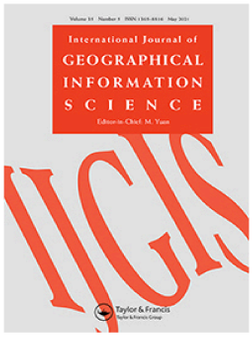

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Discovering regions of anomalous spatial co-locations

Jiannan Cai, Min Deng, Yiwen Guo, Yiqun Xie & Shashi Shekhar

To cite this article: Jiannan Cai, Min Deng, Yiwen Guo, Yiqun Xie & Shashi Shekhar (2021) Discovering regions of anomalous spatial co-locations, International Journal of Geographical Information Science, 35:5, 974-998, DOI: 10.1080/13658816.2020.1830998

To link to this article: https://doi.org/10.1080/13658816.2020.1830998

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2021, VOL. 35, NO. 5, 974–998
https://doi.org/10.1080/13658816.2020.1830998

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Discovering regions of anomalous spatial co-locations

Jiannan Cai[a,b,c], Min Deng[a], Yiwen Guo[a], Yiqun Xie[b,d,e] and Shashi Shekhar[b]

[a]Department of Geo-informatics, Central South University, Changsha, China; [b]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA; [c]Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong, China; [d]Center for Geospatial Information Science, University of Maryland, College Park, MD, USA; [e]Department of Geographical Sciences, University of Maryland, College Park, MD, USA

ABSTRACT

Regions of anomalous spatial co-locations (ROASCs) are regions where co-locations between two different features are significantly stronger or weaker than expected. ROASC discovery can provide useful insights for studying unexpected spatial associations at regional scales. The main challenges are that the ROASCs are spatially arbitrary in geographic shape and the distributions of spatial features are unknown *a priori*. To avoid restrictive assumptions regarding the distribution of data, we propose a distribution-free method for discovering arbitrarily shaped ROASCs. First, we present a multidirectional optimization method to adaptively identify the candidate ROASCs, whose sizes and shapes are fully endogenized. Furthermore, the validity of the candidates is evaluated through significance tests under the null hypothesis that the expected spatial co-locations between two features occur consistently across space. To effectively model the null hypothesis, we develop a bivariate pattern reconstruction method by reconstructing the spatial auto- and cross-correlation structures observed in the data. Synthetic experiments and a case study conducted using Shanghai taxi datasets demonstrate the advantages of our method, in terms of effectiveness, over an available alternative method.

ARTICLE HISTORY
Received 4 October 2019
Accepted 27 September 2020

KEYWORDS
Spatial data mining; anomalous spatial co-locations; region detection; pattern reconstruction; multiple significance tests

1. Introduction

Geographers have long considered and quantified spatial relationships between different features (He *et al*. 2020). One of the most fundamental relationships is the spatial association between locations of different features, denoted by the term spatial co-location, which refers to instances of different features that co-occur in close spatial proximity (Huang *et al*. 2004, Leslie and Kronenfeld 2011, Zhou *et al*. 2019). Spatial co-locations can be commonly observed in real life. For example, in transportation, taxi supply tends to co-locate with trip demand (Wang *et al*. 2013, Pei *et al*. 2015); in ecology, emerald ash borers usually co-exist with ash trees (Xie *et al*. 2018).

Because of spatial heterogeneity, spatial co-locations between different features are usually inconsistent across a geographical space (Deng *et al*. 2017). Thus, regions of anomalous spatial co-locations (ROASCs) in which the spatial co-locations are significantly

CONTACT Min Deng dengmin@csu.edu.cn

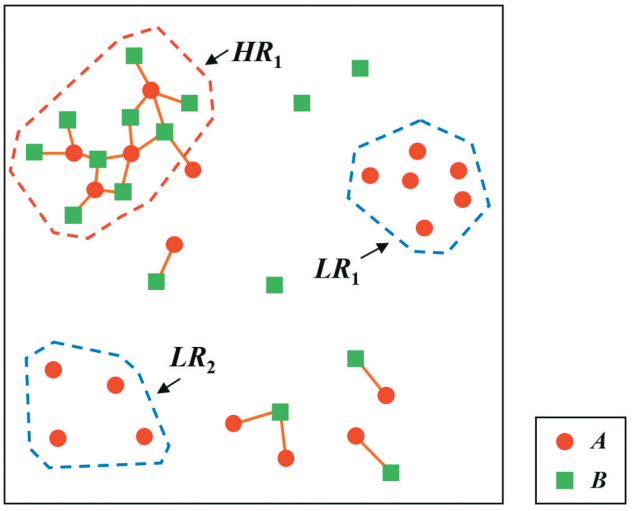

Figure 1. Example of ROASCs between features *A* and *B*.

stronger or weaker than the expected level exhibited by the dataset as a whole may occur. Consider the dataset presented in Figure 1 as an example. The lines represent the neighbor relation between features *A* and *B*. On average, each instance of *A* co-occurs with one instance of *B* across the entire study area. However, in region HR_1, *A* co-occurs with three or four instances of *B*, and in regions LR_1 and LR_2, *A* does not co-occur with any instances of *B*. Thus, HR_1, LR_1 and LR_2 may be considered as ROASCs. ROASC detection can reveal unforeseen spatial associations at regional scales; thus, it is of substantial interest to domain experts. For example, the presence of ROASCs between the supply of and demand for taxi services indicates that there is a mismatch between the two; furthermore, it can provide insights to help cities improve their transportation systems (Tang *et al*. 2019).

Previous research on analyzing spatial co-locations at regional scales generally falls into two categories, namely regional co-location discovery and spatial cross-outlier detection. The former aims to discover regions where spatial co-locations are prevalent, whereas the latter focuses on identifying anomalous instances of a feature with respect to co-occurring instances of another feature. These approaches, generally, cannot directly support the detection of ROASCs targeted in this study. In addition, with regard to making decisions, previous methods (Papadimitriou and Faloutsos 2003, Wang *et al*. 2013) usually necessitate *a priori* restrictive assumptions regarding the distribution of features, which may cause false or missing detections if assumptions are inconsistent with the underlying distribution. Consequently, this paper proposes a method without restrictive assumptions regarding the distribution of data, i.e. a distribution-free method. This method can endogenously discover arbitrarily shaped ROASCs that are exhibited by the data; it can also effectively establish the statistical significance of results by reconstructing spatial auto- and cross-correlation structures observed in the data.

The remainder of this paper is organized as follows: Section 2 reviews related work on detecting the spatial co-locations at regional scales. Section 3 outlines the proposed distribution-free strategy for ROASC detection. Section 4 details the techniques involved

in our method. Section 5 presents the synthetic experiments and a case study of Shanghai taxi datasets to compare and evaluate the performance of our method with that of an existing method. Section 6 offers closing comments on the advantages and limitations of the research.

2. Literature review

2.1. Regional co-location discovery

Initially, the problem of discovering spatial co-locations was defined for mining subsets of spatial features whose instances were frequently located together across an entire study area (Shekhar and Huang 2001, Bao and Wang 2019, Cai *et al*. 2020). However, global methods fail to discover hidden co-locations occurring in individual regions, which are common in spatial datasets where most relationships are geographically regional, rather than global (Ding *et al*. 2011). Thus, research on the discovery of regional co-locations has received increasing attention in recent years (Xie *et al*. 2017). It updates the problem of discovering global spatial co-locations to regional scales through space-partitioning or region-detection strategies.

Methods that employ space-partitioning first partition the study area into smaller regions, thereby allowing the reuse of global discovery methods in each region to extract co-locations. Space can be partitioned using quad-tree structures (Celik *et al*. 2007), multi-resolution grids (Ding *et al*. 2011), or *k*-nearest neighbor graphs (Qian *et al*. 2014). However, these user-specified schemes for partitioning are independent of the endogenic distribution of co-locations; furthermore, they may impair the discovery of the true regions with co-locations (Mohan *et al*. 2011).

The second strategy, region detection, attempts to overcome this limitation by identifying co-location regions in a data-driven manner. Generally, regions of co-locations are determined based on the co-location instances of different features (e.g. minimum orthogonal bounding rectangles of all subsets of co-location instances (Li and Shekhar 2018)). Furthermore, regions of co-location can be understood as the concentration of co-location instances, which can be identified using a neighbor graph (Mohan *et al*. 2011), the prototype-based clustering method (Eick *et al*. 2008), or the adaptive pattern clustering method (Deng *et al*. 2017); this interpretation may be of substantial interest to geographers. To further reduce subjectivity in the evaluation, Cai *et al*. (2018) developed non-parametric significance tests to validate the co-location regions.

All of the aforementioned methods can determine where the co-locations among features are prevalent; however, they cannot identify the regions where the co-locations are anomalous. Taking the dataset in Figure 1 as an example, all the aforenoted methods will report R_1 and R_2 as co-location regions because features *A* and *B* always occur together in these two regions (Figure 2(a)). However, some normal co-locations are also included in these regions, and regions with weak co-locations cannot be identified.

