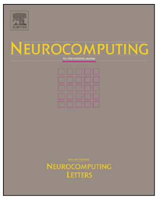
Neurocomputing 117 (2013) 161–172

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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



Spatial outlier detection based on iterative self-organizing learning model

Qiao Cai [a], Haibo He [b],*, Hong Man [a]

[a] Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[b] Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form
23 January 2013
Accepted 1 February 2013
Communicated by K. Li
Available online 16 March 2013

Keywords:
Self-organizing map
Spatial data mining
Spatial outlier detection
Iterative learning
Robust distance

ABSTRACT

In this paper, we propose an iterative self-organizing map (SOM) approach with robust distance estimation (ISOMRD) for spatial outlier detection. Generally speaking, spatial outliers are irregular data instances which have significantly distinct non-spatial attribute values compared to their spatial neighbors. In our proposed approach, we adopt SOM to preserve the intrinsic topological and metric relationships of the data distribution to seek reasonable spatial clusters for outlier detection. The proposed iterative learning process with robust distance estimation can address the high dimensional problems of spatial attributes and accurately detect spatial outliers with irregular features. To verify the efficiency and robustness of our proposed algorithm, comparative study of ISOMRD and several existing approaches are presented in detail. Specifically, we test the performance of our method based on four real-world spatial datasets. Various simulation results demonstrate the effectiveness of the proposed approach.

1. Introduction

With the continuous explosive increase of data availability in many real-world applications, computational intelligence techniques have demonstrated great potential and capability to analyze such data and support decision-making process. In general, there are five primary categories of data engineering research, including classification, clustering, regression, association, and deviation or outlier detection [1]. In this paper, our objective is to investigate the spatial outlier detection based on computational intelligence approaches.

The procedure of outlier detection can be considered as similar to the discovering of "nuggets of information" [2] in large databases. The motivation for this type of research is that in many practical situations, such outliers normally carry the most critical information to support the decision making process. Due to the wide range of application scenarios for outlier detection across different domains [19–21], such as financial industry, biomedical engineering, security and defense, to name a few, outlier detection has been an important research topic in the community for many years. For instance, an anomalous traffic pattern in a computer network might indicate the presence of malicious intrusion from unauthorized users or computers. In public health information systems, outlier detection techniques are widely employed to detect abnormal patterns in physical records that might indicate uncommon symptoms. In all of such situations, once the outliers are identified, they prompt a more focused human analysis to understand those data sets from a vast amount of original raw data.

We would like to point out that, due to the extensive research efforts in the community, there are different terminologies referring to the same or similar idea, such as outlier or anomaly detection [3,4], exception mining [5], mining rare classes [6], novelty detection [7], and chance discovery [8]. Data mining techniques concerning this issue involve both supervised and unsupervised learning paradigms. Generally speaking, supervised learning methods first establish a prediction model for regular and irregular events based on labeled data in the training set, and then make classifications for future test data. One of the shortcomings of such approaches is that they require a representative set of training data with the target function to train the model. Such labeled training data might be difficult or expensive to obtain in many real applications. Unsupervised learning, on the other hand, does not require labeled data. The performance of such approaches depends on the choice of feature selection, similarity measures, and clustering methods.

In this paper, we propose to use self-organizing map (SOM) with robust distance estimation for spatial outlier detection research in [27,28]. Although SOM was proposed a long time ago and there is a rich literature on SOM and related techniques in the community, the use of SOM specifically targeting for spatial outlier detection is a relatively new topic. With the continuous expansion of data availability in many of today's data intensive applications (the Big Data Challenge [45]), we consider the analyze of spatial

* Corresponding author.
E-mail address: he@ele.uri.edu (H. He).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.02.007

data has become more and more critical in many real world applications. Therefore, we hope the proposed SOM-based spatial outlier detection method in this work could provide important techniques and solutions to tackle the spatial data challenge. The major motivation for this approach is to take advantage of the data clustering capability of SOM to effectively detect outliers with both spatial and non-spatial features. Furthermore, to improve the learning and detection performance, we propose an iterative SOM approach with robust distance estimation for improved performance. The rest of this paper is organized as follows. Section 2 briefly introduces the development of related research on this topic. In Section 3, we present our proposed approach for spatial outlier detection. In Section 4, the detailed simulation results and analysis of our method are presented based on the U. S. Census Bureau databases for spatial outlier detection. Finally, we give a conclusion in Section 5.

2. Related work

In general, there are five major categories of approaches for outlier detection in the literature: distribution-based, clustering-based, distance-based, density-based, and depth-based methods. Distribution-based approaches are primarily concentrated on the standard statistical distribution models. Some representative distribution models like Gaussian or Poisson are frequently used to identify outliers that perform irregularly in such models [4]. In clustering-based approaches, the identification of outliers is normally considered as a side product while the primary goal of clustering is to find data cluster distributions [10]. However, these approaches have been successful in many applications, such as the CLARANS [11], DBSCAN [12], and CURE [13] approaches. Distance-based approaches rely on different distance metrics to measure the relationships between data items and to find outliers [14]. Some interesting methods have the capability of calculating full dimensional mutual distances with existing attributes [15,16] or feature space projections [3]. Density-based approaches are based on the analysis of data distribution density, such as the approach in [17], to determine a local outlier factor (LOF) for each data sample based on its corresponding local neighborhood density. In this way, those data samples with higher LOF can be considered as outliers. Finally, depth-based approaches can identify outliers based on geometric computation, which computes distinct layers of k-dimensional convex hulls [18].

An essential characteristic of spatial data analysis is that it involves both spatial attributes such as longitude, latitude and altitude, and associated non-spatial attributes, such as the population density and age distribution of each spatial point. Meanwhile, spatial data appear to be highly correlated. For example, spatial objects with the similar properties seem to cluster together in the neighboring regions. In fact, as discussed in [24], spatial autocorrelation problems involved in spatial dependency occur for all spatial objects when spatial properties are involved. The spatial relationships among the items in spatial datasets are established through a contiguity matrix, which may indicate neighboring relationships, such as vicinity or distance. Given such characteristic of spatial data mining, detection of spatial outliers aims at discovering specific data instances whose non-spatial attribute values are significantly distinct from the corresponding spatial neighbors. Informally speaking, a spatial outlier might be considered as a local instability whose non-spatial attributes are intrinsically relevant to the surrounding items, although they may be obviously distinct from the entire population. There are two major categories of outliers in spatial datasets: multi-dimensional space-based outliers and graph-based outliers [25]. The major difference between them is their spatial neighborhood definitions. Multi-dimensional space-based outliers are based on Euclidean distances, while graph-based outliers follow graph connectivity.

Most of the existing spatial outlier detection algorithms focus on identifying single attribute outliers, and could potentially misclassify normal items as outliers when genuine spatial outliers exist in their neighborhoods with extremely large or small attribute values. In addition, many practical applications involve multiple non-spatial attributes which should be incorporated into outlier detection. There are several reasons that spatial outlier detection still remains a great challenge. First of all, the definition of neighborhood is crucial to the determination of spatial outliers. Additionally, statistical approaches are required to characterize the distributions of the attribute values at various locations compared with the aggregate distributions of attribute values over all the neighboring data.

Several different approaches have been used to improve the classical definition of outlier by Hawkins [30]. Knorr and Ng [31] presented the concept of distance-based outliers for multi-dimensional datasets. Another approach for identifying distance-based outlier is to calculate the distance between certain point and its corresponding k nearest neighbors [33]. The ranked points are identified as outlier candidates based on the distance to its k nearest neighbors. In some specific models, local outliers appear to be more important than global outliers [34]. However, they are normally difficult to be identified using general distance-based techniques. The method based on a local outlier factor (LOF) was proposed to capture the phenomena that a sample is isolated from its surrounding neighborhood rather than the whole dataset. The local correlation integral (LOCI) method was also presented to discover local outliers [35]. This approach seems moderately similar to the LOF except for the definition of the local neighborhood. However, spatial attributes are not considered in these algorithms or approaches for outlier detection.

In spatial datasets, the non-spatial dimensions provide intrinsic properties of each data example, while spatial attributes describe location indices to define neighborhood boundary for spatial outlier detection. Thus, the physical neighborhood plays a crucial role in spatial data analysis. Common techniques for neighborhood characterization include KNN, grid technique, and grid-based KNN. Additionally, a variogram-cloud [22] displays spatial objects based on the neighboring relationships. For each pair of location coordinates, the square-root of the absolute difference between attribute values at the locations compared with the mutual Euclidean distance is depicted. Vicinal locations with significant attribute difference might be considered as a spatial outlier, even though attribute values at such locations may seem to be regular or normal if spatial attributes are ignored.

