WORKING PAPER
2025:75

Identifying Catalyst Technologies in Clusters with Unsupervised Machine Learning

– an application on patent clusters in the UK

Zehra Usta, Martin Andersson, Katarzyna Kopczewska and Maria Kubara

RESEARCH NETWORK DEBATE
SWEDISH ENTREPRENEURSHIP FORUM

Working Papers Series from Swedish Entrepreneurship Forum

In 2009 Swedish Entrepreneurship Forum started publishing a new series of Working Papers. These are available for download on www.entreprenorskapsforum.se, and are part of our ambition to make quality research available to a wider audience, not only within the academic world.

Scholars from different disciplines are invited to publish academic work with the common denominator that the work has policy relevance within the field of entrepreneurship, innovation and SMEs.

The working papers published in this series have all been discussed at academic seminars at the research institution of the author.

ABOUT SWEDISH ENTREPRENEURSHIP FORUM

Swedish Entrepreneurship Forum is the leading Swedish network organization for generating and transferring policy relevant research in the field of entrepreneurship and small enterprise development.

Swedish Entrepreneurship Forum is a network organization with the aim

- to serve as a bridge between the small business research community and all agents active in development of new and small enterprises.
- to initiate and disseminate research relevant to policy in the fields of entrepreneurship, innovation and SME.
- to offer entrepreneurship researchers a forum for idea sharing, to build national and international networks in the field and to bridge the gap between research and practical application.

Find out more on www.entreprenorskapsforum.se

Identifying Catalyst Technologies in Clusters with Unsupervised Machine Learning

an application on patent clusters in the UK

Zehra Usta[1], Martin Andersson[2], Katarzyna Kopczewska[3] and Maria Kubara[4]

ABSTRACT

A common proposition is that certain technologies play a catalytic role in regions by paving the way for the emergence of new related technologies, contributing to the development and diversification of technology clusters. This paper employs unsupervised machine learning algorithms with temporally informed association rule mining to identify catalytic patents in clusters in the UK. Using data spanning over 30 years (1980-2015) we show clear asymmetric relationships between patents. Some act as evident catalysts that drive future patent activity in clusters. The results point to a strong empirical relevance of asymmetric relatedness between patents in the development of clusters of technology. They also highlight the usefulness of machine learning algorithms to better understand the long-term evolution of clusters and show how temporally informed association rule mining can be used to analyses asymmetries in relatedness and to identify catalyst technologies.

Highlights

- Technologies are asymmetrically related, playing the causal roles of antecedents and consequents
- High-tech patents influence low-tech more than the opposite
- The dynamic association rules revealed asymmetric influences across technologies, highlighting key catalyst technologies
- Patents in telecommunications frequently affect each other and are important drivers of technological clusters

Key words: clusters, innovation, cluster dynamics, technological relatedness, asymmetric relatedness, innovation catalysts, patents

JEL Code: O31, O33, R12

[1] Faculty of Economic Sciences, University of Warsaw: z.usta@student.uw.edu.pl
[2] Department of Industrial Economics, Blekinge Institute of Technology (BTH) and the Swedish Entrepreneurship Forum, Stockholm: martin.andersson@bth.se // martin.andersson@entreprenorskapsforum.se
[3] Faculty of Economic Sciences, University of Warsaw: kkopczewska@wne.uw.edu.pl
[4] Faculty of Economic Sciences, University of Warsaw: maria.kubara@uw.edu.pl

1. INTRODUCTION

A central proposition in analyses of the origins and growth of industry and technology clusters is that there are often certain firms, organizations, technologies or other type of investments that act as catalysts for subsequent development. Klepper's work on the evolution of industries and industry clusters emphasizes that large resourceful firms act as catalysts by pushing local industry growth through spinoffs (Klepper 2015, 2010). Agarwal's and Cockburn's (2003) work on the "anchor-tenant hypothesis" emphasizes the local presence of large R&D-intensive firms (i.e. anchor firms) as a driving force for different types of spillovers to the local economy. Giuliani and Bell (2005) show that a few firms (often those with superior R&D or external linkages) tend to act as knowledge sources, while many others are knowledge receivers, creating a directed pattern of interaction. A similar type of phenomenon is found in analyses of technologies. For example, analyses of patent citation data routinely find that certain technology classes draw extensively on knowledge from another class, while the latter class does not rely on the former to the same extent (see e.g. Jaffe and Trajtenberg 2002). That is, some technology domains or patents are catalysts for the emergence of other technologies in the sense that they are sources of knowledge (i.e. being cited and built upon) for other technology domains or patents.

This type of dynamics is of great interest for economic geography, regional science as well as urban economics, as it is related to the processes through which regions, cities and clusters diversify and develop their industry and technology base. In the last decades, there has been a considerable growth of studies on how regions diversify in new industries, technologies and knowledge (Boschma 2017). Diversification has also been a core goal of many regional development policies, not least the EUs smart specialization strategy (Foray 2014, McCann and Ortega-Argiles 2014, Balland et al 2019). Several studies document that diversification processes in regions build on relatedness in the sense that new activities are related to existing ones and build on pre-existing industries, technologies and knowledge bases available in the region (Neffke et a 2011, Xiao et al 2018, Rigby 2015). Most analyses still treat relatedness as symmetric[5] although many relatedness patterns are likely to be asymmetric in reality – i.e. a product, technology or industry *x* may be related to *y* even though *y* is not necessarily related to *x* (Boschma 2017).

Asymmetric relatedness is akin to the idea of catalyst activities in the sense that it implies that the local presence of *y* may give rise to *x* although *x* may not give rise to *y*. From this perspective, *y* is a catalyst for *x*. Boschma (2017) argues that there is a need for more analyses of asymmetries in relatedness and that such analyses are likely to bring new insights of processes of regional diversification. Better

[5]That is, an activity (product/technology/industry) *x* is assumed to related to activity *y* as much as *y* is related to *x*.

knowledge of the processes of regional diversification may for instance facilitate the design of different types of policy schemes to support renewal of regional economies. Although the role of catalyst firms, technologies or products is well-established at a conceptual level, the empirical evidence-base of their role in driving development of clusters is still scant. Most analyses so far build on case-studies and there are few established methodologies to identify potential catalyst in the form of anchor firms, catalyst patents or other types of investments/activities (Nilsson et al 2025).

This paper contributes with an empirical analysis of catalyst technologies in technology clusters based on identification of asymmetric relatedness. We make a twofold contribution to the literature on relatedness, catalyst technologies and dynamics of clusters: First, we introduce a novel method to identify catalyst technologies that is generic in the sense that it can be applied to any type of data that has a temporal structure. The main idea is to model asymmetric relatedness by using temporal association rules in an unsupervised machine learning algorithm. The algorithm is designed to identify a form of temporally ordered co-occurrences of the type "if-then", where catalyst technologies are technologies that pre-dominantly are sources (if) to new offsprings (then) rather than offspring themselves. Second, using data on patent clusters in the UK over a 30-year period, we illustrate (i) the usefulness of temporal association rules in analyses of asymmetric relatedness and (ii) the empirical relevance of catalyst technologies in driving patent activity in technology clusters.

An additional merit of the analysis is that we identify patent clusters and analyse their evolution at a precise geographical level. Most research uses a pre-determined geographical delineation of regions, such as NUTS3, and it is well-documented that results may depend on the size and location of the regions (Arbia and Petrarca 2011, Wong 2004, Andersson et al 2016). We use geo-referenced point data that allow us to identify and analyse clusters free of pre-defined delineations.

Overall, our analyses add to our understanding of how technologies in a region evolve and how their development and spatial spread are influenced by relatedness and asymmetric relationships. We document that there are evident 'catalyst patents' that drive patent activity in related technologies in clusters. Asymmetric relatedness appears to be an important characteristic of the process by which new technologies emerge, as evidenced by the development of patents over time. Our analyses also highlight the usefulness of machine learning algorithms to better understand the long-term evolution of clusters and show how temporally informed association rule mining can be used to analyse asymmetric relatedness and to identify catalyst technologies. On a methodological note, we maintain that the use of temporal association rules in unsupervised machine learning algorithms can be extended to several different areas, such as firm-level innovation studies.

The remainder of the paper is organized as follows: section 2 provides a brief overview of technological relatedness, asymmetries and catalytic patents. Section 3 presents our methodology and data. The main emphasis is on explaining the logic of using temporal association rules in an unsupervised machine learning algorithm. Results are presented in Section 4 and Section concludes.

2. TECHNOLOGY, ASYMMETRIC RELATEDNESS AND CATALYST TECHNOLOGIES

2.1 Technological relatedness

Technological relatedness refers to the degree of similarity between different technologies in terms of their underlying knowledge base (Breschi et al. 2003). Rigby (2013) was one of the first researchers to investigate regional technological diversification through patent-relatedness, showing that technologies related to those already present in a US metropolitan area were more likely to emerge within that region. Rigby's findings proved the importance of relatedness in patents, a concept that has been explored by subsequent research. Boschma et al. (2023) showed that relatedness plays a significant role in the emergence of breakthrough inventions within regions, with the probability of breakthroughs increasing when a region's technology is related to its existing local technological knowledge base. Mascarini et al. (2023) investigated the impact of related and unrelated technologies on innovation, focusing on regional differences. Their research evidenced that knowledge exchange between related industries significantly contributes to the promotion of new innovations.

