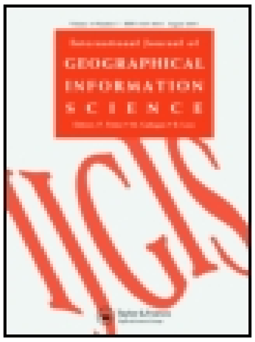

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Two-stage permutation tests for determining homogeneity within a spatial cluster

Qiliang Liu, Wenkai Liu, Jianbo Tang, Min Deng & Yaolin Liu

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2019.1608998

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Two-stage permutation tests for determining homogeneity within a spatial cluster

Qiliang Liu[a,b], Wenkai Liu[b], Jianbo Tang[b], Min Deng[b] and Yaolin Liu[c]

[a]Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring (Central South University), Ministry of Education, Changsha, P.R. China; [b]Department of Geo-informatics, Central South University, Changsha, P.R. China; [c]School of Resource and Environmental Science, Wuhan University, Wuhan, P.R. China

ABSTRACT

The discovery of spatial clusters formed by proximal spatial units with similar non-spatial attribute values plays an important role in spatial data analysis. Although several spatial contiguity-constrained clustering methods are currently available, almost all of them discover clusters in a geographical dataset, even though the dataset has no natural clustering structure. Statistically evaluating the significance of the degree of homogeneity within a single spatial cluster is difficult. To overcome this limitation, this study develops a permutation test approach Specifically, the homogeneity of a spatial cluster is measured based on the local variance and cluster member permutation, and two-stage permutation tests are developed to determine the significance of the degree of homogeneity within each spatial cluster. The proposed permutation tests can be integrated into the existing spatial clustering algorithms to detect homogeneous spatial clusters. The proposed tests are compared with four existing tests (i.e., Park's test, the contiguity-constrained nonparametric analysis of variance (COCOPAN) method, spatial scan statistic, and *q*-statistic) using two simulated and two meteorological datasets. The comparison shows that the proposed two-stage permutation tests are more effective to identify homogeneous spatial clusters and to determine homogeneous clustering structures in practical applications.

ARTICLE HISTORY
Received 12 August 2018
Accepted 15 April 2019

KEYWORDS
Spatial clustering; permutation test; homogeneity; cluster validation

1. Introduction

In real-world scenarios, spatially contiguous regions usually exist in a spatial dataset where observations are homogeneous within each region but not between regions (Dutilleul 2011, Wang *et al.* 2016), e.g., ecological regions (Rueda *et al.* 2010), climate zones (Fovell and Fovell 1993), social-economic units (Openshaw and Rao 1995); and the distributions of land use, land cover, and soil types (Jansen and Gregorio 2002, Goktepe *et al.* 2005). The detection of these homogeneous spatially contiguous regions is useful for understanding local patterns of geographical phenomena and is helpful for removing spurious data variation (Legendre 1987, Wang *et al.* 2012, Deng *et al.* 2018). Currently, the discovery of homogeneous spatially contiguous regions has played

CONTACT Jianbo Tang jianbo.tang@csu.edu.cn

a key role in exploratory spatial data analysis. For example, while delineating ecological regions, important aspects and tradeoffs to be considered involve addressing the homogeneity of the regions with respect to any implicit or explicit classification criteria and the spatial contiguity of the resulting units (Kupfer *et al*. 2012). Identification of homogeneous spatially contiguous regions from global land areas can yield important insights for landscape structural analysis (Hay *et al*. 2003).

Currently, spatial contiguity-constrained clustering is the main technique to delineate homogeneous spatially contiguous regions. This technique enforces spatial contiguity constraint during clustering (Legendre 1987). Three types of spatial contiguity-constrained clustering methods are popular in practical applications: (i) spatial partitioning clustering, e.g., the AZP method (Openshaw and Rao 1995), MaxP method (Duque *et al*. 2012), and GeoSOM (Henriques *et al*. 2012); (ii) spatial hierarchical clustering, e.g., the SKATER method (Assunção *et al*. 2006), REDCAP method (Guo 2008), and MSSC method (Mu and Wang 2008), and (iii) density-based clustering, e.g., spatial scan statistic (Kulldorff 1997), ST-DBSCAN (Birant and Kut 2007), and DBSC (Liu *et al*. 2012). A review of spatial contiguity-constrained clustering methods can be seen in Gordon (1996) and Guo and Wang (2011).

The abovementioned spatial contiguity-constrained clustering methods discover homogeneous spatially contiguous regions subject to balancing the number of clusters and within-cluster homogeneity. While many clusters can impede spatial pattern understanding, a decrease in the number of clusters can degrade the degree of homogeneity within a cluster. When the degree of homogeneity within a cluster becomes insignificant, the clustering operation should stop. Currently, the homogeneity within a single cluster is usually determined by user-specified parameters (e.g., the number of clusters or homogeneity thresholds). However, determining these parameters in practice may be difficult. Almost all the existing clustering methods discover clusters in a geographical dataset, even though the dataset may have no natural clustering structures (Park *et al*. 2009). This is one of the main limitations for users to apply the spatial contiguity-constrained clustering methods in practice (Tan *et al*. 2006). The determination of homogeneity threshold (or cluster number) is highly affected by the scale of the dataset (scale of the sampling framework utilized to produce spatial data) (Liu *et al*. 2015). Furthermore, the identification of suitable homogeneity thresholds can be helpful for determining an appropriate scale of analysis (the scale of variation or phenomenon) and will be further utilized to alleviate the modifiable areal unit problem (MAUP) (Mu and Wang 2008).

In this study, we developed a permutation test approach to determine the significance of the degree of homogeneity within a spatial cluster. The main contributions of this study include (i) developing two-stage permutation tests to statistically evaluate the significance of the degree of homogeneity within each spatial cluster, and (ii) proposing permutation tests to set the stopping criterion for a given spatial contiguity-constrained clustering method.

The rest of this article is organized as follows. Section 2 reviews the related work for determining the homogeneity within spatial clusters. Section 3 presents a new strategy for evaluating the homogeneity within a spatial cluster. Section 4 describes the two-stage permutation tests. Section 5 introduces the method for integrating the two-stage permutation tests into existing spatial contiguity-constrained clustering. Section 6 discusses an experimental evaluation, and finally, Section 7 concludes the study and highlights future work.

2. Related work

Currently, several statistical tests have been developed for assessing the statistical significance of the clustering results. They can be divided into two categories: tests for assessing the statistical significance of the degree of homogeneity within a single cluster and those for assessing the statistical significance of a spatial partition. Although this study mainly focuses on the testing of the statistical significance of a single cluster, we believe that it would be also useful to review tests for assessing the statistical significance of a spatial partition.

2.1. Methods for assessing the statistical significance of the degree of homogeneity within a single cluster

For identifying homogeneous clusters from the dendrograms obtained by hierarchical clustering, two kinds of permutation tests are currently available:

(i) Some permutation tests were proposed by comparing the dendrogram constructed for the observed data with that constructed for randomly permuted data or resampling data. Greenacre and Primicerio (2013) proposed a simple permutation test for determining homogeneous clusters. Under the null hypothesis, they expected that the cluster obtained at a given level of the dendrogram is non-homogeneous (the height computed for the observed data is larger than that computed for the randomly permuted datasets). The Monte Carlo *p*-value is computed as the proportion of the times the heights computed for the permuted datasets are smaller than or equal to those computed for the observed dataset. The distance between the merged clusters has to increase monotonically for this test. However, two merged clusters may be more similar than the pair of clusters merged in the previous step when the spatial constraint is considered. Suzuki and Shimodaira (2006) developed a permutation test based on multi-scale bootstrap resampling. For each cluster discovered from the observed dataset, the bootstrap probability is computed as the proportion of the times that the cluster is discovered from the bootstrapped samples. This test was designed for the analysis of high dimensional DNA microarray data and cannot consider the spatial contiguity constraint of the spatial data. When the test is applied to spatial data, we cannot construct a spatial proximity relationship for the resampled datasets. Consequently, the spatial clustering methods cannot be applied to the resampled datasets.

