Data Min Knowl Disc (2010) 20:221–258
DOI 10.1007/s10618-009-0147-0

Spatial neighborhood based anomaly detection in sensor datasets

Vandana P. Janeja · Nabil R. Adam · Vijayalakshmi Atluri · Jaideep Vaidya

Received: 1 August 2008 / Accepted: 14 August 2009 / Published online: 22 January 2010

Abstract Success of anomaly detection, similar to other spatial data mining techniques, relies on neighborhood definition. In this paper, we argue that the anomalous behavior of spatial objects in a neighborhood can be truly captured when both (a) *spatial autocorrelation* (similar behavior of nearby objects due to proximity) and (b) *spatial heterogeneity* (distinct behavior of nearby objects due to difference in the underlying processes in the region) are taken into consideration for the neighborhood definition. Our approach begins by generating *micro neighborhoods* around spatial objects encompassing all the information about a spatial object. We selectively merge these based on *spatial relationships* accounting for autocorrelation and *inferential relationships* accounting for heterogeneity, forming *macro neighborhoods*. In such neighborhoods, we then identify (i) *spatio-temporal outliers*, where individual sensor readings are anomalous, (ii) *spatial outliers*, where the entire sensor is an anomaly, and (iii) *spatio-temporally coalesced outliers*, where a group of spatio-temporal outliers in the macro neighborhood are separated by a small time lag indicating the traversal

Responsible editor: Sanjay Chawla.

This work is supported in part by the National Science Foundation under grants IIS-0306838 and CNS-0746943.

V. P. Janeja
University of Maryland, Baltimore County, ITE 429, 1000 Hilltop Circle, Baltimore, MD 21250, USA
e-mail: vjaneja@umbc.edu

N. R. Adam · V. Atluri · J. Vaidya (✉)
Rutgers University, 1 Washington Park, Newark, NJ 07102, USA
e-mail: jsvaidya@business.rutgers.edu; jsvaidya@rbs.rutgers.edu

N. R. Adam
e-mail: adam@adam.rutgers.edu

V. Atluri
e-mail: atluri@business.rutgers.edu

of the anomaly. We demonstrate the effectiveness of our approach in neighborhood formation and anomaly detection with experimental results in (i) water monitoring and (ii) highway traffic monitoring sensor datasets. We also compare the results of our approach with an existing approach for spatial anomaly detection.

Keywords Outlier detection · Spatial neighborhood · Sensors

1 Introduction

An outlier is an object which is vastly deviant in behavior with respect to some other objects in comparison. Lets call these comparative objects as neighbors. The basic idea in the discovery of outliers is that there is always a frame of reference, such as similarly behaving neighbors, to identifying an object as an outlier, which is in turn different from these neighbors. This behavior is quantified in terms of significant difference in attribute values of the object with its neighbors.

Outlier detection in spatial data extends the data mining process to more complex spatial objects qualified by both spatial (e.g., point, lines, polygons, location, etc.) and non-spatial data (e.g., population count, water salinity, pH, etc.). Such anomaly detection is useful in spatio temporal sensor datasets in several domains, for example: (*i*) *Environmental scientists may want to identify abnormally behaving water monitoring sensors,* (*ii*) *A customs agent may want to discover anomalies among cargo shipments with RFID tags, to identify potentially deviant shipments even before they cross the border,* (*iii*) *City officials may want to identify threats or malfunctions based on numerous sensors placed around a metropolitan area in subways and tunnels, etc.*

Anomaly detection, similar to other spatial data mining techniques, first captures the behavior of objects using neighborhood definition. Essentially, spatial neighborhood refers to a group of similarly behaving objects in a certain spatial proximity. Behavior of the objects is inherently governed by spatial dependence among the neighboring objects, specifically due to the properties of spatial autocorrelation and heterogeneity (Unwin 1982; Shekhar et al. 2001, 2002). On the one hand, nearby objects are related due to spatial autocorrelation, on the other hand, spatial heterogeneity causes attribute values of objects to vary greatly even by small changes in the spatial region where the object is located. Thus, the neighborhood (Ester et al. 1997; Miller and Han 2001; Shekhar et al. 2001) cannot be identified simply by the change in geospatial features or solely on the basis of spatial proximity (e.g., Ester et al. 1996; Kang et al. 1997; Ester et al. 1998) determined using spatial relationships such as adjacency of objects. Any spatial data mining task is directly affected by the accurate accounting of these properties.

