Computers, Environment and Urban Systems 84 (2020) 101543

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Significant spatial co-distribution pattern discovery

Jiannan Cai[a,b], Yiqun Xie[b,c,d], Min Deng[a,*], Xun Tang[b], Yan Li[b], 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] Center for Geospatial Information Science, University of Maryland, College Park, MD, USA
[d] Department of Geographic Sciences, University of Maryland, College Park, MD, USA

ARTICLE INFO

Keywords:
Spatial data mining
Spatial association
Co-distribution patterns
Spatial clustering
Significance test

ABSTRACT

Given instances (spatial points) of different spatial features (categories), significant spatial co-distribution pattern discovery aims to find subsets of spatial features whose spatial distributions are statistically significantly similar to each other. Discovering significant spatial co-distribution patterns is important for many application domains such as identifying spatial associations between diseases and risk factors in spatial epidemiology. Previous methods mostly associated spatial features whose instances are frequently located together; however, this does not necessarily indicate a similarity in the spatial distributions between different features. Thus, this paper defines the significant spatial co-distribution pattern discovery problem and subsequently develops a novel method to solve it effectively. First, we propose a new measure, dissimilarity index, to quantify the difference between spatial distributions of different features under the spatial neighbor relation and then employ it in a distribution clustering method to detect candidate spatial co-distribution patterns. To further remove spurious patterns that occur accidentally, the validity of each candidate spatial co-distribution pattern is verified through a significance test under the null hypothesis that spatial distributions of different features are independent of each other. To model the null hypothesis, a distribution shift-correction method is presented by randomizing the relationships between different features and maintaining spatial structure of each feature (e.g., spatial autocorrelation). Comparisons with baseline methods using synthetic datasets demonstrate the effectiveness of the proposed method. A case study identifying co-morbidities in central Colorado is also presented to illustrate the real-world applicability of the proposed method.

1. Introduction

Given instances (spatial points) of different spatial features (categories), the discovery of significant spatial co-distribution patterns (SSCDPs) aims to find subsets of features whose instances show statistically significant similarity in terms of spatial distribution. SSCDP discovery can facilitate the understanding of spatial associations among different features, and further help domain scientists to predict the spatial distributions of features that are otherwise difficult to obtain. For example, in spatial epidemiology, the distribution of a disease is linked to the spatial distribution of its sources or risk factors that contribute to its transmission or development (Elliott & Wartenberg, 2004), thus enabling the formation of a SSCDP between a disease and its sources or factors, e.g., {*Lyme Disease*, *Infected Host-Seeking Ixodes scapularis Nymphs*} in the Eastern United States (Pepin et al., 2012). Such a spatial pattern can contribute to the generation of a reliable disease risk map, which can be used as a guide to geographically prioritize prevention efforts (Diuk-Wasser et al., 2006). SSCDP discovery is also important in diverse applications, such as ecology, e.g., identifying the parasitic relationship between emerald ash borers and ash trees (Xie, Bao, Shekhar, & Knight, 2018), and criminology, e.g., detecting the trigger events associated with risks of criminal activities (Cai et al., 2019).

Most previous work has focused on discovering spatial co-location patterns represented by subsets of features whose instances are frequently located together in space (Shekhar & Huang, 2001). However, although different features spatially occur together with high prevalence, their spatial distributions may still vary significantly. Thus, this paper formally defines the SSCDP discovery problem to provide a novel perspective for analyzing the spatial association between features. Specifically, we want to know which spatial features often occur in close spatial proximity and also have similar spatial distributions. To

* Corresponding author.
E-mail addresses: jiannan.cai@csu.edu.cn (J. Cai), xiexx347@umn.edu (Y. Xie), dengmin@csu.edu.cn (M. Deng), tangx456@umn.edu (X. Tang), lixx4266@umn.edu (Y. Li), shekhar@umn.edu (S. Shekhar).

https://doi.org/10.1016/j.compenvurbsys.2020.101543
Received 11 November 2019; Received in revised form 10 June 2020; Accepted 10 August 2020
Available online 28 August 2020

effectively solve this problem, we propose a distribution clustering method combined with the significance test. The proposed method can effectively discover SSCDPs without assumptions about the distribution models or characteristics of features.

The rest of this paper is organized as follows. Section 2 reviews the related work and proposes a new strategy. Section 3 introduces the basic concepts and formulates the problem of SSCDP discovery. Section 4 describes and analyzes the SSCDP discovery method in detail. In Section 5, we present the experimental evaluation and a case study on a public health dataset. Section 6 concludes the paper and outlines future work.

2. Related work and a new strategy

2.1. Spatial co-location pattern discovery

Our work is closely related to the previous work on spatial co-location pattern discovery, which aims to find subsets of features whose instances are frequently co-located in close spatial proximity (Shekhar & Huang, 2001). Existing methods can be broadly categorized into three classes, namely, clustering-based, frequent-pattern-based, and statistics-based methods.

Clustering-based methods can be further divided into layer-clustering and feature-clustering methods. Layer-clustering methods (Estivill-Castro & Lee, 2001; Estivill-Castrol & Murray, 1998) first detect the cluster regions of instances in the layer of each feature, and then find co-location patterns based on the overlapping areas of cluster regions from all the layers. This strategy can only work well if the instances of each spatial feature tend to cluster in space. A feature-clustering method (Huang & Zhang, 2006), by contrast, treats each spatial feature as a clustering object, and then clusters features to represent co-location patterns. However, the similarity measure, namely density ratio, can only capture the pairwise dependence between two features. Additionally, it is difficult to determine appropriate clustering parameter(s), e.g., a meaningful stopping criterion for hierarchical clustering methods.

Frequent-pattern-based methods extend the pruning strategies used in classic association rule mining methods, e.g., Apriori method (Agrawal & Srikant, 1994), to detect the subsets of spatial features that frequently occur together. However, these classic methods cannot be directly applied to spatial datasets because transactions do not naturally exist in continuous geographic space (Huang, Shekhar, & Xiong, 2004; Yoo & Shekhar, 2006). For this reason, an event-centric model (Shekhar & Huang, 2001) is commonly used to build neighbor graphs among instances of different features. Subsequently, a participation index evaluates the prevalence of different features occurring in neighborhoods. A co-location pattern is identified as prevalent if its participation index value is not smaller than a given threshold. In practice, such prevalent co-location patterns may happen by chance, which may result in false positives or negatives regarding the association between features.

Statistics-based methods interpret a co-location pattern as a dependence among various types of spatial point processes, and determine the statistical significance through testing under the null hypothesis of independence. One of the key requirements of the test is maintaining the univariate spatial structure of each feature (Dixon, 2002). This can be done using either parametric or nonparametric methods. Parametric methods use the realizations of the fitted point-process model as a null model, such as the homogenous Poisson process used in the cross K-function (Ripley, 1976) and the Matérn cluster process used in Barua and Sander (2014). This strategy works well only if the underlying distribution is appropriately fit by the presumptive distribution model. An alternative solution is the nonparametric strategy used by Deng, He, Liu, Cai, and Tang (2017) and Cai et al. (2019), where stochastic permutations are generated by closely approximating several predefined statistical characteristics of the observed dataset. However, suitable distribution characteristics of features still need to be carefully selected in advance. In addition, due to the non-monotonic property of statistical significance, a brute-force test on all the candidate patterns is usually required, which will result in expensive computational costs.

2.2. Critical analysis of existing methods

Previous methods commonly measure the prevalence of co-location patterns by comparing local instance number (or density) of a feature co-located with others with its global number (or density) in the study area (see Table 1), and then associate different spatial features if their co-location prevalence is high enough. However, different features can frequently occur together even if their spatial distributions are significantly different, which may lead to misjudgments of the spatial association. Fig. 1 presents four motivating examples to illustrate the necessity of spatial distributions for understanding the association between different features.

Example 1. In Fig. 1(a), both the spatial locations and distributions of features A and B are clearly different from each other. Pattern $\{A, B\}$ is neither a co-location pattern nor a co-distribution pattern.

Example 2. In Fig. 1(b), all the instances of features A and B occur in neighboring spatial locations, and their spatial distributions are quite consistent. Pattern $\{A, B\}$ is both a co-location pattern and a co-distribution pattern.

Example 3. In Fig. 1(c), although the instances of features A and B are always located together, they have distinctive local densities. Pattern $\{A, B\}$ is a co-location pattern, but not a co-distribution pattern.

Example 4. In Fig. 1(d), the locations of features A and B are in close spatial proximity; however, their spatial distributions are different in the central area. Pattern $\{A, B\}$ is a co-location pattern, but not a co-distribution pattern.

Existing co-location pattern discovery methods can perform well in Examples 1 and 2; however, they may wrongly associate features in Examples 3 and 4 if the spatial distributions are not considered.

Some measures in mathematical statistics, such as the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) and the Jensen-Shannon (JS) divergence (Lin, 1991), can be incorporated into the discovery model to reduce the effect of the above limitation. However, these measures require users to assume the form of distributions beforehand or perform estimations based on independent and identically distributed samples. Such requirements are not ideal in geographic domains because of the spatial dependencies that exist among sample points. For example, to apply the KL divergence on the spatial features shown in Fig. 2(a), it is common to partition the study area into a set of grid cells (Sengstock, Gertz, & Van Canh, 2012) (see Fig. 2(b)). Subsequently, the distribution difference can be described by the difference between the probabilities of the two features, $P_A(c_i)$ and $P_B(c_i)$, in each grid cell c_i. In this example, the KL divergence between A and B is computed as $\Sigma_{i=1}^{24} P_A(c_i) \cdot \log\left(P_A(c_i)/P_B(c_i)\right) = 50$,[1] which indicates that the distributions of A and B are quite different. However, the instances of the two features are always located with similar distributions in close spatial proximity. This conflicting result is caused by the missing spatial relationships broken by the boundaries of space partitioning (Xie et al., 2017). In contrast, the proposed dissimilarity index can correctly capture the identity of these two distributions under the spatial relationship among instances (see Fig. 2(c)). A more detailed description of this index is presented in Section 3.1.

[1] Each $P(c_i)$ is smoothed by $(P(c_i)+\varepsilon)/(1+\varepsilon \cdot C)$ so that $P(c_i) > 0$ for any c_i, where ε is a tiny value and C is the number of grid cells.

Table 1
Comparison of measures for co-location and co-distribution patterns.

