International Journal of Geo-Information

MDPI

Article

Finding and Evaluating Community Structures in Spatial Networks

You Wan [1,2], Xicheng Tan [3] and Hua Shu [4,*]

[1] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; wanyou@whu.edu.cn
[2] Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, Wuhan 430079, China
[3] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[4] School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China
* Correspondence: shuh@hubu.edu.cn; Tel.: +86-188-1311-9475

Abstract: Community detection can reveal unknown spatial structures embedded in spatial networks. Current spatial community detection methods are mostly modularity-based. However, due to the lack of appropriate spatial networks serving as a benchmark, the accuracy and effectiveness of these methods have not been tested sufficiently so far. This study first introduced a spatial autoregressive and gravity model united method (SARGM) to simulate benchmark spatial networks with known regional distributions. Then, a novel spectral clustering-based spatial community detection method (SCSCD) was proposed to identify spatial communities from eight kinds of benchmark spatial networks. Comparative experiments on SCSCD and three other methods showed that SCSCD performed the best in accuracy and effectiveness. Moreover, the scale parameter and the community number setting of the SCSCD were investigated experimentally. Finally, a case study was applied to the SCSCD to demonstrate its ability to extract the internal community structure of a high-speed train network in China.

Keywords: spectral clustering; spatial community detection; spatial community evaluation; benchmark spatial network

Citation: Wan, Y.; Tan, X.; Shu, H. Finding and Evaluating Community Structures in Spatial Networks. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 187. https://doi.org/10.3390/ijgi12050187

Academic Editors: Godwin Yeboah and Wolfgang Kainz

Received: 22 March 2023
Revised: 29 April 2023
Accepted: 1 May 2023
Published: 4 May 2023

1. Introduction

Spatial networks are networks for which nodes are located in geospatial space, and edges are constructed by actual flow or virtual interactions. Community (also called subsystem in the early research [1]) is a special network structure. It is a set of nodes with more connections among themselves than with the rest [2]. Identifying the community structure in spatial networks is very important because it can reveal people's movement patterns [3,4], identify spatial clusters of transportation networks [5,6], delineate healthcare service areas [7,8], and prevent epidemics spreading [9,10].

The spatial relation constraints are the most significant difference between spatial and traditional community detection. Traditional community detection methods do not consider spatial contiguity in their community results [7]. On the other hand, spatial community detection can be a particular type of spatially constrained regionalization problem. In the traditional regionalization problems [11,12], geospatial units are aggregated into spatially contiguous and attributed homogeneous regions [13–15]. In contrast, a spatial network's distribution of unit attributes is usually heterogeneous under actual conditions. Even in the same community, the attributes of units can be quite heterogeneous. Therefore, both traditional community detection and regionalization methods cannot be directly used for spatial community detection.

Current spatial community detection methods are mainly based on modularity optimization, which divides the network by maximizing the difference between real and expected connections modeled by appropriate null models [16]. However, due to the

complexity of building spatial null models and simulating benchmark spatial networks, very little work has been done to evaluate the accuracy and effectiveness of these methods. This leads to uncertainty and problems in the results of existing methods.

This study proposed a new spatial autoregressive and gravity model united method (SARGM) to simulate spatial networks with known community distributions. SARGM first uses the spatial autoregressive model to simulate benchmark regional data with known regional numbers and spatial distributions. Then the SARGM uses the gravity model to generate benchmark spatial networks based on the simulated regional data. The spatial communities existing in the spatial network are considered to have the same number and distribution as the regions in the corresponding regional data. Therefore, their locations and scopes can be used to evaluate the accuracy of different community detection methods. In addition, a novel spectral clustering-based spatial community detection method (SCSCD) was proposed to divide the spatial network into communities. SCSCD integrates the spatial contiguity constraint into the traditional spectral clustering process, which can divide the geospatial units into spatially contiguous communities without any spatial null model assumption. Comparative experiments with three other classical community detection methods on eight kinds of simulated benchmark data have proven that the proposed SCSCD method performed best. Furthermore, the accuracy of SCSCD in finding the right communities in all simulated data is very stable and much higher than the other methods.

The rest of this article is arranged as follows. Section 2 presents the related works on spatial community detection. Section 3 describes the problem of spatially constrained network community detection and spectral clustering methods. Then, we introduce the benchmark spatial network simulation method SARGM and the spatial community detection method SCSCD. Section 4 describes experimentation. The accuracy and effectiveness of SCSCD on several benchmarks and actual regional data were evaluated. The scale parameter and the community number setting of SCSCD were also investigated experimentally. Finally, Section 5 provides the conclusions and discussions of this article.

2. Related Works

Existing spatial community detection methods can be classified into modularity-based and graph partitioning methods [16,17]. Newman [18] defined the modularity-based community as when the number of between-group edges is significantly lower than expected purely by chance. Studies on both simulated and benchmark networks have demonstrated the effectiveness of modularity on traditional complex network community detection [16,19–21]. In the spatial community detection area, some studies directly used classical community detection algorithms to discover topological structures in the network [5,22]. In addition, three extensional methods were developed to integrate spatial relationships into the detection process, which can find quite different community structures. Guo [14,23] proposed the REDCAP method, an agglomerative hierarchical clustering method with dynamic spatial contiguity constraints. Chen et al. [6] defined the geo-modularity, which added the inverse distance weights to the original edges in the network. As a result, the community structure tends to be more local since the new weights of the edges reinforce the distance decay effects among nodes. Expert et al. [24] integrated a distance decay factor into the original modularity model. Then, modified the classical community detection algorithm to detect communities that eliminate the spatial effect, revealing hidden structural similarities between nodes.

Nevertheless, several limitations of modularity-based methods have been pointed out [25,26], such as the resolution limit, the extreme degeneracy exhibited in the modularity function, and the limiting behavior of the maximum modularity [17]. Moreover, the communities found by modularity-based methods are highly affected by the construction of null reference models [21,27,28]. For example, Sarzynska et al. [29] investigated the effects of three null models (radiation, gravity, and the standard NG null model) incorporating spatial information on spatial community detection. They found that the quality of the

community results with different null models strongly depended on the network and parameter settings.

On the other hand, graph partitioning methods do not make any model assumptions on the network structure. Instead, they aim to divide the nodes into a group of predefined numbers such that the edges between groups are minimal. Many algorithms perform a bisection of the network [30]. Partitions in more than two communities are attained by iterative bisection [18,31]. However, the partitioning process cannot ensure that every iteration is correct, and the final result will be highly affected by previous partitions [18]. Therefore, using iterative bisection to split the network into more pieces is unreliable [27]. Newman and Girvan [32] proposed cutting edges with the highest betweenness iteratively. The final communities were obtained when a particular network division reached the largest modularity value. Zhang and Newman [33] presented a spectral algorithm to map modularity maximization to a vector partitioning problem, and the algorithm can directly divide the network into any number of communities. These two methods do not use modularity optimization strategies to maximize the modularity, but they still use modularity to evaluate their community results.

Spectral methods originated as graph partitioning and data clustering methods [34,35]. Researchers found that the spectrum eigenvectors of various types of graph matrices can be used to extract community structures from networks, such as adjacency, standard Laplacian, normalized Laplacian, modularity, and correlation matrices [36]. The edges between communities are commonly denoted as the graph or edge cut. Spectral clustering methods can efficiently find the best graph cut in a weighted network and are well-suited for community detection [18,33,37]. However, only a few spatial community detection studies have been performed based on spectral clustering. Existing methods do not consider the spatial contiguity constraint in the clustering process. Furthermore, they need two parameter-setting problems. The first problem is the scale parameter to construct the graph matrix between nodes. A graph matrix with a specific scale parameter is needed when the data structure is not a graph. This parameter is usually selected manually, and it has a significant effect on clustering results [38,39]. The second is the predefined number of communities [33]. This parameter could be derived by checking whether there is an integer n such that the first n eigenvalues are small and the $(n + 1)$-th is relatively large [27]. However, the number obtained by the eigengap always differs from the number of communities in the ground-truth community structures [40]. In particular, when the clusters are very mixed, it may be challenging to identify a significant gap between eigenvalues [27]. Cafieri et al. [41] defined an edge-cut ratio for a community as the ratio of the number of edges within a community to the number of cut edges with only one endpoint within that community. For different partition results, the best partition will be obtained when the edge cut ratio for that partition reaches its maximum value. Experiments on several artificial and well-known networks have shown that compared with modularity maximization, the edge-cut ratio appears not to suffer from the resolution limit problem and usually identifies more communities [41], often with similar or better precision [42]. However, experiments on benchmark spatial networks are still lacking, and their effectiveness on spatial networks is uncertain.

3. Methods

3.1. Problem Statements

Let G = {V, E} be a spatial network, where V is a set of georeferenced spatial units, and E is a set of weighted edges indicating the interaction among units. R = {1, 2, . . . , r} denotes the set of region identifiers, where r is the total number of regions. The goal of spatially constrained network community detection is to generate a partition of the units set V in spatial network G into r regions in such a way that: (1) maximizes the edge weights within each region, (2) minimizes the edge weights between regions, and (3) keeps the spatial contiguity within each region.