Recently, the ability of spatial scan methods (e.g. the spatial scan statistic (Kulldorff 1997)) has been exploited in bivariate and multivariate cases to detect statistically sound clusters of multiple features. For example, Jung *et al*. (2010) proposed a multinomial spatial scan statistic to detect clusters where the proportions of at least one of the features are significantly different from those expected. However, co-locations among different

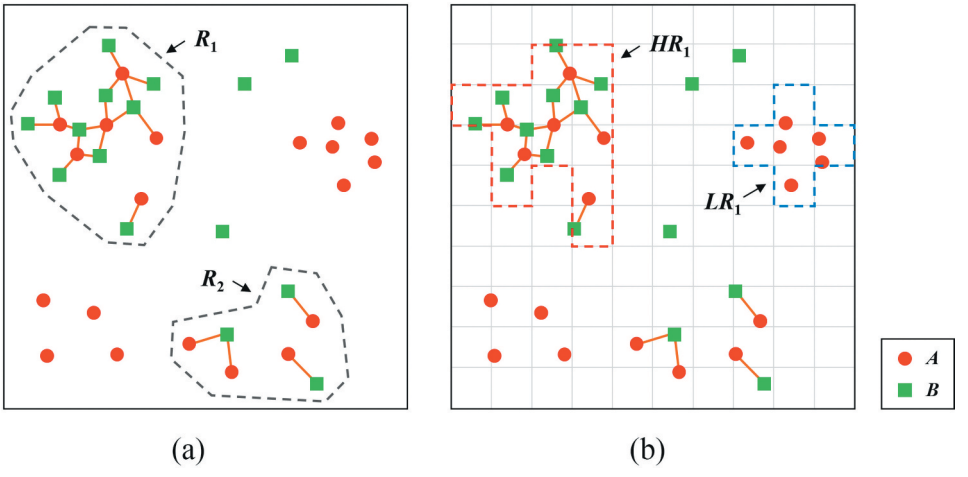

Figure 2. Illustration of existing methods for regional co-location discovery: (a) adaptive pattern clustering method and (b) scan-statistic-based method.

features cannot always be guaranteed in multinomial clusters (Leibovici *et al*. 2014). Leibovici *et al*. (2011) developed an exploratory scan approach to visualize and test clusters of multivariate associations using statistics based on local co-occurrences. The clustered associations imply local spatial dependence among features (i.e. significantly prevalent co-locations in sub-regions); thus, they do not necessarily point out anomalous co-locations. In addition, the pre-defined geometric shapes (e.g. circle) cannot well represent the natural shapes of clusters (Xie and Shekhar 2019). Wang *et al*. (2013) noticed that the scan-statistic-based method has the potential to discover co-location regions of arbitrary shapes by comparing the co-location probability of two features inside and outside a region represented by connected grid cells. However, the method treats the occurrence of both many and few co-location instances of a feature equally in the estimation of co-location probability; thus, it may also include some normal co-locations in the identified region (see HR_1 in Figure 2(b)). Although the method can easily be modified to detect a single region with a minimal co-location probability ratio (see LR_1 in Figure 2(b)), other valid regions may be missed (see LR_2 in Figure 1). In addition, to perform the significance tests, the data are assumed to follow a bivariate Poisson distribution. This assumption could be invalid in some instances.

2.2. Spatial cross-outlier detection

The second category of related work, spatial cross-outlier detection, is an extension of spatial outlier detection for a single type of feature, which aims to find anomalous instances of a single feature that deviate significantly from their neighborhoods (Shekhar *et al*. 2003). Anomalous instances of a single feature can be determined based on the values of their spatial attributes (locations) (e.g. density-based method (Breunig *et al*. 2000) and Delaunay-triangulation-based method (Shi *et al*. 2016)) or both their spatial and nonspatial attributes (e.g. distance-based method (Lu *et al*. 2003) and graph-based method (Lu *et al*. 2011)). However, the presence of extra features may cause these

methods to misidentify normal instances as anomalous (Papadimitriou and Faloutsos 2003). Therefore, some typical methods, designed for a single type of feature, are modified to discover spatial cross-outliers between two types of features.

In the detection of spatial cross-outliers, the spatial attributes of a primary feature are used to define the neighbor relation, and the number of co-occurring instances of another feature (called a reference feature) serves as the nonspatial attribute for evaluating the outlier instances of the primary feature. Typically, spatial cross-outliers can be identified using the '*k* times the standard deviation' criterion (Papadimitriou and Faloutsos 2003), constrained Delaunay triangulation (Shi *et al*. 2018), or statistical tests (Deng *et al*. 2018). However, these methods cannot directly evaluate statistically anomalous regions of co-occurring features. Furthermore, the determination of spatial cross-outliers commonly involves subjective assumptions regarding the distribution of features, such as the assumption of a normal distribution underlying the cross-outlier criterion (Papadimitriou and Faloutsos 2003) and the complete spatial randomness process used to define the null distribution of a feature (Deng *et al*. 2018), in which the spatial auto- and cross-correlation characteristics of the observed datasets are ignored.

3. A novel strategy for ROASC detection

As discussed above, discovering ROASCs remains challenging because the nature of the distribution underlying the data is unknown *a priori*. ROASCs are usually spatially irregular owing to the complex distribution of features inside the regions. To reveal the arbitrarily shaped ROASCs that are fully encapsulated in the data, we propose a detection strategy that is independent of distribution assumptions.

First, we develop a multidirectional optimization method that frees the generation of candidate ROASCs from implicit assumptions regarding the size and shape of regions. For two input spatial features, one is designated as the primary feature *pf* and the second is designated the reference feature *rf* with respect to which we investigate the co-locations (Papadimitriou and Faloutsos 2003). The designation of *pf* and *rf* depends on the semantics of the application domain. The method adaptively constructs a spatial neighbor relation among instances of *pf* and measures the co-location intensity with respect to *rf*. The collection of neighboring instances of *pf* that exhibit a higher-than-average (or lower-than-average) co-location intensity with *rf* and their co-located instances of *rf* is then considered a candidate high-value (or low-value) ROASC. Using this protocol, candidate ROASCs are discovered in a bottom-up manner by iteratively searching for interesting instances of *pf* from each seed instance, in all directions specified by the spatial neighbor relation.

Second, when determining ROASCs, we eliminate the need for restrictive assumptions regarding the distribution forms of features. To ensure that the discovered ROASCs are unlikely to occur by chance, the determination of a ROASC is modeled as a significance test problem under a null hypothesis H_0 which says that the co-location intensity between two features inside a candidate region is consistent with that expected in the entire study area, i.e. the expected co-locations exhibited by the observed dataset are spatially uniform across the whole geographical space. To model the H_0, we need to randomize the distribution of spatial co-locations constrained by all other characteristics of the observed data. In practice, this means that the null model must be conditioned on the observed auto- and cross-correlation structures of the two features, and only questions regarding

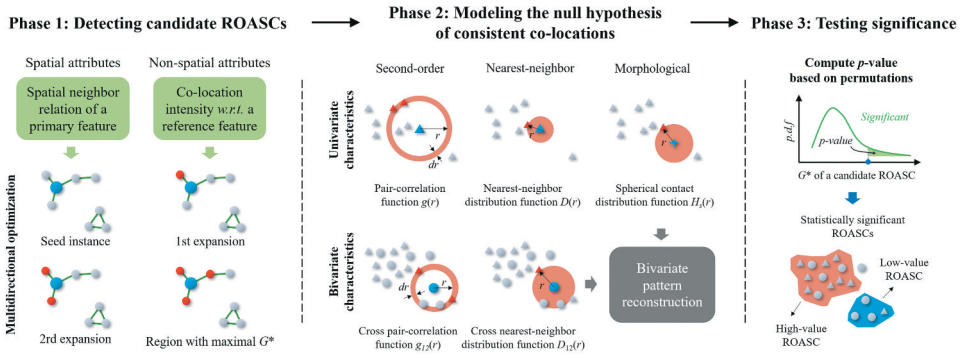

Figure 3. Framework of the distribution-free strategy for detecting arbitrarily shaped ROASCs.

the uniformity of expected co-locations are explored. Although the distribution forms of two features are unknown *a priori*, their spatial structures can be described using several summary characteristics. Considering this, a bivariate pattern reconstruction method is developed as a Monte Carlo simulator of H_0 for the significance tests by reconstructing the observed univariate and bivariate summary characteristics.

Figure 3 presents the framework of the proposed three-part strategy: (1) generation of the candidate ROASCs; (2) construction of the null hypothesis; and (3) implementation of the significance tests. The techniques involved in these three phases are detailed in the following section.

4. A distribution-free method

4.1. Multidirectional optimization for identifying candidate ROASCs

We first present a multidirectional optimization method to identify candidate ROASCs, whose sizes and shapes are fully endogenized. Fundamental to this method is a well-established spatial statistic tool, AMOEBA (Aldstadt and Getis 2006), that is used to identify spatial clusters of related areal units with high or low attribute values. The AMOEBA procedure starts with one or more seed units and then, defines a high-value (or low-value) cluster by iteratively adding neighboring units until the local spatial autocorrelation statistic is maximized (or minimized). We take advantage of its ability to guide our method toward an optimal solution with regard to ROASCs at the finest scale. However, AMOEBA is not immediately extensible to the ROASC discovery for two main reasons. First, the spatial neighbor relation, defined based on the contiguity of spatial units, is not applicable to spatial points that are usually unevenly distributed in continuous space. Second, it is designed for spatial units with one type of continuous variable, but here is for spatial points of two Boolean spatial features. Our method upgrades the original AMOEBA so that it can be used for two types of point data through an adaptive neighborhood definition and a co-location intensity measurement.

A prerequisite for this method is the definition of the spatial neighbor relation among instances of the primary feature *pf*. Here, we employ the multi-level constrained Delaunay triangulation-based method (Deng *et al*. 2011) because it provides an adaptive concept of neighborhood that can better reflect the characteristics of data. With regard to Delaunay

triangulation (*DT*) performed among instances of *pf*, the global and local long edges linked to each instance I_i^{pf} are successively removed if their length is larger than the global and local constraint statistics, $GC(I_i^{pf})$ and $LC(I_i^{pf})$, respectively, which are represented as:

$$GC\left(I_i^{pf}\right) = Mean(DT) + \frac{Mean(DT)}{Mean(E_{DT}^1(I_i^{pf}))} \cdot SD(DT) \tag{1}$$

$$LC\left(I_i^{pf}\right) = Mean(E_{SG}^2(I_i^{pf})) + Mean(SD(E_{SG}^1)) \tag{2}$$

where *Mean*(*DT*) and *SD*(*DT*) are the mean and standard deviation of the lengths of all edges in *DT*, respectively, $Mean(E_{\mathrm{DT}}^1(I_i^{pf}))$ is the mean length of edges directly linked to I_i^{pf} in *DT*, *SG* is the sub-graph containing I_i^{pf}, obtained by removing global edges from *DT*, $Mean(E_{\mathrm{SG}}^2(I_i^{pf}))$ is the mean length of the edges linked to I_i^{pf} within two paths in *SG*, and $Mean(SD(E_{\mathrm{SG}}^1))$ is the mean of the standard deviations of edges directly linked to each instance in *SG*. In Figure 4(a), the connected instances are identified as neighbors after removing the global and local long edges.