Pattern	Measure	Definition
Co-location	Cross *K*-function (Ripley, 1976)	$\frac{\text{average instance number of } f_i \text{ co}-\text{located with } f_j}{\text{instance number of } f_i \text{ in unit area}}$
	Participation index (Shekhar & Huang, 2001)	$\min\left(\frac{\text{instance number of } f_i \text{ co}-\text{located with others}}{\text{total number of instances of } f_i \text{ in the study area}}\right)$
	Density ratio (Huang & Zhang, 2006)	$\frac{\text{average instance density of } f_i \text{ co}-\text{located with } f_j}{\text{instance density of } f_i \text{ in the study area}}$
Co-distribution	Dissimilarity index (This work)	Average difference in the probabilities of f_i and other features occurring around baseline locations

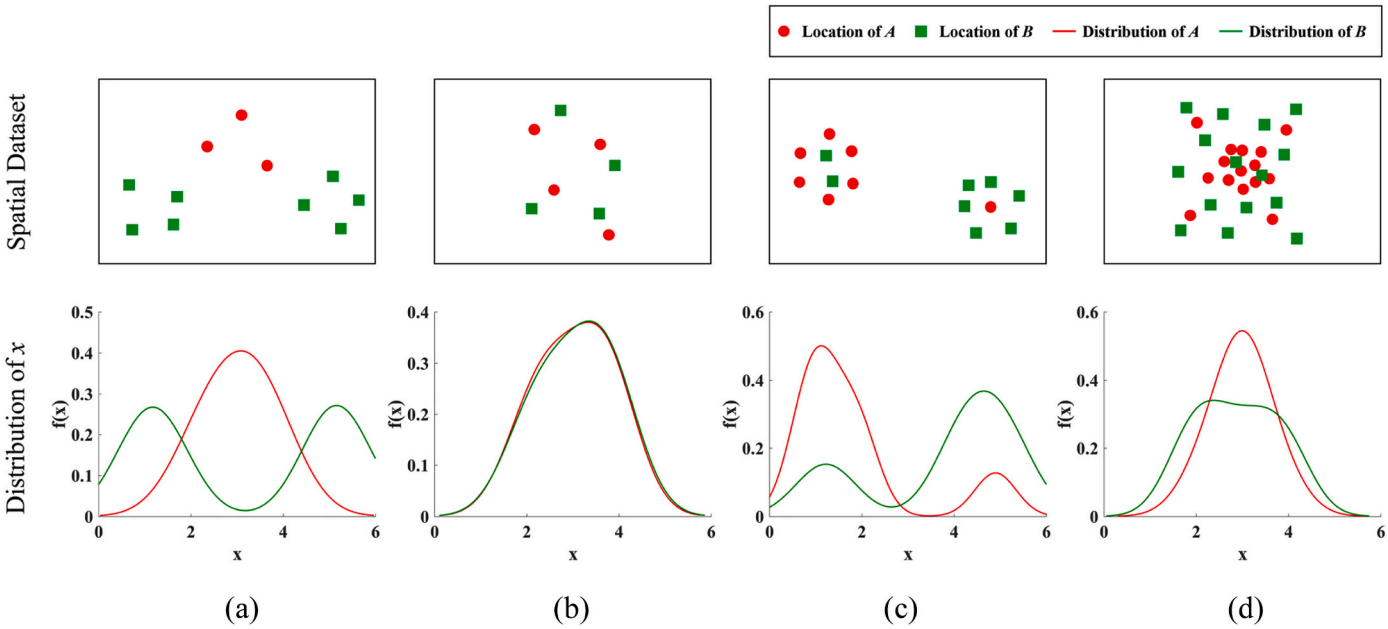

Fig. 1. Examples to distinguish co-location and co-distribution patterns: (a) example 1: not co-location and not co-distribution; (b) example 2: co-location and co-distribution; (c) example 3: co-location but not co-distribution and (d) example 4: co-location but not co-distribution.

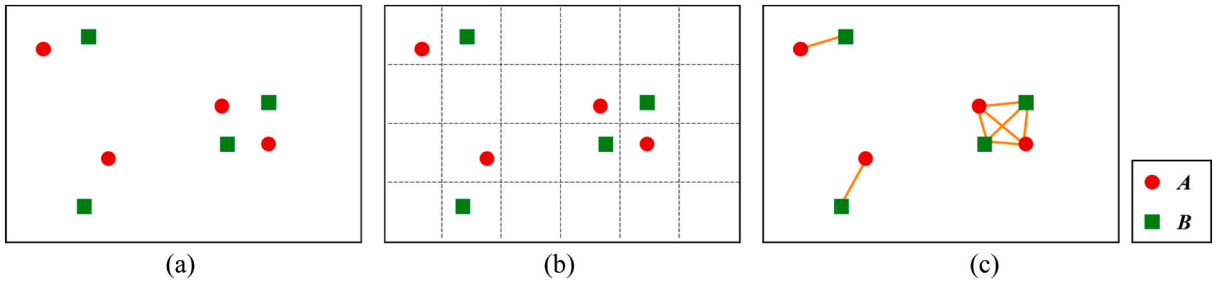

Fig. 2. Illustration of different strategies for measuring distribution similarity: (a) a spatial dataset with instances of two features; (b) space partitioning for applying KL divergence to spatial dataset and (c) neighbor graphs to handle the spatial relationship among instances.

2.3. Novel strategy for discovering SSCDPs

Motivated by the above observations, we formally define a novel problem of SSCDP discovery and propose a hybrid strategy which combines both clustering- and statistics-based strategies to solve this problem. To provide a more rigorous cognition for spatial association between features, a more interesting question is whether specific subsets of spatial features tend to have similar spatial distributions. More specifically, co-distributed features should not only occur in close spatial proximity with high prevalence but also with similar distributions (probabilities).

Based on this cognition, the proposed strategy can be described as follows:

(1) Generation of candidate patterns: First, we define a new measure, namely a dissimilarity index, to measure the difference in the spatial distributions of different features. To achieve an informative and nonredundant representation of co-distributed features, the candidate SSCDPs are represented by clusters of features, and a distribution clustering method is proposed to generate candidate SSCDPs, where the dissimilarity index is used as the similarity measure between features.
(2) Determination of significant patterns: To further remove spurious patterns that happen by chance, the evaluation of SSCDPs is modeled as a significance test problem under the null hypothesis that spatial distributions of different features are independent of each other. To construct the null hypothesis, we develop a distribution shift-correction method that can randomize the relationships

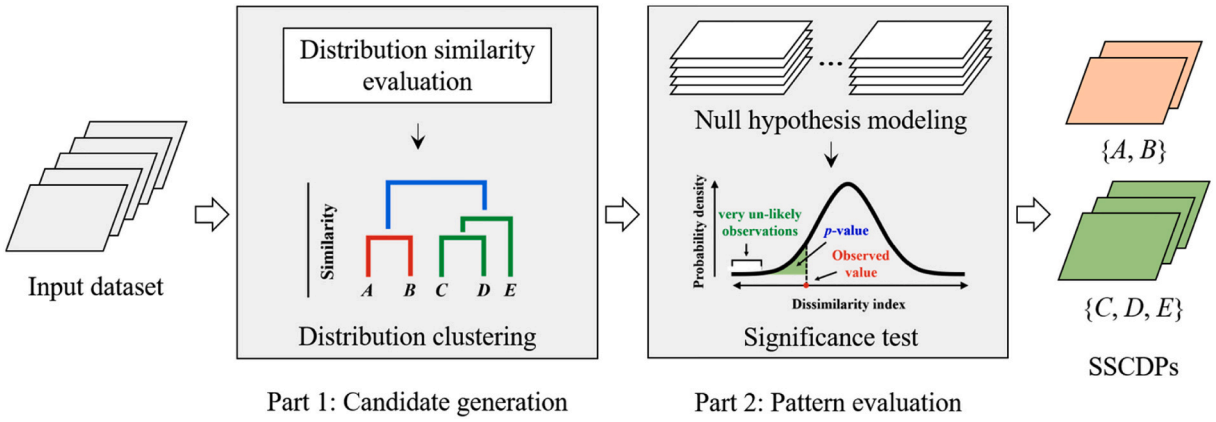

Fig. 3. Framework of the proposed strategy for discovering SSCDPs.

between different features (spatial cross-correlation), while maintaining the spatial structure of each feature (spatial auto-correlation).

Fig. 3 presents the framework of the proposed strategy. In the following section, we initially introduce the key concepts, and then formally define the SSCDP discovery problem.

3. Problem formulation

3.1. Basic concepts

A ***spatial feature*** refers to the conceptual abstraction of a set of spatial points of the same category. An ***instance of a spatial feature*** refers to a spatial point labeled with that feature category. The proximity relationship among different features is modeled using a ***spatial neighbor relation*** (Shekhar & Huang, 2001) in which the distance in space is restricted by a distance threshold (see Fig. 4). An instance of a feature f_i is considered to occur around a location l_j if this instance is a spatial neighbor to l_j. Here a location l_j belongs to a set of baseline locations where the spatial features could occur. The baseline locations can be specified with some reference points (e.g., centers of artificial grids or census units). However, the user-specified baseline locations may break the spatial neighbor relation among instances of different features around adjacent baseline locations. Thus, in this study, the ***baseline locations of a feature*** correspond to the locations that host all instances of that feature, and the ***baseline locations of a candidate SSCDP*** are the collection of baseline locations of all the component features. A ***spatial co-distribution pattern*** (SCDP) is generally a subset of spatial features whose instances not only frequently occur around the same baseline locations, but also have similar distributions (probabilities) around these locations.

It is essential to select an appropriate neighbor distance threshold to ensure meaningful SCDPs. In practice, without prior knowledge provided by domain experts, the spatial auto-correlation of the input dataset can serve as guidelines for the neighborhood determination (Yoo & Bow, 2012). It is suggested to set the neighbor distance threshold as one distance where spatial processes greatly promote clustering measured by a modified L function dealing with multi-type points, that is, a distance that shows large positive deviation between observed L function value and expected value under complete spatial randomness. Alternatively, the spatial neighbor relation can be defined using mutual k-nearest-neighbor graph, topological proximity criterion (Nilsson & Smirnov, 2017) and so on. The mutual k-nearest-neighbor graph connects two instances of different features if they are k-nearest neighbors of different categories to each other (see Fig. 5(a)). The topological proximity criterion considers an instance of feature f_i to be located relatively close to an instance of feature f_j if that instance is located within the relative attraction areas of f_j. The relative attraction areas are constructed by connecting all midpoints between each instance of f_j and the vertexes of the Thiessen polygon around that instance (see Fig. 5(b)). This criterion is a good option in cases where there is a very uneven spatial distribution of points.