Figure 1a depicts the spatial distribution of a sample urban agglomeration comprising 20 regular grid boundary cities. All cities are grayscale-coded according to their attribute

values. The higher the attribute value, the darker the symbol. The attribute values here can be any numeric data to describe the capacity of cities, e.g., GDP or population. Cities numbered 1 and 13 have the largest attribute values compared to the others, so they are considered two core cities in the sample urban agglomeration. Then, interactions between each city pair are modeled by the gravity model. The larger the attribute values and the smaller the distance, the stronger the link between the two cities. For simplification, only each city's top k (k = 5) interaction links are chosen to construct the interaction network. As shown in Figure 1a, cities numbered 1 and 13 are two core cities in the network since they have the strongest links with nearby cities. When setting the community number r equals 2, the spatially constrained network community detection method will cut the interaction network into two smaller spatial contiguous urban agglomerations, as shown in Figure 1b. According to Figure 1a, each small urban agglomeration retains the strong links between core cities and their neighbor cities. Moreover, the edges cut exist at the weak links between two small urban agglomerations.

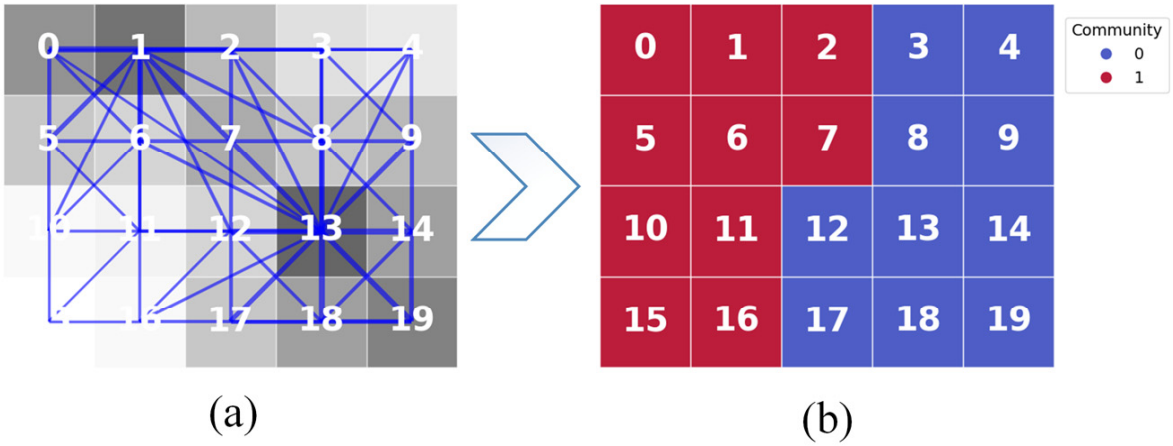


Figure 1. The spatially constrained network community detection on a sample regional dataset and its interaction network (**a**) A sample regional data and its interaction network, (**b**) Two communities found by the spatially constrained network community detection method.

3.2. Simulation of the Benchmark Spatial Interaction Networks

Although acquiring spatial data has become increasingly convenient in recent years, high-quality benchmark spatial networks that can be used to evaluate the accuracy of spatial community detection methods still need improvement. Here, a new SARGM method that unites two classical models, the spatial autoregressive (SAR) model and the gravity model (GM), is designed to simulate benchmark regional data and their interaction networks. A region in the simulated regional data has the same definition as a spatial community in the spatial network, a set of contiguous adjacent units with high interactions inside and low outside interactions. In this way, the distribution of regions in regional data equals the distribution of communities in the interaction network generated by the regional data, just as Figure 1 shows. Then, the accuracy of the community detection methods can be evaluated by comparing the community result with the original distribution of the regions.

3.2.1. Benchmark Regional Data Simulation by the SAR Model

SAR is one of the most commonly used autoregressive models for spatial data regression [43]. There are mainly four kinds of spatial autoregressive models [44] and the spatial lag model [45,46] was used here to simulate regional data. As shown in Equation (1), the spatial dependence is directly modeled on the right side of the equation using a contiguity matrix with a spatial autocorrelation coefficient parameter. In addition, the explanatory variable X can help generate the benchmark regions in simulated data.

Let $y = (y_1, \ldots, y_n)^T$ be an n-by-1 vector of observations referring to n spatial units s_1, $\ldots$, s_n; then the spatial autoregressive model can be written in the following vector form:

$$y = \rho W y + X\beta + \varepsilon \quad (1)$$

where ρ is the spatial autocorrelation coefficient parameter and ranges from -1 to 1 ($\rho = 0.9$ as default). A higher ρ value will ensure that y shows evident spatial regional patterns [11,47]. β is an n-by-1 vector of correlation coefficient parameters (β is set to all 1 as default). X is an n-by-m matrix containing the explanatory variables, and ε is an n-by-1 vector of random error term with mean equaling 0 and standard deviation equaling 1. W is an n-by-n matrix where W_{ij} represents the spatial contiguity of two units, i and j ($W_{ij} = 0$ when i equals j). Standard methods to form W include rook/queen contiguity relations and k nearest neighbor. Here, the queen contiguity relationship was chosen. The implementation of generating simulated SAR variables by using Equation (1) was based on an open-source python Package named Clusterpy [48]. Extra codes were added to their original function to generate the core units.

To simulate benchmark regions in the spatial data, X is fixed to an n-by-1 matrix with 10 as all units' default value, and a few core units have large values equal to 50. Setting the value of a few core units to 50 can distinguish them from the non-core units, which are initialized to 10. Like the small urban agglomerations in Figure 1b, the core and nearby units will generate the benchmark regions. Then, by setting the parameters above, and putting different numbers and positions of the core units, eight kinds of spatial data are generated and displayed in Figure 2. As can be seen, these spatial data all contain very simple and clear region distributions. Each core unit and its queen neighbors with large attribute values are labeled by a region number, representing a region's location and scope. The random error term defined in the Formula 1 will make the regional data different for every simulation time. Because the positions of core units in each kind of regional data are fixed, the location and scope of regions will not change in the same kind of regional data. On the other hand, the unlabeled units have relatively small attribute values, and their values are easily affected by the random error term. Therefore, assigning the correct region numbers to these units is challenging. In the end, the community results of these unlabeled units are not included in the accuracy test in the community detection experiments. Only those labeled units need to be identified as communities correctly.

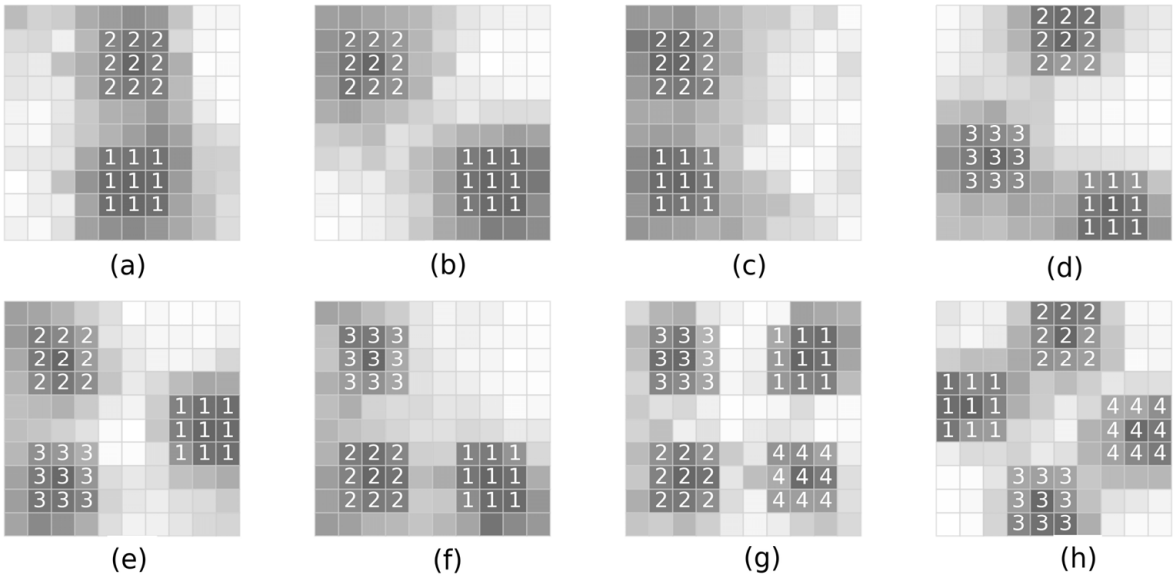


Figure 2. Eight kinds of benchmark regional data (**a**–**h**) with different regional distributions (the locations and scopes of regions are marked by different numbers).

3.2.2. Benchmark Spatial Network Generation by the Gravity Model

The popular gravity model is used to generate spatial networks based on the benchmark regional data in the previous section:

$$g(i,j) = y_i * y_j * \text{distance}(i,j)^{-\gamma} \quad (2)$$

where i and j are two spatial units and y_i and y_j are attribute values of two spatial units. Therefore, the spatial network can have the same region number and distribution as the benchmark regional data. However, the distance decay function may be quite distinct for different applications. To simplify, the calculation of the distance and value of the distance decay ratio γ are both data-driven and self-adapted. The distance between i and j is a ratio value calculated by the Euclidean distance between i and j dividing the average distance between all the unit pairs. In addition, the distance decay ratio γ is set according to the k nearest neighbor relation among units in the original network.

Since the whole network is very dense, a scale parameter is needed to simplify the network and decrease the computational demands [49]. This simplification operation is also beneficial to make the community boundaries clearer and sharper [40]. Specifically, the scale parameter can be the k nearest neighbor or ε-neighborhood (ε is a distance threshold) relation among units [35,49]. However, setting a unique distance threshold to a network is challenging, especially when outliers or clusters have different similarities inside. Thus, it is better to construct the k nearest neighbor interaction network (knn net).