Relatedness has also been shown to play an important role in the technological diversification of firms, as evidenced by analyses of patent records (Kim et al 2023). Examining patents can provide insights into how firms align their business activities with specific technologies strategically (Boschma et al. 2023). The concept of technological relatedness has been explored in several studies (e.g. Rigby 2013, Boschma 2009). Regions with diverse but related technological activities tend to benefit more from knowledge spillovers and experience higher economic growth. It is common to refer to this as a reflection of the empirical relevance of ‘related variety’ (Boschma 2009, Frenken et al 2007, Mewes and Broekel 2022, Wixe and Andersson 2017). There is thus plenty of evidence pointing to a relatedness of patents within a region playing significant role in promoting technological capabilities, regional innovation and growth (Hasan and Tucci 2010).

2.2 Importance of asymmetric technological relatedness and catalyst technologies

Traditional methods of measuring technological relatedness often assume that the relationship between technologies is symmetrical. However, this assumption ignores the directional nature of many technological relationships and relatedness between technologies can in many cases be assumed to be asymmetric (Boschma 2017). The existence of one technology in a region may lead to the emergence of another, but the reverse is not necessarily equally likely. For instance, hardware technology can

greatly influence software development, yet software development may not necessarily contribute to hardware technology in the same way. Recognising this asymmetry has significant implications for understanding regional diversification. It suggests that certain technologies can act as catalysts, facilitating the emergence of new technologies, while others play a more passive or dependent role. Identifying these directional relationships will help researchers and policymakers understand and support the evolution of regional innovation systems. As maintained by Boschma (2017), acknowledging asymmetries in analyses of technological relatedness research aligns it with broader themes in economic geography, such as unequal trade flows, brain drain and uneven spatial diffusion of knowledge, all of which emphasise the importance of asymmetric dynamics in shaping regional economic outcomes.

Recent empirical studies find that asymmetric relatedness indeed plays a role in regional development. For example, Kuusk and Martynovich (2021) analyse the evolution of inter-industry relatedness networks in Sweden over two decades. They observed that the formation and disappearance of relatedness linkages over time is a dynamic process, and that new linkages often emerge in a one-sided manner. Even when two industries become related, the intensity of their connection is seldom balanced. A key finding in their analysis is that "relatedness unfolds and disentangles asymmetrically" as industries evolve. When a new related industry pair forms, typically one of the industries is the principal driver or beneficiary of the connection. Another example is a recent study by McHale et al (2023). They study the effect of recruitment of a prominent scientist ("star") into a new region. A main concern is how relatedness between the star's expertise and the local scientists' expertise influences knowledge spillovers, as evidenced by the scientific output of local scientists' post recruitment. They find that directional relatedness from the perspective of the incumbents (i.re. how much the locals' knowledge overlaps with the star) matters: incumbent scientists that have cited the recruited star prior recruitment benefit more from the arriving star, increasing their productivity after four years by approximately 40%.

A key implication is that regional diversification opportunities are asymmetric – some industries provide a stronger platform for branching out than others. Likewise, in regional innovation networks, not all connections are mutual: who you connect from (e.g., a leading firm or an external star) versus to (a peripheral firm or local partners) makes a difference for outcomes. The concept of asymmetric relatedness adds a layer of realism to the analysis of economic evolution: it recognizes that relatedness is rarely an "even playing field" and that understanding asymmetries requires examining the direction of relationships.

As stated in the introduction, this type of reasoning has been implicit in origins and growth of industry and technology clusters as well as entrepreneurial eco-systems. A large literature shows that some large and resourceful firms act as 'catalysts' for spinoffs (Klepper 2015, 2010), as generators of various types

of spillovers that are essential for the development of local innovation systems (Agarwal and Cockburn 2003, Feldman 2003) and processes related to industrial path creation (Crescenzi et al 2022, Dimos et al 2021). Mason and Brown (2014) as well as Herzog et al (2024) argue that the development of an entrepreneurial ecosystem in a region depends critically on the local presence of large established business.[6] Likewise, some patents are catalyst patents in the sense that they breed patenting in other areas (see e.g. Jaffe and Trajtenberg 2002, Bresnahan 2010).

These are all examples of the 'catalyst phenomena' in the sense that they all refer to instances by which an activity, firm, organization or technology is catalytic and breeds the development of other parts. They also relate to asymmetric relatedness in that 'catalysts' breed new technologies, new firms or new industries, but are not necessarily influenced by those that they breed.

Despite a growing number of studies in recent years, the evidence-base of their role in driving development and diversification of clusters is still scant. Most analyses so far build on case-studies and there are few established methodologies to identify potential catalyst in the form of anchor firms, catalyst patents or other types of investments/activities (Nilsson et al 2025). In our analysis, we develop a new methodology to identify catalyst patents by applying principles of asymmetric relatedness in an unsupervised machine learning algorithms with temporally informed association rule mining.

3. DATA AND METHODOLOGY

3.1 Data

Our analyses of the diversification and development of technology clusters are based on UK patent data that include detailed inventor addresses, technological classifications from the UK Intellectual Property Office (UKIPO), as well as citation and inventor data from the European Patent Office's (EPO) PATSTAT database. This database was published alongside a paper by MacDonald and Salmanovic (2023). The main dataset comprises citation information for 1,517,797 patents over a period of 36 years, from 1980 to 2015.

As our analysis of technological relatedness builds on patent citations, we grouped the data by cited patents and obtained 26,976 original patents. For each cited patent, we use information on the technology it represents. There are eight main categories and 313 subcategories according to the International Patent Classification (IPC). Additionally, we use the EUROSTAT (2024) patent

[6]Mason and Brown (2014) states: "At the heart of an entrepreneurial ecosystem typically there is at least one, and usually several, 'large established businesses', with significant management functions (e.g. head office or divisional/ subsidiary office) as well as undertaking R&D and production activities. These businesses will also be rich in technology. They play significant roles in developing the ecosystem."

classification system to divide the patents into a high-technology class (9.3%, comprising 2,513 observations; we merged the high-tech class, comprising 2,137 observations, and the biotechnology class, comprising 376 observations) and a low-technology class (90.7%, comprising 24,463 observations).

We divide the patents into seven groups of five-year periods: 1980–1984 (t_1), 1985–1989 (t_2), 1990–1994 (t_3), 1995–1999 (t_4), 2000–2004 (t_5), 2005–2009 (t_6) and 2010–2015 (t_7). These labels are used in causal analysis, which assumes that causality requires simultaneous or prior occurrence. The number of cited patents decreases over time in the temporal distribution, as it takes time to be cited and older patents are more often cited than the newest ones.

For the spatial dimension, we grouped neighbouring patents into clusters. We assumed that patents located no more than 5 kilometres apart could be considered as neighbouring. This range is consistent with the findings of Buzard et al. (2020), who confirmed that knowledge spillovers and citations of other patents occur within a radius of up to 5 miles. Using the DBSCAN algorithm, we detected dense spatial clusters — groups comprising at least five patents within a 5 km radius and issued throughout the entire analysed period. This operationalises the importance of tacit knowledge flows and the direct interactions of innovators. Consequently, we obtained 344 spatially delineated clusters, which we then studied for temporal causal relationships. Details are presented in Section 3 and in the Appendix.

3.2 Identification of patent clusters and asymmetric relatedness

We develop a *Spatial Machine Learning* (SML) approach as an alternative to spatial econometrics and statistics. These algorithms can discover hidden patterns and relationships in spatial data in a way that is impossible for traditional methods (Kopczewska 2022). Although the potential of machine learning algorithms is increasingly recognised in many disciplines, they are still rarely applied in spatial, regional and urban analysis. This paper addresses this gap in the literature by demonstrating the value of these tools for this type of analysis. Unsupervised machine learning algorithms are designed to reveal hidden patterns and structures in data by finding clusters of similar features or relationships. This enables them to identify common features that were previously unknown, allowing observations to be categorised based on data rather than assumptions. In this study, we use two unsupervised learning algorithms: DBSCAN, which identifies neighbouring observations that can form spatial clusters, and association rules, which detect technologies that frequently co-occur in time sequences and can therefore be classified as drivers or followers. The value of this approach lies in our ability to address non-linearity and asymmetry in relatedness. These methods are described below.

Patent clusters in the UK using DBSCAN

DBSCAN (Ester et al. 1996) uses circles to inspect all geolocated points. If the radial neighbourhood of a point contains at least *minPts* points within a given radius (eps), the point is labelled as densely located. Densely located points that mutually belong to each other's neighbourhood form clusters of observations. Other points are classified as 'noise' (low-density points outside clusters), and the number of clusters is the outcome of the algorithm. In this study, we considered patents to be densely located if they were in proximity to at least five other patents within a 5 km radius (in any given period). This approach divided the dataset of 26,976 points into 344 spatial clusters, with a low noise ratio of 6.8% (1,817 observations), meaning that, on average, one cluster comprises 77 patents. The primary aim of clustering is to identify natural spatial groupings within the data, where the number of groups and their boundaries are unknown in advance. Clustering can minimise the modifiable areal unit problem (MAUP) when analysing point data because clusters are determined by the behaviour of points rather than predefined borders.

