Scientometrics
https://doi.org/10.1007/s11192-025-05430-y

Mapping regional knowledge spaces via BERTopic and BERT similarity: the semantic proximity of quantum science in the overall scientific publication landscape

Keungoui Kim[1] · **Dieter F. Kogler**[2] · **Jisoo Hur**[3]

Received: 12 March 2025 / Accepted: 16 September 2025

Abstract

Regional diversification and innovation depend on how well existing knowledge bases align with emerging scientific domains. Conventional measures of relatedness, however, rely on static classification systems that struggle to capture the interdisciplinary and dynamic nature of modern knowledge production. This study introduces *semantic proximity* as a novel, text-based approach to assess the cognitive alignment between regional scientific activity and new fields. Drawing on scientific publications from European NUTS-3 regions between 2010 and 2021, we employ BERTopic and BERT-based similarity to map regional knowledge spaces and measure their proximity to quantum science, an interdisciplinary field of high strategic importance. Using negative binomial regression models, we find that regions with higher semantic proximity to quantum science generate significantly more quantum-related publications. By moving beyond rigid taxonomies, the concept of semantic proximity provides a more flexible and fine-grained tool for analyzing regional knowledge evolution. Our findings extend evolutionary economic geography and offer actionable insights for policymakers seeking to foster smart specialization and guide regional innovation strategies in emerging technologies.

Keywords Knowledge proximity · Semantic proximity · Regional knowledge space · BERTopic · BERT-similarity · Quantum science · Evolutionary Economic Geography · Science specialization · Regional branching

✉ Jisoo Hur
jisoo.hur@tum.de

[1] Humanities, Arts, and Social Sciences Division, Underwood International College, Yonsei University, Seoul, Republic of Korea

[2] Spatial Dynamics Lab, School of Architecture, Planning & Environmental Policy & Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

[3] TUM School of Management, Technical University of Munich, Munich, Germany

Published online: 29 September 2025

Introduction

The accumulation and application of local knowledge are critical drivers of innovation and economic growth in regional economies (Feldman, 1994; Nelson, 1993). Local knowledge, embedded in regional networks comprised of actors, entities, and institutions (Kedron et al., 2020), forms the foundation for technological advancement and economic diversification (Feldman & Kogler, 2010). The principle of relatedness builds on this notion, highlighting how regions can leverage existing knowledge bases to explore new, albeit related, technological domains in order to expand their technological knowledge portfolios (Hidalgo et al., 2018). This concept asserts that the likelihood of a region successfully developing a new technology depends on its alignment with the region's existing knowledge base. Rooted in the Evolutionary Economic Geography (EEG) paradigm (Kogler et al., 2023a), this principle has informed both scientific inquiry and policy frameworks, most notably the European Union's Smart Specialization Strategy policy initiative (Kim, et al., 2024a, 2024b, 2024c; Rigby et al., 2022; Santoalha, 2019).

Empirical studies often quantify relatedness through the co-occurrence of technological classes in patent data, typically using classification systems like the International Patent Classification (IPC), Cooperative Patent Classification (CPC), or alternative schemes (Boschma et al., 2015; Kogler et al., 2013). These hierarchical knowledge classification systems enable researchers to map regional knowledge spaces, to identify technological adjacencies, and to infer the direction of knowledge production and innovation trajectories (Kogler, 2015). Typically, relatedness is measured by examining how frequently two or more technological classes co-appear within the same patent document, normalized against their overall baseline occurrence (Boschma et al., 2015; Kogler et al., 2017; Rigby, 2015). Such analyses shed light on the compatibility and recombinant frequency between technologies and provide insights into the dynamics of regional innovation systems.

Such analyses have established that technologies more closely related to a region's existing capabilities are more likely to emerge, while unrelated ones tend to fade (Boschma et al., 2015; Kogler et al., 2017). Boschma et al. (2014), for instance, demonstrated that new biotechnology topics were more likely to appear in cities already hosting semantically related scientific activity, reinforcing the principle of path dependency in regional knowledge evolution.

Despite these advancements, current approaches to measuring relatedness face several notable limitations. Foremost is their reliance on static classification systems such as the IPC and CPC, which, despite their utility, are manually curated and usually struggle to keep up with the rapid evolution of technology. These systems provide a static and sometimes incomplete snapshot of knowledge domains, potentially overlooking novel and interdisciplinary trends. Likewise, studies that utilize broad and static scientific knowledge categorizations, such as the science subject categories provided in the Web of Science database, usually fail to capture the contextual richness of scientific contributions, especially in emerging fields. This challenge is further exacerbated by the nature of modern discoveries, which frequently occur at the intersection of disciplines, making them difficult to fit within traditional classification schemas (Fagerberg et al., 2012). Yet, these interdisciplinary breakthroughs are often at the forefront of innovation, combining established knowledge into novel forms (Eisenhardt & Martin, 2000; Lee et al., 2015; Schumpeter, 1942, 2021). Another significant challenge stems from the potential biases introduced by the systemic approach in the assignment of classifications (Kogler et al., 2024). The novelty of a patent application is examined by searching prior art, and where relevant it is then the technological knowledge domains contained in prior patents, i.e., patent citations, that largely determine the various classification codes that are attributed to a new invention. Thus, there is an inherent path-dependency ingrained in the system that favours structural consistency, even if it might become outdated, over structural dynamics that would better reflect the evolution of technological change (Kogler et al., 2024). Similarly, journal categorizations are vulnerable to subjective interpretation during manual assignments. These biases can undermine the accuracy of relatedness measures, leading to distorted outcomes and in parallel raising endogeneity concerns.

To overcome these constraints, this paper introduces a new concept: semantic proximity. This approach evaluates the relatedness of a region's existing knowledge base to potential new technologies by analyzing the actual semantic content of scientific outputs rather than relying on prescribed, static categorizations. Semantic proximity provides a more dynamic and comprehensive understanding of knowledge alignment by moving beyond traditional classification systems to incorporate textual and contextual information. We apply this method to the field of quantum science, an inherently interdisciplinary field that draws from a wide range of scientific and technological domains, making traditional classification methods insufficient to fully capture its complexity. Recently, quantum science has garnered significant attention and investment at the national level, with governments worldwide recognizing its transformative potential across industries such as computing, cryptography, and materials science. This scientific area is particularly well-suited for the application of semantic proximity due to its inherently interdisciplinary nature, and equally due to recent significant government leadership and intervention actions that aim at driving rapid advancements in this growing science field.

Utilizing scientific journal publications listed in the Web of Science (WoS) database, and geocoded at the NUTS-3 level across European regions from 2010 to 2021, we identify the key knowledge domains within both, general scientific fields and quantum science.[1] Specifically, we apply unsupervised topic modelling to classify research topics from large volumes of scientific publications, enabling a systematic analysis of regional knowledge bases. By leveraging an embedded topic modelling approach, i.e., BERTopic, which is built on Bidirectional Encoder Representations from Transformers (BERT), we gain deeper insights into the composition and structure of regional expertise (Devlin et al., 2019). These identified topics are further refined using generative AI to produce precise and contextually relevant labels. Subsequently, and based on these labelled topics, we measure the semantic proximity between a region's existing knowledge base and quantum science. This, in turn, allows us to quantify the alignment between regional expertise and the emerging field of quantum science, offering valuable insights into the potential for regional specialization and development in this strategically important scientific field. In summary, this study examines and refines the stylized findings that a region's pre-existing knowledge and capabilities are pivotal in its ability to specialize in an emerging field via a novel approach to assess knowledge alignment through semantic proximity. In particular, we test the hypothesis that greater semantic proximity to a region's knowledge base is positively associated with the generation of new knowledge. To explore this in more detail, the field of quantum

[1] For further information about the European NUTS (Nomenclature of territorial units for statistics) classification system, see: https://ec.europa.eu/eurostat/web/nuts. Restrictions apply to the publication dataset used in this paper. The Web of Science data is owned by Clarivate Analytics. To obtain the bibliometric data in the same manner as authors, one needs to contact Clarivate Analytics at "https://clarivate.com/webofsciencegroup/solutions/web-of-science/contact-us/" in order to gain access.

science is employed as a case study. It is reasonable to expect that results will confirm prior findings and indicate that regions with knowledge more semantically aligned with quantum science are also the ones more likely to generate new knowledge in this emerging field. However, and in contrast to existing methods to measure knowledge relatedness, it is also to be expected that the suggested semantic proximity approach will offer a more nuanced and dynamic method to capture emerging and especially inter-/multi-disciplinary science/technology fields, and applied in the context of regional economies development trajectories will offer a significant improvement to the current state-of-the-art and thus also provide more advances insights towards policy actions aiming to advance regional innovation systems.

The article is structured as follows: The next section provides the theoretical background on regional specialization and existing measures of knowledge proximity. Sect. "Data and methods" describes the data sources and methodology in detail, including the measure of semantic proximity. Sect. "Discussions & concluding remarks" introduces and analyzes the results obtained. Finally, Sect. 5 discusses the main findings of the study, along with its limitations and suggestions for future research.