In our research, we consider a spatial outlier as a "spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood" [21]. The spatial outlier detection in the graph dataset was discussed with detailed algorithms in [21,37]. Two representative methods, i.e. Scatterplot [38] and Moran scatterplot [39], can be employed to quantitatively analyze spatial dataset and discover spatial outliers. A scatterplot illustrates attribute values on the X-axis and the average attribute values in the neighborhood on the Y-axis. A least-square regression line is used to identify spatial outliers. A positive spatial autocorrelation is suggested by the right upward sign for a scatter slope; otherwise, it turns to be negative. The purpose of a Moran scatterplot is to compare the normalized attribute values with the average normalized attribute values in the neighborhoods. In [40], several statistical outlier detection techniques were proposed, and they were compared with four algorithms. The z, iterative z, iterative r, and Median algorithms [36] were successfully used to identify spatial outliers. A measure for spatial local outliers was proposed in [25], which relates spatial autocorrelation or dependency with non-uniform distributed variance of spatial dataset. However, these methods failed to detect spatial outliers with multiple attributes. Recent work in [41] discussed the statistical method to identify spatial outliers with multiple non-spatial attributes. The spatial attributes were used to search for neighboring data instances, while the non-spatial attributes were used to determine outlier candidates based on Mahalanobis distance with chi-squared distributional cut-off threshold [21,37]. However, it ignored the situation when the spatial attributes became too complex to search for the similar properties. For instance, the approach uses a statistical test that is useful for discovering global outliers but may fail to identify local outliers. Additionally, normal distribution fitting might be unavailable in arbitrary spatial datasets in the real world applications. Furthermore, the previous methods can hardly provide robust percentile cut-offs for spatial outlier detection.

3. Integration of SOM with robust distance for spatial outlier detection

SOM [26] can be considered as a special class of neural networks based on competitive learning. The main goal of SOM is to project the input vector with higher dimensions into one or two dimensional discrete map in topologically ordered pattern. In this paper, we propose to use SOM with robust distance estimation for effective spatial outlier detection motivated by our previous research results [27,28].

Briefly speaking, the learning in SOM involves three stages: competitive phase, cooperative phase and adaptive phase. Suppose x_i is the i_{th} input vector, which is randomly selected from input space X. Here w_j represents the synaptic weight vector for the j_{th} neuron on the lattice with identical dimension as input pattern. Then the best matching unit (BMU) b_i for x_i is determined by minimizing the Euclidean distance between x_i and w_j

$$b_i = \arg\min_j \|x_i - w_j\| \tag{1}$$

The winning neuron is centered on the topological neighborhood of cooperative neurons. The lateral interaction forms the similarity between winning neuron and synaptic neuron. In Eq. (2), we select Gaussian kernel function $\Phi_{j,i}(n)$ to specify topological neighborhood for BMU b_i

$$\Phi_{j,i}(n) = \exp\left(-\frac{\|r_i - r_j\|^2}{2\sigma_0^2 \exp\left(-\frac{2n}{\tau_1}\right)}\right) \tag{2}$$

where n is the current epoch; r_j is the position index vector of jth neuron on the 2-D lattice; σ_0 is the initial width of the topological neighborhood; τ_1 is the time constant in the cooperative learning.

The feature map is self-organized by updating the synaptic weight vector of excited neuron in the network. At the discrete time $n+1$, the updated synaptic weight vector $w_j(n+1)$ is shown in Eq. (3).

$$\begin{aligned} w_j(n+1) &= w_j(n) + \eta(n)\Phi_{j,i}(n)(x_i(n) - w_j(n)) \\ &= (1 - \eta(n)\Phi_{j,i}(n))w_j(n) + \eta(n)\Phi_{j,i}(n)x_i(n) \end{aligned} \tag{3}$$

where $\eta(n)$ represents the learning rate function.

3.1. SOM with robust distance for spatial outlier detection

The proposed algorithm can effectively detect spatial outliers with multiple spatial and non-spatial attributes. SOM can explicitly locate the spatial similarities or relationships for neighboring clusters in high-dimensional space. Due to multiple attributes, the concept of minimum covariance determinant for robust distance is used in our method to determine the threshold for identifying spatial outliers. Iterative SOMRD (ISOMRD) approach is proposed based on the iterative strategy for self-organizing learning. The advantage of this method is that it can eliminate the influence of surrounding data caused by local outliers in the same cluster. The following highlights the proposed approach, with detailed discussions in the following sections.

Definition 1. The spatial attributes can be considered as two groups: explicit and implicit spatial attributes. The explicit spatial attributes include location, shape, size, orientation and relationship. For example, if there is a square object, we can define:

Location: The diagonal intersection.

Shape: Square.

Size: The length of each side.

Orientation: The angle between the diagonals and reference axis.

The implicit spatial attributes can be represented by spatial relationships. The typical case is the neighborhood relationship. The implicit spatial attributes depend on the way that specifies explicit spatial attributes.

Definition 2. Non-spatial attributes generally reflect independent information of spatial attributes. For example, if the target object is the building, the non-spatial attributes include material, color, date, style. However, some particular non-spatial attribute of the building is indirectly related with spatial attribute, e.g. the price of building at downtown area. Therefore, the relation between spatial attribute and particular non-spatial attribute cannot be ignored.

Definition 3. Suppose the target data set $O:\{o_1,\ldots,o_n\}$, the spatial attribute function $S: s_i \leftarrow S(o_i)$, the neighborhood function $G: G(s_i) = (1/|N_{s_i}|)\sum_{k=1}^{|N_{s_i}|} n_k$, where the neighborhood $N_{s_i}:\{n_1,\ldots,n_K\}$, K is the cardinality of the set N_{s_i}, non-spatial attribute function $A: a_i \leftarrow A(o_i)$. To minimize data redundancy, the normalization of non-spatial attributes can be implemented through $a_i \leftarrow (A(o_i) - \mu_A)/\sigma_A$, where μ_A and σ_A^2 are the mean and variance of non-spatial data distribution, respectively. The normalization essentially fits the data into Gaussian distribution for outlier data analysis. The comparison function measures the difference between normalized non-spatial attribute and its corresponding neighborhood function as $H: H(o_i) = a_i - G(s_i)$.

Definition 4. Let $F_{q,m-q+1}(\beta)$ denote F distribution with certain confidence level β. The threshold of outlier detection $Th = \lambda F_{q,m-q+1}(\beta)$, λ can be estimated by minimum covariance determinant (MCD). Based on MCD estimator, we can obtain the robust distance $RD:\{rd_i \leftarrow MCD(o_i)\}$.

Algorithm 1. ISOMRD Algorithm.

```
 Input:
 (1) Spatial dataset: $O:\{o_1,\ldots,o_n\}$, where $n$ is the number of
 input data
 (2) Total neuron number: $N_m$
 (3) Maximum iteration: $max\_iter$
$a_i \leftarrow \frac{A(o_i)-\mu_A}{\sigma_A}$
 repeat
   for $i=1$ to $max\_iter$ do
     Search BMU $b(a_i)$ via Eq. (1)
     for $j=1$ to $N_m$ do
       Update $w_j(i+1)$ via Eq. (3)
     end for
   end for
```

```
    Calculate the neighborhood function $G(s_i)=\frac{1}{|N_{s_i}|}\sum_{k=1}^{|N_{s_i}|} n_k$
    Calculate the comparison function $H(o_i)=a_i-G(s_i)$
    Calculate the robust distance $RD$ through $MCD$ estimator in Eq. (34)
    Select the input data with largest $RD$ and remove it from input dataset
until The $m_{th}$ outlier candidate is obtained
for $k=1$ to $m$ do
    if $rd_k^2 > Th$ then
        $o_k$ can be identified as a spatial outlier
    end if
end for
```

Computational complexity: The computational complexity of self-organizing learning procedure in ISOMRD algorithm can be calculated as follows:

For each iteration, we divide the learning procedure into three phases.

Phase 1: The computation of Euclidean distance between x_i and w_j costs $O(MN^2)$.

Phase 2: The computation for ordering Euclidean distance vector and selecting the BMU costs $O(MN \log M)$.

Phase 3: The computation for updating neuron weight via Gaussian kernel function in Eq. (3) costs $O(MN^2)$.

where N is the input data, M is the neuron number on the lattice.

Suppose that the number of iterations is I, the total computational complexity of self-organizing learning is $O(IMN^2)+O(IMN \log M)+O(IMN^2)$.

Assuming that $N \gg M$, $N \gg I$, the final computational complexity is $O(N^2)$.

Convergence analysis: In Eq. (3), we analyze the adaptive learning procedure from neuron weight vector $w_j(n)$ and input data vector $x_i(n)$. Obviously, the factors of learning rate function $\eta(n)$ and Gaussian kernel function $\Phi_{j,i}(n)$ affect the learning speed. The sum of coefficients for $w_j(n)$ and $x_j(n)$ is equal to 1, which indicates the implicit relationships between the two vectors. The learning rate function is specified as Eq. (4).

$$\eta(n)=\eta_0 \exp\left(-\frac{n}{\tau_2}\right) \tag{4}$$

where η_0 represents the initial learning rate, whereas τ_2 is the time constant in the adaptive learning.

