systems

MDPI

Article

A Patent Mining Approach to Accurately Identifying Innovative Industrial Clusters Based on the Multivariate DBSCAN Algorithm

Siping Zeng [1], Ting Wang [2], Wenguang Lin [2], Zhizhen Chen [3] and Renbin Xiao [4,*]

[1] School of Economics and Management, Xiamen University of Technology, Xiamen 362114, China; 2015000023@xmut.edu.cn
[2] School of Mechanical and Automotive Engineering, Xiamen University of Technology, Xiamen 361024, China; wt22992022@163.com (T.W.); linwg@xmut.edu.cn (W.L.)
[3] School of Business, University of Greenwich, London SE10 9LS, UK; z.chen@gre.ac.uk
[4] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
* Correspondence: rbxiao@hust.edu.cn

Abstract: Innovative Industrial Clusters (IIC), characterized by geographical aggregation and technological collaboration among technology enterprises and institutions, serve as pivotal drivers of regional economic competitiveness and technological advancements. Prior research on cluster identification, crucial for IIC analysis, has predominantly emphasized geographical dimensions while overlooking technological proximity. Addressing these limitations, this study introduces a comprehensive framework incorporating multiple indices and methods for accurately identifying IIC using patent data. To unearth latent technological insights within patent documents, Latent Dirichlet Allocation (LDA) is employed to generate topics from a collection of terms. Utilizing the applicants' names and addresses recorded in patents, an Application Programming Interface (API) map systems facilitates the extraction of geographic locations. Subsequently, a Multivariate Density-Based Spatial Clustering of Applications with Noise (MDBSCAN) algorithm, which accounts for both technological and spatial distances, is deployed to delineate IIC. Moreover, a bipartite network model based on patent geographic information collected from the patent is constructed to analyze the technological distribution on the geography and development mode of IIC. The utilization of the model and methodologies is demonstrated through a case study on the China flexible electronics industry (FEI). The findings reveal that the clusters identified via this novel approach are significantly correlated with both technological innovation and geographical factors. Moreover, the MDBSCAN algorithm demonstrates notable superiority over other algorithms in terms of computational precision and efficiency, as evidenced by the case analysis.

Keywords: bipartite network; density-based spatial clustering; innovative industrial clusters; latent Dirichlet allocation; patent analysis

Citation: Zeng, S.; Wang, T.; Lin, W.; Chen, Z.; Xiao, R. A Patent Mining Approach to Accurately Identifying Innovative Industrial Clusters Based on the Multivariate DBSCAN Algorithm. *Systems* **2024**, *12*, 321. https://doi.org/10.3390/systems12090321

Academic Editor: Sitalakshmi Venkatraman

Received: 16 July 2024
Revised: 20 August 2024
Accepted: 22 August 2024
Published: 24 August 2024

1. Introduction

Industrial clusters are pivotal in enabling regions to garner, establish, and maintain competitive advantages, acting as catalysts for economic dynamism and innovation [1]. The concept of industrial clusters was first introduced by Marshall [2] to describe the phenomenon of geographic aggregation among interrelated enterprises, suppliers, and institutions within a specific industry. Through industrial clusters, enterprises gain enormous benefits by minimizing the transportation time and production cost [3], sharing resources, and constantly collaborating [4]. In China, more than 70% of the entire supply chain products can be produced within a cluster with a 50–200 km radius, which reduces logistic costs by more than 30% [5]. In addition, Japanese research in 2001 reported that enterprises within clusters clearly outperform enterprises outside the clusters in terms of intensity, cooperation, and the number of achievements in Research and Development (R&D) [6].

In recent years, with the rapid development of new technologies such as integrated circuits, artificial intelligence, and biological sciences, technology and those resources are simultaneously concentrated, forming a new form of industrial cluster, which is defined as IIC. Unlike traditional industrial clusters, which depend on natural resources and human resources, IIC have created significantly higher profits since they increase the innovative ability and the application of intelligent resources of the enterprise. The most well-known example is Silicon Valley in California and the chemical engineering industry clusters in Boston [7,8]. The recent success of technological development in China followed a similar path, such as Beijing Zhongguancun Mobile, Zhejiang Shaoxing Textile Machinery, and Fujian Jinjiang Sports [9,10]. Therefore, they have been implementing supportive policies to continuously encourage the development of IIC. For instance, the "Guiding Opinions on Further Promoting the Development of Industrial Clusters Strategy in 2015" and the "Interim Measures for Promoting the Development of Characteristic Industrial Clusters of Small and Medium-sized Enterprises (SMEs) Strategy in 2022" aim to identify around 200 local IIC across China and assist them in becoming larger IIC from the province-level [11–13].

The competition for technological resources between countries has shifted from single-technology competition to competition for IIC. Identifying technology clusters from the perspective of patented technologies helps enterprises and R&D personnel to understand the technology structure and explore technology opportunities so as to gain advantages in technology R&D and competition. It is of great significance to promote the specialized division of labor and collaboration among enterprises, effectively allocate production factors, reduce the cost of innovation and entrepreneurship, save social resources, and promote regional economic and social development [14].

Therefore, the key to technological development and policy effectiveness depends on the accurate identification of IIC. On the one hand, the accurate identification of IIC can provide systematic suggestions for effective and efficient collaboration and coordination among different regions within the same industry by avoiding intra-cluster competition. On the one hand, effective policymaking for regional industrial development from both spatial and technological perspectives relies on the accurate identification of IIC. Government policies on clusters aim to create an ecological environment conducive to improving the efficiency and reducing the cost of collaboration for firms within clusters. However, the formulation of such policies faces great risks, because how clusters are classified affects the policy subsidies of a large number of enterprises. Traditional cluster identification methods often rely on the administrative region, ignoring the industrial cooperation between different regions and the technical relevance of the cluster. In this paper, the cluster identification method based on patent big data can provide an objective and reasonable tool for the designation of cluster policy. However, the task of precisely identifying IIC has proven to be challenging for policymakers worldwide, including those in China, the US, and EU countries, as innovative industries and enterprises are embedded in every sector of the national economy, both technologically and geographically [3]. In practice, the identification of IIC needs to meet Objectivity–Accuracy–Width–Universality (OAWU) indicators: objectivity (i.e., to avoid the influence of subjective factor), accuracy (i.e., to avoid the omission of related areas or the inclusion of irrelevant areas in the cluster), width (i.e., to ensure the data comprehensiveness), and universality (i.e., to ensure the method can be applied to multiple fields) [15].

Currently, both survey methods and data-based methods are widely used among scholars to satisfy the OAWU criteria. Although common survey methods include questionnaire surveys and personal interviews, which have strong versatility and professionalism, they have some challenges [16,17]. On the one hand, the data collected from questionnaire surveys are likely limited due to cost and time restrictions. On the other hand, expert interviews have strong professionalism, simple operation, and high universality, but they depend on interviewees' expertise and can be significantly influenced by personal subjectivity. By contrast, data-based methods using data-mining technology based on huge market information have become the mainstream, and they integrate traditional statistical analysis tools with machine learning methods. By obtaining a large amount of related data, data-mining technology identify and cluster spatially distributed objects, including enterprises, scientific research institutions, and even individuals [18,19]. Hence, with this method, researchers can provide a more comprehensive understanding of the actual situation of IIC with minimized influences from investigators.

Previous research on IIC identification through big data mainly focuses on geographical information and largely ignores the influence of specific technological factors involved in the industry, which leads to inaccurate identification results on IIC. For instance, rather than considering the overall technological performance of the cluster from a system perspective, the Moran index model, which is commonly used, can only identify clusters formed by several neighboring units with indicators above the average [20]. In addition, the data used in existing research often refer to the open resources of social and economic data, which can satisfy the requirements of high volume, velocity, variety, veracity, and value (5V). Still, they are less relevant to the technology involved in the industry, which is not likely to meet the identification expectations of IIC.

In summary, this research proposes an improved algorithm of IIC identification, which is characterized by a big-data-driven method and allows for a thorough exploration of all aspects and attributes related to IIC, thereby mitigating the limitations of data scarcity. Furthermore, this research integrates a wide range of machine learning methods, including unsupervised learning such as text mining and complex networks, providing a more fine-grained, convenient, accurate, and universal identification method for IIC from the perspective of multi-technologies. The paper is structured as follows: Section 2 provides a summary of the relevant literature. Section 3 presents the research framework and methods, followed by a research example in Section 4. Section 5 offers an analysis of the results, and Section 6 presents the conclusion.

2. Literature Review

This paper aims to propose an unsupervised IIC recognition method combining big data mining and a machine learning algorithm. This section discusses the research reviews of the IIC definition, data sources, and identification algorithms.

2.1. The Definition of IIC

The definition of IIC is yet to be universally agreed upon. Voyer [21] defined IIC as "a knowledge-based industrial cluster, which means that enterprises including manufacturers, suppliers and service providers form regional or urban clusters across multiple industries" [22]. Affected by the supply chain, the cooperation intensity of enterprises in the cluster is higher than that outside of the cluster. Innovation clusters have been identified as a means of efficiently satisfying market requirements and encouraging the development of innovative technologies, allowing knowledge-based economies to access knowledge resources more easily [23,24]. Moreover, IIC has played a crucial role in supporting high-tech enterprises in the incubation environment [25]. Knowledge acquisition is affected by spatial distance, which means that enterprises within IIC can assimilate knowledge more easily than enterprises outside, giving them a better chance to use innovative knowledge resources to develop and obtain new market share [26,27]. Meanwhile, IIC can gather a large number of technological talents, which can significantly improve the efficiency of R&D within the cluster [28]. Especially, adopting IIC is a strategy for SMEs enhancing market competitiveness by reducing the investment related to logistics and electricity [29]. For SMEs, these expenses constitute a larger proportion of their overall costs than larger enterprises. However, it should be noted that most IIC are just one chain of multiple technologies, which means the absence of certain technologies can lead to the incompleteness and inefficiency of an industry's operation. Thus, Xu et al. [30] further define IIC as "a complex and sophisticated form of cooperative innovation, which is embedded in every step of the industrial chain", thus forming an innovation system of technology integration and expansion.

Besides academic definitions, governments and authorities have also proposed different definitions of IIC. For example, the United States National Research Council [31] defines regional IIC as "the agglomeration of enterprises providing innovative products and services, as well as suppliers and research institutions". In IIC, members have gathered a large number of knowledge- and skill-related resources and benefit from cooperative exchange. More than this, China Ministry of Science and Technology [11] defines IIC as "the enterprises in the industrial chain. R&D and service institutions gather in the region and form a competitive cross-industry and cross-regional industrial organization system through divisional works and collaborative innovation".

The differentiation of IIC from science parks and traditional industrial clusters is evident in current definitions. First, IIC focus on specific industries, and its spatial distance could be far greater than that of traditional science and technological parks [32]. Meanwhile, the development of IIC depends on the knowledge and technological resources within the cluster, and this emphasizes the importance of the coordination and collaboration of diverse technologies within the industry. This aligns with the characteristics of the existing science and technology industry, which is characterized by many enterprises, knowledge centralization, and technological diversification. In contrast, traditional industrial clusters rarely involve the level of industrial technology and primarily focus on the economic or social yield [33].

2.2. Data Source of Industrial Clusters Research

The data source is the key to the accurate data-driven industrial clusters identification process. Based on the data access channels, current research on the cluster identification field mainly discusses economic, social, and scientific data. The collection of data is typically sourced from government-published yearbooks. Such data are authoritative and offer comprehensive coverage of the performance of specific industries or regions, albeit at a macro-level, which limits its ability to facilitate the detailed analysis of individual industries. Moreover, the dataset frequently pertains to a restricted number of industries, without including the cluster status of most industries.

Social data mainly include social data [34], logistics data [35] and Internet of Things data [36]. Similar to economic data, social data are also often collected from governmental reports from the internet enterprise sector from a general level. Hence, some scholars also use questionnaire survey data to compensate for the disadvantages. Nonetheless, the data collection cost is considerable, but the data size may be comparatively insignificant, thereby rendering it improbable to comprehensively reflect the conditions of the entire industry. In contrast, scientific data mainly include patent data [37] and science data [38]. Current research on scientific data tend to rely on statistical analyses of the quantity of such data within a region or industry as the basis for assessing the innovative or creative potential of the cluster. As patent data are a crucial instrument for protecting and recoding the technological development of various industries, scientific data offer a more precise depiction of the level of technological development, particularly in high-tech industries, when compared to economic or social data [39]. However, current research places greater emphasis on the abundance of literary data, rather than the content of the literature itself. Therefore, a more detailed examination of the subjects encompassed within technology is conducted to achieve accurate identification.

As is well known, unlike other industries, the high-tech industry relies on patent protection [40]. Therefore, the development of IIC is often accompanied by the emergence of a large number of patents [41]. In this context, this article uses patent big data to conduct research on the identification of IIC.

In response to the significant challenges in less-documented technological areas or countries, we are concerned that the World Intellectual Property Organization (WIPO) has so far had a total of 193 member countries, all of which have established patent systems, including major innovative countries, which are most likely to have IIC. As for countries in which a patent system is not well constructed, they are reasonably likely to have a lack of IIC and are thus out of the scope of our study. In addition, the patent literature covers almost all technical areas of human production activities. For example, the technologies covered by patents have a unified international patent classification, which are divided into 8 divisions, 20 sub-divisions, 118 categories, 620 sub-categories, 6871 main groups and 57,320 sub-groups, showing us a very wide range of technical areas.