Theoretical background

Regional specialization and evolutionary economic geography

Regional specialization and competitiveness, paired with policy relevance in the realm of science, technology, and innovation, are central tenets in the Evolutionary Economic Geography paradigm (Kogler et al., 2023a), emphasizing the role of pre-existing knowledge, capabilities, and spatial configurations in shaping economic trajectories. Regions develop distinct economic profiles through the interplay of localized skills, technological capabilities, and industrial structures. This process is almost exclusively path-dependent, meaning that existing knowledge bases and industrial landscapes significantly influence opportunities and limitations for the diversification into new activities (Grabher, 2009; Kogler, 2015; Kogler et al., 2023b; Martin & Sunley, 2006). The "stickiness" of knowledge, its tendency to remain geographically localized, reinforces these trajectories, fostering the clustering of knowledge production and innovation in specific regions and times (Feldman & Kogler, 2010; Jaffe et al., 1993; von Hippel, 1994).

While advances in transportation and information technologies are believed to have mitigated some geographical constraints in that context, spatial differences remain critical in shaping regional innovation potential and, in turn, result in uneven spatial development. Proximity to knowledge hubs determines the extent of knowledge spillovers, perpetuating disparities in innovation outcomes across regions (Jaffe et al., 1993; Sonn & Storper, 2008). These dynamics often give rise to agglomeration of specialized commercial and industrial activities due to localized and dense concentrations of interconnected firms and institutions. As a result, such clusters act as innovation ecosystems, enabling resource sharing, knowledge diffusion, and synergy creation, analogous to spatial strategic alliances (Cooke et al., 2007; Feldman & Choi, 2015; Gertler, 2003).

Within such clusters, innovation is driven not only by sheer physical proximity but also by resulting social and cognitive proximity arrangements that are considered essential for effective collaboration and knowledge exchanges to take place (Boschma, 2005). Social proximity reflects the role of interpersonal relationships and trust in fostering cooperation, while cognitive proximity pertains to the shared understanding of knowledge bases, facilitating communication and innovation. Together, these forms of proximity drive collaboration at various levels, i.e., individual, organizational, industrial, and regional, enabling actors to share risks, pool resources, and enhance innovation outcomes (Ahammad et al., 2016; Broekel, 2012; Broekel & Boschma, 2012; Huber, 2012).

Cognitive proximity is particularly crucial for regional innovation, as it enables the recombination of diverse expertise to create new technologies (Nooteboom, 2000). This principle underpins the concept of relatedness, which measures the similarity and compatibility between activities based on shared capabilities and knowledge bases. Relatedness explains how new technologies emerge through the recombination of existing knowledge, fostering diversification and driving regional economic growth, and further empirical evidence shows that regions with highly related knowledge bases are better positioned to sustain technological development and economic growth while also being more resilient to economic shocks demonstrate that regions with high relatedness in their knowledge bases exhibit greater economic resilience, stronger technological evolution, and enhanced specialization (Neffke et al., 2011; Tóth et al., 2022). Although these insights are highly relevant and have advanced the line of inquiry concerned with evolutionary development trajectories at the intersection of regional knowledge production and socio-economic outcomes, all associated studies heavily rely on methodologies that measure relatedness via the co-occurrence of knowledge domains specified in static hierarchical classification systems and thus most likely underestimate, if not ignore, the dynamic nature of regional knowledge production and innovation processes; something the proposed semantic proximity concept aims to address.

Knowledge relatedness and regional branching into new and emerging knowledge domain specializations

Historically, early attempts to measure the cognitive and relational distance or relatedness between distinct knowledge domains have been carried out across multiple disciplines, including science and technology studies (Engelsman & Van Raan, 1994; Kopcsa & Schiebel, 1998), economics (Jaffe, 1986; Teece et al., 1994), management studies (Almeida, 1996; Makri et al., 2010), and regional science (Fischer et al., 2006; Quatraro, 2010). In parallel, the relevant literature also started to engage in an number of fundamental discussions concerning "the definition and operationalization of concepts such as relatedness of scientific (or Technological subfields" (Van Raan, 1997), which continues, in particular also in the context of aspects that might result in the variations and potential implications derived from such studies (Milojević, 2015). From a territorial innovation systems perspective, and based on evolutionary thinking (Nelson & Winter, 1985), the literature encompassing Regional Innovation Systems (RIS) took particular interest in how regional technological knowledge structures determined by relatedness measure might explain uneven place-specific socio-economic outcomes (Cooke, 1998; Morgan, 1997). Here the focus has been primarily on technological relatedness and regional knowledge spaces as measured either by hierarchical or co-occurrence relatedness measures (Whittle & Kogler, 2020).[2]

[2] Whittle and Kogler (2020) offer a detailed overview and critical review regarding contemporary studies that focus on relatedness in the context of methodological approaches and associated empirical evidence.

Hierarchical relatedness approaches leverage the hierarchical structure of patent, industry, or product classification schemas to decompose observed variations into within-category (related variety) and between-category (unrelated variety) components (Frenken et al., 2007). While conceptually useful, this approach imposes top-down structures that may obscure emerging or nuanced interdependencies between technologies, while another limitation is that these models typically treat relatedness as a binary, i.e., either related or unrelated, and thus fail to capture gradations in technological proximity (Barbero et al., 2024; Santoalha, 2019). Relatedness measures based on co-occurrence patterns manage to overcome some of these limitations by employing a bottom-up approach that identifies related technologies based on their co-occurrence within the same patent document, or via the probabilities that patents in a certain technological knowledge domain, i.e., CPC class, will cite patents in other classes (Kogler et al., 2013, 2017; Rigby, 2015). Building on this foundation, the concept of relatedness density introduces a spatial perspective, evaluating how closely new or disappearing technologies align with a region's existing technological capabilities (Boschma et al., 2014; Rigby et al., 2022). A high relatedness density implies strong potential for regional adoption and specialization, while low values signal misalignment. This measure assists in tracing technological trajectories, providing insights into how regional knowledge can be recombined to support and advance future regional innovation processes. In addition to quantitative indicators, knowledge space networks offer a visual framework for exploring technological proximity. Nodes represent technologies (e.g., patent classes), and links indicate the degree of relatedness. These networks map the structure of regional innovation systems and reveal strategic pathways for knowledge recombination (Broekel et al., 2014; Kogler et al., 2023a, 2023b; Ter Wal & Boschma, 2009) Centrality measures such as weighted degree and betweenness help identify core and bridging technologies, respectively, offering insight into their role in driving regional diversification (Jung et al., 2021; Kim et al., 2019, 2024a, 2024b, 2024c).

While existing measures of regional knowledge relatedness provide valuable insights into how a region's pre-existing knowledge base aligns with potentially new to the region or even emerging technologies, they have notable limitations. First, these measures often rely on static classification systems such as CPC or IPC, which struggle to keep pace with the rapid evolution of technology. Such classification systems may fail to capture emerging trends and contextual nuances in regional knowledge bases, leading to an incomplete understanding of technological alignment. Second, patent classification is typically conducted manually by patent examiners, who rely on prior art and existing classification codes. Since these codes are largely determined by the technological knowledge domains of prior patents (i.e., patent citations), this process introduces an inherent path dependency. As a result, the system tends to favour structural consistency, even if it becomes outdated, over structural dynamics that would better capture the evolution of technological change (Kogler et al., 2023a). These classifications can be influenced by subjective interpretations, organizational strategies, or contextual factors, which may distort the accuracy of relatedness measures. Third, and concerning scientific knowledge, while scientific publications are an important reflection of regional knowledge production and technological capabilities, they are often not fully leveraged in relatedness analyses. This is because studies using scientific journal publications typically rely on broad subject categories, which may not capture the detailed nuances of emerging fields (Mongeon & Paul-Hus, 2016). For example, the categorization system used by the WoS database includes 177 broad subject categories that may fail to reflect the specificities and contextual depth of scientific contributions, particularly in rapidly evolving fields. As a result, they may not provide deep insights into regional knowledge, particularly in areas where academic research plays a critical role in driving innovation (Shin et al., 2023). These limitations underscore the need for more comprehensive and adaptive approaches to measuring regional knowledge proximity with the ability to capture the dynamic and multifaceted nature of knowledge production.

To address these challenges, this study introduces semantic proximity as a more adaptive and nuanced measure of regional knowledge alignment. Rather than depending on rigid, predefined categories, semantic proximity evaluates the textual and conceptual similarity between knowledge domains by analyzing the language used in scientific outputs. Specifically, this method enables a fine-grained assessment of relationships between existing and emerging knowledge, the recognition of interdisciplinary linkages, and a more accurate reflection of contemporary innovation dynamics.