From Eqs. (2)–(4), we can obtain

$$w_j(n+1)=(1-\eta_0 \exp)\left(-\frac{n}{\tau_2}-\frac{\|r_i-r_j\|^2}{2\sigma_0^2 \exp\left(-\frac{2n}{\tau_1}\right)}\right)w_j(n)$$
$$+\eta_0 \exp\left(-\frac{n}{\tau_2}-\frac{\|r_i-r_j\|^2}{2\sigma_0^2 \exp\left(-\frac{2n}{\tau_1}\right)}\right)x_i(n) \tag{5}$$

Suppose that

$$U=-\frac{n}{\tau_2}-\frac{\|r_i-r_j\|^2}{2\sigma_0^2 \exp\left(-\frac{2n}{\tau_1}\right)}=-\frac{n}{\tau_2}-\frac{\|r_i-r_j\|^2}{2\sigma_0^2}\exp\left(\frac{2n}{\tau_1}\right) \tag{6}$$

We can transform Eqs. (5) into (7)

$$w_j(n+1)=(1-\eta_0 \exp(U))w_j(n)+\eta_0 \exp(U)x_i(n) \tag{7}$$

Then

$$\lim_{n\to+\infty} U=\lim_{n\to+\infty}\left(-\frac{n}{\tau_2}-\frac{\|r_i-r_j\|^2}{2\sigma_0^2}\exp\left(\frac{2n}{\tau_1}\right)\right)$$
$$=-\left(\frac{1}{\tau_2}\right)\lim_{n\to+\infty} n+\frac{\|r_i-r_j\|^2}{2\sigma_0^2}\lim_{n\to+\infty}\exp\left(\frac{2n}{\tau_1}\right) \tag{8}$$

$$\lim_{n\to c}\exp(x)=\exp(c)=\begin{cases}+\infty, & c=+\infty\\ 1, & c=0\\ 0, & c=-\infty\end{cases} \tag{9}$$

where c is a specific constant.

From Eq. (9), we can know

$$\lim_{n\to+\infty} n\to+\infty$$

and

$$\lim_{n\to+\infty}\exp\left(\frac{2n}{\tau_1}\right)\to+\infty$$

Then

$$\lim_{n\to+\infty} U\to-\infty \tag{10}$$

According to Eqs. (9) and (10),

$$\lim_{n\to+\infty}\exp(U)\to\exp(-\infty)=0 \tag{11}$$

Since

$$\lim_{n\to+\infty} x_i(n)=x_i \tag{12}$$

where x_i is a constant vector during the adaptive learning for the i_{th} input data.

Suppose that

$$\lim_{n\to+\infty} w_j(n)=C \tag{13}$$

where C is a constant vector.

According to Limit Theorems, we can obtain Eq. (14)

$$\begin{aligned}\lim_{n\to+\infty} w_j(n+1)&=\lim_{n\to+\infty}(1-\eta_0 \exp(U))w_j(n)+\lim_{n\to+\infty}\eta_0 \exp(U)x_i(n)\\ &=(1-\eta_0 \exp(U))\lim_{n\to+\infty} w_j(n)+\lim_{n\to+\infty}\eta_0 \exp(U)\lim_{n\to+\infty} x_i(n)\\ &=(1-\eta_0 \exp(U))\lim_{n\to+\infty} w_j(n)\\ &=\lim_{n\to+\infty} w_j(n)-\eta_0 \exp(U)\lim_{n\to+\infty} w_j(n)\\ &=\lim_{n\to+\infty} w_j(n)\\ &=C\end{aligned} \tag{14}$$

From Eq. (14), we can verify the assumption in Eq. (13). The self-organizing learning procedure is convergent.

3.2. Optimization of kernel density estimation

Suppose that $x_1,x_2,\ldots,x_n$ are a sequence of independent and identically distributed (IID) random variables sampled from a common distribution specified by the density function $f_\sigma(x)$. The kernel density estimate can be calculated as Eq. (15)

$$\hat{f}_\sigma(x)=\frac{1}{n}\sum_{i=1}^{n}\Phi_\sigma(x-x_i) \tag{15}$$

where Φ is kernel function, σ is kernel bandwidth.

The kernel bandwidth selection has a significant effect on kernel density estimation. To optimize the kernel density estimation, we adopt the mean integrated squared error (MISE) to determine the optimal kernel bandwidth.

$$MISE(\sigma)=E\left[\int(\hat{f}_\sigma(x)-f_\sigma(x))^2\,dx\right] \tag{16}$$

Based on the weak assumptions on the kernel function and density function [44], we can know Eq. (17)

$$MISE(\sigma) = AMISE(\sigma) + o(1/(n\sigma)) + \sigma^4 \tag{17}$$

where o is little-o notation, AMISE is asymptotic MISE calculated by Eq. (18)

$$AMISE(\sigma) = \frac{T(\Phi)}{n\sigma} + \frac{1}{4}M^2(\Phi)\sigma^4 T(f''_\sigma) \tag{18}$$

where $T(\Phi) = \int \Phi^2(x)\,dx$, $M(\Phi) = \int x^2\Phi(x)\,dx$.

To optimize the AMISE, we can obtain Eq. (19)

$$\frac{\partial}{\partial\sigma}AMISE(\sigma) = -\frac{T(\Phi)}{n\sigma^2} + M^2(\Phi)\sigma^3 T(f''_\sigma) = 0 \tag{19}$$

From Eq. (19), the optimal kernel bandwidth σ^*

$$\sigma^* = \frac{[T(\Phi)]^{1/5}}{n^{1/5}[M(\Phi)]^{2/5}[T(f''_\sigma)]^{1/5}} \tag{20}$$

Since

$$\frac{\partial^2}{\partial\sigma^2}AMISE(\sigma)|_{\sigma=\sigma^*} = \frac{2T(\Phi)}{n(\sigma^*)^3} + 3M^2(\Phi)(\sigma^*)^2 T(f''_\sigma) > 0 \tag{21}$$

We can know

$$\arg\min_{\sigma^*\in\sigma} AISE(\sigma) = \sigma^* = \frac{[T(\Phi)]^{1/5}}{n^{1/5}[M(\Phi)]^{2/5}[T(f''_\sigma)]^{1/5}} \tag{22}$$

Considering the Gaussian kernel in the self-organizing learning model shown in Eq. (2), we can derive σ^* in Eq. (23).

$$\sigma^* = \left(\frac{4s^5}{3n}\right)^{1/5} = \left(\frac{4}{3n}\right)^{1/5} s \tag{23}$$

where s is sample standard deviation as Eq. (24)

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{24}$$

where $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$

When the lattice achieves the convergence state with the maximum training iteration number I_{max},

$$\sigma(n)|_{n=I_{max}} = \sigma_0 \exp\left(-\frac{I_{max}}{\tau_1}\right) \tag{25}$$

Suppose

$$\sigma(n)|_{n=I_{max}} = \sigma^* \tag{26}$$

If Eq. (26) is satisfied, we can obtain

$$\sigma_0 = s\left(\frac{4}{3n}\right)^{1/5} \exp\left(\frac{I_{max}}{\tau_1}\right) \tag{27}$$

To ensure the upper bound of initial kernel bandwidth, it should be satisfied as

$$\sigma_0 \le (\tfrac{1}{2}M)^{1/2} \tag{28}$$

The estimate neuron number M is

$$M \ge \lceil 2\sigma_0^2 \rceil = \left\lceil 2\left(s\left(\frac{4}{3n}\right)^{1/5} \exp\left(\frac{I_{max}}{\tau_1}\right)\right)^2 \right\rceil \tag{29}$$

The neuron number for self-organizing learning model can be estimated by Eq. (30)

$$\begin{aligned}\min_{\sigma^* = \sigma(n)|_{n=I_{max}}} M &= \left\lceil 2\left(s\left(\frac{4}{3n}\right)^{1/5} \exp\left(\frac{I_{max}}{\tau_1}\right)\right)^2 \right\rceil \\ &= \left\lceil 2s^2\left(\frac{4}{3n}\right)^{2/5} \exp\left(\frac{2I_{max}}{\tau_1}\right) \right\rceil \end{aligned} \tag{30}$$

3.3. Mahalanobis distance

As one of the important statistical tools in machine learning and data mining, Mahalanobis distance [43] is an efficient measure for identifying and analyzing various patterns. This can also be used to discover the similar properties from an unknown sample dataset compared with the normal data samples. This distance metric differs from the Euclidean distance in that data correlations or dependency is considered. Mahalanobis distance is generally derived from a group of items with specific mean and covariance for the given multivariate data vectors. It can also be summarized as a distinction measurement between two random vectors from the identical statistical distribution by covariance matrix

$$MD_i = \sqrt{(x_i - \overline{x})^T S^{-1}(x_i - \overline{x})} \tag{31}$$

where $\overline{x}$ denotes the mean of samples and s is the covariance matrix as the location and scatter estimates.

3.4. Robust distance based on minimum covariance determinant

In Mahalanobis distance, the shape matrix derived from the consistent multivariate shape and location estimators might have the asymptotic characters of the chi-squared distribution based on classical covariance matrix. Typically, minimum covariance determinant (MCD) [32] can be considered as a robust and high-breakdown estimator [33]. However, in a very large database, the performance of the chi-squared approximation for MCD approach can hardly satisfy the desired requirement, even though the statistical models are obtained. It is difficult to determine exact cut-off threshold values for the given dataset. The modification of the distance metric with F distribution can be considered as a robust distance to detect accurate spatial outlier candidates for various sizes of sample data. The comparison between the χ^2 and F distributions indicates two different cut-off thresholds and estimates, which might lead to divergent results for predicting outlier candidates. Generally speaking, the F distribution is more representative than the χ^2 distribution to deal with extreme data.