Asymmetric relatedness – identifying catalytic patents with unsupervised machine learning

To identify catalytic patents, we employ unsupervised machine learning algorithms with temporally informed association rule mining. Association rules, as proposed by Agrawal et al. (1993), are primarily used in the context of market baskets to identify items that frequently co-occur when shopping, i.e. items that are often purchased together by customers. In our empirical context, the aim is to detect which patent classes frequently co-occur within spatial clusters. Our study comprises 344 clusters of different technologies (patent classes), analogous to market baskets containing various products. We study the internal composition of these clusters.

Association rules are a type of unsupervised learning method. We apply this data mining technique in a spatial context to provide a new measure of relatedness. Association rules are expressed in an “if-then” format, with the antecedent and consequent components explicitly defined. Popular algorithms such as Apriori and Eclat optimise the identification of meaningful rules by pruning datasets, i.e. removing the least frequent elements and their combinations, in an iterative process. First, the algorithm identifies particularly popular elements (items), formally measured by support statistic, which must meet a minimum threshold (the minimum frequency assumed in the analysis). This cleaning procedure keeps only frequent elements. Second, the algorithm then progressively combines elements to form larger groups (itemsets) that also meet this threshold. It finds the most frequent co-occurrences, composed of frequent individual elements, based on the logic that only frequent elements can produce frequent combinations. The most crucial part of the entire process is identifying which elements (items) and combinations (itemsets) appear frequently (Shahin et al. 2021, Agrawal et al., 1993). Third, strong association rules are obtained from these frequent itemsets, which must satisfy minimum support and confidence thresholds, i.e. frequent co-occurrence (support) and frequent causal appearance

(confidence). Additionally, redundant and insignificant relationships are eliminated, effectively reducing the complexity and number of resulting rule combinations.

By fixing the right-hand side (RHS) *{then}*, researchers specify a particular outcome or label of interest, which is analogous to a dependent variable in econometric modelling. The general representation of an association rule is expressed as follows:

$$\{if \rightarrow then\} \tag{1}$$

what is equivalently denoted by:

$$\{LHS \rightarrow RHS\} \tag{2}$$

where RHS refers to the right-hand side (consequent) and LHS to the left-hand side (antecedent). We adopt this technique to the spatial analysis of relatedness using the following approach. Each spatial cluster of patents is treated as a 'basket', with the technologies associated with the patents in this cluster considered the 'items'. This enables us to detect directional relationships between technologies. For example, if technology A is present, then technology B is likely to follow. Unlike traditional measures of relatedness, this approach captures asymmetry since the rules identified by association rule mining algorithms are asymmetrical by nature. This means that the presence of one technology may lead to another, but not vice versa – having a rule 'A is associated with B' does not necessarily imply that 'B is associated with A'.

The selection of strong and relevant association rules is guided by specific metrics. The primary input parameters are minimum support (eq.3), which indicates the lowest acceptable frequency with which an item occurs within the dataset, and minimum confidence (eq.4), which represents the minimum acceptable frequency with which items co-occur. Output metrics, such as lift (eq.7), quantify the likelihood of two items co-occurring compared to occurring independently and facilitate further evaluation of the rules. These association rule metrics are mathematically defined as follows:

$$Support = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ trasanctions} = P(A \cap B) \tag{3}$$

$$Confidence = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P(A \cap B)}{P(A)} \tag{4}$$

$$Coverage = \frac{Number\ of\ transactions\ with\ A}{Total\ number\ of\ trasanctions} = P(A) \tag{5}$$

$$ExpectedConfidence = \frac{Number\ of\ transactions\ with\ B}{Total\ number\ of\ trasanctions} = P(B) \quad (6)$$

$$Lift = \frac{Confidence}{ExpectedConfidence} = \frac{P(A \cap B)}{P(A) \cdot P(B)} \quad (7)$$

Support (eq.3) is the proportion of clusters in the dataset containing both the antecedent and the consequent (i.e. co-existence of patents in specific sectors within spatial clusters) and is used to measure the significance of the rule. Higher support indicates that the rule applies to a larger proportion of the dataset. Confidence (eq.4) is the proportion of clusters containing the antecedent (if) that also contain the consequent (then). This indicates the reliability of the rule; a higher confidence value means that when the antecedent occurs, the consequent is also more likely to occur (Agrawal et al. 1993). Coverage (eq.5) is the proportion of clusters containing the antecedent (if), indicating how frequently the antecedent appears in the dataset. Lift (eq.7) is the ratio of the observed confidence (eq. 4) to the expected confidence (eq.6) if the antecedent and consequent were independent. Values greater than one suggest a positive association, indicating that the antecedent and consequent occur together more frequently than would be expected by chance (Agrawal et al. 1993). Higher values of the quality metrics — support (eq.3), confidence (eq.4) and lift (eq.7) — indicate stronger, more reliable association rules.

Even though the above approach includes a causal relationship (*if - then*), it is static over time and causality is defined by Bayesian conditional probabilities. We have extended this approach by adding real temporal dynamics, enabling us to capture temporal relationships between technologies as catalyst and follower. The data, which is available from 1980 to 2015, has been divided into seven 5-year periods (t_1 to t_7). Each observation within spatial clusters was labelled with both a patent class and a time slot. We only used temporally logical rules where past observations were the antecedents ('if') and future (or same period) observations were the consequents ('then'), such as 'A_t1 → B_t2' or 'A_t2 → B_t5'. Therefore, association rule mining offers an interesting solution to the problem of asymmetry in current relatedness methods (Shahin et al. 2021) by resolving the issue of causality and the temporal impact of one technology on another.

The dataset used for association rule analysis should be structured like transactional data, such as shopping baskets or store invoices. Each record should include labels corresponding to specific categories. In our case, each spatial cluster comprising temporally labelled patent classes is equivalent to one shopping basket (or one invoice). To allow for deeper interpretation, we also labelled each patent class as either high-tech or low-tech. Therefore, each of the 344 clusters contains elements such as 'E06B_t4_Low-Tech', showing the patent subclass, temporal label and technology label.

Asymmetric technological relatedness based on spatio-temporal patent data

Figure 1 shows the analytical framework used to identify asymmetric technological relatedness based on spatio-temporal patent data - from building the dataset to running the analysis.

The process begins with the main research question: do certain technologies act as catalysts, influencing the emergence of others in an asymmetric way in technology clusters? After acquiring the data and filtering it for original patents (i.e. those cited by others), it is segmented into time periods (t_1 to t_7) of five years each between 1980 and 2015, and classified into high-tech and low-tech categories. The DBSCAN algorithm is then employed to perform spatial clustering and group patents according to geographical proximity. These clusters then act as the input for association rules. Each cluster contains patent types and time tags to identify temporal co-occurrence patterns. Association rules are then employed to determine the frequency with which patents appear together in the same geographical clusters and whether these co-occurrences are statistically non-random. Finally, we extract directional, time-informed rules such as $A_t_1 \rightarrow B_t_2$.

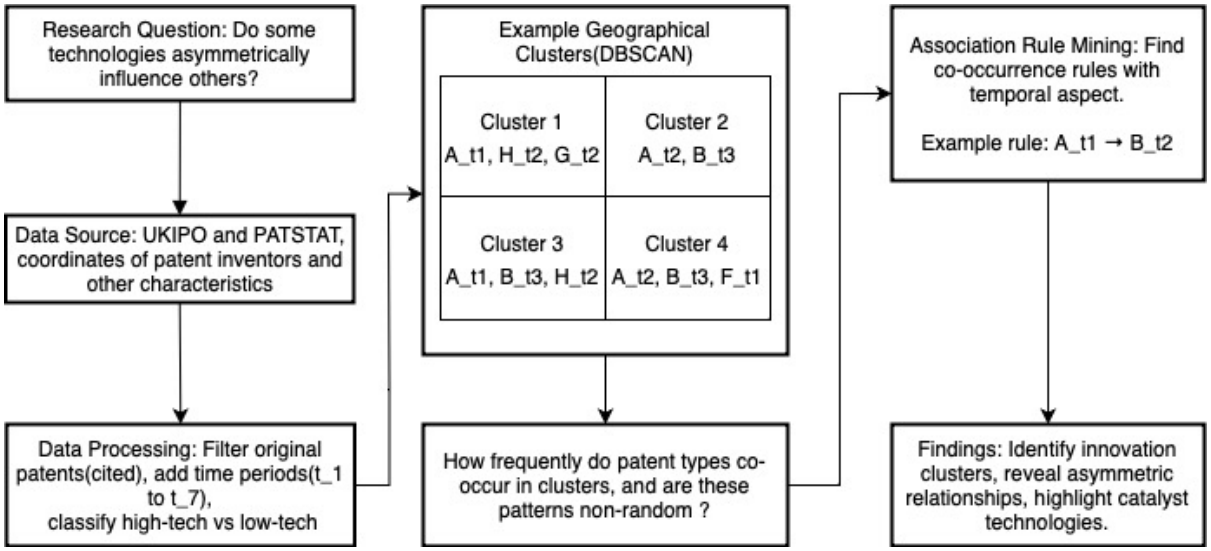


Figure 1. *Diagram showing the analytical workflow combining spatial clustering and association rule mining*

4 RESULTS

Step 1: Patent clusters in the UK

The first step of the analysis involved identifying groups of patents by inspecting the locations of spatial clusters. Figure 2 shows the DBSCAN clustering results, revealing a clear spatial concentration of patenting activity across the United Kingdom between 1980 and 2015. The largest cluster is located in London, indicating that a significant number of patents are concentrated in a major urban and industrial area. Some medium-sized clusters are found around Manchester, Liverpool, Birmingham and

Southampton. Several smaller clusters are scattered throughout the country, indicating localised hubs of innovation that are smaller than the primary centres. Overall, these patterns illustrate how DBSCAN identified specific geographic 'hot spots' of patent activity, providing insight into where inventive activity is most concentrated and where smaller yet still significant clusters of innovation exist. Most spatial clusters are local, containing up to 50 patents and covering an area of up to 100 km² (10 km x 10 km) (Figure 3).