One such spatial outlier detection technique (Shekhar et al. 2001) uses a graph based neighborhood definition. Here objects are connected in a graph based on spatial relationships, such as adjacency. The size of the neighborhood is determined based on the cardinality of the number of objects in the neighborhood (for instance 1 to 10 neighbors). In such a neighborhood, the value of one attribute is compared to the aggregate value of its neighbors. However, this and other similar approaches (Ng and Han 1994; Ester et al. 1996; Kang et al. 1997; Lu et al. 2003; Sun and Chawla 2004) have the following three issues:

– *First*, these techniques mainly consider neighborhoods which are based primarily on spatial relationships thus accommodating autocorrelation only but tend to ignore the effect of spatial heterogeneity in combination.
– *Second*, the neighborhood formation is not order invariant. For instance if we consider a neighborhood based on cardinality(or number of objects in the neighborhood) then depending on which point we start with the neighborhood formation will be different every time leading to different outlier detection outcomes.
– *Third*, the outlier detection considers the deviation in terms of a single attribute only. Recently there have been some approaches which detect outliers based on deviation in multiple attributes (Lu et al. 2003), however, they do not account for autocorrelation and heterogeneity. Moreover, multiple attributes are not considered for the neighborhood definition but only for the outlier detection.

1.1 Motivating example

Let us consider an example comprised of a set of sensors with the goal of detecting anomalous levels of toxicity levels in a water body. These are shown as solid dots in the Fig. 1. Assume the neighborhood is formed based on spatial autocorrelation, as in Fig. 1a. In this neighborhood formation, the values of the toxicity levels of sensor readings are not considered. As a result, the node with extreme toxicity value (with 80) may be considered to be part of the neighborhood as spatial autocorrelation alone is considered. This may lead to identification of frivolous anomalies. However, if we take the toxicity level also into consideration when defining the neighborhoods, they may be formed as shown in Fig. 1b. Under this, the anomaly detected earlier is no longer considered as an anomaly. Therefore, we argue in this paper that considering autocorrelation captured through adjacency relationships alone is not adequate to identify true anomalies, but one need to consider heterogeneity captured by the underlying features of the sensor readings.

In such a neighborhood we would like to identify sensor readings which are anomalous with respect to the other readings of the neighborhood objects. Secondly we would like to identify whether the objects generating these readings are malfunctioning or have an overwhelming number of anomalous readings associated with them. Finally in order to detect if there is a truly unusual phenomenon which is being monitored

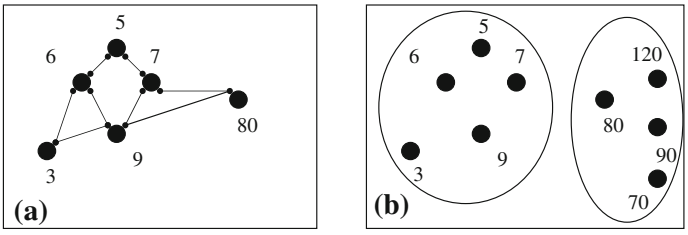

Fig. 1 Considering autocorrelation and heterogeneity for a complete spatial neighborhood definition. (**a**) Neighborhood formed based on only spatial relationships. (**b**) Possible demarcation of neighborhoods based on spatial and inferential relationships

by these sensors we would like to see if sensors spatio-temporally in relation to each other are generating anomalous readings indicating a traversing anomaly. □

Specifically, our paper makes the following contributions.

- We present an order invariant technique to more accurately identify neighborhoods for spatial objects capturing spatial autocorrelation and spatial heterogeneity in combination.
- We present an anomaly detection technique to identify different types of outliers in this refined spatial neighborhood. Specifically we discover three types of outliers in spatio-temporal sensor datasets:(*a*) *Spatio-temporal outlier:* for detecting sensor readings that are anomalous with respect to other readings in the neighborhood, (*b*) *Spatial outlier:* for the identification of sensors which are entirely anomalous with respect to other sensors in the neighborhood, (*c*) *Spatio-temporally coalesced outliers:* for detecting a set of temporally and spatially linked anomalous readings (spatio-temporal outliers) indicating the traversal of an anomaly across the region over time. Essentially we are discovering whether there is truly an anomalous phenomenon traversing the region or it is a potential malfunction due to weather fluctuations, debris and so on.
- We demonstrate the effectiveness of our approach in sensor datasets and provide detailed experimental results. Specifically, we demonstrate using water monitoring and highway traffic monitoring sensor datasets that we can refine the creation of neighborhoods and subsequently refine the discovery of outliers in the spatial neighborhood based on our approach.

Our approach has several advantages over the existing approaches. Specifically our approach is order invariant and captures the spatial properties in our neighborhood definition leading to a well refined outlier discovery in these neighborhoods as demonstrated by our results in real world datasets. Essentially the reason for capturing autocorrelation and heterogeneity is that we can produce a quantified improvement in the outliers detected leading to a decrease in the false positives and increase in the accuracy of our approach.

The rest of the paper is organized as follows. In Sect. 2 we discuss spatial autocorrelation and heterogeneity. In Sect. 3 we outline some preliminary concepts. In Sect. 4 we outline the proposed approach and describe each step in detail. In Sect. 5 we discuss anomaly detection in spatial neighborhood. In Sect. 6, we provide detailed experimental results to validate our techniques, as well as discuss the tuning of several parameters. Related work is presented in Sect. 7. Section 8 concludes the paper and discusses future work.

