The Annals of Regional Science (2026) 75:30
https://doi.org/10.1007/s00168-026-01465-y

ORIGINAL PAPER

Knowledge flows through text data: the case of Milan's tech ecosystem

Alessandro Marra[1] · **Andrea D'Isidoro**[1] · **Alessandro Sarra**[1]

Received: 23 July 2024 / Accepted: 1 February 2026
© The Author(s) 2026

Abstract
Investigating the role of proximity between firms is crucial, as it can influence performance through spillover effects. This paper makes two main contributions. First, we introduce a replicable web-based methodology for measuring different types of proximity among firms. We draw on text data from company websites to assign firms vectors of keywords reflecting their industrial specializations, knowledge and competencies, and adopted technologies. These vectors are then used to compute cosine similarity. To account for knowledge exchange, the resulting linear similarity measures are transformed into an inverted U-shaped relationship, as suggested by theory. Second, we assess and compare the spillover effects arising from non-geographical proximities and their impact on firms' performance. We construct three proximity matrices (industrial, cognitive, and technological) and employ spatial models to distinguish and compare the informational contribution of each matrix in capturing spillovers. The findings confirm that proximity, through both knowledge exchange and competitive pressures, significantly affects firm performance. This indicates that spillovers arising from non-geographical proximities are important indirect drivers of firm performance and should not be underestimated.

JEL Classification C55 · L25 · L52

✉ Alessandro Marra
alessandro.marra@unich.it

Andrea D'Isidoro
andrea.disidoro@unich.it

Alessandro Sarra
alessandro.sarra@unich.it

[1] University of Chieti-Pescara, Chieti, Italy

Published online: 18 February 2026

1 Introduction

The role of proximity between firms is increasingly discussed in the scientific literature and in the policy debate. This growing focus highlights the importance of geographical, industrial, cognitive, and technological proximities in fostering innovation, collaboration, and competitive advantage: close or similar firms can benefit from knowledge spillovers, which collectively boost productivity and economic growth (Neffke and Henning 2008; Ter Wal and Boschma, 2011). While the role of geography is widely acknowledged, attention is being directed toward the conditions under which non-geographical forms of proximity can foster a fertile ground for collaboration. The underlying assumption is straightforward yet powerful: a shared base of expertise enhances mutual understanding, knowledge exchange and fosters the generation of new ideas. However, while firms with a moderate degree of overlap are naturally positioned to cooperate, innovate, and grow, such similarity can also intensify competition as they contend in the same markets (Aghion et al. 2005). This tension is further amplified by imitation, as companies often replicate their peers' strategies and policies. The outcome is a set of evolving and dynamic interactions: on the one hand, they foster knowledge exchange that fuels firms' innovation and growth; on the other, they can trigger overtly competitive processes with inevitable spillovers, though at times these interactions may take the form of 'coopetition' (Boschma and Gianelle 2014).

The above assumption is common to different theoretical perspectives. The 'related variety' approach highlights the advantages firms gain when operating across related or complementary industries (Frenken et al. 2007). The 'absorptive capacity' framework stresses that firms need sufficient prior knowledge and internal expertise to recognize, absorb, and apply external knowledge (Cohen and Levinthal, 1990). The 'recombinant innovation' perspective emphasizes that innovation often emerges from recombining existing knowledge in new ways, a process facilitated by industrial, cognitive, or technological proximity (Li et al. 2021). Taken together, these perspectives suggest that firms with industrial, cognitive, and technological similarities are better positioned to exchange knowledge, collaborate, and innovate.

Nonetheless, the measurement of such non-geographical proximities is far from straightforward. Some scholars employ activity codes, emphasizing the advantages that arise when firms from related industries are also co-located in the same geographic area (Frenken et al. 2007; Quatraro 2010; Van Oort et al. 2015). Likewise, patent codes are a standard and widely used source of data (Jaffe 1986; Aldieri 2013). Recent applications leverage data analytics and process text data drawn from company websites, patent descriptions, product descriptions, and other sources (Nathan and Rosso 2015; Hoberg and Phillips 2016; Bekamiri et al. 2021). Another empirical challenge lies in linking proximity to performance: once proximity has been measured, the goal is to assess whether spillovers exist and affect performance: spatial econometric models are essential tools (Carboni 2013; Marra et al. 2024).

In this paper, we use text data from company websites to construct vectors of keywords for firms, capturing their industrial specializations, knowledge and competencies, and adopted technologies. Next, we employ these vectors to measure similarity

between firms and transform the estimated linear relationship into an inverted-U one, as suggested in the literature (Boschma 2005; Nooteboom et al. 2007). Accordingly, we use the three resulting proximity matrices (respectively, industrial, cognitive, and technological) for two purposes: first, to distinguish and compare the informational contribution of each matrix to capture spillovers; second, to interpret the results of the chosen spatial models in explaining firms' performance. Our contribution to the literature is twofold. First, we introduce a replicable methodology for measuring different proximities among firms using web-based data. Second, we assess and compare the potential spillover effects arising from non-geographical proximities and their impact on firms' performance.

The paper is structured as follows. Section 2 introduces the relevant literature. Section 3 presents the dataset, while Sect. 4 describes the methodology, encompassing both textual analysis and statistical modeling. Section 5 discusses the results and Sect. 6 provides some robustness checks. Finally, Sect. 7 addresses the main limitations and outlines future steps of research.

2 Literature

As well-known, several streams of research provide insights into the mechanisms of knowledge exchange. The related variety perspective projects an industrial form of proximity, underscoring the benefits firms gain from operating in sectors that share commonalities or complementarities, as such conditions foster mutual learning (Van Oort et al. 2015; Cortinovis and Van Oort, 2015). The idea is that firms operating within related industries face analogous challenges and are better positioned to exchange knowledge, learn from one another, and generate spillovers. The absorptive capacity framework typically refers to cognitive and technological proximities, positing that firms can effectively capture and exploit external knowledge only when they possess a sufficient base of prior expertise and internal capabilities (Cohen and Levinthal, 1990; Lin et al. 2012). This shared capacity accelerates the diffusion of ideas, ultimately fostering innovation and improved performance. Finally, the recombinant innovation literature, although perhaps the most distant from the approach proposed in this paper, deserves brief mention: it highlights how novel ideas and products often emerge through the recombination of existing knowledge (Guan and Yan 2016; Cecere and Ozman 2014). The underlying mechanism is that two similar firms, by sharing their own knowledge bases, can combine complementary expertise in new ways. In this view, similarity facilitates mutual understanding while maintaining sufficient differentiation to enable the creation of innovative products, processes, or services (Nan et al. 2018).