In MCD location and shape estimates, the mean and covariance matrix can be calculated to minimize the determinant of covariance matrix for d points sampled from overall n data points ($d \le n$). The maximum breakdown value of d depends on the total number of data points n and the attribute dimensions q

$$\overline{\mu}^*_{MCD} = \frac{1}{d}\sum_{i\in G} x_i \tag{32}$$

$$\Sigma^*_{MCD} = \frac{1}{d}\sum_{i\in G}(x_i - \overline{\mu}^*_{MCD})(x_i - \overline{\mu}^*_{MCD})^T \tag{33}$$

$$RD_i = \sqrt{(x_i - \overline{\mu}^*_{MCD})^T \Sigma^*_{MCD}(x_i - \overline{\mu}^*_{MCD})} \tag{34}$$

where $\overline{\mu}^*_{MCD}$ and Σ^*_{MCD} denote robust location and shape estimates, respectively; d is the greatest integer of $(n+q+1)/2$ referred as “half sample” and G represents the set of m sample points derived from n data points.

Based on location and shape estimates, the robust distance based on MCD can be expressed as Eq. (34). The following work is required to calculate the estimation of the degrees of freedom concerned with statistical distribution in order to estimate the robust cut-off threshold values. As discussed in [33], the robust distance can be estimated by

$$\frac{c(m-q-1)}{qm} RD^2 \propto F_{q,m-q-1}(\beta) \tag{35}$$

Here two parameters c and m are still unknown for the distribution of robust distance. Based on the asymptotic method, c can be determined, even if there are very small samples

$$c=\frac{n}{d}P(\chi^2_{q+2}<\chi^2_{q,d/n}) \tag{36}$$

where χ^2 is a chi-squared random variable; q denotes degree of freedom and the ratio d/n means cut-off points for χ^2

$$\alpha=\frac{n-d}{n} \tag{37}$$

TH_α can be obtained from Eq. (38).

$$1-\alpha=P(\chi^2_q\le TH_\alpha) \tag{38}$$

In order to estimate the other parameter m for asymptotic degrees of freedom, the following statistical formulas can be obtained based on the analysis in [34]:

$$c_\alpha=\frac{1-\alpha}{P(\chi^2_{q+2}\le TH_\alpha)} \tag{39}$$

$$c_2=\frac{-P(\chi^2_{q+2}\le TH_\alpha)}{2} \tag{40}$$

$$c_3=\frac{-P(\chi^2_{q+4}\le TH_\alpha)}{2} \tag{41}$$

$$c_4=3c_3 \tag{42}$$

$$b_1=\frac{c_\alpha(c_3-c_4)}{1-\alpha} \tag{43}$$

$$b_2=0.5+\frac{c_\alpha}{1-\alpha}\left(c_3-\frac{TH_\alpha}{q}\left(c_2+\frac{1-\alpha}{2}\right)\right) \tag{44}$$

The parameter m can be estimated by Eq. (48). With the parameters m and c, the outlier candidates might be identified based on the predetermined cut-off percentiles by robust distance using MCD approach.

$$v_1=(1-\alpha)b_1^2\left(\alpha\left(\frac{c_\alpha TH_\alpha}{q}-1\right)^2-1\right)$$
$$-2c_3c_\alpha^2(3(b_1-qb_2)^2+(q+2)b_2(2b_1-qb_2)) \tag{45}$$

$$v_2=n(b_1(b_1-qb_2)(1-\alpha))^2c_\alpha^2 \tag{46}$$

$$v=\frac{v_1}{v_2} \tag{47}$$

$$m=\frac{2}{c_\alpha^2 v} \tag{48}$$

4. Simulation results and analysis

4.1. Case study for spatial analysis

The principle of computational intelligence in SOM is to promote componential neurons of the network to seek similar properties for certain input patterns, which is motivated by the observation that different parts of the cerebral cortex in the human brain are responsible for processing complex visual, auditory or other sensory information. Therefore, SOM is an effective method to preserve intrinsic topological and metric relationships in large datasets for visualization of high dimensional data, which provide unique advantages when it is used to solve spatial data mining problems. Spatial data mining aims to extract implicit, novel and interesting patterns from large spatial databases [42].

To make spatial analysis based on the U.S. Census dataset [29], Fig. 1 shows the learning procedure of a typical SOM. Fig. 1 (a) shows the spatial attributes of the dataset, where the horizontal and vertical coordinates represent longitude and latitude of spatially referred objects, respectively. In Fig. 1(b), the feature map is initially obtained. The adaptive phase can be divided into two

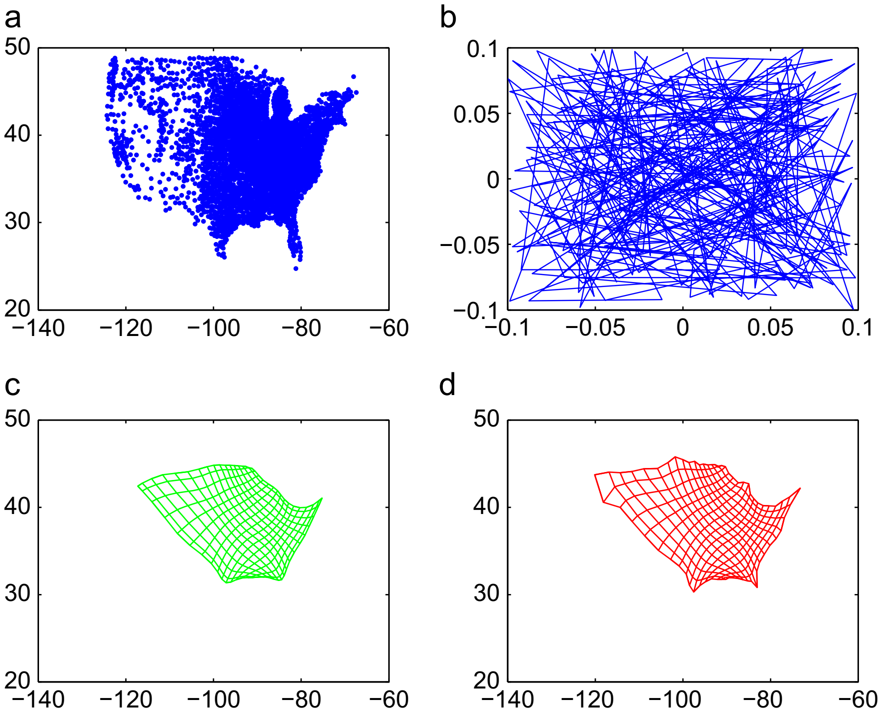


Fig. 1. The concrete procedure of self-organizing learning based on spatial attributes. (a) Spatial distribution of input data. (b) Initial condition of 2-D lattice. (c) The ordering phase. (d) The convergence phase.

steps: ordering and convergence as illustrated in Fig. 1(c) and (d), respectively. Therefore, SOM is ultimately generated to reserve explicit topological information and search similar relationships for the input data.

We also adopt MISE in Eq. (16) to optimize Gaussian kernel bandwidth for the experimental data collected from the duration in minutes for eruptions of the Old Faithful geyser in Yellowstone National Park with totally 107 data samples. From Fig. 2, we can see clearly the optimal kernel bandwidth $\sigma^{*}\approx 8$ at the red point with the minimum of $MISE = 0.5176$.

4.2. Analysis of spatial outlier detection for single non-spatial attribute dataset

To illustrate the effective model of ISOMRD, we first give an example with 40 artificial data instances to demonstrate the robustness and efficiency of searching reasonable clustering or neighborhood set based on spatial attributes. Compared with several typical algorithms in the previous works, ISOMRD can accurately detect spatial outlier regardless of the serious influence arising from some potential falsely detected outliers, which might be frequently ignored by other approaches, such as Z algorithm [37], Scatterplot [38] and Moran scatterplot [39].

In Table 1, each data item represents object with spatial attribute in *X*–*Y* coordinate plane, while *Z* coordinate means non-spatial attribute value. From this table, S1, S2, S3 and S4 are truly spatial outliers, while E1, E2 and E3 might be considered as falsely detected outliers. Fig. 3 visualizes these 40 data instances for clear presentation.

ISOMRD can be applied to detect potential spatial outliers without error data items such as E1 or E2. In Fig. 4, each synaptic neuron in the lattice means an independently self-organizing cluster or neighborhood. After ordering and convergence phase, the SOM model mapping to spatial distribution can capture the reasonable matching objects according to neuron's weight. Generally speaking, SOM neurons are representative for data clustering to seek the similar spatial relationship, which can be used to define neighborhood set to facilitate spatial outlier detection.