Given a distance threshold *r* that reflects the scale-level of interest at which one wants to investigate the spatial co-locations (termed as co-location distance), an instance of the reference feature *rf*, I_j^{rf}, is considered to co-occur with an instance of *pf*, I_i^{pf}, if the distance between them, $d(I_i^{pf}, I_j^{rf})$, is not larger than *r*. The co-location intensity of *pf* with respect to *rf* at the location of I_i^{pf} is then measured using the number of co-occurring instances of *rf*, represented as:

$$CI_i = |\{I_j^{rf} | d(I_i^{pf}, I_j^{rf}) \leq r\}| \tag{3}$$

For a region *R*, we employ the G^* statistic (Getis and Ord 1992, Duque *et al*. 2011) as the interest measure, represented as:

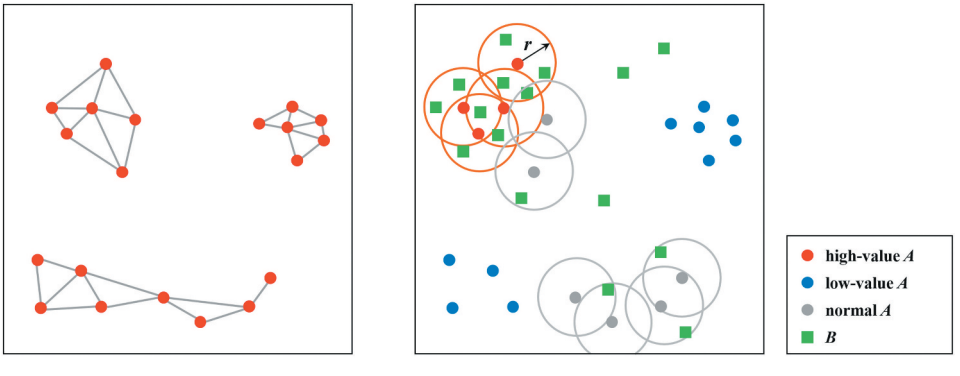

Figure 4. Preliminaries for the multidirectional optimization method: (a) neighbor relation constructed for the primary feature *A* and (b) spatial co-locations defined between the primary feature *A* and the reference feature *B*.

$$G^*(R) = \left(\sum_{I_i^{pf} \in R} CI_i - n \cdot \overline{CI}\right) / \left(S \cdot \sqrt{\frac{N \cdot n - n^2}{N-1}}\right) \tag{4}$$

where n is the number of instances of pf included in the region R, N is the total number of instances of pf, $\overline{CI}$ is the mean of all the CI_i values, and

$$S = \sqrt{\frac{\sum_{j=1}^{N} CI_j^2}{N} - \overline{CI}^2} \tag{5}$$

A positive (or negative) G^* value indicates that the co-location intensity of pf and rf in R is higher (or lower) than the average level exhibited by the dataset as a whole. Consider the primary feature A and reference feature B in Figure 4(b) as an example. The average co-location intensity $\overline{CI}$ is the average number of instances of B that co-occur with an instance of A; furthermore, it equals (4 + 3+ … +0)/20 = 20/20 = 1. The G^* value of the region that contains four high-value instances of A (represented by red points) is $((4+3+4+3)-4\cdot 1)/\left(\sqrt{\left(4^2+3^2+\ldots+0^2\right)/20-1^2}\cdot\sqrt{\left(20\cdot 4-4^2\right)/(20-1)}\right)\approx 4.06$, which indicates the stronger-than-average co-locations between A and B in that region.

Based on the above definitions, the multidirectional optimization method starts by considering each instance of the primary feature (called the primary instance) as a seed instance and iteratively expands the region from each seed instance to its neighboring primary instances in a constructive manner (Duque *et al*. 2011). This is an efficient and equivalent alternative to exhaustive evaluations on all possible neighbor combinations (Widener *et al*. 2012). The process is detailed as follows:

(1) For a region R_i^t with a positive G^*, its neighboring primary instances outside R_i^t are sorted according to their G^* values, in descending order. The variable t is the number of primary instances included in the current region and its value starts from 1, i.e. the initial region R_i^1 consists of only the seed instance I_i^{pf}.
(2) The sorted neighbors are tested one-by-one. If $G^*(R_i^{t+1}) > G^*(R_i^t)$, i.e. the region R_i^{t+1} that contains R_i^t and a neighbor is more interesting than R_i^t according to the G^* value, the region is expanded by adding that neighbor so that t becomes $t + 1$.
(3) If any neighbor is added, the neighbors that are not included are eliminated from further consideration, and step (1) is followed to test the neighbors of newly added neighbors. The process is terminated if no new neighbors are added, and R_i^t is outputted as the high-value region with the maximal G^* value, with respect to I_i^{pf}.

Figure 5(a–d) illustrates the process starting from the seed instance A_1. For regions with negative G^*, the process followed is the same, except that the goal is to minimize the value of G^*. After each seed instance is examined, non-overlapping regions with the maximal absolute G^* values are reported. The collection of primary instances in each reported region and their co-located reference instances is then identified as a candidate ROASC for the significance test (see the high-value region HR_1 and the low-value regions LR_1 and LR_2 in Figure 5(e)).

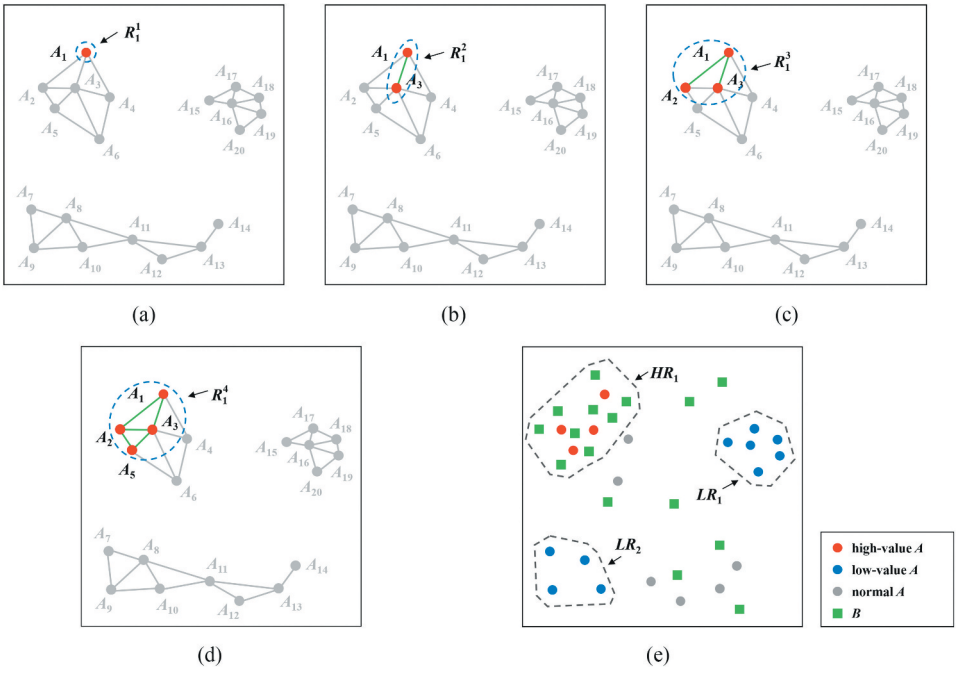

Figure 5. Multidirectional optimization method for identifying candidate ROASCs: (a) seed instance; (b) the first expansion; (c) the second expansion; (d) the third expansion and (e) candidate high- and low-value ROASCs.

4.2. Bivariate pattern reconstruction method for constructing the null hypothesis

At the outset of the significance test, we need to model the null hypothesis H_0 such that no regions with unexpected co-locations exist, i.e. the expected co-locations between two features are uniformly distributed in the study area. As analyzed in Section 3, the permutations under H_0 should consider the following properties: (1) consistent co-location intensity across space; (2) similar univariate spatial structures of each feature; and (3) similar bivariate spatial structures between two features as in the observed dataset. The first two properties are used to randomize the distribution of the co-locations between two spatially autocorrelated features. However, the potential spatial cross-correlation between features is likely to be disrupted. Therefore, the third property should also be maintained, so that the expected co-locations in the observation can be guaranteed in permutations. To generate such permutations, we propose a bivariate pattern reconstruction method, that translates a reconstruction technique in materials science (Rintoul and Torquato 1997) into the bivariate point pattern analysis. Instead of assuming the distribution forms of features, the proposed method generates permutations by fitting several univariate and bivariate summary characteristics of the observations, thus allowing subjectivity to be reduced in the modeling of H_0.

The first step is to determine a proper combination of summary characteristics because different summary characteristics usually describe different aspects of spatial structures and may also capture redundant information. To comprehensively characterize the univariate spatial structures, we select three summary characteristics, namely, the pair-correlation

function $g(r)$, nearest-neighbor distribution function $D(r)$, and spherical contact distribution function $H_s(r)$, based on the systematic comparison conducted by Wiegand *et al*. (2013). Figure 6(a) shows the schematic representation of these univariate summary characteristics. The $g(r)$ function captures the average neighborhood properties of points; furthermore, it is the most informative characteristic when used in isolation. $D(r)$ is valuable because of its up-close view of the nearest neighbor that can quantify subtle variations in local structures that are lost by $g(r)$. Finally, $H_s(r)$ can provide important additional information on the size of gaps, especially for non-stationary patterns. Similarly, the bivariate forms of $g(r)$ and $D(r)$, namely, the cross pair-correlation function $g_{12}(r)$ and cross nearest-neighbor distribution function $D_{12}(r)$, are also recommended for describing the bivariate spatial structures. The computation is analogous to that in the case of univariate functions, except that bivariate functions summarize the neighborhood properties of one feature with respect to another feature (Figure 6(b)). Unlike point-centered $g(r)$ and $D(r)$, the $Hs(r)$ function characterizes the spatial structures from the viewpoint of arbitrary locations; thus, it is difficult to present in a bivariate form to measure the cross-correlation between points of two features. Details regarding the five selected summary characteristics, $g(r)$, $D(r)$, $H_s(r)$, $g_{12}(r)$, and $D_{12}(r)$, can be found in Wiegand and Moloney (2013).