Even if the assumption underlying the above streams is broadly acknowledged in the literature, capturing and quantifying non-geographical proximity is anything but simple. With regards to industrial proximity, many argue that activity codes, while useful, can be excessively rigid and fail to reflect the dynamic nature of modern markets: as these evolve and new sectors emerge, activity codes may not capture the latest trends. Also, industrial classifications overlook the multifaceted operations of new firms, especially in the tech sectors, limiting the accuracy of economic

analyses and policy implications (Nathan and Rosso 2015). For such a reason, most studies complement activity codes with additional sources of information, such as workflows, input–output relationships, or the co-occurrence of products and services (Neffke and Henning 2013; Timmermans and Boschma 2014; Boschma and Iammarino 2009; Neffke et al. 2011). The measurement of technological proximity is a pivotal aspect of innovation studies, with patent codes serving as a standard tool: the distribution of patents across technology classes is a proxy for similarity between firms (Jaffe 1986). Numerous developments have followed: for instance, Aldieri (2013) extends this framework by incorporating the covariance between different technology classes, improving the identification of potential knowledge spillovers and opportunities for collaboration. Nevertheless, the proposed advancements come with limitations: patent classifications may not fully capture the complexity nature of modern technological progress. Additionally, the reliance on patent data assumes that firms actively patent their innovations, which may not be the case for all firms and for all industries. Therefore, patent codes should be used in conjunction with other qualitative and quantitative measures (Bekamiri et al. 2021; Marra and Baldassari 2022).

Recently, the ability to harness vast amounts of structured and unstructured digital data is transforming traditional research methods. Shi et al. (2016) use topic modeling on unstructured online text data describing business activities to measure proximity between firms. Nathan and Rosso (2015) and Papagiannidis et al. (2017) highlight how text data can deepen our understanding of industrial sectors, particularly technology-intensive ones, and help to build more accurate company profiles. Pavone and Russo (2017) and Kinne and Lenz (2021) show the potential of text analysis to reveal collaboration networks, supply chains, and innovative performance. These new approaches serve not only to provide a more accurate characterization of the observed firms but also to develop more refined measures of proximity among them.

Once proximity between firms has been measured, the goal is to assess whether reciprocal effects (positive or negative) increase in magnitude as distance decreases within the chosen space (i.e., in a linear manner), or whether they follow nonlinear patterns. Boschma (2005) cautions that the benefits of proximity may be nonlinear: when firms are either too distant or too close, the advantages diminish. If too little proximity may hinder knowledge exchange, also too much proximity can lead to redundancy and lack of new ideas (Nooteboom 2003; Nooteboom et al. 2007). On a similar line of reasoning, Cohen and Levinthal (1990). Accordingly, the knowledge exchange taking place between firms would shape as a U-inverted curve. From an empirical perspective, the evidence appears to be substantial and supported by robust analyses. Boschma et al. (2009) observe that when firms are excessively similar, the benefits diminish or even become detrimental. Along the same lines, Timmermans and Boschma (2014) highlight the need for an optimal degree of relatedness in inter-firm connections. Aarstad et al. (2016) further argue that firms embedded in related (but not identical) sectors tend to enhance their innovative capacity, while those operating in unrelated domains often suffer productivity losses due to limited opportunities for knowledge exchange. Beyond the numerous empirical confirmations, understanding how proximity influences performance remains a

major empirical challenge worth discussing. It lies at the core of a broad stream of research that addresses various research questions and carries concrete, policy-relevant implications (Guerrero et al. 2023; Audretsch et al. 2022). Spillovers generally foster innovation and growth, but only under specific conditions. For instance, Guerrero et al. (2023) emphasize that their magnitude depends on firm size and R&D intensity, while Colombelli and Quatraro (2019) show that startups particularly benefit from access to heterogeneous yet complementary expertise. Martin-Rios et al.

Table 1 Distribution of Milan's tech companies according to size (subset vs. dataset)

Size	Subset	Dataset
Micro (< 10 employees)	44.6%	47.3%
Small (10–50 employees)	38.0%	37.6%
Medium (50–250 employees)	11.3%	9.9%
Big (> 250 employees)	6.1%	5.2%

Table 2 Distribution of Milan's tech companies according to sales (subset vs. dataset)

Quartile	Subset	Dataset
Min	0.9	0.2
25%	5.2	5.1
50%	5.8	5.6
75%	6.4	6.3
Max	10.1	10.2
Mean	5.9	5.7

Table 3 Composition of Milan's tech ecosystem by industries (subset vs. dataset)

Industries	Subset	Dataset
Fintech	14.0%	12.8%
Marketing	9.9%	9.8%
Software	9.4%	9.4%
Media	7.8%	7.6%
Health	7.2%	8.6%
Food	6.2%	7.1%
Energy	4.8%	4.5%
Real estate	4.4%	4.2%
Home living	4.1%	3.7%
Transportation	3.5%	3.7%
Fashion	3.4%	4.2%
Education	3.4%	3.4%
Travel	2.7%	3.0%
Wellness	2.3%	2.2%
Sports	1.9%	1.9%
Others	15.0%	14.8%

Some tech companies belong to more than one industry

Table 4 Geographical distribution of company headquarters in Milan by zone (subset vs. dataset)

Zones	Subset	Dataset
Zone 1	30.8%	33.1%
Zone 2	4.0%	3.1%
Zone 3	9.8%	8.2%
Zone 4	10.5%	9.5%
Zone 5	7.8%	7.7%
Zone 6	6.2%	6.5%
Zone 7	7.4%	6.0%
Zone 8	13.6%	15.0%
Zone 9	9.9%	10.9%

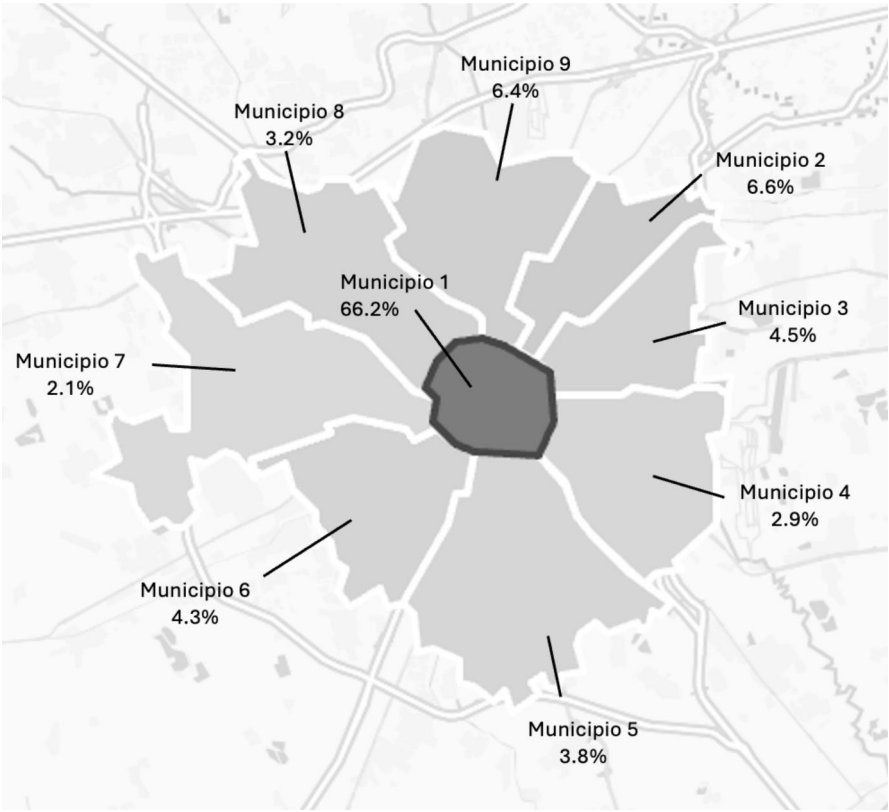


Fig. 1 Geographical distribution of companies in the City of Milan, Source: OpenStreetMap (2024)