2 Spatial autocorrelation and heterogeneity

Spatial Autocorrelation essentially embodies Tobler's First Law of Geography which states that everything is related to everything else but nearby objects are more related than distant objects. Autocorrelation (Griffith 1987; Haining 2003) literally means the correlation of a variable to itself. Spatial autocorrelation refers to the correlation of the variable with itself in space. It can be positive (spatial clusters for high–high or low–low values) and negative (high–low or low–high values). Spatial Autocorrelation can

be measured and quantified using various statistical measures which firstly identify the spatial relationships for each object with the other objects in the region. It then identifies the extent of the deviations from the mean and the variability (or similarity) of the variable across the region. These measures of autocorrelation provide an overall assessment of the region. However, within the overall autocorrelated neighborhood there are pockets of variability.

Spatial Heterogeneity is the spatially varying autocorrelation. It refers to the structural instability in the region that may generate characteristic spatial patterns. This behavior can be understood in scales of spatial variation which can be summarized in the following (Haining 2003):

Spatial data = large scale variation + medium/small scale variation + error

Large scale variation refers to the overall trends present across the region under study representing spatial autocorrelation. *Medium/small scale variation* refers to the more localized spatial structures in the data and is present as a superimposition on the large scale variation. The error may be due to measurement or independent noise. For certain cases the small scale variation may dominate leading to spatial heterogeneity in the region. Let us consider here the localized distribution based heterogeneity depicted as the pockets of varying distributions in a region due to the local processes across the region. For instance due to some underlying features such as rivers, mountains, man made features such as landfills the behavior of certain objects is different as compared to others.

In defining a neighborhood it is essential to consider this type of heterogeneity in the vicinity since most of the times the outliers are determined by such localized factors. This type of heterogeneity plays a key role in forming the neighborhood. Thus, we need to have some technique for incorporating the local knowledge in each object for defining a neighborhood and identifying outliers in it.

It is therefore, critical to identify the right neighborhood for a given object before performing the outlier detection. The focus of this paper is to conduct anomaly detection in sensor datasets considering both spatial autocorrelation and heterogeneity. Here the sensors are spatial objects and the data is spatio-temporal in nature. Of the various types of sensors we focus on environmental sensor datasets, where the sensors are associated with a spatial location and monitor the environmental variables as readings. However, this approach is generalizable to other sensor datasets as well.

3 Preliminaries

3.1 Spatial data primitives

3.1.1 Spatial object

We first begin by defining a spatial object. We assume S be the set of spatial objects, $S = \{s_1, \ldots, s_n\}$. A spatial object can be formally defined as follows.

Definition 1 (*Spatial object*) Each spatial object, $s_i \in S$ is associated with (i) spatial coordinates (x_i, y_i) representing its latitude and longitude, and (ii) a set of spatial

and non-spatial attributes $a_i = \{a_{i1}, \ldots, a_{im}\}$. Moreover each s_i is associated with a set of points $p_i = \{p_{i1}, \ldots, p_{it}\}$, which are generated by s_i specific to the spatial location.

3.1.2 Voronoi diagrams

Let us assume that we have a finite set of n distinct spatial objects S in a plane. We generate the periphery of each object in the form of a Voronoi polygon. This can be further understood from the simplest technique of creating a Voronoi diagram where two objects are connected by a line segment and the bisector of the line segment divides the plane into two half planes. As we keep adding new spatial objects, more half planes are formed and the region of influence of the object is the intersection of these half planes. For example, consider Fig. 2 where s_1 to s_6 are six spatial units. The dotted lines indicate line segments connecting these spatial objects. The solid lines represent the bisecting lines, which form a Voronoi polygon surrounding s_1. We denote the Voronoi polygon surrounding a spatial object s_i with $V(s_i)$. The Voronoi polygons form a polygonal partition of the plane—called as the Voronoi diagram $V(S)$, of the finite spatial object set S. Thus, $V(S)$ is comprised of the entire proximity information about S in an explicit and computationally useful manner.

Several algorithms for identifying Voronoi polygons have been proposed in the literature (Aurenhammer 1991; Okabe et al. 2000). The performance of these algorithms deteriorates with the increase in the number of dimensions. However, we only need Voronoi diagrams in 2 dimensional space since we only consider the x, y coordinates of the spatial object to generate a Voronoi polygon. Currently, we use the Triangle-2D mesh generator (Shewchuk 1996) to generate the Voronoi diagram.

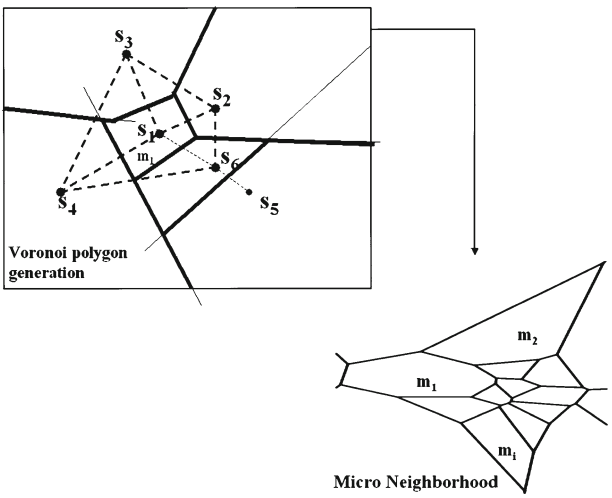

Fig. 2 Generation of Voronoi polygons

3.2 Similarity coefficients

The topic of similarity and overlap has been well studied in clustering and a number of similarity coefficients have been identified in the literature Kaufman and Rousseeuw (1990). We outline here the Jaccard Coefficient for quantifying similarity between binary valued vectors and the Silhouette Coefficient to quantify the similarity in terms of the overlap between clusters.