2.3. Identification Algorithms

The process of identifying industrial clusters involves collecting data and clustering them according to geographical or technological factors. Such data are typically characterized by a large volume, multiple dimensions, and a low density, requiring accurate and reliable algorithms to effectively classify it. According to the indicators considered, the current commonly used clustering algorithms mainly include Local Moran's I (LMI) [42], K-Means (KM) [43], Spectral Cluster (SC) [44], Hierarchical Cluster (HC) [45] and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [46].

The LMI algorithm is utilized to determine the existence of industrial clusters by examining the performance of indicators in a specific target area and its geographical adjacent areas. This algorithm assesses whether these indicators surpass the overall average value and subsequently verifies the forecasted results with a *p*-value and Z-score. LMI comprises two subgroups: bivariate [47] and multivariable local Moran index [48]. These two subgroups calculate the performance of two or more indicators within a region. By using this method, it is possible to geographically analyze the agglomeration of high or low indicators simultaneously. However, fundamentally, this approach involves separately calculating the values of the elements within the cluster and gathering individuals who meet all the conditions at the same time. Hence, this method cuts off the relationship between different elements and fails to comprehensively analyze the overall indicator performance of the cluster from a macro perspective [49]. Other methods that share similar limitations include the location entropy method [50] and principal component analysis method [51], both of which will not be included in this research.

The KM algorithm is an iterative and unsupervised clustering analysis algorithm and begins by randomly selecting k cluster centers and calculating the distance between the sample data nodes and each cluster center. A threshold value for cluster classification is then applied, and the algorithm proceeds to calculate the average value of each cluster datum to obtain a new cluster center. The process is repeated through several iterations to guide the cluster towards convergence and complete the clustering analysis. The KM algorithm has strong readability and quick iteration. However, there are several concerns regarding its application in industrial clusters identification. First, the number of clusters and threshold must be predetermined, which is challenging, as it is difficult to accurately determine the required number of groups in advance. Second, the KM algorithm primarily focuses on the spatial distance for clustering nodes in the research of industrial clusters and often neglects the clustering of industrial aspects. Lastly, the KM algorithm can only be used for cluster samples with regular shapes. In industrial clusters, clustering may encompass multiple regions and irregular shapes. Thus, the KM algorithm is unable to meet all requirements.

The SC algorithm is a node clustering method based on graphic information. It needs to preset the number of clusters to be divided, construct the adjacency matrix of different nodes, use the distance algorithm to calculate the connection strength of different nodes, and cut the graph to form multiple subgraphs. The cutting standard is that the distance between nodes in the subgraph is the smallest and the distance between nodes in the subgraph is the largest. Although the SC algorithm can process and obtain clusters of different shapes, it can be learned from the clustering principle that the SC algorithm and the KM algorithm both need to be presented with the number of clusters in advance and set the minimum number of clusters, which obviously cannot meet the requirements of the actual spatial cluster clustering [52].

The HC algorithm is divided into top-down and bottom-up according to the clustering method. Bottom-up is to first classify the original data nodes, calculate the clusters of different categories of data nodes and select the nearest category to merge, and then form a new classification layer and form the lowest number of categories by clustering layer by layer. The top-down approach is to first merge all nodes into one category and then gradually divide them into different categories until the clustering effect is below a pre-required threshold. Like other algorithms, the HC algorithm is also based on similarity, but this multilayer clustering approach is computationally complex and susceptible to individual singular values, which is clearly not sufficient for the study of industrial clusters driven by big data [53].

The DBSCAN algorithm is a conventional density-based unsupervised node clustering technique. Its methodology involves the progressive formation of new clusters by evaluating connectable samples while also considering the density of samples as a constraint. Compared to the KM algorithm, the DBSCAN algorithm eliminates the need for the manual determination of cluster quantity and can identify irregularly shaped clusters [54]. Industrial cluster research aims to cluster spatially dispersed nodes to identify clusters with improved indicators. The DBSCAN algorithm surpasses other algorithms in terms of algorithm principles, ease of use, and ability to visualize clustering effects. However, the research of the DBSCAN algorithm in industrial clusters often only considers the spatial distance and the number of single nodes. This limited approach fails to consider other important indicators of overall performance, which is insufficient for identifying multi-technology industrial clusters. While some scholars have proposed using the HDBSCAN algorithm to address the consistency of different cluster densities, this algorithm still lacks the ability to implement the multivariate input function [55].

3. Research Framework and Methodology

This research contains four phases, and the research framework is shown in Figure 1. The first phase involves processing industry patent information, which includes searching for the patent and downloading technological and geographical information, such as titles, abstracts, applicants, and addresses. The next phase involves mining technological information, which includes topic acquirement based on the LDA algorithm and technology-applicant association based on co-occurrence analysis. The following phase is about the identification and evaluation of the IIC cluster from the perspectives of technology and geography. And the final phase is visualizing and analyzing the result of the clusters with a complex network.

3.1. Patent Geographic Information Mining Based on Applicants and API Map

In Phase 1, the patents search string is formulated based on the customers' requirements and industrial characteristics. Those keywords include patent application regions, keywords, classification numbers, and the starting and ending date of the patent application. After collecting patents, we employed a process of consolidating duplicate or repeatedly submitted patents to ensure that they are not counted multiple times. For example, there will be a new patent filed for both the invention and its utility model in China. Meanwhile, there will also be authorized patents for inventions and open patents for inventions. In this case, there would be separate patents in examination and publication, even though both of these patents refer to the same technological solution. Herein, we then proceed to acquire the patent in detail: information regarding the applicant, address, abstract, claims, and other relevant textual data.

The patent contains both the name and address of the applicant. Due to name similarity and multi-location application issues of patent applicants, as well as the applicant's address showing only the approximate location, it is impossible to determine the geographic location solely on patent application information accurately. With the existing open online

API map system, the geographic distribution of applicants can be accurately estimated by combining both applicants' names and addresses. By adopting this approach, we can overcome the limitations of the existing system and accurately obtain the longitude and latitude information of the applicant.

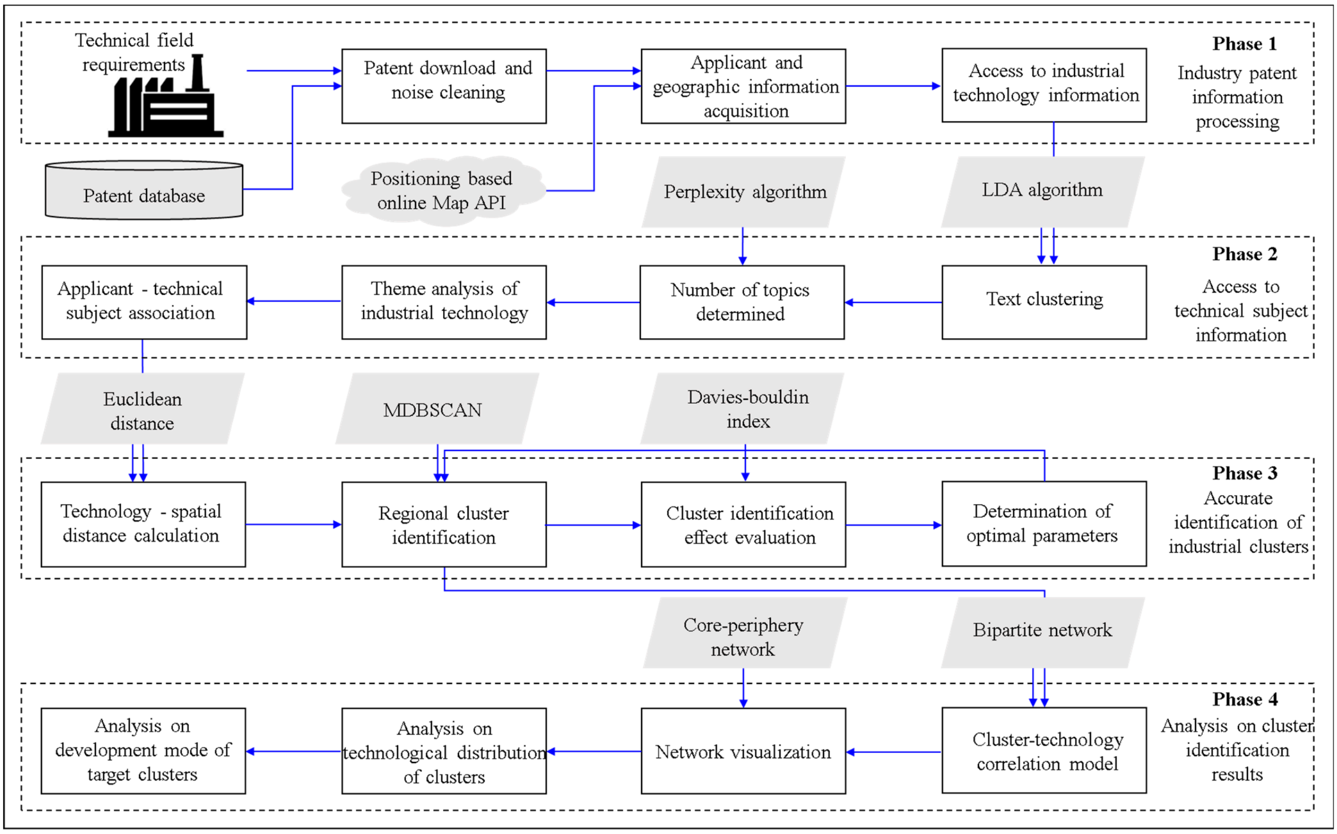

Technical field requirements
Patent download and noise cleaning
Applicant and geographic information acquisition
Access to industrial technology information
Phase 1
Industry patent information processing
Patent database
Positioning based online Map API
Perplexity algorithm
LDA algorithm
Applicant - technical subject association
Theme analysis of industrial technology
Number of topics determined
Text clustering
Phase 2
Access to technical subject information
Euclidean distance
MDBSCAN
Davies-bouldin index
Technology - spatial distance calculation
Regional cluster identification
Cluster identification effect evaluation
Determination of optimal parameters
Phase 3
Accurate identification of industrial clusters
Core-periphery network
Bipartite network
Analysis on development mode of target clusters
Analysis on technological distribution of clusters
Network visualization
Cluster-technology correlation model
Phase 4
Analysis on cluster identification results

Figure 1. Research framework.

To further improve the accuracy of obtained information of subsequent technological subject extraction, we conducted a manual review of all the patents' abstracts and claims. This allowed us to create a stop word list and dictionary by eliminating common or insignificant words, symbols, and numbers while replacing certain synonyms.

3.2. Patent Technological Subject Acquisition Based on the LDA Model

In order to extract technological topics from a vast amount of patent text data, the classification of all patents' technological texts is performed by utilizing the LDA topic analysis algorithm. The LDA algorithm is a text topic generation model and an unsupervised machine learning algorithm [56]. It adopts a multilayer probabilistic model with a three-layer structure comprising words, documents, and topics. LDA assumes that words are generated from a mixture of topics, and each topic is generated by a polynomial distribution over a fixed word list. These topics are shared across all documents in the dataset. The generation process of the LDA topic model is depicted in Figure 2.

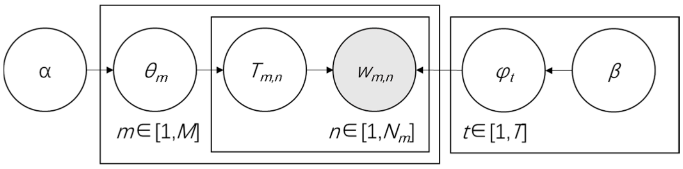


Figure 2. LDA model.

Suppose a collection of patent documents, $D = \{d_1, d_2, \ldots, d_{\mathrm{M}}\}$, N_m represents the number of words in the mth document, t denotes the number of topics, $T_{m,n}$ is the nth topic in the mth patent document, and $w_{m,n}$ is the nth word in the mth document. Here, φ_k is a parameter obeying a Dirichlet distribution with β, and θ_m is a parameter obeying a Dirichlet distribution with α. Given a set of one document, $w_{m,n}$ are the observed known variables, β and α are the a priori parameters given based on manual experience, and $T_{m,n}$, φ_k, and θ_m are unknown implicit variables. Two main parameter inference methods, variational Bayesian inference and Gibbs sampling, are used for the subject model [57]. The algorithm of variational Bayesian inference is characterized by its myopic nature, as it approximates the expressions of posterior probabilities for model parameters using a straightforward variational distribution. Through the iterative application of the EM algorithm, it aims to maximize the variational lower bound to estimate the parameters. On the other hand, Gibbs sampling is a stochastic algorithm that relies on samples drawn from a Markov chain. In comparison to variational Bayesian inference, the Gibbs sampling algorithm is considered simpler and more user-friendly. As a result, LDA topic models often favor the adoption of Gibbs sampling.

It should be noted that the superiority of the LDA topic representation results is significantly influenced by the number of different topics, which must be predetermined using the LDA algorithm [58]. Currently, various metrics exist to determine the number of topics, including coherence and perplexity. Coherence has high computational complexity and is suitable on large-scale datasets. By contrast, perplexity has a higher recall rate and efficiency, which is more prominent in long texts such as patent technology texts and is a common determination index [59]. Therefore, the perplexity index is introduced to measure the superiority of the subject modeling results, enabling the assessment of the model's predictive ability for uncertain data. The formula for calculating perplexity is as follows:

$$Perplexity(D) = \exp\left\{ -\frac{\sum\limits_{m=1}^{M} \log p(w_m)}{\sum\limits_{m=1}^{M} N_m} \right\} \tag{1}$$

where D denotes the test set in the corpus, $\sum_{m-1}^{M} N_m$ denotes the number of words in the entire test dataset, and $p(w_m)$ refers to the probability of occurrence of word w_m in the test set. It can be seen that the confusion formula is primarily based on the information entropy, and the entropy obtained from the number of topics is evaluated by calculating the probability of word occurrence across different topics. A lower perplexity indicates a more effective outcome in terms of topic clustering.