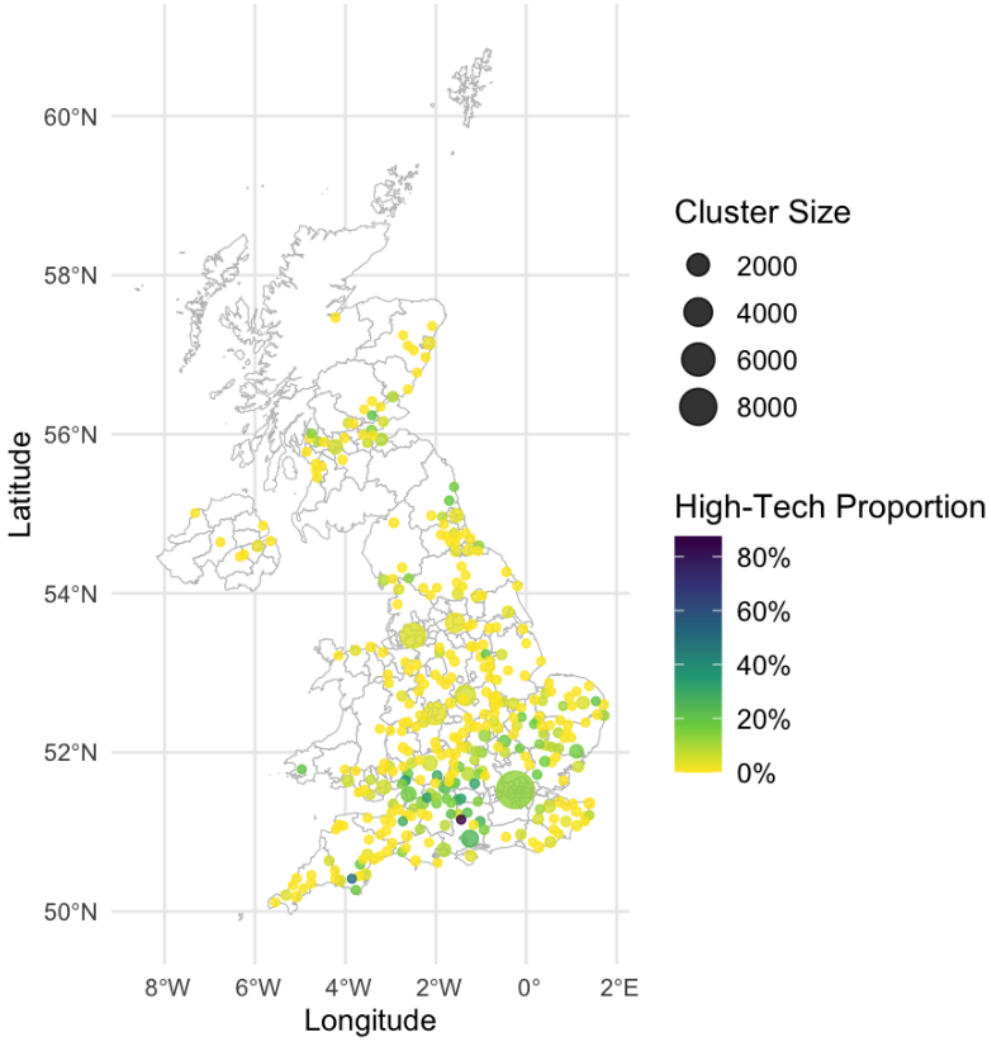


Figure 2. *DBSCAN clusters of patent activity with share of high-tech patents in clusters*

Note: *we used the DBSCAN algorithm with a radius of eps=0.045° (approximately 5 km) and a minimum density of minPts=5. This ensured that each cluster contained at least five points within a 5 km neighbourhood and yielded a total of 344 clusters. Many of the clusters are small and have a low share of high-tech technologies.*

An important element of this analysis is the high-tech/low-tech structure, which can be seen as colour in Figure 2. The results suggest a strong geographic concentration of high-tech patents in southern

England, particularly around Southampton, Bristol and London. These clusters exhibit high patent volumes and a relatively high proportion of high-tech patents, highlighting their role as core innovation hubs. In contrast, clusters in northern England, Scotland, and Wales tend to be smaller and have a lower proportion of high-tech patents. This spatial heterogeneity shows that high-tech innovation is not evenly distributed, and that cluster size does not fully correlate with high-tech specialisation. Further details can be found in the Appendix.

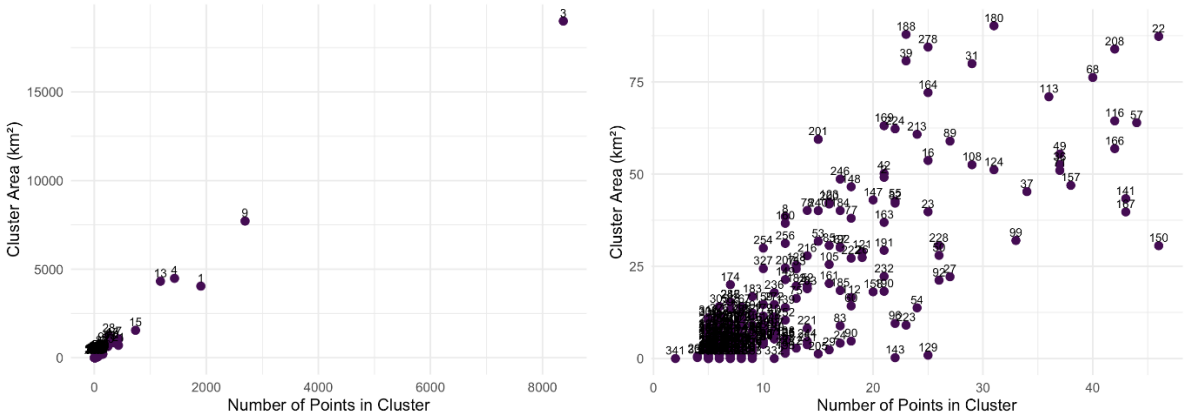


Figure 3. *Cluster size by number of patents (x-axis) and territorial coverage (y-axis): a) all clusters, b) zoom to show the smallest clusters*

Note: *The number of patents (points) in a spatial cluster is mostly up to 50 (left figure), but can be around 8,000 (right figure). Single clusters cover a territory of up to approximately 100 km² (10 km x 10 km) (left figure), while larger clusters of up to approximately 20,000 km² (100 km x 200 km) are also visible (right figure).*

Step 2: Catalytic patents

The second stage of the analysis involves dynamic association rules (see Table 1). Each rule indicates a directional relationship, whereby the presence of one patent class (if) is frequently followed by another patent class (then) within the same spatial cluster over time.

A total of 1,207 rules were created for the subsequent analysis of technologies playing a catalytic role. The first rule in Tab. 1 indicates that in period t4, the presence of patents in classes E05B (low-tech) and H04N (high-tech) in the same cluster was consistently followed by the presence of class H01L (high-tech) in period t5. This relationship occurred in five clusters (count = 5).

An overview of the listed rules reveals temporal and directional relationships. Clearly, drivers do not influence the cluster immediately; this takes several years. The analysis also hints at technological influence from high-tech to high-tech domains. The high value of lift (lift = 48.2) shows that the co-occurrence is non-random, i.e. the co-occurrence of patents is more likely than their individual appearance.

An overview of the listed rules shows that there are evident temporal and directional relationships. Drivers do not influence the cluster immediately – they require several years. It also hints at technological influence from "high-tech to high-tech" domains. The high value of lift (*lift = 48.2*) shows

that the co-occurrence is non-random, with the co-occurrence of patents being more likely to happen 48 times than their individual appearance.