3.2. Quantifying the difference in spatial distributions

To compare the spatial distributions among included features in a SCDP, the spatial distribution of each spatial feature is modeled by the

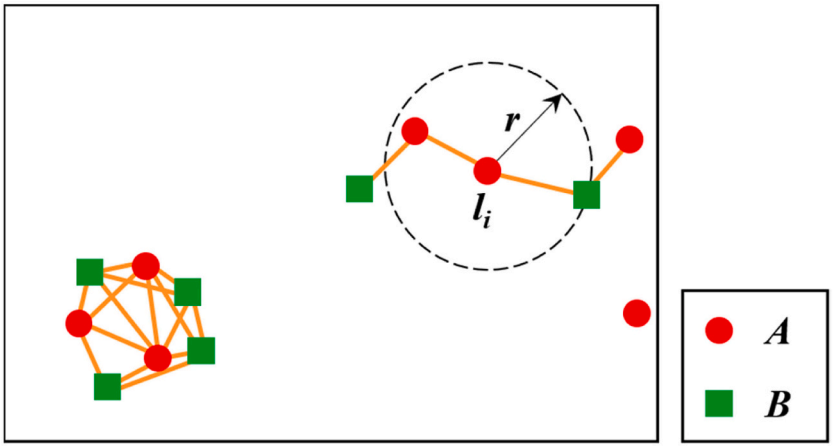

Fig. 4. Illustrative spatial co-distribution pattern {A, B}.

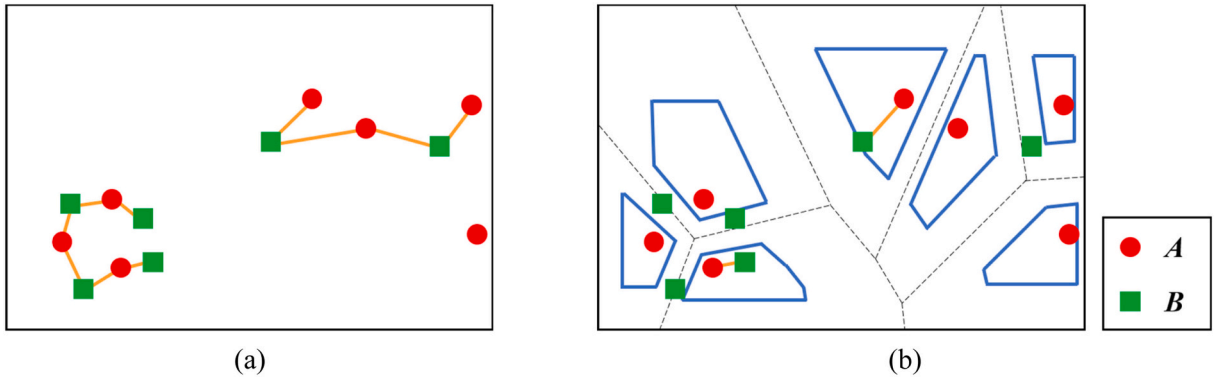

Fig. 5. Alternative definitions of the spatial neighbor relation: (a) mutual k-nearest-neighbor graph with k of 2 and (b) relative attraction areas.

rate it occurs in close spatial proximity to each baseline location of the SCDP. The dissimilarity among spatial distributions of different features in the SCDP is then measured by the difference in the occurrence rate of features around all baseline locations. These terms are formally defined as follows:

Definition 1. The ***occurrence rate*** $OR(f_i, l_j)$ of a feature f_i around a baseline location l_j is the probability of finding an instance of f_i in close spatial proximity to l_j under the spatial neighbor relation R. Formally, $OR(f_i, l_j)$ can be written as

$$OR(f_i, l_j) = \frac{|I(f_i, l_j)|}{|I(f_i)|} \tag{1}$$

where $|I(f_i, l_j)|$ is the number of instances of f_i occurring around l_j, and $|I(f_i)|$ is the number of all instances of f_i in the entire study area.

Definition 2. For a baseline location l_j, the ***local distribution difference*** $LDD(\text{SCDP}, f_i, l_j)$ between a feature f_i and all features in a $\text{SCDP} = \{f_1, f_2, \cdots, f_k\}$ is defined as the difference between the occurrence rate of f_i and the mean occurrence rate of all k features in the SCDP, represented as

$$LDD(\text{SCDP}, f_i, l_j) = OR(f_i, l_j) - \frac{\sum_{i=1}^{k} OR(f_i, l_j)}{k}, \quad f_i \in \text{SCDP} \tag{2}$$

Definition 3. The ***distribution difference*** $DD(\text{SCDP}, f_i)$ between a feature f_i and other features in a SCDP is defined as the quadratic mean of $LDD(\text{SCDP}, f_i, l_j)$ over all the baseline locations of the SCDP, $L = \{l_1, l_2, \cdots, l_n\}$, computed as

$$DD(\text{SCDP}, f_i) = \sqrt{\frac{\sum_{j=1}^{n} \{LDD(\text{SCDP}, f_i, l_j)\}^2}{n}}, \quad l_j \in L \tag{3}$$

Definition 4. The ***dissimilarity index*** $DI(\text{SCDP})$ of a $\text{SCDP} = \{f_1, f_2, \cdots, f_k\}$ is defined as the mean distribution difference between one feature and all features in the SCDP, formally written as

$$DI(\text{SCDP}) = \frac{\sum_{i=1}^{k} DD(\text{SCDP}, f_i)}{k}, f_i \in \text{SCDP} \tag{4}$$

The dissimilarity index of a SCDP is always nonnegative, and a smaller dissimilarity index value indicates more similar spatial distributions among different features. The value is zero only if all the features included in the SCDP have the same occurrence rates around all the baseline locations (see Fig. 2(c)).

Consider the illustrative dataset in Fig. 4. There are two different spatial features, A and B, which have seven and six instances, respectively. Given a distance threshold, the spatial neighbor relation R over this dataset is represented by the lines connecting the instances of A and B. The baseline locations of pattern $\{A, B\}$ are the locations of these 13 instances. For the baseline location l_i, the occurrence rates of feature A, $OR(A, l_i)$, and feature B, $OR(B, l_i)$, are 2/7 and 1/6, respectively. The mean occurrence rate of the features in pattern $\{A, B\}$ at l_i is $(2/7 + 1/6)/2 \approx 0.23$. Thus, the local distribution difference between features in $\{A, B\}$ at l_i, i.e., $LDD(\{A, B\}, A, l_i)$ and $LDD(\{A, B\}, B, l_i)$, can be computed as $2/7 - 0.23 \approx 0.06$ and $1/6 - 0.23 \approx -0.06$, respectively. After traversing all the baseline locations of $\{A, B\}$, we obtain distribution differences $LDD(\{A, B\}, A)$ and $LDD(\{A, B\}, B)$, both of which are 0.066. Thus, the dissimilarity index of $\{A, B\}$, $DI(\{A, B\})$ is $(0.066, 0.066)/2 = 0.066$.

3.3. Significance test preliminaries

Although the dissimilarity index (DI) proposed in Section 3.2 can measure the degree of difference between multivariate spatial distributions, it does not necessarily indicate positive or negative interactions among different features. In practice, SCDPs decided with a DI value threshold may associate absolutely independent features while ignoring strong dependencies. To remove such spurious patterns and enhance the statistical interpretability, the significance of a SCDP is validated through a statistical test under the null hypothesis of independence, which is described as follows.

Definition 5. To test the significance of a $\text{SCDP} = \{f_1, f_2, \cdots, f_k\}$, the ***null hypothesis of independence*** (H_0) states that the univariate spatial distribution of each feature f_i is independent of the other features in the SCDP.

Furthermore, the test statistic for this test is the measure DI. However, it is analytically intractable to determine the theoretical distribution of DI under H_0. Therefore, a more practical alternative is to estimate the empirical distribution of DI using Monte Carlo permutations. What is of interest in this significance test is the relationship among different spatial features and not the univariate spatial structure of each feature (Wiegand & Moloney, 2013). Thus, the permutations of the observed dataset generated under H_0 should satisfy the following two properties (Dixon, 2002): (1) the potential interactions among different features must be broken, and (2) the observed spatial structure of each feature (e.g., spatial auto-correlation) should be maintained. Once a sufficient number of such permutations are generated, a p-value is used to assess whether a SCDP deviates from H_0, and a significant spatial co-distribution pattern can be formally defined.

Definition 6. Given a large number N of permutations under H_0, the ***p-value*** of a SCDP, $p - \text{value}(\text{SCDP})$, is defined as the probability of finding a DI value from permutations $DI_n^{null}(\text{SCDP})$ smaller than or equal to the observed DI value $DI^{obs}(\text{SCDP})$, computed as

$$p - \text{value}(\text{SCDP}) = \frac{|DI_n^{\text{null}}(\text{SCDP}) \le DI^{obs}(\text{SCDP})| + 1}{N + 1}, \quad n = 1, 2, \cdots, N \tag{5}$$

Definition 7. Given a threshold α of a p-value (known as a significance level and conventionally specified as 0.05 or 0.01), the null hypothesis of independence is rejected and a SCDP is defined as a ***significant spatial co-distribution pattern*** (SSCDP) if its p-value is not greater than α.

3.4. Formal problem statement

Given a large number of spatial features, the number of reported SSCDPs is typically too large and unmanageable for humans to interpret, and the patterns usually contain many redundant descriptions of the correlations between features. Thus, in this study, the desired SSCDPs are represented by clusters of features to achieve an informative and nonredundant representation of the pattern collection as a whole. Based on this consideration, the SSCDP discovery problem is formally defined as follows:

Given: (1) A collection of spatial features and their instances; (2) a distance threshold r for defining the spatial neighbor relation R; and (3) a significance level α;

Find: SSCDPs represented by clusters of spatial features with p-values $\leq \alpha$;

Objective: Effectively establish the statistical significance of SSCDPs with fewer assumptions;

Constraints: The underlying spatial distribution of each feature is a priori unknown.

Fig. 6 provides an illustration of SSCDP discovery. Given a dataset containing instances of five features, a neighbor relation defined based on a distance threshold and a significance level of 0.05, two clusters of features, {A, B} and {C, D, E}, are reported as SSCDPs because they have qualifying p-values of 0.01 and 0.03, respectively.

4. SSCDP discovery method

We now describe the three main components of the SSCDP discovery method: (1) generation of candidate SSCDPs, (2) construction of the null hypothesis of independence, and (3) statistical significance test.

4.1. Distribution clustering method for generating candidate SSCDPs

Given K types of spatial features, the number of all possible co-distribution patterns is exponentially related to K. If K is small, it may be computationally feasible to perform brute-force evaluation on all possible patterns; otherwise, it is more practical to prioritize the most promising candidates whose component features share the most similar spatial distributions. For example, in Fig. 6(a), compared with other features, the spatial distribution of feature A is more similar to that of feature B. Pattern {A, B} may then be a promising SSCDP candidate. Thus, we model the generation of candidate SSCDPs as a special clustering problem for hierarchically grouping spatial features, where the similarity measure is defined based on the similarity between spatial distributions of features.

Specifically, the proposed distribution clustering method identifies clusters of spatial features in a divisive (i.e., top-down) manner. The reasons are twofold: First, the top-level clustering decisions made by divisive methods are based on more global properties, thus producing more accurate hierarchies than agglomerative (i.e., bottom-up) methods, which rely on local properties from the bottom level (Guha, Rastogi, & Shim, 2000; Steinbach, Karypis, & Kumar, 2000). Second, a top-down search allows us to stop the clustering process without searching for smaller-size patterns (i.e., clusters at lower levels), if the larger-size patterns (i.e., clusters at higher levels) are guaranteed to be significant. This is because larger-size patterns can provide a more informative description than smaller patterns regarding the co-distributions of features.

