Geoinformatica (2014) 18:501–536
DOI 10.1007/s10707-013-0188-9

On detecting spatial categorical outliers

Xutong Liu · Feng Chen · Chang-Tien Lu

Received: 11 July 2012 / Revised: 27 February 2013 /
Accepted: 6 August 2013 / Published online: 28 September 2013

Abstract Spatial outlier detection is an important research problem that has received much attentions in recent years. Most existing approaches are designed for numerical attributes, but are not applicable to categorical ones (e.g., binary, ordinal, and nominal) that are popular in many applications. The main challenges are the modeling of spatial categorical dependency as well as the computational efficiency. This paper presents the first outlier detection framework for spatial categorical data. Specifically, a new metric, named as Pair Correlation Ratio (PCR), is measured for each pair of category sets based on their co-occurrence frequencies at specific spatial distance ranges. The relevances among spatial objects are then calculated using PCR values with regard to their spatial distances. The outlierness for each object is defined as the inverse of the average relevance between an object and its spatial neighbors. Those objects with the highest outlier scores are returned as spatial categorical outliers. A set of algorithms are further designed for single-attribute and multi-attribute spatial categorical datasets. Extensive experimental evaluations on both simulated and real datasets demonstrated the effectiveness and efficiency of our proposed approaches.

X. Liu (✉)
Traffic Science, ebay Inc, One Bellevue Center,
411-108th Avenue NE, Bellevue, WA 98004, USA
e-mail: xutliu@ebay.com

F. Chen
Inter-disciplinary research center (iLab), Carnegie Mellon University,
Hamburg Hall #2105B, x8-3885, 4800 Forbes Ave, Pittsburgh, PA 15213, USA
e-mail: fchen1@cmu.edu

C.-T. Lu
Department of Computer Science, Virginia Polytechnic, Institute and State University,
7054 Haycock Road, Falls Church, VA 22043, USA
e-mail: ctlu@vt.edu

Keywords Spatial Categorical data · Spatial dependency · Pair correlation · Outlier detection

1 Introduction

With the ever-increasing volume of spatial categorical data, identifying hidden but potentially interesting patterns of anomalies has attracted considerable attentions. Spatial Categorical Outlier (SCO) analysis, which aims at detecting abnormal objects in spatial context, becomes one of the most important spatial data mining branches. The identification of SCOs can help extract important knowledge in many applications, including geological data, meteorological data, satellite image analysis, and hotspot identification.

During the past decades, numerous Traditional Categorical Outlier Detection (TCOD) algorithms [11, 14, 29] have appeared in literature. TCOD approaches can be categorized into four groups: rule-based, probability distribution-based, entropy-based and similarity-based. Rule based approaches [2, 10, 21, 22, 35, 47] mine rules from the data set, and observations which are significantly uncommon are recognized as anomalies. Typical algorithms include LERAD [10], WSARE [47], and FP-Outlier [22]. Distribution-based approaches [9, 14, 34, 36] model the normal data as a specific probability density distribution. Each object that significantly deviates the normal distribution is identified as an outlier. Representative models include the Bayesian network and dependency trees. Entropy-based methods [19, 20] define TCOD as an optimization problem. That is, identifying l objects such that after removing them, the expected entropy of the rest of data set is minimized. Similarity-based approaches combine some typical TNOD approaches [8, 37] with certain well-designed dissimilarity measures to identify TCOs. Meanwhile, some research works focus on efficiently identifying categorical outliers, including AVF [29] and MapReduce AVF [30]. When encountering Spatial Categorical Outlier Detection (SCOD), TCOD approaches sometimes can't be satisfactory with the spatial context. First, spatial objects have complex structures (e.g., points, lines, polygons and locations, etc.). Second, traditional approaches do not consider spatial dependencies when identifying anomaly patterns. As the geographic rule of thumb, "Nearby things are more related than distant things [46]" requires more considerations on spatial auto correlation in spatial analysis. Third, TCOD methods treat spatial and non-spatial attributes equally, which should be considered separately for spatial anomaly identification.

Recently, a number of algorithms [1, 3, 12, 17, 45] have been proposed to identify outliers in spatial databases [39, 40]. There are three basic classes, namely, visualization-based, statistic-based, and graph-based. Visualization-based approaches utilize visualization techniques to highlight outlying objects. Representative algorithms include scatterplot [18] and Moran scatterplot [3]. Statistic-based approaches apply statistical tests to measure the local inconsistencies. Typical methods include Z [42], median-based Z [32], iterative-Z [32], and GLS-SOD [13] approaches. Graph-based methods [28, 31, 43] detect spatial outliers by designing a function to compute the difference between specific observation and its neighboring points. Other works identified outliers by studying the property of specific spatial

data. Zhao et al. proposed a wavelet-based method to detect region outliers [49]. Lu et al. presented a multi-scale approach to detecting spatial temporal outliers [33]. Adam et al. introduced an approach that considers both the spatial and semantic relationship among neighbors [1]. A local outlier measure [45] was proposed by Sun and Chawla to capture the local behaviors of data in their spatial neighborhood. However, most of the aforementioned techniques concentrated on continuous real-valued data attributes. There is no mechanism for processing spatial categorical data with no implicit ordering.

In real world, the non-spatial attributes of spatial data are usually category-typed, where attributes have no intrinsic order. A typical example is Rock whose values include Igneous, Sedimentary, and Metamorphic. This special property makes anomaly detection in spatial categorical domain more complicated than that in numerical one. Currently, there is a lack of Spatial Categorical Outlier Detection (SCOD) approaches. When encountering categorical dataset, some introduce Spatial Numerical Outlier Detection (SNOD) methods by directly mapping the categorical attributes to continuous ones. However, there are several critical issues: (1) **Mis-utilization:** statistically, the definition of an SCO is different with that of a Spatial Numerical Outlier (SNO). Although both of them focus on the identification of spatial abnormal behaviors, SCOD is determined by the co-occurrence infrequency, while SNOD focuses on the numerical differences; (2) **Complicated function:** the mapping process is not straightforward, especially for nominal attributes; (3) **Swamping and masking problems:** without estimating outlying degrees accurately, some true outliers may be missed and normal ones misclassified as outliers.

In the past decades, there are some association rule based researches [24, 25, 41, 48] which focus on mining spatial co-location patterns. *A spatial co-location pattern is a set of spatial events that are frequently located together in spatial proximity* [41]. Most of the works were interested in identifying a collection of boolean features (e.g., bird, drought) which have higher co-occurrence frequency. Spatial association rules was first discussed in [27]. Huang et al. [25] proposed a general framework of mining spatial co-location patterns where an instance join-based algorithm was introduced. Furthermore, Yoo and Shekhar [48] designed a partial-join and joinless method to improve its performances. Meanwhile, Huang et al. [24] introduced a novel methodology to mine the rare set of spatial events. The co-location pattern identification process aims at mining association rules among different types of features in close geographic proximity. Its objective is to identify the group of events from different types with higher co-occurrence frequency. However, spatial categorical outlier detection focuses on identifying the spatial objects which behave abnormally with respect to its spatial neighbors in the same types of features. The co-occurrence frequency within the same types of attributes is computed at different spatial distances. The distinct objective and application make the methodologies of frequency computation quite different in these two areas.

Pair Correlation Function (PCF) has been proven very effective [26] to capture how observations are packed together, which could be utilized to estimate the relevance among spatial categorical objects. It is a probability measure to find a unit at a distance of d away from a reference unit. PCF techniques have been widely used to analyze the behavioral characteristics of the individual objects in a variety of natural systems, e.g., electrostatic, magnetic and biological application. This paper

investigates the benefits of PCF techniques on SCOD, and design algorithms for spatial datasets with single and multiple categorical attributes. First, PCF techniques are utilized to measure the Pair Correlation Ratio (PCR) between any pair of category sets as a function of spatial distances. Second, the discrete relevance between a reference object and its neighbors is fitted by the PCR function. Third, the outlying degree for each object is computed as the inverse of average PCR between the object and its neighbors. Finally, the top l objects with high outlierness values are identified as SCOs. The key contributions of this paper include:

- **Formalization of the SCOD problem**. This is the first work that specifically focuses on Spatial Categorical Outlier Detection(SCOD). The SCOD problem is differentiated from the SNOD one: an SCO is identified as a spatial observation which occurs infrequently with regard to its spatial neighbors.
- **Design of two SCOD algorithms for single attribute dataset**. We first present a PCF-SCOD (Pair Correlation Function based Spatial Categorical Outlier Detection) algorithm to identify SCOs by investigating the capability of PCF techniques of calculating the Pair Correlation Ratios (PCRs) for each pair of categories at specific distances. Further, considering the computational cost of PCF-SCOD, a kNN-SCOD(k Nearest Neighbor based Spatial Categorical Outlier Detection) approach is proposed to approximate the outlier scores. It allows for efficient SCOD when memory and processor resources are issues.
- **Design of one SCOD algorithm for multi-attribute dataset**. The kNN-SCOD work is extended to the SCOD issue in multi-attribute domain. By mapping the kNN relationship from the raw dataset into a well-defined pair object dataset, the PCR of possible pair category sets is computed to capture the relevance among objects which are spatial neighbors with each other.
- **Comprehensive experiments to validate the effectiveness and efficiencies of the proposed techniques**. The proposed approaches were evaluated by extensive experiments on simulated and real datasets. The results demonstrated that PCF series of algorithms outperformed 14 existing techniques for both single and multiple attribute dataset.

The paper is organized as follows. Section 2 provides fundamental definitions used in SCOD, and introduces a general SCOD framework. Section 3 presents two SCOD approaches to identifying SCOs with single attribute, named PCF-SCOD and kNN-SCOD. kNN-SCOD work is extended to detect the SCOs with multiple attributes in Section 4. Experimental evaluations on both simulated and real life datasets are presented in Section 5. The paper concludes with a summary of the research in Section 6.

2 Preliminary concept

This section introduces PCF techniques, defines key notations used, and examines the SCOD problem. The deficiencies of existing methods are also discussed.

2.1 Pair correlation function

In mathematical mechanics, PCF, $g(r)$, is defined as the observed probability of finding an object at a given distance, r, from a fixed reference particle [38]. The mathematical definition of $g(r)$ is

$$g(r) = \frac{dn(r)/N}{dv(r)/V} = \frac{dn(r)}{dv(r)} \cdot \frac{V}{N} = \frac{dn(r)}{4\pi r^2 dr} \cdot \frac{V}{N} \tag{1}$$

Where N and V denote the number of units and the volume of the entire system, respectively; $dn(r)$ and $dv(r)$ represent those in the shell-region; r is the distance from reference unit to the shell of interest. Figure 1 depicts the 2D-projection of a typical example which describes the PCF computation in Eq. 1. In this paper, the relevances among spatial objects are determined by the frequency of co-occurrence of a pair of categories at specific distances. PCF is capable of estimating how observations are packed together, which could be utilized to capture the relationship among spatial categorical objects.

2.2 Preliminary definition

To formalize the SCOD framework, we need to understand some basic definitions.

Definition 1 (Spatial categorical dataset) Let s denote a spatial location on a domain S of the d dimensional Euclidean space R^d. Let $A_1, \cdots, A_m$ be a set of categorical attributes and $\mathcal{C}_1, \cdots, \mathcal{C}_m$ non-empty sets over these attributes where $\mathcal{C}_i \bigcap \mathcal{C}_j = \phi$ for $i \neq j$.

A set $\mathcal{D} \subseteq S \times \mathcal{C}_1 \times \cdots \times \mathcal{C}_m$ is called a spatial categorical dataset over the domains, $S, \mathcal{C}_1, \cdots, \mathcal{C}_m$. Each record $r_i \in \mathcal{D}$ $(i \in 1, \cdots, n)$ can be denoted as a vector $(r.s, r.A_1, \cdots, r.A_m)'$, where $r.A_i \in \mathcal{C}_i$. The number of categorical attributes, m, is also referred as the dimensionality of the spatial data set.

Categorical attributes can be classified into two types: ordinal and nominal ones. The key characteristic of nominal attributes is that different values in an attribute domain are absolutely not inherently ordered, e.g., color. The issue of distance or dissimilarity for nominal data is not as straightforward as for ordinal or numerical one. Thus it is difficult to directly compare two nominal values. This paper is focused on such type of data sets as consist solely of nominal attributes. However, our approach could be directly applied to spatial categorical data set with ordinal attributes.

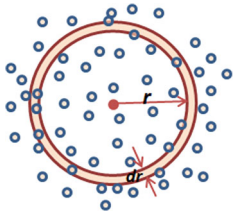


Fig. 1 PCF using a spherical shell of thickness dr

Informally, an anomalous behavior in spatial domain can be truly captured by the local difference, which is determined by the irrelevance between a specific object and its spatial neighbors. A spatial neighbor query refers to identifying the k spatial objects nearest to specific points. An classical set of queries is the class of k Nearest Neighbor (NN) queries. Typical methods include simple kNN queries [6], approximate kNN queries [15], reverse NN queries [44], and kNN join queries [23]. Meanwhile, Voronoi diagram method can be utilized to identify spatial neighborhood by partitioning the plane into N (the number of spatial objects) polygons. Hence, the nearest neighbors of any query point inside a Voronoi polygon are the generators of those polygons. Normally, an observation can not have more than six Voronoi neighbors on average due to the search space [4].

In the paper, simple k-Nearest Neighbor (kNN) is utilized to construct the neighborhood relationship.

Definition 2 (Spatial neighborhood) Given a dataset $\mathcal{D}$ with n points and parameter k, for $r_i \in \mathcal{D}$, its spatial neighborhood is constructed by the top k points according to its spatial Euclidean Distance vector with the rest of observations in the dataset, such that $\forall j \in 1, ..., n, j \neq i, r_j \in kNN(r_i) : d^E(r_i, r_j) \leq d_k^E(r_i)$, where $d_k^E(r_i)$ represents the distance between r_i and its k^{th} spatial neighbor.