(ii) Some permutation tests have been developed based on comparing the within-cluster structure of the observed dataset with that of the sample datasets by permuting the cluster membership of objects. Under the null hypothesis, two clusters are expected to be combined to form a homogeneous cluster. Different test statistics have been constructed to characterize the within-cluster structure. For example, Park *et al*. (2009) used the within-cluster variance as the test statistic to measure the clustering quality, and Bruzzese and Vistocco (2015) defined a relative cost measure as the test statistic for measuring the similarity between two clusters. During each aggregation step, a large number of random samples are obtained by permuting the elements between the two clusters and preserving

> the original number of elements in both these clusters. If the test statistic computed for the observed dataset is similar with that computed for the permuted clusters, then the null hypothesis cannot be rejected. In existing methods, the Monte Carlo *p*-value is computed as the proportion of the times that the test statistic calculated for the observed dataset is greater than or equal to that calculated for the permuted clusters. It can be seen that the similarity between the statistics calculated for the observed dataset and the random samples is not well measured.

The spatial scan statistic and its variants (Kulldorff 1997, Pei *et al*. 2011) are also able to assess the statistical significance of a single cluster. However, these spatial scan statistics are designed to identify significant hot or cold spots, and the degree of the homogeneity within a single cluster cannot be evaluated.

2.2. Methods for assessing the statistical significance of a spatial partition

A spatial partition or a stratification of heterogeneity is a partition of a study area, where spatial units are similar within each region (or stratum) but not between regions (Wang *et al*. 2016). The spatial partition is usually obtained by the spatial partitioning clustering method, and each spatial unit should be assigned to a certain region.

Some clustering validity indices that are defined based on within-cluster similarity and between-cluster difference can be used as indicators to select the optimal clustering results or spatial partition (Halkidi *et al*. 2001, Salvador and Chan 2004). However, the significance of these indicators cannot be evaluated statistically, and the selection of optimal clustering results remains difficult. Additionally, these indices are usually not suitable for evaluating arbitrarily shaped spatial clusters and are not robust to noise.

Certain analysis of variance methods designed for spatially autocorrelated data can be used to judge the significance of a spatial partition (Sokal *et al*. 1993), such as Griffith's method (Griffith 1978) and COCOPAN method (Legendre *et al*. 1990). The COCOPAN method can also test the homogeneity of each cluster. This method randomly partitions the study area into contiguous regions, corresponding in size to the observed regions (or clusters). For each cluster, a number of pseudo areas that are approximately similar to an overserved cluster (with the same number of spatial units and approximately the same shape) can be obtained. The sum of squares within a cluster (SSW) is used as the test statistic, and the Monte Carlo *p*-value is computed as the proportion of the times that the SSW computed for the observed dataset is greater than or equal to that computed for the pseudo areas. We argue that the SSW (or variance) is not sensitive enough for measuring the homogeneity within each cluster. In the experimental results given in Section 6, it is seen that the COCOPAN method usually false rejects the null hypothesis.

Recently, Wang *et al*. (2016) proposed a *q*-statistic method for measuring the degree of spatial stratified heterogeneity and for testing its significance. The exact probability density function of the *q*-statistic can be derived; therefore, the *q*-statistic method does not require the time-consuming Monte Carlo simulation. When the number of regions in a spatial partition is fixed, the *q*-statistic can be used to select the best partition from the partitions obtained by different spatial clustering methods.

Table 1. Application conditions of different methods.

Methods	Apply to spatial contiguity-constrained clustering	Test the statistical significance of a single cluster	Test the statistical significance of a spatial partition	Requirement of Monte Carlo simulation
Greenacre's test (Greenacre and Primicerio 2013)	×	√	×	√
Suzuki's test (Suzuki and Shimodaira 2006)	×	√	×	√
Park's test (Park *et al*. 2009)	√	√	×	√
Bruzzese's test (Bruzzese and Vistocco 2015)	×	√	×	√
Spatial scan statistic (Kulldorff 1997)	√	√	×	√
Griffith's method (Griffith 1978)	√	×	√	×
The COCOPAN method (Legendre *et al*. 1990)	√	√	√	√
The *q*-statistic (Wang *et al*. 2016)	√	×	√	×
The two-stage permutation test	√	√	×	√

In Table 1, the application conditions of different methods are summarized. Based on the above analysis, we can conclude that although some reliable methods (e.g., the *q*-statistic) have been proposed for testing the statistical significance of a spatial partition, there is still a lack of a powerful statistical test for assessing the significance of the degree of homogeneity within a spatial cluster. In this study, we present a new strategy for evaluating the homogeneity of a spatial cluster based on the local variance and cluster member permutation and develop two-stage permutation tests for identifying homogenous spatial clusters.

3. Measuring homogeneity within a spatial cluster based on local variance and cluster member permutation

To determine whether a spatial cluster is homogenous, the homogeneity within it should be properly measured. In this study, the spatial contiguity constraints of the spatial units are considered to measure the homogeneity of a cluster. We think that a homogeneous spatial cluster should meet the following two conditions:

(i) Each part of the cluster should be homogeneous, i.e., each unit in a spatial cluster should have a similar non-spatial attribute value with its spatial neighbors in the same cluster.
(ii) Different parts of the cluster should be similar, i.e., each unit in a spatial cluster should have a non-spatial attribute value similar to its non-adjacent units in the same cluster.

To satisfy the above two conditions, in this study, the homogeneity within a spatial cluster is measured from two aspects.

(i) The basic part of a cluster is defined as each spatial unit and its first-order spatial neighbors, and the local variance is used to measure the similarity between a spatial unit O_i and its first-order spatial neighbors. This is represented as follows:

$$LV(O_i) = \frac{1}{n_i - 1} \sum\nolimits_{O_j \in N(O_i)} (A(O_j) - M_i)^2 \tag{1}$$

where $\boldsymbol{N}(O_i)$ represents the first-order spatial neighborhood of O_i (including O_i), n_i is the number of spatial units in $\boldsymbol{N}(O_i)$, $A(O_j)$ represents the non-spatial attribute value of the j^{th} spatial unit in $\boldsymbol{N}(O_i)$, and M_i represents the mean non-spatial attribute value of the spatial units in $\boldsymbol{N}(O_i)$. The local variance is somewhat similar to the local indicator of spatial autocorrelation (e.g. local Geary's C); therefore, the conditional permutation approach can be used to test the statistical significance of $LV(O_i)$ (Anselin 1995). In the mth conditional permutation, the non-attribute value of O_i is fixed, and the remaining non-attribute values in the dataset are randomly permuted. The permuted first-order spatial neighborhood of O_i is represented as $\boldsymbol{N}_m(O_i)$, and the local variance of the units in $\boldsymbol{N}_m(O_i)$ is calculated by using Equation (1), denoted as $LV^m(O_i)$.

(ii) A cluster member permutation strategy is proposed to measure the similarity between a spatial unit O_i and its non-adjacent units in the same cluster. If each unit O_i has a similar non-spatial attribute value with its non-adjacent units in the same cluster $\boldsymbol{C}$. $LV(O_i)$ should also be small when the non-adjacent units of O_i in $\boldsymbol{C}$ are randomly assigned to $\boldsymbol{N}(O_i)$. Based on this observation, the non-spatial attribute value of O_i is fixed, and the non-spatial attribute values of the remaining spatial units in $\boldsymbol{C}$ are permuted H times. It can be seen that after the cluster member permutation, the non-adjacent units of O_i in $\boldsymbol{C}$ will be located in $\boldsymbol{N}(O_i)$. The local variance $LV(O_i)$ is calculated for each permutation; the list of local variances $\boldsymbol{LV}(O_i) = \{LV_1(O_i), LV_2(O_i), \ldots, LV_H(O_i)\}$ can be used to indicate the similarity between O_i and its non-adjacent units in the same cluster.