Based on the selected summary characteristics $F_m(r)$ (where $m = 1, 2, \ldots, M$), the bivariate pattern reconstruction pattern method produces permutated datasets with fixed instances of the primary feature *pf*, while reconstructing the instances of the reference feature *rf*. The reconstruction of *rf* starts with a random pattern that has the same number of instances of *rf* as in the observed dataset ω. The permutated dataset ϖ is, then, iteratively modified to minimize the deviations of $F_m(r)$ values between ω and ϖ, calculated as

$$\Delta F(\varpi) = \sum_{m=1}^{M} k_m \cdot \sqrt{\frac{1}{|r|} \cdot \sum_{r=r_{min}}^{r_{max}} \left[F_m^{\omega}(r) - F_m^{\varpi}(r)\right]^2} \Big/ \sum_{m=1}^{M} k_m \tag{6}$$

where $|r|$ is the number of distances r ($r_{min} \leq r \leq r_{max}$) at which the $F_m(r)$ values are evaluated, and k_m is the weight of $F_m(r)$ that is used to balance the importance of different

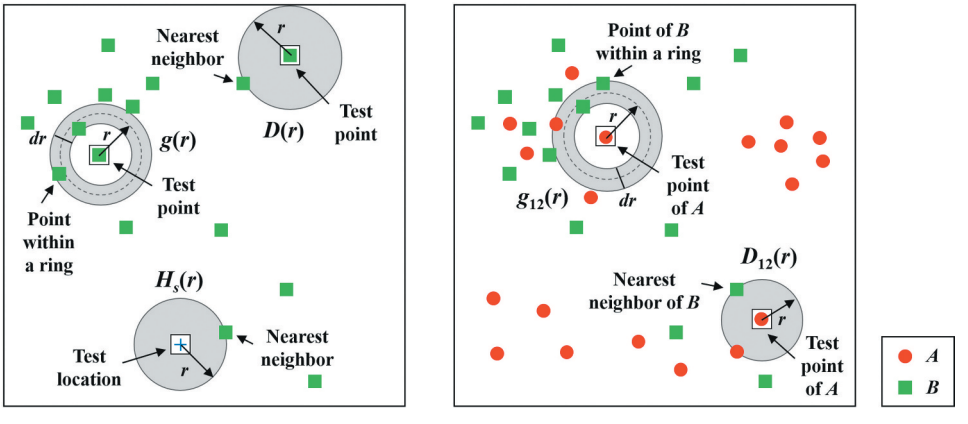

Figure 6. Schematic representation of univariate and bivariate summary characteristics: (a) $g(r)$, $D(r)$, and $H_s(r)$ and (b) $g_{12}(r)$ and $D_{12}(r)$.

summary characteristics. Note that the univariate $F_m(r)$ (e.g. $g(r)$, $D(r)$, and $H_s(r)$) is calculated for *rf* and the bivariate $F_m(r)$ (e.g. $g_{12}(r)$ and $D_{12}(r)$) is calculated by considering *rf* as the reference feature. In each modification step *t*, a randomly selected instance of *rf* in the last permutated dataset ϖ_{t-1} is tentatively replaced with a new point with random coordinates. The modified dataset ϖ_t is accepted only if ϖ_t is more similar to ω than ϖ_{t-1}, i.e. $\Delta F(\varpi_t) < \Delta F(\varpi_{t-1})$. Otherwise, another modification is considered. This process proceeds until the $\Delta F(\varpi_t)$ value becomes smaller than a tiny value (0.01 in this study) or a sufficient number of steps (40, 000 in this study) is reached.

Figure 7(b) shows a permutated dataset of the dataset in Figure 7(a), produced using the bivariate pattern reconstruction pattern method. Figure 7(c–g) displays the curves of $g(r)$, $D(r)$, $H_s(r)$, $g_{12}(r)$, and $D_{12}(r)$ values calculated for feature *B* in the observed dataset and 99 permutated datasets. Clearly, the method randomizes the distribution of spatial co-locations between two features while maintaining the observed univariate and bivariate spatial structures.

4.3. Monte Carlo tests for ROASCs adjusted for the multiple testing problem

In the decision-making step, tests of ROASCs employ the G^* statistic as the test statistic. G^* is asymptotically distributed as a standard normal variate (Duque *et al*. 2011). However, the normality of G^* maybe lost in practice, in which case tests based on the normal

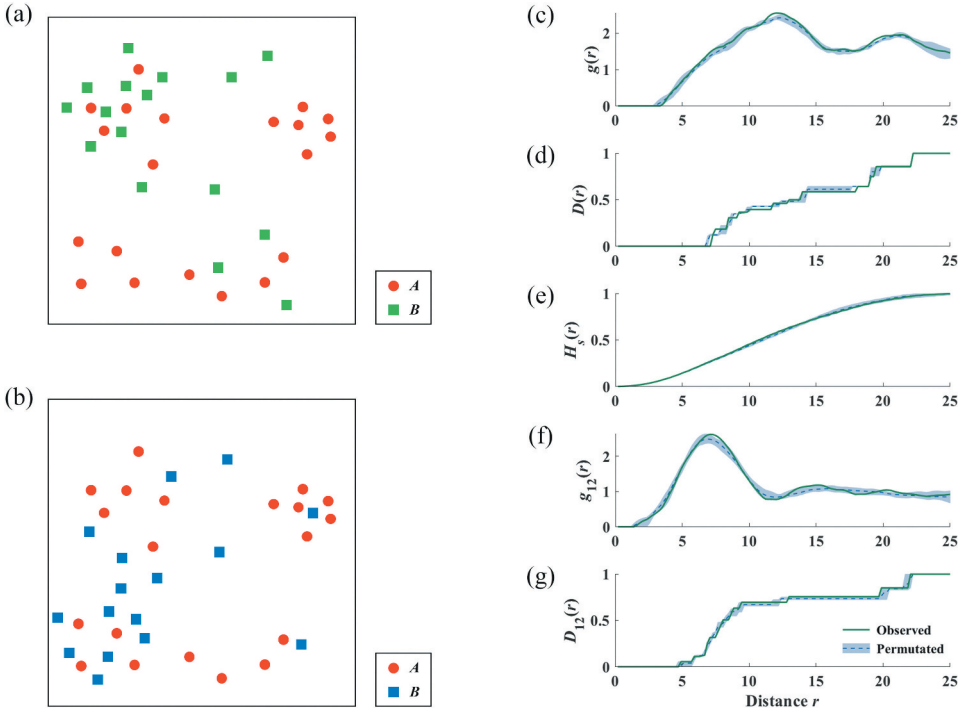

Figure 7. Bivariate pattern reconstruction pattern method for modeling the null hypothesis of consistent co-locations: (a) observed dataset; (b) an example of the permutated datasets; (c)–(e) curves of the $g(r)$, $D(r)$, $H_s(r)$, $g_{12}(r)$ and $D_{12}(r)$.

approximation will be inappropriate (Getis and Ord 1992). To obtain more objective decisions with regard to ROASCs, G^* is assessed via its empirical distribution estimated using Monte Carlo permutations, wherein *a priori* assumptions on the null distribution are not required.

For a high-value (or low-value) ROASC, *HR* (or *LR*), we rank the observed G^* value, G^{obs}, in descending (or ascending) order amongst a corresponding set of values, $G_n^{null}(n = 1, 2, \ldots, N)$, calculated for a large number N of Monte Carlo permutated datasets. The *p*-value of *HR* (or *LR*) is, then, calculated as the rank divided by $N + 1$, represented as

$$p - value(HR) = (|G_n^{null}(HR) \geq G^{obs}(HR)| + 1)/(N + 1) \tag{7}$$

$$p - value(LR) = (|G_n^{null}(LR) \leq G^{obs}(LR)| + 1)/(N + 1) \tag{8}$$

where $|G_n^{null}(HR) \geq G^{obs}(HR)|$ is the number of G_n^{null} values of *HR* that exceed the observed value and $|G_n^{null}(LR) \leq G^{obs}(LR)|$ is analogous.

In practice, the dataset usually has more than one ROASC. The unguarded use of multiple tests will result in an increased false-positive rate (i.e. the probability of falsely identifying a region as a significant ROASC). To alleviate the multiple testing problem, we adjust the given significance level α (a cutoff value of *p*-value which is 0.01 or 0.05 by convention) using the false discovery rate method (Benjamini and Hochberg 1995). Let *p*-value(R_1) $\leq$ *p*-value(R_2) $\leq \cdots \leq$ *p*-value(R_K) be the ordered *p*-values of K candidate ROASCs. The adjusted significance level α_{adj} is *p*-value(R_i), where i is the largest index in {1, 2, …, K} for which

$$p - value(R_i) \leq \frac{i}{K} \cdot \alpha \tag{9}$$

If the *p*-value for an *HR* (or *LR*) is not larger than α_{adj}, we reject the null hypothesis and conclude that the spatial co-locations between features in *HR* (or *LR*) are significantly stronger (or weaker) than expected; moreover, the *HR* (or *LR*) is identified as a significant ROASC of high (or low) value.

4.4. Implementation and analysis of the distribution-free method

Given (1) an observed dataset containing instances of two spatial features (primary feature *pf* and reference feature *rf*; (2) a distance threshold r for defining spatial co-locations; and (3) a significance level α, the distribution-free method detects all non-overlapping ROASCs with maximal absolute G^* values and qualified *p*-values, using the following steps:

(1) Construct the spatial neighbor relation among n instances of *pf*. The multi-level constrained Delaunay triangulation-based method approximately requires $O(n \cdot logn)$ time.
(2) Calculate co-location intensity and G^* for each instance of *pf*. This requires approximately $O(n \cdot logm)$ time, where m is the number of instances of *rf*.
(3) Identify candidate ROASCs with maximal absolute G^* values using the multidirectional optimization method. This requires a maximum of $O(n^2)$ time when the

neighbor graph of primary instances is completely connected and the entire set of primary instances is explored, for each seed instance.

(4) Generate a set of N permutated datasets using the bivariate pattern reconstruction method. Because each candidate permutation differs only by one point from that of the last iteration, only the part of an estimator of a summary characteristic that is affected by the exchange needs to be updated. This process requires $O(m)$ time for functions $g(r)$ and $D(r)$, $O(t)$ time for $H_s(r)$, and $O(n)$ time for $g_{12}(r)$ and $D_{12}(r)$ at each distance. This whole step requires a maximum of approximately $O(N \cdot R \cdot S \cdot (n + m + t)$ time when all the permutated datasets are modified for the maximum number S of times. Here, t is the number of test locations used in the $H_s(r)$ function, and R is the number of distances at which summary characteristics are evaluated.

(5) Conduct the Monte Carlo tests and report the statistically significant ROASCs. This requires approximately $O(N \cdot n \cdot logm)$ time.

As discussed above, the time complexity of the distribution-free method mainly depends on Steps 3–5, which require a maximum time of $O(n^2) + O(N \cdot R \cdot S \cdot (n + m + t)) + O(N \cdot n \cdot logm)$.

5. Experimental evaluation and case study

We evaluated the performance of the distribution-free method using both synthetic and real-world taxi datasets. For comparison, the scan-statistic-based method (Wang *et al.* 2013) was also applied because of its similarity to our method in the problem formulation. For both methods, the co-location distance was predefined in the synthetic experiments and estimated as 500 m in the case study, in accordance with the spatial auto-correlation method (Yoo and Bow 2012) which recommends the use of a distance at which spatial processes substantially promote clustering. The significance level was set to 0.05, and the number of permutated datasets was set to 99. For better visualization, the α-shape algorithm (Edelsbrunner *et al.* 1983) was used to delineate the boundary of the ROASCs identified using our method.