Moreover, since spectral clustering needs a symmetric matrix to extract the eigenvectors, all directed connections in the knn net must be converted to undirected connections. There are two ways to do the conversion. One simply ignores the directions of edges, and the other connects two nodes if they both exist in each other's k nearest neighbor sets. Luxburg compared the two kinds of knn nets and the ε-neighborhood net, then suggested choosing the first knn conversion method as a general situation [35]. Compared with non-core units in the spatial interaction network, core units are more likely to become the top-k connected units of their neighboring units because of their larger attribute values. Therefore, after the conversion, some weak connections that do not exist in the top-k strongest connections of core units are retained, resulting in the connection number of core units larger than k. Furthermore, this will enhance the centrality of core units, especially for small γ values. On the other hand, if γ is set to a very large value, the core units' effect on the nearby units will decrease too fast, reducing the network's centrality. Therefore, too small or large γ values may both affect community detection. In the end, a ratio between the largest number of connections of the unit and k is defined to measure the fitness of γ. The ratio is called the largest connection ratio (LCR):

$$\text{LCR} = \max(\text{number of connections of each unit})/k \quad (3)$$

where the connections of each unit are calculated based on the undirected knn net. The LCR value indicates the maximum centrality accepted in a knn net. An empirical range of LCR can be set between 2 and k. Other values out of this range indicate that the distance decay ratio γ is not fitful.

When the distance decay ratio γ is set to 1.5, the LCRs of each regional data's four knn nets (k equals 3, 6, 9, and 12) are listed in Table 1. The LCR values are all between 2 and the corresponding k in each row. In addition, when γ is set to 1 or 2, some of the LCR values will be larger than k or smaller than 2. Therefore, the value of γ in all the knn nets is set to 1.5 as the default value for all the knn nets.

3.3. The Spectral Clustering-Based Spatial Community Detection Method

Spectral clustering is a clustering algorithm that relies on the eigendecomposition of feature similarity matrices to determine the cluster membership of its data points. Let $\{x_1, x_2, \cdots, x_N\}$ be a set of points to be clustered. To apply spectral clustering, we first compute a similarity matrix S between every pair of data points. Then, the similarity matrix is used

to construct an undirected weighted graph G = (V, E), and the weight of each edge in E is given by the similarity between the corresponding pair of data points. The Laplacian matrix of the graph is defined as L = D − S, where D is a diagonal matrix whose diagonal elements correspond to the weighted degree of each unit. Then, a set of *k* eigenvectors corresponding to the first *k* eigenvalues extracted from L provides a low-dimensional embedding over simple clustering methods, such as K-means or hierarchical partitioning, and can cluster more efficiently in lower dimensions [34].

Table 1. The largest connection ratio (LCR) of the regional data's knn nets.

γ	k	LCR of the Eight Regional Data's Knn Nets							
		a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
1.5	3	2.67	2.67	2.67	2.67	2.67	2.67	2.67	2.67
	6	3.50	3.67	3.67	2.83	3.67	3.67	3.83	2.83
	9	2.67	2.67	2.67	2.89	2.67	2.67	2.67	2.22
	12	2.92	2.67	2.42	2.58	2.58	2.42	2.50	2.42

Traditional spectral clustering methods do not consider the spatial contiguity constraint. In most cases, its clustering results cannot ensure spatial contiguity. A novel SCSCD method integrating a spatially constrained k-means algorithm was proposed to add the spatial constraint into spectral clustering, which can group eigenvectors into contiguous spatial regions. The spatially constrained k-means method is a special class of k-means clustering methods. Various constraints can be added to the clustering process, like must-link and cannot-link [50], spatial contiguity constrain [51], distance threshold [52] et al. Here, we implemented a spatial contiguity constrained k-means method. A detailed description of the SCSCD method is shown in Figure 3.

Given a spatial network G and a community number *r*, step 1 uses a spectral embedding method to find r eigenvectors of the spatial network. Step 2 uses the spatial constraint k-means algorithm on the eigenvector matrix to cut the network at its minimum weighted edges and to find highly connected and spatially contiguous regions. Two special parts in step 2 ensure the contiguity of k-means clustering results. In 2.7, the candidate moveable nodes are restricted to the boundary of each region. Only these nodes can be moved to their neighboring regions, which have a smaller distance than their current region. In 2.10, since the movement of the boundary nodes may occasionally destroy the contiguity of regions, each region's contiguity will be checked after each iteration of k-means clustering. Only the largest connected part is kept if a region is not fully connected. Other parts will be set to unlabeled nodes and assigned to their nearest adjacent region.

3.4. Determination of the Community Number

The community number is a predefined parameter in the SCSCD method. Unfortunately, most of the time, the best number of communities is unknown. Newman [18] pointed out that the spectral method for modularity optimization can be regarded as a principal components analysis for networks. The approach is similar in concept to the standard technique of principal components analysis (PCA) used to reduce high-dimensional datasets to low dimensions by focusing on the eigendirections along which the variance about the mean is greatest and ignoring directions that contribute little [18]. Inspired by this, the cumulative variance plot was used here to automatically find the best number of communities in the SCSCD method.

The cumulative variance plot is a popular method in principal component analysis. It can select a few derived projections of the dataset that conserve those characteristics that contribute the most to the variance. The larger the variance, the better projection it is. When the community number parameter is larger than the actual number, newly generated communities will not exist in densely connected areas. In contrast, they are located in weakly connected areas since the deviation operation tends to cut as few edges

as possible. Compared to real communities, these new communities usually do not contain interesting information since they have lower node degrees and lower degree variance. Here, the degree variances of all communities were ranked in descending order. Then, the cumulative variance list was added and plotted as a curve. Finally, the community that ranks at the elbow of the curve indicates the correct number of communities.

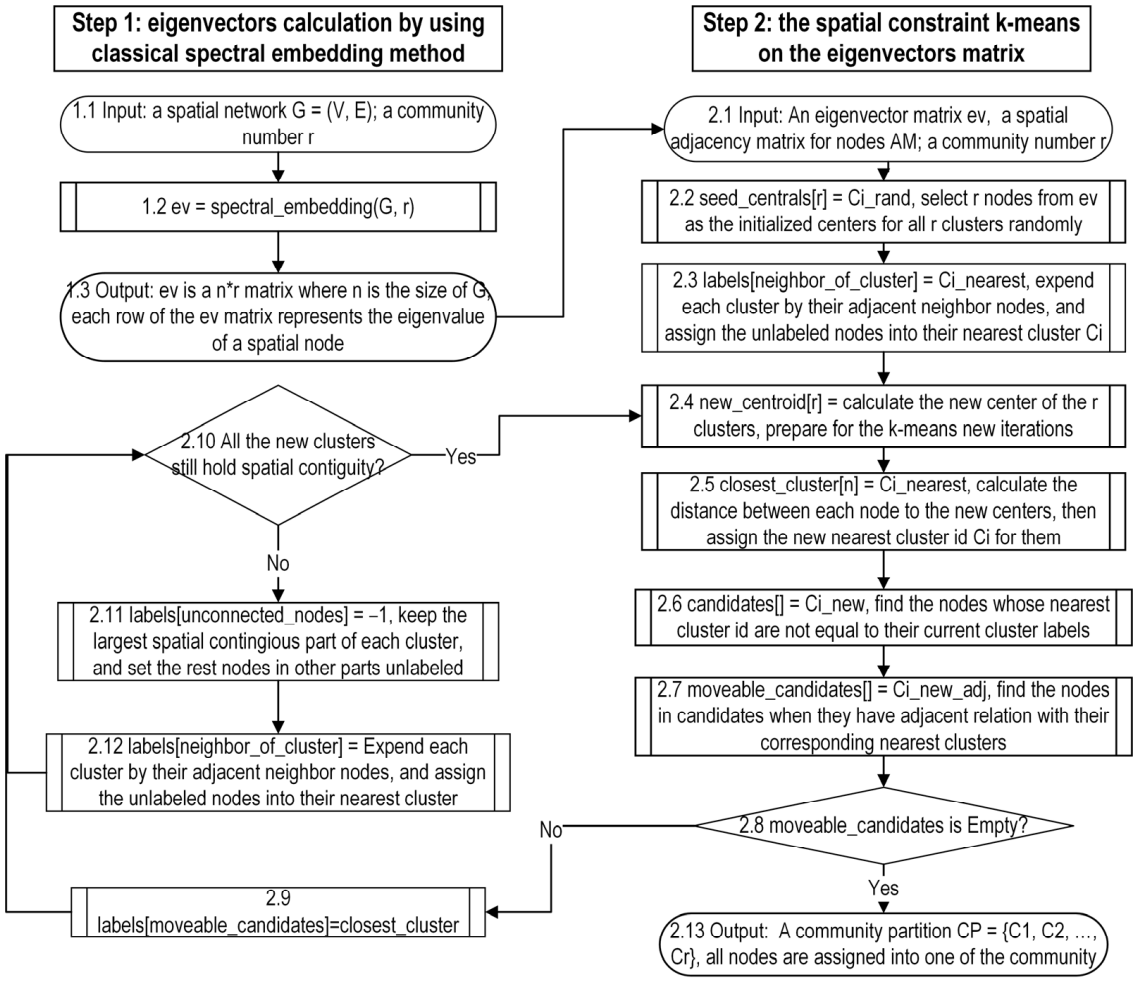


Figure 3. Spectral clustering based spatial community detection (SCSCD) method.