The distribution clustering method can be regarded as a recursive application of partitioning clustering at multiple levels. The method starts at the topmost level with all spatial features in one cluster. Subsequently, each feature cluster FC at the ith level is identified as a candidate SSCDP whose validity will be verified using the significance test presented in the following sections. If FC is insignificant, we implement a distribution bisection method to split the FC into two sub-clusters as follows:

(1) Initial the first representative feature rf_1 as the one which has the smallest sum of the dissimilarity index (DI) values to all other features in FC. Taking the dataset in Fig. 6 for example, feature C will be selected as the initial rf_1.
(2) Tentatively choose each unselected feature as the second representative feature rf_2. For any other non-representative feature nf_i, assign it to its most similar representative feature rf_j (j = 1 or 2) if $DI(\{nf_i, rf_j\})$ is smaller than the DI value between nf_i and another representative feature. Measure the quality of the clustering result with the total DI value (TDI) of two feature sub-clusters, FSC_1 and FSC_2, represented as:

$$TDI = DI(FSC_1) + DI(FSC_2) \quad (6)$$

Then, determine the initial rf_2 as the one that brings the smallest TDI value. As shown in Fig. 7, after testing all the unselected features (i.e., A, B, D and E), feature A or B will be selected as the initial rf_2 because both of them can produce the best clustering quality.

(3) Swap each non-representative feature, nf_i, tentatively with the representative feature rf_j (j = 1 or 2) to which it is assigned and reassign rf_j and other non-representative features to their most similar new representative features. Record the TDI value after the reassignment using each nf_i.
(4) Finally, check whether the clustering result can be improved, i.e., whether the current TDI value can be reduced. If so, carry out the

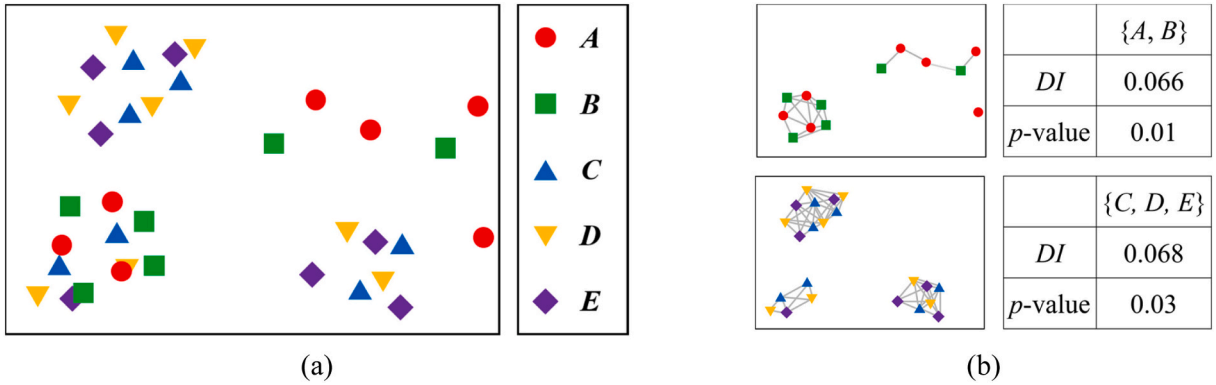

Fig. 6. Illustration of the SSCDP discovery problem: (a) input: a simulated dataset and (b) output: two clusters of features: {A, B} and {C, D, E}.

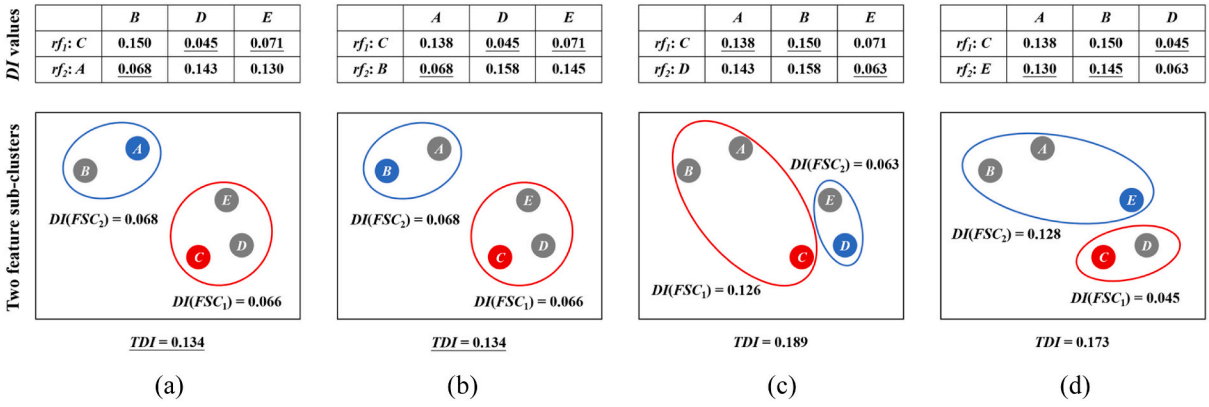

Fig. 7. Initialization of the second representative feature: (a) testing feature *A*; (b) testing feature *B*; (c) testing feature *D* and (d) testing feature *E*.

reassignment using the feature nf_i which reduces the current *TDI* value the most, and return to Step 3; otherwise, terminate the clustering process, and output the two sub-clusters FSC_1 and FSC_2.

The distribution bisection method is repeatedly implemented on each insignificant feature cluster to find the most promising candidate SSCDPs at the next level unless only one or two spatial features are contained in that cluster. In this study, significant feature clusters will not be further split because we want to find the most informative and nonredundant set of patterns that can statistically explain the spatial correlations between features. If users are interested in more local information, the distribution bisection method can also be implemented on significant feature clusters to identify lower-level clusters with fewer features.

4.2. Distribution shift-correction method for modeling the null hypothesis of independence

Modeling the null hypothesis of independence is the premise for the development of our significance test on candidate SSCDPs. The key here is to generate permutations of the observed dataset that satisfy the two properties described in Section 3.3 (i.e., the randomization of relationships between different features and the maintenance of the spatial structure of each feature). It is straightforward to fulfill the first property by repositioning the instances of each feature regardless of the others. To ensure the second property, predefined point-process models or statistical properties are usually required. A simple alternative is to entirely shift all the points by considering the study area as a torus (Lotwick & Silverman, 1982); however, this may produce some artificial structures at the edges. To overcome these limitations, we propose a distribution shift-correction method.

For each spatial feature f_i, the distribution shift-correction method generates permutations in three major phases: (1) shifting, (2) relocation and (3) correction, described as follows:

(1) *Shifting*. In the shifting phase, all the instances $I(f_i)$ of f_i are shifted as a whole across the study area $S = [0, X] \times [0, Y]$ by adding a fixed random vector (dx, dy) to each location of $I(f_i)$ (Fig. 8(b)).
(2) *Relocation*. The relocation phase aims to relocate the instances of f_i that are shifted outside the study area S by following the rules of toroidal geometry. This is achieved by subtracting a value of X from the x-coordinates of instances lying within the right area of S (see regions R_2 and R_4 in Fig. 8(b)), and subtracting a value of Y from the y- coordinates of instances lying within the upper area of S (see regions R_3 and R_4 in Fig. 8(b)).
(3) *Correction*. This phase is to correct the univariate spatial structures of each feature under the null hypothesis because they are not of interest in the test of dependence among different features. In spatial statistics, univariate spatial structures are quantified with summary statistics (e.g., *K*-function (Ripley, 1976)) by summarizing the statistical properties of each feature. The spatial neighbor relation is the basis of these summary statistics. So, the real focus here is on the correction of artificial spatial neighbor relation produced by the first two phases, including missing neighbor relation or emerging neighbor relation.
   - Missing structures lie close to the edges of S and are artificially split apart in the shifted and relocated dataset (e.g., the neighbor relation represented by dotted lines in Fig. 8(b)). Such structures will cause underestimation of the local statistical properties of f_i. Thus, the spatial structures near the edges need to be corrected based on the fixed neighborhoods (see yellow regions in Fig. 8(d)).
   - Emerging structures occur near the borders between shifted and relocated regions (e.g., the neighbor relation represented by the black line in Fig. 8(c)). Such structures will lead to overestimation of the local statistical properties of f_i. Thus, the neighbor relation in these emerging artifacts needs to be broken to refine the statistical properties of f_i (see the green regions in Fig. 8(d)).

It should be noted that the distribution shift-correction method does not directly generate permutations that have the same spatial structure as the observation; rather, it allows the structure to be recovered from the permutations through structure correction (see Appendix A for details). In the following section, we explain the role played by the recovered spatial structure of each feature in the estimation of distribution similarity between different features under the null hypothesis.

4.3. Statistical significance test for identifying SSCDPs

To test the statistical significance of each candidate SSCDP, we need to determine the null distribution of the test statistic (i.e., the probability distribution of the dissimilarity index (*DI*) when the null hypothesis is true). To provide a good estimator for the null distribution, a sufficient number of permuted datasets, each of which contains instances of different features, are generated using the distribution shift-correction method described above. The null distribution is then

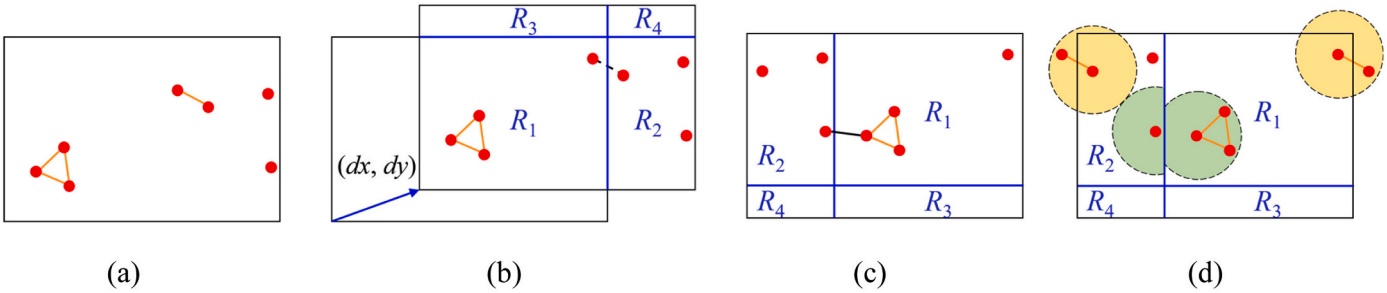

Fig. 8. Generation of a permutation based on the distribution shift-correction method: (a) observed dataset; (b) shifted dataset; (c) relocated dataset and (d) permuted dataset with corrected neighbor relation.

estimated by evaluating the *DI* values in these permuted datasets. However, the artificial structures of each feature in the permuted datasets will also cause significant bias in the estimator of permuted *DI* values. In light of this, we implement a distribution recovery method to recover the observed spatial structure of each feature in the estimator of permuted *DI* values.