The datasets in Figure 1 are used to illustrate the measurement of the similarity between a spatial unit and its non-adjacent units in the same cluster. In the homogeneous cluster $\boldsymbol{C}_1$, each unit O_i has similar non-spatial attribute values with its first-order neighbors. In Figure 1(a), the local variance of the units in the red square is 0.17. After cluster member permutation, for each unit O_i, the other non-adjacent units become the first-order neighbors of O_i. If each unit O_i is similar to its non-adjacent units, each element in $\boldsymbol{LV}(O_i)$ should be low. In Figure 1(b), the local variance of the units in the red square is 0.17. Cluster $\boldsymbol{C}_2$ represents an inhomogeneous cluster in which the non-spatial values of neighboring units have small differences (in Figure 1(c), the local variance of the units in the red square is 0.84), and the values of the border units in the cluster are significantly different from those of the other border units in the opposite side. After the cluster member permutation, one can find that the local variance of the

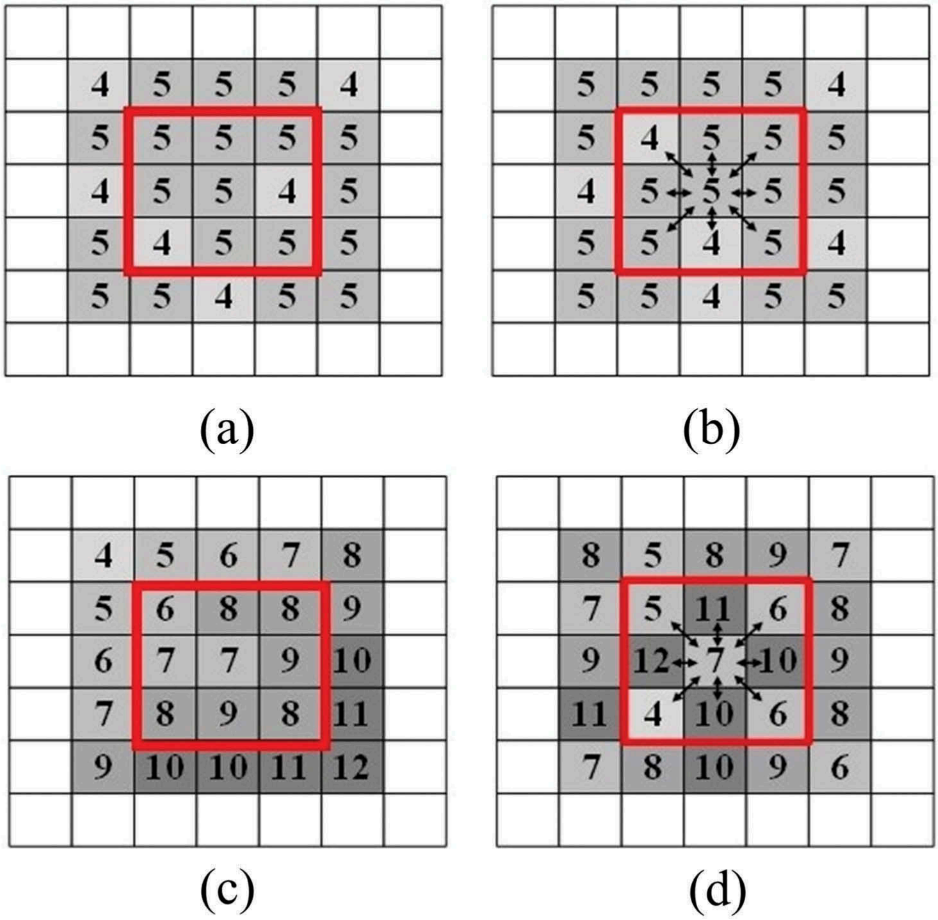

Figure 1. Measuring the homogeneity within spatial clusters (spatial neighbors are identified based on the queen contiguity). (a) Homogeneous cluster C_1, (b) cluster member permutation in C_1, (c) inhomogeneous cluster C_2, and (d) cluster member permutation in C_2.

units in the red square depicted in Figure 1(d) increases remarkably (LV = 7.43). Therefore, it can be concluded that different parts of cluster $\boldsymbol{C}_2$ are dissimilar.

After the homogeneity within a cluster is measured, the performance of the permutation tests will be addressed.

4. Two-stage permutation tests for identifying statistically significant spatial clusters

Based on the strategy for measuring the homogeneity of a cluster introduced in the Section 3, to test whether a cluster $\boldsymbol{C}_i$ is homogeneous, whether each part of the cluster is homogeneous and whether different parts of the cluster are significantly similar need to be tested. To achieve this purpose, the following two permutation tests are developed.

4.1. Testing the significance of the similarity between each unit and its spatial neighbors

Under the null hypothesis, we expect that in cluster $\boldsymbol{C}_i$, there is at least one spatial unit O_i ($N(O_i) \subset C_i$) whose first-order spatial neighbors have different non-spatial attribute values. For each unit O_i in a cluster $\boldsymbol{C}_i$ ($N(O_i) \subset C_i$), the local variance (represented in Equation (1)) is used as the test statistic. The non-spatial attribute values of the dataset are permuted M times, and the Monte Carlo p-value is calculated as the proportion of the times the observed value $LV(O_i)$ is greater than the permuted values $LV^m(O_i)$ (m = 1, 2,..., M). It is given by Equation (2) as follows:

$$p(O_i) = \frac{\sum_{m=1}^{M} I_m}{M} \tag{2}$$

where $p(O_i)$ is the p-value calculated for unit O_i by using the conditional permutation approach; I is an indicator variable. After the mth permutation, if $LV(O_i) > LV^m(O_i)$, then I_m = 1, otherwise, I_m = 0.

When the above permutation test is performed for k spatial units, the multiple and dependent testing problem should be considered. Benjamini and Hochberg (1995) developed a false discovery rate (FDR) approach to control the multiple testing problem for independent test statistics. Benjamini and Yekutieli (2001) further demonstrated that the FDR approach can handle the multiple and dependent testing problem. The experimental analysis by Caldas de Castro and Singer (2006) shows that the FDR approach is the most suitable method for controlling the multiple and dependent tests in the local statistics of spatial association. Therefore, in this study, the FDR approach is used to manage both the multiple and dependent testing problems. Given a significance level α, the adjusted significance level can be obtained as follows:

Step 1: The p-values calculated for k spatial units are ordered in ascending order such as $p(O_1) \le p(O_2) \le \ldots \le p(O_k)$.

Step 2: Starting from $p(O_k)$, the first $p(O_i)$ that satisfies the following equation is found and used as the adjusted significance level α_{adj}:

$$\alpha_{adj} = p(O_i) \le \frac{i}{k} \cdot \alpha \tag{3}$$

The adjusted significance level α_{adj} is used to determine the significance of the p-values calculated by Equation (2). For cluster $\boldsymbol{C_i}$, if the local variance of each unit O_i ($N(O_i) \subset C_i$) is significantly small ($p(O_i) \le \alpha_{adj}$), then the null hypothesis should be rejected, and it can be concluded that each part of the cluster is homogeneous.

Further, the significance of the degree of similarity between a spatial unit O_i and its non-adjacent spatial neighbors in $\boldsymbol{C_i}$ is tested. In this study, we propose another permutation test based on the cluster member permutation.

4.2. Testing the significance of the similarity between each unit and its non-adjacent units

The null hypothesis of the test states that there is at least one spatial unit O_i ($N(O_i) \subset C_i$) whose non-spatial attribute value is different from that of its non-adjacent units in cluster $\boldsymbol{C}_i$. For each unit O_i ($N(O_i) \subset C_i$), the non-spatial attribute value of O_i is fixed, and the non-spatial attribute values of the other spatial units in cluster $\boldsymbol{C_i}$ are permuted H times. The list of local variances $\boldsymbol{LV}(O_i) = \{LV_1(O_i), LV_2(O_i), \ldots, LV_H(O_i)\}$ can be obtained. If unit O_i has a similar non-spatial attribute value with its non-adjacent unit in cluster $\boldsymbol{C_i}$, each local variance $LV_j(O_i)$ ($j = 1, 2, \ldots, H$) should be small. To determine whether $LV_j(O_i)$ is small, we can compare $LV_j(O_i)$ with the permuted values $LV^m(O_i)$ ($m = 1, 2, \ldots, M$) used in Equation (2). As calculated by Equation (3), if the proportion of the times that $LV_j(O_i)$ is greater than $LV^m(O_i)$ is higher than the adjusted significance level α_{adj}, then it can be concluded that $LV_j(O_i)$ is not small. The Monte Carlo p-value can be calculated as the proportion of times when $LV_j(O_i)$ is not small:

$$p_{within}(O_i) = \frac{\sum_{j=1}^{H} I_j}{H} \quad (4)$$

Where $p_{within}(O_i)$ is the p-value of O_i calculated based on cluster member permutation; I is an indicator variable. After the jth cluster member permutation, if $LV_j(O_i)$ is not small, then $I_j = 1$, otherwise, $I_j = 0$.