Still taking sample regional data in Figure 1 for example, when the distance decay ratio γ is set to 1, the LCR of the sample regional data's 5nn net is three. Given that the community number equals four, the community result and the cumulative variance plot are displayed in Figure 4. As the green curve shows, the largest degree of variance of the community is 0.66, and the community locates to the lower right corner of the SCSCD community result subfigure. The second largest degree of variance of the community is 0.2, and the community locates to the upper left corner of the SCSCD community result subfigure. The remaining two communities are located at the upper right and lower left corner of the SCSCD community result subfigure. They both have very low degree variances, which are 0.08 and 0.06. Then, the elbow point of the cumulative variance curve is automatically found at community two, indicating that the best number of communities is two. Further experimental analysis on the cumulative variance plot is presented in the next section.

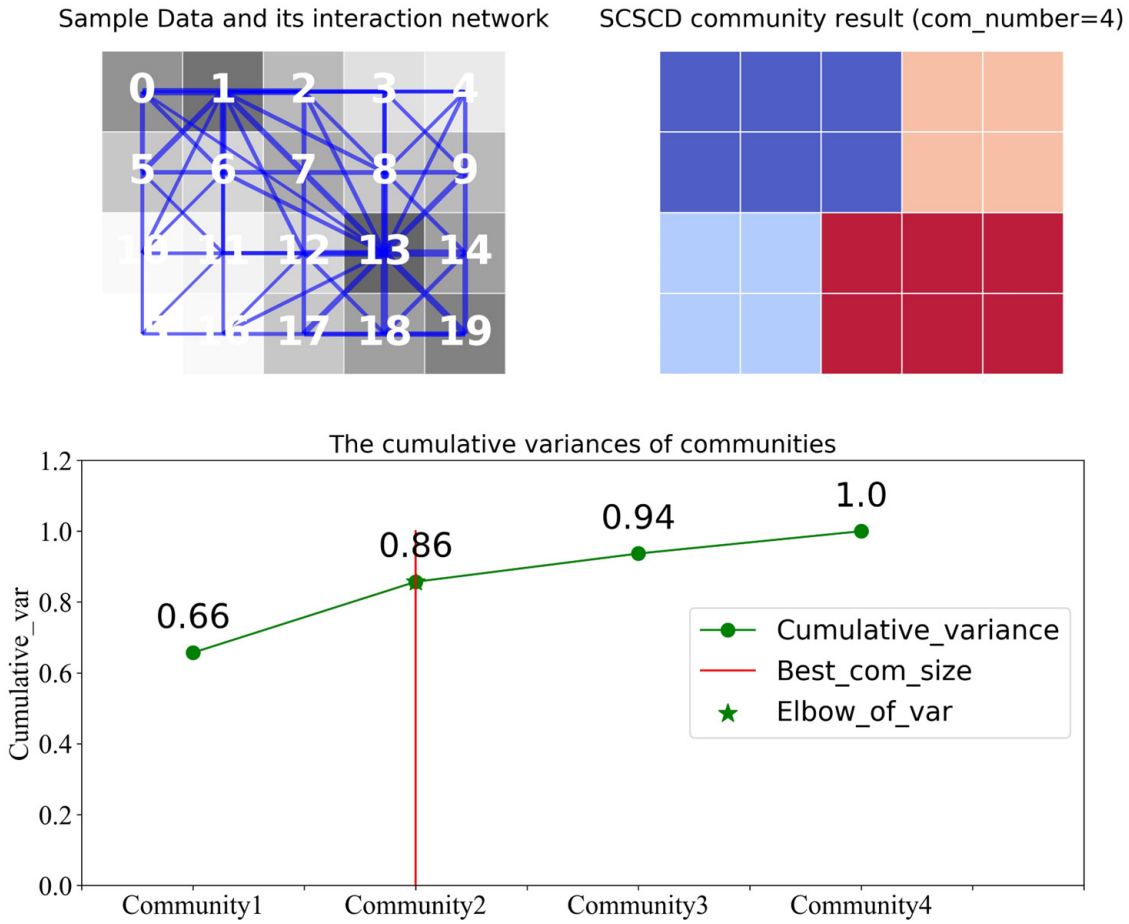


Figure 4. The cumulative variance plot of the communities of the sample regional data.

4. Experiments

This section tries to test the accuracy and effectiveness of SCSCD on both benchmark and real spatial networks. To avoid the random error effect during the simulation of benchmark regional data, each kind of regional data was simulated 100 times. Then, four different knn nets were generated from each regional data to perform the community detection experiments.

Three other modularity-based community detection methods were used to make a comparison with SCSCD. The first is a popular modularity optimization community detection method, Louvain [53]. The second is the geo-modularity optimization community detection method [6]. The distance decay ratio in the geo-modularity model is set to 1.5, the same as all simulated benchmark networks. The strategy to maximize geo-modularity is the same as Louvain, referred to as geo-Louvain. The third is a popular regionalization method REDCAP [14,23]. The full-order and average link constraint was chosen to generate spatially contiguous regions, and the objective function of REDCAP was also modularity maximizing.

4.1. Scale Parameter Experiments

For all simulated regional data, a scale parameter k is used to simplify their spatial networks. Therefore, the detection method can extract high-quality community structures accurately. However, different k values generate different knn nets, leading to different community results. Therefore, Luxbrug suggests trying out the appropriate k value "by foot" for small or medium size networks [35]. This section evaluates the stability and accuracy of SCSCD and three other methods on eight kinds of regional data when setting k values from 3 to 12 by 3 increments. The distance decay ratio γ in all the knn nets is set

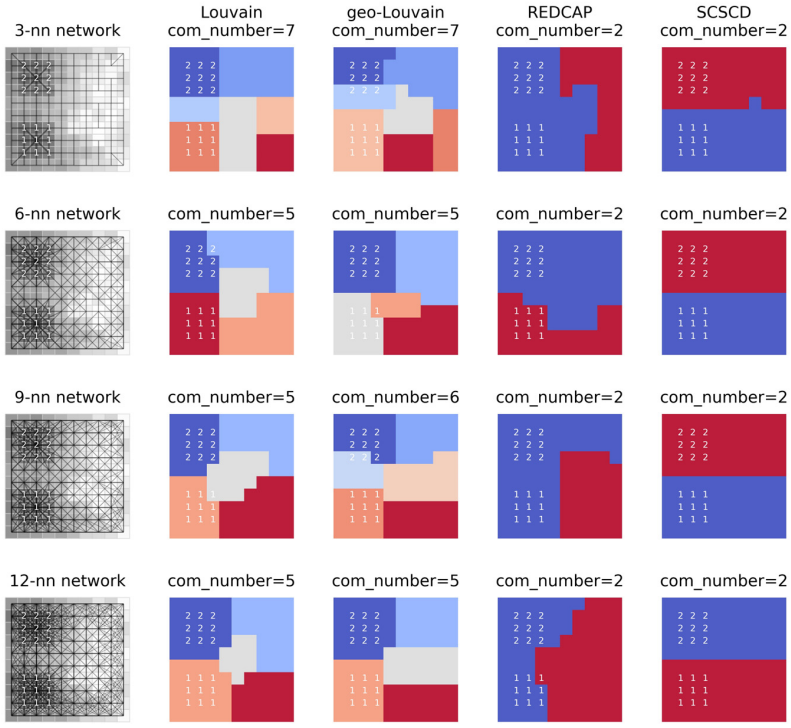
to 1.5 as default. In addition, to simplify the control experiments, the community number parameter is set to the actual number of regions of the simulated data as default in REDCAP and SCSCD.

Each simulated regional data generates four knn nets (k equals 3, 6, 9, and 12) and obtains four community results. In addition, the distributions of community results are usually different in each simulated regional data due to the randomness of data and detection process. The spatial distribution of one of the community results for the four methods on four knn nets of the regional data (c), (f), and (h) are drawn in Figure 5. They have covered region numbers from 2 to 4.

From Figure 5 and the other community results that are not shown here due to space limitations, some conclusions can be drawn:

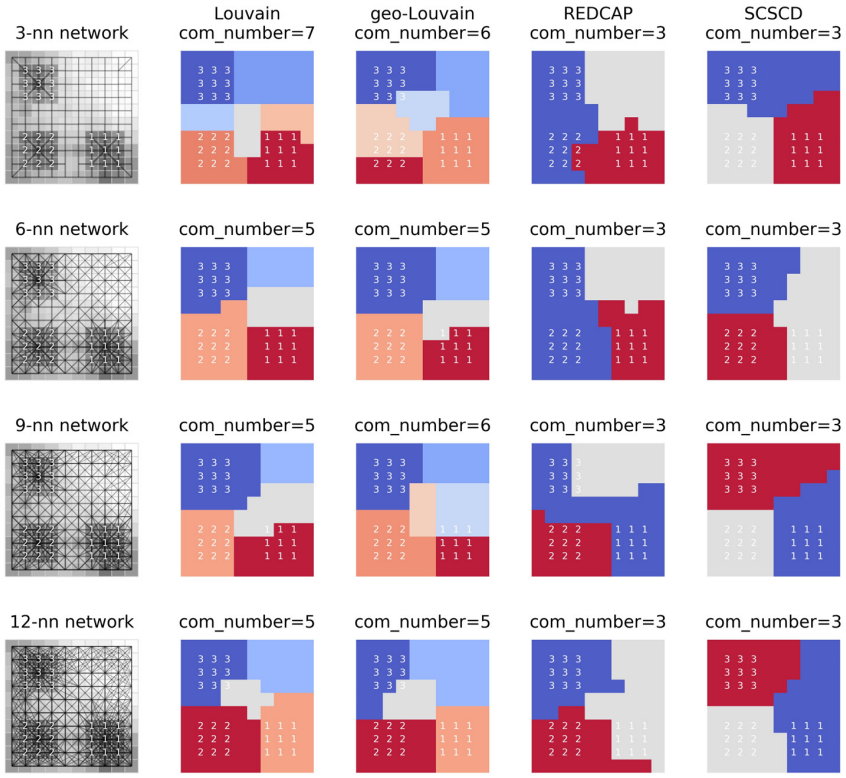
(1) As the scale parameter k increases, SCSCD obtains the most stable community results among the four methods. Moreover, the communities of SCSCD always have the same distribution as the simulated regional data. In contrast, the distribution of communities changes dynamically in the Louvain, geo-Louvain, and REDCAP methods. Therefore, they cannot catch the actual distribution of communities.