(2022) highlight the specific role of small- and medium-sized enterprises in innovation dynamics. In this regard, a firm's size may not only reflect its resource availability, but the number of employees can also serve as a proxy for its knowledge base or absorptive capacity (Bloom et al. 2013). It is important to note that high dynamism

Table 5 Distribution of Milan's tech companies by firm size and industry (subset)

	Firm size (by employees)			
Industry	≤ 10 employees	11–50 employees	51–250 employees	> 250 employees
Fintech	39.5%	42.1%	15.8%	2.6%
Marketing	46.4%	41.1%	7.1%	5.4%
Software	45.3%	39.6%	11.3%	3.8%
Media	50.0%	29.6%	15.9%	4.6%
Health	48.8%	36.6%	12.2%	2.4%
Food	38.2%	35.3%	14.7%	11.8%
Energy	41.7%	50.0%	4.2%	4.2%

Table 6 Distribution of Milan's tech companies by firm age and industry (subset)

	Age classes (by years)			
Industry	2–4 years	5–7 years	8–10 years	> 10 years
Fintech	28.4%	40.5%	16.22%	14.9%
Marketing	18.8%	41.7%	22.92%	16.7%
Software	26.7%	35.6%	26.67%	11.1%
Media	22.5%	20.0%	32.50%	25.0%
Health	41.5%	36.6%	19.51%	2.4%
Food	22.6%	35.5%	25.81%	16.1%
Energy	33.3%	45.8%	8.33%	12.5%

of startups and early-growth firms might reflect in the speed of workforce expansion (Coad et al. 2016). Finally, the presence of technologically similar and physically neighboring firms emphasizes the role of geographical proximity (Rao-Nicholson et al. 2019).

3 Data

We collected data on the Milan's tech companies from Dealroom.co, a commercial database combining machine learning and data engineering with user-submitted data and verification processes, in March 2024 (Dealroom 2024). We started with the 3,600 units available in the database. Based on the completeness score (or 'Dealroom Signal') provided by the data provider, we retained only firms with enough reported data, reducing the sample to 1,200 units to ensure adequate information for analysis. From these 1,200 firms, we further focused on a subset of 505 units for which were available data on sales and number of employees over the last two years and data for geographical localization.

We verified that the selected sample is representative in terms of firm size (Table 1), sales (Table 2), sectoral composition (Table 3), and geographic

distribution (Table 4 and Fig. 1), thereby ensuring the robustness and generalizability of our findings.

All firms in our dataset are technology-based, with technology representing their defining characteristic. For example, a startup in the software sector can exhibit a level of technological intensity comparable to that of a firm operating in the health or food sectors. Moreover, it is worth noting that the companies' age ranges from 1 to 16 years, with the majority (84%) still at the startup stage. For the dataset used in this study, we also included two cross-tabulations, restricted to the six main sectors, which account for approximately 60% of the observed firms, to combine the sectoral dimension with firm size (Table 5) and firm age (Table 6).

Milan is Italy's business and financial capital, with over 800,000 firms and a regional GDP exceeding 390 billion euros. The city's sectoral strengths include life sciences, agribusiness, fintech, advanced manufacturing, and fashion and design. With more than 5,000 multinational companies that have their headquarters in Milan, the city has also produced its first two homegrown unicorns: Scalapay and Satispay.

4 Methodology

4.1 Text analysis

We calculated various measures of proximity to capture knowledge flows between firms. To this end, we begin with the descriptive texts available on company websites to extract an initial set of keywords (or 'entities'). This step relies on publicly accessible online content, where firms typically present their core activities, products, and services. Through text mining, we identify the most relevant entities that capture a firm's specializations, competencies, and technologies (Marra and Baldassari 2022). We recognize the concern regarding the heterogeneity of information between startups and more established firms: in our procedure, websites are considered informative only if they provide sufficient textual content; otherwise, firms with 'empty boxes' are discarded from the analysis. Next, we employ topic modeling (Latent Dirichlet Allocation, LDA) on the extracted entities to generate additional keywords (or 'topics') for each firm. LDA uncovers latent information by clustering entities into coherent 'topics', which effectively serve as supplementary keywords (Shi et al. 2016). Where necessary, we semantically refine and enrich these keywords by adding contextual information, ensuring that the terms are both relevant and meaningful within the industry. These two additional steps are specifically designed to mitigate potential heterogeneity in the available textual content. The outcome is a comprehensive firm profile that captures both the explicit and latent dimensions of its specializations, competencies, and technologies. All conventional operating steps have been taken (Gentzkow et al. 2019). We removed stop words and proceeded to stemming and lemmatization. Following normalization, text data underwent filtering to remove symbols, numbers, and other non-alphabetic characters. Also, we removed generic words or those too common across the firms' vectors of keywords, which helped in reducing redundancy and focusing on more distinctive terms.

4.2 From similarity to proximity matrices

Each firm is assigned three final vectors of keywords, one for each relevant dimension (industrial, cognitive, and technological). The vectors of keywords serve as unique identifiers, encapsulating the company's distinct characteristics, and are used to calculate the cosine similarity between every pair of firms. As well-known, the cosine similarity measures how similar two vectors are by calculating the cosine of the angle between them in the defined multi-dimensional space: unlike other similarity measures, it focuses only on the angle between vectors rather than their magnitudes or the number of keywords included. The value ranges from -1 to 1, where 1 indicates identical orientation, 0 indicates orthogonality, and -1 indicates opposite orientation.

To shift from cosine similarity to our proximity measure, meant to replicate the chance of knowledge exchange between firms, we model the attained linear relationships into an inverted-U shaped one. Such a shape suggests that the probability of knowledge exchange is highest at modest levels of proximity and decreases at the extremes (Boschma 2005; Nooteboom et al. 2007). Accordingly, we build spatial econometric models using different adjacency matrices ($\boldsymbol{W}$) that are assumed as exogenous and constructed according to a hyperbola: if two firms have very low or very high cosine similarity, w_{ij} is zero; higher values of the cosine similarity may have a more pronounced impact on the extent of proximity or knowledge exchange. Hence, w_{ij} is expected to increase up to one (the maximum), if the cosine similarity grows toward the median. The values on the main diagonal are zero.