For each spatial unit O_i ($O_i \in \boldsymbol{C_i}$ and $N(O_i) \subset C_i$), if $p_{within}(O_i)$ is equal to or less than a given significance level β, then we can conclude that the degree of similarity between a spatial unit O_i and its non-adjacent spatial neighbors in $\boldsymbol{C}$ is statistically significant. When the above test is performed for several units in a cluster, the adjusted significance level β_{adj} should be computed by using the FDR approach to alleviate the multiple and dependent testing problem. If each spatial unit O_i ($N(O_i) \subset C_i$) in cluster $\boldsymbol{C_i}$ meets condition $p_{within}(O_i) \leq \beta_{adj}$, then the null hypothesis should be rejected. It can further be concluded that different parts of the cluster are similar.

If each unit O_i in cluster $\boldsymbol{C_i}$ ($N(O_i) \subset C_i$) satisfies the conditions $p(O_i) \leq \alpha_{adj}$ and $p_{within}(O_i) \leq \beta_{adj}$, then cluster $\boldsymbol{C_i}$ is recognized as a homogeneous cluster.

5. Integration of the two-stage permutation tests into spatial contiguity-constrained clustering

The proposed two-stage permutation tests can not only be performed following a given spatial clustering method but also be integrated into existing spatial contiguity-constrained clustering methods.

It is easy to perform the two-stage permutation tests following a given spatial clustering method: after a clustering result $\boldsymbol{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_i, \ldots, \boldsymbol{C}_{max}\}$ is obtained using a certain clustering method, the two permutation tests introduced in Sections 4.1 and 4.2 should be performed for each cluster $\boldsymbol{C}_i$. For each spatial unit O_i ($N(O_i) \subset C_i$) in cluster $\boldsymbol{C_i}$, if $p(O_i) \leq \alpha_{adj}$ and $p_{within}(O_i) \leq \beta_{adj}$ ($p(O_i)$ and $p_{within}(O_i)$ are calculated using Equations (2) and (3), respectively), then $\boldsymbol{C_i}$ should be identified as a homogeneous cluster.

The two-stage permutation tests can also be embedded into existing spatial contiguity-constrained clustering methods and can be used to guide a clustering method to find homogenous clusters. In this study, we will present an example to show how the two-stage permutation tests can be integrated into spatial hierarchical clustering methods because the existing comparative study observed that spatial hierarchical clustering methods usually perform the best while identifying homogeneous spatially contiguous regions (Guo and Wang 2011).

Step 1: Construct the spatial proximity relationship among spatial units based on the topological relationship or graph-based methods (e.g., trimmed Delaunay triangulation method (Liu *et al*. 2012)).

Step 2: Identify within-clusters spatial units. As illustrated in Section 3, each unit in a homogenous cluster should have a similar non-spatial attribute value with its spatial neighbors in the same cluster. We can easily deduce that only the units in a homogeneous spatial neighborhood will be used to construct homogenous clusters. Therefore, we can first use the permutation test given in Section 4.1 to identify within-clusters units:

For each spatial unit O_i in the dataset, calculate the *p*-value $p(O_i)$ using Equation (2). Set the significance level α and calculate the adjusted significance level α_{adj} using the FDR approach. If unit O_i meets the condition$p(O_i) \leq \alpha_{adj}$, the units in $\boldsymbol{N}(\mathrm{O_i})$ will be identified as within-clusters units.

Step 3: Cluster within-cluster units by using a certain spatial hierarchical clustering method. After two clusters $\boldsymbol{C}_j$ and $\boldsymbol{C}_k$ are combined to form a new cluster $\boldsymbol{C}_i$, the permutation test given in Section 4.2 is used to test the significance of the similarity between each spatial unit and its non-adjacent units:

For each spatial unit O_i in $\boldsymbol{C}_i$ $(N(O_i) \subset C_i)$, calculate the *p*-value $p_{within}(O_i)$ using Equation (4). Set the significance level β and calculate the adjusted significance level β_{adj} using the FDR approach. If each unit O_i $(N(O_i) \subset C_i)$ in cluster $\boldsymbol{C_i}$ meets the condition $p_{within}(O_i) \leq \beta_{adj}$, cluster $\boldsymbol{C_i}$ is identified as homogeneous. Otherwise, $\boldsymbol{C}_j$ and $\boldsymbol{C}_k$ should not be combined, and the spatial proximity relationship between $\boldsymbol{C}_j$ and $\boldsymbol{C}_k$ should be deleted.

Step 4: Repeat Step 3 until no homogeneous cluster can be obtained, and output all identified homogeneous clusters.

6. Experimental evaluation

For the proposed two-stage permutation tests, the number of Monte Carlo simulations M and H are both set to 999, and the significance levels α and β are both set to 0.05. For comparison, Park's permutation test (Park *et al*. 2009), the COCOPAN method (Legendre *et al*. 1990), spatial scan statistic (Kulldorff 2011), and *q*-statistic (Wang *et al*. 2016) that were developed for spatial data are also used to determine the significance of the degree of homogeneity within spatial clusters. In the following experiments, the number of Monte Carlo simulation is set to 999 for Park's permutation test, the COCOPAN method, and spatial scan statistic. For all the four comparison methods, the significance level is set to 0.05.

In this study, the spatial contiguity-constrained Ward's method is selected as the spatial clustering method (Guo and Wang 2011). The proposed two-stage permutation test is integrated into the spatial contiguity-constrained Ward's method using the strategy given in Section 5. For Park's test, the COCOPAN method, and *q*-statistic, the cluster discovered in each aggregation step is evaluated. When the obtained cluster is identified as inhomogeneous, the clustering operation is stopped.

Most of the tests (the COCOPAN method, *q*-statistic, and proposed tests) evaluated in this study are designed for univariate spatial data. Therefore, the five tests are only evaluated by using univariate spatial data (two simulated and two real-life datasets) so that the clustering results can be examined visually (Guo 2008). Currently, intuitively identifying clusters using human eyes is still acknowledged as the best benchmark for evaluating clustering results (Baatz and Schäpe 2000, Drăgut *et al*. 2011).

6.1. Experiments on simulated datasets

Two simulated datasets SD_1 and SD_2 are shown in Figure 2. Four clusters are predefined in SD_1. The non-spatial attribute values of the spatial units in each cluster follow a uniform distribution. The non-spatial attribute values of the other units follow a uniform distribution in [1,100]. In SD_2, two-level clusters are designed. At the high-level, four clusters are predefined. At the low-level, 16 clusters can be identified, and the non-spatial attribute values of the spatial units in each cluster follow a normal distribution. To mimic a more realistic geographical pattern (boundaries between clusters are blurred), the two datasets were further spatially smoothed using a Gaussian filter (the size of the filter is [3 3] and the standard deviation is 0.5).