For a candidate pattern $CP = \{f_1, f_2, \cdots, f_k\}$ in a permuted dataset that is embedded in the study area $S = [0, X] \times [0, Y]$, the distribution recovery method first corrects the permuted occurrence rates (*OR*) of each included feature f_i around the baseline locations of *CP* (i.e., the locations of permuted instances of all features included in *CP*) in two cases:

- For the permuted baseline locations L^i of f_i, the permuted $OR(f_i, l_m)$ ($l_m \in L^i$) is directly replicated from that around the corresponding location in the observed dataset.
- For the permuted baseline locations L^o of other features in *CP*, L^o are relocated to the context of the observed instances of f_i, and the permuted $OR(f_i, l_n)$ ($l_n \in L^o$) is corrected by estimating the *OR* value of observed f_i around the corresponding location in the relocated L^o. The relocation of L^o is a reverse process of the generation of permuted instances of f_i. This is achieved by adding a value of X to the x-coordinates of locations whose x-coordinates lie within $[0, dx]$ (see locations in regions R_2 and R_4 in Fig. 9(b)), adding a value of Y to the y-coordinates of locations whose y-coordinates lie within $[0, dy]$ (see locations within regions R_3 and R_4 in Fig. 9(b)), and then entirely shifting L^o according to the vector $(-dx, -dy)$ (Fig. 9(c)), where (dx, dy) is the random vector used for generating permutations of f_i.

Subsequently, the *DI* value of *CP* in the permuted dataset is refined using the corrected *OR* values of all included features. For example, Fig. 9(a) shows a permuted dataset containing shifted instances of features *A* and *B* and Fig. 9(d) presents the dataset containing recovered instances of *A* and *B* according to vector (dx, dy) for shifting observed instances of *A*. The permuted *OR* values of *A* are corrected with the *OR* values of *A* computed at the corresponding recovered locations in the dataset shown in Fig. 9(d). Similarly, the permuted *OR* values of *B* can be corrected by applying the distribution recovery method according to the vector (dx', dy') for shifting observed instances of *B*. The permuted *DI* value of $\{A, B\}$ can be further refined by considering the observed spatial structures of *A* and *B*.

After obtaining the permuted *DI* values of *CP* in a sufficient number of permuted datasets, the statistical significance of *CP* is determined according to the *p*-value in Eq. (5). In practice, to further accelerate the computation, some unnecessary evaluations of permuted *DI* values can be terminated early if the permuted *DI*(*CP*), which is not higher than the observed *DI*(*CP*), has already been found in so many permuted datasets that the significance level α cannot be met.

4.4. Implementation and analysis of the SSCDP discovery method

4.4.1. Algorithm description

As shown in Algorithm 1, the proposed method identifies SSCDPs using the following steps:

Step 1: Generate *N* permuted datasets that conform to the null hypothesis using the distribution shift-correction method introduced in Section 4.2 (line 1).

Step 2: Identify the pattern formed by all spatial features $\{f_1, f_2, \ldots, f_K\}$ as the candidate CP_1 at the first level and test the statistical significance of that pattern as described in Section 4.3 (lines 2–3).

Step 3: If any insignificant pattern containing more than two features is identified from CP_n at the *n*th level, generate the candidate CP_{n+1} at the next level using the distribution bisection method introduced in Section 4.1 (line 5); otherwise, terminate the algorithm and output all the significant patterns (line 14).

Step 4: Test the statistical significance of candidates in CP_{n+1} (line 6). Assign $n + 1$ to n and return to Step 3 to check whether the algorithm needs to continue.

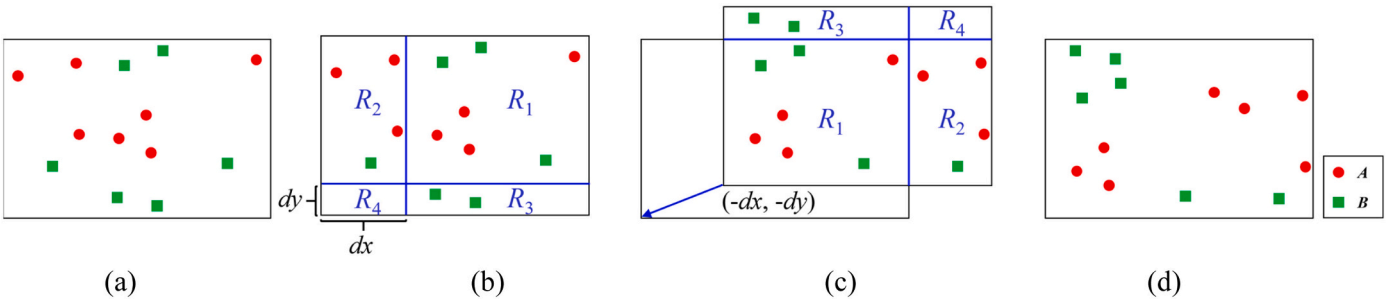

Fig. 9. Distribution recovery method for refining the permuted occurrence rates of feature *A*: (a) a permuted dataset containing shifted instances of *A* and *B*; (b) subregions for relocation; (c) relocated instances for shifting and (d) a dataset containing recovered instances.

Algorithm 1. SSCDP discovery

Input:

(a) An observed dataset OD containing instances of K spatial features $f_1, f_2, \ldots, f_K$;

(b) A distance threshold r for defining the spatial neighbor relation R;

(c) A significance level α.

Output:

SSCDPs represented by clusters of features with p-values $\leq \alpha$

Variables:

n: the level of clustering;

PD_s: permuted datasets, each of which contains permutations of K features;

CP_n: set of candidate patterns at the nth level of clustering;

SP_n: set of significant patterns at the nth level of clustering.

IP_n: set of insignificant patterns at the nth level of clustering

Method:

1. Generate PD_s using the distribution shift-correction method;
2. $n = 1$; $CP_1 = (\{f_1, f_2, \ldots, f_K\})$; // Initialization
3. Identify SP_1 and IP_1 using the significance test;
4. **while** (isempty(IP_n) = FALSE) **do**
5. Generate CP_{n+1} from IP_n using the distribution bisection method;
6. Identify SP_{n+1} and IP_{n+1} using the significance test;
7. **for each** pattern c in IP_{n+1} **do**
8. **if** size(c) $\leq$ 2 **then**
9. Delete c from IP_{n+1};
10. **end if**
11. **end for**
12. $n = n + 1$;
13. **end while**
14. Return union($SP_1, \ldots, SP_n$).

Consider the dataset in Fig. 6 as an example. First, 99 Monte Carlo permutations are generated for each feature using the distribution shift method. The first candidate for significance test is the pattern that contains all the features, pattern {A, B, C, D, E}. That pattern is found to be insignificant (p-value = .96) and rejected. At the next level, two candidate patterns {A, B} and {C, D, E} are identified using the distribution bisection method. Both {A, B} and {C, D, E} are further reported as SSCDPs at the significance level of 0.05 because p-value({A, B}) = 0.01 and p-value({C, D, E}) = 0.03.

4.4.2. Computational complexity analysis

The distribution shift in Step 1 does not need to iteratively modify the parameters of the point-process model or locations of points in the permutation. It generates each permutation at once. Thus, the cost for generating N permuted datasets of K spatial features can be simplified to $O(N \cdot K)$. The cost of Step 2 is mainly due to the construction of neighbor relation R between instances of all the features in both the observed and permuted datasets for estimating the test statistic, which requires $O(N \cdot M \cdot logM)$ time. Here, M is the total number of instances of all spatial features. Then, for each l-size insignificant pattern in CP_n, Step 3 requires $O(r \cdot l^2)$ time to obtain two sub-patterns, where r is the number of iterations in the swapping phase to find two optimal representative features. In Step 4, the neighbor relation R does not need to be rebuilt. Thus, this step only requires $O(N \cdot X)$ time, where X is the number of candidates in CP_{n+1}. Assuming that the mean size of the candidates is $\overline{L}$ and the mean number of iterations for splitting a candidate pattern is $\overline{R}$, then the proposed method has a time complexity of approximately $O(N \bullet M \bullet logM) + O(K \bullet \overline{R} \bullet \overline{L}^2) + O(N \bullet K)$ in the worst case, i.e., when all the identified candidates are evaluated using the significance test.

5. Experimental evaluation and case study

5.1. Experimental evaluation using synthetic datasets

In the experimental evaluation, our aim was to answer the following four questions:

Q1: How well does our method capture the similarity between spatial distributions of different features compared to the state-of-the art methods?

Q2: Does our method outperform other methods in the effectiveness of modeling the null hypothesis of independence?

Q3: How do different choices of input parameters affect the

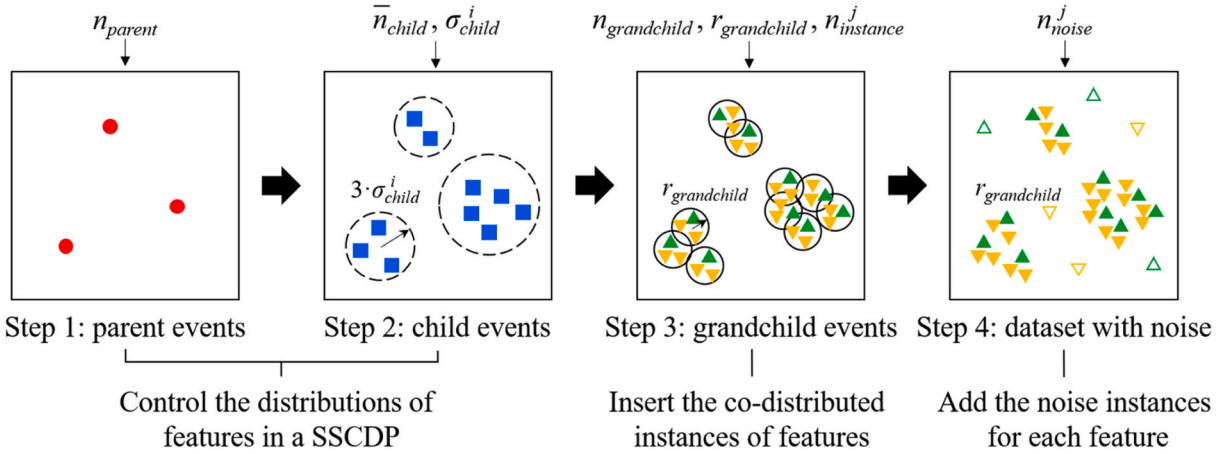

Fig. 10. Generation of a predefined SSCDP.