While LDA can provide a substantial number of technological topics, there are often duplicates among these topics [59]. To address this, manual detection is employed to identify and merge similar topics, followed by the individual marking of technological classifications for each patent. Since the number of topics is usually not extensive, the manual approach requires less time and yields more accurate results.

3.3. Accurate Identification of Industrial Clusters Based on MDBSCAN

Patents involve the latest and most active technical information in almost all related technical fields. Effectively analyzing and utilizing the complex correlations between patents can not only provide direction for future research but also help organizations such as enterprises and institutions to identify the best partners. At present, academics generally believe that the relationship between patents is mainly divided into four kinds: competing, blocking, complementary, and unrelated [60]. Among them, the first two relationships mainly reflect the similarity between patents when conducting patent relevance analysis. The research on patent technology complementarity appears relatively weak, and few studies at home and abroad have synthesized the theory, method, and application related to patent technology complementarity. In fact, major technological innovation generally relies on the introduction of complementary technologies, and through technological cooperation between enterprises or organizations with complementary patents, they can absorb each other's advantageous technologies to make up for their own technological deficiencies while maintaining their own core technological strengths so as to improve the success rate of innovation, reduce the costs and risks of technological innovation, and obtain a more desirable investment value. In order to determine the complementarity between applicants within a cluster, we propose the following definition of technical distance: when two patents belong to two different subfields in the same technical field, it means that they belong to the same field in terms of technical content and can realize complementary advantages through their differences.

Scholars have emphasized that spatial concentration is a key feature of industrial clusters [61,62]. In essence, clusters reflect proximity, i.e., the spatial concentration of firms from a micro-geographic perspective. In order to identify the clusters, it is necessary to consider both spatial and technological distances based on the calculation needs of the applicant. An algorithm compatible with both distances needs to be introduced due to the variations in how different distances are represented. One such similarity calculation method is the Euclidean distance, which calculates similarity by cumulatively calculating the difference between different variables of the nodes [63]. Compared with similarity calculation methods such as cosine similarity and Jaccard similarity, the Euclidean distance is well suited for numerical similarity calculation. It imposes no constraints on the value range, with larger values indicating greater distances. Moreover, it provides an absolute distance measure that effectively captures the differences between different variables of the nodes, making it suitable for both spatial and technological distance calculations. If there exists a set of variables $x = \{x_1, x_2, \ldots, x_\mathrm{n}\}$ and $y = \{y_1, y_2, \ldots, y_\mathrm{n}\}$, the Euclidean distance of the two variables is computed as follows:

$$d_{ij} = \sqrt{\sum_{n=1}^{N} (x_{in} - x_{jn})^2} \tag{2}$$

The calculation of the distance between various applicants can establish a geographical foundation for subsequent cluster identification. Currently, the methods of calculating spatial distance are divided into linear distance and spherical distance [64].

$$sd_{ij} = d_s \times \sqrt{(ln_i - ln_j)^2 + (la_i - la_j)^2} \tag{3}$$

where d_s refers to the plane distance of the unit longitude and latitude and is determined uniformly based on the node's location. The spatial distance increases as the difference between the longitude and latitude of different nodes becomes greater, as indicated by Formula (3). This relationship aligns with real-world geographical scenarios.

To calculate the technological distance between different applicants who are involved in multiple technologies simultaneously, a combination of the Euclidean distance method and the bipartite network algorithm can be employed. The technological distance is computed as follows:

$$td_{ij} = \sqrt{\sum_{t=1}^{T} (e_{it} - e_{jt})^2} \tag{4}$$

where td_{ij} represents the technological distance between the applicant nodes e_i and e_j, and t represents the presence or absence of technology in the same field identified by the LDA method, with applicants having two or more patents in technology t recorded as 1 and those with none recorded as 0.

The process of identifying industrial clusters involves clustering and evaluating clusters, and it directly impacts relevant industrial policies and the economic interests of those clusters. To minimize artificial intervention, this paper utilizes an unsupervised identification method based on machine learning. The DBSCAN algorithm, a classical density-based clustering algorithm, is employed. Unlike the KM algorithm, DBSCAN has the advantage of automatically determining the number of clusters and identifying clusters with arbitrary shapes. This characteristic makes it well suited for clustering nodes located at various positions within a spatial network [65]. In addition to massive data, DBSCAN can find outliers while clustering and is insensitive to outliers in the dataset. However, the traditional DBSCADN only considers the maximum distance sd_{min} between nodes within a cluster and the minimum number of nodes x_{min} as clustering constraints, i.e., it requires the need to satisfy $((td \leq td_{min}) \cap (x \leq x_{min}))$ among cluster members, without considering the complementarity or dissimilarity between technologies within a cluster. Consequently, this limitation fails to meet the requirements of actual industrial clusters.

To address this issue, we propose an enhancement to the existing DBSCAN algorithm, with the name MDBSCAN by adding the minimum technology distance variate td_{min}. This enhancement ensures that during the clustering process, cluster members must satisfy the condition $((td \leq td_{min}) \cap (x \leq x_{min}) \cap (sd \leq sd_{min}))$. In this regard, the MDBSCAN algorithm is employed in this study to facilitate the fusion of multiple variables. By doing so, it enables the identification of clusters with both a high spatial density and significant technological complementarity. The flow algorithm of the whole algorithm and its pseudo-code can be found in Appendix A. It is worth stating that various studies have employed KM, HC, and SC methods to combine multiple variables during the clustering process. However, the underlying principle primarily involves computing the Euclidean distances of multiple variables concurrently during the initial clustering stage; these distances are then consolidated into a unified distance value using weights. Subsequently, different elements are clustered based on a distance threshold [66–68]. Since different weights can have a significant impact on the final clustering effect, the determination of the weights often relies on manual assignment, making such algorithms susceptible to subjective experiences. In contrast, MDBSCAN uses technological distance and spatial distance to filter and classify elements separately, thereby avoiding these issues.

Industrial clusters are formed by utilizing big data and require the evaluation of their effectiveness with different combinations of input variables. The evaluation methods for clusters can be mainly classified into internal and external evaluation techniques [69]. Internal evaluation involves calculating metrics that measure the intra-cluster and inter-cluster elements, supplying an assessment of clustering effectiveness without the use of real labels. This method is applicable for evaluating clusters without real labels. On the other hand, external evaluation is conducted using real labels, where the clustering results are compared against the actual labels to determine the evaluation outcome. In the process of industrial clusters evaluation, true labels are not usually available; thus, the internal evaluation methods are generally adopted.

There are three commonly used internal evaluation methods, namely, the Silhouette Coefficient Index (SCI) [70], Calinski–Harbasz Score (CHS) [71], and Davies–Bouldin Index (DBI) [72]. The SCI algorithm yields evaluation scores ranging from −1 to 1. The CHS provides the fastest clustering evaluation but is primarily applicable to the evaluation of clusters of spherical data. However, it is less accurate compared to SC and DBI when evaluating the clustering results obtained based on the density algorithm [73]. The DBI integrates the intra-class sample similarity and inter-class sample difference, resulting in superior efficiency and accuracy. The evaluation score for the DBI value is unbounded and can range from zero to any positive number. A higher DBI value indicates a better clustering effect, and it is also suitable for joint evaluation of multiple indicators. The following formula stands for the DBI, assuming that the applicant dataset A is divided into k clusters:

$$DBI = \frac{1}{C}\sum_{i=1}^{C} \max_{i \neq j}\left(\frac{\overline{S_i} + \overline{S_j}}{||w_i - w_j||_2}\right) \quad (5)$$

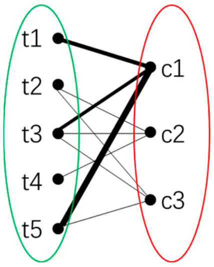
where $\overline{S_i}$ represents the average distance from all internal elements of the ith cluster to the center of the cluster, which also indicates the degree of dispersion of sample data within the cluster. $||w_i - w_j||_2$ represents the cluster center distance from the ith to jth clusters. The formula shows that the smaller the DBI index is, the smaller the distance within the cluster, the larger the inter-cluster clustering, and the better the clustering effect.

In order to calculate the applicant's final clustering index, it is essential to consider both spatial and technological distances with two DBI indices, namely, the spatial distance clustering index DBI^{sd} and technological distance clustering index DBI^{td}. Normalization is required to mitigate the impact of varying ranges of values on joint settlement results and prevent substantial differences between values. To achieve this, the chosen approach involves the utilization of a logarithmic function conversion method. Consequently, the overall calculation process for the complete Dissimilarity-Based Index can be outlined as follows:

$$DBI = \frac{1}{2}\{\log[DBI^{td}] + \log[DBI^{sd}]\} \tag{6}$$

3.4. Analysis of Cluster Identification Results with Bipartite Network

Various clusters are obtained through MDBSCAN and DBI. To further analyze the relationship between clusters and technological topics, a bipartite network is introduced to represent their relationship. Unlike an ordinary network, a bipartite network can be compatible with nodes of both natures, while there is no association between nodes of the same type. Moreover, a bipartite network exhibits interconnected edges solely between nodes of distinct types, where the strength of the association between a pair of nodes is denoted by the weight of the connected edge [74]. The characteristic form of a bipartite network is shown in Figure 3.



	c1	c2	c3
t1	e_{11}	0	0
t2	0	e_{22}	e_{32}
t3	e_{13}	e_{23}	e_{33}
t4	0	e_{24}	0
t5	e_{15}	0	e_{35}

Figure 3. Bipartite network.

The existence of a weighted technology-applicant bipartite network $G = (C, T, E)$ is assumed, where $C = \{c_1, c_2, \ldots, c_C\}$ represents the cluster set, $T = \{t_1, t_2, \ldots, t_T\}$ represents the technology topics set, and $E \in (C \times T)$ is the edge set of G. The weights are calculated as follows:

$$e_{ij} = \begin{cases} pn_{ij}, e \in (c_i \times t_j) \\ \;0 \quad, e \notin (c_i \times t_j) \end{cases} \tag{7}$$

where pn_{ij} denotes the number of patents involving both the cluster c_i and technology t_j.

The clustering capabilities of MDPSCAN encompass both geographical and technological aspects; however, the resulting clusters can only be represented on a two-dimensional map, leaving the relationship between clusters and different technologies undisclosed. In order to improve the visualization of the cluster and technology network model, a network layout configuration is required. Specifically, a force-guided layout approach is employed for the cluster–technology relationship network, wherein the merging of nodes within clusters aims to reduce the total number of network nodes. This layout method employs the concept of spring force to calculate pairwise forces between nodes, thus centering the important nodes within the network. Consequently, this arrangement allows for a more visually coherent display of the significant clusters [75]. Based on this, different clusters are divided with Core–Periphery Network (CPN) structure facilitation. Within the CPN framework, the core denotes a set of nodes located in the center and tightly intricately connected, whereas the periphery refers to a set of nodes encircling the center and loosely connected to the core. With regard to the IIC, the utilization of CPN enables the precise analysis of the technology distribution of the target cluster and its position within the overall cluster.

In addition to the technology distribution, the cluster development model can also be analyzed. According to Markuse [76], the current cluster development models encompass four main types, namely, Italian-style industrial clusters, satellite-based industrial clusters, axle-wheel industrial clusters, and national power-dependent industrial clusters. Italian-style clusters are dominated by SMEs, exhibiting strong specialization without leading enterprises. Satellite clusters are also dominated by SMEs; however, their existence relies mainly on enterprises affiliated with other clusters. Axis-wheel-type clusters are mostly dominated by both large-scale local enterprises and SMEs, exhibiting a distinct hierarchical structure. State power-dependent industrial clusters, also referred to as government-led industrial clusters, materialize as a result of the state's support via industrial policies.

Based on the definitions of the four patterns, this paper utilizes a patent analysis method to identify cluster patterns. The patent applicants encompass various entities, including enterprises, universities, and research institutes. Notably, universities and research institutes often serve as state-owned institutions in numerous countries, such as China. At the same time, the number of patents contained in different institutions within a cluster exhibits considerable variation, thereby facilitating the identification of Italian-style industrial clusters and axle-wheel-style industrial clusters. In addition, there are differences in the spatial distance of different clusters; for example, a few clusters are close to large clusters and belong to the typical satellite development pattern.

4. Empirical Research

4.1. Data Collection

Flexible electronics (FE), also known as flex circuits, is technology for assembling electronic circuits by mounting electronic devices on flexible substrates. FE can replace multiple rigid boards or connectors and is ideal for dynamic or high-flex applications, and it has attracted thousands of enterprises and institutions, forming a complete industry, FEI, in China. More than this, FEI is a typical high-tech industry, which possesses the characteristics of many participating enterprises, many patents, and complex industrial chains. Taking this industry as the research case, it can provide effective methods and tools for identifying other high-tech industry clusters.