5.1. Experiments using synthetic data

5.1.1. Data generation

Figure 8 illustrates the synthetic data generator designed to predefine the ROASCs. The generator first produced the instances of primary feature A, including n_{high} high-value instances, n_{low} low-value instances, and n_{normal} normal instances. The high- and low-value instances were divided into several groups, and the normal instances were randomly distributed. Instances of reference feature B were, then, randomly located within the predefined distance r of instances of A. The mean number of instances of B that co-occurred with each high-value, low-value, and normal instance of A is μ_{high}, μ_{low}, and μ_{normal}, respectively. Here, $\mu_{\text{normal}} = (\mu_{\text{high}} \cdot n_{\text{high}} + \mu_{\text{low}} \cdot n_{\text{low}})/(n_{\text{high}} + n_{\text{low}})$, so that the average co-location intensity between A and B can be controlled to μ_{normal}. The region formed by each group of high-value (low-value) instances of A and the corresponding co-occurring instances of B is known

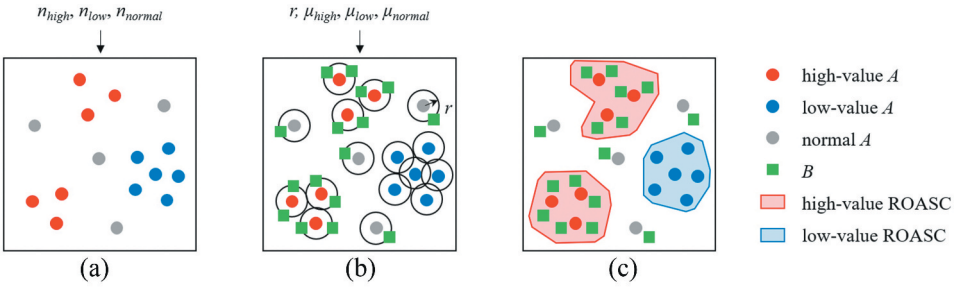

Figure 8. Experimental setup of the synthetic dataset: (a) generate instances of the primary feature; (b) generate instances of the reference feature and (c) predefined ROASCs.

as a high-value (low-value) ROASC. Using this generator, we obtained a synthetic dataset containing four predefined ROASCs, of which two were high-value and two were low-value, in a study area $S = [0, 100]^2$ (Figure 9(a)). The total number of instances of primary feature A and reference feature B was 160 and 800, respectively. Here, $n_{high}=n_{low}= 30, n_{normal}= 100, r = 2, \mu_{high}= 9, \mu_{low}= 1$, and $\mu_{normal}= 5$. Thus, in the synthetic dataset, each primary instance has an average of nine, one, and five neighboring reference instances in the high-value ROASCs, low-value ROASCs, and the entire study area, respectively.

5.1.2. Performance metrics

We evaluated two aspects of the detection methods: (1) the extent to which each method can correctly find the known ROASCs and (2) the extent to which the known ROASCs can be completely uncovered by each method. Since a ROASC is determined by the instances of the primary feature A inside it, the predefined groups of high-value and low-value instances of A were employed to serve as the benchmark for assessing the performance of both methods, using the metrics of precision, recall, and F1 score, defined as

$$\text{precision} = |TP|/(|TP|+|FP|) \tag{10}$$

$$\text{recall} = |TP|/(|TP|+|FN|) \tag{11}$$

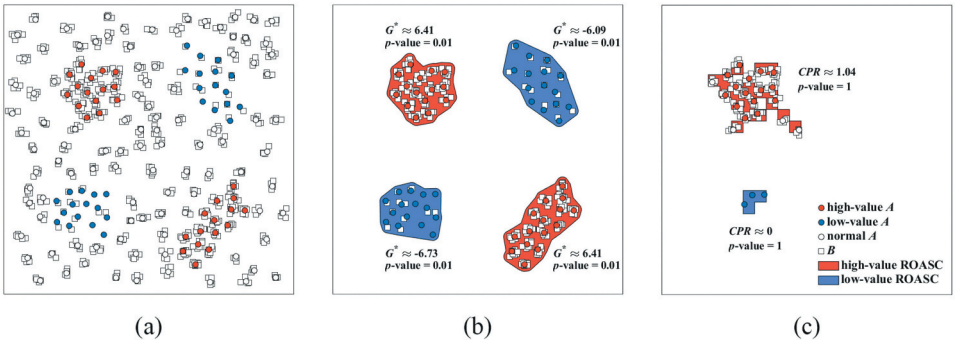

Figure 9. Synthetic data and discovered ROASCs: (a) predefined high- and low-value primary instances; (b) our method; and (c) scan-statistic-based method.

$$\text{F1 score} = 2 \cdot \text{precision} \cdot \text{recall} / (\text{precision} + \text{recall}) \tag{12}$$

where $|TP|$, $|FP|$, and $|FN|$ are the number of true positive, false positive, and false negative instances of A, respectively. Precision is also referred to as the positive predictive value; moreover, it represents the fraction of detections that are truly positive, indicating the correctness of results. Recall is also referred to as sensitivity; furthermore, it represents the fraction of true cases that are successfully detected, indicating the completeness of results. The F1 score is a comprehensive metric that takes both correctness and completeness into account. The larger the values of these metrics, the better the results.

5.1.3. Comparison and analysis

Figure 9(b) displays the ROASCs detected by our method and reports their G^* values and p-values. The precision, recall, and F1 score of the results are all equal to 100%. These results demonstrate that our method detected all the predefined ROACSs without any false-positive or false-negative errors in the synthetic dataset. It was able to do this because the shapes and sizes of ROASCs are adaptively determined using the multi-directional optimization method, and the casual ROASCs that occur by chance can be effectively removed using the significance tests.

By contrast, Figure 9(c) shows the regions with maximal and minimal co-location probability ratio (*CPR*) obtained by the scan-statistic-based method. As discussed in Section 2.1, the high-value region incorrectly includes some normal instances of A near the predefined ROASCs. In addition, the low-value region omits some predefined low-value instances of A. This happened because the minimal *CPR* of the regions in this dataset was zero. Thus, adding any instances of A that co-occurred with B would increase the *CPR* of the reported region, even though the number of co-occurring instances of B was significantly smaller than the average. Furthermore, both regions are reported as statistically insignificant by the significance tests on the observed *CPR*. This was because the co-location probability embedded in the *CPR* only considers whether the instances of A are co-located with instances of B; that is, it cannot capture the co-location intensity between the two features. In this synthetic dataset, almost all instances of A (154 out of 160) co-occur with instances of B. Thus, it is not surprising to observe a higher or equal *CPR* in the replicas of the bivariate Poisson distribution, where the co-location rate of two features is learned from observations. Similarly, we can commonly detect one region with a *CPR* of zero in the replicates. Therefore, the scan-statistic-based method is not suitable for datasets where two features frequently co-occur across the entire study area.

5.1.4. Effect of co-location distance

Figure 10 shows the effect of co-location distance on the performance of our distribution-free method. As can be seen, co-location distances somewhat smaller or larger than the preset distance ($r = 2$) will result in undesirable performance in terms of precision, recall, and F1 score. Some meaningful co-location instances of the reference feature may be missed at a smaller distance, which will, then, lead to the underestimation of co-location intensity. Similarly, ROASCs discovered at a larger distance could include certain distant reference instances that are weakly correlated to the primary feature.

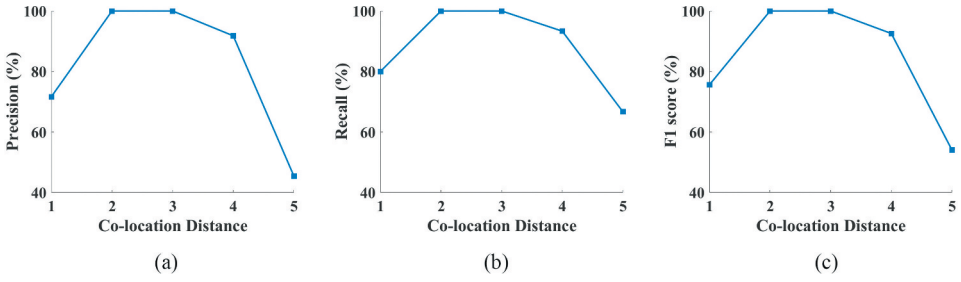

Figure 10. Effect of co-location distance on the performance of our distribution-free method: (a) precision; (b) recall and (c) F1 score.

5.2. Case study: detecting regions of taxi demand-supply mismatch in Shanghai

5.2.1. Data description

We assessed the practicality and effectiveness of the distribution-free method via a case study of taxi data pertaining to Shanghai, China. Shanghai is one of the most densely populated Chinese cities, with more than 24 million permanent residents. However, only approximately 50,000 taxis are registered in the city. A report from the Didi Media Research Institute and CBNData (2016) concluded that Shanghai is the most challenging regarding hailing taxis among all the cities in the Yangtze Delta area. The background population is one of the most apparent factors influencing the demand for taxis (Qian and Ukkusuri 2015). The built environment (e.g. residential and commercial buildings) is a key determinant of the daily activities of individuals (Sung and Oh 2011), and this causes dynamic variation in population size and density across space and time. Thus, the distribution of taxi demand is usually uneven in space and varies with time, making it difficult for taxis to satisfy the demand. Detecting regional mismatches of taxi demand and supply is of critical importance to the provision of responsive taxi services and the facilitation of passenger commutes.

The taxi data used in this case study were collected once every 10 seconds from 48,806 taxis in Shanghai on Monday, 2 December 2013. These taxis, mostly registered with five taxi companies (Dazhong, Jinjiang, Bashi, Qiangsheng, and Haibo), account for more than 97% of all the taxis registered. The data recorded regarding each taxi included the time, location, and status (0 for empty and 1 for occupied). We investigated the supply of and demand for taxi services at 8:00 a.m., 6:00 p.m., 8:00 p.m., and 11:00 p.m. (Figure 11). For each time t that we investigated, the locations corresponding to taxi demand are indicated by the pick-up locations (i.e. the locations where the taxi status switches from 0 to 1) within 10 minutes before t, and the locations of taxi supply are represented by the locations of empty taxis at t. It is noteworthy that we only considered the satisfied demand here, as unmet demand cannot be revealed from these taxi data. In the context of taxi services, we want to know whether the supply of taxis can adequately match the demand. Thus, the demand for taxis is considered as the primary feature and the supply is the reference feature. High-value ROASCs indicate regions of oversupply of taxi services and low-value ROASCs indicate regions of undersupply. Figure 12 shows the distribution of 10 places of interest in Shanghai, which are used to explain the distribution of discovered ROASCs.