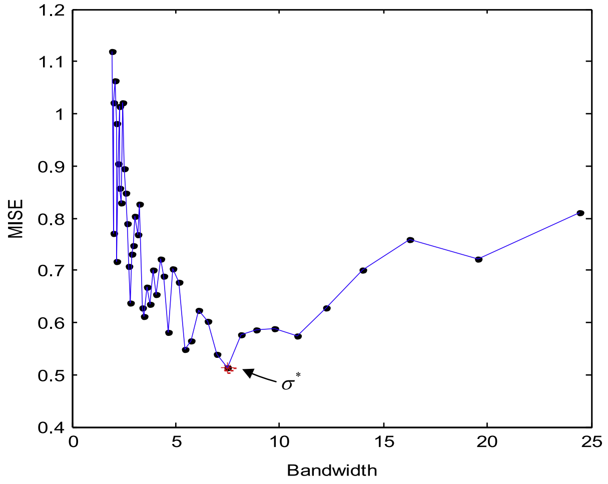


Fig. 2. The kernel bandwidth optimization.

Table 1
Coordinates of potential spatial outliers.

Data	Spatial attribute-(*x*, *y*)	Non-spatial attribute-*z*
S1	(40, 40)	200
S2	(60, 40)	104
S3	(30, 90)	20
S4	(100, 30)	10
E1	(30, 50)	14
E2	(30, 40)	11
E3	(30, 30)	19

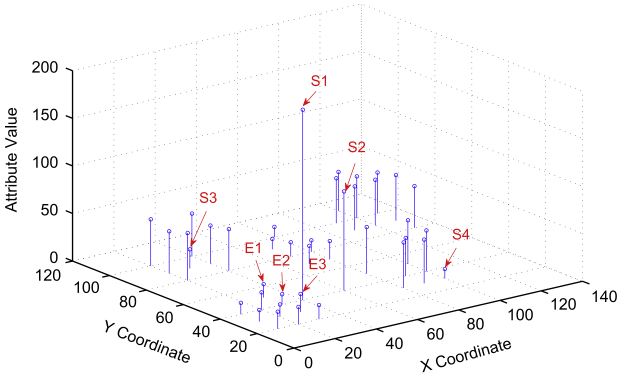


Fig. 3. Data visualization of spatial and non-spatial attributes.

The simulation results from this example based on all these algorithms can be summarized in Table 2. Comparison of top 4 spatial outliers indicates that ISOMRD is a robust model for spatial outlier detection with high accuracy and efficiency. Due to its competitive learning mechanism, the proposed model can perform efficiently to extract crucial features from spatial data with explicit visualization.

4.3. Multiple non-spatial attribute dataset introduction and description

In this paper, all datasets are derived from the U.S. Census Bureau [29] to demonstrate the performance of our approach. The purpose of our experiments is to investigate spatial and non-spatial attributes for each specific dataset and select the top 25 counties as spatial outlier candidates for the corresponding category items including house, education, crime and West Nile Virus (WNV). According to practical requirements in our simulation, some representative attributes or features are taken into account so as to minimize the unexpected errors in raw data resource.

The location information of 3108 counties (Alaska and Hawaii excluded) in the U.S. are collected to construct spatial attributes. With the values of longitude and latitude, it is available to locate the exact position for spatial data items. The spatial data can be used to search spatial clusters by SOM, as illustrated in Fig. 5. The clustering characteristics that depend on intrinsic properties of spatial attributes and competitive learning mechanisms can reflect implicit spatial patterns of arbitrary spatial datasets. Due to the number of 3108 counties, the optimal neuron number σ^{*} can be obviously calculated according to Eq. (30). The size of the feature map also depends on the total input data. Compared with the other common methods, the unsupervised learning techniques facilitate the procedure to further discover spatial attributes with more reasonable manner, which can essentially reflect the principles of spatial auto-correlation [23].

The formation of neighboring clusters via SOM seems to be naturally organized and by virtue of competitively computational procedure and topological preservation. The analysis of cluster density can help us to understand the quantity of spatial data with similar spatial patterns. Furthermore, the histogram of spatial clusters can also be employed to display the neighborhood based on the feature map. The corresponding neurons with the exact two-dimensional index in the feature map appear clearly in the histogram of SOM cluster density as shown in Fig. 6.

The neighboring clusters based on spatial relationships are crucial for spatial outlier detection. Nevertheless, the outlier

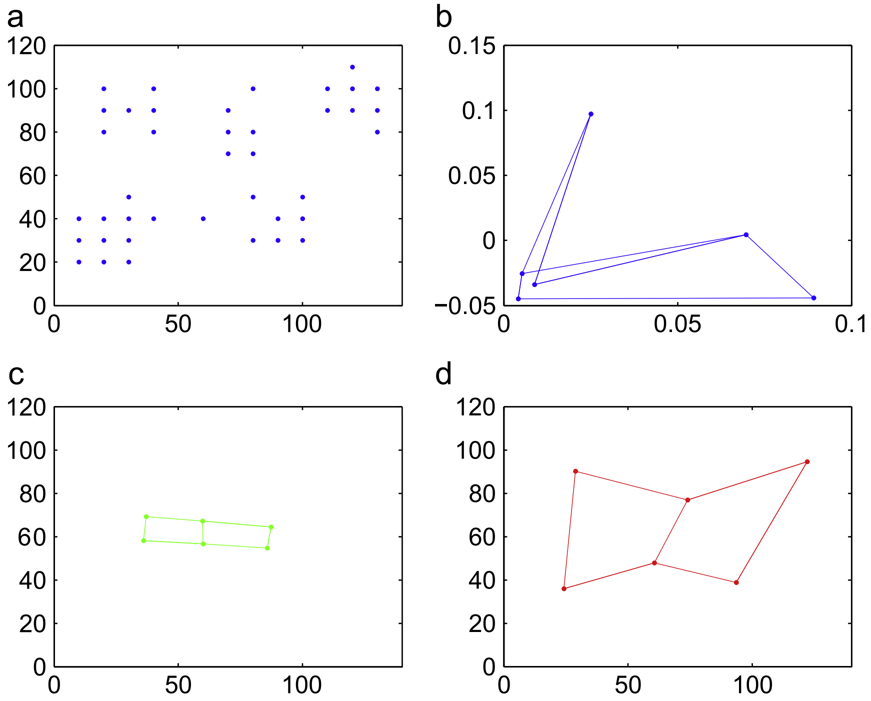


Fig. 4. The application of ISOMRD model in spatial outlier detection. (a) Spatial attributes. (b) Initial weight of neurons. (c) SOM neuron ordering. (d) Convergence for updating SOM neuron's weight.

Table 2
The top 4 potential spatial outliers detected by ISOMRD, Z algorithm, Moran scatterplot, and Scatterplot.

Rank	ISOMRD	Z algorithm	Moran scatterplot	Scatterplot
1	S1	S1	E2	S1
2	S2	E2	S1	S2
3	S4	E1	E1	E2
4	S3	S4	E3	E3

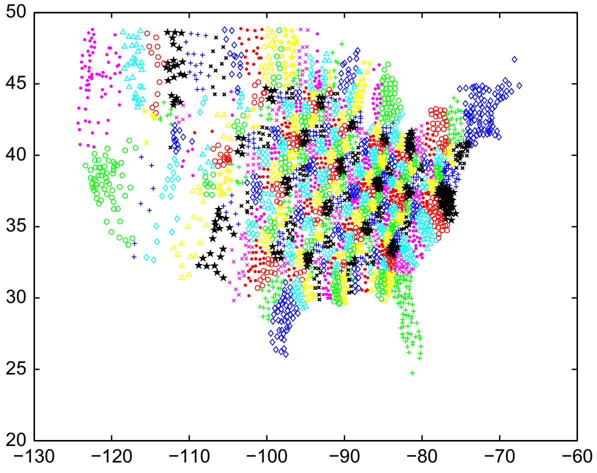


Fig. 5. Spatial clusters: The data points in the clusters with continuously identical marks and colors share common spatial properties in the neighborhood. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

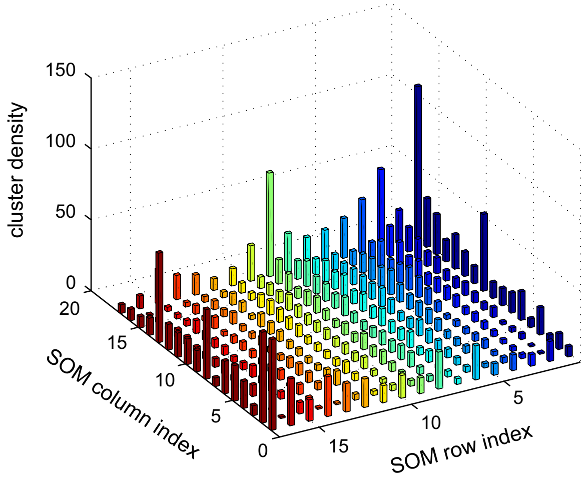


Fig. 6. The histogram of cluster density in SOM.

candidates depend on non-spatial attributes or patterns in spatial dataset.

As mentioned at the beginning of this section, several datasets including house, education, crime and West Nile Virus (WNV), become the main targets in the simulation. The "house" dataset primarily focuses on the housing units and building permits in the United States. It collects the detailed information about the housing or building ownerships and distribution density. The non-spatial attributes with 5 dimensions include house units in 2000, house units net change percentage from 2000 to 2005, house units per square mile of land area in 2005, housing units in owner-occupied percentage in 2000 and housing units in multi-unit structures percentage in 2000.