One of the open issues of kNN algorithm is the selection of the optimal value of k beforehand. In the area of spatial outlier detection, if k is too large, the abnormal behavior of an outier might be masked by the average differences. On the other hand, if k is too small, the normal observation might be errorly identified as an outlier since it is more sensitive to the existence of outlying neighbors. The objective of this paper is to measure the similarities among categorical observations, and further identify the spatial categorical outliers. In the experiment, we evaluated different k values on spatial outlier detection and found it was enough to set k as around 8 in different sizes of data sets.

In numerical domain, an SNO is defined as the one whose non-spatial attributes are significantly different with those of its neighbors. Such definition is not applicable in categorical domain. For example, as shown in Fig. 2, based on the idea of SNOD approaches, the object A will be recognized as an outlier since it has the categorical attribute, T, which is very different with its neighbors', Fs. However, the contrary is the case in categorical domain. This is because the pair of attributes, $< T, F >$ or $< F, T >$ occurs normally at the spatial distance, d. Object A should be treated as a normal observation. In this sense, the definition of spatial outliers in categorical domain is totally different with that of SNOs.

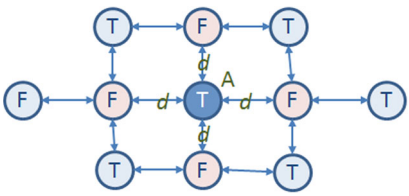


Fig. 2 An example of differentiating an SNO and an SCO

Definition 3 (SCO) Let r_i be an observation in $\mathcal{D}$, and $r_{i\_1}, \cdots, r_{i\_k}$ be its spatial neighbors. Its outlierness, for $k \geq 1$,is defined as

$$OutScore(r_i) = -\frac{\sum_{j=1}^{k} PCR(r_i.A, r_{i\_j}.A, d^E(r_i, r_{i\_j}))}{k} \quad (2)$$

The first l observations with higher $OutScore$ are considered as spatial categorical outliers. $PCR(r_i.A, r_{i\_j}.A, d^E(r_i, r_{i\_j})))$ denotes the co-occurrence frequency of the pair category sets, $< r_i.A, r_{i\_j}.A >$, of objects, r_i and r_{i_j}, at the specified distance, $d^E(r_i, r_{i\_j})$. Here, l is decided by the cut-off θ so that these l observations satisfy $OutScore >= \theta$, while the rest of observations $OutScore < \theta$. Assume that the $OutScore$s follow a normal distribution, the cut-off θ is calculated by the mean and standard deviation. This paper declares θ is computed as the $OutScore$ whose p-value is equal to 0.01.

In summary, an SCO is an observation which has lower co-occurrence frequency with its spatial neighbors. PCF is capable of estimating such frequencies as how objects are packed together, which could be utilized to calculate PCR, further $OutScore(r_i)$. Sections 3 and 4 will discuss the PCR computation in spatial categorical dataset with single and multiple attributes, respectively.

Based on the above definitions, the SCOD problem can be modeled as follows.

Given:

- $\mathcal{D}$ is a set of spatial objects r_1, ..., r_n with single or multiple categorical attributes.
- $\boldsymbol{k}$ is an integer denoting the number of adjacent data objects which form the spatial neighborhood.
- $\boldsymbol{l}$ is the number of outliers to be identified, generally, $l \ll n$.

Objective:

- Design a mapping function $f : \mathcal{D} \times \mathcal{D} \longrightarrow R^+$, which estimates PCR for each pair of objects as a function of spatial distances.
- Estimate the $OutScore$ for each observation, and identify a set of O_1, ..., $O_l \in \mathcal{D}$ with higher values as SCOs.

3 Spatial categorical outlier detection in single attribute dataset

Intuitively, given a spatial data set, a normal observation is the one that behaves normally with regard to its spatial neighborhood. In single categorical domain, this corresponds to the high frequency of co-occurrence of a pair of categories at a specified distance. The categorical outlier has rarely occurring category attributes with regard to the ones of its neighborhood. This section presents two SCOD approaches to detecting SCOs with single attribute, namely PCF-SCOD (Pair Correlation Function based Spatial Categorical Outlier Detection) and kNN-SCOD-S (k Nearest Neighbor based Spatial Categorical Outlier Detection in Single attribute dataset).

3.1 Pair correlation function based SCOD

We investigate the benefits of PCF techniques to capture the rare behaviors of SCOs. The major components are described as follows.

- **PCR (Pair Correlation Ratio) estimation**. PCR is defined to characterize the co-occurrence frequency of each pair of categories at different specified distances. With the set of discrete points in a 2-D space, determined by PCR values against spatial distances, we can statistically learn a continuous PCR function which can easily estimate the PCRs among spatial objects.
- **Neighborhood formulation and outlierness computation**. The spatial neighborhood for each object can be formed using kNN. And the outlier degree for each object is computed by the mean of PCRs between itself and its spatial neighbors.
- **Outlier identification**. The outer scores are ranked in an descending order and the top l objects are marked as SCOs.

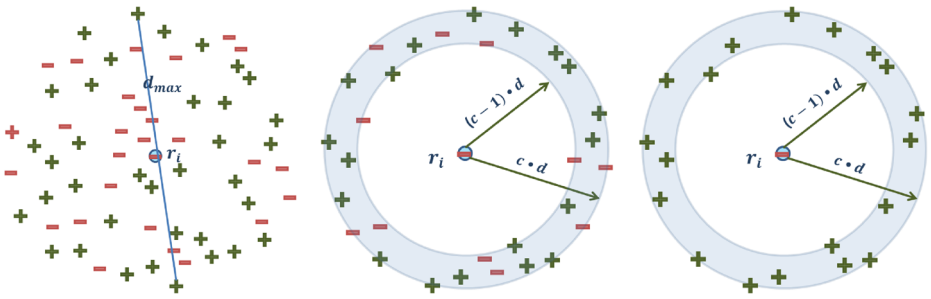
For the above components, the first one is critical since it determines the estimation quality of the relevance among observations. Section 3.1.1 introduces PCR computation in particular, and then PCF-SCOD algorithm is described in Section 3.1.2.

3.1.1 Pair correlation ratio computation

For each random variable r, $r.A$ is a multilevel categorical variable taking values in $\mathcal{C} = A^1, ..., A^L$. We denote Eq. 3 as the frequency of observing category A^l in the dataset,

$$Freq(A^l) = P\left[A(r_i.A) = A^l\right] = \frac{n^{A^l}}{n} \tag{3}$$

where n^{A^l} represents the number of objects whose non-spatial attribute are A^ls, and n the number of objects in the whole dataset. Figure 3a depicts a small spatial categorical dataset which consists of 64 objects, of which 37 ones take category "+", and 27 ones "−". With Eq. 3, we get $Freq(+) = 37/64$ and $Freq(-) = 27/64$.



(a) An example dataset (b) $\mathcal{D}^c$ identification w.r.t. r_i (c) $\mathcal{D}^c_{-+}$ identification w.r.t. r_i

Fig. 3 An example of identifying B-PD and B-PC-PD

Let $SPF(< A^l, A^{l'} >, d^E(r_i.X, r_j.X))$ denote Spatial Pair Frequency associated to two objects, r_i and r_j, where $r_i.A = A^l$ and $r_j.A = A^{l'}$. Then PCR can be defined as follows:

Definition 4 (Pair Correlation Ratio-PCR) Considering a spatial pair correlation process in which there are two observations, r_i, r_j in $\mathcal{D}$, each of them is tagged with one category, A^l and $A^{l'}$, respectively. The PCR of r_i,r_j is defined as the normalized spatial pair frequency of the pair of categories, $< A^l, A^{l'} >$ at r_i and r_j.

The mathematical definition of PCR is

$$PCR(r_i, r_j) = \frac{SPF\left(< A^l, A^{l'} >, d^E(r_i.S, r_j.S)\right)}{Freq(A^l) \cdot Freq(A^{l'})} \tag{4}$$

As shown in Eq. 4, the PCR value between two spatial objects, is determined by the co-occurrence frequency of categories and their spatial Euclidean distance, not their specific spatial locations. In the following, SPF computation is discussed by utilizing the example in Fig. 3.

– **Distance division**. Compute the Euclidean distance for each pair of spatial objects, identify the maximal and minimal ones, $d^E{}_{\text{Max}}$ (as shown in Fig. 3a) and $d^E{}_{\text{Min}}$ (set as 0), and divide the distance into b small bins whose sizes are computed as:

$$d = \frac{d^E{}_{\text{Max}}}{b} \tag{5}$$

As we know, it is common for spatial objects to be autocorrelated at shorter distances. It is not necessary to take the pair correlation at longer distances into considerations. Simply, $d^E{}_{\text{Max}}$ can be approximated by

$$d^E{}_{\text{Max}} = \frac{1}{2}\max\Big\{\left|\max(Proj_x(r_i.S)) - \min(Proj_x(r_i.S))\right|, \\ \left|\max(Proj_y(r_i.S)) - \min(Proj_y(r_i.S))\right|\Big\}, i, j = 1, ..., n \tag{6}$$

where $Proj_x(\cdot)$ and $Proj_y(\cdot)$ represent the projection operations of S location on X, Y coordinates, respectively. It is reasonable for such approximation since SCOD focuses on the local relevance estimation.

– **Identification of Bin based Pair Dataset (B-PD)**. Based on the spatial distance, map each pair of objects into their corresponding distance bin.

$$\mathcal{D}^c = \left\{< r_i, r_j >, (c-1) \cdot d \le d^E(r_i.S, r_j.S) < c \cdot d, c \in [1, b]\right\} \tag{7}$$

For example, for the reference object r_i shown in Fig. 3, based on its spatial distances from others, we can identify 30 objects, $\{r_j\}_{j=1}^{30}$, which make $< r_i, r_j > \in \mathcal{D}^c$ since they satisfy the condition: $(c-1) \cdot d \le d^E(r_i.S, r_j.S) < c \cdot d$. As depicted in Fig. 3b, the 30 objects are contained in the shaded circular ring.

– **Identification of Bin and Pair Category based Pair Dataset (B-PC-PD)**. By scanning the identified B-PD, we map each pair of objects into its pair category based

subset so that their categorical attributes are the specified pair of categories. That is, we can construct $\mathcal{D}^c_{A^l A^{l'}}$ as follows:

$$\mathcal{D}^c_{A^l A^{l'}} = \Big\{\langle r_i, r_j\rangle, \Big[(r_i.A == A^l)\&\&(r_j.A == A^{l'})\Big] \Big\| \Big[(r_i.A == A^{l'})\&\&(r_j.A == A^l)\Big], < r_i, r_j > \in \mathcal{D}^c, c \in [1, b]\Big\} \quad (8)$$

In Fig. 3b, in the identified objects $\{r_j\}_{j=1}^{30}$ with regard to reference object r_i in $\mathcal{D}^c$, we can map 11 pairs of $\{< r_i, r_j >\}_{j=1}^{11}$ into $\mathcal{D}^c_{--}$, and the other 19 pairs into $\mathcal{D}^c_{-+}$ based on their corresponding categorical attributes. Figure 3c depicts the pair objects in $\mathcal{D}^c_{-,+}$ with regard to the reference object r_i.

– **Spatial Pair Frequency computation**. The SPF of the pair of categories in the c^{th} bin can be computed by

$$SPF(\langle A^l, A^{l'}\rangle, [(c-1)\cdot d, c\cdot d]) = \frac{|\mathcal{D}^c_{A^l A^{l'}}|}{|\mathcal{D}^c|} \quad (9)$$

where $|\mathcal{D}^c_{A^l A^{l'}}|$ and $|\mathcal{D}^c|$ represent the number of pair objects in $\mathcal{D}^c_{A^l A^{l'}}$ and $\mathcal{D}^c$, respectively. Overall, for each pair of $\langle A^l, A^{l'}\rangle$, we can estimate b pair frequency values corresponding with b bins. Based on the b discrete points in a 2-D space, we can statistically learn a pair frequency function $SPF(\langle A^l, A^{l'}\rangle, d^E)$ by polynomial and curve fitting, subjecting to the following constraints:

1. $SPF(\langle A^l, A^{l'}\rangle, d^E) = SPF(\langle A^{l'}, A^l\rangle, d^E)$

   Proof By definition, it is easy to prove this constraint. □

2. $SPF(\langle A^l, A^{l'}\rangle, 0) = \begin{cases} Freq(A^l) & A^l = A^{l'} \\ 0 & A^l \neq A^{l'} \end{cases}$

   Proof If $A^l = A^{l'}$, $SPF(\langle A^l, A^{l'}\rangle, 0) = \frac{|\mathcal{D}^0_{A^l A^{l'}}|}{|\mathcal{D}^0|} = \frac{n^{A^l}}{n} = Freq(A^l)$, and if $A^l \neq A^{l'}$, $PF(\langle A^l, A^{l'}\rangle, 0) = \frac{|\mathcal{D}^0_{A^l A^{l'}}|}{|\mathcal{D}^0|} = \frac{0}{n} = 0$. □

3. $\sum_{l'=1}^{L} SPF(\langle A^l, A^{l'}\rangle, d^E) = Freq(A^l)$.