(2) Louvain and geo-Louvain do not have a predefined community number parameter, and the numbers of their communities are usually larger than the actual scenario. As a result, the unlabeled units located at the edge of the actual regions have small attribute values, and they are misidentified as communities in Louvain and geo-Louvain. Moreover, they cannot merge into labeled units since the merge operation will decrease the modularity values in the two methods.

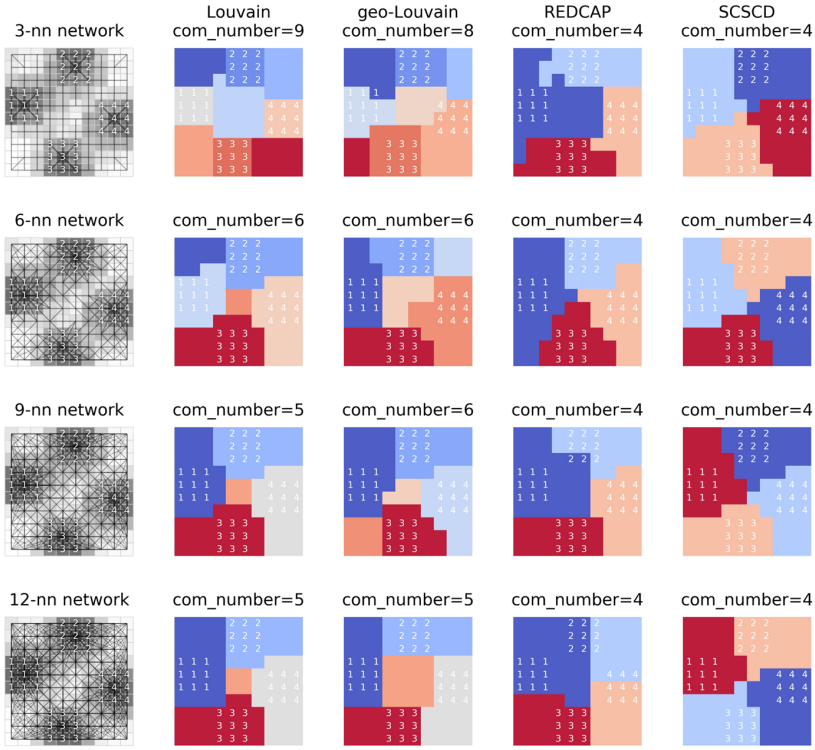


(**a**)

Figure 5. *Cont.*



(**b**)



(**c**)

Figure 5. Community results of the four methods on different knn nets of regional data (**a**) Results of regional data c, (**b**) Results of regional data f, (**c**) Results of regional data h.

The F1 score was used to qualify the accuracy of the four methods. To simplify, only labeled units in the regional data will be used to calculate F1 scores. The classifications of the other units will be omitted. If the community results of one method divide all labeled units correctly, the F1 score of that method obtains its largest value which equals one. Two other indicators, community number, and modularity value, are also calculated here to compare with the modularity-based methods. Moreover, the original modularity of the geo-Louvain results was calculated for comparison with the other methods.

For 100 simulations of each regional data (c), (f), and (h), the statistical analysis of the community number and modularity of the four methods on the four different knn networks are listed in Tables 2 and 3. The best community numbers and largest modularity values in each row are in bold font. Louvain and geo-Louvain always obtained much larger region numbers than the actual values. Moreover, Louvain and geo-Louvain always obtained larger modularity values than REDCAP and SCSCD, except for the 12 nn net of regional data (h). However, the F1 scores of Louvain and geo-Louvain are always smaller than or equal to SCSCD. This indicates that the largest modularity values do not correspond to correct community partitions in the simulated regional data and their knn nets.

Table 2. The average community number of the four methods on regional data (c, f, h) over 100 simulation times.

Regional Data	K Value	Average Community Number over 100 Simulation Times			
		Louvain	Geo-Louvain	REDCAP	SCSCD
(c)	3	6.9	7.1	**2**	**2**
	6	5.6	5.9	**2**	**2**
	9	5.3	5.7	**2**	**2**
	12	4.8	5.3	**2**	**2**
(f)	3	6.8	6.7	**3**	**3**
	6	5.1	5.6	**3**	**3**
	9	5.2	5.4	**3**	**3**
	12	4.7	5.1	**3**	**3**
(h)	3	8.8	7.9	**4**	**4**
	6	5.5	6.5	**4**	**4**
	9	5.0	6.0	**4**	**4**
	12	4.5	5.5	**4**	**4**

Table 3. The average modularity of the four methods on regional data (c, f, h) over 100 simulation times.

Regional Data	K Value	Average Modularity over 100 Simulation Times			
		Louvain	Geo-Louvain	REDCAP	SCSCD
(c)	3	**0.65**	0.63	0.37	0.43
	6	**0.61**	0.60	0.35	0.43
	9	**0.59**	0.58	0.35	0.44
	12	**0.55**	0.54	0.34	0.42
(f)	3	**0.66**	0.64	0.51	0.58
	6	**0.62**	0.61	0.51	0.58
	9	**0.60**	0.59	0.50	0.56
	12	**0.57**	0.56	0.48	0.54
(h)	3	**0.68**	0.66	0.59	0.63
	6	**0.62**	0.60	0.57	**0.62**
	9	**0.60**	0.59	0.56	**0.60**
	12	0.55	0.54	0.53	**0.56**

The F1 score comparison between SCSCD and the other three methods is given in Table 4. Numbers in bold font in each cell indicate the best result of the four methods on

the same regional data's knn net. SCSCD usually obtained 99% to 100% accuracy in all regional data's 6nn, 9nn, and 12nn nets, which performed the best. Louvain performs the second best among the four methods. Specifically, it has obtained the three best results in the 3nn net. However, Louvain produced too many fake communities in its results, as seen in Figure 5. Those fake communities were not involved in the F1 score calculation, but they decreased the accuracy of the Louvain. Therefore, a conclusion can be drawn that the best community results will be obtained when SCSCD with the scale parameter *k* equals or is larger than six. The number six is just the average number of adjacent neighbors of all the units.

Table 4. Number of times when F1 score equals 1 over 100 simulation times.

Regional Data	Number of Times When F1 Score Equals 1 over 100 Simulation Times (SCSCD vs. the Largest Value among Louvain, Geo-Louvain and Redcap)			
	k = 3	**k = 6**	**k = 9**	**k = 12**
(a)	**100**/99(louvain)	**100**/50(louvain)	**100**/45(redcap)	**100**/70(redcap)
(b)	**100**/96(louvain)	**100**/56(redcap)	**100**/69(redcap)	**100**/89(redcap)
(c)	76/**96(louvain)**	**98**/48(louvain)	**100**/39(louvain)	**100**/73(louvain)
(d)	**100**/98(louvain)	**100**/51(louvain)	**100**/39(louvain)	**100**/58(louvain)
(e)	**100**/97(louvain)	**100**/64(redcap)	**100**/74(redcap)	**100**/87(redcap)
(f)	92/**97(louvain)**	**100**/46(louvain)	**100**/45(louvain)	**100**/86(louvain)
(g)	**100**/99(louvain)	**99**/98(louvain)	**99**/95(louvain)	**100**/100(louvain)
(h)	**100/100(louvain)**	**100**/83(louvain)	**100**/94(louvain)	**99**/87(louvain)

Based on the exploratory spatial and quantitative analysis above, the modularity model and its maximizing strategy cannot obtain the correct number and spatial distribution in the eight kinds of simulated regional data and their knn nets. Therefore, they are not fit for the community detection task on the knn nets of regional data. In addition, experimental results on simulated regional data show that the accuracy of the SCSCD method is very high, regardless of whether the scale parameter *k* is small or large. However, as the community results of the 3nn net of regional data (c) and (f) show, a k value that is too small may affect the accuracy of the results. Therefore, a recommended and data-driven k value is all the units' average number of adjacent neighbors.

4.2. Community Number Determination Experiments

The upper experiments of the SCSCD use the actual number of regions as the default value of the community size parameter and obtain very high accuracy. However, most of the time, the best number of communities is unknown. An unsuitable size parameter will lead to quite different community results and affect the accuracy. This section will investigate the ability of the cumulative variance plot to indicate the correct number of communities automatically. To make a comparative analysis, two other indicators, eigengaps and the edge-cut ratio [40], were added to the experiments.