The "education" dataset includes some relevant data to describe the situation of educational degrees for the U.S. residents, which facilitates us to recognize highly educated regions or poorly educated regions. Furthermore, the educational levels can also be observed and distinguished by various categories of degrees listed as non-spatial attributes, which consist of less than high school degree in 2000, high school degree only in 2000, some college degree in 2000, college degree or above (at least a 4 year degree) in 2000.

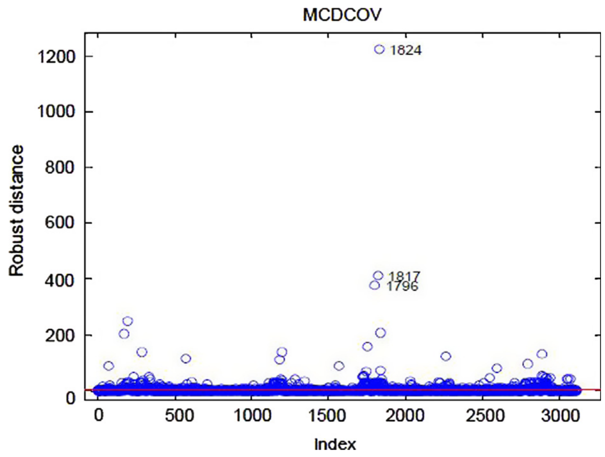


Fig. 7. The robust distance on House dataset. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

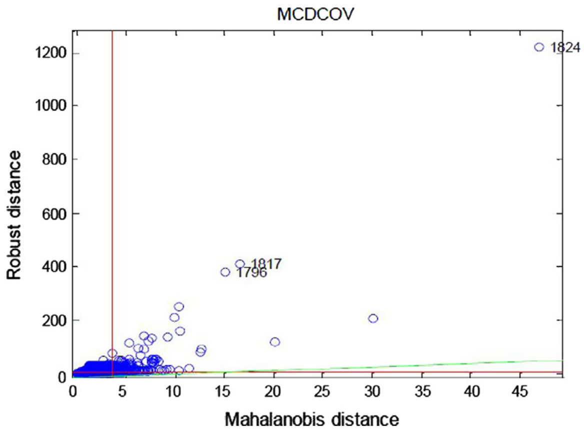


Fig. 9. The comparison between robust distance and Mahalanobis distance on House dataset.

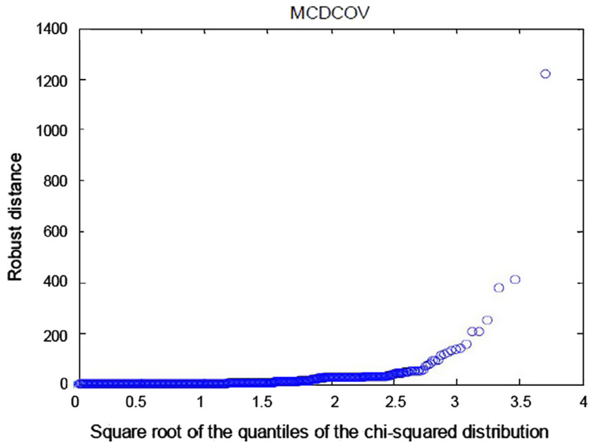


Fig. 8. The robust distance based on the square root of the quantiles of chi-squared distribution on House dataset.

Spatial outliers or anomalies can be used to discover the increasing rate of crime or the "hot" spots with higher crime rate. The selected categories of crime in 2004 can be summarized as murder and non-negligent man-slaughter, forcible rape, robbery, aggravated assault, burglary, larceny-theft as well as motor vehicle theft.

With rapid improvement in the standard of living, the public health issues are becoming more and more sensitive topics. Some diseases can be spread through animals in certain regions of the country. To resolve this problem, spatial outliers can be associated with disease control and prevention. The "West Nile Virus" is an example for such challenging cases. The spatial dataset incorporates the concrete number of the "West Nile Virus" infected birds in 2002, the infected veterinaries in 2002, the infected birds in 2003 and the infected veterinaries in 2003.

4.4. Simulation performance analysis

All spatial datasets involved in the simulation share two-dimensional spatial attributes. With specific spatial properties, the purpose of neighborhood clustering is to discover the implicit or unexpected spatial similarities or relationships. The comparisons of several neighboring cluster approaches are tested.

The robust distance in the "house" dataset illustrated in Fig. 7 can be used to efficiently detect spatial outliers. The red line indicates a reasonable distance threshold. It can be observed that several data items have large distance values, such as the counties with the index "1824", "1817" or "1796". However, there are also a large amount of spatial outlier candidates with relatively short distances although above the red line. Most of these data items cannot be easily detected as outliers by the other algorithms, but SOM with robust distance performs better.

To accurately estimate robust MCD [33] mean and covariance matrix, the toolbox LIBRA [9] can be employed. The robust distance depends on the quantiles of chi-squared distribution. From Fig. 8, the top outliers are obviously responsible for higher quantile values, which can provide more useful information to select the robust cut-off percentiles in the procedure of outlier detection.

The comparison between two types of distances indicates that the robust distance might provide more robust cut-off percentile as shown in Fig. 9, since chi-squared percentile cut-off discards too many points that might be outlier candidates. Furthermore, the outlier candidates with larger values in robust distance metric can facilitate us to accurately identify truly definite outliers based on asymptotic analysis in previous discussion.

Fig. 10 illustrates ISOMRD algorithm in detecting spatial outliers in "house" dataset. The top outlier candidates with high robust distance can be visualized in three-dimension as given in Fig. 10(a), whereas Fig. 10(b) exhibits the spatial location distribution. For explicit observation, longitude or latitude is selected separately to specify some data points with outstanding robust distance shown in Fig. 10(c) and (d).

From Fig. 11, the quantitative comparisons among several approaches demonstrate that ISOMRD has superior capability for spatial outlier detection due to its robust distance estimation, while other methods using Mahalanobis distance can only detect less number of potential outlier candidates. However, the KNN and SOMMD perform better than Grid and Grid based KNN, since Grid restricts the spatial neighborhood definition regardless of spatial dependency, but it is a highly efficient outlier detection method.

In order to make a detailed comparison of various methods, Table 3 demonstrates the top 25 potential outlier counties identified by several algorithms. The detection results of several other methods seem to be less effective than ISOMRD, although the top outlier counties might appear similarly.

ISOMRD displays hierarchical structure in distance distribution while other methods with Mahalanobis distance can hardly reach this point for the hypothetical properties brought by asymptotic cut-offs. More importantly, the asymptotic cut-offs used in robust distance also show better performance in the large size of dataset. Due to the conservative characteristic of asymptotic methods in

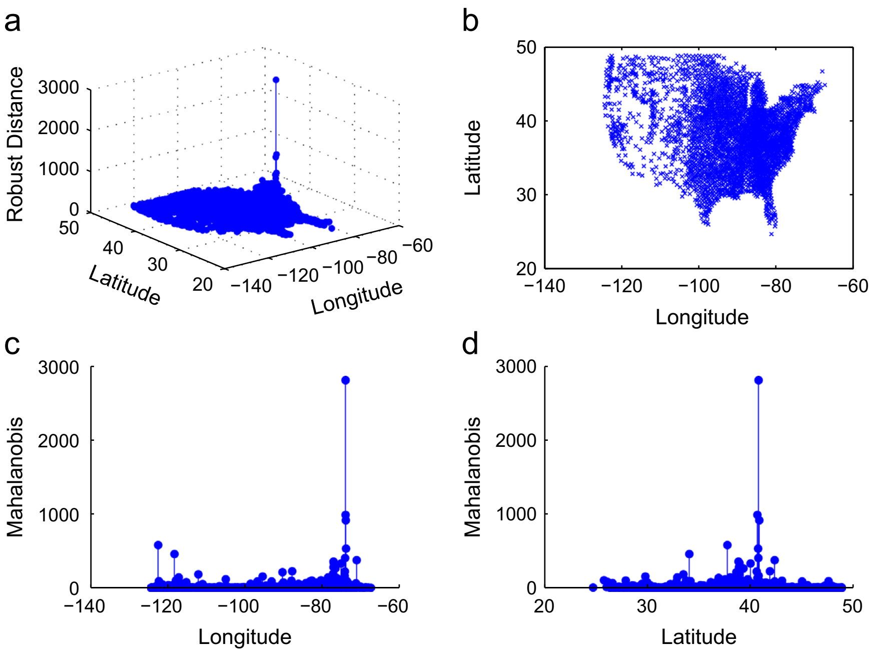


Fig. 10. Spatial outlier detection by ISOMRD. (a) Spatial outlier detection, (b) Spatial data Set, (c) Spatial outlier detection(X–Z) and (d) Spatial outlier detection(Y–Z).