The experimental results of SD_1 are shown in Figure 3. By using the proposed permutation tests, the clustering process is stopped when there are four clusters (shown in Figure 3(a)). All the four predefined clusters can be seen to be well discovered. By using Park's test, the clustering operation is stopped when there are 26 clusters (shown in Figure 3(b)). It can be seen that the predefined clusters are wrongly

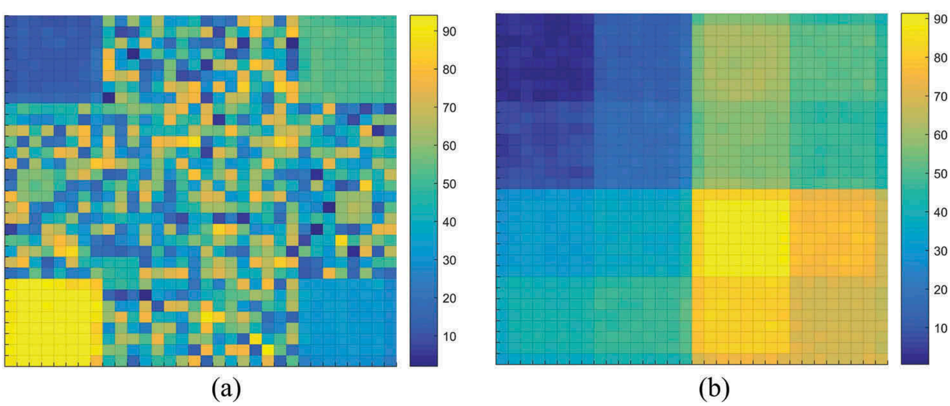

Figure 2. Simulated datasets (spatial neighbors are identified based on queen contiguity). (a) $SD_1(n = 1024)$, (b): $SD_2(n = 1024)$.

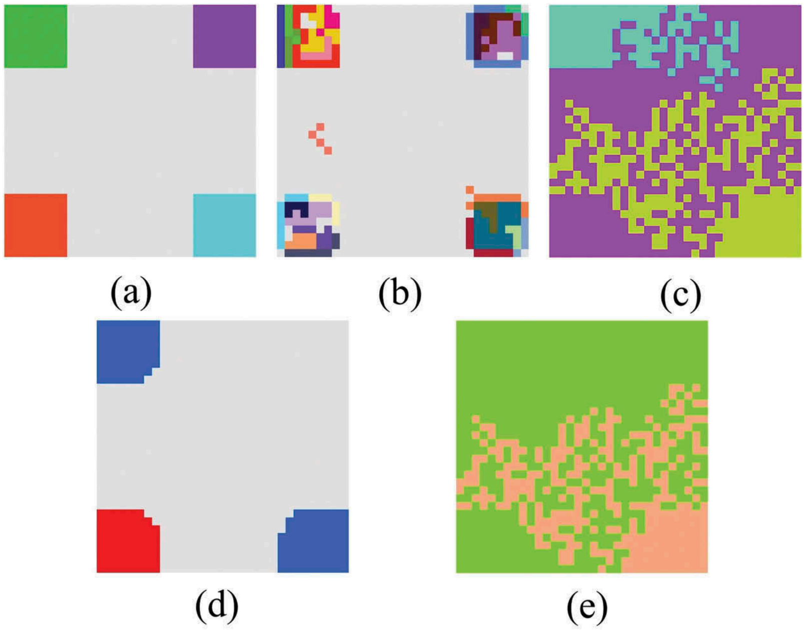

Figure 3. Experimental results of SD_1. (a) Homogeneous clusters identified by the proposed test, (b) homogeneous clusters identified by Park's test, (c) homogeneous clusters identified by the COCOPAN method, (d) clusters identified by spatial scan statistic (blue cluster: cold spot; red cluster: hot spot), and (e) significant spatial partition identified by the q-statistic.

segmented into several small parts. By using the COCOPAN method, the clustering operation is stopped when there are three clusters (shown in Figure 3(c)). A large amount of noise is wrongly clustered. As seen in Figure 3(d), the spatial scan statistic can only discover three clusters (two cold spots and one hot spot) and some noise is wrongly clustered. By using the q-statistic, in case of two clusters, the stratified heterogeneity is still significant (shown in Figure 3(e)). However, the identified clusters are not homogenous.

In Figure 4, the experimental results of SD_2 are shown. By using the proposed permutation tests, the clustering process is stopped when there are four clusters (shown in Figure 4(a)). We can see that all the high-level clusters are discovered correctly. Then, the proposed tests are used to evaluate the homogeneity of clusters in each high-level cluster, and all 16 low-level clusters can be identified satisfactorily (shown in Figure 4(b)). The clusters identified by Park's test are shown in Figure 4(c). The clustering process is stopped when there are 28 small clusters. All the predefined clusters are over-segmented. In Figure 4(d,e), the two-level clusters identified by the COCOPAN method are depicted. It can be seen that all the predefined clusters are under-segmented. In Figure 4(f), the spatial scan statistic can only discover three clusters, and the cluster with medium non-spatial attribute value is missing. In Figure 4(g), although the spatial partition evaluated using the q-statistic is statistically significant, the two identified clusters are not homogenous.

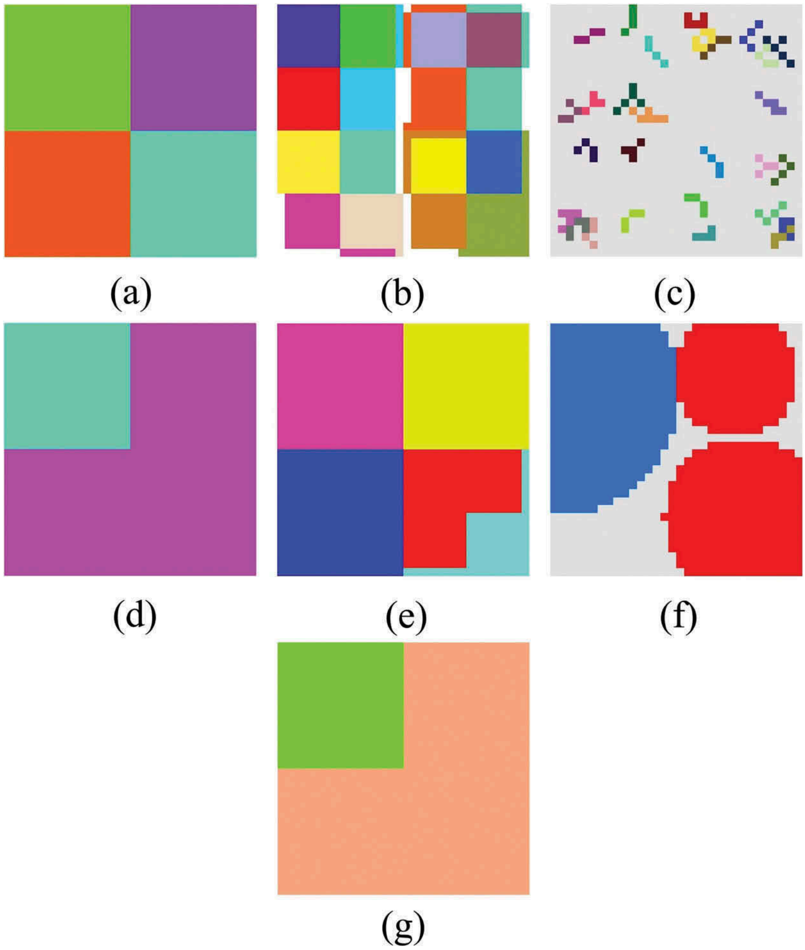

Figure 4. Experimental results of SD_2. (a) High-level homogeneous clusters identified by the proposed tests, (b) low-level homogeneous clusters identified by the proposed tests, (c) homogeneous clusters identified by Park's test, (d) high-level homogeneous clusters identified by the COCOPAN method, (e) low-level homogeneous clusters identified by the COCOPAN method, (f) clusters identified by the spatial scan statistic (blue cluster: cold spot; red cluster: hot spot), and (g) significant spatial partition identified by the *q*-statistic.

6.2. Experiments on climate datasets

The proposed permutation tests and the four existing tests are further evaluated using the annual average temperature and the annual precipitation of 554 stations in mainland China in the year 2009. In Figure 5, the spatial distribution of the stations is displayed and the spatial proximity relationship among these stations is constructed by using the trimmed Delaunay triangulation method (Liu *et al*. 2012). To evaluate the clustering results visually, the spatial distributions of the annual average temperature and annual precipitation are mapped by using the ordinary kriging method after removing the second-order trend, as shown in Figures 6 and 7.