3.2.1 Jaccard coefficient

Jaccard Coefficient (JC) is used for identifying the similarity between two vectors where each bit of the vector can contain values 1 or 0. The Jaccard coefficient is computed as follows:

$$\text{JC} = \frac{\#\,\text{positive match}}{\#\,\text{positive match} + \#\,\text{mismatch}} \tag{1}$$

For example, consider Table 1 that gives the similarity matrix for two objects A and B representing the matches and mismatches between the individual bit values of the vector objects. The Jaccard coefficient $\text{JC} = 1/(1 + 1 + 2) = 0.25$.

3.2.2 Silhouette coefficient

The Silhouette coefficient (SC) has been used in the literature to identify the quality of clustering results in terms of structure and its silhouette (shadow) or overlap on other clusters (Ng and Han 1994).

First, some definitions: Given a point x in a cluster A, we define $a(x)$ to be the average distance between the point x and the other points in A and $b(x)$ to be the average distance between the point x and the points in the second closest cluster B. The Silhouette of x is then defined as (Kaufman and Rousseeuw 1990):

$$S(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{2}$$

3.3 Spatial outlier detection

Spatial outliers are data objects that are significantly different in attribute values from collection of data objects among the spatial neighborhood. Graph based spatial outlier

Table 1 Similarity for two vector objects

		Object B	
		1	0
Object A	1	1	1
	0	2	6

detection approach Shekhar et al. (2001) addresses the identification of neighborhood based spatial outliers. It finds outliers based on their distance (a spatial statistic) from the neighbors. The paper depicts an experiment based on data from traffic monitoring sensors located on lanes in the highways. Each sensor is associated with certain attribute values. The sensors in a highway network form the nodes and are connected by edges. Their technique determines the neighborhood as one based on fixed graph distance and the neighborhood aggregate function is mean, variance, autocorrelation. It is based on a statistic which compares attribute value of a data record $f(x)$ with the average attribute value of neighbors of $f(x)$ which is $f(y)$. The distance statistic $S(x)$ is given by: $S(x) = f(x) - f(y)$.

Based on the above statistic, the outlier is: $(S(x) - \mu_s)/\sigma_s > \theta$, i.e., $S(x)$—mean of all $S(x)$/SD should be greater than a threshold value (θ).

The threshold value is determined by the user, however, the confidence interval is determined by the assumption that the attributes follow a normal distribution.

4 The proposed approach

The goal of our approach is to identify spatial neighborhoods accommodating both autocorrelation and heterogeneity in the region and subsequently identify different types of outliers successfully, with a higher accuracy. In the context of our approach, we refer to each sensor as a spatial object and each individual reading (for a sensor) as a point. Our overall approach depicted in Fig. 3 has the following distinct steps:

1. We begin by generating the immediate neighborhood of an object, which we call as the **micro neighborhood**. A micro neighborhood captures the entire knowledge about the immediate neighborhood of this object such as the presence or absence of spatial features (e.g., landfill) in its proximity and attributes of the object.
2. We identify **spatial relationships** (such as adjacency) using the spatial attributes of the micro neighborhoods, to accommodate for autocorrelation.
3. In order to capture heterogeneity we identify **inferential relationships** between the non-spatial attributes and features in the micro neighborhoods. The inferential relationship quantifies the similarity of features and overlap of the points across micro neighborhoods. This facilitates the determination of the local heterogeneous patterns across neighborhood in terms of the attributes associated with each spatial object.

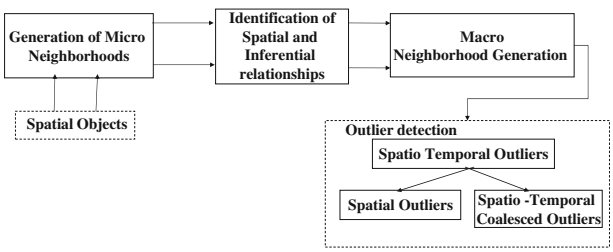

Fig. 3 Proposed approach to spatial neighborhood based anomaly detection

4. We then use both spatial and inferential relationships to merge appropriate micro neighborhoods to form larger **macro neighborhoods**. Thus, the macro neighborhood captures both autocorrelation and heterogeneity in the neighborhood definitions by not only considering proximity but also the change in the localized features and attributes in the region.
5. We next perform anomaly detection in these macro neighborhoods to discover **Spatio-temporal outliers**, **Spatial outliers** and **Spatio-temporally coalesced outliers**. Spatio-temporal outliers are the anomalous points which do not conform to the behavior of other points in the neighborhood in terms of the distance between them. A spatial outlier is the object which has a high number of spatio-temporal outliers associated with it. Thus, it is abnormal as compared to other spatial objects in the neighborhood in terms of the high number of spatio-temporal outliers it is generating. In addition, there may be certain spatio-temporal outliers which are from different micro neighborhoods within the macro neighborhood but are separated by a small time lag, forming a spatio-temporally coalesced outlier traversing through the region.