Table 1. First 5 association rules out of 756 rules

No	Rules {if→then}	support	confidence	coverage	lift	count
1	{E06B_t4_Low-Tech,H04N_t4_High-Tech} => {H01L_t6_High-Tech}	0.021	1.00	0.021	48.20	5
2	{E06B_t4_Low-Tech,G06F_t2_High-Tech} => {H01L_t6_High-Tech}	0.021	1.00	0.021	48.20	5
3	{E06B_t4_Low-Tech,H04Q_t5_High-Tech} => {H01L_t6_High-Tech}	0.021	1.00	0.021	48.20	5
4	{E06B_t4_Low-Tech,H04M_t4_High-Tech} => {H01L_t6_High-Tech}	0.021	1.00	0.021	48.20	5
5	{G06F_t2_High-Tech,G06F_t4_High-Tech} => {A01G_t5_Low-Tech}	0.021	1.00	0.021	48.20	5

Note: This table presents the first 5 rules, that were identified with A priori algorithm. The 'transaction' is the spatial cluster, while 'items' are patents. The listed outcomes are the most often occurring relationships. Explanations of patents codes are in Appendix.

A case-by-case analysis of the rules reported in the table is challenging due to the large amount of output, so a more synthetic overview is needed. Figure 4 shows the most frequent rules between the patents. Most causal frequent relationships occur between high-tech patents.

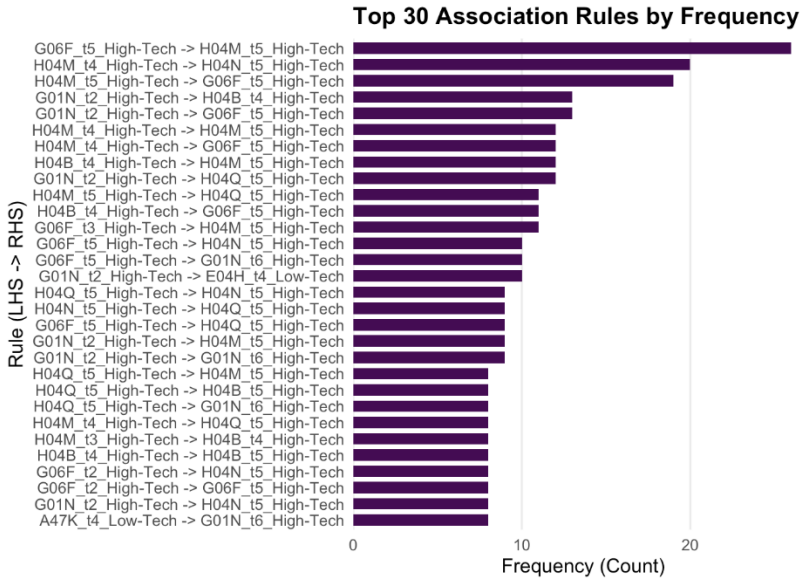


Figure 4. *Most influential technologies in association rules*

Note: *Association rules may produce outcomes in which the left-hand side (LHS) {if} has many elements. This figure shows the frequency of one-to-one associations to highlight the most common ones.*

The core of the study is to identify the presence of asymmetry, i.e. whether the frequency of occurrence of patent classes is the same on both sides of the rules (i.e. on the *{if}* and *{then}* sides), or whether some patent classes occur more frequently on one side or the other, acting as a driver (*{if}*, LHS) or a consequence (*{then}*, RHS). We therefore checked the frequency of occurrence on both sides and

visualised the difference between the LHS and the RHS. Positive values (LHS > RHS) indicate drivers, while negative values (LHS < RHS) indicate followers (Figure 5).

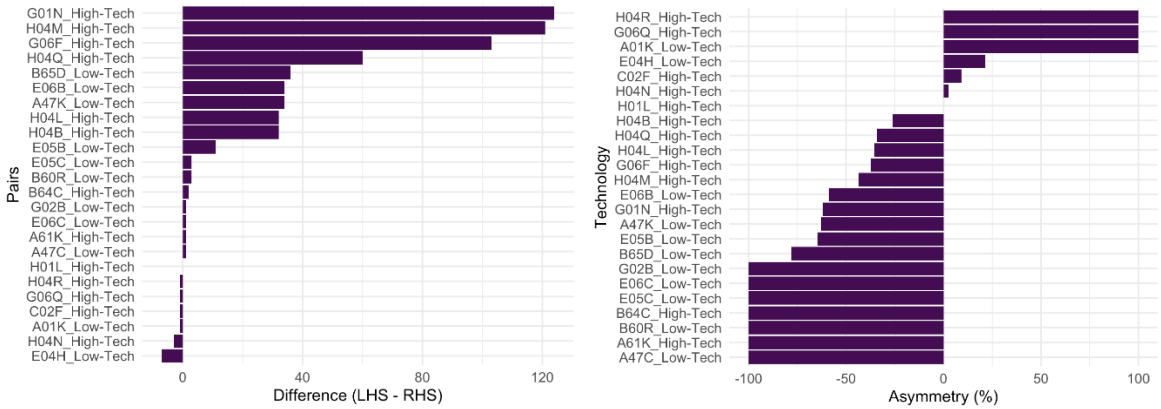


Figure 5. *Most influential technologies in association rules*

Note: *Among 1,207 rules we checked how many times particular patent class appears as {if} and as {then} and calculated a difference between those counts. The top patent classes are the most influential drivers of patents in future. The figure reports data aggregated over time (no time labels included).*

The top-listed patents are key drivers in terms of technological relatedness, indicating that they are likely to influence the occurrence of other technologies. A higher positive value indicates that a technology appears more frequently as an initiator in the rules (if). Patents at the top of the list play a catalytic or influential role in the emergence of other technologies. Technologies such as G01N, H04M and H06F are among the top contributors on the LHS *{if}* side. Figure 5a shows the absolute values for all created rules – the top-listed patents appear 120 times more frequently as drivers than as followers. Figure 5b illustrates the relative frequency, expressed as (LHS-RHS)/(LHS+RHS) and understood as the excess of the *{if}* role in all occurrences. The top-listed patents in Figure 5b (with a relative value of 100%) are technologies that never appear as *{then}* – they always act as *{if}*.

Interestingly, both high-tech (e.g. G01N and H04M) and low-tech (e.g. B65D and E06B) technologies appear as influential LHS components. This suggests that influence depends not only on technological classification, but also on the specific role that a technology plays within the innovation system, such as infrastructural, foundational, or integrative. This asymmetry in positioning within association rules provides meaningful insight into the directionality of innovation flows and helps identify potential catalyst technologies. This finding supports the notion of asymmetric technological relatedness, whereby some technologies act as catalysts while others tend to follow.

A more in-depth analysis can be achieved by examining the intensity of the interactions between technologies (Figure. 6) and the number of pairs of drivers and followers. The heatmap reveals notable clustering patterns, showing that specific technologies frequently co-occur in rules and form strong

associations. For example: H04Q (selecting), H04N (pictorial communication, e.g. television), H04M (telephonic communication) and H04L (transmission of digital information, e.g. telegraphic communication) frequently appear together in rules, indicating dense interaction networks among these technologies.

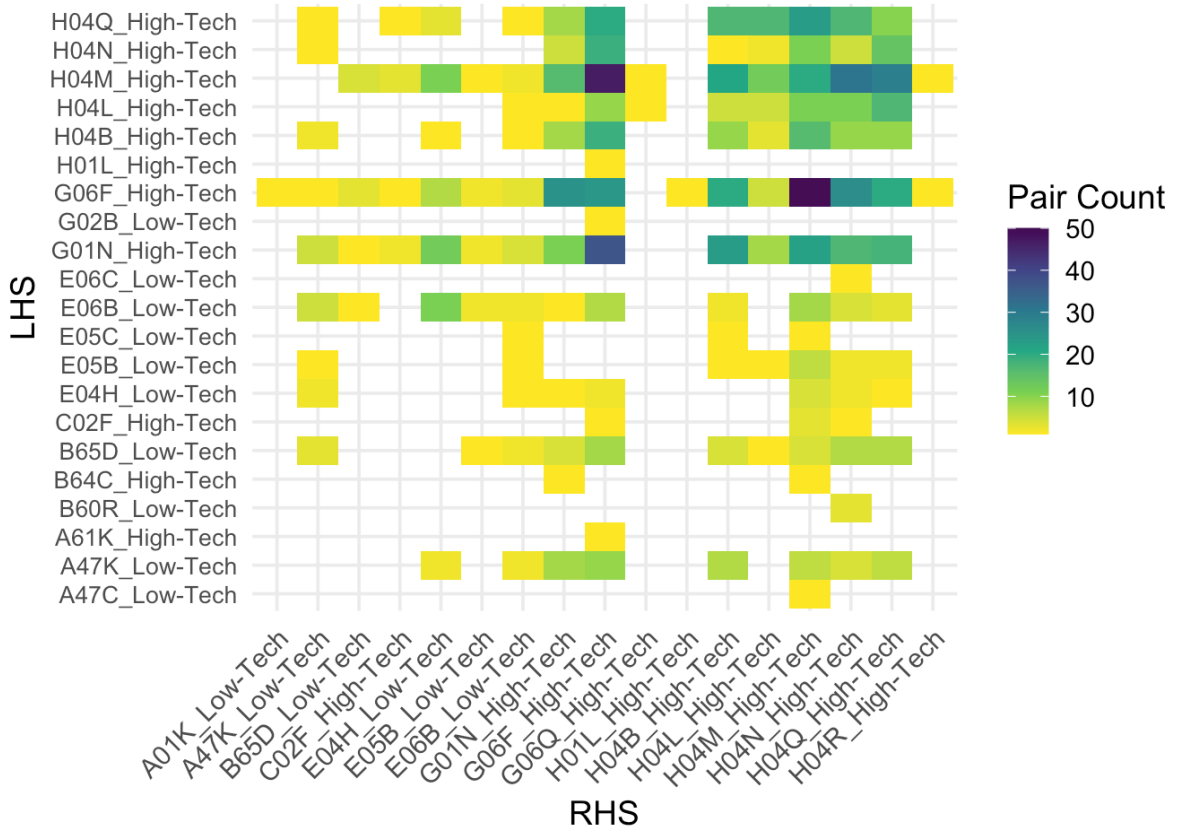


Figure 6. *Heatmap of pairwise interactions between technologies appearing as drivers (LHS) and followers (RHS)*