Table 2
Parameters used to generate the synthetic SSCDPs.

Parameter	Definition	Pattern1	Pattern2	Pattern3
n_{parent}	The number of parents	4	3	2
$\bar{n}_{child}$	The average number of children around each parent	50	65	100
σ_{child}^{i}	The standard deviation to generate children around ith parent	[5, 5, 5, 5]	[6, 8, 10]	[6, 10]
$n_{grandchild}$	The number of grandchildren around each child	3	3	3
$r_{grandchild}$	The radius to generate grandchildren around each child	2	2	2
$n_{instance}^{j}$	The number of instances of jth feature around each child	[1,2,3]	[1,2,3]	[1, 2, 3]
n_{noise}^{j}	The number of noise instances added for jth feature	[50, 50, 50]	[50, 50, 50]	[50, 50, 50]

Table 3
Description of different features in the synthetic dataset.

Feature ID	Description	Instance number of each feature
[1, 2, 3]	Co-distributed with each other	[250, 450, 650]
[4, 5, 6]	Co-distributed with each other	[245, 440, 635]
[7, 8, 9]	Co-distributed with each other	[250, 450, 650]
[10,11,12]	Randomly distributed	[300,300,300]

performance of our method?

Q4: How does the performance of our method vary with the size of the input dataset?

To answer these questions, we compared our Co-Distribution discovery method using distribution Shift and Correction (denoted by CDSC) with the Co-Location pattern discovery method using Model Fitting (denoted by CLMF) (Barua & Sander, 2014) because of their similar purposes to identify statistically significant spatial associations among multiple features. In addition, to determine the independent

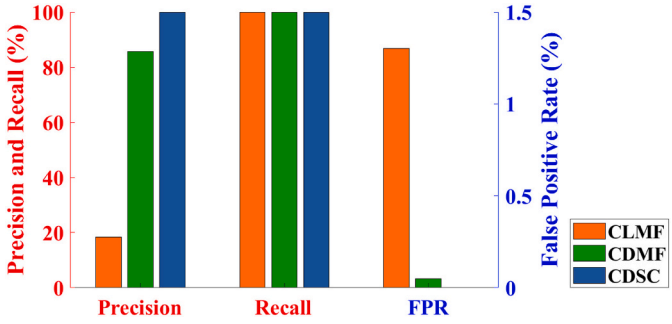

Fig. 12. Precision, recall and false positive rate of different methods on the synthetic dataset.

improvements on the test statistic and null model, we also tested a variant of the CLMF method (denoted by CDMF) using our dissimilarity index as the test statistic to discover Co-Distribution patterns. For all the methods, the significance level was set to 0.05 following the

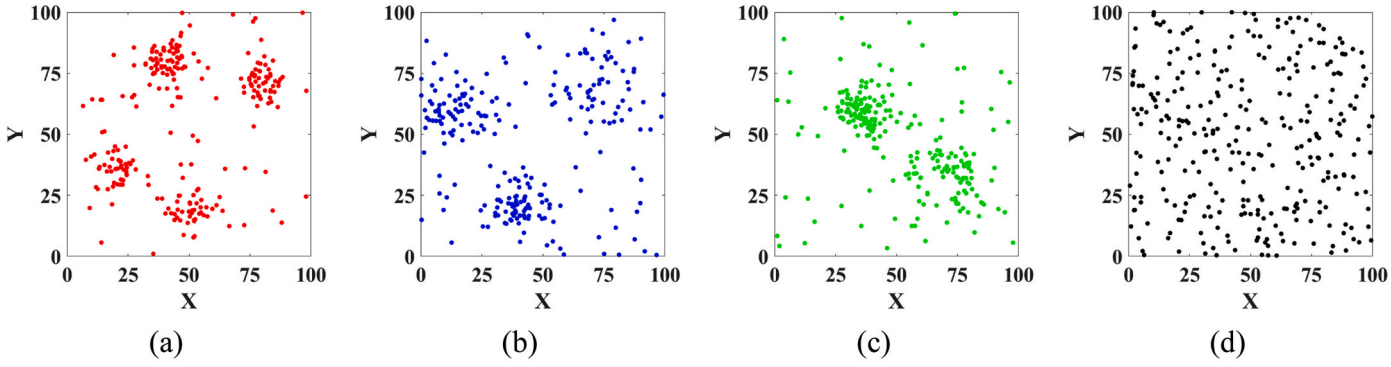

Fig. 11. Distributions of representative features in the synthetic dataset: (a) feature 1; (b) feature 4; (c) feature 7 and (d) feature 10.

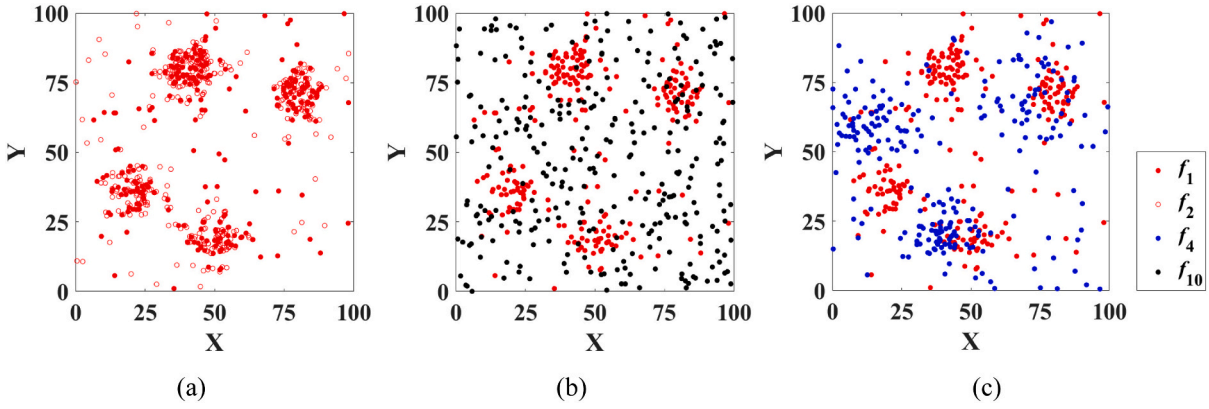

Fig. 13. Examples of co-location and co-distribution patterns detected from the synthetic dataset: (a) co-location and co-distribution; (b) not co-location and not co-distribution and (c) co-location but not co-distribution.

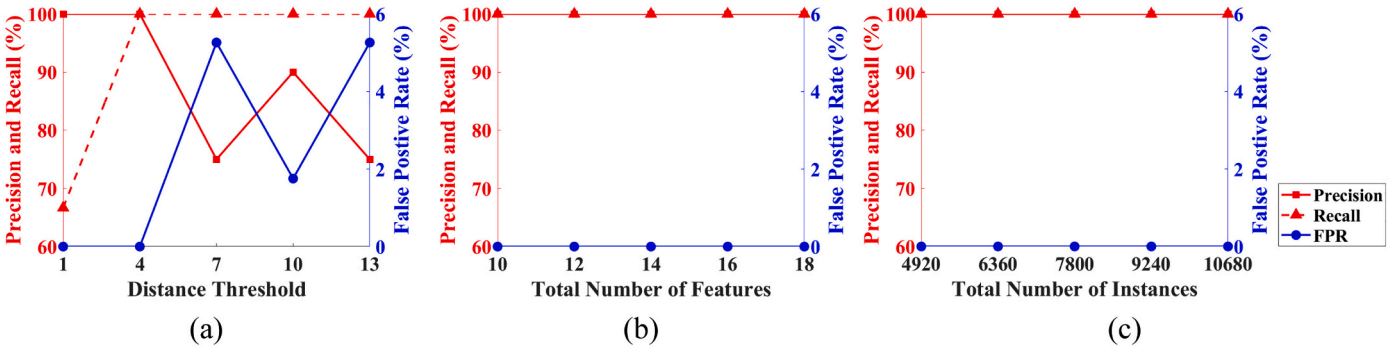

Fig. 14. Precision, recall and false positive rate of our method varied with different factors: (a) distance threshold; (b) number of features and (c) number of instances.

Table 4
Overview and explanation of the public health dataset in central Colorado.

ID	Feature type	Definition	Instance number
1	*Asthma*	Highest quintile of percent of adults who currently have asthma	136
2	*BingeDrink*	Highest quintile of percent of adults who are binge drinking	135
3	*DelayedMC*	Highest quintile of percent of adults who delayed medical care because of cost	130
4	*Diabetes*	Highest quintile of percent of adults ever diagnosed with diabetes	128
5	*HeartDisease*	Highest quintile of percent of adults ever diagnosed with heart disease	121
6	*HeavyDrink*	Highest quintile of percent of adults who are heavy drinking	129
7	*MentalDistress*	Highest quintile of percent of adults with frequent mental distress	136
8	*NoCheckUp*	Highest quintile of percent of adults with no routine medical checkup	137
9	*NoPhysAct*	Highest quintile of percent of adults that did not report doing physical activity	122
10	*Obese*	Highest quintile of percent of adults who are obese (BMI ≥ 30)	130
11	*Overweight*	Highest quintile of percent of adults who are overweight or obese (BMI ≥ 25)	134
12	*PhysDistress*	Highest quintile of percent of adults with frequent physical distress	124
13	*PoorHealth*	Highest quintile of percent of adults with fair or poor health status	125
14	*Smoking*	Highest quintile of percent of adults who currently smoke cigarettes	138

BMI: Body Mass Index is a person's weight in kilograms divided by the square of height in meters.

convention in statistics, and the number of Monte Carlo permutations was set to $5/\alpha - 1 = 99$ according to Besag and Diggle (1977).

5.1.1. Experimental setup

The SSCDPs in the synthetic dataset were predefined using the generator, as shown in Fig. 10. The first two steps set the main distribution of features for each SSCDP. Here, the child sets in different clusters follow a bivariate normal distribution with different standard deviation ${\sigma_{child}}^i$. Then, the instances of $n_{grandchild}$ spatial features are designed to be co-distributed at a distance of $2r_{grandchild}$. Finally, ${n_{noise}}^j$ noise instances are also included for each feature.

Using the generator, we obtained three predefined SSCDPs: {1,2,3}, {4, 5, 6} and {7, 8, 9}. Tables 2 and 3 summarize the generation parameters and statistical information of these patterns, respectively. Fig. 11 presents the spatial distributions of one representative feature included in each SSCDP. In addition, three randomly distributed features with 300 instances, i.e., features 10, 11 and 12, were also inserted into the synthetic dataset to interfere with the identification of meaningful patterns.