By incorporating semantic analysis, we propose a framework that better captures the evolving, multidimensional nature of regional innovation systems. This approach not only contributes methodologically but also offers practical insights for policymakers and regional planners aiming to foster innovation. In that context, we apply this approach to the field of quantum science, an interdisciplinary domain characterized by rapid growth and high strategic value. Quantum science exemplifies the kinds of challenges static classification systems fail to address, given that its development draws on physics, materials science, computing, and engineering, making it ideal for testing the power of semantic-based methods. We hypothesize the following:

Hi. The semantic proximity of target scientific knowledge (e.g., quantum science) to a region's scientific knowledge base is positively associated with the development of new scientific knowledge in that field.

This hypothesis links the conceptual framework of semantic proximity to regional innovation potential, enabling us to test how well textual and contextual alignment predicts scientific specialization and knowledge emergence.

Data and methods

Semantic proximity

This section outlines the fundamental process and associated principles underlying our suggested novel measurement approach. The proposed semantic proximity measure aims to capture the proximity between a target science field vis-à-vis a region's entire spectrum of practiced science fields, incorporating a contextual understanding of science itself. To empirically capture this concept, we suggest a measurement approach that includes three stages: topic modelling, label generation, and similarity measurement (Fig. 1).

First, topic modelling is applied to abstracts from two data sets: the region's overall science and the region's target science. In the context of this study, topic modelling plays a critical role in identifying semantically coherent and interpretable topics that represent the latent structure of scientific activity in each region. Rather than analyzing documents or keywords in isolation, the clustering step enables us to derive topic distributions that characterize a region's knowledge profile. More importantly, generated topics serve as a structured basis for comparing how a region's emerging scientific domains align with its overall scientific landscape. In this way, topic modelling is a necessary step for transforming unstructured text data into interpretable representations that allow us to quantify and compare regional specialization patterns and semantic alignment.

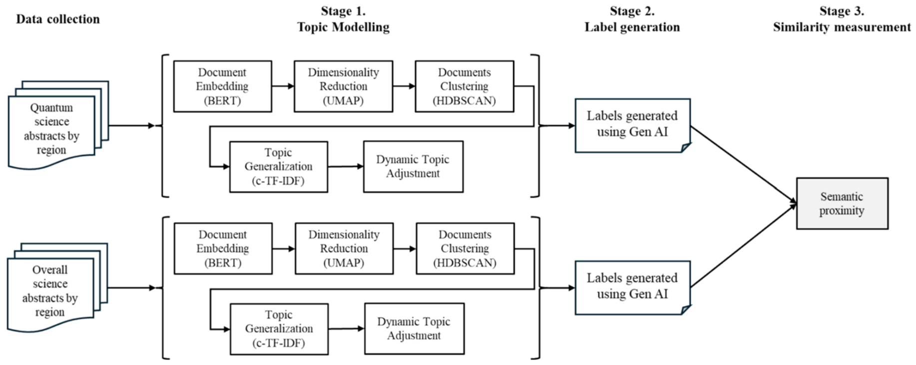

Fig. 1 Process of measuring semantic proximity

In our study, we used BERTopic, a topic modelling technique that integrates BERT-based document embedding, dimensionality reduction, clustering, and class-based TF-IDF (c-TF-IDF) (Grootendorst, 2022). Compared to the conventional Bag of Words (BoW) approaches that rely solely on term frequency within documents, this method enables a deeper contextual understanding of the focal documents. This is particularly advantageous when analyzing a large corpus of text, like scientific publications, that requires a nuanced semantic interpretation (Kim et al., 2024a, 2024b, 2024c). In this study, "all-MiniLM-L6-v2", a lightweight BERT-based transformer model, is used for the document embedding process, as its performance and application quality have been proven in a previous study (Kim et al., 2024a, 2024b, 2024c). Prior to clustering, Uniform Manifold Approximation and Projection (UMAP) is applied to reduce the dimensionality of the embeddings while preserving the essential semantic structure. UMAP has been widely used in topic modeling pipelines for its ability to retain both local and global relationships in high-dimensional embedding spaces. While dimensionality reduction inevitably involves a degree of abstraction, we addressed potential concerns of semantic oversimplification by fine-tuning UMAP parameters. This preliminary step enhances the coherence and accuracy of the derived topic clusters. Subsequently, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is employed to group documents that share latent topics, enabling the identification of semantically coherent themes across the documents of interest. Finally, representative keywords for each topic are extracted using c-TF-IDF, which enhances interpretability by highlighting terms that are not only frequent within a topic but also distinctive compared to others. The c-TF-IDF quantifies the importance of words within a specific topic by considering their frequency within a topic and their rarity across other topics, allowing for the identification of a topic-specific vocabulary. Employing this measure, the top 10 keywords with the highest c-TF-IDF scores are selected in descending order for each topic to capture its semantic characteristics. Like in any unsupervised learning approach, optimizing hyperparameters is an essential task. In this regard, we set the maximum number of clusters to a value of 10, and the minimum topic size to 1.5% of all publications. Following several rounds of testing to determine these optimized parameter settings, it became conclusive that this setup allows us to obtain a minimum of 3 to a maximum of 7 topics while also minimizing the number of outliers. Through this process, a list of topic keywords is generated for each dataset, categorized by region and six distinct two-year periods ranging from 2010 to 2021.

Using these sets of topic modelling keywords, prompt engineering techniques with generative AI are applied to create labels for each topic. While topic modelling provides a prioritized list of keywords representing each topic, summarizing these keywords into coherent and descriptive labels is a necessary step. When dealing with numerous topic modelling results and multiple trials, maintaining consistency and quality of the outcomes becomes crucial. By utilizing prompt engineering tailored to this task, we generated structured and well-summarized representations of the topic keywords. Among the various existing generative AI models, we selected the GPT-4-o model to implement our prompts,[3] successfully generating topic labels for all identified topic keywords. These generated topic labels can therefore be regarded as concise summaries of both the region's overall scientific landscape and its target science, reflecting their contextual backgrounds.

In a final step, we measured the semantic proximity (Eq. 1) by calculating the similarity between a region's overall science labels and target science labels. This comparison allows us to assess the extent to which the target science is semantically embedded within the broader scientific landscape of the region. To assess the semantic relationship between the two sets of topics, BERT-based similarity, which computes cosine similarity within a contextualized embedding space, is used. Unlike frequency-based distance measures, it allows for a more nuanced comparison of text by capturing semantic meaning beyond surface-level word co-occurrence. For topic embedding, "all-MiniLM-L6-v2" is employed. The semantic proximity of a region i in period t is measured as the cosine similarity between two embedding vectors: the region's overall science (S_1), and its target science (S_2). This is calculated as the dot product of S_1 and S_2, divided by the product of their magnitudes.

$$Semantic\,proximity_{i,t} = \frac{\sum_{j=1}^{n} S_{1j} \cdot S_{2j}}{\sqrt{\sum_{j=1}^{n} S_{1j}^2} \cdot \sqrt{\sum_{j=1}^{n} S_{2j}^2}} \tag{1}$$

Data

To enable the empirical analysis, it was necessary to collect and process information concerning scientific publications, and quantum science publications in particular, that were produced across European regions from the WoS database. The initial step was to determine if a publication record was associated with a European institution. To do so, the address information provided in the WoS database was geocoded by converting institutional addresses into geolocation data (latitude and longitude). These geolocation data were then mapped to European NUTS classification codes, allowing us to identify and analyze publication records from researchers affiliated with European institutions between 2010 and 2021. Considering the time typically required for publications to be processed and indexed, it appears reasonable to focus on two-year time intervals in the subsequent analysis. In parallel, additional information regarding each publication, derived from various WoS sub-tables that offer details on authors, institutions, funding, and science subject categories, was collected as well. In a second step, the aim was to identify publication records that are directly

[3] The following prompt has been used with ChatGPT (GPT-4): "The list of keywords is the outcome of topic modelling on journal articles related to quantum science. Based on the keywords, generate a single label that best describes the topic of the articles. No descriptions are required."

related to quantum science. For this purpose, a list of quantum keywords from the Korea Institute of Science and Technology Information (KISTI) report,[4] which also used WoS to analyze publication patterns in the field of quantum science, was applied. In a further step, the unit of analysis was determined by further converting the geolocation data into a NUTS-3 level regional dataset. This process included creating NUTS-3 level science-related variables and adding socio-economic regional data obtained from the Annual Regional Database of the European Commission's Directorate General for Regional and Urban Policy (ARDECO). One challenge during this process was the finding that not all regions had produced quantum science publications; perhaps not surprising given the highly complex and advanced nature of this field. However, in the analysis that follows, this may cause unexpected results for two key reasons. First, regions without quantum science publications will have a semantic proximity value of 0. Second, a topic modeling analysis based on very few publications may lead to misinterpretation. To address these issues, only the top 100 NUTS-3 regions (see Appendix 1), ranked by the number of quantum science publications, were selected for further analysis.