Note: *Each box represents the co-occurrence of patents in spatial clusters. The colour of the box reflects the strength of the relationship, i.e. how often a pair appeared in association rules. The most frequent boxes are in the top right corner and are coloured green. These relate to H04Q (selecting), H04N (pictorial communication, e.g. television), H04M (telephonic communication) and H04L (transmission of digital information, e.g. telegraphic communication).*

These clusters suggest technological areas in which joint innovation is common. The presence of these dense interaction zones lends weight to the idea that certain technological fields evolve as interconnected clusters rather than in isolation. These technological clusters may reflect shared knowledge bases or functional interdependencies. Such findings support the concept of related variety (Boschma and Iammarino 2009), whereby diversity within a technologically coherent domain enhances innovation and knowledge spillover opportunities.

These results can provide an overview by aggregating drivers and followers into high-tech and low-tech categories. Table 2 shows the occurrence of technology types (high-tech and low-tech) in the detected 'if' and 'then' rules. This reveals a clear asymmetry in technological influence between high-tech and low-tech patents. When a high-tech patent appears first, another high-tech patent follows in 68% of cases (2/3 of rules), whereas a low-tech follow-up occurs in only 8.2% of cases (1/12). Conversely, low-tech patents are followed by high-tech patents in 18.2% of cases and by other low-tech patents in just 5.6% of cases. These findings suggest that high-tech patents are self-reinforcing and play a central role in shaping the innovation landscape. Their influence extends beyond their own domain, reinforcing their position as catalyst technologies. In contrast, low-tech patents tend to play a more passive role, rarely driving subsequent innovation within or outside their category. This supports the notion of asymmetric relatedness, whereby certain technologies — particularly high-tech ones — influence the emergence and trajectory of low-tech patents.

Table 2. Proportion of {if→then} rules in the division into high-tech and low-tech sectors

		THEN	
		High-tech	**Low-tech**
IF	**High-tech**	68.0%	8.2%
	Low-tech	18.2%	5.6%

Note: The total of shares in this table is 100%.

The detected relationships can naturally be presented as a network graph. Figure 7 shows a directed network of patent technologies, where the direction of influence is indicated by the arrows and is based on association rules derived from spatial clusters. Each node represents a distinct technology, and each edge represents a directional rule, whereby the originating technology precedes and likely influences the receiving technology. The colour intensity and thickness of each arrow correspond to the frequency of the associated rule: darker, thicker arrows indicate stronger and more frequently observed co-occurrence patterns. The distance between the nodes represents the number of connections a patent has: the closer the nodes, the more connections.

Both networks (Figure 7a and Figure 7b) reveal a dense core of high-tech technologies with strong directional flows, particularly within the H04 series (communication-related technologies such as H04Q, H04M, H04L and H04N) and G06F (data processing). These technologies interact frequently with one another and serve as initiators in multiple rules, as evidenced by the large number of outgoing arrows. This pattern suggests that they play a catalytic role in driving innovation across the cluster. The directed edges further confirm the presence of asymmetric technological relatedness: while some technologies consistently lead innovation pathways, others tend to follow. This visual representation reinforces the study's key insight. Innovation is not evenly distributed among technologies; rather, certain domains act as hubs or drivers of technological evolution within regional clusters.

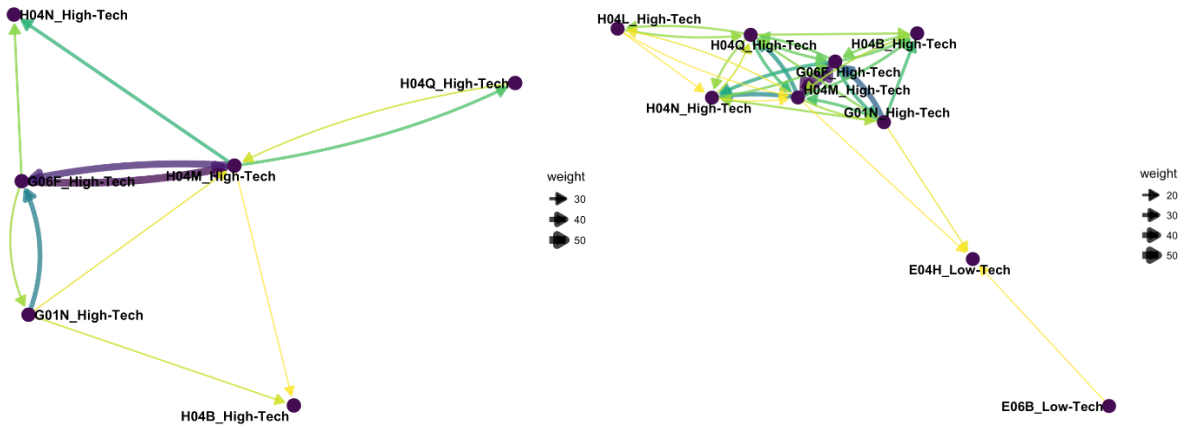


Figure 7. *Network graph of frequent patent interactions: a) for patents occurring at least in 20 rules, b) for patents occurring at least in 10 rules.*

Note: *Edge thickness and colour represent the frequency of co-occurrence; the graph shows directional influence based on association rule mining. Distance indicates the number of connections a patent has. The more observations required, the less patents fulfil the conditions.*

Taken together, the results show that technological relatedness is not always symmetrical. Some technologies consistently emerge earlier in certain regions and are associated with a wider variety of others, suggesting a directional influence. This lends weight to the concept of 'catalyst technologies' that drive innovation pathways in regional technology clusters.

5 CONCLUSIONS

We study the empirical relevance of catalytic technologies (patents) as a driving force for the emergence of new technologies (patents) in technology clusters using data from the UK. While previous research has often assumed symmetrical relationships between technologies (Boschma 2017), our findings demonstrate that certain technologies (particularly high-tech patents) influence the emergence of others.

By applying spatial clustering (DBSCAN) and temporally ordered association rule mining, we address a methodological shortcoming in the literature and provides a new approach to identifying catalyst technologies. The method uses point data instead of regional data and uses a dynamic approach instead of a static one. Thereby, our analyses overcome the well-known problem of loss of information about how patent structures change over time. We introduce an original unsupervised machine learning framework that uses clustering and association rule mining that overcome many limitations of existing studies.

Our results highlight the importance of considering directionality in innovation pathways, suggesting that policy efforts to stimulate regional innovation should prioritise technologies with high asymmetric influence. We document that there are evident 'catalyst patents' that drive patent activity in related technologies in clusters. Asymmetric relatedness appears to be an important characteristic of the process by which new technologies emerge, as evidenced by the development of patents over time. Using unsupervised machine learning techniques such as DBSCAN clustering and association rule mining has provided valuable insights into the spatio-temporal dynamics of patent clusters. Understanding these patterns enables the development of strategies to enhance regional innovation capabilities, promote economic growth and ensure resilience against technological and economic shifts.

The use of temporal association rules in unsupervised machine learning algorithms can be extended to several different areas, such as firm-level innovation studies. The methodological framework developed is highly adaptable and can be replicated across a wide range of datasets, not just patent records. Where data contain geographic coordinates to identify spatial relationships and time stamps to capture the temporal aspect, the combination of spatial clustering (DBSCAN) and association rule mining can be employed to reveal asymmetric relationships. This makes the approach applicable to various domains, such as scientific publications, start-up formation, firm-level innovation and research collaborations. The flexibility of unsupervised learning techniques, such as the DBSCAN algorithm and the Apriori algorithm, allows researchers to apply this framework across different scales, sectors or countries. This makes it a versatile tool for detecting directional influence patterns in spatial and temporal innovation processes.

The analyses contribute to spatial analysis by offering new perspectives on the impact of technological relatedness on regional development but have some limitations. Firstly, it relies on historical data, which may not fully reflect the latest trends in patent clustering. Second, the use of citation-based data means that fewer recent patents are included, as they are cited less frequently. While this reduction could complicate the analysis, focusing on cited patents strengthens the results by excluding those with limited impact. Third, while the unsupervised learning approach provides new insights, it also poses challenges in terms of interpreting and validating the identified clusters. Fourth, the geographic concentration of patenting activities could skew the results towards more active regions and cause emerging technological hubs to be overlooked. Despite these limitations, the research makes valuable contributions and lays the groundwork for future studies to build upon and refine our understanding of spatial and technological dynamics.