Therefore, we analyze from the perspective of the flexible electronics industry applied in China since 2000 so as to better grasp the pattern of technological innovation in the IC industry. The search terms utilized in this study are as follows: TA_ALL: ("Flexible Electronic" OR "Flexible Printed Circuit" OR "flex circuits" OR "Flexible Hybrid Electronics" OR "flexible displays" OR "Flexible Printed Sensor" OR "stretchable electronic") OR CPC: G09F9/301 OR H01L51/0097 OR G06F1/1652 OR H01L2251/5338 OR H04M1/0268 OR G09G2380/02 OR G05B2219/25321 OR G05B2219/25439 OR H01H2229/038 OR H01R12/59 OR H01R12/61 OR H01R12/77 OR H01R12/78 OR H05K1/028 OR H05K1/118 OR H05K1/147 OR H05K3/361 OR H05K2201/2009 OR H05K2201/046 OR H05K2201/2027 OR H05K3/4635 OR H05K2201/09445 OR G09G3/035 OR G02F1/133305 OR G06F2203/04102 OR G06F1/1616 OR H01F2017/006 OR H01H2001/5816 OR H01H2001/5827 OR H01G9/2095 OR H01 L23/4985 OR G05B2219/23358 OR H05K2201/05 OR H05K1/148) AND APD:[20000101 TO 20201231]. A total of 860,701 Chinese patents were obtained, and the specific information is shown in Table 1. The provided information includes various elements, such as the title, application number, filing date, applicant, applicant's address, and abstract of the patents. For some of the patents jointly applied by different applicants, the patents are divided into two distinct applications. Additionally, irrelevant terms within the abstract are eliminated using the HIT stop word list to enhance the effectiveness of patent subject extraction.

Table 1. Information of patent applicants.

Application No.	Patent Title	Applicant	Applicant Address	Longitude and Latitude
CN202011550473.0	A method of making a conductive circuit board	Shenzhen Bairou New Material Technology Co.	No.8 Baoqing Road, Baolong Community, Baolong Street, Longgang District, Shenzhen, Guangdong	118.09644, 24.48541
CN202010943557.4	The production method of circuit board with embedded conductive lines	Pengding Holdings (Shenzhen) Co.	Building A1 to Building A3, Peng Ding Park, Song Luo Road, Yan Luo Community, Yan Luo Street, Baoan District, Shenzhen City, Guangdong	113.86367, 22.79640
CN202010195388.0	A circuit board and its manufacturing method	Yancheng Wixin Co.	No.999, Yandu Road, Yandu District, Yancheng City, Jiangsu	120.18987, 33.34369
…	…	…	…	
CN201410149990.5	Preparation of copper-zinc-tin-sulfur films on flexible substrates using magnetron sputtering	Guangdong University of Technology	No.100 Waihuan West Road, Guangzhou University City, Panyu District, Guangzhou City, Guangdong	113.39960, 23.04570

To ensure the accuracy of the analysis results and mitigate the influence of individual applicants who filed only one patent, a criterion is set to analyze applicants with two or more applications. Consequently, a total of 5610 applicants who had submitted multiple patents in the field of flexible electronics were selected for the analysis. This was combined with Baidu map API (https://lbsyun.baidu.com/, accessed on 23 October 2022) to obtain the latitude and longitude information of the applicants. Simultaneously, the correlation information between the applicants and technologies was constructed, as shown in Table 1.

4.2. Obtaining Technology Topic and Keywords

The application of LDA enables the extraction of topics from patent abstracts, referring to the related literature [58,77,78]. Additionally, the number of iterations for Gibbs sampling is set to 1000, with K representing the number of topics present in the corpus. Based on Formula (1), we calculate the perplexity values for all numbers of topics from the range (0.149) and found that the lowest value is 194.8483 when the number of topics is 94, as shown in Figure 4.

A set of 20 technological topic keywords is established, with the top-weighted keywords being selected as the display objects. At the same time, experts' experience is utilized to interpret and filter various combinations of technological keywords, resulting in the merging of duplicate topics. Finally, the process resulted in 95 topics in total, as shown in Appendix B.

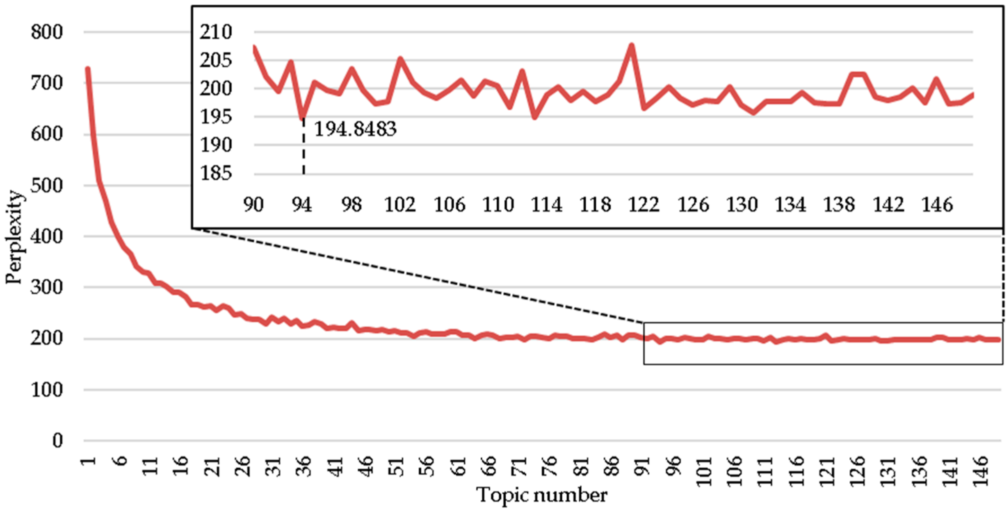


Figure 4. Perplexity calculation results.

4.3. FEI Clusters

Referring to the criteria for the identification of industrial clusters of SMEs in China [13] and existing literature [79–81], the range of the distance for the applicants is set between 25 km and 250 km, with 25 km being one unit of measurement, and the range for the number of patent application enterprises is set between 10 and 100, with 5 as a measurement unit; the range of technological distance is set between 20 and 1200, with 20 as a measurement unit. Through the partial DBI results, the index floor of the joint DBI is set to 10. By employing the python program to develop the MDBSCAN algorithm and incorporating a matplotlib plug-in to display the algorithm results, the calculation results are shown in Figure 5. In this figure, the darker color of the ball signifies a lower DBI value, indicating a superior clustering effect, whereas a lighter color corresponds to a higher DBI value. Through the joint analysis of the three input indicators, it can be found that some of the indicators within the value range are significantly better than those in other regions. For the number of patent application enterprises, when the number of applicants within x_{min} = (5, 10), the DBI index will be significantly lower than the indexes in other intervals. In terms of spatial distance, the DBI indicator is significantly lower than that in the other zones when the spatial distance is equal to 25, 200, and 225 values. For example, when the number of applicants is less than 20 and the spatial distance is equal to 25 and 225, an interesting observation emerges: as the technological distance tends to approach 1200, the DBI decreases. Conversely, in other indicators, there seems to be no significant disparity in the technological distance; in fact, larger distances correspond to larger indicators. This example underscores the notable influence of the technological distance, spatial distance, and number of participating enterprises on the cluster evaluation effect yielded by the MDBSCAN algorithm. If only the number of participating enterprises and spatial distance are employed as constraints, it becomes apparent that the approach does not align with the actual situation.

The selection of clusters considered the number of clusters in the evaluation of DBI. To further obtain the optimal parameters in this case, considering the range of the obtained DBI values (0.397533376, 1.222981034) and the number of distributions, combinations lower than 0.45 are selected as secondary screening objects, and a total of 13 groups are obtained, as shown in Table 2. In this table, the ranks of 1, 2, and 3 variable combinations have low DBI values, but the numbers of clusters are all small, making the granularity of cluster division too large, resulting in too many participating enterprises within the cluster, which is not conducive to the implementation of cluster policy. sd_{min} = 25, td_{min} = 350, and x_{min} = 10 are selected as cluster identification indicators, and a total of 44 clusters are obtained.

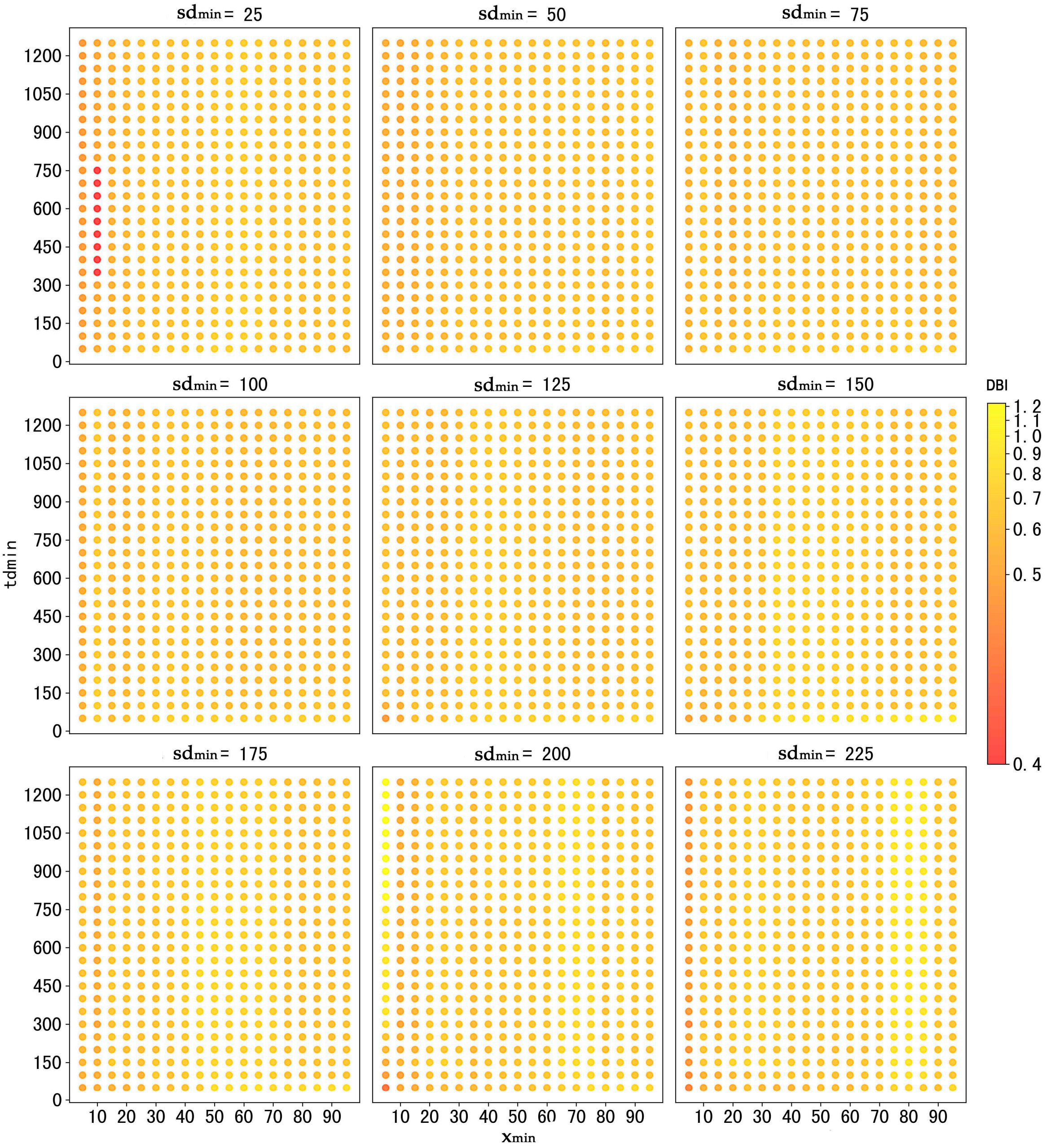


Figure 5. Multivariate oriented evaluation effect of FEI clusters.

The application of the MDBSCAN algorithm under specific constraints of a minimum technological distance, spatial distance, and number of nodes results in the identification of a total of 44 clusters. These clusters are then assigned numerical labels based on their cluster sizes, as illustrated in Figure 6. The dominant clusters in China's flexible electronics industry are primarily concentrated in three key regions: the Yangtze River Delta (c1), the Pearl River Delta (c2), and Beijing (c3). These clusters not only have the characteristics of many participating enterprises but also tend to span multiple administrative regions. For example, the flexible electronics industry cluster in Shenzhen extends to the adjacent Guangzhou and Dongguan, and the industry cluster in Shanghai extends to its adjacent Suzhou. In order to validate the presence of cross-regional industry clustering and facilitate the effective enhancement of overall cluster competitiveness, it is imperative for multiple administrative units to collaborate during the implementation of industry cluster support. Moreover, these clusters demonstrate the occurrence of cross-regional clustering, alongside irregular shapes. For example, the Yangtze River Delta shows a V-type cluster, which extends to two adjacent provinces through Shanghai. The Pearl River Delta, on the other hand, is a △-type cluster, with several cities developing in parallel. Beijing represents a cluster of the ∗-type, characterized by a radial shape that extends to various regions within the city, primarily due to its central location. The accurate delineation of these clusters proves challenging when employing traditional clustering models based on an administrative area division or KM algorithm.

Table 2. List of combinations with DBI values below 0.45 and their calculation results.