   Proof We can identify $D^c_{A^l}$ as $D^c_{A^l} = \{\langle r_i, r_j\rangle, (r_i.A == A^l)\&((c-1)\cdot d \leq d^E(r_i.S, r_j.S) < c\cdot d), < r_i, r_j > \in \mathcal{D}, c \in [1, b]\}$. □

   There is a deduction as follows: $\sum_{l'=1}^{L} SPF(\langle A^l, A^{l'}\rangle, d^E) = \frac{\sum_{l'=1}^{L} |\mathcal{D}^c_{A^l A^{l'}}|}{|\mathcal{D}^c|} = \frac{n^{A^l}}{n} = Freq(A^l)$

SPF takes input as the spatial distance and output the pair frequency. Because spatial distance is continuous, we can only sample a limited number of spatial distances and calculate the corresponding spatial pair frequencies based on the data set. What we need is a parametric form for this function. Therefore, we need to further fit the sampled values to a curve of parametric form. Based on our observation, we can conduct regular nonlinear regression process to fit a polynomial curve of order two

Table 1 Main parameters used in this paper

Parameters	Description
S	A dataset storing the spatial attributes
A	A dataset storing the non-spatial categorical attributes
b	The number of bins to divide the distance values
m	The number of categorical attributes
n	The number of spatial objects in the dataset
k	The number of spatial neighbors
l	The number of SCOs

to minimize the mean squared error (MSE). We implemented the fitting process using the standard Matlab function "polyfit".

3.1.2 PCF-SCOD algorithm

The proposed PCF-SCOD algorithm for single attribute domain has 6 input parameters, S, A, b, n, k and l, described in Table 1. Algorithm 1 describes this approach as the following 4 steps.

Step 1 (*lines: 1–3*) **Formalization of spatial neighborhood**. First, we construct distance matrix, $DistMat$, in which the i^{th} row records the spatial distances between r_i and the rest of objects in the dataset. With it, the spatial neighborhood matrix, $Neighbor$, can be identified for each spatial object.

Step 2 (*lines: 4–17*) **Computation of SPFs among spatial objects**.

a. (*lines: 4–5*) `Distance division`. With the stored values in $DistMat$, identify its maximum and minimum values (0). Then, the size of unit bin, d, can be computed using Eq. 5.

b. (*line: 6*) `Computation of category frequency`. We construct the frequency array, $Freq_A$, which records the occurrence frequencies for all the observing category, and the pair category array, PC_Arr, which stores all the possible pairs of categories (N_p represents the number of possible pairs of categories) in the dataset.

c. (*lines: 7–14*) `SPF computation`. This step includes three important procedures: bin based pair set identification, bin and pair category based pair set identification and the discrete SPF computations. At step 8, we use the function, B_PD_Iden, to extract all the pair objects for $\mathcal{D}^c$, which satisfy certain distance conditions, $Cond^{d,c}$ (in Eq. 7). At step 11, function $B\_PC\_PD\_Iden$, is used to construct $\mathcal{D}^c_{A^l A^{l'}}$ by scanning the pair objects in $\mathcal{D}^c$, which satisfy category attribute condition, $Cond^{A^l A^{l'}}$ (in Eq. 8). With the above two pair sets, the b discrete SPF values for each pair of categories can be computed at step 12.

d. (*lines: 15–17*) `Learn of continuous SPF function`. With the above discrete SPF values against b different distance range, we statistically learn a continuous SPF function for each pair of categories using the polynomial and curve fitting.

Step 3 (*lines: 18–24*) **Construction of PCR matrix**. Utilizing Eq. 4, with SPF function, the relevance scores (PCR) can be simply calculated for each pair of spatial objects. In addition, PCR matrix, $PCRMat$, is constructed as

Algorithm 1 PCF-SCOD-S Approach

1: **for** $i = 1$ to n **do** {Calculate the neighborhood and distance matrix}
2: $[Neighbor(i,:), DistMat(i,:)] = kNN(S, r_i.S, k)$
3: **end for**
4: $d^E_{Max} = max(DistMat)$;{Identify the maximum spatial distance}
5: $d = \frac{d^E_{Max}}{b}$;{Calculate the size of unit bin}
{Computation of category frequency, pair category array and its sizes.}
6: $[Freq_A, PC\_Arr, N_p] = CateFreq(A)$;
7: **for** $c = 1$ to b **do** {Identify B-PD}
8: $\mathcal{D}^c = B\_PD\_Iden(DistMa, Cond^{d,c})$;
9: **for** $p = 1$ to N_p **do** {Identify B-PC-PD}
10: $A^l A^{l'} = PC\_Arr(p)$;
11: $\mathcal{D}^c_{A^l A^{l'}} = B\_PC\_PD\_Iden(\mathcal{D}^c, Cond^{A^l A^{l'}})$;
12: $SPF(A^l, A^{l'}, c \cdot d) = \frac{|\mathcal{D}^c_{A^l A^{l'}}|}{|D^c|}$ {Calculate its corresponding spatial pair frequency.}
13: **end for**
14: **end for**
15: **for** $c = 1$ to N_p **do** {Model continuous PCR function}
16: $SPF(A^l, A^{l'}, d^E) = FitModel(\{SPF(A^l, A^{l'}, [(c-1) \cdot d, c \cdot d])\}_{c=1}^{b})$;
17: **end for**
18: **for** $i = 1$ to n **do** {Calculate PCR matrix between spatial object and its neighbors}
19: **for** $j = 1$ to k **do**
20: $f = Neighbor(i, j)$;
21: $PCRMat(i, j) = \frac{SPF(r_i.A, r_f.A, DistMat(i,f))}{|Freq_{r_i.A}| \cdot |Freq_{r_f.A}|}$;
22: **end for**
23: **end for**
24: $RelevanceArr = mean(PCRMat)$; {Compute relevances for spatial objects}
25: $RankList = Rank(RelevanceMat, ascend)$; {Rank objects with ascending relevance values}
26: $O_l = Outlier(RankList, 1 : l)$ {Mark the outliers}

the mean of the PCR values between the reference observation and its neighbors.

Step 4 (*lines: 25–26*) **Outlier identification**. Finally, the objects are sorted with ascending PCR values, and the l objects with lower relevance scores are recognized as outliers.

Computational complexity To form the distance and neighborhood matrices will take $O(n^2)$. It takes $O(n)$ to construct the category frequency array and pair category array. Identifying $\mathcal{D}^c$ and $\mathcal{D}^c_{A^l A^{l'}}$ takes around $O(b \cdot N_p \cdot |\mathcal{D}^c| \cdot n^2)$. Finally, computing the PCR matrix costs $O(k \cdot n)$. In summary, assuming $n \gg k$, $n \gg b$, $n \gg N_p$ and $n \gg |\mathcal{D}^c|$. The total computational complexity of PCF-SCOD approach is $O(n^2)$=($O(n^2)$+$O(n)$+$O(b \cdot N_p \cdot |\mathcal{D}^c| \cdot n^2)$)+$O(k \cdot n)$).

3.2 k nearest neighbor based SCOD in single attribute dataset

For the PCF-SCOD method, in a larger-size dataset, it is a time-consuming process to estimate PCR values for each pair of categories as a function of distances. To solve this issue, we propose a kNN based PCR approximation which can help identify SCOs more efficiently.

3.2.1 kNN based PCR computation

*k*NN-based estimation is an approximate computation of PCR value which only utilizes the *k*NN relationship to capture the relevances among objects. First, a *k*NN mapping function is defined to map the *k*NN information of the raw dataset into a pair dataset.

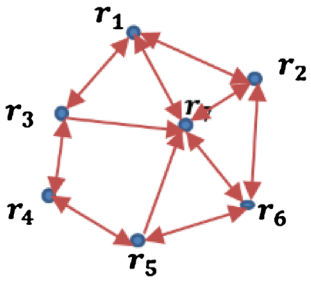
Definition 5 (*k*NN Mapping Function) Given a dataset $\mathcal{D}$ with n observations, let $g : \mathcal{D} \times \mathcal{D} \longrightarrow F$ be a mapping function which maps any pair of objects with $(m+1)$ dimensions in $\mathcal{D}$ into one point with $(2m+1)$ dimensions in F. This mapping function captures the spatial relationships among the objects into the domain F over space $\mathcal{F}^{(2m+1)}$. Here, each observation contains $2m$ categorical attributes and one continuous attribute that stores the distance between the pair of objects in $\mathcal{D}$. We can simply map the pair objects which are spatial neighbors with each other in $\mathcal{D}$ into the points in F^k, since SCOD only focuses on the local co-occurrence frequency.

By definition, only *k*NN relationships are captured to approximate PCR. And we know F^k is a dataset with at most $k \cdot n$ observations which stores k nearest neighbor relationships in $\mathcal{D}$. Therefore, as derived from a single attribute dataset $\mathcal{D}$, F^k is a 2-dimension dataset.

Considering a sample spatial categorical dataset with *7* observations, we set k as *3*, and the *3*NN neighborhood is depicted in Fig. 4. For each object, it takes one of categorical values in $\mathcal{C} = \{T, F\}$. Utilizing the mapping function, g, the *3*NN relationships are mapped into the observations of F^3, shown in Table 2. In F^3, there are three types of data: $< T, T >$, $< T, F >$ $(=< F, T >)$ and $< F, F >$. With dataset F^k, we approximate PCR values as follows:

– **Identification of pair category based subset,** $F^k_{A^l A^{l'}}$. Scan F^k and identify all pair objects whose attributes are specified pair categories. That is, the subset $F^k_{A^l A^{l'}}$ is identified as the one in which each object, f_i, has pair attributes $< A^l, A^{l'} >$ or $< A^{l'}, A^l >$.

$$F^k_{A^l A^{l'}} = \{ f_i, [(f_{i1} == A^l)\&\&(f_{i2} == A^{l'})] \| [(f_{i1} == A^{l'})\&\&(f_{i2} == A^l)],\ f_i \in F^k \} \quad (10)$$



ID	Attr.1	Attr.2	3NN		
r_1	T	$\{T, P\}$	r_2	r_3	r_7
r_2	F	$\{F, Q\}$	r_1	r_6	r_7
r_3	F	$\{F, P\}$	r_1	r_4	r_7
r_4	F	$\{F, Q\}$	r_3	r_5	r_7
r_5	T	$\{T, P\}$	r_4	r_6	r_7
r_6	F	$\{F, Q\}$	r_2	r_5	r_7
r_7	F	$\{F, P\}$	r_1	r_2	r_6

Fig. 4 A sample of spatial categorical dataset. (Attr.1 means the observed attributes in single attribute domain, which is used in Section 3.2; Attr.2 means the observed attributes in multiple attribute domain, which is used in Section 4)

Table 2 Observations in F^3

ID	Pair observation in $\mathcal{D}$	Observation in F^3
f_1	$< r_1, r_3 >$	$< F, T >$
f_2	$< r_1, r_2 >$	$< F, T >$
f_3	$< r_1, r_7 >$	$< F, T >$
f_4	$< r_2, r_6 >$	$< F, F >$
f_5	$< r_2, r_7 >$	$< F, F >$
f_6	$< r_3, r_4 >$	$< F, F >$
f_7	$< r_3, r_7 >$	$< F, F >$
f_8	$< r_4, r_5 >$	$< F, T >$
f_9	$< r_4, r_7 >$	$< F, F >$
f_{10}	$< r_5, r_6 >$	$< F, T >$
f_{11}	$< r_5, r_7 >$	$< F, T >$
f_{12}	$< r_6, r_7 >$	$< F, F >$

For example, we can identify $F^3_{FT} = \{f_1, f_2, f_3, f_8, f_{10}, f_{11}\}$ based on the observations shown in Table 2.

– **Pair frequency computation**. Calculate the frequency of each category pair using Eq. 11.

$$Freq(A^l, A^{l'}) = \frac{|F^k_{A^l A^{l'}}|}{|F^k|} \tag{11}$$

Where $|\cdot|$ means the number of the observations in the corresponding dataset.

– **PCR computation**. kNN-based PCR approximation can be computed by Eq. 12

$$PCR^k(A^l, A^{l'}) = \frac{Freq(A^l, A^{l'})}{Freq(A^l) \cdot Freq(A^{l'})} = \frac{|F^k_{A^l A^{l'}}|/|F^k|}{Freq(A^l) \cdot Freq(A^{l'})} \tag{12}$$

Utilizing the sample in Fig. 4, we calculate the PCR value between objects, r_1 and r_7, which take category T and F, respectively. First, the frequencies of T and F in $\mathcal{D}$ are computed, $Freq(F) = 0.71(5/7)$ and $Freq(T) = 0.29(2/7)$. Next, after the mapping process, we scan F^3 and identify the subsets, F^3_{FF}, F^3_{FT}, F^3_{TT} using Eq. 10. In the following, the pair frequency values of $< F, F >$, $< F, T >$ and $< T, T >$ are computed as $0.5 = 6/12$, $0.5 = 6/12$, and $0 = 0/12$. Further, PCR value between r_1 and r_7 is equal to $2.4283(= 0.5/(0.71 * 0.29))$, as shown in Table 3.