4.3 Statistical models

We aim to explain firm performance, under the assumption that a greater exchange of knowledge occurs between firms that are moderately proximate in the industrial, cognitive, and technological space. Before modeling the interaction between firms using spatial econometric specifications, we define a basic ordinary least square (OLS) model of firms' performance as:

$$y_i = \alpha + x_i^t \beta + \varepsilon_i \tag{1}$$

Table 7 Summary statistics of variables (subset)

Variable	Min	20%	40%	60%	80%	Max	Median	Mean
Δ_*sales*	-0.95	0.00	0.01	0.18	0.35	2.18	0.03	0.09
*Sales_t*0	0.90	5.12	5.58	5.99	6.48	10.1	5.82	5.94
Δ_*emp*	-0.89	-0.25	0.11	0.40	0.85	1.49	0.21	0.33
Avg_emp	1	4.86	9.84	16.02	38.86	22,940	13.34	178.66
Age	2	5	6	8	11	17	7	10.86
Geo	10	18	21	33	42	50	23	29.14

where y_i is the growth rate of sales (Δ_sales) for the $i = 1 \ldots N$ firms between 2022 and 2023, calculated as the log difference between sales in 2023 and 2022 (Coad and Rao 2008); α is the intercept; β is the parameters' vector related to each of the covariates; ε_i is the error term. $\boldsymbol{X}$ is a matrix of $i = 1 \ldots N$ units and $r = 1 \ldots P$ variables, including: the log of revenues in 2022 (*Sales_t*0), as a proxy for the firms' initial scale (Uhlaner et al. 2013); the average number of employees between 2022 and 2023 (*Avg_emp*), which represents the firms' knowledge base (Bloom et al. 2013); the growth rate of employees between 2022 and 2023 (Δ_emp), to account for the speed of firms' knowledge base expansion, given the prevalence of startups and early-growth firms; a variable (*Age*) considering how old is the firm (Rydehell et al. 2019; Guerrero et al. 2023); the number of neighboring firms within an area of one square kilometer radius (*Geo*), which serves to account for the 'physical' or geographical space (Rao-Nicholson et al. 2019).

The summary statistics for the variables included in the model are presented in Table 7.

It should be noted that sales growth reflects a startup's ability to generate revenue, signaling actual demand for its products or services, while also facilitating access to funding, partnerships, and talent. Sales growth has several advantages: it is widely used in the literature, and sales data are generally available, allowing comparisons across startups in different sectors or stages (Coad and Rao 2008; LeBrasseur et al. 2003). However, sales growth also has limitations (Modina et al. 2023; Bronzini et al. 2020). Sales growth may be accompanied by low margins or operational losses, particularly if driven by heavy investment, and may not capture other important dimensions of performance, such as innovation (Coad and Rao 2008). Moreover, for startups, sales growth primarily signals market traction and scalability, often providing more meaningful information than profitability. In contrast, for mature firms, sales growth reflects competitive performance and market expansion and should be interpreted alongside profitability, as revenues tend to be more stable. To account for potential heterogeneity, Table 8 shows the average sales growth by firm age and size. We do not find any significant misalignment among the firms in the dataset.

Two standard spatial econometric models are considered: Spatial Lag Model (SLM) and Spatial Error Model (SEM). We consider the SLM as:

Table 8 Average sales growth by firm age and size (subset)

Age classes (by years)	Firm size (by employees)			
	≤ 10 employees	11–50 employees	51–250 employees	> 250 employees
2–4 years	0.10	0.08	0.30	–
5–7 years	0.04	0.08	0.09	0.04
8–10 years	−0.02	0.06	0.07	0.01
> 10 years	0.05	0.02	0.06	0.09

$$y_i = \alpha + x_i^t \beta + \rho \sum_{j=1}^{N} w_{ij} y_j + \varepsilon_i \; with \; j \neq i \tag{2}$$

where ρ is the autocorrelation parameter for the dependent variable indicating the magnitude of the mutual influence between neighbors. This means that if two firms are close according to the chosen similarity ($w_{ij} \neq 0$), their outcome values influence each other proportionally to the spatial parameter ρ. Through the estimation of a SLM we can measure direct and indirect impacts. In spatial models including endogenous interaction effects, such as the SLM, the interpretation of parameters is not straightforward: direct and indirect impacts should be computed by calculating the matrix of partial derivates of y on each covariate (LeSage and Pace 2009). Direct effects can be defined as the impact of an explanatory variable, in one region, over the dependent variable of the same region. We can find these direct effects on the principal diagonal of the matrix containing the partial derivatives of the expected value of y with respect to the r^{th} explanatory variable, for all units. Instead, the off-diagonal elements of the matrix capture spatial spillovers in the form of indirect effects that is the impacts due to a change in the covariates of a neighboring firm. The total impacts are obtained by summing the direct and indirect impacts. LeSage and Pace (2009) proposed two synthetic measures to describe these effects, called respectively the 'average direct impact' (of the r^{th} variable) as the average of the matrix diagonal elements, and the 'average indirect impact' as the average of the row sums, or the column sums, of the off-diagonal element in the same matrix. In our case, indirect impacts can be considered as a measure of spillovers: for example, neighbors with a higher value for the variables related to the number of employees (the extent of the firms' knowledge base or the speed of its expansion) are expected to have a greater chance to exchange knowledge with close firms (Ter Wal and Boschma, 2011; Guerrero et al. 2023).

Another spatial model considered in our work is the SEM, which takes the following form:

$$y_i = \alpha + x_i^t \beta + u_i \; with \; u_i = \lambda \sum_{j=1}^{N} w_{ij} u_j + \varepsilon_i \; and \; j \neq i \tag{3}$$

where λ is a parameter for the level of the autocorrelation embedded in the error component. In this model, the error term u_i is the sum of an autocorrelated component plus an independently and identically distributed term ε_i. In this model, the spatial dependence between the dependent variables is inserted in the error term, thereby modeling the interaction effects between disturbance terms. Such a specification allows the interpretation of the coefficients as in the OLS case.

Spatial tests.

We tested for the spatial autocorrelation in the residuals of the OLS model, based on the Moran's I statistic (Moran 1950). Starting from a generalization of this statistic, Cliff and Ord (1972) developed the Moran's I test for spatial autocorrelation in the regression error term of an OLS model. This well-known test

has many properties (King 1981). It is based on the Moran's I calculated for the regression error term as follows:

$$I_{\varepsilon} = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\varepsilon_i - \overline{\varepsilon}\right)\left(\varepsilon_j - \overline{\varepsilon}\right)}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right) \sum_{i=1}^{n} \left(\varepsilon_i - \overline{\varepsilon}\right)^2} \tag{4}$$

where n is the number of units, ε_i is the error term for unit i, $\overline{\varepsilon}$ is the average of the error term and w_{ij} is the weighting between two units as the generic element of the spatial weights' matrix. Under the hypothesis of no spatial autocorrelation, the expected value of the Moran's index is: $E(I) = 1/(n-1)$.