Experiments were still performed on eight kinds of benchmark regional data by using the SCSCD method. For the cumulative variance plot and eigengaps, the community number is set to eight, larger than the actual community number in each regional data. After calculating and ranking each community's degree variance and eigenvalue, the best number of communities will be obtained automatically at the elbow point and the largest eigengap point of each indicator's sequence. For the edge-cut ratio, the ratio value was calculated when the community number was set from three to eight. A reciprocal operation was performed to convert the range of the edge cut ratio to (0, 1). The best number of communities will be achieved when the edge-cut ratio reaches its minimum value.

Figure 6 shows the three indicator results on three kinds of regional data (c), (f), and (h) by using SCSCD. The abscissa value of each colored star represents the best community number for each indicator. In addition, the SCSCD results when the community number equals eight were drawn in each subfigure of Figure 6 to show the results when a large

community number was used. The F1 score was also used to measure the accuracy of SCSCD when the community number changed from two to eight. The red vertical line in each subfigure indicates the real number of communities of the regional data, and the F1 score of SCSCD at that number always equals one.

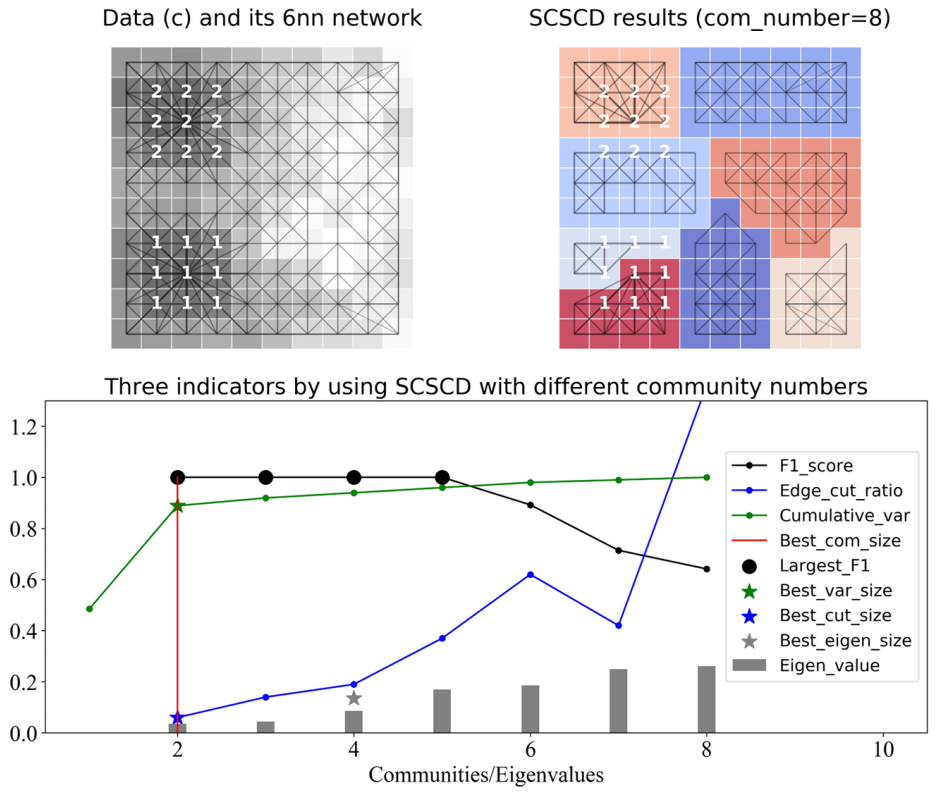
Table 5 shows the times the indicator value equals the corresponding region number. The largest equal times in each row are in bold font.

Table 5. The times the indicator values equal the regional data's region number over 100 simulation times.

Regional Data	The Times When Indicator Value Equals the Corresponding Region Number		
	Edge Cut Ratio	**Eigen-Gap**	**Cumulative Variance Plot**
(a)	**100**	0	96
(b)	**100**	0	**100**
(c)	99	0	**100**
(d)	0	5	**99**
(e)	1	0	**100**
(f)	20	0	**100**
(g)	11	**100**	**100**
(h)	0	1	**91**

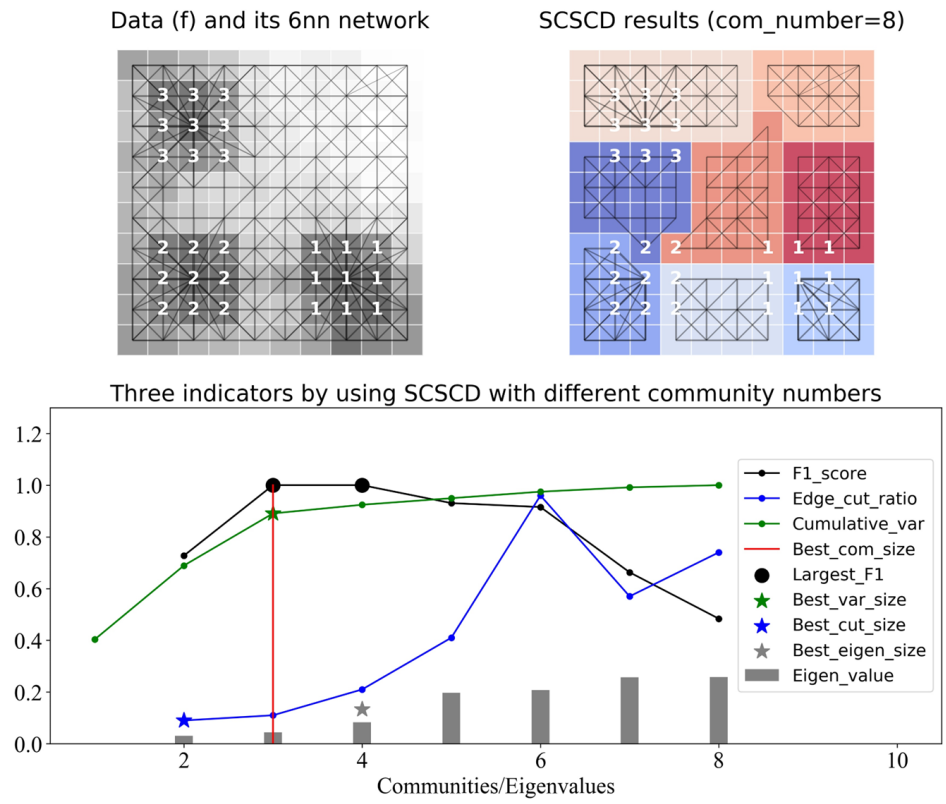
From Figure 6 and Table 5, some conclusions can be drawn:

(1) According to Figure 6, the number of the largest F1 scores in each subfigure decreases when the real number of communities increases. Therefore, finding the best number of communities is important to obtain the correct community results.

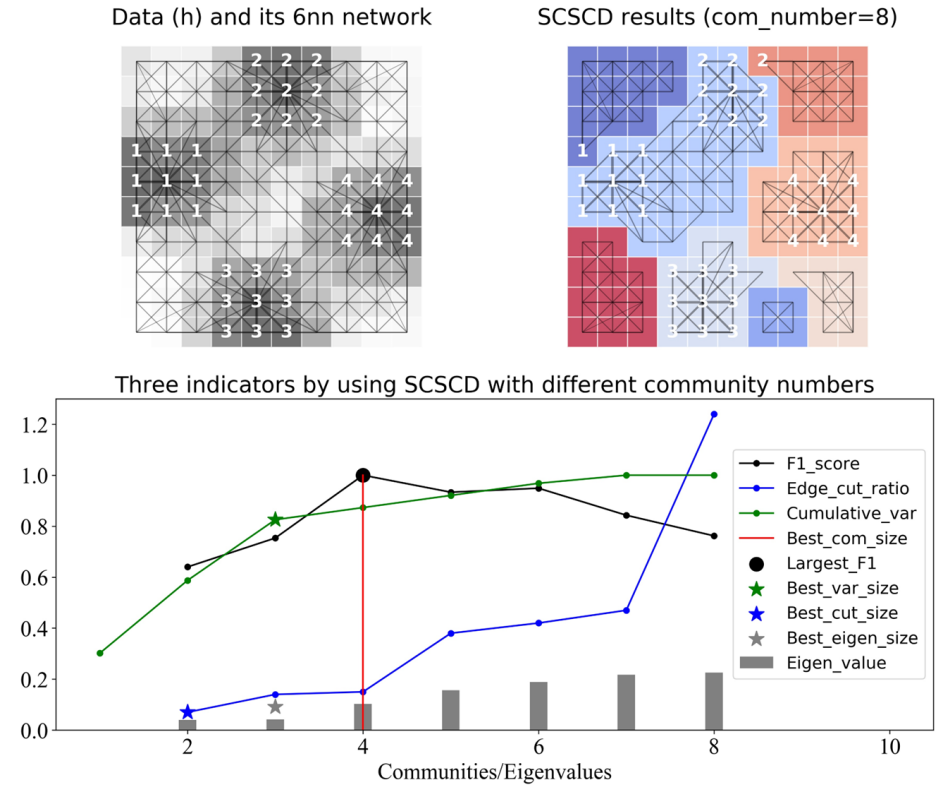


(**a**)

Figure 6. *Cont.*



(**b**)



(**c**)

Figure 6. Community number determination experiments on the 6 nn net of regional data. (**a**) Experimental results on the 6nn net of regional data c, (**b**) Experimental results on the 6nn net of regional data f, (**c**) Experimental results on the 6nn net of regional data h.