Rank	sd_{min}	td_{min}	x_{min}	Number of IIC	DBI
1	225	100	5	13	0.442539712
2	225	300	5	7	0.435385726
3	225	50	5	23	0.42753337
4	200	50	5	25	0.412094281
5	25	350	10	44	0.40390312
6	25	400	10	44	0.40390312
7	25	450	10	44	0.40390312
8	25	500	10	44	0.40390312
9	25	550	10	44	0.40390312
10	25	600	10	44	0.40390312
11	25	650	10	44	0.40390312
12	25	700	10	44	0.40390312
13	25	750	10	44	0.40390312

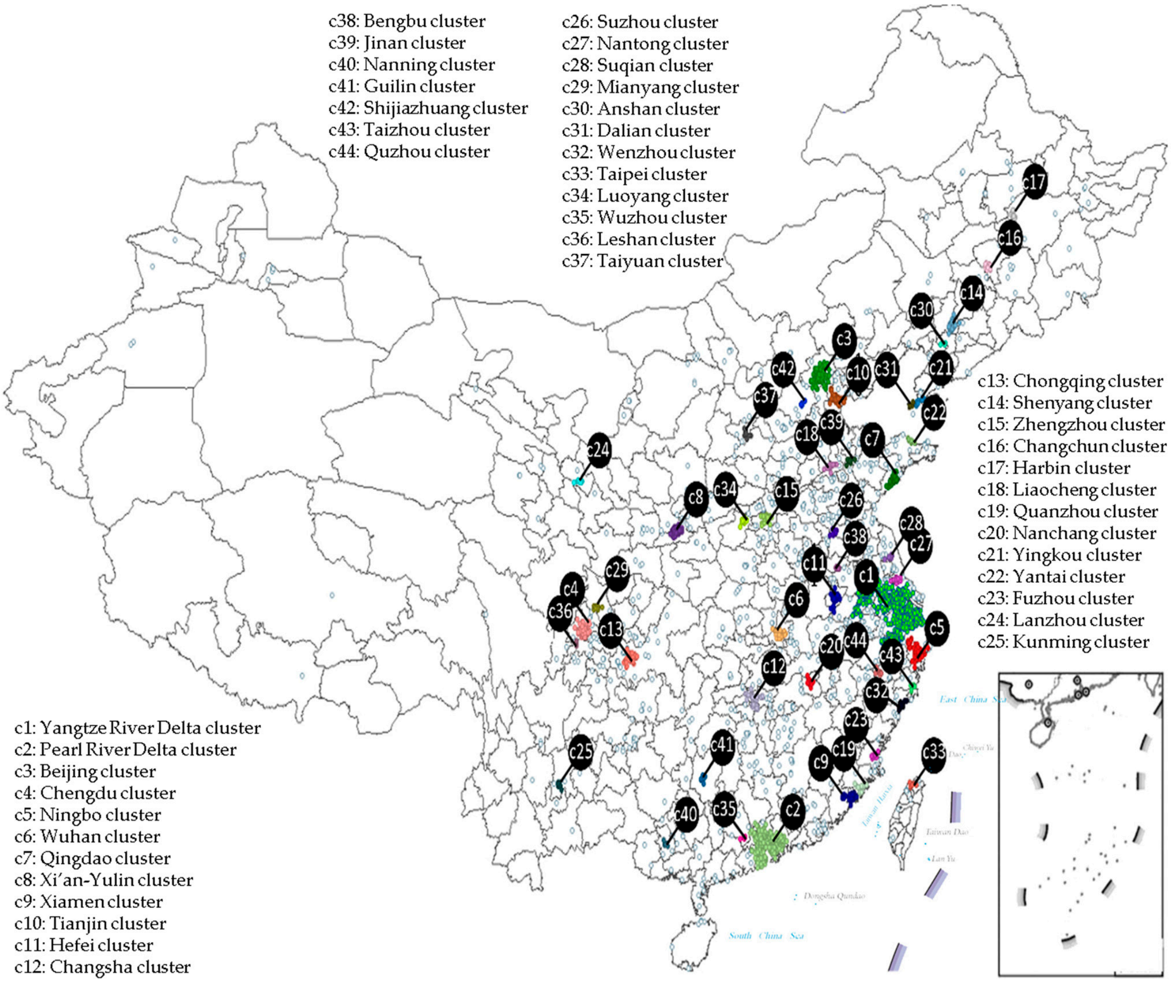


Figure 6. China's FEI clusters identification result.

4.4. Clusters Identification Results Analysis

In Figure 7, the technology distribution of the clusters was further analyzed based on the correlation between the clusters labeled with blue nodes and technology labeled with green nodes. It can be seen that the distribution of different clusters shows a core-periphery network structure. The center is the Yangtze River Delta cluster of c1, the Pearl River Delta cluster of c2, and the Beijing-Tianjin-Hebei cluster of c3. The technology clusters between the abovementioned locations are relatively close to each other, with stronger technological complementarity. In contrast, other clusters tend to be in the peripheral structure and only involve individual technology nodes, mainly because of the smaller technological distance and higher technological overlap among the enterprises within the cluster. For example, the Wuhan cluster, located in central China and designated as Cluster No. 6, primarily revolves around two technologies: organic display (T80) and display (T72). These technologies are characterized as flexible display technologies, suggesting that the cluster predominantly focuses on a single technology and lacks significant industrial competitiveness. By considering the core–periphery network structure, it becomes evident that the MDBSCAN-based algorithm can effectively identify clusters that exhibit greater technological complementarity, a capability that is not achievable with the current DBSCAN algorithm.

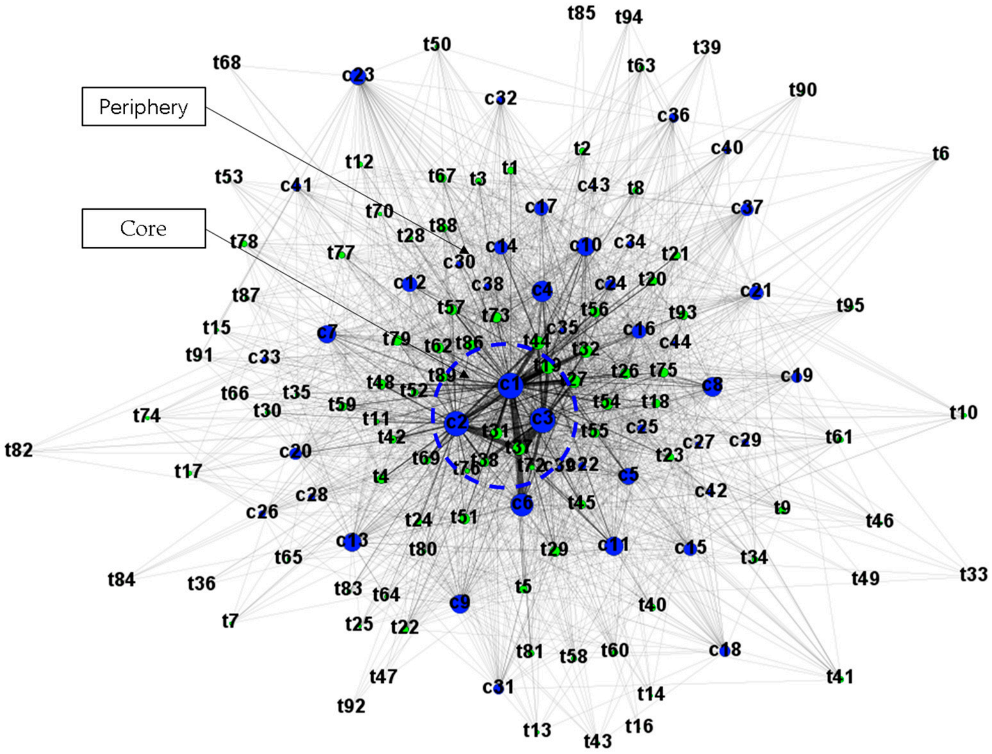
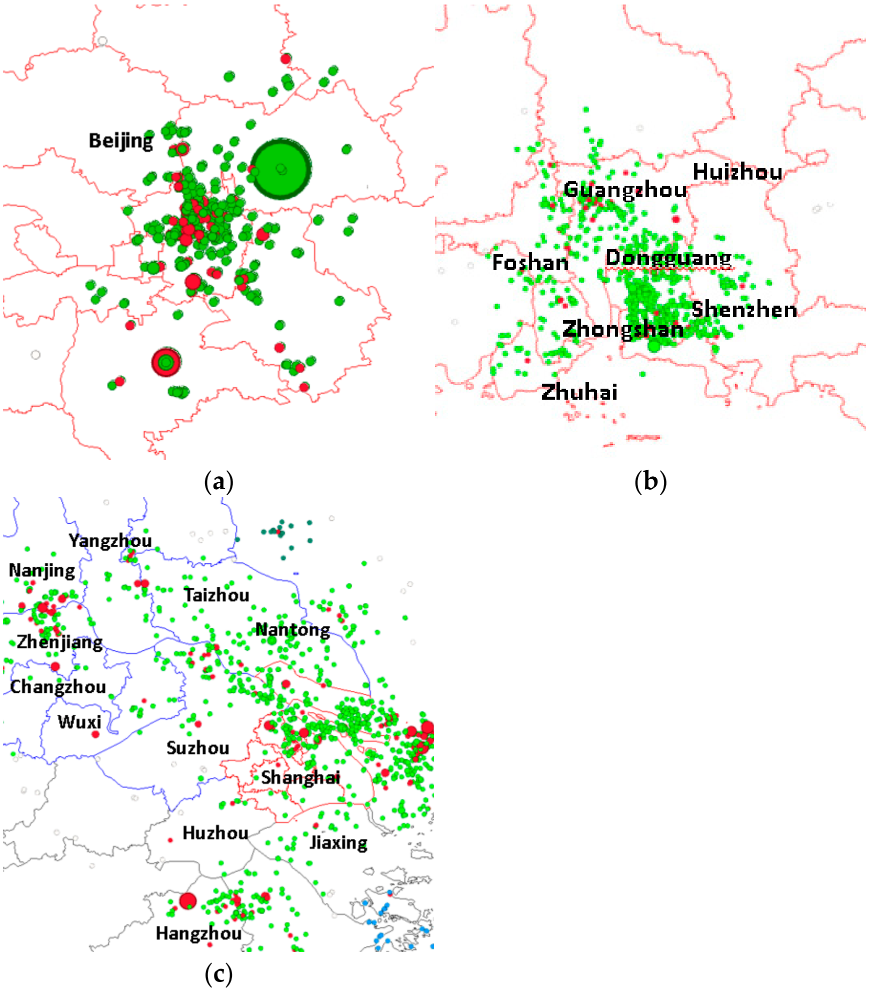


Figure 7. Technology distribution of clusters.

The identification of development patterns in the target clusters can be facilitated by the examination of patent texts alongside econometric data, as demonstrated in Figure 8. By distinguishing the number of patents associated with various applicant nodes within the cluster, it becomes apparent that the Pearl River Delta cluster conforms to the typical Marshall-type industrial cluster model. This model is characterized by a predominance of SMEs, a pronounced specialization in specific fields, a significant complementarity of technological capabilities among enterprises, and a notable level of industrial competitiveness. On the other hand, the Yangtze River Delta cluster exhibits a combination of a government-led industrial cluster model and Marshall-type industrial clusters. This hybrid configuration involves a greater presence of institutions, which are represented by red dots, alongside a higher proportion of SMEs. The Beijing–Tianjin–Hebei cluster is mainly a mixture of the hub-and-spoke industry cluster model and the government-led industry cluster model. Apart from the large number of universities and research institutes in the cluster, there is a clear distinction in the hierarchy due to significant differences in technological expertise among applicants, where the node diameter represents the number of patent applications by the applicants. As typical satellite platform-type industrial clusters (shown in Figure 8), c27, c28, c35, and c42 are established in areas at a certain distance from large clusters, and the spatial distance can ensure the realization of resource sharing while reducing the operating costs. It is also another advantage of industrial cluster identification based on patent big data.



(**a**) (**b**) (**c**)

Figure 8. Identification of development patterns of clusters. (**a**) Beijing; (**b**) Pearl River Delta; (**c**) Yangtze River Delta.

5. Discussion

5.1. Comparison with the Previous Approaches

To further verify the accuracy of the proposed algorithm, a comparison is conducted with four commonly used spatial clustering algorithms to account for the variations in clustering effects across different algorithms and clusters. The joint DBI is utilized for this purpose, and the comparative results are presented in Figure 9. Notably, when the minimum number of applicants is set to 10, DBSCAN exhibits the highest DBI index. The results demonstrate that the MDBSCAN algorithm surpasses the HC and SC algorithms when the number of clusters exceeds 15 and outperforms the KM algorithm when the number of clusters exceeds 19, except for cases where the number of clusters is missing.

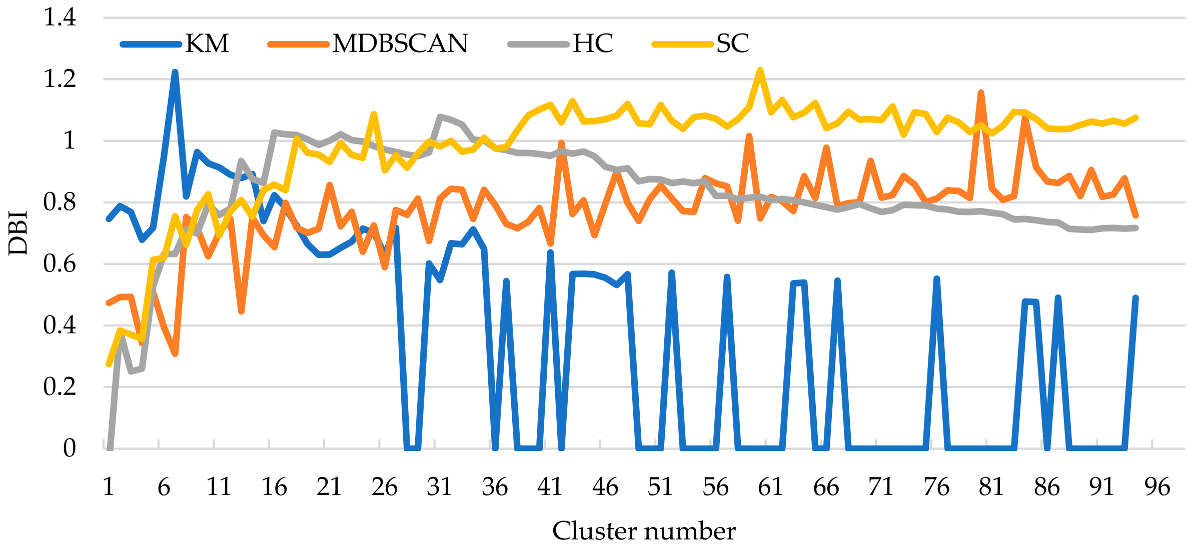


Figure 9. Comparison of clustering effects of different algorithms.