Starting from the computed value I_{ε}, Cliff and Ord (1981) demonstrated that the Moran's I test, being n sufficiently high, is asymptotically normal distributed as:

$$Z = \frac{I_{\varepsilon} - E\left(I_{\varepsilon}\right)}{\sigma_{I_{\varepsilon}}} \sim N(0,1) \tag{5}$$

The null hypothesis of the test is the absence of spatial autocorrelation between the error terms of the OLS regression. Instead, values significantly higher than expected lead to the rejection of the no correlation hypothesis in favor of the presence of positive spatial autocorrelation. Although the Moran's I test is the most used spatial diagnostic model test, it is susceptible to potential misspecifications and cannot be used to discriminate between the two specifications, SLM and SEM (Anselin and Rey 1991). With this regards, Anselin (1988) developed a Lagrange Multiplier (LM) statistic to test for SLM and SEM. However, since the LM_{Lag} and the LM_{Err} test are not mutually independent to each other, Bera and Yoon (1993) proposed a robust version of these tests (RLM). In their robust tests the hypothesis considers both the parameters as follows:

$$\mathrm{RLM}_{\mathrm{Lag}} \sim \chi^2_{(1)} \; with \; H_0 : \rho = 0; H_1 : \rho \neq 0 \; and \; \lambda = 0 \tag{6}$$

$$\mathrm{RLM}_{\mathrm{Err}} \sim \chi^2_{(1)} \; with \; H_0 : \lambda = 0; H_1 : \lambda \neq 0 \; and \; \rho = 0$$

Starting from the error term of the OLS model, the Moran's I test is used to detect the presence of spatial autocorrelation due to the proximity structure chosen and the RLM tests are employed to identify both the most appropriate spatial weights matrix

Table 9 Rank correlation between the three lists of keywords by investigated dimension

	Industrial	Cognitive	Technological
Industrial	1	0.484	0.580
Cognitive	0.484	1	0.530
Technological	0.580	0.530	1

Table 10 Results of Moran's I tests

W matrix	Moran I's statistic	Expectation	*P* value
Industrial	0.040	−0.002	0.000
Cognitive	0.011	−0.002	0.000
Technological	0.053	−0.002	0.000
Geographical	0.009	−0.002	0.039

and the most suitable model between SLM and SEM (Anselin and Rey 1991). Therefore, we adopt the following procedure:

a) we choose each time a spatial weights' matrix ($\boldsymbol{W}$) according to the selected proximity structure (specializations, competencies, technologies, and geography);
b) we perform the Moran's *I* test to detect the presence of spatial influence in relation to the selected proximity structure;
c) we use *RLM* tests for the model selection, to discriminate between SLM and SEM;
d) we estimate the best spatial regression model according to the *RLM* tests.

5 Results

Before presenting the results related to the Moran's *I* test and *RLM*, we report on the rank correlation on the lists of keywords, for each dimension, ordered from the most frequent keyword to the least used one. Table 9 shows results: the three rankings underlying the proximity measures show low correlation, close to the significance threshold of 0.5, indicating that the informational contribution provided by each list, on which the proximity matrices are built, is different. In particular, the list of keywords related to specializations is positively correlated with both the list of competencies and the list of technologies, though with the relatively low values of 0.48 and

Table 11 Significance of RLM tests

W matrix	RLM err p value	RLM lag *p* value
Industrial	0.833	0.008
Cognitive	0.975	0.197
Technological	0.008	0.969
Geographical	0.109	0.608

Table 12 Correlation between covariates

	Avg_emp	Δ*_emp*	*Sales_t*0	*Geo*	*Age*
Avg_emp	1	−0.015	0.445	−0.036	0.553
Δ*_emp*	−0.015	1	0.154	−0.017	−0.059
*Sales_t*0	0.445	0.154	1	−0.043	0.583
Geo	−0.036	−0.017	−0.043	1	0.086
Age	0.553	−0.059	0.583	0.086	1

Table 13 Estimated parameters for all the models

	OLS	SLM industrial	SLM cognitive	SEM technological
intercept	0.7045***	0.5027***	0.7179***	0.5432**
*Sales_t*0	−0.1124***	−0.0926***	−0.1042***	−0.0989***
Δ*_emp*	0.1530***	0.1501***	0.1495***	0.1529***
Avg_emp	0.0003*	0.0003*	0.0003*	0.0003*
Age	−0.0053	−0.0078	−0.0053	−0.0009
Geo	0.0376**	0.0303*	0.0344**	0.0319*
ρ		0.5964***		0.2942
λ			0.7242***	
R-squared	0.1767	0.2228	0.2423	0.2029

p values * < 0.1; ** < 0.05; *** < 0.01

0.58, respectively. Finally, the list corresponding to competencies is also positively correlated with technologies, with the rather low value of 0.53.

Regarding the Moran's *I* test, it is noteworthy that even though the statistic is close to 0, the p value indicates that there is still a significant deviation from the expectation of no spatial autocorrelation. The geographical structure, tested on the model without the *Geo* covariate to avoid endogeneity, suggests some type of spatial autocorrelation and supports its inclusion in the model (Table 10).

Concerning the *RLM* test, we recognize that the p values associated with RLM_{Err} (error) and RLM_{Lag} (lag) provide a measure of the statistical significance of spatial autocorrelation in the regression residuals and lag terms, respectively: if the p value associated with each is less than a certain level of significance, it indicates that the tested model is the most appropriate (Table 11).

Before discussing the results, we check for the absence of relevant correlation between the covariates (Table 12).

It is worth noting that sales and the number of employees are not strongly correlated. This pattern is largely due to the high share of startups: in technology-intensive industries, it is common to observe small firms achieving relatively high sales or, conversely, large firms generating little or no sales at early stages. As a result, the observed correlation is weaker than might be expected in a sample with a larger proportion of mature firms.

Table 13 shows the estimates for all the models, indicating that the covariates related to the initial firms' size (negative sign), the firms' knowledge base, represented by the average number of employees and its growth rate (both with positive signs) are statistically significant and have the expected signs, in line with the relevant literature (Guerrero et al. 2023; Bloom et al. 2013).