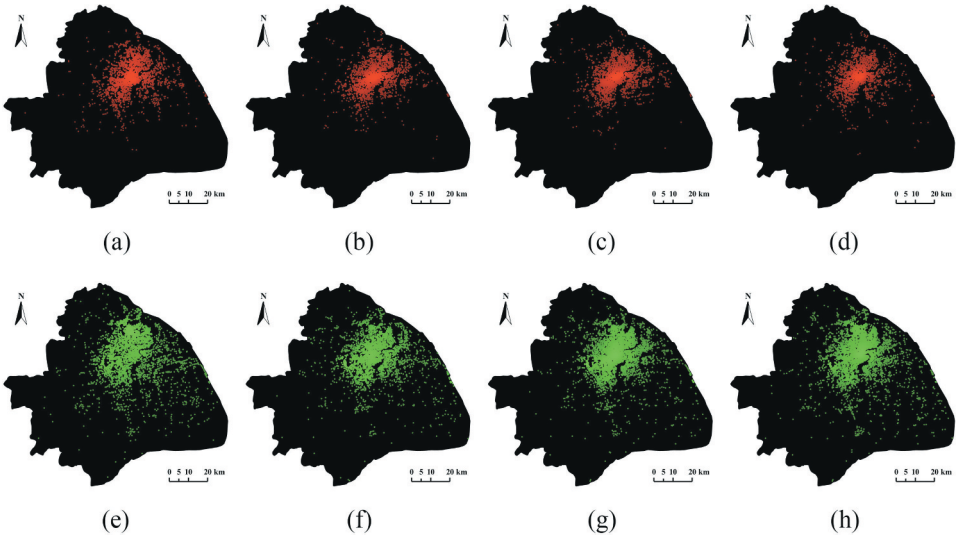

Figure 11. Distribution of taxi data in Shanghai: (a)–(d) demand at 8:00 a.m., 6:00 p.m., 8:00 p.m., and 11:00 p.m., respectively and (e)–(h) supply at 8:00 a.m., 6:00 p.m., 8:00 p.m., and 11:00 p.m., respectively.

5.2.2. ROASCs detected using different methods

Figure 13 presents the spatial distribution of ROASCs discovered by the two methods for the taxi data at 8:00 a.m. Table 1 summarizes the numerical results of the distribution-free method, which include 19 ROASCs (eight high-value regions and eleven low-value regions), sorted in descending order according to their absolute G^* values. By contrast, the scan-statistic-based method only identified one high-value ROASC, where the co-

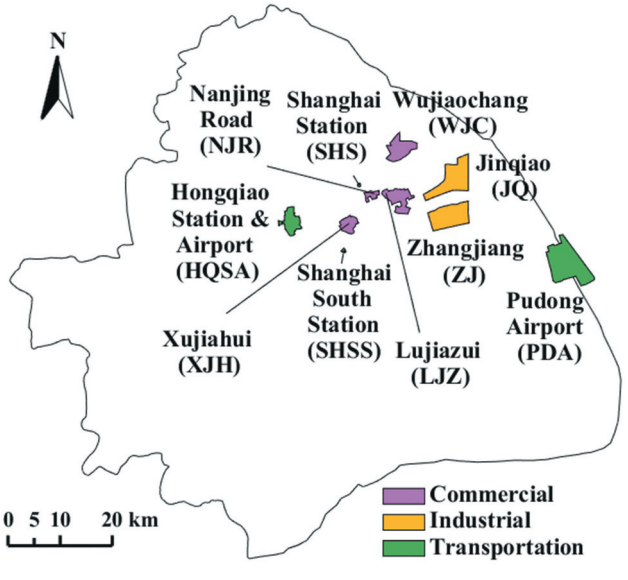

Figure 12. Places of interest in Shanghai.

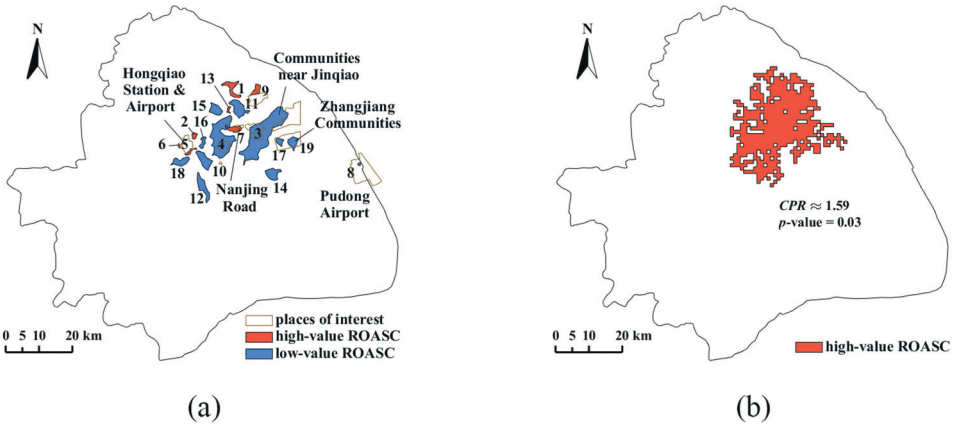

Figure 13. ROASCs at 8:00 a.m. detected by different methods: (a) our method and (b) scan-statistic-based method.

Table 1. Numerical results of ROASCs at 8:00 a.m. detected by our method.

Region ID	Region type	G^* value	p-value
1	high-value	15.80	0.01
2	high-value	11.40	0.01
3	low-value	−9.22	0.01
4	low-value	−7.84	0.01
5	high-value	7.65	0.01
6	high-value	7.46	0.01
7	high-value	6.68	0.01
8	high-value	6.29	0.01
9	high-value	4.45	0.01
10	low-value	−4.35	0.01
11	low-value	−4.13	0.01
12	low-value	−3.83	0.01
13	high-value	3.47	0.01
14	low-value	−3.44	0.01
15	low-value	−2.55	0.01
16	low-value	−2.28	0.01
17	low-value	−2.22	0.01
18	low-value	−2.21	0.01
19	low-value	−2.21	0.01

location probability ratio (*CPR*) was maximized. This region also covered some normal and weak co-locations identified by our method because spatial co-locations weaker than or equal to the average can also help increase the *CPR*. In addition, no significant low-value ROASCs were identified because the *CPR* of the low-value region was zero, i.e. no taxi supply co-occurred with demand. This phenomenon can also be regularly observed in a sub-region under the null hypothesis.

5.2.3. ROASCs detected at different times

In this subsection, we explored the variations in ROASCs discovered by our method at other times (6:00 p.m., 8:00 p.m. and 11:00 p.m.) during the day. Figure 14 shows the taxi travel flows among 10 places of interest listed in Figure 12 and other regions in Shanghai leaving at four different times. Figure 16 overlaps the detected ROASCs with these places.

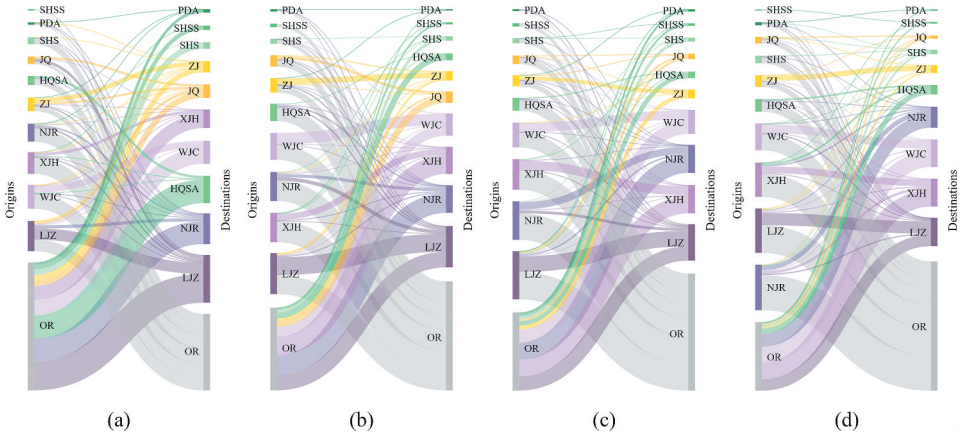

Figure 14. Travel flows among places of interest and other regions (OR) in Shanghai at different times: (a) 8:00 a.m.; (b) 6:00 p.m.; (c) 8:00 p.m. and (d) 11:00 p.m. (For better visualization, flows between other regions are not displayed.).

Some common phenomena can be observed at these times. For example, a high-value ROASC can always be detected near Pudong International Airport (see region 8 in Figure 13(a), region 2 in Figure 15(a), and region 13 in Figure 15(b)), except at 11 p.m., when the airport system is about to terminate its schedule. Figure 16 shows the time and distance distributions of taxis that traveled from Pudong Airport at different times. Although the travel times at 8:00 a.m. and 6:00 p.m. (working hours) tend to be higher than those at 8:00 p.m. and 11:00 p.m. (non-working hours), the distributions of travel distances at these times are similar. In particular, the peak distances at these times are between 36 km and 39 km (Table 2), which may reflect travel to downtown Shanghai (e.g. Nanjing Road). Such long-distance trips can make taxi services more profitable. At these four times, an average of 35.2% of the taxis carrying passengers to Pudong Airport did not leave within 15 minutes, suggesting that these taxis may have waited to pick up customers in the airport queue, thus creating an oversupply at the airport. Furthermore, high-value ROASCs were usually found in the center of the city (see region 7 in Figure 13(a), region 8 in Figure 15(a), region 2 in Figure 15(b), and region 1 in Figure 15(c)). This is because alternative

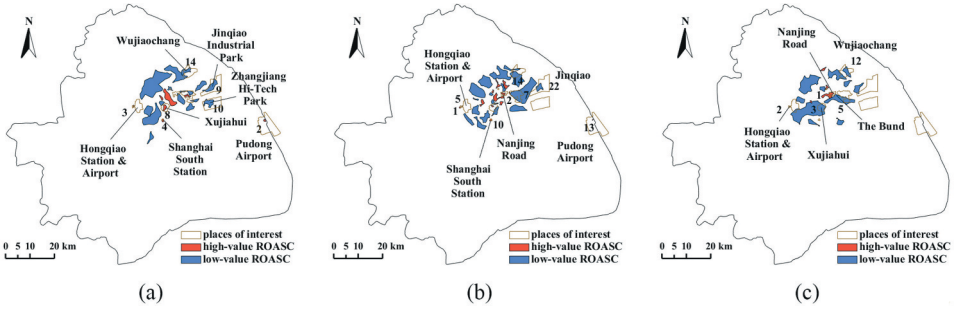

Figure 15. ROASCs detected at different times: (a) 6:00 p.m.; (b) 8:00 p.m. and (c) 11:00 p.m. (Regions are numbered in descending orders of absolute G^* values. For better visualization, some numbers are not displayed.).