In Fig. 4, there is a typical example to describe the different ways which estimate the relevance scores among objects by utilizing SNOD and SCOD approaches. For SNOD, $< F, F >$ and $< T, T >$ must have higher correlation than that of $< F, T >$. However, for SCOD, the case is not always true. In Table 4, PCR($< T, T >$) is equal to 0 since there is no such case as pair objects whose pair categories are both Ts, which means $< T, T >$ never co-occur locally together. Although $< F, F >$ is from the same category, F, and $< F, T >$ is from different ones, F and T, PCR($< F, F >$)

Table 3 PCR computation

Pair categories	Freq.	Prob.	PCR
$< F, F >$	6	0.50	0.9919
$< F, T >$	6	0.50	2.4283
$< T, T >$	0	0	0
Comment: $Freq(F) = 0.71$; $Freq(T) = 0.29$			

Table 4 Observations for PAS $\{< A_1, A_2 >, < A_1, A_2 >\}$ in F^3

Pair objects	$\{< A_1, A_2 >, < A_1, A_2 >\}$
$< r_1, r_3 >$	$< \{T, P\}, \{F, P\} >$
$< r_1, r_2 >$	$< \{T, P\}, \{F, Q\} >$
$< r_1, r_7 >$	$< \{T, P\}, \{F, P\} >$
$< r_2, r_6 >$	$< \{F, Q\}, \{F, Q\} >$
$< r_2, r_7 >$	$< \{F, Q\}, \{F, P\} >$
$< r_3, r_4 >$	$< \{F, P\}, \{F, Q\} >$
$< r_3, r_7 >$	$< \{F, P\}, \{F, P\} >$
$< r_4, r_5 >$	$< \{F, Q\}, \{T, P\} >$
$< r_4, r_7 >$	$< \{F, Q\}, \{F, P\} >$
$< r_5, r_6 >$	$< \{T, P\}, \{F, Q\} >$
$< r_5, r_7 >$	$< \{T, P\}, \{F, P\} >$
$< r_6, r_7 >$	$< \{F, Q\}, \{F, P\} >$

is smaller than PCR($< F, T >$). That means, for pair categories, F and F, they have the same co-occurrence frequency(0.5) with that of pair categories, F and T. However, category T is more infrequent in the whole dataset, and each object whose categorical attribute is T all co-occurs with objects with F in $\mathcal{D}$. To capture such spatial correlation characteristic, we normalize the PCR by the frequencies of observed categories of the pair objects. Equation 12 assigns a higher PCR to the pair objects of which the categories are more infrequent than that of which if those are more frequent.

3.2.2 Algorithm of kNN-SCOD-S

We generalize the above procedures as *k*NN-SCOD-S algorithm. The proposed approach has *5* input parameters, S, A, n, k and l (see Table 1). Algorithm 2 describes *k*NN-SCOD-S as the following *4* steps.

Step 1 (*lines: 1–4*) **Construction of spatial neighborhood and mapping process of *k*NN relationships**. We identify the k spatial neighbors for each observation in $\mathcal{D}$, and map such *k*NN relationship into F^k.

Step 2 (*lines: 5–19*) **PCR computation**.

a. (*line: 5*) `Category frequency computation and pair category identification`. The occurrence frequency of each observed category is computed, which is stored in $Freq_A$. All the possible pair categories are identified as PC_{Arr}, and N_p represents the number of pair categories.

b. (*lines: 6–14*) `Identification of pair category based pair object set`. By utilizing Eq. 12, the subset of F^k for each pair of categories is identified.

c. (*lines: 15–19*) `PCR computation`. The co-occurrence frequency and PCR value are calculated for each pair of categories.

Step 3 (*lines: 20–25*) **Computation of relevance among spatial objects.** PCR matrix is constructed, which stores PCRs between each reference object and its k spatial neighbors. The local relevance of each observation is calculated by the mean of the PCR values.

Algorithm 2 kNN-SCOD-S Approach

1: **for** $i = 1$ to n **do** {Construct the neighborhood matrix.}
2: $[Neighbor(i,:)] = kNN(S, r_i.S, n-1)$;
3: **end for**
4: $F^k = MapFunction(Neighbor, A)$; {Map the kNN relationship into dataset F^k.}
{Compute category frequency, identify pair category set and its size.}
5: $[Freq_A, PC\_Arr, N_p] = CateFreq(A)$;
6: **for** $i = 1$ to n **do** {Identify $F^k_{A^l A^{l'}}$ for each pair categories $A^l A^{l'}$.}
7: $A^l = A(i)$;
8: **for** $j = 1$ to k **do**
9: $A^{l'} = A(Neighbor(i,j))$;{Get the attribute for its j^{th} neighbor.}
{Add identified pair objects into $F^k_{A^l A^{l'}}$.}
10: **if** $(!FindIn(F^k_{A^l A^{l'}}, < r_{Neighbor(i,j), r_i} >))$ **then**
11: $F^k_{A^l A^{l'}} = AddIn(F^k_{A^l A^{l'}}, < r_i, r_{Neighbor(i,j)} >)$;
12: **end if**
13: **end for**
14: **end for**
15: **for** $l = 1$ to N_p **do** {Compute PCR value for each pair categories.}
16: $A^l A^{l'} = PC\_Arr(p)$;
17: $Freq(A^l, A^{l'}) = \frac{|F^k_{A^l A^{l'}}|}{|F^k|}$;;
18: $PCR_k(A^l, A^{l'}) = \frac{Freq(A^l, A^{l'})}{Freq(A^l) \cdot Freq(A^{l'})}$;
19: **end for**
20: **for** $i = 1$ to n **do** {Calculate PCR matrix between spatial object and its neighbors.}
21: **for** $j = 1$ to k **do**
22: $PCRMat(i,j) = PCR_k(A(i), A(Neighbor(i,j)))$;
23: **end for**
24: **end for**
25: $RelevanceArr = mean(PCRMat)$; {Compute relevances for spatial objects.}
26: $RankList = Rank(RelevanceMat, ascend)$; {Rank objects with ascending PCR values.}
27: $O_l = Outlier(RankList, 1:l)$ {Mark the outliers.}

Step 4 (*lines: 26–27*) **Outlier detection.** Finally, the objects are sorted with ascending PCR values, and the top l objects with lower relevance scores are recognized as outliers.

Computational complexity To form the neighborhood, it will take $O(n \log n)$ for kNN (Space partitioning) construction. It takes $O(n)$ to identify pair category array and compute the category frequencies. Identifying $F^k_{A^l A^{l'}}$ takes around $O(k \cdot n)$. And computing the PCR values for all pair categories takes around $O(N_p)$. Finally, constructing the relevance vector costs $O(k \cdot n)$. In summary, assuming $n \gg k$ and $n \gg N_p$, the total time complexity of kNN-SCOD-S is $O(n \log n)(= O(n \log n) + O(n) + O(k \cdot n) + O(N_p) + O(k \cdot n))$.

4 Spatial categorical outlier detection in multiple attribute dataset

The work of kNN based PCR approximation can be extended to solve the SCOD issue in multi-attribute domains. That is, given a spatial dataset, an outlying observation is the one whose non-spatial attribute set occurs infrequently with regard to

those of its spatial neighborhood. The calculation of PCR is computed by the frequency of co-occurrence of a pair of category sets at a specific spatial distance. Since PCF based approach is a time-consuming process, we only introduce how to compute PCR values in *k*NN-SCOD-M (*k* Nearest Neighbor based Spatial Categorical Outlier Detection in Multiple attribute dataset).

4.1 PCR computation in multi-attribute dataset

Similarly, *k*NN-SCOD-M approach first extracts the *k*NN relationship from the raw dataset $\mathcal{D}$ and maps it into F^k with $2m$ dimensions. In F^k, each data frame contains the attribute set, $\mathbb{A} = \{< A_1, .., A_m >, < A_1, ..., A_m >\}$, $A_i \subseteq \mathcal{C}_i, i \in [1, m]$. First, we need to learn two important concepts about attribute subset in $\mathcal{D}$ and F^k.

Definition 6 (Attribute Subset-AS) Given a dataset $\mathcal{D}$ with m categorical attributes, $A = \{A_1, .., A_m\}$, $A_i \subseteq \mathcal{C}_i$, its AS is defined as follows:

$$\mathcal{AS} = \left\{A^*, \{A^* = \{A_x, ..., A_y\}\}, 1 \leq x \leq y \leq m, A_i \subseteq D_i, i \in [x, y]\right\} \tag{13}$$

Considering *Attr*.2 of the sample spatial dataset as shown in Fig. 4, there are two category attributes: $A_1 = \{T, F\}$ and $A_2 = \{P, Q\}$. By definition, we can generate all of its ASs:

$$\mathcal{AS} : \left\{\{A_1\}, \{A_2\}, \{A_1, A_2\}\right\}.$$

Figure 4 also describes the *3*NN relationship among objects. By utilizing Definition 5, we can map it into a dataset set F^3 with 4-dimension attributes, as shown in Table 4.

Definition 7 (Pair Attribute Subset-PAS) Given a dataset F^k with $2m$ categorical attributes, $\mathbb{A} = \{< A_1, .., A_m >, < A_1, ..., A_m >\}$, $A_i \subseteq \mathcal{C}_i$, its PAS is defined as follows:

$$\begin{aligned}\mathcal{PAS} = \Big\{< A^*, A^{*'} >, \{A^* = \{A_x, ..., A_y\}\} \&\& \{A^{*'} = \{A_x, ..., A_y\}\} \\ \&\& \{|A^*| == |A^{*'}|\}, 1 \leq x \leq y \leq m, A_i \subseteq \mathcal{C}_i, i \in [x, y]\Big\}\end{aligned} \tag{14}$$

Apparently, A^* and $A^{*'}$ always exist in pairs in PAS. They are originated from the same attribute domain in dataset $\mathcal{D}$. For pair attribute set $\mathbb{A}$, there are $(2^m - 1)$ pair subsets. Similarly, we can enumerate all the PASs for the sample dataset in Fig. 4 as follows.

$$\mathcal{PAS} : \{\{< A_1 >, < A_1 >\}, \{< A_2 >, < A_2 >\}, \{< A_1, A_2 >, < A_1, A_2 >\}.$$

In this paper, we call AS and PAS observations as **AS and PAS combinations** which are generalized by the following concepts.

Definition 8 (AS Combination) Given a dataset $\mathcal{D}$ with m categorical attributes, $A = \{A_1, .., A_m\}$, $A_i \subseteq \mathcal{C}_i$. Suppose for each attribute A_i, it takes value in $\mathcal{L}_i$=$\{A_i^1, ..., A_i^{L_i}\}$. Therefore, an AS Combination (ASC) is one of the possible

category examples of existing attributes in the AS. The ASC of A^* (defined in Definition 7) is,

$$\mathcal{ASC}_{A^*} = \left\{\mathcal{A}, \mathcal{A} = \left\{A_x^{l_x}, ..., A_y^{l_y}\right\}, A_i^{l_i} \subseteq \mathcal{L}_i, i \in [x, y]\right\} \tag{15}$$

As we can see, each observation may take one of values $\{T, F\}$ for A_1, and $\{P, Q\}$ for A_2. $\{A_1\}$ is one of its ASs, then $ASC_{A_1} = \{\{T\}, \{F\}\}$. Similarly, $ASC_{\{A_1, A_2\}} = \{\{T, P\}, \{T, Q\}, \{F, P\}, \{F, Q\}\}$.

Definition 9 (PAS Combination) Given a dataset F^k with $2m$ category attributes, $\mathbb{A} = \{< A_1, .., A_m >, < A_1, ..., A_m >\}$, $A_i \subseteq \mathcal{C}_i, i \in [1, m]$. Similarly, it takes value in $\mathcal{L}_i$=$\{A_i^1, ..., A_i^{L_i}\}$. Therefore, an PAS Combination (ASC) is one of the possible pair category examples of existing attributes in the PAS. That is,

$$\begin{aligned}\mathcal{PASC}_{<A^*, A^{*'}>} = \Big\{< \mathcal{A}, \mathcal{A}' >, \left\{\mathcal{A} = \left\{A_x^{l_x}, ..., A_y^{l_y}\right\}\right\} \& \left\{\mathcal{A}' = \left\{A_x^{l_{x'}}, ..., A_y^{l_{y'}}\right\}\right\} \\ \& \left\{|\mathcal{A}| == |\mathcal{A}'|\right\}, 1 \le x \le y \le m, A_i^{l_i}, A_i^{l_{i'}} \subseteq \mathcal{L}_i, i \in [x, y]\Big\}\end{aligned} \tag{16}$$

For example, there are three possible PASCs: $\{< T >, < T >\}$, $\{< T >, < F >\}$ and $\{< F >, < F >\}$ for PAS $\{< A_1 >, < A_1 >\}$ in F^3. And, all the observed PASCs for $\{< A_1, A_2 >, < A_1, A_2 >\}$ in Fig. 4 are enumerated in the second column in Table 4.

We can compute the PCR value for each PASC of a specified PAS $< A, A' >$ by utilizing the following 5 steps.

- **Dataset identification for specific ASC**. For a specific ASC $\mathcal{A} = \{A_x^{l_x}, ..., A_y^{l_y}\}$, we scan the dataset $\mathcal{D}$ and identify $\mathcal{D}_\mathcal{A}$ as follows.

$$\mathcal{D}_\mathcal{A} = \left\{r_i, \left(r_i.A_x = A_x^{l_x}\right) \& \cdots \& \left(r_i.A_y = A_y^{l_y}\right), r_i \in \mathcal{D}, A_j^{l_j} \subseteq \mathcal{L}_j, j \in [x, y]\right\} \tag{17}$$

- **Frequency computation of ASC.** Calculate the frequency for $\mathcal{D}_\mathcal{A}$ using Eq. 18.

$$Freq(\mathcal{D}_\mathcal{A}) = \frac{|\mathcal{D}_\mathcal{A}|}{|\mathcal{D}|} \tag{18}$$

  For ASC $\{T\}$ in Fig. 4, $D_T = \{r_1, r_5\}$, and its frequency is $2/7 = 0.29$.
- **Dataset identification for specific PASC**. Suppose $< \mathcal{A}, \mathcal{A}' > =< \{A_x^{l_x}, ..., A_y^{l_y}\}, \{A_x^{l_{x'}}, ..., A_y^{l_{y'}}\} >$ is one of the PASCs of $< A^*, A^{*'} >(\{A^* = \{A_x, ..., A_y\}\}, \{A^{*'} = \{A_x, ..., A_y\}\})$. Scan the whole set F^k and identify all observations which satisfy the following conditions.

$$\begin{aligned}\mathcal{F}^k_{\{\mathcal{A}, \mathcal{A}'\}} = \Big\{f_i, \Big[\left(f_i.A^*.A_x = A_x^{l_x}\right) \& ... \& \left(f_i.A^*.A_y = A_y^{l_y}\right)\Big\} \\ \& \left\{\left(f_i.A^{*'}.A_x = A_x^{l_{x'}}\right) \& ... \& \left(f_i.A^{*'}.A_y = A_y^{l_{y'}}\right)\right] \\ || \left[\left(f_i.A^{*'}.A_x = A_x^{l_x}\right) \& ... \& \left(f_i.A^{*'}.A_y = A_y^{l_y}\right)\right\} \\ \& \left\{\left(f_i.A^*.A_x = A_x^{l_{x'}}\right) \& ... \& \left(f_i.A^*.A_{y'}\, y = A_y^{l_{y'}}\right)\right], f_i \in F^k\Big\}\end{aligned} \tag{19}$$

  In $F^k_{\{\mathcal{A}, \mathcal{A}'\}}$, it stores all pairs of objects which have the same PASC $< \mathcal{A}, \mathcal{A}' >$.