We next describe the individual steps of our approach in details.

4.1 Generation of micro neighborhood

Traditionally, a spatial object has been associated with the spatial coordinates and non-spatial attributes (for example, toxicity level measured by a sensor). In addition, there may be features present in the vicinity of these objects (such as a landfill near the sensor), whose impact on the spatial object has not been captured, but really is critical to their behavior reflected in the attributes. We want to create a neighborhood around each object, which we call as *micro neighborhood*, that considers both of the attributes associated with the object and features in the vicinity of a spatial object. Our micro neighborhood definition is based on the concept of Voronoi polygons (Aurenhammer 1991; Okabe et al. 2000).

4.1.1 Features

Let us first consider the features in the vicinity of the object. The Voronoi polygons exhibit an important property such that each feature in a Voronoi polygon is associated with the object in that polygon implicitly as its neighborhood. Using this property we can associate a set of features with each spatial object s_i.

Figure 4 shows a simple example of the Voronoi polygons formed around three sensors A, B, C placed in a cross section of the river. As shown in the figure, every feature q (such as a chemical factory) can be represented by a geometric shape (point, circle or polygon). Now, the feature can be associated with the corresponding (possibly multiple) micro neighborhoods simply by finding out the polygons with which it intersects. If at least one part of feature q lies in some polygon $V(S_i)$ corresponding to object s_i, we can associate q with s_i and not s_j, i.e., for a feature q we say $d(q, s_i) \leq d(q, s_j)$, where d is the Euclidean distance function. For instance in Fig. 4 the feature q, a factory, is close to B and feature z, a railway line, is close to B and

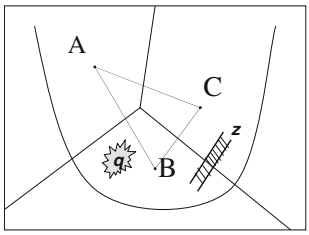

Fig. 4 Sensors in the cross section of a river

C. In essence a Voronoi polygon creates a region of dominance of one object over the other in terms of the membership of features in the polygon.

4.1.2 Attributes

Now let us consider the attributes $a_i = \{a_{i1}, \ldots, a_{im}\}$ of each s_i. There are two types of attributes associated with a spatial object **Characterizing Attributes and Time Variant Attributes**. Characterizing attributes capture the general characteristics of the spatial location where the object is placed. For example, a spatial object can be a water monitoring sensor characterized by certain attributes that capture the underlying natural or man made spatial processes such as mercury content, presence of landfill etc. These attributes are transformed into categorical values (binary) to form a feature vector $f_i = \{f_{i1}, \ldots, f_{im}\}$. If the attributes are continuous numerical values then they are converted to binary values of 1, 0. One way to do this is to check if the average value of the attribute is above a threshold value, it is assigned the value 1 else 0. This conversion may possibly lead to some information loss, however, we consider several such attributes and not relying on one of the attributes.

In addition to the above attributes each object may generate a set of time variant attributes which we call as point set for distinction. These attributes measure a phenomenon over a period of time. For instance, a set of readings measured by each sensor at the specific location. We utilize this information to capture the distinct behavior of the neighborhoods not only in terms of the spatial characteristics but also in terms of its behavior over time.

We now define micro neighborhood this knowledge associated with the spatial object.

Definition 2 (*Micro neighborhood* m_i) Given a set of spatial objects $S = \{s_1, \ldots, s_n\}$, a micro neighborhood m_i is the polygon bounded by a Voronoi polygon $V(s_i)$, representing the region of dominance for each s_i such that each s_i is associated with a feature vector $f_i = \{f_{i1}, \ldots, f_{im}\}$ and a point set $p_i = \{p_{i1}, \ldots, p_{it}\}$.

Here the point set is the points directly generated by the spatial object, for example in the case of water monitoring sensor these points are the readings generated by the sensor and the feature set such as chemical factory, a river, a rail track, pH level, oxygen level etc. Each of these would be associated with the sensors they affect. Together, these would form the characterization for that micro neighborhood.

The identification of the micro neighborhood based on Voronoi diagrams is used as an initial step to neighborhood formation. In order to identify the extended

neighborhood for spatial objects, to capture both spatial autocorrelation and heterogeneity, we need to consider the spatial and inferential relationships among the micro neighborhood. This issue is addressed in following sections.