Research model

In this study, we econometrically examine the relationship between semantic proximity and the creation of new knowledge at the regional level, using NUTS-3 region-period data (see Table 1). To operationalize new knowledge creation, we focus on the field of quantum science, an emerging, interdisciplinary domain that has gained substantial global policy attention and research investment over the past decade. Quantum science was chosen deliberately because it remains at a relatively early stage of development compared to other domains, such as Artificial Intelligence (AI) or more general Information and Communication Technologies (ICT), which have already become widely adopted general-purpose technologies. These more mature fields are now so embedded across virtually all scientific and industrial domains that measuring their relatedness to regional knowledge bases yields little meaningful variation. In contrast, quantum science retains distinct boundaries, allowing us to capture how strongly its emergence is shaped by alignment with existing regional knowledge capabilities. As such, it provides a suitable empirical setting for testing whether and how semantic proximity facilitates regional specialization in newly emerging domains.

The dependent variable, *quantum knowledge*, is measured by the number of quantum-related journal publications in each region for a given period. This metric serves as a proxy for new knowledge creation in the field of quantum science. The choice of this dependent variable is substantively grounded in both the nature of quantum science and the broader literature on regional innovation and relatedness. First, in emerging and frontier scientific domains such as quantum science, journal publications are often the most direct and timely outputs of new knowledge generation, preceding commercialization or patenting activities. Unlike patents, which are more common in applied or engineering-driven domains, quantum science remains highly academic and exploratory, making publications a more accurate indicator of innovation and regional capacity-building in this field. Second, our focus

[4] https://astinet.kr/reports/data-insights/289658

Table 1 Description of variables

	Variables	Description	Source
Dependent variable	Quantum knowledge	Number of quantum-related publications	WoS
Independent variables	Semantic proximity	Semantic proximity between quantum-related and general knowledge	WoS
Control variables	Publications	Natural logarithm of the total number of publications	WoS
	Subjects	Number of science subjects	WoS
	Quantum funding	Ratio of quantum-related funding to total funding	WoS
	GDP	Natural logarithm of regional gross domestic product per capita	ARDECO
	Population	Natural logarithm of the total population	ARDECO
	Metro	Dummy: 1 if metropolitan region; 0 otherwise	Eurostat
	Adjacent regions	Number of adjacent regions	Constructed by authors
	Period	Dummy for each two-year period (2010–2021), controlling for time-fixed effects	WoS

on publication counts aligns with established research in evolutionary economic geography and relatedness studies, which frequently use publication data to track the emergence of new scientific activities (e.g., Boschma et al., 2014; Heimeriks & Boschma, 2014).

To test our hypothesis, we include *semantic proximity* as our key independent variable. This variable captures the alignment between a region's general scientific expertise and the domain of quantum science, as measured using BERT-based semantic similarity between topic model labels generated from each body of research. We expect that higher semantic proximity is positively associated with quantum knowledge production. To account for regional heterogeneity in the factors influencing scientific knowledge production, we include several control variables. First, the *number of publications* in a region serves as a proxy for the overall research output and baseline scientific activity. Second, the *number of distinct scientific subjects* present in a region captures the breadth of its research landscape, reflecting the potential for interdisciplinary knowledge recombination. Third, *quantum funding intensity*, defined as the share of funding allocated to quantum-related publications relative to total research funding, reflects the degree of financial support specifically targeted at quantum science. We also include *GDP per capita* to control for overall regional economic development, which may influence research capacity and infrastructure. Fifth, *population* is included to account for the scale of the regional labor force and demand-side factors that may influence knowledge production. Sixth, a *metropolitan* dummy variable distinguishes urban from non-urban regions, accounting for the typically greater availability of research infrastructure, collaboration opportunities, and institutional density in metropolitan areas. Additionally, the *number of adjacent regions* is included to capture potential spatial spill-over effects, under the assumption that neighboring regions may influence each other's scientific development through proximity-based collaboration and knowledge diffusion (Hoekman et al., 2009; Jaffe et al., 1993). Lastly, *period* fixed effects (two-year dummies from 2010 to 2021) are included to control for time-specific influences such as global trends in quantum research activity and scientific publishing. The variables for publications, GDP per capita, and population are log-transformed to reduce skewness and account for their wide numerical range.

Table 2 presents descriptive statistics and correlations. Considering that some variables in Table 2 exhibit relatively high correlations, we assessed variance inflation factors (VIFs) to check for multicollinearity. The analysis shows that all VIF values are below 5, suggesting minimal multicollinearity issues. Additionally, we performed supplementary analyses excluding the highly correlated variables, and the findings were consistent.

A negative binomial regression model is employed for econometric estimation since our dependent variable is a count outcome. This approach accounts for the discrete and non-negative nature of the dependent variable while addressing potential overdispersion in the data. Time-period dummy variables are included in all models to account for variations in quantum-related journal publications across different time periods. In addition, to confirm the robustness of our results, we conducted additional analyses using robust standard errors and cross-sectional time-series feasible generalized least squares (FGLS) regression with the ratio of quantum publications, i.e., the share of quantum publications relative to overall science publications.

Table 2 Descriptive statistics and correlations of variables

		1	2	3	4	5	6	7	8	9
1	Quantum knowledge	1.000								
2	Semantic proximity	0.037	1.000							
3	Publications	0.654	−0.054	1.000						
4	Subjects	0.305	−0.122	0.770	1.000					
5	Quantum funding	0.103	−0.027	−0.428	−0.454	1.000				
6	GDP	0.221	−0.039	0.278	0.160	−0.008	1.000			
7	Population	0.315	−0.077	0.567	0.490	−0.306	−0.301	1.000		
8	Metropolitan	0.005	−0.040	0.201	0.219	−0.057	0.220	0.244	1.000	
9	Adjacent regions	0.145	−0.036	0.118	0.064	−0.152	−0.013	0.311	−0.110	1.000
Obs		600	600	600	600	600	516	516	600	600
Min		12	−0.025	1317	155	0.001	11,456.9	103,679	0	1
Max		328	0.168	48,902	254	0.049	621,284	6,757,025	1	11
Mean		77.387	0.048	10,199.6	232.378	0.011	57,201.5	1,018,489	0.85	4.47
S.D		54.681	0.029	7520.2	18.592	0.008	64,403.7	1,130,296	0.357	2.608

Untransformed values are used for descriptive statistics. The VIF values range from 1.04 to 4.02, with an average of 2.01

Results

Semantic proximity results

Employing the proposed semantic proximity measurement, semantic proximity to quantum science in each of the 100 European NUTS-3 regions is measured[5] while Table 3 presents detailed results of the case of CH013 (Geneva) in 2020–2021. As shown in Table 3, topic words are first generated, and then summarized labels are created using those words. These labels reflect the main topics extracted from the abstracts using BERTopic analysis, which applies topic modeling techniques to identify prominent themes within the texts. The labels were then used to calculate BERT-based similarity, enabling a comparison of semantic overlaps between overall science and quantum science abstracts. When averaging by region, the semantic proximity was found to be around 0.057, with Geneva having the highest value at 0.099. Although the semantic proximity values appear to be quite low, this can be considered reasonable given the diversity of scientific activities within each region and the fact that quantum science and technology is still in its early stages. Considering the presence of several quantum science research institutes in Geneva, such as the Geneva Quantum Center, Quantum Information & Communication at the University of Geneva, the Open Quantum Institute, and Swiss Quantum Centers and Initiatives, our measurement certainly reflects present realities.

Providing a more universal, pan-European perspective on the regional distribution of quantum science knowledge production and the level of semantic proximity amongst the top 100 regions of interest, Fig. 2 offers further insights. The map to the left highlights the proportion of quantum-focused publications as a share of the total number of publications over the entire observed time period, 2010–2021. Of the over 1350 European regions investigated, one-third did not produce any quantum publications, while the share of quantum publications amongst approximately 750 regions was less than 1% if all their scientific publication outputs are considered. This leaves only 100 European regions where the quantum science publication share was at least 1% or more based on their entire scientific publication outputs, respectively. The region with the highest share among those latter groups of NUTS3 regions is AT332 Innsbruck, Austria. This might be surprising given that Innsbruck is only a medium-sized city situated in the middle of the Alps, but given the relative importance of the University of Innsbruck as the most important research and educational institution in western Austria, combined with the fact that the institution's Physics department with particular focus on quantum and particle physics is highly ranked in Europe and globally, makes this finding indeed plausible. The largest city near the top of the list of places that have the highest proportion of quantum science outputs in their entire scientific publication portfolio is Munich, Germany. The map to the right in Fig. 2 highlights the derived BERT-based similarity measures for each of the focal 100 regions. The top-ranked region, i.e., CH013, Geneva, was already described in detail above. The list of places that follow in the ranking is composed of a mix of German and Italian cities, e.g., Trieste, Darmstadt, Padova, and Karlsruhe. At the bottom of the ranking, i.e., places that, despite having produced a notable amount of quantum science publications, display low semantic proximity to quantum science when considering their entire regional scientific knowledge portfolio, are Gothenburg, Lille, Leeds, Leiden, and Basel. The top 10 regions in the

[5] All results are available at the following link: https://awekim.shinyapps.io/SemanticProximity_shiny/