The algorithm performance was evaluated based on the precision, recall and false positive rate (*FPR*) of results. These measures were computed as:

$$precision = \frac{|TP|}{|TP| + |FP|} \tag{7}$$

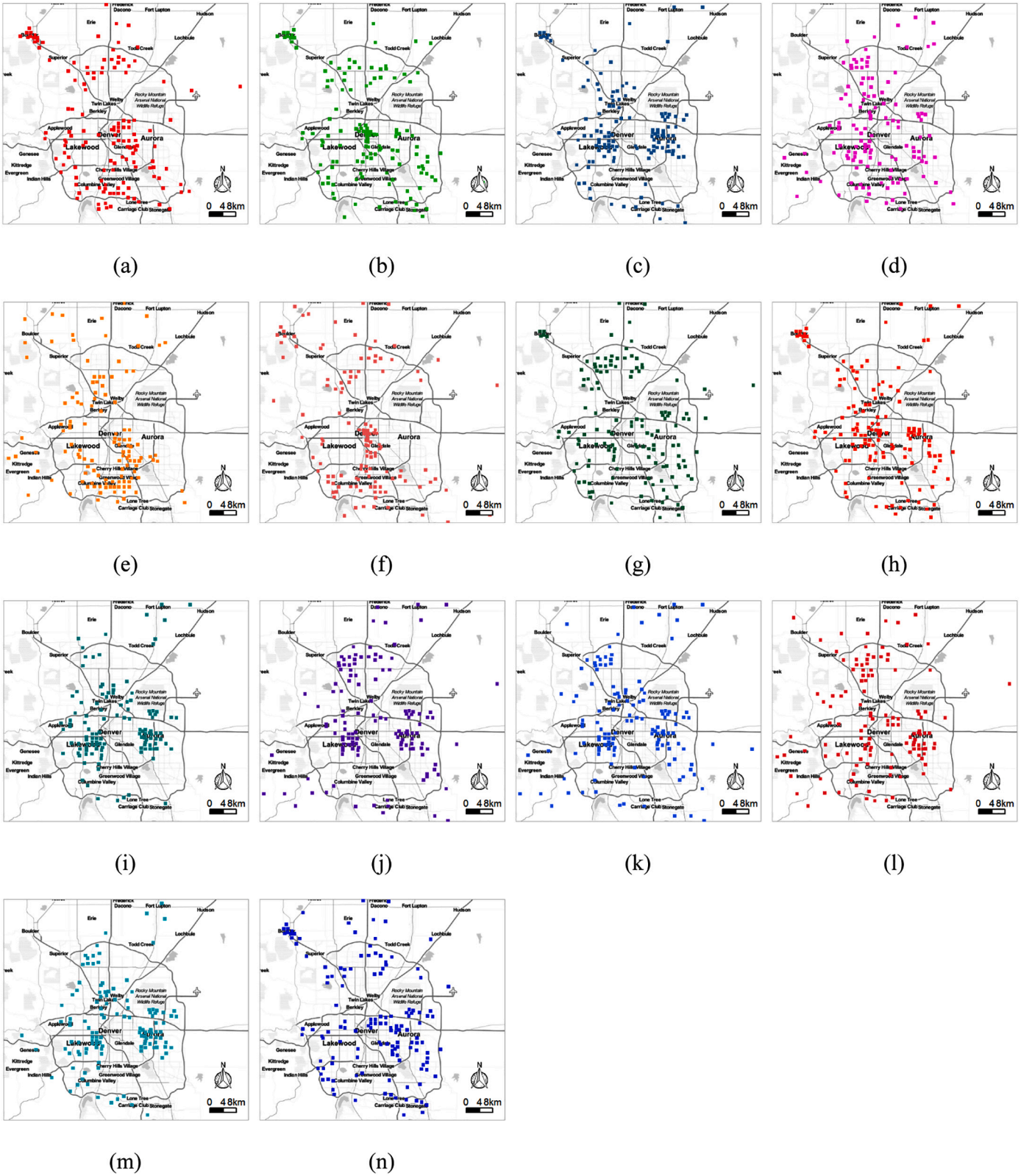

Fig. 15. Distributions of public health dataset: (a) *Asthma*; (b) *BingeDrink*; (c) *DelayedMC*; (d) *Diabetes*; (e) *HeartDisease*; (f) *HeavyDrink*; (g) *MentalDistress*; (h) *NoCheckUp*; (i) *NoPhysAct*; (j) *Obese*; (k) *Overweight*; (l) *PhysDistress*; (m) *PoorHealth* and (n) *Smoking*.

Table 5
SSCDPs detected from the public health dataset.

Pattern size	SSCDPs	*DI*	*p*-value
2	{*BingeDrink*, *HeavyDrink*}	7.7×10^{-3}	0.01
3	{*DelayedMC*, *NoPhysAct*, *PoorHealth*}	6.9×10^{-3}	0.01
6	{*Diabetes*, *MentalDistress*, *Obese*, *Overweight*, *PhysDistress*, *Smoking*}	9.9×10^{-3}	0.01

$$recall = \frac{|TP|}{|TP| + |FN|} \tag{8}$$

$$FPR = \frac{|FP|}{|FP| + |TN|} \tag{9}$$

where, $|TP|$, $|FP|$, $|FN|$ and $|TN|$ are the number of true positives, false positives, false negatives and true negatives, respectively, compared with the ground truth. For the CLMF and CDMF methods, the ground truth was all subsets of patterns {1, 2, 3}, {4, 5, 6} and {7, 8, 9}. For our CDSC method, we wanted to know if all the clustering structures of features were correctly identified. Thus, the ground truth is the pairs of features in the same predefined clusters of features. *TP* and *FP* are pairs of features that are correctly and incorrectly assigned to the same cluster, respectively. *FN* and *TN* are pairs of features that are incorrectly and correctly assigned to different clusters, respectively.

5.1.2. Comparative analysis

5.1.2.1. Performance on capturing the distribution similarity (Q1). To see the performance of different methods on capturing the distribution similarity, we compared the results obtained by CLMF and our modified CDMF method, which use the participation index and dissimilarity index as the test statistic, respectively. As shown in Fig. 12, both methods identify all the predefined SSCDPs as co-location patterns or co-distribution patterns with recall of 1 (e.g., pattern {1,2} in Fig. 13(a)). Both methods also do a good job of ignoring the interference from random features (e.g., pattern {1,10} in Fig. 13(b)). However, CLMF outputs many other spurious co-distribution patterns as co-location patterns (e.g., pattern {1, 4} in Fig. 13(c)). This is due to the high overlaps of included features. In comparison, by using the dissimilarity index, our modified CDMF method can distinguish the spatial distributions of these features with much higher precision of 85.7% and lower *FPR* of 0.05%.

5.1.2.2. Performance on modeling the null hypothesis (Q2). To investigate the effect of different null models on the discovered patterns, we show the evaluation results of CDMF and our CDSC method, which model the null hypothesis using model fitting and distribution shift-correction method, respectively. As can be seen in Fig. 12, our CDSC method can correctly and completely discover all predefined SSCDPs. By contrast, the CDMF method still reports some incorrect patterns. This might be due to the effect of model fitting errors. In addition, the predefined Matérn cluster process model assumes that the points are randomly distributed in each cluster, which is not consistent with the distribution characteristics of the observed dataset.

5.1.3. Sensitivity analysis

We also evaluated the robustness and scalability of our CDSC method by assessing the impact of distance threshold r, number of features K, and number of instances M on the precision, recall and *FPR*. To illustrate the independent effect, each factor was tested by keeping the other two the same as the default settings used in the above section, where $r = 4$, $K = 12$, and $M = 4920$.

5.1.3.1. Effect of the distance threshold (Q3). As shown in Fig. 14(a), our method performs worse in terms of precision, recall and *FPR* if a somewhat low or high distance threshold is used. This is because a lower or higher distance threshold cannot accurately capture the interactions between features that exist at a predefined scale (distance = 4).

5.1.3.2. Effect of the number of features (Q4). As we can see from Fig. 14(b), both the precision and recall of our method remain 1, and the *FPR* remains 0 on the datasets consisting of different numbers of features. This is because the correct clustering structure of features can always be effectively identified, and further validated by the significance test.

5.1.3.3. Effect of the number of instances (Q4). Similarly, as shown in Fig. 14(c), our method can always detect the complete and correct clusters of features with no false positives even with an increasing number of total instances in the dataset. This can be done because the co-distribution relations are also predefined among the instances added for each feature in a SSCDP.

5.2. Case study on a public health dataset

We validated the applicability of the proposed method using a case study aimed at identifying co-morbidities in central Colorado, USA. Co-morbidity, which is the presence of one or more additional disorders co-occurring with a primary condition, is widespread among patients and has important implications for treatment (Valderas, Starfield, Sibbald, Salisbury, & Roland, 2009). The discovery of co-morbidities can facilitate the understanding of interactions between illnesses and risk factors, which can, in turn, enhance the prevention of the occurrence of diseases and disorders.

5.2.1. Data description

The public health dataset was provided by the Colorado Department of Public Health and Environment. The original dataset contains samples for the prevalence of 14 important health conditions and risk behavior indicators in Colorado from 2013 to 2016. For each health indicator, all the samples in Colorado were divided into five groups based on the prevalence. Table 4 summarizes the 14 spatial features used in

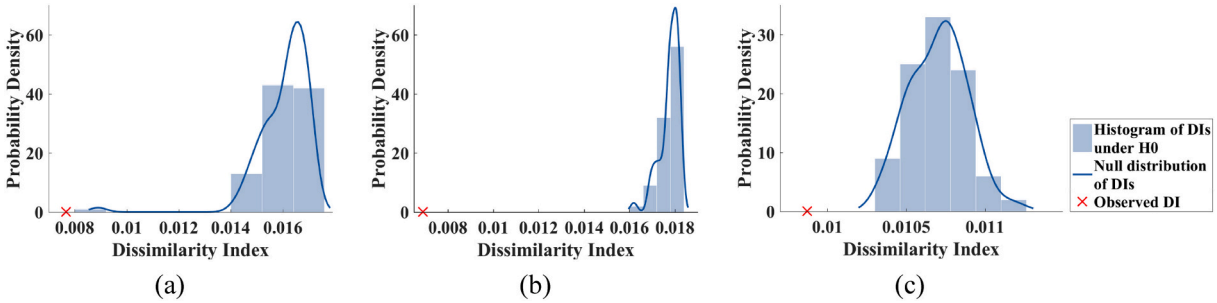

Fig. 16. Null distributions of *DI* values of SSCDPs in the public health dataset: (a) {*BingeDrink*, *HeavyDrink*}; (b) {*DelayedMC*, *NoPhysAct*, *PoorHealth*} and (c) {*Diabetes*, *MentalDistress*, *Obese*, *Overweight*, *PhysDistress*, *Smoking*}.

this study, which were defined as the health indicators with prevalence in the highest quintile of the state. The study area was located in the central area of Colorado, where the sampling density is much higher than elsewhere. Fig. 15 presents the spatial distribution of each feature.