Table 5 PCR computation

Pair Categories	Freq.	Prob.	PCR
$< \{T, P\},\{F, Q\} >$	3	0.25	2.00
$< \{T, P\},\{F, P\} >$	3	0.25	2.00
$< \{F, Q\},\{F, Q\} >$	1	0.083	0.99
$< \{F, Q\},\{F, P\} >$	4	0.34	4.04
$< \{F, P\},\{F, P\} >$	1	0.083	0.99
$Freq(\{F, P\}) = 0.29; Freq(\{F, Q\}) = 0.29$			
$Freq(\{T, P\}) = 0.43$			

– **Frequency computation of PASC**. Calculate the frequency for each PASC $< \mathcal{A}, \mathcal{A}' >$.

$$Freq(< \mathcal{A}, \mathcal{A}' >) = \frac{|F^k_{\{\mathcal{A},\mathcal{A}'\}}|}{|F^k|} \tag{20}$$

For example, for PASC $\{< T, P >, < F, P >\}$, $\mathcal{F}^3_{\{<T,P>,<F,P>\}} = \{< r_1, r_3 >, < r_1, r_7 >, < r_5, r_7 >\}$. Its frequency is equal to $3/12 = 0.25$.

– **PCR computation**. Finally, PCR for a PASC can be calculated as Eq. 21.

$$PCR_k(< \mathcal{A}, \mathcal{A}' >) = \frac{Freq(< \mathcal{A}, \mathcal{A}' >)}{Freq(\mathcal{D}_\mathcal{A}) \cdot Freq(\mathcal{D}_{\mathcal{A}'}))} = \frac{|F^k_{\{\mathcal{A},\mathcal{A}'\}}|/|F^k|}{(|\mathcal{D}_\mathcal{A}| \cdot |\mathcal{D}_{\mathcal{A}'}|)/|\mathcal{D}|^2} \tag{21}$$

Therefore, we can compute PCR on $\{< T, P >, < F, P >\}$ among object r_1 and r_3 as $PCR(< r_1, r_3 >) = 0.25/(0.29 * 0.43) = 2.00$. Table 5 shows all the PCR computation of all possible PASC for Fig. 4 on the PAS $\{< A_1, A_2 >, < A_1, A_2 >\}$. In the same way, we can compute the PCR values of other PASCs with respect to different PAS.

By scanning the dataset, we can determine, for each pair of spatial objects, there are at most $2^m - 1$ PCR scores which correspond to $2^m - 1$ PASC. After that, we choose the smallest one as their final relevance score. We identify relevance among objects in this way because, sometimes, an outlier only exists in the subspace of multiple attributes. Exhaustively estimating outlier scores in different PASC will help identify SCOs more effectively. With the smallest PCR vector, we can construct a PCR matrix (n-by-k). Further, the outlierness value can be computed for each object using the mean of neighborhood relevance.

4.2 Algorithm of *k*NN-SCOD-M

We generalize the *k*NN-SCOD-M approach to detect multi-attribute outliers. There are *6* input parameters, S, A, n, k, m and l, which are described in Table 1. As shown in Algorithm 3, identifying SCOs with multiple attributes includes the following *5* steps.

Step 1 (*lines: 1–4*) (**Construction of spatial neighborhood and mapping process of *k*NN relationships**). For each data observation r_i, identify its k spatial neighbors and map *k*NN relationship into F^k.

Step 2 (*lines: 5–18*) **PCR computation**.

a. (*line: 5*) `Identification of AS and PAS`. Algorithm 4 describes this function in detail. Intuitively, there are $\binom{m}{i}$ ASs which consist of i attributes, and $\binom{m}{i}$ PASs with size as $2 * i$. Therefore, in each loop, $\binom{m}{i}$ ASs and PASs are identified, respectively.

Algorithm 3 ***k*NN-SCOD-M Approach**

1: **for** $i = 1$ to n **do** {Construct the neighborhood matrix.}
2: $[Neighbor(i,:)] = kNN(S, r_i.S, k)$;
3: **end for**
4: $F^k = MapFunction(Neighbor, A)$; {Map the kNN relationship into dataset F^k.}
5: $[PAS, AS] = ASIdentify(A)$; {Identify all possible PASs and ASs.}
{Identify all possible PASCs and ASCs for each PAS and AS, and then extract their corresponding subsets.}
6: $[PASC, ASC, \mathcal{D}_{ASC}, \mathcal{F}^k_{PASC}] = ASCIdentify(A, PAS, AS)$;
{Compute the frequency vectors for each $\mathcal{D}_{ASC}, \mathcal{F}^k_{PASC}$.}
7: $[Freq_{asc}, Freq_{pasc}] = FreqCompute(\mathcal{D}_{ASC}, F^k_{PASC})$;
8: **for** $i = 1$ to n **do** {Calculate PCR matrix between spatial object and its neighbors.}
9: $A_i = r_i.A$;
10: **for** $j = 1$ to k **do**
11: $r_j = Neighbor(i, j)$;
12: $A_j = r_i.A$;
13: **for** $a = 1$ to $2^m - 1$ **do**
14: $curPAS = PAS(a); curAS = AS(a)$; {With the information from r_i and r_j, identify its corresponding PASC and ASCs.}
15: $[curPASC, curASC_1, curASC_2]$=$PASCIdenfity(curPAS, curAS, r_i, r_j)$;
{Compute the PCR values for neighbor objects $< r_i, r_j >$ in current PAS space.}
16: $PCRMat_a(i,k) = \frac{Freq_{pasc}(curPASC)}{Freq_{asc}(curASC_1) \cdot Freq_{asc}(curASC_2)}$;
17: **end for**
18: **end for**
19: **for** $a = 1$ to $2^m - 1$ **do** {Identify the smallest one of PCRs in all the PAS space as its final PCR value.}
20: **if** a==1 **then**
21: $PCR(i) = mean(PCRMat_a(i,:))$;
22: $tempValue = PCR(i)$;
23: **else**
24: $PCR(i) = min\{mean(PCRMat_a(i,:)), tempValue\}$;
25: $tempValue = PCR(i)$;
26: **end if**
27: **end for**
28: **end for**
29: $RelevanceMat = PCR$; {Compute relevances for spatial objects.}
30: $RankList = Rank(RelevanceMat, ascend)$; {Rank objects with ascending PCR values.}
31: $O_l = Outlier(RankList, 1:l)$ {Mark the outliers.}

Algorithm 4 PAS and AS Identification
$[PAS, AS] = ASIdentify(A)$

1: $Label = 0; m = |A|$;
2: **for** $i = 1$ to m **do** {Identify AS,PAS whose size are i, $2 * i$, respectively.}
3: $AS((i + Label) : (\binom{m}{i} + Label)), :) = \{AS^i, (AS^i \subseteq A)\ \&\ (|AS^i| = i)\}$;
4: $PAS((i + Label) : (\binom{m}{i} + Label)), :) = \{< AS^i, AS^i >, (AS^i \subseteq A)\ \&\ (|AS^i| = i)\}$;
5: $Label = \binom{m}{i} + Label$;
6: **end for**

b. (*lines: 6–7*) `Frequency computation of` $\mathcal{D}_{ASC}$ `and` F^k_{PASC}. We first operate the identification of possible ASCs and PASCs for each AS^* and PAS^* Meanwhile, their corresponding subset, $\mathcal{D}_{\mathcal{ASC}^*}$ and $\mathcal{F}^k_{\mathcal{PASC}^*}$ are identified by scanning $\mathcal{D}$ and $\mathcal{F}^k$. As shown in Algorithm 5, for each object in $\mathcal{D}$, any subset of its attributes can be mapped as one of the possible ASCs (Step 11 and 14). In the same way, for each pair observations in F^k, any subset of their attributes originated from

Algorithm 5 Identification of $PASC$, ASC, $\mathcal{D}_{ASC}$ and $\mathcal{F}_{PASC}$.
$[PASC, ASC, \mathcal{D}_{ASC}, \mathcal{F}_{PASC}] = ASCIdentify(A, PAS, AS, \mathcal{D})$

1: $PASC = \{\}; ASC = \{\}; \mathcal{D}_{ASC} = \{\}; \mathcal{D}_{PASC} = \{\};$
2: **for** $i = 1$ to n **do**
3: $A_j = r_i.A$;
4: **for** $j = 1$ to k **do**
5: $r_j = Neighbor(i, j)$;
6: $A_j = r_j.A$;
7: **for** $a = 1$ to $2^m - 1$ **do**
8: $\mathcal{AS} = AS(a)$;{Identify current AS .}
9: $\mathcal{PAS} = PAS(a)$;{Identify current PAS.}
{Generate one of the ASCs based on the attributes of object r_i.}
10: $ASC_i^a = A_i\{\mathcal{AS}\}$;
11: $ASC_{\mathcal{AS}} = Add\_ASC(ASC_{\mathcal{AS}}, ASC_i^a)$;
{Add object r_i into its corresponding ASC subset.}
12: $\mathcal{D}_{ASC_{\mathcal{AS}}} = AddObjInASCSet(D_{ASC_{\mathcal{AS}}}, r_i)$;
{Generate one of the ASCs based on the attributes of object r_j.}
13: $ASC_j^a = A_j\{\mathcal{AS}\}$;
14: $ASC_{\mathcal{AS}} = Add\_ASC(ASC_{\mathcal{AS}}, ASC_j^a)$;
{Add object r_j into its corresponding ASC subset.}
15: $\mathcal{D}_{ASC_{\mathcal{AS}}} = AddObjInASCSet(D_{ASC_{\mathcal{AS}}}, r_j)$;
{Generate one of the PASCs based on the pair attribute sets of $< r_i, r_j >$.}
16: $PASC_{\mathcal{PAS}} = Add\_PASC(PASC_{\mathcal{PAS}}, < ASC_i^a, ASC_j^a >)$;
{Add $< r_i, r_j >$ into its corresponding PASC subset.}
17: $\mathcal{F}^k_{PASC_{\mathcal{PAS}}} = AddObjInPASCSet(F^k_{PASC_{\mathcal{PAS}}}, < r_i, r_j >)$;
18: **end for**
19: **end for**
20: **end for**

the same domains can be recognized as one of the possible PASCs (Step 16). After that, we map each object into its corresponding ASC subset (Steps 12, 15), and each pair of the object and its current spatial neighbor into the corresponding PASC subset. Finally, the frequencies of all the $\mathcal{D}_{ASC}$ and F^k_{PASC} are computed by using Eqs. 18 and 20.

c. (*lines: 8–18*) `PCR computation for specific PASC`. Compute the PCR values between reference object and its *k*NN neighbors for each possible PASC.

Step 3 (*lines: 19–28*) **Computation of Relevances among objects**. Use the mean of *k* PCRs as the relevance value in each PAS subspace. And, the smallest one of the $2^m - 1$ PCR values is recognized as its final relevance score.

Step 4 (*lines: 30–32*) **Outlier detection**. Finally, the objects are sorted with ascending relevance values, and the top *l* objects with lower relevance scores are recognized as outliers.

Computational complexity To form the neighborhood, it will take $O(n \log n)$ for *k*NN (Space partitioning) construction and mapping process. As shown in Algorithm 4, it takes around $O(2^m - 1)$ to identify all possible PASs and ASs. Algorithm 5 demonstrates that it takes $O(n * k * (2^m - 1))$ to detect all possible PASCs, ASCs and their corresponding subsets. Finally, computing the final PCR value for each observation takes $O(n * k * (2^m - 1))$. In summary, assuming $n \gg k$ and $n \gg m$, the total computational complexity of *k*NN-SCOD-M approach is $O(n * (2^m - 1))(= O(n \log n) + O(2^m - 1) + O(n * k * (2^m - 1)) + O(n * k * (2^m - 1)))$.

5 Experiment results and analysis

We conducted extensive experiments on both simulated and real datasets to compare the performances of PCF-SCOD and *k*NN-SCOD, with other popular outlier detection approaches [7, 8, 14, 32, 37].

5.1 Experiment settings

This subsection introduces simulation and real life datasets, the outlier detection methods, and performance metrics.

5.1.1 Simulation and real dataset

For experiments in the single attribute domain, we applied all the approaches into one simulated and three real datasets. For those in the multiple attribute domain, since there was no public baseline dataset, we evaluated them on simulation datasets.