Table 3 Example of labels generated for overall science and quantum science abstracts

Category	Generated Topic Words	Generated Labels	BERT-based similarity
Overall science (except quantum science)	[‘patients’, ‘health’, ‘treatment’, ‘clinical’, ‘countries’, ‘ci’, ‘disease’, ‘care’, ‘19’, ‘related’]	International Clinical Health Care and Treatment of Disease in Patients with COVID-19	0.099
	[‘mass’, ‘energy’, ‘measurements’, ‘proton’, ‘detector’, ‘beam’, ‘large’, ‘galaxies’, ‘lhc’, ‘star’]	Particle Physics and Astrophysics Measurements in High-Energy Collisions	
Quantum science	[‘cng’, ‘gpps’, ‘computing’, ‘lectures’, ‘applications’, ‘sequential’, ‘plasmon’, ‘plasma’, ‘interband’, ‘polaritons’]	Quantum Plasmonics and Sequential Computing Applications	
	[‘bell’, ‘correlations’, ‘steering’, ‘independent’, ‘state’, ‘states’, ‘network’, ‘entangled’, ‘channels’, ‘measurement’]	Quantum Correlations and State Steering in Entangled Networks	
	[‘key’, ‘qkd’, ‘security’, ‘source’, ‘rate’, ‘photon’, ‘single’, ‘distribution’, ‘channel’, ‘detector’]	Quantum Key Distribution and Photon Detection in Secure Communication Channels	
	[‘temperature’, ‘topological’, ‘materials’, ‘transition’, ‘insulator’, ‘correlated’, ‘metal’, ‘electronic’, ‘monopoles’, ‘superconductivity’]	Quantum Materials and Phase Transitions in Topological Insulators and Correlated Systems	
	[‘mml’, ‘cooling’, ‘laser’, ‘antihydrogen’, ‘mspace’, ‘collisions’, ‘proton’, ‘cross’, ‘antimatter’, ‘section’]	Antimatter Cooling and Collision Dynamics in Quantum Physics	
	[‘spin’, ‘storage’, ‘rare’, ‘memory’, ‘optical’, ‘earth’, ‘cef’, ‘long’, ‘ms’, ‘spins’]	Quantum Spin Memory Storage in Rare Earth Optical Materials	
	[‘poisson’, ‘operators’, ‘feynman’, ‘cft’, ‘eft’, ‘gauge’, ‘operator’, ‘ads’, ‘loops’, ‘bulk’]	Quantum Field Theory in AdS/CFT Correspondence with Effective Field Theories and Poisson Operators	
	[‘stochastic’, ‘production’, ‘entropy’, ‘thermodynamic’, ‘driven’, ‘steady’, ‘height’, ‘ssep’, ‘coherences’, ‘fluctuation’]	Stochastic Thermodynamics of Steady-State Fluctuations in Quantum Systems	
	[‘body’, ‘tdvp’, ‘semiclassical’, ‘regular’, ‘interacting’, ‘mixed’, ‘variational’, ‘mbl’, ‘thermalizing’, ‘initial’]	Variational Approaches in Semiclassical Analysis of Many-Body Localized States and Thermalization	
	[‘magnetoresistance’, ‘strain’, ‘jacutingaite’, ‘weight’, ‘carrier’, ‘engineering’, ‘indicate’, ‘ev’, ‘monolayer’, ‘band’]	Strain-Tunable Magnetoresistance in Monolayer Jacutingaite for Carrier Engineering	

*Case of CH013, Geneva in 2020–2021

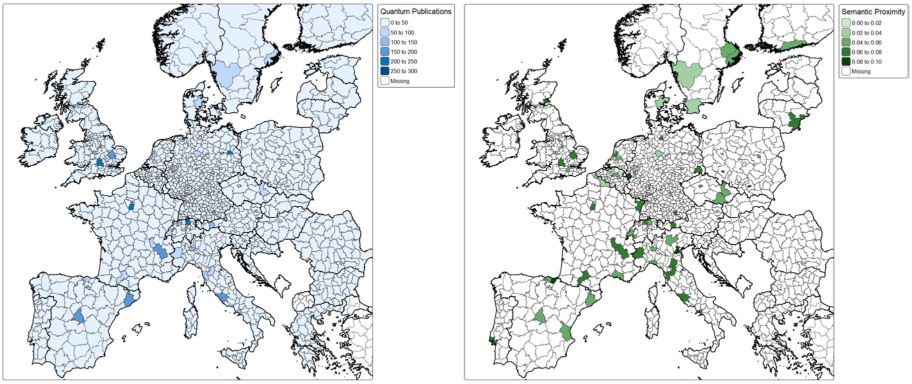

Fig. 2 Map of EU quantum science (left—quantum publication count; right – semantic proximity)

sample in terms of quantum publication outputs, as well as semantic proximity values, are highlighted in Table 4. All 100 regions included in the analysis are listed in Table 1 that is featured in the Appendix.

Regression results

Table 5 presents the results from the negative binomial regression models. The first column presents results only with the control variables. The number of *publications* has a positive and statistically significant effect on quantum-specific journal publications, suggesting that regions with higher overall scientific output are more likely to contribute to emerging fields like quantum science. The coefficients for the number of *subjects* are significantly negative in all models, suggesting that regions with a broader disciplinary spread may be less focused in their research efforts, potentially diluting resources away from specialized fields like quantum science. This finding implies that while diversity in research domains

Table 4 Top 10 regions of quantum publication count and semantic proximity

Quantum publication count			Semantic Proximity		
Rank	NUTS3	Region name	Rank	Region	Region name
1	FR101	Paris	1	CH013	Geneva
2	FR104	Essonne	2	ITH44	Trieste
3	CH040	Zürich	3	DE711	Darmstadt
4	DE300	Berlin,	4	ITH36	Padova
5	UKJ14	Oxfordshire	5	DE122	Karlsruhe
6	ES511	Barcelona	6	ES212	Gipuzkoa
7	ES300	Madrid	7	DE263	Würzburg
8	UKH12	Cambridgeshire	8	DEA26	Düren
9	FR714	Isère	9	DEA2D	Städteregion Aachen
10	DED21	Dresden	10	PL213	Miasto Kraków

Table 5 Regression results (standard errors)

Dependent variable: Regression model:	Quantum knowledge Negative binomial regression	
	(1)	(2)
Semantic proximity		1.930*** (0.509)
Publications (ln)	1.063*** (0.041_	1.045*** (0.041)
Subjects	−0.008*** (0.001)	−0.008*** (0.001)
Quantum funding	40.218*** (2.387)	40.893*** (2.350)
GDP (ln)	−0.064* (0.036)	−0.052 (0.035)
Population (ln)	−0.106*** (0.027)	−0.096*** (0.027)
Metro	−0.129*** (0.045)	−0.132*** (0.045)
Adjacent regions	0.020*** (0.006)	0.020*** (0.006)
Constant	−1.727*** (0.488)	−2.104*** (0.491)
Period	Included	Included
Observations	516	516
LR χ^2	702.79***	717.06***
Pseudo R^2	0.1318	0.1345
Log likelihood	−2315.1264	−2307.9936

Standard errors are shown in parentheses below the coefficients. * p < 0.1, ** p < 0.05, *** p < 0.01

can foster interdisciplinarity, it may also reduce the intensity of specialization required to advance knowledge in highly complex and emerging scientific areas. As expected, the positive and significant coefficient for *Quantum funding* underscores the importance of strong policy initiatives and financial support in fostering advancements in quantum knowledge. Interestingly, the *Metro* variable exhibits a statistically significant negative coefficient, suggesting that quantum knowledge is more prevalent in non-metropolitan regions than metropolitan ones. However, it is worth noting that out of the 100 NUTS3 regions, only 15 are non-metropolitan, which include regions that have benefited from specialized research institutes, such as Cambridge (UKH12), Oxford (UKJ14), and Leuven (BE242).

The second column introduces our key variable, *Semantic proximity*, which shows positive and statistically significant coefficients. This indicates that the alignment between a region's quantum-specific knowledge and its overall scientific knowledge plays an important role in the development of quantum knowledge. In other words, regions with a closer alignment between their general scientific expertise and quantum-specific knowledge are more likely to have more substantial potential for advancing quantum-related innovations.

To ensure that our results are not affected by heteroskedasticity, we present the results with robust standard errors in Table 6. As shown, our key variable, *Semantic proximity*,

Table 6 Regression results (robust standard errors)

Dependent variable: Regression model:	Quantum knowledge Negative binomial regression	
	(1)	(2)
Semantic proximity		1.930*** (0.534)
Publications (ln)	1.063*** (0.040)	1.045*** (0.041)
Subjects	−0.008*** (0.001)	−0.008*** (0.001)
Quantum funding	40.218*** (3.068)	40.893*** (2.959)
GDP (ln)	−0.064** (0.031)	−0.052* (0.031)
Population (ln)	−0.106*** (0.024)	−0.096*** (0.024)
Metro	−0.129** (0.051)	−0.132*** (0.050)
Adjacent regions	0.020*** (0.005)	0.020*** (0.005)
Constant	−1.727*** (0.445)	−2.104*** (0.457)
Period	Included	Included
Observations	516	516
LR χ^2	1447.34***	1450.85***
Pseudo R^2	0.1318	0.1345
Log likelihood	−2315.1264	−2307.9936

Robust standard errors are shown in parentheses below the coefficients. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

continues to show consistent results in Model 2, further reinforcing its important role in the creation of quantum knowledge.