5.2.2. Results analysis

To define the neighbor relation in the public health dataset, we set the distance threshold to 3000 m, which is an appropriate scale according to the modified L function (Yoo & Bow, 2012). Table 5 summarizes the results detected by our method, including three SSCDPs formed by two, three, and six spatial features, respectively. Fig. 16 shows the probability distributions of the DI values of these three SSCDPs under the null hypothesis. As can be seen, for each SSCDP, the DI values under the null hypothesis are generally larger than the observed value. Thus, the spatial distributions of features included in each reported pattern are significantly similar to each other.

Next, we compared our results with those of an analysis by the Colorado Department of Public Health (Williford & White, 2017). The department analyzed the correlation between smoking and other health indicators in Colorado using the Spearman correlation coefficient (Spearman, 1904) and found that smoking is most correlated with mental distress, then with no physical activity during leisure time, and finally with obesity. This conclusion is similar to our detected pattern {*Diabetes*, *MentalDistress*, *Obese*, *Overweight*, *PhysDistress*, *Smoking*}), except for the indicators *Diabetes* and *PhysDistress*. The reasons for the differences are twofold: (1) Our study only focused on the central area of Colorado with dense samplings. Additional socioeconomic (e.g., household income) and environmental (e.g., water pollution) factors need to be included to explore the regional differences in co-morbidities; (2) the spatial relationships among instances of features were not considered in the Spearman correlation coefficient, which may have led to underestimation of correlations between different features.

We also involved domain experts and scientific findings in public health to help verify and explain the results. Smoking is the leading preventable cause of death and disease. Studies have confirmed that adults with mental illness or substance use disorders are more likely to smoke cigarettes than adults without these disorders. The 2014 Surgeon General's Report has found that smoking is linked to abdominal obesity or belly fat, and is also related to increased risks of inflammation, oxidative stress and cortisol, which can, in turn, cause diabetes (USDHHS, 2014). The evidence has also shown that smokers with diabetes have higher risks of serious complications related to physical distress (e.g., peripheral neuropathy that can cause numbness, pain, weakness, and poor coordination) (CDC, 2018). Our detected spatial co-distribution pattern {*Diabetes*, *MentalDistress*, *Obese*, *Overweight*, *PhysDistress*, *Smoking*} is consistent with these findings.

The detected SSCDPs can provide useful insights into the multi-faceted health service needs of patients to treat and prevent co-morbidities; addressing one morbidity may help to address others. For example, the pattern {*DelayedMC*, *NoPhysAct*, *PoorHealth*} signifies that routine medical care and frequent physical exercise can help improve health status. The pattern {*Diabetes*, *MentalDistress*, *Obese*, *Overweight*, *PhysDistress*, *Smoking*} implies that better control of cigarette smoking can contribute to the management and prevention of its associated conditions, such as diabetes, obesity, as well as mental and physical distress.

6. Conclusion and future work

This paper formally defines a novel problem of discovering SSCDPs. This problem is different from the state-of-the-art research on spatial co-location pattern discovery, which associates features based on the prevalence of different features occurring together. In contrast, SSCDPs can provide a novel perspective for understanding spatial association by capturing the similarity of spatial distributions of different features under the spatial neighbor relation. SSCDP discovery is vital to many real-world applications, such as detecting comorbid diseases in medicine and identifying symbiotic species in ecology.

To effectively solve the SSCDP discovery problem, we first propose a distribution clustering method to extract the candidates and then develop a distribution shift-correction method to establish the statistical significance of the results. The null hypothesis underlying the test can be modeled without any a priori assumptions about the distribution model or characteristics of the features. Experiments validate the greater effectiveness of the proposed method over baseline methods, and a case study using a public health dataset shows that it can detect patterns that are of interest to domain experts that other method miss.

In future work, three issues will be considered. First, in this study, the distribution shift-correction method assumes that the observation window of a dataset is a rectangular region, so that the instances of features can be shifted on the periodic torus. Modified methods applicable for irregularly shaped observation windows need to be further studied. Second, the distribution shift-correction method maintains the observed interpoint distances exactly, with no stochastic variability. Future work will be devoted to producing local stochastic replicates of instances conditioning on the same statistical properties of each feature. Third, the rapidly growing sources of data pose novel computational and analytical challenges for SSCDP discovery. Parallel formulations are required to explore the emerging realities of big data (Prasad et al., 2017).

Author statement

The authors transfer all copyright ownership of this manuscript to Computers, Environment and Urban Systems, in the event the work is published. The authors warrant the article is original, does not infringe upon any copyright or other proprietary right of any third party, is not under consideration for publication by any other journal, and has not been published previously. The authors have reviewed the manuscript and approved this submission.

Acknowledgements

This study was funded through support from the National Natural Science Foundation of China (NSFC) [41730105, 41471385]; National Key Research and Development Foundation of China [2017YFB0503503]; 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. We would like to thank 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 article.

Appendix A. Spatial auto-correlation structure recovered from permuted datasets

Consider K-function $K(r)$, which is one of the most common second-order statistics for characterizing the spatial auto-correlation structure, as an example. Fig. A1 shows the $K(r)$ curves of both observed dataset and 99 permuted datasets of feature A in Fig. 6. One can find that the permuted $K(r)$ curves calculated based on corrected spatial neighbor relation can exactly fit the observed curve.

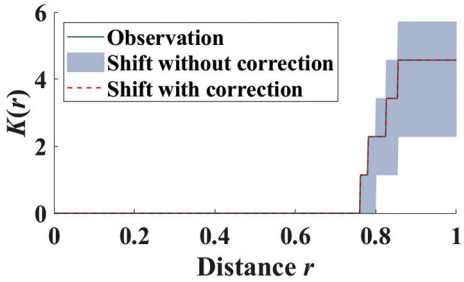

Fig. A1. *K*-function calculated for the observed dataset and 99 permuted datasets of feature *A*.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Databases* (pp. 487–499). VLDB.

Barua, S., & Sander, J. (2014). Mining statistically significant co-location and segregation patterns. *IEEE Transactions on Knowledge and Data Engineering, 26*(5), 1185–1199.

Besag, J., & Diggle, P. J. (1977). Simple Monte Carlo tests for spatial patterns. *Journal of the Royal Statistical Society: Series C: Applied Statistics, 26*(3), 327–333.

Cai, J., Deng, M., Liu, Q., Chen, Y., He, Z., & Tang, J. (2019). A statistical method for detecting spatiotemporal co-occurrence patterns. *International Journal of Geographical Information Science, 33*(5), 967–990.

Cai, J., Deng, M., Liu, Q., He, Z., Tang, J., & Yang, X. (2019). Nonparametric significance test for discovery of network-constrained spatial co-location patterns. *Geographical Analysis, 51*(1), 3–22.

Centers for Disease Control and Prevention (CDC) (2018). Smoking and diabetes. https://www.cdc.gov/tobacco/campaign/tips/diseases/diabetes.html.

Deng, M., He, Z., Liu, Q., Cai, J., & Tang, J. (2017). Multi-scale approach to mining significant spatial co-location patterns. *Transactions in GIS, 21*(5), 1023–1039.

Diuk-Wasser, M. A., Gatewood, A. G., Cortinas, M. R., Yaremych-Hamer, S., Tsao, J., Kitron, U., ... Fish, D. (2006). Spatiotemporal patterns of host-seeking Ixodes scapularis nymphs (Acari: Ixodidae) in the United States. *Journal of Medical Entomology, 43*(2), 166–176.

Dixon, P. M. (2002). Ripley's K function. *Encyclopedia of Environmetrics,* 1796–1803.

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives, 112*(9), 998–1006.

Estivill-Castro, V., & Lee, I. (2001). Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. *Proceedings of the 6th International Conference on Geocomputation,* 24–26.

Estivill-Castrol, V., & Murray, A. T. (1998). Discovering associations in spatial data—An efficient medoid based approach. *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 110–121). Berlin, Heidelberg: Springer.

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems, 25*(5), 345–366.

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

Huang, Y., & Zhang, P. (2006). On the relationships between clustering and spatial co-location pattern mining. *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence* (pp. 513–522). IEEE.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151.

Lotwick, H. W., & Silverman, B. W. (1982). Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society: Series B: Methodological,* 406–413.

Nilsson, I. M., & Smirnov, O. A. (2017). Clustering vs. relative location: Measuring spatial interaction between retail outlets. *Papers in Regional Science, 96*(4), 721–741.

Pepin, K. M., Eisen, R. J., Mead, P. S., Piesman, J., Fish, D., Hoen, A. G., ... Diuk-Wasser, M. A. (2012). Geographic variation in the relationship between human Lyme disease incidence and density of infected host-seeking Ixodes scapularis nymphs in the eastern United States. *The American Journal of Tropical Medicine and Hygiene, 86*(6), 1062–1071.

Prasad, S. K., Aghajarian, D., McDermott, M., Shah, D., Mokbel, M., Puri, S., ... Wang, F. (2017). Parallel processing over spatial-temporal datasets from geo, bio, climate and social science communities: A research roadmap. *Proceedings of the IEEE International Congress on Big Data* (pp. 232–250). IEEE.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability, 13*, 255–266.

Sengstock, C., Gertz, M., & Van Canh, T. (2012). Spatial interestingness measures for co-location pattern mining. *Proceedings of the IEEE 12th International Conference on Data Mining Workshops* (pp. 821–826). IEEE.

Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. *Proceedings of the International symposium on spatial and temporal databases* (pp. 236–256). Berlin, Heidelberg: Springer.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *Proceedings of the KDD workshop on text mining* (pp. 525–526). .

U.S. Department of Health and Human Services (USDHHS) (2014). *The health consequences of smoking—50 years of Progress: A report of the surgeon general.* Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., & Roland, M. (2009). Defining comorbidity: Implications for understanding health and health services. *The Annals of Family Medicine, 7*(4), 357–363.

Wiegand, T., & Moloney, K. A. (2013). *Handbook of spatial point-pattern analysis in ecology.* CRC Press.

Williford, D., & White, B. (2017). CDPHE Community Level Estimates: Mapping out the correlations between adult cigarette smoking and obesity, alcohol consumption, mental distress, and physical activity. https://arcg.is/1q8nHX.

Xie, Y., Bao, H., Shekhar, S., & Knight, J. (2018). A TIMBER framework for mining urban tree inventories using remote sensing datasets. *Proceedings of the IEEE International Conference on Data Mining* (pp. 1344–1349). IEEE.

Xie, Y., Eftelioglu, E., Ali, R., Tang, X., Li, Y., Doshi, R., & Shekhar, S. (2017). Transdisciplinary foundations of geospatial data science. *ISPRS International Journal of Geo-Information, 6*(12), 395.

Yoo, J. S., & Bow, M. (2012). Mining spatial colocation patterns: A different framework. *Data Mining and Knowledge Discovery, 24*(1), 159–194.

Yoo, J. S., & Shekhar, S. (2006). A joinless approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering, 18*(10), 1323–1337.