Simulation dataset The simulation categorical datasets were generated by discretization from some numerical simulation datasets. Denote a numerical dataset S as $\{Z(s_1),...,Z(s_n)\}$, $Z(s_i) \in R^m$ $(i = 1 \cdots n)$, where m is the number of non-spatial attributes. The simulation datasets $S_1, ..., S_m$ were generated by a Gaussian random field model defined as follows:

$$[Z(s_1)^T, ..., Z(s_n)^T]^T \sim N\left(0, \begin{bmatrix} \sum_{11}(\theta_{11}) & \cdots & \sum_{1n}(\theta_{1n}) \\ \vdots & \ddots & \vdots \\ \sum_{n1}(\theta_{n1}) & \cdots & \sum_{nn}(\theta_{nn}) \end{bmatrix}\right)$$

$$[\theta_{ij}]_1 \sim Uniform(1.17, 1.85), [\theta_{ij}]_2 \sim Uniform(2.00, 3.24), i = j$$

$$[\theta_{ij}]_1 \sim Uniform(1.00, 1.44), [\theta_{ij}]_2 \sim Uniform(2.30, 2.80), i \neq j$$

$$s_1, \cdots, s_n \sim Uniform(0, 5). \tag{22}$$

where $Z(s_i) = [z_1(s_i), \cdots, z_m(s_i)]^T$, $\sum_{ij}(\theta_{ij}) = Var(Z(s_i), Z(s_j)), \theta \in R^2$. $[\sum_{ij}(\theta_{ij})]_{km}$ is defined by an exponential model as

$$\left[\sum_{ij}(\theta_{ij})\right]_{km} = [\theta_{ij}]_1 \cdot e^{\frac{\|s_i - s_j\|}{|\theta_{ij}|_2}} \tag{23}$$

$[\theta_{ij}]_1$ and $|\theta_{ij}|_2$ are named as sill and range parameters, respectively. The above simulation model parameters were determined based on the distribution of a benchmark data set *97data.dat* available in GSLIB software (http://www.gslib.com/), which has two attributes. It was fitted by the Gaussian random field model which has a quadratic trend (mean), and the cross covariance functions were approximated by exponential models with sill and range parameters (1.85, 2.00) for the first attribute, (1.17, 3.24) for the second attribute, and (1.22, 2.55) for the cross covariance between the two attributes. Note that, in our simulation model, we did not consider any trend, and the data distribution was determined purely by the cross covariance functions, which potentially increases the complexity of the distribution. We did not fix the sill

Table 6 Three simulation dataset

Dataset	Size	Dimension	The number of observing categories in each dimension
Syn_1	4000	1	A_1: 3
Syn_2	4000	3	A_1: 3; A_2: 3; A_3: 3;
Syn_3	4000	3	A_1: 3; A_2: 3; A_3: 3;
Syn_4	1000	4	A_1: 3; A_2: 4; A_3: 5; A_3: 6;
Syn_5	4000	2	A_1: 5; A_2: 8;

and range parameters, but instead sampled the parameters from uniform distributions around the estimated sill and range parameters for *97data.dat*. Our model was able to flexibly generate spatial simulation data sets with multiple attributes. After the numerical data sets were generated, we applied a discretization process to convert the numerical data into categorical data. To illustrate the discretization strategy, suppose we need to convert a numerical attribute data $\{Z_1(s_1), \cdots, Z_1(s_n)\}$ into 3 categories data, we sorted the values and then separated them into three groups such that the orders among the three groups were preserved. This means, the objects in group 2 is always larger than those in group 1. The similar situation is for group 2 and 3. Since we focused on nominal data, we assigned each data a label with a unique category so that the data attributes could not be ordered.

For the five generated simulation data sets, shown in Table 6, Syn_1 was utilized in the experiments for the single attribute domain, while Syn_2, Syn_3, Syn_4 and Syn_5 were used in those for the multiple attribute one.

Real dataset We also executed the SCOD approaches on three real datasets with single attribute to further demonstrate their effectiveness. The three datasets include *Jura*, $Soil_1$ and $Soil_2$. *Jura* data is a well-known categorical dataset from Pieere Goovarerts book[16]. In the original dataset, five rock types are available. Following Bel et al. [5], Portlandian is grouped with Quanternary into category 4, because its frequency of occurrence is very low (1.2 %), which makes those observations taking Portlandian general outliers. $Soil_1$ and $Soil_2$ dataset were both extracted from Harmonized World Soil Database (http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/). Table 7 describes their detailed information. Figure 5 provides the data distribution of parts of raw soil datasets which were utilized in our experiment. $Soil_1$ data seems to distribute uniformly, but that of *Soil*2 is more complicated, which needs higher identification qualities for SCOD approaches.

Table 7 Three real datasets

Dataset	Size	Dimension	The observing categories
Jura	359	1	A^1:Argovian; A^2:Kimmeridgian; A^3:Sequanian; A^4:Quaternary;
$Soil_1$	1000	1	A^1:LP-Leptosol; A^2:CL-alcisol; A^3:RK-utcrops; A^4:DS-Sand Dunes;
$Soil_2$	3000	1	A^1:LV-Luvisol; A^2:LP-Leptosol; A^3:PT-Plinthosol; A^4:VR-Vertisol; A^5:NT-Nitisol; A^6:LX-Lixisol; A^7:FL-Fluvisol;

Fig. 5 Data distribution of three real-life datasets.(Left:*Jura*;Middle:$Soil_1$;Right:$Soil_2$)

5.1.2 Outlier detection approach

We compared the performances of our proposed methods, denoted as PCF-SCOD and *k*NN-SCOD, to other existing methods introduced in this subsection.

Univariate detection methods

TCOD Considering the local correlation property of spatial data, we chose NN (Nearest Neighbor) based techniques for SCOD tasks. Typical approaches include *k*NN [37] and LOF [8] methodologies. To compute the similarities among nominal data, we used Lin and OF measurements which showed high performances [7, 11]. We combined the NN based techniques with categorical similarity measurements together to get overall 4 different comparable "TCOD" approaches: LOF-Lin, LOF-OF, *k*NN-Lin and *k*NN-OF.
It is noted that *k*NN-SCOD is not related to *k*NN-Lin and *k*NN-OF, although they have the same prefix "*k*NN". As we introduced above, *k*NN-Lin and *k*NN-OF were generated from one of the most popular traditional numerical outlier detection approaches: *k*NN [37], while *k*NN-SCOD is proposed in this paper by introducing a novel *k*NN mapping function process as an effective and efficient approximation of the general PCR computation.

SCOD Z-test is one of the most typical methods to identify SNOs. When operating it on categorical data, we integrated Z-test with Lin and OF measurements. As a result, there were 2 comparable "SCOD" approaches: Z-OF and Z-Lin. Also, we directly applied Z-test into the categorical datasets by assuming that the nominal categories can be ordered, denoted by Z-SNOD.

Multivariate detection methods

TCOD Several advanced TCOD methods have been proposed for multivariate categorical data, including Bayes Net Method, Marginal Method, LERAD, Conditional Test, Conditional Test-Combining Evidence, and Conditional Test-Partitioning. Experiments had shown that Conditional Test and its two variants outperformed all other methods [14]. Therefore, we focused on the comparison of our method with the three best methods, denoted as Conditional Test, Conditional Test-Combining Evidence, and Conditional Test-Partitioning.

SCOD For the competing methods in SNOD group, we chose Multivariate Z-SNOD, which is an extension version of single attribute Z-test, by considering Mahalanobis distance and MCD (Minimum Covariance Determinant) techniques. In addition, we integrated the preceding method with multiple categorical similarity measurements, Lin and OF, named as Multivariate Z-Lin, and Multivariate Z-OF.

5.1.3 Performance metrics

We generated synthetic outliers in both simulation and real datasets, which enable us to analyze the effectiveness of outlier detection approaches in a controllable way. We assumed the raw dataset as a ground truth, and contaminated around α percent of the data records as outliers. In our paper, for each dataset, including both simulation and real life ones, we randomly selected 2, 3 and 5 % of the data to be anomalies by modifying them from its original category to anyone of others. For each contamination rate (*2*, *3* and *5*), the synthetic outliers were generated 10 times and the mean and standard deviation of accuracies were calculated for each method.

To compare the accuracies among all methods, we used the common evaluation measures: *precision* (detection rate), i.e., *the fraction of examples labeled as outliers that are true outliers*, and *recall* (detection precision), i.e., *the fraction of true outliers that are correctly identified*. The precision is plotted against recall, and the curves that are higher and farther to the right denote better performances since it corresponds to a higher precision for a given recall. Each point corresponds precision and recall when a specified number of outlier is predefined, from 1 to n (the number of objects in the whole dataset). As another measure of accuracy, *average precision* was computed to approximate the area under the precision-recall curve.

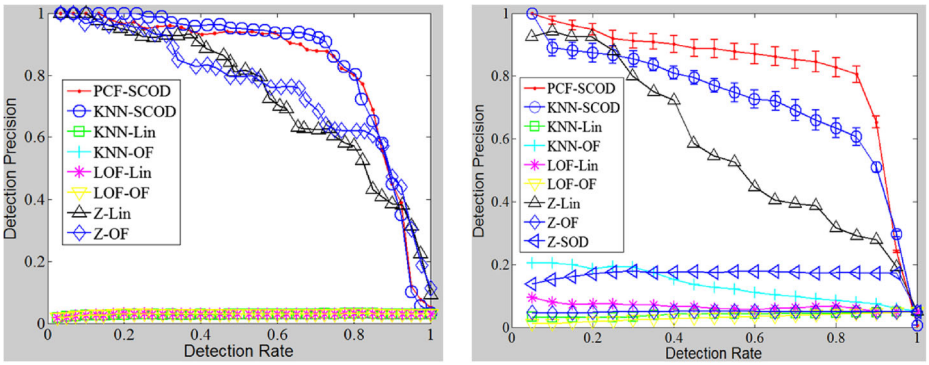
All the experiments were conducted in a PC with Intel (R) Core (TM) Duo CPU, CPU 2.80 GHz, and 2.00 GB memory. The development tool was MATLAB 2008.

5.2 Experiment results and analysis

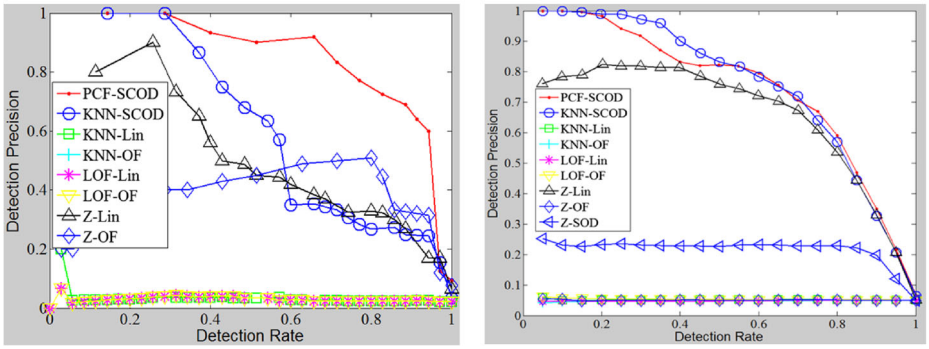
This section presents experimental evaluations for the above approaches on simulation and real datasets. We compared the SCOD accuracies among different methods based on different parameter combinations.

5.2.1 Results on single attribute datasets

Detection accuracy Figure 6 depicts the comparison of our methods against the other 7 existing approaches on the single attribute datasets. The contamination rate α was set as *3*, *5*, *2* and *5* in $Soil_1$, $Soil_2$, *Jura* and Syn_1, respectively. Each point in the curves corresponds to the average performance over *10* randomly generated datasets for each algorithm. We observed that both PCF-SCOD and *k*NN-SCOD methods achieved 20–40 % improvement over Z-OF and Z-Lin, and 60–70 % over LOF-Lin, LOF-OF, *k*NN-Lin, *k*NN-OF and Z-SNOD(Z-SOD). From these results, we found that *k*NN and LOF can't handle the categorical outlier detection in spatial context. After integrating Z-test with OF and Lin similarity measurements together, the outlier identification quality was increased. Z-Lin was always better than Z-OF.



(a) $Soil_1$, $\alpha = 3$ (b) $Soil_2$, $\alpha = 5$



(c) $Jura$, $\alpha = 2$ (d) Syn_1, $\alpha = 5$

Fig. 6 Comparison of algorithm performances for the spatial dataset with single attribute

As introduced in [7], OF and Lin compute similarities for categorical attributes in different ways:

$$Sim_{\mathrm{OF}}(X, Y) = \begin{cases} 1 & \text{if } X = Y; \\ \frac{1}{1+\log(N/f_k(X_k))\times\log(N/f_k(Y_k))} & \text{otherwise.} \end{cases}$$
$$\omega_k = 1/d \tag{24}$$

$$Sim_{\mathrm{Lin}}(X, Y) = \begin{cases} 2\log p_k(X_k) & \text{if } X = Y; \\ 2\log(p_k(X_k) + p_k(Y_k)) & \text{otherwise.} \end{cases}$$
$$\omega_k = \frac{1}{\Sigma_{i=1}^{d} \log p_i(X_i) + \log p_i(Y_i)} \tag{25}$$

where $f_k(X_k)$ denotes the number of times attribute to take the value X in the k^{th} dimension, $p_k(X_k)$ the sample probability to take the value X_k in the dataset, and ω_k is the weight value of the k^{th} dimension. When identifying the relevance score among objects with the same categorical attribute, OF always assigns a constant value 1 to it, while Lin computes the value based on the occurrence probability of the category. When two objects have different categorical attributes, OF assigns lower

relevance to the objects with higher frequencies, while Lin assigns higher values. Consequently, Lin could better capture the spatial relationship than OF by more accurately computing spatial relevance among objects. If the category of one of the pair objects occurs frequently in the dataset, it means the higher probability to co-occur with another category in the whole dataset. That is why the methods integrating with Lin always achieved better performance than those with OF. However, when identifying the spatial categorical outliers, Lin and OF measurements are based on the category frequencies that are determined by the whole data distribution, not the co-occurrence frequencies which take spatial dependency into considerations. That was why the performances of Z-Lin and Z-OF were worse than those of PCF-SCOD and *k*NN-SCOD.

