Halvarsson, D., & Mihaescu, O. (2015). High-growth firms : Not so vital after all? *Int Rev Entrep [Internet]., 14*(4), 1–30. http://www.diva-portal.org/smash/record.jsf?pid=diva2:881367.

Ejdemo T, Örtqvist D (2020) Related variety as a driver of regional innovation and entrepreneurship: A moderated and mediated model with non-linear effects. Res Policy [Internet] 49(7). Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087587251&doi=10.1016%2Fj.respol.2020.104073&partnerID=40&md5=33cbf6396d9d59cd8ab55420208a2768

Essleztbichler, J. (2015). Relatedness, industrial branching and technological cohesion in US metropolitan areas. *Regional Studies, 49*(5), 752–766.

European Commission, Directorate-General for Employment, Social Affairs and Inclusion. ESCO handbook: European skills, competences, qualifications and occupations. Available at: https://publications.europa.eu/en/publication-detail/-/publication/ce3a7e56-de27-11e7-a506-01aa75ed71a1; Brussels; 2018.

Falcioglu, P. (2011). Location and determinants of productivity: The case of the manufacturing industry in Turkey. *Emerging Markets Finance and Trade, 47*(Suppl. 5), 86–96.

Fallah, B., Partridge, M. D., & Rickman, D. S. (2014). Geography and high-tech employment growth in US counties. *Journal Economics Geographers [Internet]., 14*(4), 683–720. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84902584138&doi=10.1093%2Fjeg%2Flbt030&partnerID=40&md5=26499fea098fbccbc8bc8da6f40734c5.

Fan, J. P. H., & Lang, L. H. P. (2000). The measurement of relatedness: An application to corporate diversification. *Journal of Business, 73*, 629–660.

Federmanager. (2016). Competenze Manageriali–Disciplinare per la valutazione e la certificazione delle competenze manageriali [Guidelines for the Evaluation and Certification of Managerial Competencies]. Available at: http://www.federmanager.it

Feldman, M. P., Francis, J., & Bercovitz, J. (2005). Creating a cluster while building a firm: Entrepreneurs and the formation of industrial clusters. *Regional Studies, 39*(1), 129–141.

Fischer, M. M., Scherngell, T., & Jansenberger, E. (2006). The geography of knowledge spillovers between high-technology firms in Europe: Evidence from a spatial interaction modeling perspective. *Geographical Analysis, 38*(3), 288–309.

Fitjar, R. D., & Timmermans, B. (2017). Regional skill relatedness: Towards a new measure of regional related diversification. *European Planning Studies, 25*(3), 516–538.

Frankl, N., Kupavskii, A., & Swanepoel, K. J. (2020). Embedding graphs in Euclidean space. *Journal of Combinatorial Theory, Series A [Internet], 171*, 105146.

Frenken, K., Van Oort, F. G., & Verburg, T. (2007). Related variety, unrelated variety and regional economic growth. *Regional Studies, 41*(5), 685–697.

Fung, M. K. (2003). Technological proximity and co-movements of stock returns. *Economics Letters, 79*(1), 131–136.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–74. https://doi.org/10.1257/jel.20181020

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics, 91*(3), 645–670.

Giotopoulos, I., Kontolaimou, A., Korra, E., & Tsakanikas, A. (2017). What drives ICT adoption by SMEs? Evidence from a large-scale survey in Greece. *Journal of Business Research, 81*(August), 60–9. https://doi.org/10.1016/j.jbusres.2017.08.007

Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Syst [internet]., 151*, 78–94.

Guerrero, A. J., Heijs, J., & Huergo, E. (2023). The effect of technological relatedness on firm sales evolution through external knowledge sourcing. *The Journal of Technology Transfer [Internet], 48*(2), 476–514. https://doi.org/10.1007/s10961-022-09931-3

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

Hartog, M., Boschma, R., & Sotarauta, M. (2012). The impact of related variety on regional employment growth in Finland 1993–2006: High-tech versus medium/lowtech. *Industry and Innovation, 19*(6), 459–476.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science, 317*, 482–487. https://doi.org/10.1126/science.1144581

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy, 124*(5), 1423–1465. https://doi.org/10.1086/688176

Jaffe, A. B. (1986). Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents. *Profits and Market Value, American Economic Review, 76*(5), 984–1001.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *Quarterly Journal of Economics, 108*(3), 577–598.

Jansen, J. J. P., Van Den, B. F. A. J., & Volberda, H. W. (2005). Managing potential and realized absorptive capacity : How do organizational antecedents. *Acadamy Manag J., 48*(6), 999–1015.

Kelejian, H., & Piras, G. (2017). *Spatial econometrics*. Academic Press.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLoS ONE, 16*(4), e0249071.

Kinne, J., & Resch, B. (2018). Generating big spatial data on firm innovation activity from text-mined firm websites. *GI_Forum, 6*(1), 82–9.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Taylor & Francis.

Losurdo, F., Marra, A., Cassetta, E., Monarca, U., Dileo, I., & Carlei, V. (2019). Emerging specializations, competences and firms'' proximity in digital industries: The case of London. *Papers in Regional Science, 98*(2), 737–753.

Lu, R., Song, Q., Xia, T., Lv, D., Reve, T., & Jian, Z. (2021). Unpacking the U-shaped relationship between related variety and firm sales: Evidence from Japan. *Papers in Regional Science, 100*(5), 1136–1157.