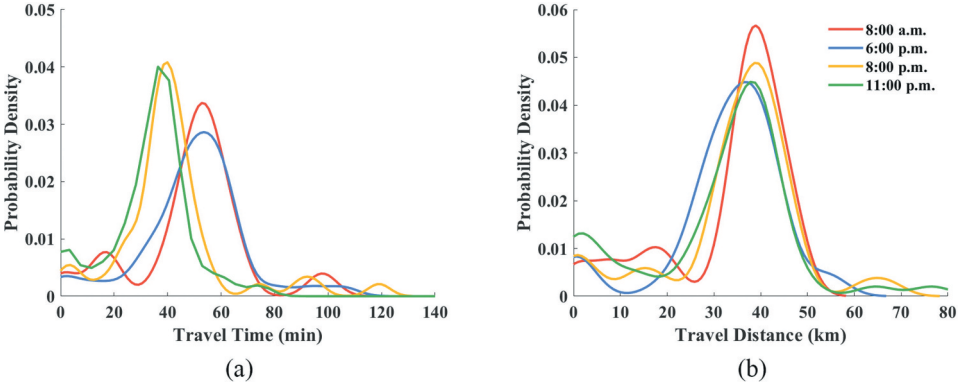

Figure 16. Distribution of the time and distance traveled from Pudong Airport at different times: (a) time distribution and (b) distance distribution.

Table 2. Peak distances and concentrative destinations traveled from Pudong Airport at different times.

Time	Peak distance (km)	Destinations around peak distance
8:00 a.m.	38.93	Nanjing Road, Wujiaochang
6:00 p.m.	36.46	Nanjing Road, Lujiazui
8:00 p.m.	38.67	Nanjing Road
11:00 p.m.	37.79	Nanjing Road, Xujiahui

transportation services, including subways, buses, and shared bicycles, are accessible in these commercial areas (e.g. the Nanjing Road Pedestrian Street and the Xujiahui district). The convenient accessibility of other transportation services likely increases the risk of oversupply of taxis in these areas.

Some interesting temporal dynamics in the attributes, locations, and shapes of ROASCs are apparent. For example, at 8:00 a.m., low-value ROASCs tend to occur in residential areas (see regions 3 and 19 in Figure 13(a)), whereas at 6:00 p.m., the low-value ROASCs shift to workplaces, such as the Jinqiao Industrial Park and Zhangjiang Hi-Tech Park (see regions 9 and 10 in Figure 15(a)). Such a phenomenon likely results from the fact that the importance of residential areas and workplaces with regard to taxi demand differs across the day. During the morning rush hour, the demand for taxis is mainly associated with residential areas, whereas during the evening rush hour, the demand is more substantial at workplaces. During non-working hours, the proportion of travel flows starting from the two workplaces, Jinqiao and Zhangjiang, decreased (Figure 14(c)–(d)). Therefore, smaller or even no low-value ROASCs were identified in these places (Figure 15(b–c)). At 8:00 p.m., low-value ROASCs were usually observed outside of business districts (e.g. region 7 around Lujiazui in Figure 15(b)). The reason for this could be the increasing number of taxi departures in these areas for leisure activities, such as dinner, shopping, and sightseeing. At 11:00 p.m., a few new low-value ROASCs were identified in busy nightlife areas (e.g. region 5 near the Bund, a famous attraction in Shanghai, in Figure 15(c)). The main reasons for this are twofold: first, the metro system is closed during this period, suggesting more demand for taxis (Wu *et al*. 2017); second, most people in these regions need to return home or to hotels, owing to their daily schedule. Factors related to the generation of

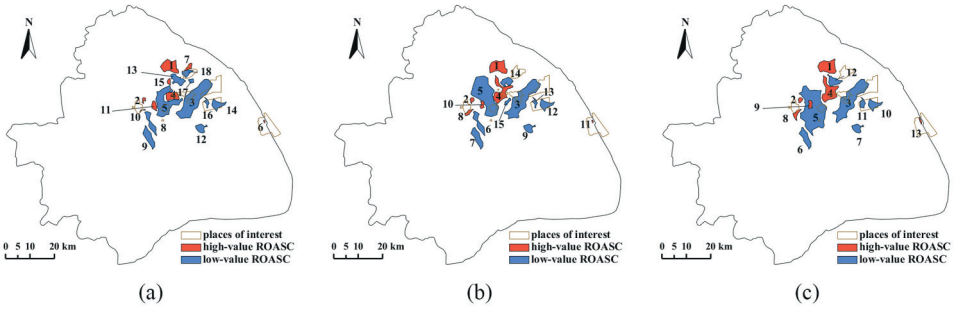

Figure 17. ROASCs detected at 8:00 a.m. using different co-location distances: (a) 750 m; (b) 1000 m and (c) 1250 m. (Regions are numbered in descending orders of absolute G^* values.).

demand may further trigger the undersupply of taxis (Tang *et al*. 2019), thereby contributing to the formation of low-value ROASCs.

5.2.4. ROASCs detected using different co-location distances

In a real-world dataset, there may be no single correct co-location distance. Thus, we further implemented our method at multiple co-location distances (750 m, 1000 m, and 1250 m) for the taxi data at 8:00 a.m. With increasing co-location distances, some ROASCs may become smaller or even disappear (e.g. regions 17 and 18 in Figure 17(a) cannot be identified in Figure 17(b–c)), while some are likely to become larger (e.g. region 10 in Figure 17(a) gradually becomes larger in Figure 17(b–c)). A larger distance may or may not associate more reference instances with primary instances; furthermore, both the number of high- and low-value primary instances may increase or decrease at a larger distance, leading to diverse changes in the spatial scope of ROASCs. However, some ROASCs were relatively stable across different distances, such as the two regions in Zhangjiang. These ROASCs are more valid according to the popular belief in the multi-scale analysis (Witkin 1984, Leung *et al*. 2000); furthermore, they should be considered as typical regions of mismatched taxi demand and supply that will require substantial efforts to improve the efficiency of the taxi network in these regions.

6. Conclusions and future work

In this paper, we define a novel problem of ROASC discovery. ROASCs are different from regional co-locations, which represent prevalent co-locations occurring in sub-regions, and differ from spatial cross-outliers, which are points of a feature that arouse suspicions with respect to points of another feature. By comparison, ROASCs refer to regions with unexpected co-locations between different features, thus providing new and valuable insights for studying surprising spatial associations at regional scales. ROASC discovery is of significant practical interest in many domains, including detecting mismatching regions of taxi demand and supply in transportation and identifying areas of high crime risk in criminology.

With regard to the discovery of ROASCs, this paper presents a distribution-free method that does not entail restrictive assumptions regarding the distribution of data. The main advantages of the method are twofold: first, the method is capable of adaptively

determining the sizes and shapes of ROASCs based on endogenous spatial neighbor relations; second, a bivariate pattern reconstruction technique enables the method to objectively establish the statistical significance of the results by reconstructing the spatial auto- and cross-correlation structures observed in the data. Experiments, with both synthetic and real-world taxi datasets, revealed that the distribution-free method can discover valid ROASCs more effectively in sharp contrast to the existing method.

The distribution-free method has several limitations to be considered in the future. First, this approach does not allow the identification of ROASCs having more than two spatial features. It is important to generalize the approach to multivariate cases where the spatial structures among more than two features need to be explored. Second, analysis based on the assumption of Euclidean space could be ill-suited for human-mobility-related phenomena as these are commonly constrained by road networks (*Yu et al*. 2017, Cai *et al*. 2019). The proposed method will be extended by using network distance to identify network-constrained ROASCs. Third, the computational performance of the distribution-free method is significantly constrained by Monte Carlo-type permutation tests. Parallel computational implementations will be studied to improve the scalability of the method to large datasets (Prasad *et al*. 2017).

Acknowledgments

The authors thank the editors, the reviewers, and the members of the spatial computing research group at the University of Minnesota for their helpful comments. We also thank Kim Koffolt for improving the readability of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the National Natural Science Foundation of China (NSFC) [41730105, 41471385]; National Key Research and Development Foundation of China [2016YFB0502303]; U.S. National Science Foundation (NSF) [1737633, 0940818, 1029711, 1541876, IIS-1218168, IIS-1320580]; Advanced Research Projects Agency - Energy, U.S. Department of Energy [DE-AR0000795]; U.S. Department of Defense [HM0210-13-1-0005, HM1582-08-1-0017]; U.S. Department of Agriculture [2017-51181-27222]; U.S. National Institute of Health [KL2 TR002492, TL1 TR002493, UL1 TR002494]; OVPR Infrastructure Investment Initiative, University of Minnesota; Minnesota Supercomputing Institute (MSI), University of Minnesota.

Notes on contributors

Jiannan Cai received his PhD in GIScience at the Central South University and was a visiting PhD student at the University of Minnesota, Twin Cities. He is currently a Postdoctoral Fellow of the Institute of Space and Earth Information Science at The Chinese University of Hong Kong. His research interests focus on spatial data science and its smart-city applications.

Min Deng is currently a Professor and Associate Dean of the School of Geosciences and Info-physics at the Central South University. His research interests are map generalization, spatio-temporal data analysis and mining.

Yiwen Guo is a PhD candidate at the Central South University and her research focuses on spatio-temporal association rule mining.

Yiqun Xie received his PhD in Computer Science at the University of Minnesota, Twin Cities. He is currently an Assistant Professor in the Department of Geographical Sciences and Center for Geospatial Information Science at the University of Maryland, College Park. His research focuses on developing novel and cutting-edge techniques for spatial data science and artificial intelligence. His work has received multiple best paper awards and was highlighted by the Great Innovative Ideas program at the Computing Community Consortium.

Shashi Shekhar is a McKnight Distinguished University Professor in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. He is an IEEE Fellow and AAAS Fellow. Earlier, he served as the President of the University Consortium for GIS, and on many National Academies' committees. His research focuses on spatial data science, spatial databases and GIS.