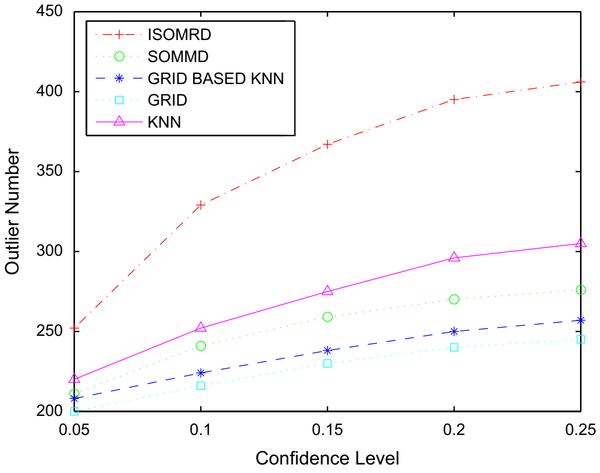


Fig. 11. The relationship between outlier number and confidence level.

the small sample size, minor adaptive adjustments might be required to handle this problem.

The non-spatial attribute values in these tables are normalized. Several counties such as New York (NY) and Los Angeles (CA) appear in the highest ranking positions while most of other counties in the top 25 are in different ordering. Some counties occur sparsely as outlier candidates, like Grafton (NH) by ISOMRD or Wayne (MI) by SOMMD with much lower ranking. This phenomenon is caused by the distinct definitions of neighborhood in given methods.

From these tables one can see that, those top outlier candidates, which have outstanding robust distance or Mahalanobis distance, can be ascribed to one or multiple irregular values in non-spatial attributes, especially when compared with their neighboring counties. For example, in ISOMRD, New York (NY) is the No. 1 outlier county affected by the number of house units per square mile in 2005 (44.77) and its neighboring counties with lower density of house units as Bergen—NJ (1.7219), Bronx—NY (14.8914), Queens—NY (9.4402) and Kings—NY (16.7425), etc. Among those neighboring counties, some might be also identified as spatial outliers due to the comparison with non-spatial properties of their surrounding counties. Another typical case is Cook (IL), which is also one of top 25 outlier candidates with significant value (18.5294) in the attribute category of house units in 2000. But it behaves normally in other attributes (e.g. the percentage of units net change) compared with its neighborhood counties Lake—IL (1.7005), Kenosha—WI (0.2064), and Racine—WI (0.3389), etc.

Additional simulations for the rest of spatial datasets including “education”, “crime” as well as “WNV” dataset are performed. The experiments show that the SOM approach is an effective tool to detect spatial outliers. It differs from the traditional machine learning and data mining techniques, which are mostly concentrated on Euclidean distance on spatial attributes. The competitive learning in SOM promotes synaptic neurons to collaborate with each other and adaptively form feature maps with the property of topological ordering. By visual geometric computation, the proposed method can ultimately acquire important information of the inherent connection for spatial data mining.

Comparisons of several algorithms (ISOMRD, SOMMD, KNN, Grid and Grid based KNN) to search neighboring clusters indicate that ISOMRD can accurately discover potential spatial outliers corresponding neighborhood locations. The iterative procedure can eliminate influence of local outliers, whereas SOMMD might ignore too many potential outlier candidates.

From Table 4, we can see that the runtime of both ISOMRD and SOMMD depends on the neuron number and maximum iteration. When $M=25$ and $I_{max}=1000$, the runtime is similar to the time of Grid, which is faster than other methods. As the two key parameters increase, the difference of results becomes larger. However, as mentioned before, Grid might lose much spatial relationships and reduce the detection accuracy. KNN is obviously time-consuming when K increases. Grid based KNN is faster than KNN, but the runtime of Grid based KNN mainly depends on K. It also lacks detection accuracy when the spatial attribute becomes complex, which is similar to Grid. Based on the simulation results, we observe that ISOMRD algorithm presents an effective approach for such challenging spatial outlier detection applications.

Table 3
The top 25 spatial outlier candidate for House dataset.

Rank	ISOMRD		SOMMD		Grid Based KNN		Grid		KNN	
	County	RD	County	MD	County	MD	County	MD	County	MD
1	New York, NY	1222.9	New York, NY	47.0005	New York, NY	47.1676	New York, NY	46.0212	New York, NY	39.3546
2	Kings, NY	411.6	Los Angeles, CA	30.1396	Los Angeles, CA	33.1943	Los Angeles, CA	28.9938	Los Angeles, CA	30.9864
3	Bronx, NY	378.6	Cook, IL	20.1359	Cook, IL	19.9999	Cook, IL	21.4275	Cook, IL	20.5746
4	San Francisco, CA	252.6	Kings, NY	16.6126	Kings, NY	17.2194	Bronx, NY	15.9446	Nassau, NY	17.4671
5	Queens, NY	207.9	Bronx, NY	15.0883	Bronx, NY	15.6216	Kings, NY	13.8582	Westchester, NY	14.7750
6	Los Angeles, CA	204.7	Maricopa, AZ	12.6488	Harris, TX	12.4259	Maricopa, AZ	13.3140	Rockland, NY	13.2724
7	Suffolk, MA	157.1	Harris, TX	12.5962	Maricopa, AZ	12.3113	Harris, TX	13.2662	Harris, TX	13.2003
8	Hudson, NJ	139.8	Flagler, FL	11.3892	Flagler, FL	11.2978	Ventura, CA	11.6329	Bergen, NJ	13.1807
9	Alexandria, VA	138.3	Hudson, NJ	10.4354	Hudson, NJ	10.8455	San Francisco, CA	10.5558	Maricopa, AZ	12.7732
10	Philadelphia, PA	132.4	San Francisco, CA	10.3692	Queens, NY	10.6102	Chattahoochee, GA	10.1661	Richmond, NY	12.2410
11	DC	123.6	Chattahoochee, GA	10.3348	San Francisco, CA	10.1129	Suffolk, MA	9.9969	Passaic, NJ	11.3938
12	Cook, IL	118.0	Queens, NY	9.8641	Chattahoochee,GA	10.0224	Flagler, FL	9.7714	Flagler, FL	11.1929
13	Arlington, VA	113.2	Paulding, GA	9.4858	San Diego, CA	9.9802	Hudson, NJ	9.1554	Hudson, NJ	11.0518
14	Baltimore city, MD	96.3	Suffolk, MA	9.1950	Suffolk, MA	9.9380	Loudoun, VA	8.9596	San Francisco, CA	10.643
15	Maricopa, AZ	90.8	Loudoun, VA	9.0758	Loudoun, VA	9.0355	King, TX	8.8284	Loudoun, VA	9.9641
16	St. Louis city, MO	90.7	King, TX	8.8654	Henry, GA	9.0182	Dallas, TX	8.6168	Chattahoochee, GA	9.3750
17	Harris, TX	80.6	Kenedy, TX	8.6339	Paulding, GA	8.9215	Kenedy, TX	8.5594	Queens, NY	9.1776
18	San Diego, CA	74.9	Rockwall, TX	8.4423	Orange, CA	8.8396	Kendall, IL	8.3657	Suffolk, MA	9.0571
19	Richmond, NY	70.0	Dallas, TX	8.2839	Kenedy, TX	8.5330	Wayne, MI	8.3007	Ventura, CA	8.9987
20	Charlottesville, VA	54.1	Eureka, NV	8.0410	Alexandria, VA	8.4137	Rockwall, TX	8.2880	King, TX	8.8110
21	Falls Church, VA	53.2	Nye, NV	7.9376	King, TX	7.9374	Alexandria, VA	8.2109	Kenedy, TX	8.3252
22	Miami-Dade, FL	52.3	Miami-Dade, FL	7.8266	Summit, CO	7.9361	Paulding, GA	8.1900	Dallas, TX	8.2756
23	Essex, NJ	52.2	Broward, FL	7.7952	Newton, GA	7.9178	Henry, GA	8.1568	Fredericksburg, VA	8.2017
24	Union, NJ	52.0	Lincoln, SD	7.7079	Dallas, TX	7.8112	Philadelphia, PA	7.8916	Summit, CO	8.1489
25	Grafton, NH	51.6	Wayne, MI	7.6844	Philadelphia, PA	7.7990	Douglas, CO	7.6760	Lincoln, SD	8.1181

Table 4
Comparison of runtime in the training phase.

Method	Parameters		Time (s)
ISOMRD	M	I_{max}	T
	25	1000	3.8186
	100	2000	8.7639
	400	5000	20.6095
SOMMD	M	I_{max}	T
	25	1000	3.2872
	100	2000	8.3853
	400	5000	20.3967
Grid	N		T
	25		1.8268
	100		5.5922
	400		12.9817
KNN	K		T
	5		12.6891
	10		22.7566
	20		45.1219
Grid based KNN	N	K	T
	400	5	9.2896
	100	10	11.0768
	25	20	15.4325

5. Conclusion

In this work, we propose the ISOMRD algorithm with robust cutoffs and adaptive thresholds for spatial outlier detection, and we tested this approach on the spatial datasets derived from U.S. census database. The experimental results and comparative analysis demonstrate the effectiveness and adaptability of this method. We believe our research not only provides some novel techniques and solutions for spatial outlier detection, but also new insights to a wide range of spatial data mining applications. In our future work, we will extend the spatial outlier detection approach to the dynamical analysis of spatio-temporal data.