4.2 Identification of spatial relationships

We first identify spatial relationships between the micro neighborhoods to account for autocorrelation. The relationships could be topological, distance or direction based. A spatial relationship can be formally defined as follows (Ester et al. 1997):

Definition 3 (*Spatial relationship*) A spatial relationship $\text{sp}(m_i, m_j)$ exists between two micro neighborhoods m_i and m_j for the corresponding spatial objects s_i and s_j if there exists either a topological, direction or distance relationship between m_i and m_j, determined using the spatial coordinates of the spatial objects.

A combination of two or more of these relationships forms a complex spatial relationship. Here we limit our discussion to only one type of relationship—the topological relationship of adjacency. Two micro neighborhoods are adjacent if they share an edge. If two Voronoi polygons share an edge they are adjacent to each other. The Triangle-2D mesh generator (Shewchuk 1996), which we utilize to generate the Voronoi polygons, generates an edge list for the Delaunay triangulation, as a precursor to the polygon generation, which identifies the adjacent objects.

The adjacency relationship can be extended to consider other complex relationships, such as a combination of adjacency and direction (north, south, north east etc.) relationship or distance and adjacency. Spatial relationships identify the proximity structure of the neighborhood for accommodating autocorrelation only; we next consider the inferential relationships for accounting of both autocorrelation and heterogeneity in combination.

4.3 Identification of inferential relationships

We extend the definition of spatial neighborhoods beyond spatial relationships based, to include inferential relationships as well. In order to account for spatial heterogeneity, we identify inferential relationships on the basis of two properties of the neighborhoods. First, remember that each micro neighborhood is characterized by the features such as a factory, bridge, railroad, stream, certain type of vegetation, etc. within the neighborhood. Such information can be accumulated with the help of domain experts. Indeed, in many cases such studies precede sensor placement. With the NASQAN (2002) data used to test our approach, the "qualities of the region" determine the placement of the sensors. Secondly, since each object captures time variant attributes such as readings from a sensor, each neighborhood also has an associated point set which can essentially be treated as a cluster of data points. Both of these can be used to identify inferential relationships.

We measure similarities, across the micro neighborhoods, between feature sets using the Jaccard coefficient and between point sets using the Silhouette coefficient

(computation of the coefficients explained in Sect. 3). This facilitates the quantification of the heterogeneity in the neighborhood resulting from the impact of the various features and reflected in the various attribute values. We next discuss this in detail.

The feature set f_i of a micro neighborhood m_i is a binary feature vector over the global list of features. When computing the similarity of two feature vectors, we are more interested in the similarity of features (1–1 match) than a non-similarity of features (0–0) match, unlike the matching or m-coefficient, which gives equal importance to both. Therefore, we use the Jaccard Coefficient (JC) to formalize the similarity in terms of features associated with each spatial object.

Silhouette Coefficient (SC) has been used in the literature to determine the degree of overlap of clusters. We use this characteristic of SC to quantify the overlap in the micro neighborhoods in terms of the point sets, i.e., readings of the associated sensors. If we consider the micro neighborhoods to be clusters, then based on the equation outlined in Sect. 3, the range of the Silhouette coefficient is $[-1, 1]$. Let us say we have two micro neighborhoods A and B a point x in A can have an SC interpreted as follows (Kaufman and Rousseeuw 1990; Ng and Han 1994):

- $\mathrm{SC}(x) = -1$ denotes highly overlapping structure, that is x on average is closer to members of micro neighborhood B.
- $\mathrm{SC}(x) = 0$, x is equally similar to micro neighborhoods A and B.
- $\mathrm{SC}(x) = 1$ indicates a good assignment of x to its micro neighborhood A.

The SC value for every point in a micro neighborhood is compared to the threshold to determine if a positive vote exists for merging the two micro neighborhoods. If a majority of points vote positively (show a strong affinity), the two micro neighborhoods should be merged. We now define a inferential relationship in terms of JC and SC.

Definition 4 (*Inferential relationship*) Given two micro neighborhoods m_i and m_j, and the corresponding feature vectors f_i and f_j, and the point sets p_i and p_j, there exists a inferential relationship $\mathrm{sm}(m_i, m_j)$ between m_i and m_j, iff $\mathrm{JC}(f_i, f_j) \geq \mathrm{th_{jc}}$ and $\mathrm{SC}(p_i, p_j) \geq \mathrm{th_{sc}}$.

Here $\mathrm{th_{jc}}$ is user defined threshold for JC and $\mathrm{th_{sc}}$ is the user defined threshold for SC.

4.4 Macro neighborhood formation

We now define the macro neighborhood based on spatial and inferential relationships between micro neighborhoods.

Definition 5 (*Macro neighborhood*) Given two micro neighborhoods m_i, and m_j, we say that they are part of a macro neighborhood M_i if there exists $\mathrm{sp}(m_i, m_j)$ and $\mathrm{sm}(m_i, m_j)$.