REFERENCES

Agrawal, A., & Cockburn, I. (2003). The anchor tenant hypothesis: exploring the role of large, local, R&D-intensive firms in regional innovation systems. *International journal of industrial organization*, *21*(9), 1227-1253.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference. In *Sigmod Record* (Vol. 22, p. 207).

Andersson, M., Klaesson, J., & Larsson, J. P. (2016). How local are spatial density externalities? Neighbourhood effects in agglomeration economies. *Regional studies*, *50*(6), 1082-1095.

Arbia, G., & Petrarca, F. (2011). Effects of MAUP on spatial econometric models. *Letters in Spatial and Resource Sciences*, *4*(3), 173–185.

Balland, P. A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Regional studies*, *53*(9), 1252-1268.

Boschma, R. (2017). Relatedness as driver of regional diversification: A research agenda. *Regional Studies*, *51*(3), 351–364.

Boschma, R., & Iammarino, S. (2009). Related Variety, Trade Linkages, and Regional Growth in Italy. *Economic Geography*, *85*, 289–311.

Boschma, R., Miguelez, E., Moreno, R., & Ocampo-Corrales, D. B. (2023). The Role of Relatedness and Unrelatedness for the Geography of Technological Breakthroughs in Europe. *Economic Geography*, *99*(2), 117–139.

Breschi, S., Lissoni, F., & Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy*, *32*(1), 69–87.

Bresnahan, T. (2010). General purpose technologies. *Handbook of the Economics of Innovation*, *2*, 761-791.

Buzard, K., Carlino, G. A., Hunt, R. M., Carr, J. K., & Smith, T. E. (2020). Localized knowledge spillovers: Evidence from the spatial clustering of R&D labs and patent citations. *Regional Science and Urban Economics*, *81*, 103490.

Crescenzi, R., Dyèvre, A., & Neffke, F. (2022). Innovation catalysts: how multinationals reshape the global geography of innovation. *Economic Geography*, *98*(3), 199-227.

Dimos, C., Fai, F. M., & Tomlinson, P. R. (2021). The attractiveness of university and corporate anchor tenants in the conception of a new cluster. *Regional Studies*, *55*(8), 1473-1486.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Knowledge Discovery and Data Mining.

EUROSTAT (2024), Eurostat indicators on High-tech industry and Knowledge – intensive services, Appendix 6, https://ec.europa.eu/eurostat/cache/metadata/Annexes/htec_esms_an_6.pdf

Feldman, M. (2003). The locational dynamics of the US biotech industry: knowledge externalities and the anchor hypothesis. *Industry and innovation*, *10*(3), 311-329.

Foray, D. (2014). *Smart specialisation: Opportunities and challenges for regional innovation policy*. Routledge.

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

Fritsch, M., & Wyrwich, M. (2023). Entrepreneurship in the long-run: Empirical evidence and historical mechanisms. *Foundations and Trends® in Entrepreneurship*, *19*(1), 1-125.

Giuliani, E., & Bell, M. (2005). The micro-determinants of meso-level learning and innovation: evidence from a Chilean wine cluster. *Research policy*, *34*(1), 47-68.

Hasan, I., & Tucci, C. L. (2010). The innovation–economic growth nexus: Global evidence. *Research policy*, *39*(10), 1264-1276.

Herzog, S., Mason, C., & Hruskova, M. (2024). The role of large corporations in entrepreneurial ecosystems–a case study of Munich. *European Planning Studies*, *32*(6), 1295-1317.

Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. MIT press

Kim, S. H., Jun, B., & Lee, J. D. (2023). Technological relatedness: how do firms diversify their technology?. *Scientometrics*, *128*(9), 4901-4931.

Klepper, S. (2010). The origin and growth of industry clusters: The making of Silicon Valley and Detroit. *Journal of urban economics*, *67*(1), 15-32.

Klepper, S. (2015). *Experimental capitalism: The nanoeconomics of American high-tech industries*. Princeton University Press.

Kopczewska, K. (2022). Spatial machine learning: New opportunities for regional science. *The Annals of Regional Science*, *68*(3), 713–755.

Kuusk, K., & Martynovich, M. (2021). Dynamic nature of relatedness, or what kind of related variety for long-term regional growth. *Tijdschrift voor economische en sociale geografie*, *112*(1), 81-96.

MacDonald, P., & Selmanovic, S. (2024). How localised are knowledge spillovers? Evidence from microgeographic data on UK patent citations. *Economics of Innovation and New Technology*, *33*(2), 323–343.

Mascarini, S., Garcia, R., Costa, A., Santos, E., & Araujo, V. (2023). Regional heterogeneity and the effects of related and unrelated variety on innovation. *Regional Science Policy & Practice*, *15*.

Mason, C., & Brown, R. (2014). Entrepreneurial ecosystems and growth oriented entrepreneurship. *Final report to OECD, Paris*, *30*(1), 77-102.

McCann, P., & Ortega-Argilés, R. (2014). The role of the smart specialisation agenda in a reformed EU cohesion policy. *Scienze regionali: Italian Journal of regional Science: 13, 1, 2014*, 15-32.

McHale, J., Harold, J., Mei, J. C., Sasidharan, A., & Yadav, A. (2023). Stars as catalysts: an event-study analysis of the impact of star-scientist recruitment on local research performance in a small open economy. *Journal of Economic Geography*, *23*(2), 343-369.

Mewes, L., & Broekel, T. (2022). Technological complexity and economic growth of regions. *Research Policy*, *51*(8), 104156.

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.

Nilsson, M., Schubert, T., & Miörner, J. (2025). The productivity effects of regional anchors on local firms in Swedish regions between 2007 and 2019—evidence from an expert-informed machine learning approach. *Journal of Economic Geography*, *25*(2), 215-234

Rigby, D. (2013). Technological Relatedness and Knowledge Space: Entry and Exit of US Cities from Patent Classes. *Regional Studies*, *49*, 1–16.

Rigby, D. L. (2015). Technological relatedness and knowledge space: Entry and exit of US cities from patent classes. *Regional Studies*, *49*(11), 1922-1937.

Shahin, M., Pious, S., Sharma, R., Kaushik, M., Ben Yahia, S., Shah, S., & Draheim, D. (2021). *Big Data Analytics in Association Rule Mining: A Systematic Literature Review* (p. 49).

Wixe, S., & Andersson, M. (2017). Which types of relatedness matter in regional growth? Industry, occupation and education. *Regional studies*, *51*(4), 523-536.

Xiao, J., Boschma, R., & Andersson, M. (2018). Industrial diversification in Europe: The differentiated role of relatedness. *Economic Geography*, *94*(5), 514-549.

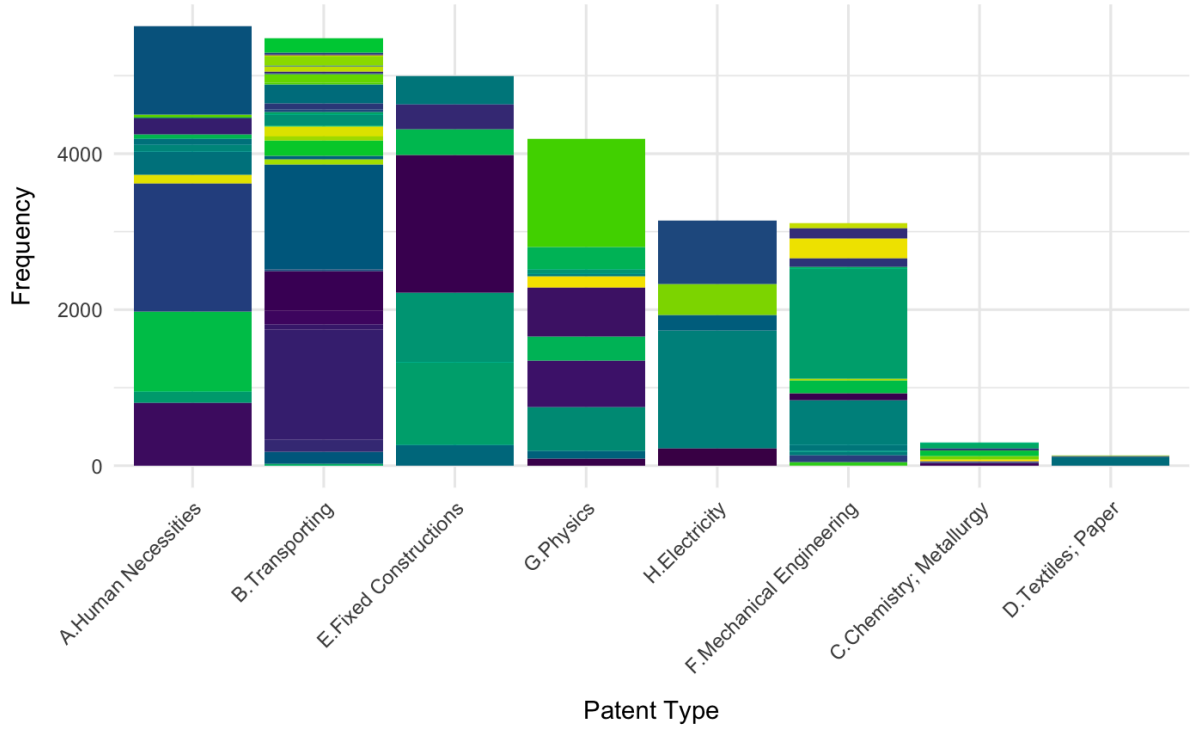
Data Availability: This study uses the same dataset as in paper by MacDonald & Salmanovic (2023) (https://www.tandfonline.com/doi/abs/10.1080/10438599.2023.2184809). Data are available at https://figshare.com/articles/dataset/Replication_data_for_How_localised_are_knowledge_spillovers_Evidence_from_microgeographic_data_on_UK_patent_citations/16771195. We used EUROSTAT (2024) classification to divide patents into high- and low-tech – classification is available at: https://ec.europa.eu/eurostat/cache/metadata/Annexes/htec_esms_an_6.pdf. Patent classification can be checked at: https://www.uspto.gov/web/patents/classification/

Reproducibility of study: The full R code for this study is available at RPubs at: https://rpubs.com/zehrausta/1307953

Financing: The paper is a part of the project "Tandems for excellence" financed within IDUB programme (Inicjatywa Doskonałości – Uczelnia Badawcza / Initiative of Excellence – Research University) at the University of Warsaw, Poland.