The DBSCAN algorithm has the disadvantage of a long running time when dealing with large amounts of data, and this needs to be considered when adding computational variables. In relation to this matter, the running times of the MDBSCAN and DBSCAN algorithms are computed individually for varying proprietary data volumes and subsequently compared, as depicted in Figure 10. It is evident that for a data volume of 250,000, the run times of both algorithms are roughly equivalent. However, when the data volume reaches 500,000, there is a notable disparity of 264 s in the running time between the two algorithms. Given that MDBSCAN has 96 additional variables compared to DBSCAN in this specific scenario, the disparity in execution time is comparatively reduced. Furthermore, the investigation of clusters is inherently interconnected with the advancement of the entire industry, emphasizing the primacy of accuracy over computation time. Therefore, when comparing the results of the two metrics, the small amount of time added to the computation process based on the MDBSCAN algorithm is acceptable.

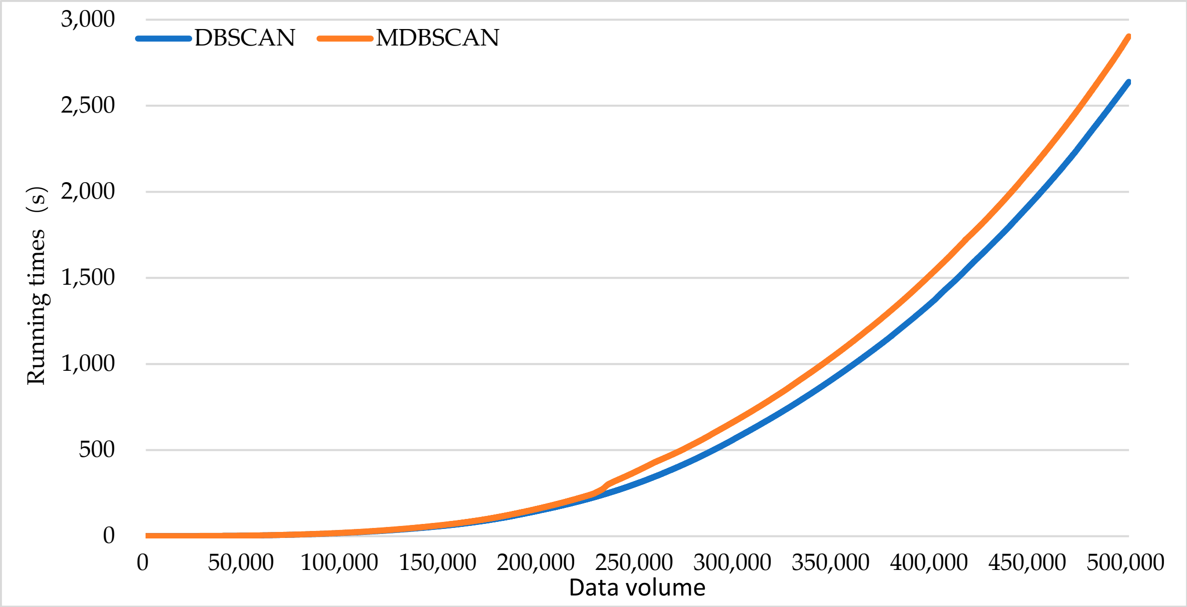


Figure 10. Comparison of algorithm running time.

5.2. Impact on Industry Cluster Research

The proposed method leverages patent data and their associated benefits to facilitate the accurate identification of IIC in terms of technological and spatial dimensions. It operates in an unsupervised manner and effectively determines the boundaries of regional clusters while also providing a systematic analysis of their technological status and patterns. This research serves as a foundation for the subsequent measurement of regional industrial clusters and the formulation of relevant policies by scholars and government authorities.

Government agencies and researchers can use this method to precisely understand the distribution of industrial clusters in various technology industries. Since a patent is the important vehicle for existing enterprises seeking to protect their technology and an indispensable tool for technology enterprises in various industries seeking to achieve market competitiveness, it can be applied to most technology industries with the help of patent big data. Analyzing the development of the industry with the help of patents is the mainstream method of existing technology management. At the same time, with the help of patent applicants and address texts, the identification of clusters can be narrowed down from a national level to a certain region, such as the one shown in Figure 8, and can be focused on the city level or county level, or even smaller, through the map API that can accurately obtain the longitude and latitude positioning of enterprises. This is unmatched by other available data. It should be added that some scholars have also used patent classification numbers to classify technologies [82]. However, for most industries, the technological content that patent classification numbers can provide is still too broad, while lacking annotation for many technologies, such as manufacturing processes and materials in the flexible electronics industry, which do not have corresponding classification numbers and obviously cannot meet the needs of industrial technology mining [83].

From an algorithmic standpoint, the present paper introduces the MDBSCAN algorithm, which effectively clusters spatial groups by incorporating additional indicators. The implemented code demonstrates the algorithm's capability to accommodate multiple indicators for clustering purposes. While the computation time of MDBSCAN may be higher than that of alternative algorithms, it not only preserves the original algorithm's ability to identify intricate contours but also addresses the limitations of indicator scarcity. Furthermore, the algorithm exhibits superior accuracy compared to other approaches, thus offering substantial advantages that outweigh any drawbacks. Consequently, this algorithm represents a valuable contribution, furnishing novel analytical tools and conceptual insights for cluster analysis research. Not only that, this paper also chooses Chinese flexible electronics patents as the research object. On the one hand, it is because Baidu API can automatically obtain the latitude and longitude based on the detailed address information provided by Chinese patents, and on the other hand, flexible electronics have been highly valued by Chinese governments at all levels in recent years. More than this, the Baidu news index (https://index.baidu.com/v2/main/index.html#/crowd/oled?words=oled, accessed on 31 December 2022) shows that Beijing, Yangtze River Delta, Pearl River Delta, and Sichuan (cluster number c4) are currently hot regions of FEI in China, as depicted in Figure 11, in which the blue shading represents the exploratory result. Therefore, this also further proves the effectiveness of the method proposed in this article. Of course, the method is also applicable to the analysis of patents and their industrial clusters in other countries, if the API platform supports it.

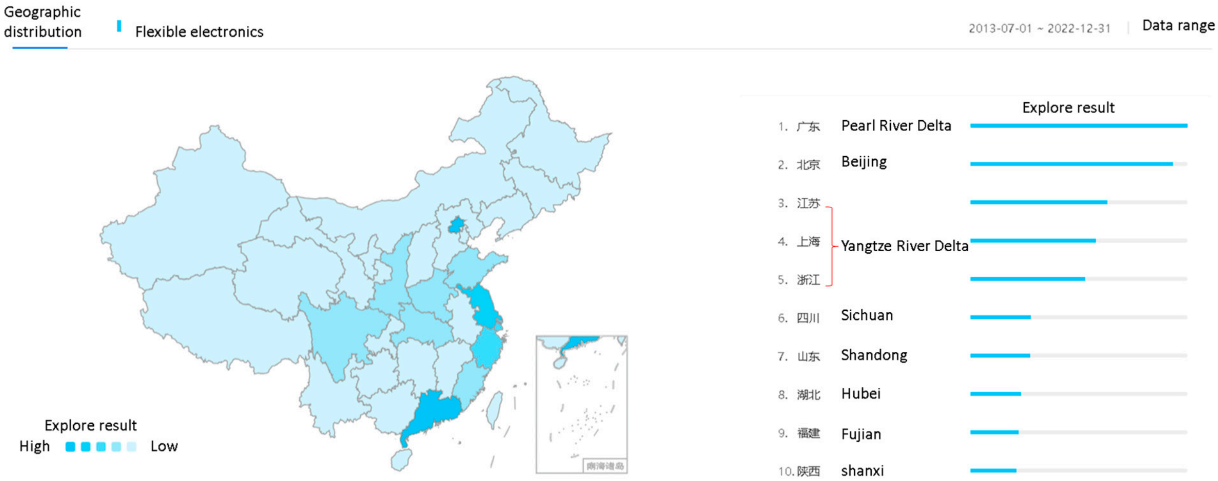


Figure 11. Explore results based on Baidu news index. 南海诸岛: South China Sea Islands.

6. Conclusions and Limitations

Since Marshall proposed industrial clusters, the accurate identification of those clusters has been a great concern of academics. The solution to this problem requires not only sufficient data but also a scientific approach. Although the existing methods can cluster individuals to be considered as industrial clusters, they lack the overall perspective of cluster screening. The main concerns include that the indicators of clustering are over simplified or ignore the technology factors, so they cannot be applied to the identification of IIC with multiple technologies. We propose a new algorithm using patent big data, combined with a text theme mining algorithm, a complex network, and MDBSCAN, for studying the clustering of the whole industry. The method adopts an unsupervised approach in both multi-technology theme mining and industry cluster identification to avoid the interference of human subjective factors, which can provide a more accurate, objective, and comprehensive analysis for the overall macroeconomic development.

Certainly, this paper also has some limitations, which can be seen as future research opportunities. First, the identification of industrial clusters is a complex problem that requires the consideration of various factors, such as economic, political, and even environmental factors. Relying on patent data alone can only provide a perspective from the technological perspective. Hence, future studies can aim to conduct a comprehensive analysis by integrating more data sources. Moreover, compared with the KM and DBSCAN algorithms, the processing time of the MDBSCAN-based algorithm is longer due to the integration of more indicators. Thus, future research should improve the efficiency of this algorithm. Lastly, the dynamic change and prediction of clusters have also been the focus of research in recent years, but the patents used in this paper do not consider the potential influences of the timeframe, so follow-up research can consider time series and analyses of the evolution of industrial clusters from a dynamic perspective.

Author Contributions: Conceptualization, S.Z., W.L. and R.X.; methodology, W.L.; software, W.L. and T.W.; validation, W.L.; formal analysis, S.Z. and W.L.; investigation, S.Z. and W.L.; resources, W.L.; data curation, T.W.; writing—original draft preparation, S.Z., W.L. and T.W.; writing—review and editing, W.L. and Z.C.; visualization, W.L.; supervision, S.Z. and W.L.; project administration, S.Z., W.L. and R.X.; funding acquisition, S.Z., W.L. and R.X. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is supported by the National Natural Science Foundation of China (No.52275249), the Social Science Foundation of Fujian Province (No. FJ2021B128, FJ2023B077), and the Education Reform Project of Xiamen University of Technology (No. JG202226).

Data Availability Statement: The data presented in this study are available upon request.

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

Appendix A. The Pseudo Code for MDBSCAN Preprocessing

Algorithm A1: MDBSCAN

```
Input: Patent data DB
Output: A set of clusters.
/*Step 1: Extraction of space data and technological data*/
Data_space = data['space']
Data_technological = data['lda_topic']
/*Step 2: Calculation of space distance and technological distance*/
Import Euclidean_Distance algoritham
sd = Euclidean_Distance(Data_space)
td = Euclidean_Distance(Data_technological)
/* Step 3: Run MDBSCAN algorithm*/
for $sd_{min}$ in range(sd):
    for $td_{min}$ in range(td):
        for $x_{min}$ in range(x):
            if(($sd \leq sd_{min}$)&($td \leq td_{min}$)):
                C = 0;
                for each point P in database DB:
                    if label(P) != undefined
                        continue
                    Neighbors N = RangeQuery(DB, $p$, $sd_{min}$, $td_{min}$)
                    if |N| < $x_{min}$: /* if number of neighbors less than $x_{min}$, P is set as noise*/
                        label(P) = Noise
                        continue
                    C = C + 1
                    label(P) = C /*Initiation of clusters label*/
                    Seed set S = N \{P}
                    for each point Q in S{
                        if label(Q) = Noise:
                            label(Q) = C
                        if label(Q) != undefined:
                            continue
                        label(Q) = C
                        Neighbors N = RangeQuery(DB, p, $sd_{min}$, $td_{min}$)
                        if N $\geq$ $x_{min}$:
                            S = S.append(N) /* neighbors set N are append in seed set S */
                    end
                end
            end
        end
    end
end
```

Appendix B. LDA Topic Classification Results

No.	Topic	Keywords
T1	Polyacrylate	Tubular furnace, bottom plate layer, coating layer, polyacrylate, RNN
T2	Metal nanotubes	Driving electrode, dielectric layer, gold nanorod, cladding layer, cathode body
T3	Hydrogel	Preparation, conductivity, sintered product, support matrix, polyacrylamide
T4	Conductive polymer	Polymer materials, conductivity, polymeric materials, polymers
T5	polyurethane	Polyurethane, adhesive, insulation pad, curing adhesive, PU
T6	Polyethylene terephthalate	Polyethylene terephthalate, fiber layer, PET, device
T7	Organic materials	Thiophene polymer, protein film, fiber bundle, nano carbon coating, naphthalene tetramethylene diamine
T8	Conductive adhesive	Bonding, conductive adhesive, epoxy resin, corrosion resistance, flexibility