Compared with *k*NN-SCOD, PCF-SCOD had better performance. Especially, when applied to the datasets with more complicated distribution, like *Jura* and $Soil_2$ (as shown in Fig. 5), PCF-SCOD can accurately capture the relationships among objects by considering PCRs among pair of categories at different spatial distances. Furthermore, it can get the highest precision of identifying spatial categorical outliers. Also, with the increasing data size, *k*NN-SCOD got better approximations of PCF-SCOD, like in $Soil_1(1000)$, $Soil_2(3000)$ and $Syn_1(4000)$, since with the more objects in the dataset, *k*NN-SCOD could capture sufficient mapping information from the raw dataset, which helps accurately approximate PCRs for pair objects. Finally, we found that the identification qualify of PCF-SCOD was not affected by different contamination rates. For each contamination value, PCF-SCOD always achieves higher accuracy with stable process abilities, as shown by its small standard deviations of detection precisions in Fig. 6b.

The average precision values are also given in Table 8 for all the SCOD approaches in single attribute domain. Note that for most datasets, PCF-SCOD and *k*NN-SCOD achieved higher accuracy than other approaches. We notice that the performance of the methods also depends on the detection tasks. For example, *k*NN-SCOD has comparable or better performance than PCF-SCOD in $Soil_1$ and Syn_1. In $Soil_1$, only 3 % of data are contaminated which makes the outlying behavior more obvious based on the information derived from the normal objects. There are 3 categories in Syn_1, and only 6 possible pair attributes. It is sufficient for 4000 observations to extract the normal pair attributes which co-occur frequently by analyzing those behaviors. On the contrary, *k*NN-SCOD can't work as well as PCF-SCOD in *Jura*. There are only 359 observations which take four different categories, which means there are overall 10 pair attributes. The neighborhood information

Table 8 Average precision (normalized area under precision-recall curve) for spatial categorical datasets with single attribute, comparing PCF-SCOD, *k*NN-SCOD-S and other 7 approaches

Approach for single attribute dataset	$Soil_1$	$Soil_2$	*Jura*	Syn_1
PCF-SCOD	0.7805	0.7822	0.7481	0.7646
*k*NN-SCOD-S	0.7811	0.7763	0.6521	0.7148
*k*NN-Lin	0.0279	0.0389	0.0276	0.0489
*k*NN-OF	0.0284	0.1261	0.0279	0.0454
LOF-Lin	0.0284	0.0621	0.0279	0.0455
LOF-OF	0.0284	0.0300	0.0279	0.0502
Z-Lin	0.6781	0.5407	0.4362	0.6298
Z-OF	0.6966	0.0473	0.3668	0.0477
Z-SNOD	0.1845	0.1603	0.0807	0.2072

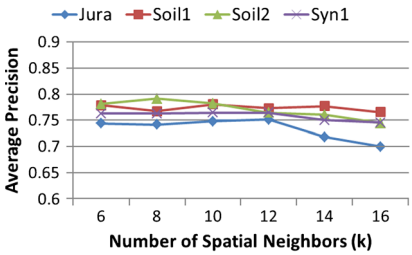


Fig. 7 Average precisions of PCF-SCOD by varying k value

in these 359 observations can't provide substantial information to help derive the normal behaviors. On the other hand, by observing Fig. 5, we notice that the distribution of the whole dataset is not uniform. It seems that there are various kinds of pair attributes co-occurring in near distances, such as, blue-blue, blue-orange, red-red, red-orange, red-while, and white-orange, etc. In this sense, only considering the neighborhood based information is not sufficient to mine the normal patterns. Although some pair attributes often co-occur within a near region, but they maybe not within a little distant one. PCF-SCOD can extract the co-occurrence frequencies of pair attributes at different distances. That is why PCF-SCOD performs well in $Jura_1$.

Impacts of neighborhood sizes We also evaluated the anomaly detection performances of proposed approaches by varying the sizes of spatial neighborhood. Figure 7 shows various k values from 6 to 16, respectively. The curves depict the effects of varying the number of spatial neighbors on the average precisions of PCF-SCOD on 4 different sizes of datasets. In general, its anomaly detection performance seems stable as the neighborhood size increases. In $Soil_1$, $Soil_2$ and Syn_1, the optimal k values are around 8 to 14. But for *Jura*, which is a small-size data set, higher k value leads to a worse performance. This is because, higher neighborhood size makes distant objects involve in evaluating the specified observation behaviors, which violates the spatial correlation theory. This same situation occurred in the results generated by kNN-SCOD method, as shown in Fig. 8. Since kNN-SCOD work is based on the neighborhood information, which makes it more sensitive to the k value. For the dataset with larger data size, 8-16 neighborhood size is appropriate to collect the co-occurrence information of pair attributes. This is proved by the curves of $Soil_1$, $Soil_2$ and Syn_1. For small dataset, both lower and higher k values result in worse identification performances.

Impact of bin sizes The effects of bin sizes were examined on the performances of PCF-SCOD method. Figure 9 shows its performances keep impressive by varying b values from 6 to 20. Apparently, SCOD identification quality achieves stable after

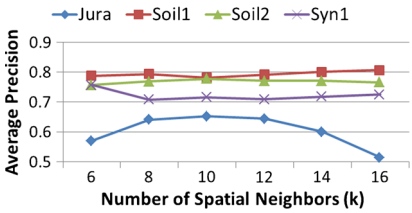


Fig. 8 Average precisions of kNN-SCOD-S by varying k value

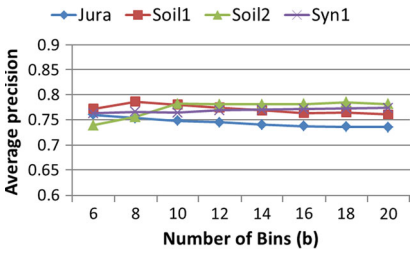


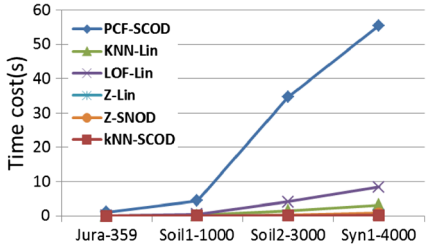
Fig. 9 Average precisions of PCF-SCOD by varying b value

the points at 10. PCF-SCOD computes the pair attribute frequencies at different bins. If b is smaller, e.g., $b < 8$, the pair observations in more distant region are mixed with those in nearer region that we are interested. This results in the incorrect computation of the co-occurrence frequency of pair attributes, which further leads to the worse identification performances. On the contrary, higher b value helps compute the pair frequency more accurately. However, too much bins will cost a lot, and normally, it is sufficient to set b as 10 which is demonstrated by the curves in Fig. 7.

Computational cost analysis Finally, we showcase the speed and respective scalability of the algorithms. Figure 10 contains the runtime performances of algorithms in the datasets with varying number of data observations. As observed, the methods based on Lin have similar runtime with those based on OF. Therefore, we only show the Lin based approaches. As shown in Fig. 10, PCF-SCOD finished execution at around 1.14 s for *Jura* data, while kNN-SCOD had a running time of 0.01 s. And, in Syn_2 dataset, PCF-SCOD is at around 55.5 s, while kNN-SCOD is at only 0.18 s. By analyzing the results in Fig. 10, kNN-SCOD approximated the accuracy of PCF-SCOD very well, while it outperformed PCF-SCOD for larger-size datasets. For other compared approaches, although they finished running more quickly than PCF-SCOD, they had lower identification accuracies.

5.2.2 Results on multiple attribute datasets

Detection accuracy Figure 11 shows the performances when contamination rate was 5 % in simulation datasets. Obviously, kNN-SCOD-M still had very preceding performance increases. The curves demonstrate that kNN-SCOD performs the best, followed by the Multivariate Z-Lin and Multivariate Z-SNOD. The series of Conditional Test perform very poorly in comparison. The worst one is Multivariate Z-OF, since OF measurement can't handle well the similarities among nominal data



Data Size	359	1000	3000	4000
PCN-SCOD	1.14	4.41	34.6	55.5
kNN-SCOD	0.01	0.02	0.11	0.18
kNN-Lin	0.01	0.14	1.48	3.14
LOF-Lin	0.04	0.44	4.11	8.39
Z-Lin	0.001	0.02	0.12	0.23
Z-SOD	0.001	0.02	0.16	0.75

Fig. 10 Runtime in seconds for datasets with varying size

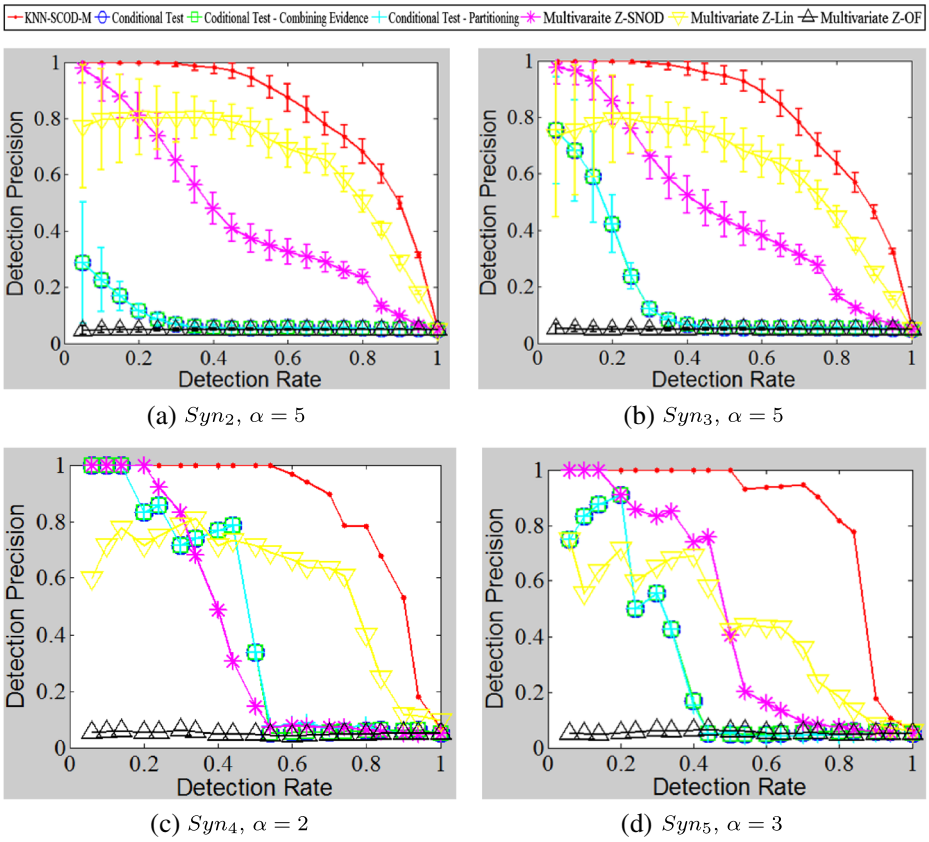


Fig. 11 Comparison of algorithm performances for the spatial data with multiple attributes

with multiple attributes. Meanwhile, the curves of all methods are also depicted with the standard variances of precision values of the *10* randomly generated datasets. The smaller standard variance of *k*NN-SCOD indicates that it has more stable performance to detect spatial multivariate categorical outliers.

Similarly, TCOD approaches didn't achieve competitive results when applied to the spatial context, as shown by the PR(Precision-Recall) curves generated by Conditional Test, Combining Evidence, and Conditional Test-Partitioning. The performance of Multivariate Z-Lin was still much better than that of Multivariate Z-OF in the multiple dimension domain. Besides the different ways of similarity computation for a specific attribute domain, OF and Lin applied different weight values when calculating the final similarities by combining the different values from different attribute domains. OF assigns the same weight to different attributes, while Lin computes the corresponding weigh based on the category distribution in it. As shown in Eq. 25, Lin measure gives higher weight to the same categories with frequent values, and lower weight to different categories with infrequent values. Such a way could better reflect the case that if two objects having the same category co-occur with a higher frequency, they will have higher relevance. Whereas, if they co-occur frequently with different categories, their relevance score will be lower

Table 9 Average Precision for spatial categorical datasets with multiple attribute datasets, comparing *k*NN-SCOD-M and other 6 out of state approaches

Approach for multi-attribute dataset	Syn_2	Syn_3	Syn_4	Syn_5
*k*NN-SCOD-M	0.7282	0.7862	0.8003	0.7886
Conditional Test	0.0526	0.2084	0.3895	0.2509
Conditional Test-Combining Evidence	0.0526	0.2084	0.4084	0.2509
Conditional Test-Partitioning	0.0526	0.2082	0.3948	0.2468
Multivariate Z-SNOD	0.3984	0.5536	0.3596	0.4257
Multivariate Z-Lin	0.5950	0.6413	0.5498	0.4068
Multivariate Z-OF	0.0442	0.0512	0.0476	0.0512

since they are assigned a lower weight. However, such measurement to capture the relevance between pair objects with different categories is significant inconsistent with the concept of SCOs. That is why the performance of Multivariate-Lin is much worse than that of kNN-SCOD. It is worthy to note that Z-SNOD approach performs well compared with TCOD methods since it takes the characteristic of spatial auto-correlation into considerations when identifying SCOs, although it treats the categorical attributes as numerical ones.

The average precision values are given in Table 9 for all the SCOD approaches in the multi-attribute domain. Note that for most datasets,*k*NN-SCOD achieved much higher accuracy, from 0.7282 to 0.8003, than other approaches, from 0.0442 for Multivariate Z-OF, to 0.6413 for Multivariate Z-Lin, and 0.5536 for Multivariate Z-SNOD. Similarly, the performance of the methods also depends on the detection tasks. In Syn_4 and Syn_5, the contamination rate are 2 and 3, respectively, which means there exist less outliers in the whole data sets. The lower contaminated data alleviated the side-effects of outlying behaviors on the identification quality of *k*NN-SCOD-M approach. This is also demonstrated by most of the higher average precisions generated by other methods, like the Conditional Test series. For Multivariate Z series, they all performs poorer in Syn_4 compared in other datasets, since there are 4 dimensions in it. It is apt to lose information when computing the spatial relevance by integrating with Mahalanobis distance if there exist more attributes in datasets (Fig. 11).