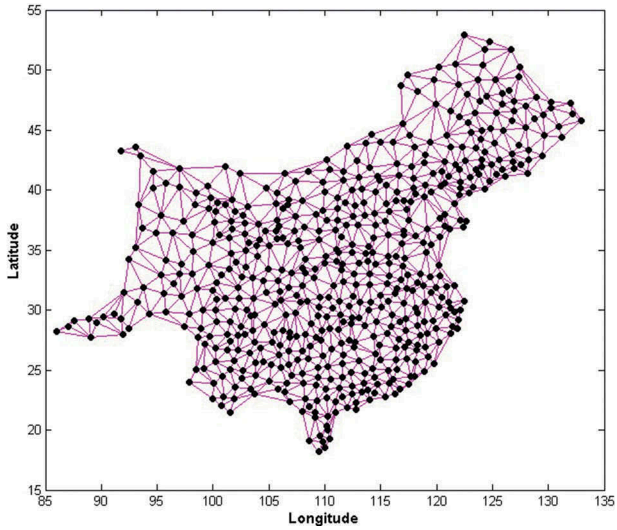

Figure 5. Spatial distributions of the meteorological stations in mainland China.

For the annual average temperature dataset, the clustering process is stopped when there are 12 clusters. The identified homogeneous clusters are displayed in Figure 6(a). The minimum, maximum, and mean temperature values; and the standard deviation of each cluster are listed in Table 2. It can be seen that the mean temperature of the clusters discovered from north to south have distinct differences and a rising trend. The standard deviation within each cluster is usually very low. The visual evaluation based on the interpolation result shows that there is an obvious difference between the spatially adjacent clusters. The clusters are significantly consistent with the main temperature zones in mainland China (Zheng *et al*. 2010). For instance, Cluster 1 represents the cold temperature zone, Clusters 2 and 3 represent the mid-temperature zone, Cluster 6 represents the warm temperature zone, Cluster 8 represents the northern subtropical and mid-subtropical zones, Clusters 10 and 11 represent the south subtropical zone, Cluster 12 represents the tropics, and Cluster 4 represents the plateau temperate zone.

The four comparison tests were also applied to the temperature dataset. Using Park's test, the clustering operation is stopped when there are 30 small clusters (shown in Figure 6(b)). The spatial pattern of temperature cannot be revealed from the clustering result. In Figure 6(c), the COCOPAN method identifies only two homogeneous temperature clusters. In Figure 6(d), the spatial scan statistic can only detect three clusters. In Figure 6(e), by using the *q*-statistic, in the case of two clusters, the spatial partition remains significant. The clustering patterns identified by the COCOPAN method, spatial scan statistic, and *q*-statistic reveal the temperature difference between South and North China; however, the local characteristics of the spatial distribution of temperature in mainland China are not reflected.

For the annual precipitation dataset, the clustering process is stopped when there are 15 clusters. The identified homogeneous clusters are displayed in Figure 7(a). In Table 3, the minimum, maximum, and mean precipitation values and the standard deviation of each cluster are listed. The differences and homogeneity among clusters can be visually

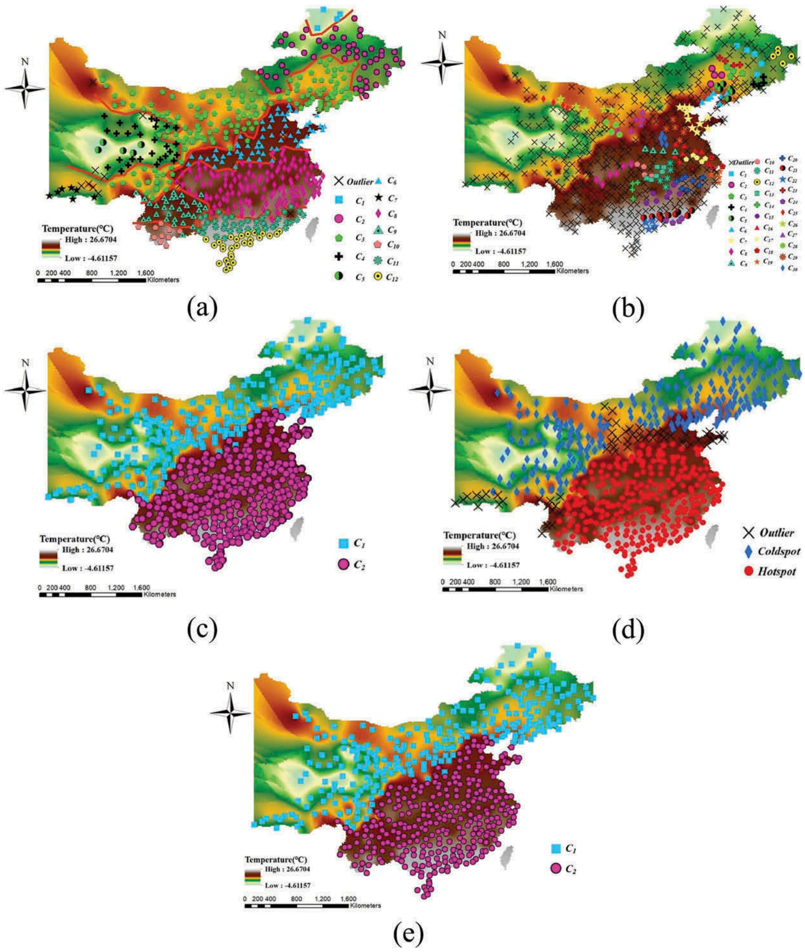

Figure 6. Experimental results of the temperature dataset. (a) Homogeneous temperature clusters identified by the proposed tests, (b) homogeneous temperature clusters identified by Park's test, (c) homogeneous temperature clusters identified by the COCOPAN method, (d) temperature clusters identified by the spatial scan statistic (blue cluster: cold spot; red cluster: hot spot), and (e) significant spatial partition identified by the *q*-statistic.

evaluated using the interpolation result. We can see that the homogeneous clusters with different precipitation values are clearly classified. The local characteristics of the spatial distribution of precipitation in mainland China are well reflected by the clusters. For example, the southern boundary of Cluster 2 (L_1) is highly consistent with the widely accepted boundary between the semi-arid and semi-humid regions in mainland China (400-mm precipitation contour). Lines L_2 and L_3 are highly consistent with the 800- and 1600-mm precipitation contours, respectively.

In Figure 7(b–e), the homogeneous clusters identified by four comparison tests are shown. It is seen that the significant spatial partition identified by the *q*-statistic and the two clusters detected by the spatial scan statistic reflect the difference of precipitation between South and North China. Although the homogeneous clusters identified by the COCOPAN method can reveal the local spatial pattern of

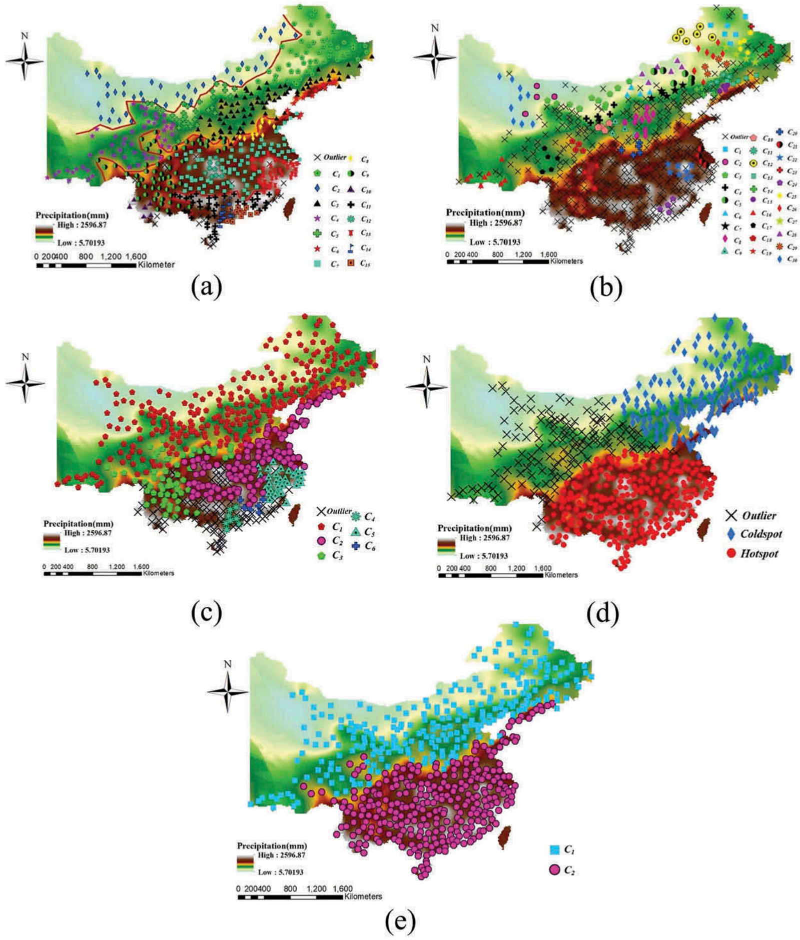

Figure 7. Experimental results of the precipitation dataset. (a) Homogeneous precipitation clusters identified by the proposed tests, (b) homogeneous precipitation clusters identified by Park's test, (c) homogeneous precipitation clusters identified by the COCOPAN method, (d) precipitation clusters identified by the spatial scan statistic, and (e) significant spatial partition identified by the *q*-statistic.