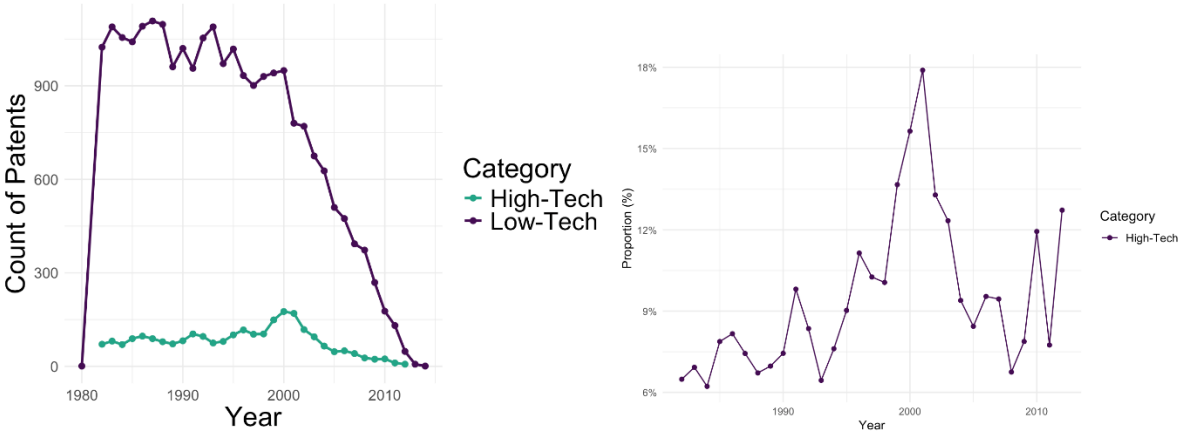
Appendix

Figure A1: Composition of dataset by patent classes (8 categories) and subclasses (313 subcategories) (sorted decreasingly by main class)



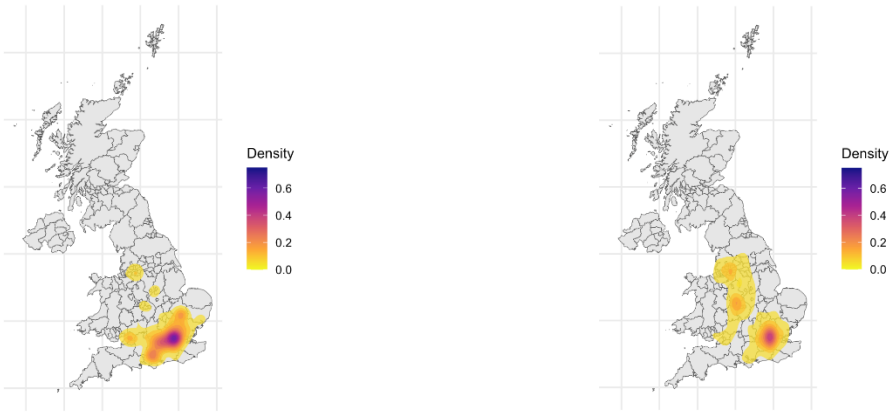
Note: This figure shows that classes and subclasses of patent activity differ by size, from the largest class A to the smallest class D.

Figure A2: Distribution of cited patents by time, in division into low-tech and high-tech: a) nominal values, b) share of high-tech patents

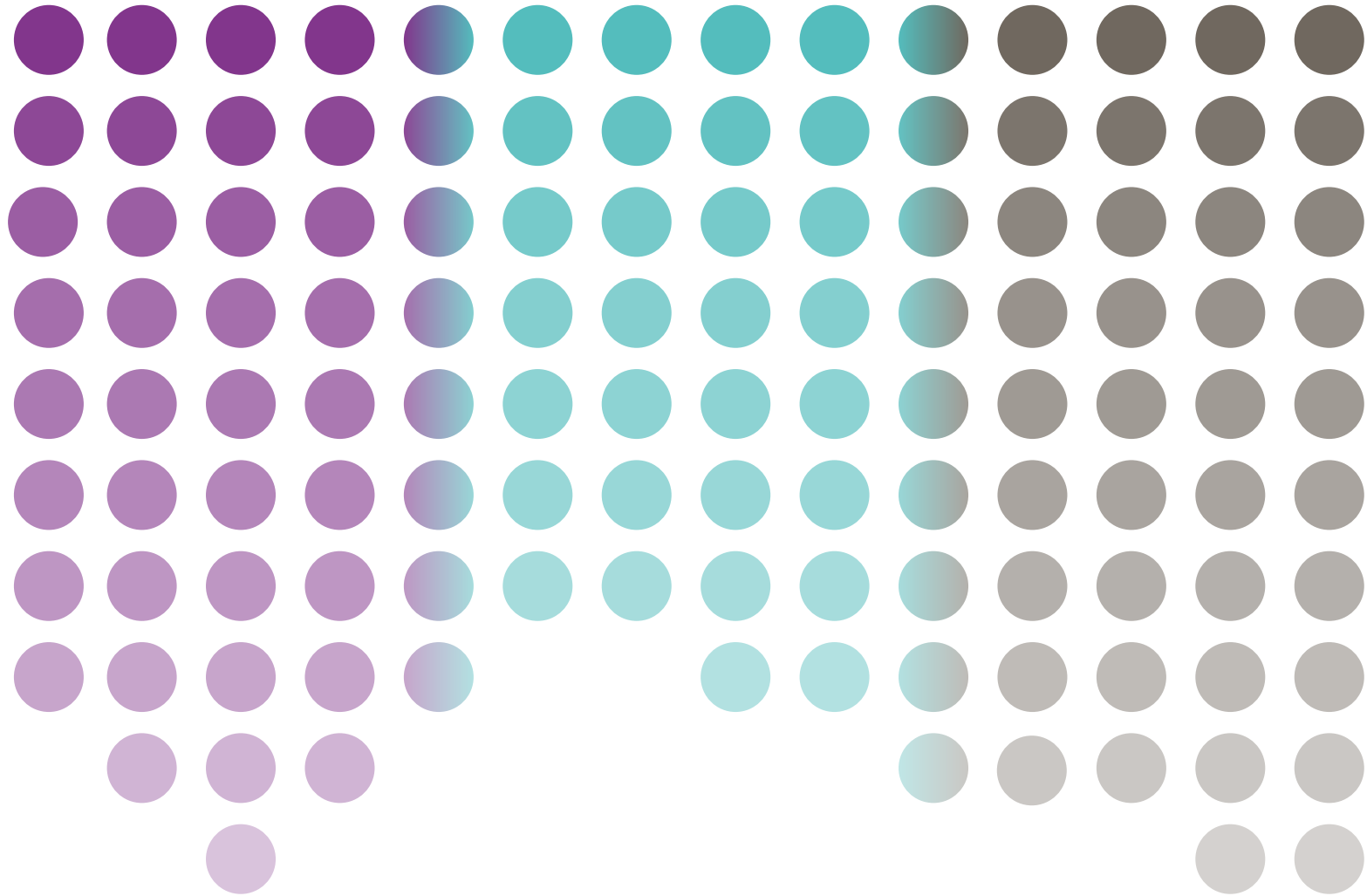


Note: The number of cited patents decreases over time as it takes time for patents to become cited. Share of high-tech patents increases over time.

Figure A3: Distribution of cited patents by space, in division into low-tech and high-tech: a) frequency of high-tech patents, b) frequency of low-tech patents



Note: The frequency of patent appearance was calculated with Kernel Density Estimation. Left panel shows that high-tech patents are highly concentrated around London, Southampton and Bristol. Right panel shows that there are three hubs of low-tech patents, around London, Birmingham and Manchester.



RESEARCH NETWORK DEBATE

SWEDISH ENTREPRENEURSHIP FORUM

WWW.ENTREPRENORSKAPSFORUM.SE