Lastly, we further tested the robustness of our main findings by incorporating lagged independent variables by one period, which helps mitigate endogeneity concerns, establish temporal causality, and account for potential delayed effects in quantum knowledge production (Table 7). In Model 2, *Semantic proximity* remains consistently positive even when using lagged values, reinforcing the stability of our findings over time. In Models 3 and 4, we use generalized least squares (GLS) models, where the dependent variable is the ratio of quantum publications—the share of quantum-related papers relative to total scientific publications. This approach helps to account for variations in the overall volume of scientific output. The robustness check confirms that our main results remain consistent, showing that the positive relationship between *Semantic proximity* and quantum knowledge production holds even when an alternative modeling approach is applied, and this in turn further strengthens the reliability of the presented findings.

Table 7 Regression results (lagged)

Dependent variable: Regression model:	Quantum knowledge Negative binomial regression		Quantum knowledge share Panel generalized least squares regression	
	(1)	(2)	(3)	(4)
Semantic proximity		2.700*** (0.606)		0.011*** (0.003)
Publications (ln)	1.016*** (0.046)	0.993*** (0.046)	0.000 (0.000)	0.000 (0.000)
Subjects	−1.766*** (0.288)	−1.581*** (0.285)	−0.018*** (0.003)	−0.017*** (0.003)
Quantum funding	35.297*** (2.707)	36.220*** (2.638)	0.233*** (0.024)	0.248*** (0.024)
GDP (ln)	−0.065 (0.041)	−0.051 (0.040)	−0.000 (0.000)	−0.000 (0.000)
Population (ln)	−0.095*** (0.031)	−0.081*** (0.030)	−0.001*** (0.000)	−0.001*** (0.000)
Metro	−0.157*** (0.052)	−0.156*** (0.051)	−0.002*** (0.000)	−0.002*** (0.000)
Adjacent regions	0.016** (0.007)	0.017** (0.007)	0.000*** (0.000)	0.000*** (0.000)
Constant	6.465*** (1.444)	5.176*** (1.444)	0.118*** (0.014)	0.109*** (0.014)
Period	Included	Included	Included	Included
Observations	430	430	430 (86)	430 (86)
LR χ^2	516.80***	536.44***		
Pseudo R^2	0.1157	0.1201		
Log likelihood	−1975.485	−1965.6678		
Wald χ^2			459.91***	494.89***
BP χ^2 for heteroscedasticity			131.10***	134.88***
Wooldridge test for autocorrelation			7.454***	8.562***

All independent variables are lagged by one period. Standard errors are shown in parentheses below the coefficients. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

Discussions & concluding remarks

This study advances the measurement of regional knowledge bases by introducing the concept of *semantic proximity*, which relies on the textual analysis of scientific publications rather than static classification systems that struggle to capture novel and interdisciplinary fields. By applying advanced NLP techniques, specifically BERT-based embeddings combined with BERTopic, we identify latent cognitive alignments that remain hidden in conventional taxonomies, thereby offering a richer and more adaptive framework for analyzing regional knowledge spaces.

The key analytical contribution is the demonstration that semantic proximity, the alignment between a region's general scientific activity and an emerging field such as quantum science, has a positive and statistically significant effect on the creation of new quantum knowledge. This relationship is robust across multiple specifications, even after accounting for research output, funding, economic development, and spatial context. Importantly, although overall levels of semantic proximity between quantum science and general scientific activity were modest, regions with even partial alignment produced substantially more quantum-related publications. This suggests that complete overlap is not essential; rather, even partial semantic alignment can serve as a meaningful indicator of a region's potential to engage in emerging scientific domains. In this sense, semantic proximity reveals latent and sometimes tacit capabilities that traditional measures, reliant on codified knowledge or citation-based spillovers, fail to capture. Such insights are particularly valuable for identifying regions that, while not yet specialized, possess the foundational knowledge needed to diversify into frontier domains.

Building on these empirical findings, the study contributes to theoretical debates in economic geography and innovation studies by extending the concept of relatedness. While traditional approaches emphasize explicit overlaps in classification systems, semantic proximity highlights subtle affinities that can facilitate breakthrough innovations. It shows that knowledge recombination does not require strong, direct overlap but can also emerge from underlying semantic connections that enable novel combinations of knowledge. This perspective enriches evolutionary economic geography by providing a more flexible lens for understanding how regional knowledge bases evolve and diversify. By focusing on scientific publications rather than patents, the study also underscores the pivotal role of academic knowledge production in shaping regional innovation trajectories, particularly in science-driven sectors where early-stage research is essential (Kogler et al., 2024). This emphasis enhances the academic discourse on the interplay between science and regional economic development, underscoring the value of integrating bibliometric and semantic data in innovation research.

Policymakers can harness semantic proximity as a powerful diagnostic and strategic planning tool to better align innovation policies with regional strengths and emerging opportunities. For instance, by measuring the semantic proximity between government policy documents and regional scientific or technological knowledge outputs (such as publications and patents), policymakers can assess how well current policies align with a region's existing capabilities and emerging specializations. This analysis helps to identify gaps where policies may need to be adjusted to better support nascent fields or regions showing potential for growth. Additionally, semantic proximity can be used to track the effectiveness of policy interventions over time by examining changes in the alignment between policy focus areas and regional innovation outputs. This enables dynamic and evidence-based policy adjustments that respond to evolving regional knowledge landscapes. Moreover, integrating semantic proximity into smart specialization strategies allows policymakers to move beyond traditional metrics of specialization (e.g., patent counts) and incorporate richer, context-driven insights into regional knowledge structures. This approach helps in selecting priority areas for investment by identifying regions with moderate to high semantic proximity to target fields, ensuring resources are directed where the likelihood of successful knowledge recombination and innovation is greatest. Policymakers can also apply semantic proximity analyses to foster cross-sectoral and interdisciplinary collaborations by identifying thematic overlaps between different industries, research institutions, and policy domains. This can support the creation of innovation ecosystems that bridge knowledge silos and accelerate the development of frontier technologies.

Funding and economic development agencies can derive substantial value from insights based on semantic proximity indicators. At a system-wide level, these measures can help identify interdisciplinary opportunity spaces within the scientific and technological knowledge landscape, informing the design of more targeted and effective funding instruments. Simultaneously, they can offer strategic guidance on the feasibility of sector-specific development strategies within particular national or regional contexts. Semantic proximity thus serves as a diagnostic benchmark, indicating whether regional path-creation policies are likely to succeed or fail, depending on the alignment (or misalignment) between existing scientific or technological capabilities and a prospective domain of interest. Importantly, such indicators are applicable across all regional innovation systems but enable differentiated policy design based on regional capacity. For instance, regions exhibiting high proximity to a broad array of scientific domains may pursue science-technology optimization strategies, leveraging their strong existing capabilities. In contrast, regions with uniformly low proximity values may need to concentrate limited resources on narrowly defined areas that are most aligned with their current knowledge base, thereby maximizing the potential for effective knowledge recombination and innovation.

While the proposed method addresses several limitations of traditional classification-based approaches, it also presents certain challenges. First, relying on textual data may introduce biases related to language use and publication practices, which can lead to the overrepresentation of some scientific fields or regions. This imbalance may skew the analysis, particularly disadvantaging regions with lower scientific output or non-English publications. Second, scientific journal articles capture only a portion of regional knowledge production and may miss other critical forms of knowledge production, such as unpublished research, or private-sector innovation activities. Future work could address these gaps by integrating multiple textual sources into a unified semantic framework, providing a more comprehensive view of regional knowledge dynamics.

Despite these challenges, semantic proximity holds significant potential for advancing both research and policy. It can be applied to diverse textual corpora such as patent documents, government policy reports, and industry papers, offering a more comprehensive understanding of knowledge production, policy alignment, and knowledge diffusion. For example, future research could employ this approach to explore the science-technology interface in greater depth, identifying emerging technology clusters grounded in regional scientific strengths. By capturing subtle and often hidden connections between knowledge domains, semantic proximity equips researchers and policymakers with a dynamic and forward-looking tool to guide regional innovation systems in an era of rapid technological change and growing interdisciplinarity.