Table 2. The minimum, maximum, and mean temperature values and standard deviation of each cluster (unit: °C).

Cluster	Min	Max	Mean	Standard deviation	Cluster	Min	Max	Mean	Standard deviation
C_1	−3.08	−1.08	−2.08	1.05	C_7	0.79	7.93	4.60	2.90
C_2	−0.65	6.7	3.83	1.68	C_8	9.11	19.72	17.61	1.49
C_3	2.75	13.33	8.51	1.93	C_9	11.42	20.81	14.77	2.05
C_4	−1.74	6.48	2.74	2.02	C_{10}	17.96	23.47	20.22	1.9
C_5	−4.68	−1.08	−2.9	1.32	C_{11}	19.12	22.58	21.05	0.93
C_6	12.26	16.41	14.35	1.06	C_{12}	22.75	26.78	23.63	0.99

precipitation in South China, the precipitation differences in northern China cannot be described. Park's test can only identify a large number of small clusters; however, these clusters are not useful for describing the spatial distribution pattern of precipitation in mainland China.

Table 3. The minimum, maximum, and mean precipitation values and standard deviation of each cluster (unit: mm).

Cluster	Min	Max	Mean	Standard deviation	Cluster	Min	Max	Mean	Standard deviation
C_1	194.4	512.3	344.6	77.83	C_9	626.6	1222.6	859	135.8
C_2	17.5	342.4	145.7	72.780	C_{10}	1322.5	1784.8	1530.2	188.6
C_3	337.8	806.9	568.7	104.99	C_{11}	841.8	2027.7	1458.9	224.7
C_4	286.6	701.7	476.6	91.93	C_{12}	1389.3	1844.6	1533.1	157.1
C_5	214.7	517.3	332.8	87.09	C_{13}	1109.6	1513	1337.8	102.3
C_6	746.1	1234.1	933.7	133.98	C_{14}	1069.6	1252.4	1163.2	54.4
C_7	884.9	1489.1	1131.2	130.99	C_{15}	1370.3	2033	1737.6	227.2
C_8	779.1	1014.9	877.2	91.23					

6.3. Discussion

Experimental results of simulated and real-life datasets show that the proposed two-stage permutation tests are more efficient for identifying homogeneous spatial clusters. Theoretical comparisons of the methods tested in this study are given as follows:

(i) Although Park's test was designed to evaluate the significance of the degree of homogeneity within a single spatial cluster, determining whether the difference between the statistics calculated for the observed dataset and the random samples is small enough is difficult. From the experimental results shown in Figures 3(b) and 4(c), one can see that a slight variation within a homogeneous cluster may be over-segmented.

(ii) For the COCOPAN method, the SSW is not sensitive enough for measuring homogeneity within each cluster. When some noise is merged into a homogenous cluster, the value of the variance of this inhomogeneous cluster will be still smaller than a pseudo area or cluster constructed by the COCOPAN method. Figures 3(c), 4(d,e), 6(c) and 7(c) show that some inhomogeneous clusters were wrongly reported.

(iii) The spatial scan statistic is suitable for identifying significant circular or elliptical hot (or cold) spots (shown in Figures 3(d) and 4(f)). The shapes of discovered clusters are limited by the circular or elliptical scanning windows. The spatial scan statistic is not designed for detecting homogeneous clusters; therefore, some clusters with medium non-spatial attribute values are usually missed.

(iv) Although the q-statistic for assessing the statistical significance of a spatial partition is compared in this study, it does not indicate that the proposed test outperforms the q-statistic. The experimental results only indicate that the tests for assessing the statistical significance of a spatial partition cannot be directly used for testing the statistical significance of the degree of homogeneity within a single cluster. One reason is that the q-statistic is defined on stratum but not on spatial continuous regions. We think that the proposed tests and the q-statistic complement each other for different application purposes rather than being competing methods.

(v) For the proposed two-stage permutation tests, we find that homogenous clusters with different shapes can be detected easily and correctly by integrating the proposed tests into certain spatial clustering methods because the homogeneity of a spatial cluster can be well measured by using local variance and cluster member permutation strategy.

The time complexity of the five methods can be analyzed as follows. The time complexity of the spatial scan statistic is $O(N^4)$, where N is the number of units in the dataset. Indeed, the spatial scan statistic cannot be applied to large datasets. The time complexity of Park's test is $O(HN^2)$, where H is the time of cluster member permutation, whereas that of the COCOPAN method is $O(RNlogN)$, where R is the number of Monte Carlo simulations. The time complexity of the q-statistic is $O(N)$. For the proposed tests, the time complexity of the first stage test in Section 4.1 is $O(MN)$ and the time complexity of the second stage test in Section 4.2 is $O(HN^2)$. Thus, the total time complexity of the proposed tests is $O(MN+ HN^2)$.

7. Conclusions

In this study, the homogeneity within a cluster is measured based on the local variance and cluster member permutation, and two-stage permutation tests are proposed to statistically evaluate the significance of the degree of homogeneity within spatial clusters. The proposed tests can be integrated into existing spatial contiguity-constrained clustering methods to guide them to find homogenous clusters. Experiments on both simulated and real-life datasets show that the proposed permutation tests are more efficient than the existing statistical methods for identifying homogeneous clusters. The proposed permutation tests have been applied to the annual average temperature and annual precipitation datasets in mainland China, and the local characteristics of the spatial distribution of the temperature and precipitation can be reflected satisfactorily by the identified homogeneous clusters.

It should be noted that the two-stage permutation tests are only designed for testing the statistical significance of a single cluster. For assessing the statistical significance of a spatial partition, users should select the q-statistic or the COCOPAN method. Two limitations of the proposed tests should be considered in the future. First, the Monte Carlo simulation used for estimating the empirical distribution of the test statistic is computationally prohibitive, and the computational efficiency needs to be further improved. Second, the two-stage permutation tests are designed only for univariate spatial data, and they should be further extended to a multivariate context.

Acknowledgments

The authors gratefully acknowledge the comments from the editor and the reviewers.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the The Innovation-Driven Project of Central South University [2018CX015]; National Key Research and Development Foundation of China [2017YFB0503601]; National Science Foundation of China (NSFC) [41601410 and 41730105].

Notes on contributors

Qiliang Liu received the Ph.D. degree in geographical information science from The Hong Kong Polytechnic University in 2015. He is currently an associate professor at Central South University, Hunan, China. His research interests focus on multi-scale spatio-temporal data mining and spatio-temporal statistics. He has published more than 30 peer-reviewed journal articles in these areas.

Wenkai Liu is currently a Ph.D. candidate at Central South University and his research interests focus on spatio-temporal clustering and association rule mining.

Jianbo Tang received the Ph.D. degree from Central South University and his research interests focus on spatio-temporal clustering and road data updating.

Min Deng is currently a professor at Central South University and the associate dean of School of Geosciences and info-physics. His research interests are map generalization, spatio-temporal data analysis and mining.

Yaolin Liu is currently a Professor at Wuhan University. His research interests include Geographic Information Science, geographic data mining, and spatial analysis and decision making.