No.	Topic	Keywords
T9	Carbon based materials	Carbon nano, tube-based, foam layer, graphene, polydimethylsiloxane
T10	Carbon nanomaterials	Film remover, carbon nanofiber membrane, carbon nanotube, organic transistor, hydrogen film
T11	Inorganic materials	Inorganic materials, silk fibroin, indium antimonide, perovskite, inorganic nanometer
T12	Metal foil	Foil, copper foil, emitter, flexible circuit, embedded
T13	Copper indium gallium selenium	Silicon-based substrate, copper indium gallium sulfide selenium sensitized layer, copper indium gallium sulfide selenium sensitized semiconductor, hydrogen ion, solar cell
T14	III-V family	III-V family, gallium arsenide, indium phosphide, gallium nitride, coated surface
T15	Metal nanowires	Conductive film, preparation, silver paste circuit layer, platinum-based bimetallic, platinum-based bimetallic nanowires
T16	Dimethyl siloxane	Dimethyl siloxane, thermal permeation method, wet chemical method, dipole moment, isolation layer
T17	biodegradable	Green, fiber bundle, natural, biological, protein film
T18	liquid metal	Microflow pipeline, melting point, temperature, solidification, robustness
T19	Structured conductive polymer	Structural type, conductivity, conductivity, polyacetylene, carbon nanomaterials
T20	Magnetron sputtering	Compensation meter, mask diagram, touch panel, Magnetron sputtering metal-plated electrode, hardening film
T21	Stretchable	Dense circuit, stretch type, bow tie type, aluminum silver alloy, non-device area
T22	Graphene	Graphene conductive electrode, high membrane-based bonding strength, data cable, amorphous silicon film, output terminal
T23	Cu2SnS3	Single membrane, Cu2SnS3, hydrophilicity, fuel cell, silicon-coated carbon particles
T24	Flexible	Digital signals, protective devices, plasma elements, organic material layers, bending resistance properties
T25	Co-polyester	Micro nano particles, fiber optic communication, Z-resin, transparency, Co-polyester
T26	developable	Circuit layer, flexibility, folding, electrolyte, size
T27	Nanoparticles	Polymer, conductivity, Magnetron sputtering coater, metal, UV laser
T28	Semiconductor type carbon nanotubes	Generation tube, charge, titanium nitride film layer, carbon nanotube optoelectronic device, enrichment method
T29	Flexible polymer	Elasticity, structural layer, dimethyl carbonate, substrate, printed circuit board
T30	Memory attribute	Photosensitive sensors, gold nanoparticles, nanoimprinting technology, memory alloys, Bragg gratings
T31	Polyimide PI	Tin-based nanocrystals, gallium-based indium tin, silver paste circuit layer, etching solution, polyimide-based
T32	Resistive type	Signal processing circuit, flow meter, pressure, piezoresistive stress sensor, corresponding stress
T33	Lift off process	Graphene glass carbon sheet, ultrasonic induction layer, linear movement, deposited thin film material, organic adhesive film
T34	Photon welding	Zinc oxide nanotubes, nano photons, resin-like vacuum deposition, peripheral circuit, UV curing
T35	Low-temperature soldering	Non-contact circuit, high temperature resistance, solder paste, welding, polymer material fiber mesh
T36	Evaporative deposition	Plasma chemical vapor deposition machine, thin film resistor, tin sol, deposition machine, induction board
T37	Soft etching	Single-material film, fuel cell, corrosion resistance, crystal drying, photolithography and etching, prefabricated film
T38	reactive sputtering	Pre sputtering chamber, buried resistance material layer, impedance tester, flap valve, Magnetron sputtering deposition chamber
T39	Sputter deposition	Magnetron magnetic plating, Magnetron sputtering coating source, deposition particles, roll-to-roll vacuum deposition machine, deposition coating
T40	Atmospheric chemical vapor deposition	Ion chemical vapor deposition, optoelectronics, atmospheric pressure chemical vapor deposition devices, micro/nano optics, conductive sheets

No.	Topic	Keywords
T41	Arc evaporation	Super hydrophobicity, zinc oxide nanowires, arc ion plating, arc ion plating, DC arc spraying method
T42	Plasma enhanced chemical deposition	Silicon oxide micro ring core cavity, chemical vapor deposition cavity, deposition insulation layer, pulse power supply, chemical vapor deposition reaction chamber
T43	screen printing	Substrate film, ink, pattern, semiconductor tube, scraper
T44	Additive manufacturing	3D printing, gel electrolyte, main arc power supply, UV curing, lamination
T45	Electron beam evaporation	Zinc oxide nanocrystals, polymer-based composites, passivation alloys, near-infrared reflectance, electron beam current,
T46	RF sputtering	Vacuum conditions, thin films, direct current, power, temperature
T47	Laser pulse evaporation	Dielectric layer, nano plating, laser pen, pulsed light, alkaline solution
T48	Piezoelectric method	Ultrasonic motor, integrated module, piezoelectric coefficient, electromechanical resonator, consistency
T49	Inkjet printing	Hydrogen film, substrate, carbon ink, organic liquid source, polymethyl methacrylate,
T50	Transfer Integration	Integrated variable torque sensor, seal, substrate, heating element, circuit board heat transfer printing
T51	Nanoimprinting	Electron beam, template, transparent strip, flexible circuit strip, nano imprinted substrate
T52	Dry preparation	Surface nanostructure, photoresist, semiconductor device, dry etching, roughness
T53	Wet preparation	Electromagnetic shielding film, wet etching machine, drilling, high-frequency mixed pressure, electroplating
T54	Low pressure chemical vapor deposition	Plasma chemical vapor deposition machine, atomic flow, thin film, temperature, pressure
T55	Photolithography	Corrosion resistance, laser, photolithography, concentric ring, grating
T56	Capacitive type	Sensors, capacitors, nanowires, pressure, conductivity
T57	Hot bubble method	Heat dissipation, metal nano ink, high-temperature sintering, particle-free copper ink, thermoplastic ink powder
T58	Roll to roll preparation	Thermal conductive layer, deposited particles, roll-to-roll vacuum deposition machine, array, carbon nanotubes
T59	DC sputtering	Nanoimprint adhesive, deposition layer, metal frame, DC sputtering, current
T60	Resistive evaporation	Passive resistance film, coating machine, thin film, hole lithography, inflatable pump
T61	Chemical vapor deposition	Chemical vapor deposition, photonic crystal period, cathode to ground, source plate, preparation method
T61	Porous deposition	Porous, thin film, deposition system, agglomeration device, ion beam
T63	Sol gel method	Bare electrode, sol gel method, synthetic rubber, magnetic absorber, silicon substrate
T64	Inorganic display	Inorganic electroluminescence, substrate, luminescent material, display screen, coating
T65	Blood oxygen	Blood oxygen signal, sensor, health, parameters, measuring instrument
T66	Mechanical energy generation	Power generation film, generator, motor, energy, mechanism
T67	Electroencephalogram	Sensitivity, recognition, intention, EEG signals, head-mounted
T68	Piezoelectric type	Thermoelectric materials, pressure sensors, arrays, touch, sensitivity
T69	temperature	Integrated sensing, ambient temperature, variable shape, sensor, thermal interface
T70	Organic semiconductor	Micro lens, electrode block, organic insulation layer, electroplated nickel, crystalline silicon solar energy
T71	pressure	Adhesive, artificial intelligence, pressure sensing, direct current method, die-casting mold
T72	display	Organic field-effect transistor, large amplitude, bipolar plate, display screen, curved screen
T73	chronic disease	Chronic diseases, physiological monitoring sensors, deposition rate, powder cavity, nanoliposomes

No.	Topic	Keywords
T74	Silicon thin film battery	Silicon film, nanoribbons, solar energy, insulation board, sic
T75	humidity	Flow generation, conductive mesh, temperature and humidity, water treatment, dampers
T76	Dye sensitized battery	Photosensitive materials, solar cells, gate metal electrodes, sensitivity, transformers
T77	Communication device NFC	Magnetron sputtering metal plating electrode fixture, real-time communication, auxiliary substrate, isolator, digital signal transmission
T78	Perovskite battery	Solar cells, transparent electrodes, formamidine perovskite, nano copper aerosols, zirconium targets
T79	Optoelectronics	Copper wire layer, insulation layer, light trough, photodetector, conductive silicone adhesive layer
T80	Organic display	Registers, organic luminescent material films, prepackaging, prepackaging layers, wire racks
T81	Thin film solar cells	Electrode block, organic insulating layer, electroplated nickel, crystalline silicon solar energy, flexible circuit board
T82	ultra-thin glass	Float glass, high vitrification, chemical vapor deposition chamber, deposition insulation layer, pulse power supply
T83	strain	Regulator, strain gauge, torsion wheel, torque sensor, display end
T84	Inorganic semiconductor	Silicon dioxide layer, titanium dioxide photocatalytic network, target base, Raman spectroscopy, semiconductor materials
T85	clothing	Medical clothing, work clothes, flexible sensors, intelligence, functionality
T86	automobile	Film-forming agent, humanized automotive parts, electrode part, flexible roll
T87	Energy storage	Flexible lithium-ion batteries, flexible electrolytes, carbon nanotubes, porous carbon nanofiber films, electrolytes
T88	fingerprint	Sensors, sensing circuits, signals, fingerprint modules, bonding effects
T89	packing	RFID, high mechanical strength, accommodating parts, labels, circuits
T90	Energy collection	Battery, energy, electrode, preparation method, positive electrode
T91	Industry 4.0	Industry 4.0, intelligent online, wireless, sensor, portable
T92	fault diagnosis	Pressure, capacitive, sensor, load, equipment failure
T93	Wearable	Intelligent device, flexible display screen, bracelet, touch signal, cover glass
T94	medical care	Flexible paddles, polypropylene, substrate holder, healthcare, water absorption

References

1. Zhuang, Z.; Fu, S.; Lan, S.; Yu, H.; Yang, C.; Huang, G.Q. Research on economic benefits of multi-city logistics development based on data-driven analysis. *Adv. Eng. Inform.* **2021**, *49*, 101322. [CrossRef]
2. Marshall, A. *Principles of Economics*; Macmillan and Co.: London, UK; New York, NY, USA, 1890.
3. Porter, M.E. *The Competitive Advantage of Nations*; Free Press: New York, NY, USA, 1990; 855p.
4. Zhuang, T.; Zhao, S.L. Collaborative innovation relationship in Yangtze River Delta of China: Subjects collaboration and spatial correlation. *Technol. Soc.* **2022**, *69*, 101974.
5. Huang, Q. Reconstruction of the global industrial chain under the epidemic—Develop an industrial chain cluster combining horizontal division of labor and vertical integration. *China Econ. Wkly.* **2020**, *780*, 24–29.
6. Nishimura, J.; Okamuro, H. R&D productivity and the organization of cluster policy: An empirical evaluation of the Industrial Cluster Project in Japan. *J. Technol. Transf.* **2011**, *36*, 117–144.
7. Guzman, J.; Stern, S. Where is Silicon Valley? *Science* **2015**, *347*, 606–609. [CrossRef]
8. Best, M.H. Greater Boston's industrial ecosystem: A manufactory of sectors. *Technovation* **2015**, *39*, 4–13. [CrossRef]
9. Filatotchev, I.; Liu, X.; Lu, J.; Wright, M. Knowledge spillovers through human mobility across national borders: Evidence from Zhongguancun Science Park in China. *Res. Policy* **2011**, *40*, 453–462. [CrossRef]
10. Yongsheng, X.; Xiaole, Z.; Wei, W. Coupling or lock-in? Co-evolution of cultural embedders and cluster innovation-exploratory case study of Shaoxing textile cluster. *Technol. Soc.* **2021**, *67*, 101765. [CrossRef]
11. China Ministry of Science and Technology, Measures for Pilot Certification of Innovative Industrial Clusters. Available online: https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qtwj/qtwj2013/201307/t20130702_106869.html (accessed on 20 March 2024).
12. China Ministry of Industry and Information Technology. Guiding Opinions of the Ministry of Industry and Information Technology on Further Promoting the Development of Industrial Clusters. Available online: https://www.gov.cn/gongbao/content/2015/content_2975894.htm (accessed on 20 March 2024).
13. China Ministry of Industry and Information Technology. Measures for Promoting the Development of Characteristic Industrial Clusters of Small and Medium Enterprises. Available online: https://www.gov.cn/zhengce/zhengceku/2022-09/14/content_5709725.htm (accessed on 20 March 2024).