Marra, A., & Baldassari, C. (2022). Using text data instead of SIC codes to tag innovative firms and classify industrial activities. *PLoS ONE, 17*(6), e0270041. https://doi.org/10.1371/journal.pone.0270041

Marra, A., Antonelli, P., & Pozzi, C. (2017). Emerging green-tech specializations and clusters – A network analysis on technological innovation at the metropolitan level. *Renewable and Sustainable Energy Reviews, 67*(C), 1037–1046. https://doi.org/10.1016/j.rser.2016.09.086

Marra, A., Carlei, V., & Baldassari, C. (2020). Exploring networks of proximity for partner selection, firms'' collaboration and knowledge exchange. The case of clean-tech industry. *Bus. Strat. Environ., 29*, 1034–1044.

Morris, D., Vanino, E., & Corradini, C. (2020). Effect of regional skill gaps and skill shortages on firm productivity. *Environment and Planning a: Economy and Space, 52*(5), 933–952.

Mowery, D. C., Oxley, J. E., & Silverman, B. S. (1998). Technological overlap and interfirm cooperation: Implications for the resource-based view of the firm. *Research Policy, 27*(5), 507–523.

Nathan, M., & Rosso, A. (2015). Mapping digital businesses with big data: Some early findings from the UK. *Research Policy, 44*, 1714–1733.

National Research Council. (2010). *A database for a changing economy: Review of the Occupational Information Network (O* NET)*. National Academies Press.

Neffke, F., & Henning, M. (2008). *Revealed relatedness: Mapping industry space*. Papers in Evolutionary Economic Geography, No. 8.19, Utrecht, the Netherlands: Urban and Regional Research Centre, University of Utrecht.

Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal, 34*(3), 297–316. https://doi.org/10.2307/23362658

Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic Geography, 87*(3), 237–265.

Nooteboom, B. (2003). *Interfirm Collaboration, Learning and Networks, an Integrated Approach Routledge*.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Res Policy [Internet], 36*(7), 1016–34. https://www.sciencedirect.com/science/article/pii/S0048733307000807.

OECD. (2013). Measuring the internet economy: a contribution to the research agenda. In: OECD Digital Economy Papers 226. OECD Publishing. https://doi.org/10.1787/5k43gjg6r8jf-en

Oikawa, K., (2017). Inter-firm technological proximity and knowledge spillovers. *Public Policy Review, Policy Research Institute, Ministry of Finance Japan, 13*(3), 305–324.

Papagiannidis, S., See-To, E. W. K., Assimakopoulos, D. G., & Yang, Y. (2017). Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the Internet age?, Computers & Operations Research, Volume 98, 2018. *ISSN, 355–366*, 0305–0548.

Park, H., Yoon, J., & Kim, K. (2013). Identification and evaluation of corporations for merger and acquisition strategies using patent information and text mining. *Scientometrics, 97*(3), 883–909.

Pavone, P., & Russo, M. (2017). Clusters of Specializations in the Automotive Supply Chain in Italy. An Empirical Analysis Using Text Mining, DEMB, Working Paper Series nr. 116. https://doi.org/10.25431/11380_1146316

Petruzzelli, A. (2011). The impact of technological relatedness, priorities, and geographical distance on university-industry collaborations: A joint-patent analysis. *Technovation, 31*, 309–319.

Piore, M. J., & Sabel, C. F. (1984). *The Second Industrial Divide*. Basic Books.

Porter, M. E. (1998). *Clusters and competition: New agendas of companies, government and institutions*. Harvard Business School Press, Boston, MA.

Postiglione, P., Andreano, M. S., & Benedetti, R. (2017). Spatial clusters in EU productivity growth. *Growth and Change, 48*(1), 40–60.

Quatraro, F. (2010). Knowledge coherence, variety and economic growth: Manufacturing evidence from Italian regions. *Research Policy, 39*(10), 1289–1302.

Raspe, O., & van Oort, F. (2008) Firm growth and localized externalities. *The Journal of Regional Analysis and Policy, 38*(2), 100–116.

Saviotti, P. P., & Frenken, K. (2008). Export variety and the economic performance of countries. *Journal of Evolutionary Economics, 18*(2), 201–218.

Schildt, H., Maula, M., & Keil, T. (2005). Explorative and exploitative learning from external corporate ventures. *Entrepreneurship Theory and Practice, 29*(4), 493–515.

Shi, Z. M., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly: Management Information Systems, 40*(4), 1035–1056.

Stuart, T. E., & Podolny, J. M. (1996). Local Search and the Evolution of Technological Capabilities. *Strategic Management Journal, 17*(S1), 21–38.

Ter Wal, A. L. J., & Boschma, R. (2011). Co-evolution of firms, industries and networks in space. *Regional Studies, 45*(7), 919–933.

Timmermans, B., & Boschma, R. (2014). The effect of intra- and inter-regional labour mobility on plant performance in Denmark: The significance of related labour inflows. *Journal of Economic Geography, 14*(2), 289–311.

Uhlaner, L. M., van Stel, A., Duplat, V., & Zhou, H. (2013). Disentangling the effects of organizational capabilities, innovation and firm size on SME sales growth. *Small Business Economics, 41*(3), 581–607.

Van Oort, F., de Geus, S., & Dogaru, T. (2015). Related variety and regional economic growth in a cross-section of european urban regions. *European Planning Studies, 23*(6), 1110–1127.

Wales, W. J., Parida, V., & Patel, P. C. (2013). Too much of a good thing? Absorptive capacity, firm performance, and the moderating role of entrepreneurial orientation. *Strateg Manag J, 34*(5), 622–633. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84875680649&doi=10.1002%2Fsmj.2026&partnerID=40&md5=6374ea8bc3df9a0776c8d7fdca2fb6f9.

World Economic Forum. (2018). *Towards a Reskilling Revolution: A Future of Jobs for All*. World Economic Forum.

Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics, 90*(2), 445–461. DOI: 10.1007/s11192-011-0543-2

Yoon, J., Park, H., & Kim, K. (2013). Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis. Scientometrics, 1–19. https://doi.org/10.1007/s11192-012-0830-6