Appendix

See Table8

Table 8 List of NUTS3 regions included in the analysis

Rank	NUTS3	# of quantum-related journal publications	Region name	Rank	NUTS3	# of quantum-related journal publications	Region name
1	FR101	1986	Paris	51	DEA2D	349	Städteregion Aachen
2	FR104	1412	Essonne	52	FR716	342	Rhône
3	CH040	1256	Zürich	53	PL514	333	Miasto Wrocław
4	DE300	1232	Berlin	54	ITF33	325	Napoli
5	UKJ14	1225	Oxfordshire	55	DK042	324	Østjylland
6	ES511	1133	Barcelona	56	DEB35	323	Mainz, Kreisfreie Stadt
7	ES300	1130	Madrid	57	CZ071	306	Olomoucký kraj
8	UKH12	1091	Cambridgeshire	58	DEG03	302	Jena, Kreisfreie Stadt
9	FR714	1038	Isère	59	UKJ32	295	Southampton
10	DED21	1007	Dresden, Kreisfreie Stadt	60	FI1B1	290	Helsinki-Uusimaa
11	ITI43	908	Roma	61	DE911	287	Braunschweig, Kreisfreie Stadt
12	DE21H	878	München, Landkreis	62	ES212	287	Gipuzkoa
13	AT130	855	Wien	63	IE021	287	Dublin
14	PL127	855	Miasto Warszawa	64	PT170	286	Lisbon Metropolitan Area
15	UKI32	737	Westminster	65	UKE32	285	Sheffield
16	ITC4C	704	Milano	66	ITH36	285	Padova
17	NL333	679	Delft and Westland	67	BE100	285	Arr. de Bruxelles-Capitale/Arr. Brussel-Hoofdstad
18	DE212	664	München, Kreisfreie Stadt	68	DE131	284	Freiburg im Breisgau, Stadtkreis
19	AT332	641	Innsbruck	69	UKD33	274	Manchester
20	DE111	634	Stuttgart, Stadtkreis	70	PL415	270	Miasto Poznań
21	ITI17	599	Pisa	71	UKE42	267	Leeds
22	CH013	588	Geneva	72	PL633	267	Trójmiejski
23	UKI31	574	Camden and City of London	73	BE242	256	Arr. Leuven
24	ITH44	563	Trieste	74	UKM22	251	Clackmannanshire and Fife
25	CH011	543	Vaud	75	NL414	248	Zuidoost-Noord-Brabant

Table 8 (continued)

Rank	NUTS3	# of quantum-related journal publications	Region name	Rank	NUTS3	# of quantum-related journal publications	Region name
26	DK011	539	Byen København	76	DEA51	247	Bochum, Kreisfreie Stadt
27	DE600	530	Hamburg	77	FR824	246	Bouches-du-Rhône
28	SE110	520	Stockholms län	78	ITC33	245	Genova
29	ITI14	507	Firenze	79	DE711	244	Darmstadt, Kreisfreie Stadt
30	HU101	504	Budapest	80	SE224	241	Skåne län
31	DE144	501	Ulm, Stadtkreis	81	ES523	238	Valencia/València
32	DE252	491	Erlangen, Kreisfreie Stadt	82	NL337	238	Leiden and Bollenstreek
33	DE125	462	Heidelberg, Stadtkreis	83	FR421	236	Bas-Rhin
34	CZ010	460	Hlavní město Praha	84	DED51	235	Leipzig, Kreisfreie Stadt
35	UKK11	450	Bristol	85	CH033	234	Aargau
36	ES213	440	Bizkaia	86	DE232	233	Regensburg, Kreisfreie Stadt
37	UKM34	431	Glasgow	87	DED2C	226	Brautzen
38	NL326	429	Groot-Amsterdam	88	LT00A	222	Vilnius
39	UKF14	429	Nottingham	89	ITH55	221	Bologna
40	DE263	424	Würzburg, Kreisfreie Stadt	90	ITH20	221	Trento
41	PL213	419	Miasto Kraków	91	DEC01	216	Regionalverband Saarbrücken
42	DE929	412	Region Hannover	92	DEA23	213	Köln, Kreisfreie Stadt
43	CH031	399	Basel-Stadt	93	UKJ25	208	West Surrey
44	UKM25	382	Edinburgh	94	SE121	206	Uppsala län
45	ITC11	379	Torino	95	UKE21	204	York
46	SE232	378	Västra Götalands län	96	DE712	202	Frankfurt am Main, Kreisfreie Stadt
47	FR623	374	Haute-Garonne	97	BE211	200	Arr. Antwerpen
48	DK012	360	Københavns omegn	98	CZ064	197	Jihomoravský kraj
49	DEA26	356	Düren	99	FR301	195	Nord

Table 8 (continued)

Rank	NUTS3	# of quantum-related journal publications	Region name	Rank	NUTS3	# of quantum-related journal publications	Region name
50	DE122	353	Karlsruhe, Stadtkreis	100	NL310	195	Utrecht

Funding Yonsei University,2025-22-0125,Keungoui Kim

Declarations

Competing interests 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

Ahammad, M. F., Tarba, S. Y., Liu, Y., & Glaister, K. W. (2016). Knowledge transfer and cross-border acquisition performance: The impact of cultural distance and employee retention. *International Business Review, 25*(1, Part A), 66–75. https://doi.org/10.1016/j.ibusrev.2014.06.015

Almeida, P. (1996). Knowledge sourcing by foreign multinationals: Patent citation analysis in the U.S. semiconductor industry. *Strategic Management Journal, 17*(S2), 155–165. https://doi.org/10.1002/smj.4250171113

Barbero, J., Diukanova, O., Gianelle, C., Salotti, S., & Santoalha, A. (2024). Technologically related diversification: One size does not fit all European regions. *Research Policy, 53*(3), Article 104973. https://doi.org/10.1016/j.respol.2024.104973

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

Boschma, R., Balland, P.-A., & Kogler, D. F. (2015). Relatedness and technological change in cities: The rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change, 24*(1), 223–250. https://doi.org/10.1093/icc/dtu012

Boschma, R., Heimeriks, G., & Balland, P.-A. (2014). Scientific knowledge dynamics and relatedness in biotech cities. *Research Policy, 43*(1), 107–114. https://doi.org/10.1016/j.respol.2013.07.009

Broekel, T. (2012). Collaboration intensity and regional innovation efficiency in Germany—A conditional efficiency approach. *Industry and Innovation, 19*(2), 155–179. https://doi.org/10.1080/13662716.2012.650884

Broekel, T., Balland, P.-A., Burger, M., & van Oort, F. (2014). Modeling knowledge networks in economic geography: A discussion of four methods. *The Annals of Regional Science, 53*(2), 423–452. https://doi.org/10.1007/s00168-014-0616-2

Broekel, T., & Boschma, R. (2012). Knowledge networks in the Dutch aviation industry: The proximity paradox. *Journal of Economic Geography, 12*(2), 409–433. https://doi.org/10.1093/jeg/lbr010

Cooke, P. (1998). Regional Innovation Systems Book - Introduction: Origins of the Concept. *"Regional Innovation Systems" Book*.

Cooke, P., De Laurentis, C., Tödtling, F., & Trippl, M. (2007). Regional Knowledge Economies: Markets, clusters and innovation. *Edward Elgar Publishing*. https://doi.org/10.4337/9781847206930

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Association for Computational Linguistics*. https://doi.org/10.18653/v1/N19-1423

Eisenhardt, K. M., & Martin, J. A. (2000). Dynamic capabilities: What are they? *Strategic Management Journal, 21*(10–11), 1105–1121. https://doi.org/10.1002/1097-0266(200010/11)21:10/11%3c1105::AID-SMJ133%3e3.0.CO;2-E

Engelsman, E. C., & van Raan, A. F. J. (1994). A patent-based cartography of technology. *Research Policy, 23*(1), 1–26. https://doi.org/10.1016/0048-7333(94)90024-8

Fagerberg, J., Landström, H., & Martin, B. R. (2012). Exploring the emerging knowledge base of 'the knowledge society.' *Research Policy, 41*(7), 1121–1131. https://doi.org/10.1016/j.respol.2012.03.007

Feldman, M. P., Choi, J (2015) Harnessing the Geography of Innovation. The Handbook of Global Science, Technology, and Innovation, John Wiley & Sons Ltd.,

Feldman, M. P. (1994). *The geography of innovation* (Vol. Vol. 2). Springer. https://doi.org/10.1007/978-94-017-3333-5

Feldman, M. P., & Kogler, D. F. (2010). Chapter 8—Stylized Facts in the Geography of Innovation. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the Economics of Innovation.* North-Holland.