References

Anselin, L., 1995. Local indicators of spatial association-LISA. *Geographical Analysis*, 27 (2), 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Arbelaiz, O., *et al*., 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46 (1), 243–256. doi:10.1016/j.patcog.2012.07.021

Assunção, R.M., *et al*., 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20 (7), 797–811. doi:10.1080/13658810600665111

Baatz, M. and Schäpe, A., 2000. Multiresolution segmentation- an optimization approach for high quality multi-scale image segmentation. *In*: J. Strobl, T. Blaschke, and G. Griesebner, eds. *Angewandte geographische informatiuonseverarbeitung*. Heidelberg, Germany: Wichmann-Verlag, 12–23.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57 (1), 289–300.

Benjamini, Y. and Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 (4), 1165–1188.

Birant, D. and Kut, A., 2007. ST-DBSCAN: an algorithm for clustering spatial-temporal data. *Data & Knowledge Discovery*, 60 (1), 208–221.

Bruzzese, D. and Vistocco, D., 2015. DESPOTA: dEndrogram slicing through a pemutation test approach. *Journal of Classification*, 32 (2), 285–304. doi:10.1007/s00357-015-9179-x

Caldas de Castro, M. and Singer, B., 2006. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis*, 38 (2), 180–208. doi:10.1111/gean.2006.38.issue-2

Deng, M., *et al*., 2018. Heterogeneous space–time artificial neural networks for space–time series prediction. *Transactions in GIS*, 22 (1), 183–201. doi:10.1111/tgis.2018.22.issue-1

Drăgut, L., Eisank, C., and Strasser, T., 2011. Local variance for multi-scale analysis in geomorphometry. *Geomorphology*, 130 (3–4), 162–172. doi:10.1016/j.geomorph.2011.03.011

Duque, J.C., Anselin, L., and Rey, S.J., 2012. The max-p-regions problem. *Journal of Regional Science*, 52 (3), 397–419. doi:10.1111/jors.2012.52.issue-3
Dutilleul, P.R.L., 2011. *Spatio-temporal heterogeneity: concepts and analysis*. UK: Cambridge University Press.
Fovell, R.G. and Fovell, M., 1993. Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate*, 6 (11), 2103–2135. doi:10.1175/1520-0442(1993)006<2103:CZOTCU>2.0.CO;2
Goktepe, A.B., Altun, S., and Sezer, A., 2005. Soil clustering by using Fuzzy c-means algorithm. *Advances in Engineering Software*, 36 (10), 691–698. doi:10.1016/j.advengsoft.2005.01.008
Gordon, A.D., 1996. A survey of constrained classification. *Computational Statistics and Data Analysis*, 21 (1), 17–29. doi:10.1016/0167-9473(95)00005-4
Greenacre, M. and Primicerio, R., 2013. *Multivariate analysis of ecological data*. Bilbao: Fundación BBVA.
Griffith, D.A., 1978. A spatially adjusted ANOVA model. *Geographical Analysis*, 10 (3), 296–301. doi:10.1111/j.1538-4632.1978.tb00661.x
Guo, D., 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22 (7), 801–823. doi:10.1080/13658810701674970
Guo, D. and Wang, H., 2011. Automatic region building for spatial analysis. *Transactions in GIS*, 15, 25–45. doi:10.1111/j.1467-9671.2011.01269.x
Halkidi, M., Batistakis, Y., and Vazirgiannis, M., 2001. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17 (2–3), 107–145. doi:10.1023/A:1012801612483
Halkidi, M., Batistakis, Y., and Vazirgiannis, M., 2002. Clustering validity checking methods: part II. *ACM Sigmod Record*, 31 (3), 19–27. doi:10.1145/601858
Hay, G.J., *et al*., 2003. A comparison of three image-object methods for multi-scale analysis of landscape structure. *ISPRS Journal of Photogrammetry & Remote Sensing*, 57 (5–6), 327–345. doi:10.1016/S0924-2716(02)00162-4
Henriques, R., Bacao, F., and Lobo, V., 2012. Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, 36 (3), 218–232. doi:10.1016/j.compenvurbsys.2011.11.003
Jansen, L.J.M. and Gregorio, A.D., 2002. Parametric land cover and land-use classifications as tools for environmental change detection. *Agriculture, Ecosystems & Environment*, 91 (1–3), 89–100.
Kulldorff, M., 1997. A spatial scan statistic. *Communications in Statistic-Theory and Methods*, 26 (6), 1481–1496. doi:10.1080/03610929708831995
Kulldorff, M., 2011. *SaTScan v9.0: software for the spatial and space-time scan statistics*. Rockville, MD: Information Management Services, Inc.
Kupfer, J., Gao, P., and Guo, D., 2012. Regionalization of forest pattern metrics for the continental United States using contiguity constrained clustering and partitioning. *Ecological Informatics*, 9, 11–18. doi:10.1016/j.ecoinf.2012.02.001
Legendre, P., 1987. Constrained clustering. *In*: P. Legendre and L. Legendre, eds. *Developments in numerical ecology*. AFI series, vol G 14, 289–307. Berlin, Germany: Springer.
Legendre, P., *et al*., 1990. Approximate analysis of variance of spatially autocorrelated regional data. *Journal of Classification*, 7 (1), 53–75. doi:10.1007/BF01889703
Liu, Q.L., *et al*., 2012. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences*, 46, 296–309. doi:10.1016/j.cageo.2011.12.017
Liu, Q.L., *et al*., 2015. Modeling the effect of scale on clustering of spatial points. *Computers, Environment and Urban Systems*, 52, 81–92. doi:10.1016/j.compenvurbsys.2015.03.006
Mu, L. and Wang, F.H., 2008. A scale-space clustering method: mitigating the effect of scale in the analysis of zone-based data. *Annals of the Association of American Geographers*, 98 (1), 85–101. doi:10.1080/00045600701734224
Openshaw, S. and Rao, L., 1995. Algorithms for re-engineering 1991 census geography. *Environment and Planning A*, 27 (3), 425–446.

Park, P.J., *et al.*, 2009. A permutation test for determining significance of clusters with applications to spatial and gene expression data. *Computational Statistics and Data Analysis*, 53 (12), 4290–4300.

Pei, T., *et al.*, 2011. Detecting arbitrarily shaped clusters using ant colony optimization. *International Journal of Geographical Information Science*, 25 (10), 1575–1595. doi:10.1080/13658816.2010.533674

Pham, D.L., 2001. Spatial models for fuzzy clustering. *Computer Vision and Image Understanding*, 84 (2), 285–297. doi:10.1006/cviu.2001.0951

Rueda, M., Rodriguez, M.A., and Hawkins, B.A., 2010. Towards a biogeographic regionalization of the European biota. *Journal of Biogeography*, 37 (11), 2067–2076. doi:10.1111/j.1365-2699.2010.02388.x

Salvador, S. and Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *In*: *Proceedings of the Sixteenth IEEE International Conference on Tools with Artificial Intelligence. Institute of Electrical and Electronics Engineers*. Piscataway, NJ, 576–584.

Sokal, R., *et al.*, 1993. Testing for regional differences in means: distinguishing inherent from spurious spatial autocorrelation by restricted randomization. *Geographical Analysis*, 25 (3), 199–210. doi:10.1111/gean.1993.25.issue-3

Suzuki, R. and Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22 (12), 1540–1542. doi:10.1093/bioinformatics/btl117

Tan, P.-N., Steinbach, M., and Kumar, V., 2006. *Introduction to data mining*. London: Pearson Education, Inc.

Wang, F., Guo, D., and Mclafferty, S., 2012. Constructing geographical areas for cancer data analysis: a case study on late-stage breast cancer risk in Illinois. *Applied Geography*, 35 (1–2), 1–11. doi:10.1016/j.apgeog.2012.04.005

Wang, J.F., Zhang, T.L., and Fu, B.J., 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67, 250–256. doi:10.1016/j.ecolind.2016.02.052

Zheng, J.Y., Yin, Y.H., and Li, B.Y., 2010. A new scheme for climate regionalization in China. *Acta Geographica Sinica*, 65 (1), 3–12.