The results indicate contrasting effects of firm characteristics on performance via spatial spillovers. First, firm size appears to capture competitive dynamics: larger firms may prioritize business expansion over R&D or innovation, thereby limiting potential benefits for neighboring firms (Marra et al. 2024). This is consistent with the expectation that when larger and similar firms focus on growth and market capture, they may crowd out opportunities for other companies to innovate or leverage

Table 14 Impacts of SLM based on the proximity matrix of specializations

	Direct impacts	Indirect impacts	Total impacts
*Sales_t*0	−0.0943***	−0.1352**	−0.2295***
Δ_*emp*	0.1524***	0.2185***	0.3709***
Avg_emp	0.0003*	0.0005*	0.0008*
Age	−0.0079	−0.0014	−0.0093
Geo	0.0301**	0.0043	0.0344**

p values * < 0.1; ** < 0.05; *** < 0.01

complementary knowledge (Aghion et al. 2005). In contrast, a firm's knowledge base generates positive spillovers, consistent with knowledge diffusion theory: Firms that are similar in their activities or technologies can benefit from each other's expertise (Audretsch and Feldman 1996). Proximity thus facilitates learning and imitation, amplifying innovation and growth across the relevant industrial or technological networks (Malmberg and Maskell 2002; Boschma 2005). These dimensions appear to capture firm differences more effectively than the cognitive dimension, which reflects the underlying knowledge and competences associated with the firms' specializations and technologies. As a result, it seems to dilute the distinctive features that characterize individual firms, making them appear closer in the constructed non-geographical space.

Moreover, it is important to notice that the signs are confirmed across the different models. The significance of the spatial proximity based on specializations and technologies, measured respectively through the parameters ρ and λ, is substantial. Interestingly, the significance levels of the (positive) coefficients associated with the geographical variable differ across models (Rao-Nicholson et al. 2019). This may suggest the limited relevance of geography in the modern world, where digital technologies reduce the benefits of physical proximity. However, it should be noted that this result could be influenced by the metropolitan context under study, where distance may play a smaller role compared to analyses covering larger geographical areas: in other words, excessive geographical proximity can lead to spatial lock-in, in which over-embeddedness within a local system reduces exposure to diverse ideas and constrains innovative capacity (Martínez et al. 2024; Capó-Vicedo et al. 2008; Molina-Morales et al. 2011). This suggests investigating new methodologies that allow the simultaneous use of different proximity matrices, especially when each dimension captures distinct and complementary types of information.

Table 14 shows the direct and indirect impacts in the SLM with reference to the dimension of specializations. It can be observed that both impacts are significant: a direct impact on the dependent variable when covariates change, and an indirect impact, or spillover effect, which is the impact that a covariate has on the dependent variable of neighboring units through its direct effect. Accordingly, the dependent variable for one unit interacts with the dependent variable for other units by means of the spatial parameter ρ. The direct effect of the initial dimension (*Sales_t*0) is confirmed to be negative, as larger firms tend to grow less (Uhlaner et al. 2013). Additionally, the indirect impact shows that such a negative (direct) effect also

Table 15 Direct and indirect impacts of SLM with industrial adjacency matrix (the adopted model, Model 1), SLM model estimated via IV (Model 2), and SLM with the addition of two variables (Model 3)

	Model 1	Model 2	Model 3
Direct impacts			
*Sales_t*0	−0.0943***	−0.1007***	−0. 0949***
Δ_*emp*	0.1524***	0.1557***	0.1493***
Avg_emp	0.0003*	0.0003**	0.0003*
Age	−0.0079	−0.0087	−0.0079
Geo	0.0301**	0.0326*	0.0355**
Startup			−0.0410
B2B			0.0104
Indirect impacts			
*Sales_t*0	−0.1352**	−0.1433***	−0.0915**
Δ_*emp*	0.2185***	0.2217***	0.1441***
Avg_emp	0.0005*	0.0005*	0.0003*
Age	−0.0014	−0.0012	−0.0070
Geo	0.0043	0.0046*	0.0043*
Startup			−0.0394
B2B			0.0101

p values * < 0.1; ** < 0.05; *** < 0.01

Table 16 Estimated parameters of SEM with technological adjacency matrix (the adopted model, Model A), SEM model estimated via IV (Model B), and SEM with the addition of two variables (Model C)

	Model A	Model B	Model C
intercept	0.7179***	0.6887***	0.7959***
*Sales_t*0	−0.1042***	−0.0919***	−0.1075***
Δ_*emp*	0.1495***	0.1545***	0.1486***
Avg_emp	0.0003*	0.0003*	0.0003*
Age	−0.0053	−0.0104	−0.0552
Geo	0.0344**	0.0426***	0.0344**
Startup			−0.0604
B2B			0.0452
λ	0.7242***	0.6391**	0.6834***

p values * < 0.1; ** < 0.05; *** < 0.01

spreads to neighboring firms: possible interpretations may relate to a sort of competitive effect among companies that, similar in terms of specializations, may target the same markets and, consequently, exert downward pressure on each other's growth prospects (Marra et al. 2024). Significant direct and indirect impacts of the variables related to employees (*Avg_emp* and Δ_*emp*) confirm that the more people are potentially connected, the greater the benefit for firms' growth, both directly and indirectly. This result is consistent not only regarding the literature on the resource-based view of the firm, but also with the literature on knowledge spillovers (Guerrero et al. 2023; Bloom et al. 2013).

6 Robustness checks

In empirical studies, the validity of estimated statistical models is typically assessed through robustness checks, which aim to evaluate the stability of core coefficients under alternative model specifications. It is advisable to examine the insensitivity of the main coefficients to different estimation methods and the inclusion of additional (potentially redundant) variables (Lu and White, 2014).

Following this approach, we re-estimate the two main models (i.e., the SLM using the industrial matrix and the SEM using the technological adjacency matrix) by including two additional independent variables: a dummy variable capturing the firm's growth stage (*Startup*) and a dummy variable for the business model (*B2B*). The first control variable is introduced to complement the discussion in Sect. 4.3 on the statistical models, specifically to account for potential differences in growth performance between startups and more mature firms. The second variable serves as an additional control to capture potential differences in how firms with different business models may grow in terms of sales, as Business-to-Consumer (B2C) firms might achieve higher performance due to a larger consumer base (Cappelli and Cucculelli 2024). The results are provided, respectively, with Model 3 and Model C in Table 15 and Table 16.

Additionally, we perform an Instrumental Variable (IV) estimation by lagging the covariates (Model 2 and Model B in Tables 15 and 16) to address potential endogeneity concerns. The instruments employed are WX, W^2X, and W^3X, as recommended in the literature (Kelejian and Prucha 1998; Baltagi et al. 2014).

All estimations yield coefficients with the same sign, similar magnitudes, and comparable levels of statistical significance. Although the chosen instruments are theoretically recommended, we further perform tests for over-identification and weak instruments to verify their validity. The results of these tests are reported in Table 17.

Table 17 Weak instruments

Type of test	Test value	
Weak instruments tests	*Sales_t*0	5.463***
	Δ_*emp*	4.823***
	Avg_emp	13.803***
	Age	7.856***
	Geo	3.778***
Over-identification test	Sargan test	7.853 (*p* value: 0.346)

Tests show that the null hypothesis of weakness is rejected in all the cases, while the over-identification test cannot reject the null hypothesis of uncorrelation between instruments and error term, proving the validity of the chosen instruments; *p* values * < 0.1; ** < 0.05; *** < 0.01

7 Conclusions

Since proximity affects firms' performance through spillover effects in both geographical and non-geographical spaces, it is important to examine its role from multiple research perspectives. The methodology proposed in this paper builds on text data collected from the web, which are used to assign firms vectors of keywords capturing their industrial specializations, knowledge and competencies, and adopted technologies. These vectors are then employed to measure inter-firm similarity. Following theoretical insights, the estimated linear similarities are transformed into a nonlinear, inverted U-shaped relationship.