Impacts of neighborhood sizes In the same way, we showcase the effects of neighborhood size on the performance of *k*NN-SCOD-M on multiple attribute domain. Figure 12 depicts its identification quality is not sensitive to different sizes, from 6 to 16, of spatial neighborhood. The sizes of these four data sets are 1000 and

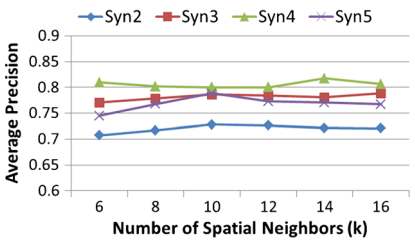


Fig. 12 Average precisions of kNN-SCOD-M by varying *k* value

4000. It is sufficient to set the k as around 8 to perform the computation of spatial relevance scores.

5.2.3 Analysis and discussion

Based on the above experimental evaluations, PCF techniques have shown to be very effective in modeling the relevances among spatial category objects in both single and multiple attribute datasets. As a result, PCF-SCOD and kNN-SCOD demonstrated superior identification qualities over the competing techniques in both simulated and real datasets. The evaluations verify two observations: 1) first, SCOs are identified in a different way with that of SNOs. Two objects taking different attributes are not necessarily irrelevant with each other. Sometimes, their frequent co-occurrence exactly illustrates their higher spatial correlation. This can be demonstrated by comparisons of Z-SNOD against PCF series of methods; 2) when identifying SCOs, the existing TCOD and SCOD approaches can't avoid the well-known swamping and masking problems. TCOD approaches treat spatial and non-spatial attributes equally and don't take the spatial dependency and spatial auto-correlation into considerations, which are the specific properties of spatial data. Z-OF and Z-Lin outperformed TCOD methods since they differentiate spatial and non-spatial attributes. However, they performed worse than PCF-SCOD and kNN-SCOD since their dissimilarity computation is based on the global frequencies, not local frequencies (Fig. 6).

Notice that there might be white noise in the original data set. Considering that data noise is usually uniformly distributed over the space, our defined pair correlation ratio is able to capture the spatial correlation as a small but nontrivial ratio value between noise observations and normal observations as long as there exist some correlation patterns between them based on their spatial distances. The pair correlation ratio will increase if the signal-noise ratio decreases. In the situation with high signal-noise ratio, noise will not be identified as outliers. In the situation with low signal-noise ratio, some noise observations may be identified as outliers, but should not be highly ranked outliers. In the extreme case where the signal-noise ratio is very high, then all the noise observations will be returned as top ranked outliers, since they are rare observations and should be regarded as outliers.

This paper assumes that the spatial locations are uniformly distributed. The case of sparsely distributed data may refer to two scenarios. The first scenario refers to the situation where the data set has a very low signal-to-noise ratio. In this case, noise will be handled well as explained above. The second scenario refers to the situation where some categorical types are rare. This still depends on the sample size of these rare categorical types is still sufficient to calculate the stable pair correlation ratios. If it is not sufficient, we may need to remove them in the pre-processing step, since we are not able to calculate stable statistics for them. However, note that sparse distribution is not common in spatial categorical data. The datasets that we collected are all not sparsely distributed.

6 Conclusion

This paper investigates the benefits of PCF technique on the SCOD, and designs three algorithms which can identify SCOs with single and multiple attributes. The

general idea in PCF-SCOD is that, first, for each pair of categories, we compute its Pair Correlation Ratios (PCR) as a function of distances. Then, the outlier scores are computed by the mean of estimated PCR values between each object and its spatial neighbors. Finally, the top l objects with higher infrequent behaviors are recognized as SCOs. We propose two kNN based estimators which utilize kNN neighborhood information to estimate the co-occurrence frequency of pair objects in single and multiple attribute domains, respectively. The proposed approaches have several advantages: (1) they can identify SCOs with both single and multiple categorical attributes; (2) they can process not only ordinal, but nominal categorical datasets; (3) compared with existing approaches, they can better avoid swamping and masking issues. The experiments conducted on the synthetic and real datasets demonstrated PCF series of approaches significantly outperformed other existing popular approaches.

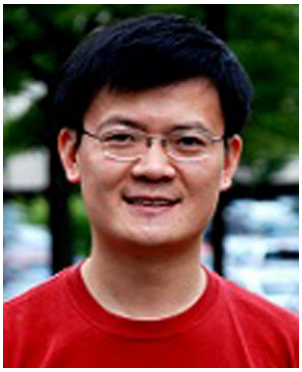
References

1. Adam NR, Janeja VP, Atluri V (2004) Neighborhood based detection of anomalies in high dimensional spatio-temporal sensor datasets. In: Proceedings of the 2004 ACM symposium on applied computing, pp 576–583
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large data bases, VLDB '94, pp 487–499
3. Anselin L (1995) Local indicators of spatial association-lisa. Geogr Anal 27(2):93–115
4. Aurenhammer F (1991) Voronoi diagrams: a survey of a fundamental geometric data structure. ACM Comput Surv 23(3):345–405
5. Bel L, Allard D, Laurent JM, Cheddadi R, Bar-Hen A (2009) Cart algorithm for spatial data: application to environmental and ecological data. Comput Statist Data Anal 53(8):3082–3093
6. Berchtold S, Ertl B, Keim DA, Kriegel HP, Seidl T (1998) Fast nearest neighbor search in high-dimensional space. In: Proceedings of the 14th international conference on data engineering, pp 209–218
7. Boriah S, Chandola V, Kumar V (2008) Similarity measures for categorical data: a comparative evaluation. In: SDM, pp 243–254
8. Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data, pp 93–104
9. Bronstein R, Das J, Duro M, Friedrich R, Kleyner G, Mueller M, Singhal S, Cohen I, Kleyner G, Mueller M, Singhal S, Cohen I (2001) Self-aware services: using bayesian networks for detecting anomalies in internet-based services. Northwestern University and Stanford University, pp 623–638
10. Chan PK, Mahoney MV, Arshad MH (2003) A machine learning approach to anomaly detection. Technical Report
11. Chandola V, Boriah S, Kuman V (2008) Understanding categorical similarity measures for outlier detection. Technical report, University of Minnesota
12. Chen D, Lu C-T, Kou Y, Chen F (2008) On detecting spatial outliers. Geoinformatica 12(4): 455–475
13. Chen F, Lu C-T, Boedihardjo AP (2010) Gls-sod: a generalized local statistical approach for spatial outlier detection. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1069–1078
14. Das K, Schneider J (2007) Detecting anomalous records in categorical datasets. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '07, pp 220–229
15. Ferhatosmanoglu H, Tuncel E, Agrawal D, Abbadi AE (2001) Approximate nearest neighbor searching in multimedia databases. In: Proceedings of the 17th international conference on data engineering. IEEE Computer Society, 2–6 Apr 2001. Heidelberg, Germany, pp 503–511

16. Goovaerts P (1997) Geostatistics for natural resources evaluation. Applied geostatistics series, Oxford University Press
17. Grekousis G, Fotis YN (2012) A fuzzy index for detecting spatiotemporal outliers. Geoinformatica 16(3):597–619
18. Haining R (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press
19. He Z, Deng S, Xu X, Huang JZ (2006) A fast greedy algorithm for outlier mining. In: Proceedings of the 10th Pacific–Asia conference on knowledge and data discovery, pp 567–576
20. He Z, Xu X, Deng S (2005) An optimization model for outlier detection in categorical data. CoRR, abs/cs/0503081
21. He Z, Xu X, Huang JZ, Deng S (2004) A frequent pattern discovery method for outlier detection. In: WAIM, pp 726–732
22. He Z, Xu X, Huang JZ, Deng S (2005) Fp-outlier: frequent pattern based outlier detection. Comput Sci Inf Syst 2(1):103–118
23. Hjaltason GR, Samet H (1998) Incremental distance join algorithms for spatial databases. In: SIGMOD conference, pp 237–248
24. Huang Y, Pei J, Xiong H (2006) Mining co-location patterns with rare events from spatial data sets. Geoinformatica 10(3):239–260
25. Huang Y, Shekhar S, Xiong H (2004) Discovering colocation patterns from spatial data sets: a general approach. IEEE Trans Knowl Data Eng 16(12):1472–1485
26. Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns. Int Stat Rev 76:458
27. Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. Analysis 6:47–66
28. Kou Y, Lu C-T, Santos RFD (2007) Spatial outlier detection: a graph-based approach. In: 19th IEEE international conference on tools with artificial intelligence, ICTAI '07, Patras, Greece, pp 281–288
29. Koufakou A, Ortiz EG, Georgiopoulos M, Anagnostopoulos GC, Reynolds KM (2007) A scalable and efficient outlier detection strategy for categorical data. In: Proceedings of the 19th IEEE international conference on tools with artificial intelligence, vol 02, ICTAI '07, pp 210–217
30. Koufakou A, Secretan J, Reeder J, Cardona K, Georgiopoulos M (2008) Fast parallel outlier detection for categorical datasets using mapreduce. In: IEEE world congress on computational intelligence (WCCI)
31. Liu X, Lu C-T, Chen F (2010) Spatial outlier detection: random walk based approaches. In: ACM SIGGIS, pp 370–379
32. Lu C-T, Chen D, Kou Y (2003) Algorithms for spatial outlier detection. In: ICDM, pp 597–600
33. Lu C-T, Chen D, Kou Y (2003) Detecting spatial outliers with multiple attributes. In: ICTAI, pp 122–128
34. Mingming NY (2000) Probabilistic networks with undirected links for anomaly detection. In: Proceedings of IEEE systems, man, and cybernetics information assurance and security workshop, pp 175–179
35. Otey ME, Ghoting A, Parthasarathy S (2006) Fast distributed outlier detection in mixed-attribute data sets. Data Min Knowl Discov 12:203–228
36. Pelleg D (2004) Scalable and practical probability density estimators for scientific anomaly detection. PhD thesis, Carnegie Mellon University
37. Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data, pp 427–438
38. Reed T, Gubbins K (1973) Applied statistical mechanics: thermodynamic and transport properties of fluids. Butterworth-Heinemann reprint series in chemical engineering. Butterworth-Heinemann
39. Shekhar S, Chawla S (2003) Spatial databases—a tour. Prentice Hall
40. Shekhar S, Chawla S, Ravada S, Fetterer A, Liu X, Lu CT (1999) Spatial databases: accomplishments and research needs. IEEE Trans Knowl Data Eng 11:45–55

41. Shekhar S, Huang Y (2001) Discovering spatial co-location patterns: a summary of results. In: Proceedings of the 7th international symposium on advances in spatial and temporal databases, SSTD '01. Springer, London, pp 236–256
42. Shekhar S, Lu C-T, Zhang P (2001) Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: KDD, pp 371–376
43. Shekhar S, Lu C-T, Zhang P, Shekhar S, Lu CT, Zhang P (2003) A unified approach to spatial outliers detection. GeoInformatica 7:139–166
44. Stanoi I, Agrawal D, Abbadi AE (2000) Reverse nearest neighbor queries for dynamic databases. In: In ACM SIGMOD workshop on research issues in data mining and knowledge discovery, pp 44–53
45. Sun P, Chawla S (2004) On local spatial outliers. In: IEEE international conference on data mining, pp 209–216
46. Tobler WR (1979) Cellular geography, pp 379–389. Reidel, Dordrecht, Netherlands
47. Wong W-K, Moore A, Cooper G, Wagner M (2002) Rule-based anomaly pattern detection for detecting disease outbreaks. In: Eighteenth national conference on Artificial intelligence, pp 217–223
48. Yoo JS, Shekhar S (2006) A joinless approach for mining spatial colocation patterns. IEEE Trans Knowl Data Eng 18(10):1323–1337
49. Zhao J, Lu C-T, Kou Y (2003) Detecting region outliers in meteorological data. In: Proceedings of the 11th ACM international symposium on advances in geographic information systems, pp 49–55

Xutong Liu received the ME degree in computer science from Jinan University, GuangZhou, China in 2006 and the Ph.D. degree in computer science from Virginia Polytechnic Institute and State University in 2013. She is an applied researcher at ebay. Her research interests include machine learning, data mining and information retrieval, with an emphasis on prediction and anomaly detection.



Feng Chen is a postdoctoral research fellow at Carnegie Mellon University. He received his B.S. from Hunan University, China, in 2001, M.S. degree from Beihang University, China, in 2004, and Ph.D. degree from Virginia Polytechnic Institute and State University in 2012, all in Computer Science. He has published 25 refereed articles in major data mining venues, including ACM-SIGKDD, ACM-CIKM, ACM-GIS, IEEE-ICDM, and IEEE-INFOCOM. He holds two U.S. patents on human activity analysis filed by IBM's T.J. Watson Research Center. His research interests are in the areas of statistical machine learning and data mining, with an emphasis on spatio-temporal analysis, social media analysis, and energy disaggregation.

Chang-Tien Lu received the MS degree in computer science from the Georgia Institute of Technology in 1996 and the PhD degree in computer science from the University of Minnesota in 2001. He is an associate professor in the Department of Computer Science, Virginia Polytechnic Institute and State University and is the founding director of the Spatial Lab. He served as Program Co-Chair of the 18th IEEE International Conference on Tools with Artificial Intelligence in 2006, and General Co-Chair of the 20th IEEE International Conference on Tools with Artificial Intelligence in 2008 and 17th ACM International Conference on Advances in Geographic Information Systems in 2009. He is also serving as Vice Chair of the ACM Special Interest Group on Spatial Information (ACM SIGSPATIAL). His research interests include spatial databases, data mining, geographic information systems, and intelligent transportation systems.