(2) The cumulative variance plot performs the best in finding the right number of communities automatically. It achieved 91% to 100% accuracy in all regional data. On the other hand, the edge-cut ratio and eigengap both failed in several regional data. The edge-cut ratio gets the community number equal to two in over 80% of simulated regional data when their region numbers are larger than two. The eigengap always obtained a community number equal to four when the region number is two or three, and it obtained three in 99% of simulated regional data (h).

4.3. Real Data Experiment on a High-Speed Train Network in China

The real data experiment focused on cities in three Chinese provinces: Hubei, Hunan, and Jiangxi. They covered all cities in the urban agglomeration of the middle reaches of the Yangtze River. Furthermore, the network was generated by the number of high-speed trains that go through each city pair. Figure 7 shows the spatial distribution of the high-speed train network. The grey value of each city polygon represents the weighted degree of each city node in the network. As can be seen, three provincial capitals: Wuhan, Changsha, and Nanchang, are the central and hub nodes in the network. Their weighted degrees are all larger than 1000. Then, the SCSCD was used to explore their spatial community structure.

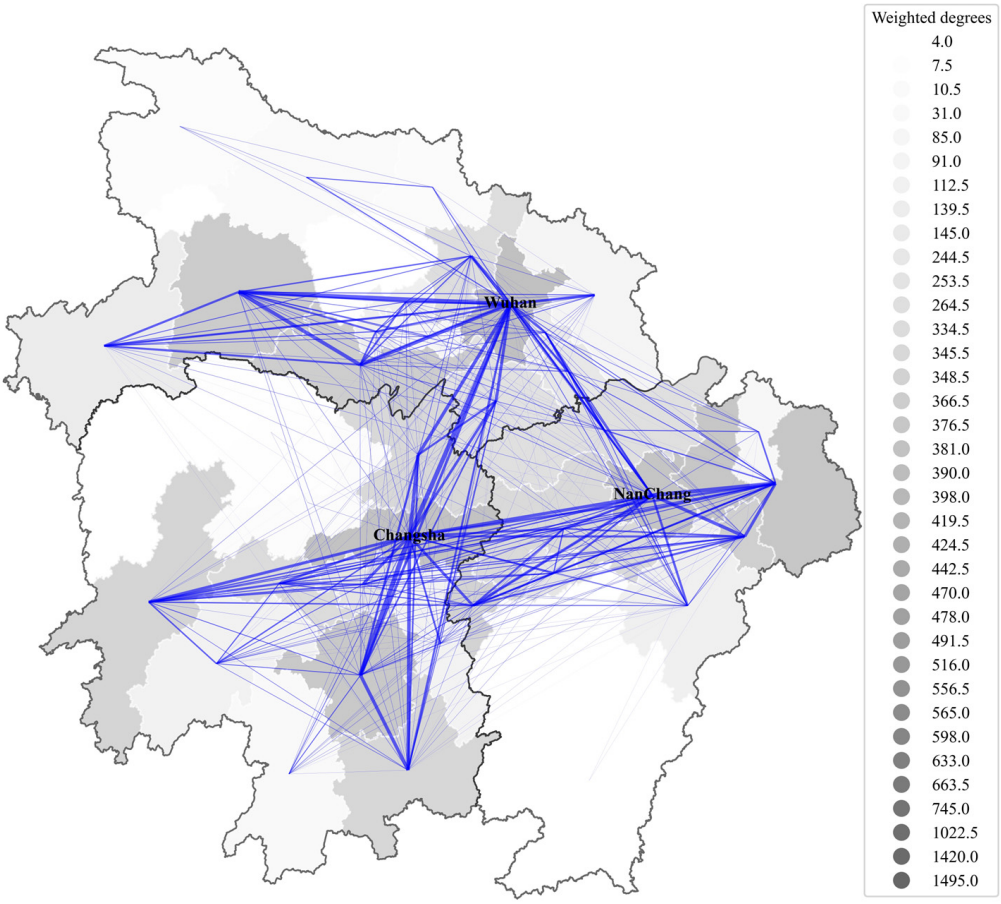


Figure 7. The spatial distribution of the high-speed train network.

The community number was set from two to five, and the community results and three indicators' plots are shown in Figure 8. As can be seen, the best community number of eigengap, edge cut ratio, and cumulative variance plot are one, two, and three, respectively. The eigengap bar fails to find a reasonable community number since the largest gap exists between eigenvalues one and two. In the cumulative variance curve, the largest gap of two clusters' variance exists between clusters 1 and 2. However, the elbow point is identified at cluster 3. So, it indicates that the best community number is three, which

equals the original province number. All three provincial capitals are separated into one of the communities, and most of the other cities belong to their corresponding provincial communities. Two cities that have no high-speed train passing by are colored white, and they do not belong to any community.

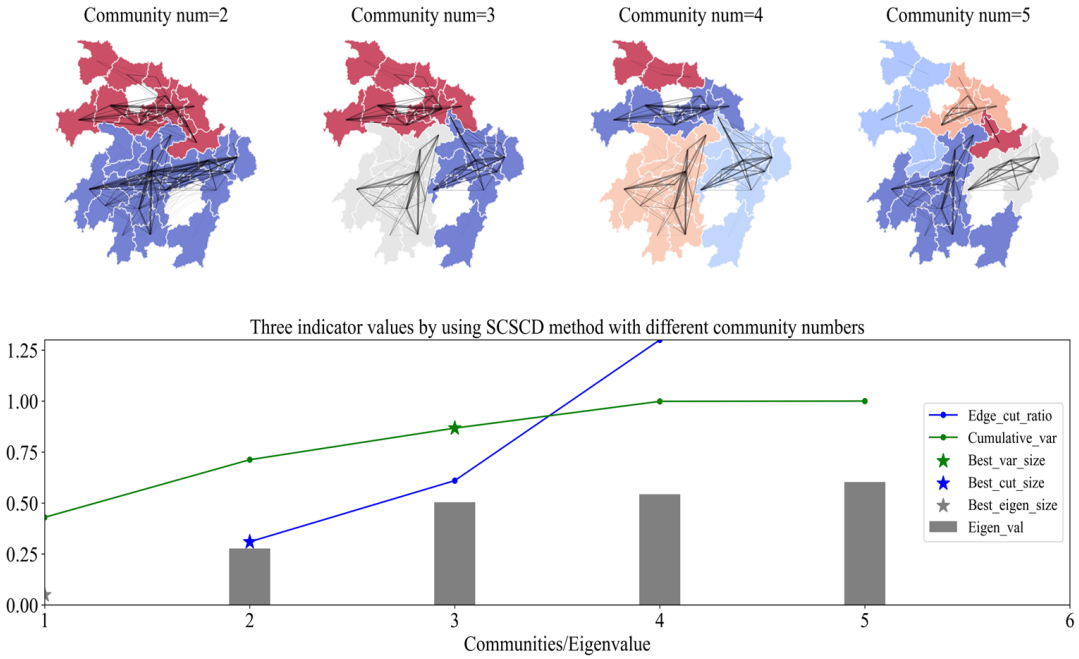


Figure 8. Community results of the high-speed train network and the three indicators' plot by using SCSCD.

Usually, the high-speed train network is constructed by using provincial cities as central points, and these central points connect all the peripheral cities in the same provinces. So, the community results obtained by SCSCD are reasonable.

5. Conclusions and Discussion

The community structure of spatial networks is a very interesting spatial pattern that has been extensively studied. However, due to the lack of appropriate benchmark spatial networks, the accuracy and effectiveness of current community detection methods have yet to be tested sufficiently.

This paper introduced a new SARGM method to generate benchmark regional data and spatial networks. SARGM is a combination of the spatial autoregressive model and the gravity model. The spatial autoregressive model with user-defined central distributions was first used to simulate benchmark regional data with known region numbers and spatial distributions. Then, a gravity model with data-driven and self-adaptive parameters was used to generate interaction networks from the benchmark regional data. The largest connection ratio indicator was also defined to help set the distance decay parameter in the gravity model. In this way, SARGM can easily simulate benchmark spatial networks with known community numbers and spatial distributions.

Moreover, current spatial community detection methods are mainly based on modularity optimization strategies. However, according to experiments on several benchmark regional data and their corresponding spatial networks, modularity-based methods are unsuitable for spatial network community detection. The maximization of modularity will not correspond to the best community structure. Therefore, a spectral clustering-based spatial community detection method (SCSCD) was presented to detect communities from spatial networks. The spectral clustering based SCSCD method has well-established theoretical properties, which allows it to perform better than the other community detection

methods. Experiments on eight kinds of benchmark regional data and their corresponding interaction networks also proved that the SCSCD method performs the best in both accuracy and effectiveness compared to modularity-based methods and the dynamically spatially constrained agglomerative hierarchical clustering method REDCAP.

In addition, the scale parameter and the community number setting of the SCSCD were investigated experimentally. In particular, the scale parameter of the interaction network was converted into the *k* nearest neighbor setting problem. The recommended k value is the average number of neighbors in the original regional data. The cumulative variance plot, widely used to determine the number of principal components in PAC analysis, was used to help decide the best number of communities. Experimental results showed that the cumulative variance contribution plot is highly accurate in all simulated spatial networks. It performed much better than the edge-cut ratio and eigengap.

Finally, an experiment on a high-speed train network in China was also performed to explore the spatial structure of the city network. By setting the recommended parameters, interesting results were obtained, showing the good effectiveness of the SCSCD.