We leveraged three proximity matrices (respectively, industrial, cognitive, and technological) to distinguish and compare the informational contribution of each matrix to capture spillovers; second, to interpret the results of the chosen spatial models in explaining firms' performance. The results confirm that 'not too much' and 'not too little' proximity between firms enhances performance through spillovers. Both industrial and technological proximities show substantial significance. Initial firm size and firms' knowledge base display the expected signs (negative and positive, respectively) in line with the relevant literature. More specifically, these patterns suggest the presence of competitive pressures on the one hand, and knowledge spillovers on the other. Finally, we observe that the significance levels of the coefficients associated with the geographical variable vary across models, indicating the need for further investigation.

We consider Milan, the metropolitan area of this first application, as an excellent starting point since it represents a clear example of a metropolitan tech ecosystem. Its strong technology-intensive profile and the dynamism of its startup and innovation landscape position Milan as one of the most vibrant technological hubs in Europe, making it particularly well suited as a benchmark for studying innovation ecosystems at the city level. The evidence obtained from this analysis can serve as a benchmark for future applications aimed at assessing the relevance of non-geographical proximities and the existence of indirect effects related to firm size, age, and knowledge base.

From a policy perspective, these results are relevant. Depending on the technological ecosystem to be supported, policymakers may choose to foster different types of inter-firm networks (industrial or technological) to enhance spillover effects and stimulate growth. The findings also highlight the value of policies that promote related and complementary activities, rather than focusing narrowly on specific industrial specializations. Furthermore, attracting firms that may not appear immediately functional to the existing cluster could prove beneficial in the long run, as they may contribute to the emergence of new technological trajectories and new market trends.

This paper has several limitations. Even if the commercial database on startups and tech firms has several advantages, the number of variables that can be effectively applied in the statistical model are limited, primarily due to the scarce data availability that is common when firms are in the seed or early stages of growth.

These limitations had implications for our choice of variables and, consequently, in the model configuration.

Moving forward in the research program, our next step is to explore alternative functional forms that can help us understand knowledge spillovers beyond the conventional inverted U-shape. This investigation aims to uncover more relationships and dynamics, as proposed or discussed in the literature, within our data. Furthermore, we plan to employ models capable of capturing the indirect effects of covariates on the performance of neighboring firms. Future research will also integrate semantic similarity measures with geographical space (Autant-Bernard & LeSage 2011). By combining geographical and non-geographical proximity matrices, we aim to develop a more comprehensive understanding of the mutual impacts that firms exert on one another. These extensions will strengthen our ability to analyze and interpret inter-firm interactions, providing useful insights for both academic research and policy design.

Author Contribution Conceptualization: A.M., A.D., and A.S. Data curation: A.M., A.D., and A.S. Formal analysis: A.M., A.D., and A.S. Investigation: A.M., A.D., and A.S. Methodology: A.M., A.D., and A.S. Software: A.M., A.D., and A.S. Supervision: A.M., A.D., and A.S. Validation: A.M., A.D., and A.S. Writing—original draft: A.M., A.D., and A.S. Writing—review and editing: A.M., A.D., and A.S.

Funding Open access funding provided by Università degli Studi G. D'Annunzio Chieti Pescara within the CRUI-CARE Agreement. This work is part of a research program funded by the Italian Ministry of University and Research as a Project of Relevant National Interest (PRIN 2022). The project aims to use textual data to define new measures of proximity through which to measure the spillover effects on firm performance (Grant Prot. 20224EATBN)

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aarstad J, Kvitastein OA, Jakobsen SE (2016) Related and unrelated variety as regional drivers of enterprise productivity and innovation: a multilevel study. Res Policy 45(4):844–856

Aghion P, Bloom N, Blundell R, Griffith R, Howitt P (2005) Competition and innovation: an inverted-U relationship. Q J Econ 120(2):701–728. https://doi.org/10.1093/qje/120.2.701

Aldieri L (2013) Knowledge technological proximity: evidence from US and European patents. Econ Innov New Technol 22:807–819. https://doi.org/10.1080/10438599.2013.788838

Anselin L (1988) Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. Geogr Anal 20(1):1–17
Anselin L, Rey S (1991) Properties of tests for spatial dependence in linear regression models. Geogr Anal 23(2):112–131
Audretsch DB, Feldman MP (1996) R&D spillovers and the geography of innovation and production. Am Econ Rev 86(3):630–640
Audretsch DB, Belitski M, Guerrero M (2022) The dynamic contribution of innovation ecosystems to schumpeterian firms: a multi-level analysis. J Bus Res 144:975–986. https://doi.org/10.1016/j.jbusres.2022.02.037
Autant-Bernard C, LeSage JP (2011) Quantifying knowledge spillovers using spatial econometric models. J Reg Sci 51:471–496. https://doi.org/10.1111/j.1467-9787.2010.00705.x
Baltagi BH, Fingleton B, Pirotte A (2014) Spatial lag models with nested random effects: an instrumental variable procedure with an application to English house prices. J Urban Econ 80:76–86
Bekamiri H, Hain D, Jurowetzki R (2021) PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. Technol Forecast Soc Change. https://doi.org/10.1016/j.techfore.2024.123536
Bera AK, Yoon MJ (1993) Specification testing with locally misspecified alternatives. Econometr Theory 9(4):649–658
Bloom N, Schankerman M, Van Reenen J (2013) Identifying technology spillovers and product market rivalry. Econometrica 81(4):1347–1393
Boschma R (2005) Proximity and Innovation: A Critical Assessment. Reg Stud 39(1):61–74. https://doi.org/10.1080/0034340052000320887
Boschma R, Iammarino S (2009) Related variety, trade linkages, and regional growth in Italy. Econ Geogr 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. J Econ Geogr 9(2):169–190
Boschma R, Gianelle C (2014) Regional Branching and Smart Specialisation Policy, S3 Policy Brief Series, 06. jcr Technical Reports, European Commission.
Bronzini R, Caramellino G, Magri S (2020) Venture capitalists at work: a diff-in-diff approach at late-stages of the screening process. J Bus Ventur 35:105968. https://doi.org/10.1016/j.jbusvent.2019.105968
Capó-Vicedo J, Expósito-Langa M, Molina-Morales FX (2008) Improving SME competitiveness reinforcing interorganisational networks in industrial clusters. Int Entrep Manag J 4(2):147–169
Cappelli R, Cucculelli M (2024) The role of business models in explaining differences in firm performance. Edward Elgar Publishing, In Unpacking Innovation
Carboni O (2013) A spatial analysis of R&d: the role of industry proximity. Econ Innov New Technol 22:820–839. https://doi.org/10.1080/10438599.2013.788871
Cecere G, Ozman M (2014) Innovation, recombination and technological proximity. J Knowl Econ 5:646–667. https://doi.org/10.1007/s13132-014-0209-4
Cliff A, Ord K (1972) Testing for spatial autocorrelation among regression residuals. Geogr Anal 4(3):267–284
Cliff A, Ord K (1981) Spatial Processes, Models and Applications. Pion, London
Coad A, Rao R (2008) Innovation and firm growth in high-tech sectors: a quantile regression approach. Res Policy 37(4):633–648
Coad A, Nielsen K, Timmermans B (2016) My first employee: an empirical investigation. Small Bus Econ 48:25–45. https://doi.org/10.1007/s11187-016-9748-3
Colombelli A, Quatraro F (2019) Green start-ups and local knowledge spillovers from clean and dirty technologies. Small Bus Econ 52(4):773–792
Cohen W M, Levinthal D A (1990) Absorptive capacity: A new perspective on learning and innovation. Administrative science quarterly 35(1), 128-152.
Dealroom (2024). Data on the Milan's tech ecosystem. Date of the last extraction: March 23rd, 2024. Available from: https://innovationmap.yesmilano.it/dashboard
Frenken K, Van Oort FG, Verburg T (2007) Related variety, unrelated variety and regional economic growth. Reg Stud 41(5):685–697
Gentzkow M, Kelly B, Taddy M (2019) Text as data. J Econ Lit 57(3):535–574. https://doi.org/10.1257/jel.20181020
Guan J, Yan Y (2016) Technological proximity and recombinative innovation in the alternative energy field. Res Policy 45:1460–1473. https://doi.org/10.1016/J.RESPOL.2016.05.002