14. Karreman, B.; Burger, M.J.; van Eenennaam, F. Revealed competition between cluster organizations: An exploratory analysis of the European life sciences sector. *Environ. Plan. A* **2018**, *51*, 705–723. [CrossRef]
15. Wang, X.; Liu, J.; Ma, C. A research on the cluster competitiveness evaluation of the Chinese automobile industry based on cuckoo-AHP. *Chin. Manag. Stud.* **2016**, *10*, 746–769. [CrossRef]
16. Geum, Y.; Kim, M.; Lee, S. How industrial convergence happens: A taxonomical approach based on empirical evidences. *Technol. Forecast. Soc. Chang.* **2016**, *107*, 112–120. [CrossRef]
17. Sun, C.C.; Lin, G.T.R.; Tzeng, G.H. The evaluation of cluster policy by fuzzy MCDM: Empirical evidence from HsinChu Science Park. *Expert. Syst. Appl.* **2009**, *36*, 11895–11906. [CrossRef]
18. Dimos, C.; Fai, F.M.; Tomlinson, P.R. The attractiveness of university and corporate anchor tenants in the conception of a new cluster. *Reg. Stud.* **2021**, *55*, 1473–1486. [CrossRef]
19. Xiao, R.B. Four development stages of collective intelligence. *Front. Inf. Technol. Electron. Eng.* **2024**, *25*, 903–916. [CrossRef]
20. Anselin, L. Local Indicators of Spatial Association-LISA. *Geogr. Anal.* **2010**, *27*, 93–115. [CrossRef]
21. Voyer, R. Knowledge-Based Industrial Clustering: International Comparisons. In *Local and Regional Systems of Innovation*; Springer: Boston, MA, USA, 1998; pp. 81–110.
22. Simmie, J.; Sennett, J. Innovative clusters: Global or local linkages? *Natl. Inst. Econ. Rev.* **1999**, *170*, 87–98. [CrossRef]
23. Xiong, W.; Li, J. The Knowledge Spillover Effect of Multi-Scale Urban Innovation Networks on Industrial Development: Evidence from the Automobile Manufacturing Industry in China. *Systems* **2024**, *12*, 5. [CrossRef]
24. Guo, B.; Guo, J. Patterns of technological learning within the knowledge systems of industrial clusters in emerging economies: Evidence from China. *Technovation* **2011**, *31*, 87–104. [CrossRef]
25. Engel, J.S.; Del-Palacio, I. Global networks of clusters of innovation: Accelerating the innovation process. *Bus. Horiz.* **2009**, *52*,493–503. [CrossRef]
26. Molina-Morales, F.X.; Martinez-Fernandez, M.T. Social Networks: Effects of Social Capital on Firm Innovation. *J. Small Bus. Manag.* **2010**, *48*, 258–279. [CrossRef]
27. Nie, L.; Wang, Y. Spatial Effects of Service Industry's Heterogeneous Agglomeration on Industrial Structure Optimization: Evidence from China. *Systems* **2024**, *12*, 85. [CrossRef]
28. Shi, J.; Lai, W. Fuzzy AHP approach to evaluate incentive factors of high-tech talent agglomeration. *Expert. Syst. Appl.* **2023**, *212*, 118652. [CrossRef]
29. McLoughlin, F.; Duffy, A.; Conlon, M. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl. Energy* **2015**, *141*, 190–199. [CrossRef]
30. Xu, Y.; Li, X.; Tao, C.; Zhou, X. Connected knowledge spillovers, technological cluster innovation and efficient industrial structure. *J. Innov. Knowl.* **2022**, *7*, 100195. [CrossRef]
31. National Research Council; Policy and Global Affairs. *Growing Innovation Clusters for American Prosperity: Summary of a Symposium*; National Academies Press: Washington, WA, USA, 2011.
32. Corrocher, N.; Lamperti, F.; Mavilia, R. Do science parks sustain or trigger innovation? Empirical evidence from Italy. *Technol. Forecast. Soc. Chang.* **2019**, *147*, 140–151. [CrossRef]
33. Delgado, M.; Porter, M.E.; Stern, S. Defining clusters of related industries. *J. Econ. Geogr.* **2016**, *16*, 1–38. [CrossRef]
34. Yang, T.; He, M. Study on the Features of Textile Industry Cluster in Guangzhong. *Int. J. Bus. Manag.* **2010**, *6*, 243.
35. Lan, S.; Yang, C.; Huang, G.Q. Data analysis for metropolitan economic and logistics development. *Adv. Eng. Inform.* **2017**, *32*, 66–76. [CrossRef]
36. Zhao, W.; Watanabe, C.; Griffy-Brown, C. Competitive advantage in an industry cluster: The case of Dalian Software Park in China. *Technol. Soc.* **2009**, *31*, 139–149. [CrossRef]
37. Liu, C. The effects of innovation alliance on network structure and density of cluster. *Expert. Syst. Appl.* **2011**, *38*, 299–305. [CrossRef]
38. Catini, R.; Karamshuk, D.; Penner, O.; Riccaboni, M. Identifying geographic clusters: A network analytic approach. *Res. Policy* **2015**, *44*, 1749–1762. [CrossRef]
39. Lin, W.; Yu, W.; Xiao, R. Measuring Patent Similarity Based on Text Mining and Image Recognition. *Systems* **2023**, *11*, 294. [CrossRef]
40. Liu, C.Y.; Yang, J.C. Decoding Patent Information Using Patent Maps. *Data Sci. J.* **2008**, *7*, 14–22. [CrossRef]
41. Li, X.; Fan, M.; Zhou, Y.; Fu, J.; Yuan, F.; Huang, L. Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining. *Nano Energy* **2020**, *71*, 104636. [CrossRef]
42. Tiefelsdorf, M.; Boots, B. A Note on the Extremities of Local Moran's I is and Their Impact on Global Moran's I. *Geogr. Anal.* **1997**, *29*, 248–257. [CrossRef]
43. Trappey, A.J.C.; Pa, R.J.S.; Chen, N.K.T.; Huang, A.Z.C.; Li, K.; Hung, L.P. Digital transformation of technological IP portfolio analysis for complex domain of satellite communication innovations. *Adv. Eng. Inform.* **2023**, *55*, 101879. [CrossRef]
44. Kagawa, S.; Suh, S.; Kondo, Y.; Nansai, K. Identifying environmentally important supply chain clusters in the automobile industry. *Econ. Syst. Res.* **2013**, *25*, 265–286. [CrossRef]
45. Argüelles, M.; Benavides, C.; Fernández, I. A new approach to the identification of regional clusters: Hierarchical clustering on principal components. *Appl. Econ.* **2014**, *46*, 2511–2519. [CrossRef]

46. Zhao, Z.; Zhao, Z.; Zhang, P. A new method for identifying industrial clustering using the standard deviational ellipse. *Sci. Rep.* **2023**, *13*, 578. [CrossRef]
47. Souris, M.; Bichaud, L. Statistical methods for bivariate spatial analysis in marked points. Examples in spatial epidemiology. *Spat. Spatiotemporal Epidemiol.* **2011**, *2*, 227–234. [CrossRef]
48. Anselin, L.; Li, X. Tobler's Law in a Multivariate World. *Geogr. Anal.* **2020**, *52*, 494–510. [CrossRef]
49. Guo, J.; Lao, X.; Shen, T. Location-Based Method to Identify Industrial Clusters in Beijing-Tianjin-Hebei Area in China. *J. Urban. Plan. Dev.* **2019**, *145*, 04019001. [CrossRef]
50. Billings, S.B.; Johnson, E.B. The location quotient as an estimator of industrial concentration. *Reg. Sci. Urban. Econ.* **2012**, *42*, 642–647. [CrossRef]
51. Lai, Y.L.; Hsu, M.S.; Lin, F.J.; Chen, Y.; Lin, Y. The effects of industry cluster knowledge management on innovation performance. *J. Bus. Res.* **2014**, *67*, 734–739. [CrossRef]
52. Li, C.Z.; Xu, Z.B.; Qiao, C.; Luo, T. Hierarchical clustering driven by cognitive features. *Sci. China-Inf. Sci.* **2014**, *57*, 1–14. [CrossRef]
53. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 86–97. [CrossRef]
54. Pavlis, M.; Dolega, L.; Singleton, A. A Modified DBSCAN Clustering Method to Estimate Retail Center Extent. *Geogr. Anal.* **2018**, *50*, 141–161. [CrossRef]
55. Neto, A.C.A.; Sander, J.; Campello, R.J.G.B.; Nascimento, M.A. Efficient Computation and Visualization of Multiple Density-Based Clustering Hierarchies. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 3075–3089. [CrossRef]
56. Blei, D.M.; Ng, A.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
57. Park, H.; Park, T.; Lee, Y. Partially collapsed Gibbs sampling for latent Dirichlet allocation. *Expert. Syst. Appl.* **2019**, *131*, 208–218. [CrossRef]
58. Yi, K.; Zhou, Z.; Wu, Y.; Zhang, Q.; Li, X. Empathic connectivity of exhibition technology and users in the digital Transformation: An integrated method of social network analysis and LDA model. *Adv. Eng. Inform.* **2023**, *56*, 102019. [CrossRef]
59. Newman, D.; Asuncion, A.; Smyth, P.; Welling, M. Distributed Algorithms for Topic Models. *J. Mach. Learn. Res.* **2009**, *10*, 1801–1828.
60. Andewelt, R.B. Analysis of patent pools under the antitrust laws. *Antitrust Law. J.* **1984**, *53*, 611–639.
61. Liu, Z.; Chen, X.; Xu, W.; Chen, Y.; Li, X. Detecting industry clusters from the bottom up based on co-location patterns mining: A case study in Dongguan, China. *Env. Plan. B-Urban. Anal. City Sci.* **2021**, *48*, 2827–2841. [CrossRef]
62. Malmberg, A.; Maskell, P. The Elusive Concept of Localization Economies: Towards a Knowledge-Based Theory of Spatial Clustering. *Environ. Plan. A* **2016**, *34*, 429–449. [CrossRef]
63. Flores-Sintas, A.; Cadenas, J.M.; Martin, F. Detecting homogeneous groups in clustering using the Euclidean distance. *Fuzzy Sets Syst.* **2001**, *120*, 213–225. [CrossRef]
64. Monticone, L.C.; Snow, R.E.; Box, F. Minimizing Great-Circle Distance Ratios of Undesired and Desired Signal Paths on a Spherical Earth. *IEEE Trans. Veh. Technol.* **2009**, *58*, 4868–4877. [CrossRef]
65. Shang, S.; Zheng, K.; Jensen, C.S.; Yang, B.; Kalnis, P.; Li, G.; Wen, J. Discovery of Path Nearby Clusters in Spatial Networks. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1505–1518. [CrossRef]
66. Fouedjio, F. A hierarchical clustering method for multivariate geostatistical data. *Spat. Stat.* **2016**, *18*, 333–351. [CrossRef]
67. Fouedjio, F. A spectral clustering approach for multivariate geostatistical data. *Int. J. Data Sci. Anal.* **2017**, *4*, 301–312. [CrossRef]
68. García, J.; Crawford, B.; Soto, R.; Castro, C.; Paredes, F. A k-means binarization framework applied to multidimensional knapsack problem. *Appl. Intell.* **2018**, *48*, 357–380. [CrossRef]
69. Aliguliyev, R.M. Performance evaluation of density-based clustering methods. *Inf. Sci.* **2009**, *179*, 3583–3602. [CrossRef]
70. Song, W.; Wang, Y.; Pan, Z. A novel cell partition method by introducing Silhouette Coefficient for fast approximate nearest neighbor search. *Inf. Sci.* **2023**, *642*, 119216. [CrossRef]
71. Ertunç, E.; Karkınlı, A.E.; Bozdağ, A. A clustering-based approach to land valuation in land consolidation projects. *Land Use Policy* **2021**, *111*, 105739. [CrossRef]
72. Xiao, J.; Lu, J.; Li, X. Davies Bouldin Index based hierarchical initialization K-means. *Intell. Data Anal.* **2017**, *21*, 1327–1338. [CrossRef]
73. Katarya, R.; Saini, R. Enhancing the wine tasting experience using greedy clustering wine recommender system. *Multimed Tools Appl.* **2022**, *81*, 807–840. [CrossRef]
74. Zhou, T.; Ren, J.; Medo, M.; Zhang, Y. Bipartite network projection and personal recommendation. *Phys. Review. E Stat. Nonlinear Soft Matter Phys.* **2007**, *76*, 46115. [CrossRef]
75. Hu, Y. Efficient, high-quality force-directed graph drawing. *Math. J.* **2005**, *10*, 37–71.
76. Markusen, A. Sticky Places in Slippery Space: A Typology of Industrial Districts. *Econ. Geogr.* **1996**, *72*, 293–313. [CrossRef]
77. Bagheri, A.; Saraee, M.; de Jong, F. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. *J. Inf. Sci.* **2014**, *40*, 621–636. [CrossRef]
78. Liu, J.; Wei, J.; Liu, Y.; Jin, D. How to channel knowledge coproduction behavior in an online community: Combining machine learning and narrative analysis. *Technol. Forecast. Soc. Chang.* **2022**, *183*, 121887. [CrossRef]

79. Kukalis, S. Agglomeration Economies and Firm Performance: The Case of Industry Clusters. *J. Manag.* **2010**, *36*, 453–481. [CrossRef]
80. Haque, N. Mapping prospects and challenges of managing sludge from effluent treatment in Bangladesh. *J. Clean Prod.* **2020**, *259*, 120898. [CrossRef]
81. Someda, H.; Akagi, T.; Kajikawa, Y. An analysis of the spillover effects based on patents and inter-industrial transactions for an emerging blockchain technology. *Scientometrics* **2022**, *127*, 4299–4314. [CrossRef]
82. Lee, J.; Ko, N.; Yoon, J.; Son, C. An approach for discovering firm-specific technology opportunities: Application of link prediction to F-term networks. *Technol. Forecast. Soc. Chang.* **2021**, *168*, 120746. [CrossRef]
83. Ma, T.; Zhou, X.; Liu, J.; Lou, Z.; Hua, Z.; Wang, R. Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. *Technol. Forecast. Soc Chang.* **2021**, *173*, 121159. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.