However, this paper only tests limited kinds of benchmark regional data and spatial networks. The accuracy of the SCSCD method on more complex spatial networks still needs further investigation. The effectiveness of the recommended parameter setting of the SCSCD method also needs further verification.

Author Contributions: Conceptualization, You Wan and Xicheng Tan; methodology, You Wan; software, You Wan; validation, Hua Shu; formal analysis, You Wan; investigation, You Wan; resources, You Wan; data curation, You Wan; writing—original draft preparation, You Wan; writing—review and editing, You Wan and Hua Shu; visualization, You Wan. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Plan of China (grant number: 2017YFB0503601); the Hubei Provincial Natural Science Foundation of China (grant number: 2022CFB067); and the National Science Foundation of China (grant number: 41871312, 42271425, 42101431).

Data Availability Statement: All the figures and data (including the experimental results of all the eight kinds of simulated data) that support the findings of this study are openly available in "figshare": https://doi.org/10.6084/m9.figshare.18025034.

Acknowledgments: Thanks to all authors for their contributions, and anonymous reviewers for helpful comments and suggestions. Thanks for the funding support.

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

References

1. Pandit, K. Differentiating between subsystems and typologies in the analysis of migration regions: A US example. *Prof. Geogr.* **1994**, *46*, 331–345. [CrossRef]
2. Barthélemy, M. Spatial networks. *Phys. Rep.* **2011**, *499*, 1–101. [CrossRef]
3. Zaltz Austwick, M.; Brien, O.; Strano, E.; Viana, M. The structure of spatial networks and communities in bicycle sharing systems. *PLoS ONE* **2013**, *8*, e74685. [CrossRef] [PubMed]
4. Liu, X.; Gong, L.; Gong, Y.; Liu, Y. Revealing travel patterns and city structure with taxi trip data. *J. Transp. Geogr.* **2015**, *43*, 78–90. [CrossRef]
5. Guimera, R.; Mossa, S.; Turtschi, A.; Amaral, L.N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7794–7799. [CrossRef]
6. Chen, Y.; Xu, J.; Xu, M. Finding community structure in spatially constrained complex networks. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 889–911. [CrossRef]
7. Wang, C.; Wang, F.; Onega, T. Network optimization approach to delineating health care service areas: Spatially constrained Louvain and Leiden algorithms. *Trans. GIS* **2020**, *25*, 1065–1081. [CrossRef]
8. Wang, C.; Wang, F. GIS-automated delineation of hospital service areas in Florida: From Dartmouth method to network community detection methods. *Ann. GIS* **2022**, *28*, 93–109. [CrossRef]
9. Stegehuis, C.; Van Der Hofstad, R.; Van Leeuwaarden, J.S. Epidemic spreading on complex networks with community structures. *Sci. Rep.* **2016**, *6*, 1–7. [CrossRef]

10. Wang, S.; Gong, M.; Liu, W.; Wu, Y. Preventing epidemic spreading in networks by community detection and memetic algorithm. *Appl. Soft Comput.* **2020**, *89*, 106118. [CrossRef]
11. Duque, J.C.; Church, R.L.; Middleton, R.S. The p-regions problem. *Geogr. Anal.* **2011**, *43*, 104–126. [CrossRef]
12. Duque, J.C.; Anselin, L.; Rey, S.J. The max-p-regions problem. *J. Reg. Sci.* **2012**, *52*, 397–419. [CrossRef]
13. Assuncao, R.M.; Neves, M.C.E.A.; C A Mara, G.; Da Costa Freitas, C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 797–811. [CrossRef]
14. Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 801–823. [CrossRef]
15. Wolf, L.J. Spatially—Encouraged spectral clustering: A technique for blending map typologies and regionalization. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 2356–2373. [CrossRef]
16. Newman, M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef]
17. Wan, Y.; Liu, Y. DASSCAN: A density and adjacency expansion-based spatial structural community detection algorithm for networks. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 159. [CrossRef]
18. Newman, M.E. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **2006**, *74*, 036104. [CrossRef]
19. Newman, M.E. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **2004**, *69*, 066133. [CrossRef] [PubMed]
20. Lancichinetti, A.; Fortunato, S. Community detection algorithms: A comparative analysis. *Phys. Rev. E* **2009**, *80*, 056117. [CrossRef]
21. Schaub, M.T.; Delvenne, J.-C.; Rosvall, M.; Lambiotte, R. The many facets of community detection in complex networks. *Appl. Netw. Sci.* **2017**, *2*, 1–13. [CrossRef] [PubMed]
22. Barber, M.J.; Fischer, M.M.; Scherngell, T. The community structure of R&D cooperation in Europe. Evidence from a social network perspective. *Geogr. Anal.* **2011**, *43*, 415–432.
23. Guo, D. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Trans. Vis. Comput. Graph.* **2009**, *15*, 1041–1048.
24. Expert, P.; Evans, T.S.; Blondel, V.D.; Lambiotte, R. Uncovering space-independent communities in spatial networks. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7663–7668. [CrossRef]
25. Good, B.H.; De Montjoye, Y.-A.; Clauset, A. Performance of modularity maximization in practical contexts. *Phys. Rev. E* **2010**, *81*, 046106. [CrossRef] [PubMed]
26. Lancichinetti, A.; Fortunato, S. Limits of modularity maximization in community detection. *Phys. Rev. E* **2011**, *84*, 066122. [CrossRef]
27. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75–174. [CrossRef]
28. Chakraborty, T.; Dalmia, A.; Mukherjee, A.; Ganguly, N. Metrics for community analysis: A survey. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–37. [CrossRef]
29. Sarzynska, M.; Leicht, E.A.; Chowell, G.; Porter, M.A. Null models for community detection in spatially embedded, temporal networks. *J. Complex Netw.* **2016**, *4*, 363–406. [CrossRef]
30. Kernighan, B.W.; Lin, S. An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.* **1970**, *49*, 291–307. [CrossRef]
31. Suaris, P.R.; Kedem, G. An algorithm for quadrisection and its application to standard cell placement. *IEEE Trans. Circuits Syst.* **1988**, *35*, 294–303. [CrossRef]
32. Newman, M.E.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113. [CrossRef] [PubMed]
33. Zhang, X.; Newman, M.E. Multiway spectral community detection in networks. *Phys. Rev. E* **2015**, *92*, 052808. [CrossRef]
34. Ng, A.Y.; Jordan, M.I.; Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2002**, *14*, 849–856.
35. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
36. Shen, H.-W.; Cheng, X.-Q. Spectral methods for the detection of network community structure: A comparative analysis. *J. Stat. Mech. Theory Exp.* **2010**, *2010*, P10020. [CrossRef]
37. Chauhan, S.; Girvan, M.; Ott, E. Spectral properties of networks with community structure. *Phys. Rev. E* **2009**, *80*, 056114. [CrossRef]
38. Filippone, M.; Camastra, F.; Masulli, F.; Rovetta, S. A survey of kernel and spectral methods for clustering. *Pattern Recognit.* **2008**, *41*, 176–190. [CrossRef]
39. Jia, H.; Ding, S.; Xu, X.; Nie, R. The latest research progress on spectral clustering. *Neural Comput. Appl.* **2014**, *24*, 1477–1486. [CrossRef]
40. Cheng, J.; Li, L.; Leng, M.; Lu, W.; Yao, Y.; Chen, X. A divisive spectral method for network community detection. *J. Stat. Mech. Theory Exp.* **2016**, *2016*, 033403. [CrossRef]
41. Cafieri, S.; Hansen, P.; Liberti, L. Edge ratio and community structure in networks. *Phys. Rev. E* **2010**, *81*, 026105. [CrossRef] [PubMed]
42. Guo, D.; Jin, H.; Gao, P.; Zhu, X. Detecting spatial community structure in movements. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1326–1347. [CrossRef]
43. Ord, K. Estimation methods for models of spatial interaction. *J. Am. Stat. Assoc.* **1975**, *70*, 120–126. [CrossRef]
44. Anselin, L. *Spatial Econometrics: Methods and Models*; Springer Science & Business Media: Cham, Switerland, 1988; Volume 4.

45. Getis, A. Spatial autocorrelation. In *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*; Springer: Cham, Switerland, 2009; pp. 255–278.
46. LeSage, J.P.; Pace, R.K. Spatial econometric models. In *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*; Springer: Cham, Switerland, 2009; pp. 355–376.
47. Wei, R.; Rey, S.; Knaap, E. Efficient regionalization for spatially explicit neighborhood delineation. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 135–151. [CrossRef]
48. Duque, J.C.; Betancourt, A. Python Package Clusterpy. Available online: https://github.com/clusterpy/clusterpy (accessed on 29 April 2023).
49. Alshammari, M.; Stavrakakis, J.; Takatsuka, M. Refining a k -nearest neighbor graph for a computationally efficient spectral clustering. *Pattern Recognit.* **2021**, *114*, 107869. [CrossRef]
50. Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*; pp. 577–584. Available online: https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf (accessed on 1 January 2020).
51. Damiana Costanzo, G. A constrained k-means clustering algorithm for classifying spatial units. *Stat. Methods Appl.* **2001**, *10*, 237–256. [CrossRef]
52. Miranda, L.; Viterbo Filho, J.; Bernardini, F.C. RegK-Means: A clustering algorithm using spatial contiguity constraints for regionalization problems. In Proceedings of the 2017 Brazilian Conference on Intelligent Systems (BRACIS), Uberlandia, Brazil, 2–5 October 2017; pp. 31–36.
53. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]