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. https://doi.org/10.1111/j.1538-4632.2006.00687.x

Frenken, K., Van Oort, F., & Verburg, T. (2007). Related variety, unrelated variety and regional economic growth. *Regional Studies, 41*(5), 685–697. https://doi.org/10.1080/00343400601120296

Gertler, M. S. (2003). Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of Economic Geography, 3*(1), 75–99. https://doi.org/10.1093/jeg/3.1.75

Grabher, G. (2009). Yet another turn? The evolutionary project in economic geography. *Economic Geography, 85*(2), 119–127. https://doi.org/10.1111/j.1944-8287.2009.01016.x

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Heimeriks, G., & Boschma, R. (2014). The path- and place-dependent nature of scientific knowledge production in biotech 1986–2008. *Journal of Economic Geography, 14*(2), 339–364. https://doi.org/10.1093/jeg/lbs052

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S., & Zhu, S. (2018). The Principle of Relatedness. In A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, & Y. Bar-Yam (Eds.), *Unifying Themes in Complex Systems IX* (pp. 451–457). Springer International Publishing. https://doi.org/10.1007/978-3-319-96661-8_46

Hoekman, J., Frenken, K., & van Oort, F. (2009). The geography of collaborative knowledge production in Europe. *The Annals of Regional Science, 43*(3), 721–738. https://doi.org/10.1007/s00168-008-0252-9

Huber, F. (2012). On the role and interrelationship of spatial, social and cognitive proximity: Personal knowledge relationships of R&D workers in the Cambridge information technology cluster. *Regional Studies, 46*(9), 1169–1182. https://doi.org/10.1080/00343404.2011.569539

Jaffe, A. B. (1986). *Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value* (Working Paper No. 1815). National Bureau of Economic Research. https://doi.org/10.3386/w1815

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations*. *The Quarterly Journal of Economics, 108*(3), 577–598. https://doi.org/10.2307/2118401

Jung, S., Kim, K., & Lee, C. (2021). The nature of ICT in technology convergence: A knowledge-based network analysis. *PLoS ONE, 16*(7), Article e0254424. https://doi.org/10.1371/journal.pone.0254424

Kedron, P., F. Kogler, D., & Rocchetta, S. (2020). Mind the gap: Advancing evolutionary approaches to regional development with progressive empirical strategies. *Geography Compass*, *14*(9), e12501. https://doi.org/10.1111/gec3.12501

Kim, K., Ferrante, C., & Kogler, D. F. (2024a). Smart specialisation strategies and regional knowledge spaces: How to bridge vision and reality. *Regional Studies*. https://doi.org/10.1080/00343404.2024.2355985

Kim, K., Jung, S., & Hwang, J. (2019). Technology convergence capability and firm innovation in the manufacturing sector: An approach based on patent network analysis. *R & D ManagemeNt, 49*(4), 595–606. https://doi.org/10.1111/radm.12350

Kim, K., Kogler, D. F., & Maliphol, S. (2024b). Identifying interdisciplinary emergence in the science of science: Combination of network analysis and BERTopic. *Humanities and Social Sciences Communications, 11*(1), 603. https://doi.org/10.1057/s41599-024-03044-y

Kim, K., Kogler, D. F., & Zabetta, M. C. (2024c). New collaborations and novel innovations: The role of regional brokerage and collaboration intensity. *European Planning Studies, 32*(6), 1251–1272. https://doi.org/10.1080/09654313.2023.2291087

Kogler, D. F. (2015). Editorial: Evolutionary economic geography – Theoretical and empirical progress. *Regional Studies, 49*(5), 705–711. https://doi.org/10.1080/00343404.2015.1033178

Kogler, D. F., Brenner, T., Celebioglu, F., & Shin, H. (2024). The science-innovation nexus in a regional context—Introduction to the special issue, policy and future research directions. *Review of Regional Research, 44*(2), 141–149. https://doi.org/10.1007/s10037-024-00212-0

Kogler, D. F., Davies, R. B., Lee, C., & Kim, K. (2023a). Regional knowledge spaces: The interplay of entry-relatedness and entry-potential for technological change and growth. *The Journal of Technology Transfer, 48*(2), 645–668. https://doi.org/10.1007/s10961-022-09924-2

Kogler, D. F., Essletzbichler, J., & Rigby, D. L. (2017). The evolution of specialization in the EU15 knowledge space. *Journal of Economic Geography, 17*(2), 345–373. https://doi.org/10.1093/jeg/lbw024

Kogler, D. F., Evenhuis, E., Giuliani, E., Martin, R., Uyarra, E., & Boschma, R. (2023b). Re-imagining evolutionary economic geography. *Cambridge Journal of Regions, Economy and Society, 16*(3), 373–390. https://doi.org/10.1093/cjres/rsad029

Kogler, D. F., Rigby, D. L., & Tucker, I. (2013). Mapping knowledge space and technological relatedness in US cities. *European Planning Studies, 21*(9), 1374–1391. https://doi.org/10.1080/09654313.2012.755832

Kopcsa, A., & Schiebel, E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science, 49*(1), 7–17. https://doi.org/10.1002/(SICI)1097-4571(1998)49:1%3c7::AID-ASI3%3e3.0.CO;2-W

Lee, W. S., Han, E. J., & Sohn, S. Y. (2015). Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. *Technological Forecasting and Social Change, 100*, 317–329. https://doi.org/10.1016/j.techfore.2015.07.022

Makri, M., Hitt, M. A., & Lane, P. J. (2010). Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal, 31*(6), 602–628. https://doi.org/10.1002/smj.829

Martin, R., & Sunley, P. (2006). Path dependence and regional economic evolution. *Journal of Economic Geography, 6*(4), 395–437. https://doi.org/10.1093/jeg/lbl012

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics, 9*(4), 962–973. https://doi.org/10.1016/j.joi.2015.10.005

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Morgan, K. (1997). The learning region: Institutions, innovation and regional renewal. *Regional Studies, 31*(5), 491–503. https://doi.org/10.1080/00343409750132289

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. https://doi.org/10.1111/j.1944-8287.2011.01121.x

Nelson, R. R. (1993). *National Innovation Systems: A Comparative Analysis* (SSRN Scholarly Paper No. 1496195). Social Science Research Network. https://papers.ssrn.com/abstract=1496195

Nelson, R. R., & Winter, S. G. (1985). *An Evolutionary Theory of Economic Change*. Harvard University Press.

Nooteboom, B. (2000). Learning by interaction: Absorptive capacity, cognitive distance and governance. *Journal of Management & Governance, 4*(1), 69–92.

Quatraro, F. (2010). Knowledge coherence, variety, and economic growth: Manufacturing evidence from Italian regions. *Research Policy, 39*(10), 1289–1302. https://doi.org/10.1016/j.respol.2010.09.005

Rigby, D. L. (2015). Technological relatedness and knowledge space: Entry and exit of US cities from patent classes. *Regional Studies, 49*(11), 1922–1937. https://doi.org/10.1080/00343404.2013.854878

Rigby, D. L., Roesler, C., Kogler, D., Boschma, R., & Balland, P.-A. (2022). Do EU regions benefit from smart specialisation principles? *Regional Studies, 56*(12), 2058–2073. https://doi.org/10.1080/00343404.2022.2032628

Santoalha, A. (2019). New indicators of related diversification applied to smart specialization in European regions. *Spatial Economic Analysis, 14*(4), 404–424. https://doi.org/10.1080/17421772.2019.1584328

Schumpeter, J. A. (1942). *Capitalism, Socialism, and Democracy* (SSRN Scholarly Paper No. 1496200). Social Science Research Network. https://papers.ssrn.com/abstract=1496200

Schumpeter, J. A. (2021). The Theory of Economic Development. *Routledge*. https://doi.org/10.4324/9781003146766

Shin, H., Kim, K., Lee, J., & Kogler, D. F. (2023). Inventors, firms and localities: Insights into the nexus that forms and alters the evolution of regional knowledge spaces. *Cambridge Journal of Regions, Economy and Society*. https://doi.org/10.1093/cjres/rsad019

Sonn, J. W., & Storper, M. (2008). The increasing importance of geographical proximity in knowledge production: An analysis of US patent citations, 1975–1997. *Environment and Planning a: Economy and Space, 40*(5), 1020–1039. https://doi.org/10.1068/a3930

Teece, D. J., Rumelt, R., Dosi, G., & Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization, 23*(1), 1–30. https://doi.org/10.1016/0167-2681(94)90094-9

Ter Wal, A. L. J., & Boschma, R. A. (2009). Applying social network analysis in economic geography: Framing some key analytic issues. *The Annals of Regional Science, 43*(3), 739–756. https://doi.org/10.1007/s00168-008-0258-3

Tóth, G., Elekes, Z., Whittle, A., Lee, C., & Kogler, D. F. (2022). Technology network structure conditions the economic resilience of regions. *Economic Geography, 98*(4), 355–378. https://doi.org/10.1080/00130095.2022.2035715

Van Raan, A. F. J. (1997). Scientometrics: State-of-the-art. *Scientometrics, 38*(1), 205–218. https://doi.org/10.1007/bf02461131 von Hippel, E. (1994). "Sticky information" and the locus of problem solving: Implications for innovation. *Management Science, 40*(4), 429–439. https://doi.org/10.1287/mnsc.40.4.429

Whittle, A., & Kogler, D. F. (2020). Related to what? Reviewing the literature on technological relatedness: Where we are now and where can we go? *Papers in Regional Science, 99*(1), 97–114. https://doi.org/10.1111/pirs.12481