Acknowledgment

This work was supported by the National Science Foundation (NSF) under Grant ECCS 1053717 and CNS 1117314, Army Research Office (ARO) under Grant W911NF-12-1-0378, Defense Advanced Research Projects Agency (DARPA) under Grant FA8650-11-1-7148 and Grant FA8650-11-1-7152. This work is also partially supported by the Rhode Island NASA EPSCoR Research Infrastructure and Development (EPSCoR-RID) via NASA EPSCoR grant number NNX11AR21A (through a sub-award from Brown University).

References

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine, 17 (1996) 37–54.
[2] D. Larose, Discovering Knowledge in Data: an Introduction to Data Mining, Wiley, John & Sons, 2004.
[3] C. Aggarwal, P. Yu, Outlier detection for high dimensional data, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, Santa Barbara, CA, May 2001.
[4] V. Barnett, T. Lewis, Outliers in Statistical Data, John Wiley and Sons, New York, NY, 1994.
[5] E. Suzuki, J. Zytkow, Unified algorithm for undirected discovery of exception rules, in: Proceedings of the Principles of Data Mining and Knowledge Discovery, Fourth European Conference, PKDD-2000, Lyon, France, September 13–16, 2000, pp. 169–180.
[6] N. Chawla, A. Lazarevic, L. Hall, K. Bowyer, SMOTEBoost: improving the prediction of minority class in boosting, in: Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003, Cavtat, Croatia, September 2003.
[7] M. Markou, S. Singh, Novelty detection: a review-part1: statistical approaches, Signal Process. 83 (December (12)) (2003) 2481–2497.
[8] P. McBurney, Y. Ohsawa, Chance Discovery Advanced Information Processing, Springer, 2003.
[9] S. Verboven, M. Hubert, LIBRA: a Matlab library for robust analysis, Chemometrics Intell. Lab. Syst. 75 (1996) 127–136.
[10] A. Jain, M. Murty, P. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[11] R. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, in: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp. 144–155.

[12] M. Ester, H-P. Kriegel, J. Sander, X. Xu, Clustering for mining in large spatial databases, KI J. Artif. Intell. Spec. Issue Data Min. 12 (1) (1998) 18–24.

[13] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithms for large databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, pp. 73–84.

[14] E. Knorr, R. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, J. Very Large Data Bases 8 (3–4) (2000) 237–253.

[15] E. Knorr, R. Ng, Algorithms for mining distance based outliers in large data sets, in: Proceedings of the Very Large Databases (VLDB) Conference, New York City, NY, August 1998.

[16] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, in: Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.

[17] M. Breunig, H-P. Kriegel, R. Ng, J. Sander, LOF: identifying density-based local outliers, in: ACM SIGMOD International Conference on Management of Data, Dallas, TX, 2000, pp. 93–104.

[18] I. Ruts, P. Rousseeuw, Computing depth contours of bivariate point clouds, J. Comput. Stat. Data Anal. 23 (1996) 153–168.

[19] R. Gting, An introduction to spatial database systems, Int. J. Very Large Data Bases 3 (1994) 357–399.

[20] J. Roddick, M. Spiliopoulou, A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research, ACM SIGKDD Explorations Newsletter, 1999.

[21] S. Shekhar, S. Chawla, A Tour of Spatial Databases, Prentice Hall, 2002.

[22] N. Cressie, Statistics for Spatial Data, John Wiley and Sons, New York, 1993.

[23] A. Cliff, J. Ord, Spatial Autocorrelation, London Pion, 1973.

[24] W. Tobler, A computer movie simulating urban growth in the Detroit region, Econ. Geogr. 46 (2) (1970) 234–240.

[25] P. Sun, S. Chawla, On local spatial outliers, in: Proceedings of the IEEE International Conference on Data Mining, 2004, pp. 209–216.

[26] T. Kohonen, Self-Organizing Maps, Springer, 2001.

[27] Q. Cai, H. He, H. Man, SOMSO: a self-organizing map approach for spatial outlier detection with multiple attributes, in: Proceedings of the International Joint Conference on Neural Networks, 2009, pp. 425–431.

[28] Q. Cai, H. He, H. Man, J. Qiu, IterativeSOMSO: an iterative self-organizing map for spatial outlier detection, in: Proceedings of the ISNN (1), 2010, pp. 325–330.

[29] U.S. Census Bureau, United States Department of Commerce. ⟨http://www.census.gov/⟩.

[30] D. Hawkins, Identification of Outliers, Chapman and Hall, 1980.

[31] E. Knorr, R. Ng, A unified notion of outliers: properties and computation, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1997, pp. 219–222.

[32] P. Rousseeuw, K.V. Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.

[33] J. Hardin, D. Rocke, The distribution of robust distances, J. Comput. Graph. Stat. 14 (2005) 928–946.

[34] X. Jia, J. Richards, Fast k-NN classification using the clustering—space approach, IEEE Geosci. Remote Sensing Lett. 2 (April (2)) (2005).

[35] S. Papadimitriou, H. Kitagawa, P. Gibbons, C. Faloutsos, Loci: fast outlier detection using the local correlation integral, in: Proceedings of the International Conference on Data Engineering, March 2003, pp. 315–328.

[36] D. Chen, C. Lu, Y. Kou, F. Chen, On detecting spatial outliers, GeoInformatica 12 (4) (2008) 455–475.

[37] S. Shekhar, C. Lu, P. Zhang, A unified approach to spatial outliers detection, in: GeoInformatica, vol. 7, Springer, 2003, pp. 139–166.

[38] L. Anselin, Exploratory spatial data analysis and geographic information systems, in: New Tools for Spatial Analysis, Eurostat, Luxembourg, 1994.

[39] L. Anselin, Local Indicators of Spatial Association-LISA, Geographical Analysis, 1995.

[40] C. Lu, D. Chen, Y. Kou, Algorithms for spatial outlier detection, in: Proceedings of the Third IEEE International Conference on Data Mining, 2003, pp. 597–600.

[41] C. Lu, D. Chen, Y. Kou, Detecting spatial outliers with multiple attributes, in: Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, 2003, pp. 122–128.

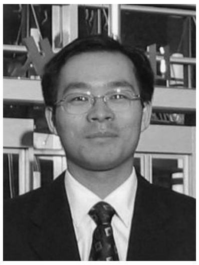
[42] K. Koperski, J. Adhikary, J. Han, Spatial data mining: progress and challenges, in: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, June 1996.

[43] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, The MIT Press, 2001, pp. 276–277.

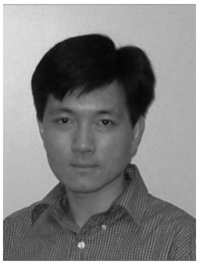
[44] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall/CRC, London, 1995.

[45] Big Data, Nature, http://www.nature.com/news/specials/bigdata/index.html, Sep. (2008).

Qiao Cai received the B.S. degree in electrical engineering from Wuhan University, China, in 2005, and the M.S. degree in electrical engineering from Huazhong University of Science and Technology, China, in 2008. He is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, Stevens Institute of Technology. His current research interests include machine learning, data mining, computational intelligence, and pattern recognition. He is a student member of IEEE.



Haibo He received the B.S. and M.S. degrees in electrical engineering from Huazhong University of Science and Technology, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from Ohio University in 2006. He is currently an Associate Professor at the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island. From 2006 to 2009, he was an Assistant Professor at the Department of Electrical and Computer Engineering, Stevens Institute of Technology. His current research interests include computational intelligence, self-adaptive systems, machine learning, data mining, embedded intelligent system design (VLSI/FPGA), and various applications such as smart grid, cognitive radio, sensor networks, and others. He has published 1 sole-author research book with Wiley, edited 6 conference proceedings with Springer, and authored and co-authored over 100 peer-reviewed journal and conference papers. His research results have been covered by national and international medias such as The Wall Street Journal, Yahoo!, Providence Business News, among others. He has delivered numerous invited talks. He is currently a member of various IEEE Technical Committees, and also served as various conference chairship positions for numerous international conferences. He has been a Guest Editor for several journals including IEEE Computational Intelligence Magazine, Cognitive Computation (Springer), Applied Mathematics and Computation (Elsevier), Soft Computing (Springer), among others. Currently, he is an Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems, and IEEE Transactions on Smart Grid, and also serves on the Editorial Board for several international journals. He received the National Science Foundation (NSF) CAREER Award (2011) and Providence Business News (PBN) “Rising Star Innovator” Award (2011). He is a Senior Member of IEEE.



Hong Man received his Ph.D. degree in electrical engineering from Georgia Institute of Technology in 1999. He joined Stevens Institute of Technology in 2000. He is currently an associate professor and the director of the Computer Engineering Program in the Electrical and Computer Engineering Department, and the director of the Visual Information Environment Laboratory at Stevens. He served on the Organizing Committees of IEEE ICME (2007 and 2009), IEEE MMSP (2002 and 2005), and the Technical Program Committees of various IEEE conferences. His research interests include signal and image processing, pattern recognition and data mining. He is a senior member of IEEE.