Guerrero AJ, Heijs J, Huergo E (2023) The effect of technological relatedness on firm sales evolution through external knowledge sourcing. J Technol Transf 48(2):476–514. https://doi.org/10.1007/s10961-022-09931-3
Hoberg G, Phillips G (2016) Text-based network industries and endogenous product differentiation. J Polit Econ 124(5):1423–1465
Jaffe AB (1986) Technological opportunity and spillovers of R&d: evidence from firms' patents, profits and market value. Am Econ Rev 76(5):984–1001
Kelejian HH, Prucha IR (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. J Real Estate Finance Econ 17:99–121
King ML (1981) A small sample property of the Cliff-Ord test for spatial correlation. J R Stat Soc Ser B Stat Methodol 43(2):263–264
Kinne J, Lenz D (2021) Predicting innovative firms using web mining and deep learning. PLoS ONE 16(4):e0249071
LeBrasseur R, Zanibbi L, Zinger T (2003) Growth momentum in the early stages of small business start-ups. Int Small Bus J 21:315–330. https://doi.org/10.1177/02662426030213004
LeSage JP, Pace RK (2009) Introduction to Spatial Econometrics. Taylor & Francis
Li D, Heimeriks G, Alkemade F (2021) Recombinant invention in solar photovoltaic technology: can geographical proximity bridge technological distance? Reg Stud 55(4):605–616
Lin C, Wu Y, Chang C, Wang W, Lee C (2012) The alliance innovation performance of R&D alliances—the absorptive capacity perspective. Technovation 32:282–292. https://doi.org/10.1016/J.TECHNOVATION.2012.01.004
Malmberg A, Maskell P (2002) The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering. Environ Plann A 34(3):429–449. https://doi.org/10.1068/a3457
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, Cucculelli M, Cartone A (2024) So far, yet so close Using networks of words to measure proximity and spillovers between firms. Eurasian Bus Rev. https://doi.org/10.1007/s40821-024-00270-x
Martínez M, Zouaghi F, García M (2024) Geographical and cognitive proximity effects on innovation performance: which types of proximity for which types of innovation? Eur Manage Rev. https://doi.org/10.1111/emre.12641
Martin-Rios C, Erhardt NL, Manev IM (2022) Interfirm collaboration for knowledge resources interaction among small innovative firms. J Bus Res 153:206–215. https://doi.org/10.1016/j.jbusres.2022.08.024
Modina M, Capalbo F, Sorrentino M, Ianiro G, Khan M (2023) Innovation ecosystems: a comparison between university spin-off firms and innovative start-ups Evidence from Italy. Int Entrep Manag J. https://doi.org/10.1007/s11365-023-00857-2
Molina-Morales F, García-Villaverde P, Parra-Requena G (2011) Geographical and cognitive proximity effects on innovation performance in SMEs: a way through knowledge acquisition. Int Entrep Manag J 10:231–251. https://doi.org/10.1007/s11365-011-0214-z
Moran PA (1950) Notes on continuous stochastic phenomena. Biometrika 37(1/2):17–23
Nan D, Liu F, Ma R (2018) Effect of proximity on recombination innovation in R&d collaboration: an empirical analysis. Technol Anal Strateg Manag 30(8):921–934
Nathan M, Rosso A (2015) Mapping digital businesses with big data: some early findings from the UK. Res Policy 44:1714–1733
Neffke F, Henning M (2013) Skill relatedness and firm diversification. Strateg Manag J 34(3):297–316
Neffke F, Henning M, Boschma R (2011) How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. Econ Geogr 87(3):237–265
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.
Nooteboom B (2003) Interfirm Collaboration, Learning and Networks, an Integrated Approach Routledge. Routledge, London
Nooteboom B, Van Haverbeke W, Duysters G, Gilsing V, van den Oord A (2007) Optimal cognitive distance and absorptive capacity. Res Policy 36(7):1016–1034

Papagiannidis S, See-To EWK, Assimakopoulos DG, 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

Pavone P, Russo M (2017) Clusters of Specializations in the Automotive Supply Chain in Italy. An Empirical Analysis Using Text Mining, DEMB

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

Rao-Nicholson R, Khan Z, Akhtar P (2019) Nature of technology and location effects on firm performance in the US medical device industry. Econ Innov New Technol 28(5):498–517

Rydehell H, Isaksson A, Löfsten H (2019) Business networks and localization effects for new Swedish technology-based firms' innovation performance. J Technol Transf 44(5):1547–1576. https://doi.org/10.1007/s10961-018-9668-2

Shi ZM, Lee GM, Whinston AB (2016) Toward a better measure of business proximity: topic modeling for industry intelligence. MIS Quart Manag Info Sys 40(4):1035–1056

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. J Econ Geogr 14(2):289–311

Uhlaner LM, van Stel A, Duplat V, Zhou H (2013) Disentangling the effects of organizational capabilities, innovation and firm size on SME sales growth. Small Bus Econ 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. Eur Plan Stud 23(6):1110–1127

Wal T, Anne LJ, Boschma R (2011) Co-evolution of firms, industries and networks in space. Reg Stud 45(7):919–933

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.