ORCID

Jiannan Cai http://orcid.org/0000-0003-4752-0153

Data and codes availability statement

The synthetic data and codes that support the findings of this study are available in 'figshare.com' with the identifier: https://doi.org/10.6084/m9.figshare.12993146. The Shanghai taxi data cannot be made publicly available due to third party restrictions. Mocked taxi data are provided at the link to show how the codes work.

References

Aldstadt, J. and Getis, A., 2006. Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38 (4), 327–343.

Bao, X. and Wang, L., 2019. A clique-based approach for co-location pattern mining. *Information Sciences*, 490, 244–264. doi:10.1016/j.ins.2019.03.072

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 (Methodological)*, 57 (1), 289–300.

Breunig, M.M., *et al*., 2000. LOF: identifying density-based local outliers. *In*: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, USA: ACM, 93–104

Cai, J., *et al*., 2018. Adaptive detection of statistically significant regional spatial co-location patterns. *Computers, Environment and Urban Systems*, 68, 53–63. doi:10.1016/j.compenvurbsys.2017.10.003

Cai, J., *et al*. 2019. Nonparametric significance test for discovery of network-constrained spatial co-location patterns. *Geographical Analysis*, 51 (1), 3–22. doi:10.1111/gean.12155

Cai, J., *et al*., 2020. Significant spatial co-distribution pattern discovery. *Computers, Environment and Urban Systems*, 84, 101543. doi:10.1016/j.compenvurbsys.2020.101543

Celik, M., Kang, J.M., and Shekhar, S., 2007. Zonal co-location pattern discovery with dynamic parameters. *In*: *Proceedings of the 7th IEEE International Conference on Data Mining*, 28–31 October, Omaha NE.

Deng, M., *et al*. 2011. An adaptive spatial clustering algorithm based on Delaunay triangulation. *Computers, Environment and Urban System*, 35 (4), 320–332. doi:10.1016/j.compenvurbsys.2011.02.003

Deng, M., *et al*. 2017. Multi-level method for discovery of regional co-location patterns. *International Journal of Geographical Information Science*, 31 (9), 1846–1870. doi:10.1080/13658816.2017.1334890

Deng, M., *et al*. 2018. A non-parametric statistical test method to detect significant cross-outliers in spatial points. *Transactions in GIS*, 22 (6), 1462–1483. doi:10.1111/tgis.12481

Didi Media Research Institute and CBNData., 2016. Yangtze River Delta city intelligent travel big data report. https://www.cbndata.com/report/354/detail

Ding, W., *et al*. 2011. A framework for regional association rule mining and scoping in spatial datasets. *Geoinformatica*, 15 (1), 1–28. doi:10.1007/s10707-010-0111-6

Duque, J.C., *et al*. 2011. A computationally efficient method for delineating irregularly shaped spatial clusters. *Journal of Geographical Systems*, 13 (4), 355–372. doi:10.1007/s10109-010-0137-1

Edelsbrunner, H., Kirkpatrick, D., and Seidel, R., 1983. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29 (4), 551–559. doi:10.1109/TIT.1983.1056714

Eick, C.F., *et al*., 2008. Finding regional co-location patterns for sets of continuous variables in spatial datasets. *In*: *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, Irvine, California: ACM.

Getis, A. and Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24 (3), 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x

He, Z., *et al*., 2020. Mining spatiotemporal association patterns from complex geographic phenomena. *International Journal of Geographical Information Science*, 34 (6), 1162–1187. doi:10.1080/13658816.2019.1566549

Huang, Y., Shekhar, S., and Xiong, H., 2004. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16 (12), 1472–1485. doi:10.1109/TKDE.2004.90

Jung, I., Kulldorff, M., and Richard, O.J., 2010. A spatial scan statistic for multinomial data. *Statistics in Medicine*, 29 (18), 1910–1918. doi:10.1002/sim.3951

Kulldorff, M., 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26 (6), 1481–1496. doi:10.1080/03610929708831995

Leibovici, D.G., 2011. Spatially clustered associations in health related geospatial data. *Transactions in GIS*, 15 (3), 347–364. doi:10.1111/j.1467-9671.2011.01252.x

Leibovici, D.G., *et al*. 2014. Local and global spatio-temporal entropy indices based on distance-ratios and co-occurrences distributions. *International Journal of Geographical Information Science*, 28 (5), 1061–1084. doi:10.1080/13658816.2013.871284

Leslie, T.F. and Kronenfeld, B.J., 2011. The colocation quotient: a new measure of spatial association between categorical subsets of points. *Geographical Analysis*, 43 (3), 306–326. doi:10.1111/j.1538-4632.2011.00821.x

Leung, Y., Zhang, J., and Xu, Z., 2000. Clustering by scale-space filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (12), 1396–1410. doi:10.1109/34.895974

Li, Y. and Shekhar, S., 2018. Local co-location pattern detection: a summary of results. *In*: *Proceedings of the 10th International Conference on Geographic Information Science*. Melbourne, Australia: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lu, C.T., *et al*., 2003. Detecting spatial outliers with multiple attributes. *In*: *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, Sacramento, California, USA: IEEE, 122–128.

Lu, C.T., *et al*. 2011. A graph-based approach to detect abnormal spatial points and regions. *International Journal on Artificial Intelligence Tools*, 20 (4), 721–751. doi:10.1142/S0218213011000309

Mohan, P., *et al*., 2011. A neighborhood graph based approach to regional co-location pattern discovery: A summary of results. *In*: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, 1–4 November, Chicago, IL, 122–132.

Papadimitriou, S. and Faloutsos, C., 2003. Cross-outlier detection. *In*: *International Symposium on Spatial and Temporal Databases*, Santorini, Greece: Springer, Berlin, Heidelberg, 199–213.

Pei, T., *et al*. 2015. Density-based clustering for data containing two types of points. *International Journal of Geographical Information Science*, 29 (2), 175–193. doi:10.1080/13658816.2014.955027

Prasad, S.K., *et al*., 2017. Parallel processing over spatial-temporal datasets from geo, bio, climate and social science communities: a research roadmap. *In*: *2017 IEEE International Congress on Big Data*. Honolulu, HI, USA: IEEE, 232–250.

Qian, F., *et al*., 2014. Mining regional co-location patterns with kNNG. *Journal of Intelligent Information Systems*, 42 (3), 485–505. doi:10.1007/s10844-013-0280-5

Qian, X. and Ukkusuri, S.V., 2015. Spatial variation of the urban taxi ridership using GPS data. *Applied Geography*, 59, 31–42. doi:10.1016/j.apgeog.2015.02.011

Rintoul, M.D. and Torquato, S., 1997. Reconstruction of the structure of dispersions. *Journal of Colloid and Interface Science*, 186 (2), 467–476. doi:10.1006/jcis.1996.4675

Shekhar, S., *et al*. 2003. A unified approach to detecting spatial outliers. *GeoInformatica*, 7 (2), 139–166. doi:10.1023/A:1023455925009

Shekhar, S. and Huang, Y. 2001. Discovering spatial co-location patterns: a summary of results. *In*: *International symposium on spatial and temporal databases*, Springer Berlin Heidelberg, 236–256.

Shi, Y., *et al*., 2016. Adaptive detection of spatial point event outliers using multilevel constrained Delaunay triangulation. *Computers, Environment & Urban Systems*, 59, 164–183. doi:10.1016/j.compenvurbsys.2016.06.001

Shi, Y., *et al*., 2018. A graph-based approach for detecting spatial cross-outliers from two types of spatial point events. *Computers, Environment and Urban Systems*, 72, 88–103. doi:10.1016/j.compenvurbsys.2018.05.011

Sung, H. and Oh, J.T., 2011. Transit-oriented development in a high-density city: identifying its association with transit ridership in Seoul, Korea. *Cities*, 28 (1), 70–82. doi:10.1016/j.cities.2010.09.004

Tang, J., *et al*. 2019. Identification and interpretation of spatial–temporal mismatch between taxi demand and supply using global positioning system data. *Journal of Intelligent Transportation Systems*, 23 (4), 403–415. doi:10.1080/15472450.2018.1518137

Wang, S., *et al*., 2013. Regional co-locations of arbitrary shapes. *In*: *International Symposium on Spatial and Temporal Databases*. Munich, Germany: Springer, Berlin, Heidelberg, 19–37.

Widener, M.J., Crago, N.C., and Aldstadt, J., 2012. Developing a parallel computational implementation of AMOEBA. *International Journal of Geographical Information Science*, 26 (9), 1707–1723. doi:10.1080/13658816.2011.645477

Wiegand, T., He, F., and Hubbell, S.P., 2013. A systematic comparison of summary characteristics for quantifying point patterns in ecology. *Ecography*, 36 (1), 92–103. doi:10.1111/j.1600-0587.2012.07361.x

Wiegand, T. and Moloney, K.A., 2013. *Handbook of spatial point-pattern analysis in ecology*. Boca Raton, Florida: CRC Press.

Witkin, A., 1984. Scale-space filtering: a new approach to multi-scale description. *In*: *Proceedings of the 1984 IEEE International Conference on Acoustics*, Speech, and Signal Processing. San Diego, CA, USA, 150–153.

Wu, H., Fan, H., and Wu, S., 2017. Exploring spatiotemporal patterns of long-distance taxi rides in Shanghai. *ISPRS International Journal of Geo-Information*, 6 (11), 339. doi:10.3390/ijgi6110339

Xie, Y., *et al*. 2017. Transdisciplinary foundations of geospatial data science. *ISPRS International Journal of Geo-Information*, 6 (12), 395. doi:10.3390/ijgi6120395

Xie, Y., *et al*., 2018. A TIMBER framework for mining urban tree inventories using remote sensing datasets. *In*: *2018 IEEE International Conference on Data Mining*. Singapore: IEEE, 1344–1349.

Xie, Y. and Shekhar, S., 2019. Significant DBSCAN towards statistically robust clustering. *In*: *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*. Vienna, Austria: ACM, 31–40.

Yoo, J.S. and Bow, M., 2012. Mining spatial colocation patterns: a different framework. *Data Mining and Knowledge Discovery*, 24 (1), 159–194. doi:10.1007/s10618-011-0223-0

Yu, W., *et al*., 2017. Spatial co-location pattern mining of facility points-of-interest improved by network neighborhood and distance decay effects. *International Journal of Geographical Information Science*, 31 (2), 280–296. doi:10.1080/13658816.2016.1194423

Zhou, M., *et al*. 2019. A visualization approach for discovering colocation patterns. *International Journal of Geographical Information Science*, 33 (3), 567–592. doi:10.1080/13658816.2018.1550784