In the above definition, $\mathrm{sp}(m_i, m_j)$ refers to the spatial relation between the micro neighborhoods m_i and m_j, $\mathrm{sm}(m_i, m_j)$ refers to the inferential relationship between

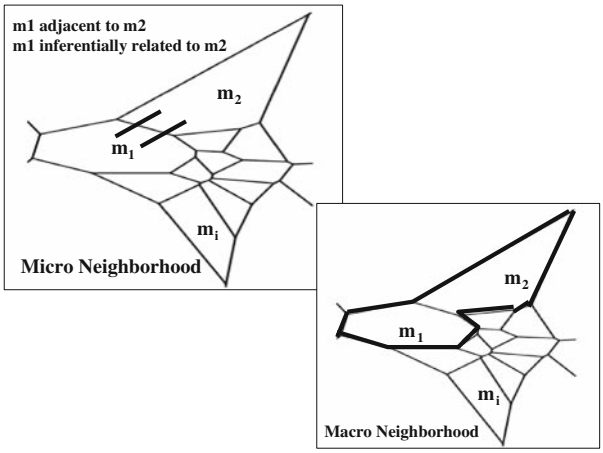

Fig. 5 Macro neighborhood formation

them. Essentially, the contour of the macro neighborhood is formed by eliminating the common edge(s) between the adjacent micro neighborhoods and considering only their outer edges. Note that the macro neighborhood is also a polygon. For example, the two adjacent micro neighborhoods, m_1 and m_2, are merged to form a macro neighborhood M_1, as shown in Fig. 5. We next outline the algorithm to identify the macro neighborhood.

Algorithm 1 gives the complete algorithm for spatial neighborhood formation. The complexity of the algorithm is $O(S^2)$ where S is the number of spatial objects. The algorithm starts by identifying the *micro neighborhood* m_i for each object s_i. The set of *macro neighborhoods* is generated by appropriately merging the micro neighborhoods M. This merging takes place based on the presence of spatial and inferential relationships. Firstly, spatial relationships are identified between every pair of micro neighborhoods. In lines 7–20, we iterate over every pair of objects (s_i, s_j). If s_i and s_j are spatially connected, we then identify if there exists inferential relationships. This is done by computing similarity coefficients for the pair, namely the Jaccard coefficient (JC) and the Silhouette coefficient (SC). Line 12 tests whether the computed coefficients are greater than the thresholds and appropriately combines the two conditions. Note that $\otimes$ implies either the logical AND or the logical OR operator. If the neighborhoods should be merged, lines 14–16 update the global set of neighborhoods (by replacing the two older neighborhood sets with a single merged one). Here the ith macro neighborhood identifier is replaced by the identifier of the jth micro neighborhood. The individual micro neighborhoods for s_i and s_j are removed and the next comparison for jth micro neighborhood will be with the combined s_{ij} neighborhood. Essentially this process is iterative to check the affinity between the various micro neighborhoods. At the end of the loop, the set N contains the merged macro neighborhoods. Our algorithm has the desirable property of order invariance which we discuss next.

Algorithm 1 The neighborhood generation algorithm

Require: Set of objects S with their spatial co-ordinates
Require: A threshold th_{jc} for the Jaccard coefficient
Require: A threshold th_{sc} for the Silhouette coefficient
Require: N, the set of macro neighborhoods, where each micro neighborhood forms its own Macro Neighborhood
Ensure: The set N containing the coalesced macro neighborhoods M
1: **for** each object s_i **do**
2: {Let $m(s_i) = m_i$ represent the neighborhood to which s_i is assigned}
3: Generate the micro neighborhood m_i and assign it to its own set $[m_i]$
4: **end for**
5: {Now generate the set of macro neighborhoods by appropriately merging the micro neighborhoods}
6: Identify spatial relationships between every pair of micro neighborhoods m_i
7: **for** $i = 1 \ldots |S| - 1$ **do**
8: **for** $j = i + 1 \ldots |S|$ **do**
9: **if** $\text{sp}(m(s_i), m(s_j))$==true **then**
10: $\text{jc} = \text{JC}(m(s_i), m(s_j))$
11: $\text{sc} = \text{SC}(m(s_i), m(s_j))$
12: **if** $\text{jc} \geq \text{th}_{\text{jc}} \otimes \text{sc} \geq \text{th}_{\text{sc}}$ **then**
13: {Merge $m(s_i)$ and $m(s_j)$ to form the merged neighborhood M}
14: $M = m(s_i) \bigcup n(s_j)$
15: $\{\forall s_k \in m(s_i) \vee m(s_j)$, associate s_k with the macro neighborhood $M\}$
16: Remove $m(s_i)$ and $m(s_j)$ from N
17: Add M_{ij} to N
18: **end if**
19: **end if**
20: **end for**
21: **end for**
22: The set N now contains the coalesced macro neighborhoods M

4.5 Order invariance in macro neighborhood formation

One nice property of the neighborhood generation algorithm is that it is **order invariant**. Regardless of the order of the micro neighborhoods in the input sequence, the same set of macro neighborhoods is formed. This is critical for a neighborhood based knowledge discovery task such as outlier detection, since the membership of the neighborhood determines how different is the outlier from the neighbors. We formally prove that our approach is indeed order invariant:

Theorem 1 *For a given set of micro neighborhoods, Algorithm 1 creates the same set of macro neighborhoods regardless of the order of the micro neighborhoods in the input sequence.*

Proof The proof is by contradiction. Assume that the macro neighborhoods created are different for some two orderings of the micro neighborhoods. If the macro neighborhoods created are different, there must be some two micro neighborhoods m_p, m_q which were earlier in the same macro neighborhood and are now in different macro neighborhoods or vice versa. Let m_p, m_q be in the positions p_1, q_1 in the first sequence and p_2, q_2 in the second sequence respectively. Note that m_p, m_q can only be merged at line 14 in Algorithm 1. Therefore, in the first case, when $i = \min(p_1, q_1)$ and $j = \max(p_1, q_1)$, since m_p, m_q are in the same macro neighborhood, $\text{sp}(m_p, m_q)$ &&

$(\text{jc} \geq \text{th}_{\text{jc}} \otimes \text{sc} \geq \text{th}_{\text{sc}})$ is true (where $\otimes$ signifies Logical AND or Logical OR). However, when the order is different, in the second case, when $i = \min(p_2, q_2)$ and $j = \max(p_2, q_2)$, m_p and m_q are not merged. This implies that $(\text{sp}(m_p, m_q)$ && $(\text{jc} \geq \text{th}_{\text{jc}} \otimes \text{sc} \geq \text{th}_{\text{sc}})$ is false. However, all the three functions involved—sp, JC and SC, are themselves order invariant. Thus, in either ordering of m_p and m_q, the evaluation should be exactly the same. This leads to a contradiction, thus concluding our proof. □

Note that there is no overlap in the spatial neighborhoods discovered. The micro neighborhoods are iteratively merged relabeling the merged neighborhoods to form the macro neighborhoods.

5 Outlier detection

o An outlier is an object that sufficiently varies from other neighboring objects so that it appears to be generated by a different process from the one governing its neighbors. The current literature defines an outlier and its difference from neighbors in terms of distance, density, etc. We first consider outliers in the point set. For our approach we consider a distance based approach defined by Knorr and Ng (1998):

Definition 6 (*Outlier*) A point O in a dataset T is a DB(p, D) outlier if at least a fraction p of the points in T are at least at distance D from O.

To identify anomalies in sensor datasets, important questions that come to mind are *why do sensors have spikes in their readings, why do they produce anomalous readings, why do they malfunction?*. Sensors sense or monitor certain parameters that they are designed to observe. For instance a water monitoring sensor placed in a water body will monitor various parameters such as toxicity in the water. The sensors gather this data and register a certain pattern in the readings over a period of time across a region. However, due to various reasons the readings may deviate from this pattern. There may be several reasons for this. For instance, weather such as ice formation, sudden temperature fluctuations causing extreme hot or cold temperature, strong impacts of shocks, proximity to other electronic or magnetic devices, toxic dispersed in the vicinity of water monitoring sensor, formation of a coating on sensing surface, aeration or bubbles trapped, debris around sensor.

In these circumstances, the sensor readings may temporarily become unusual as compared to readings measured by the sensors (spatio-temporal context) or as compared to other sensors in the neighborhood (spatial context). Another outcome may be that the sensor will malfunction due to one or more of the above reasons. Lastly it is very much possible that the sensor is reflecting a true phenomenon but appears to be unusual. For instance if there is a toxic spill in the vicinity of the water monitoring sensor, it is actually monitoring a true phenomenon.

Our aim here is to first, discover anomalous readings which are unusual in the spatio-temporal context. Second, we would like to discover if this sensor and all its associated readings are primarily anomalous, i.e., it is a spatial anomaly with respect to other sensors in the neighborhood. One might wonder why this sensor is part of the neighborhood in the first place if it is anomalous with respect to the other sensors

in the neighborhood. This is due to the fact that the neighborhood that we consider is based on not only the behavior of the objects but also the features located around these sensors. Thus, it is very much possible that two sensors may be put in the same neighborhood due to the high similarity in terms of the features in their proximity, but less of a similarity in terms of their readings.

Third, we want to discover whether this is a true anomaly such as a toxic spill. If only the readings of one sensor are anomalous it is highly likely that this is a malfunction as compared to the other normally behaving sensors in the neighborhood. However, if the anomaly is traversing the neighborhood such that there is a small time lag between the anomalous readings between two sensors, then there is a true phenomenon being measured, such as a toxic spill flowing across the sensors.

We would like to discover different types of outliers in the neighborhood, which facilitate the identification of these phenomenon.

- To identify spatio-temporal outliers: sensor readings that are outliers.
- To identify spatial outliers: completely malfunctioning sensors.
- To identify spatio-temporally coalesced outliers: link chains of outlying spatio-temporal sensor readings.

Our identification techniques are based on the macro neighborhood generation which accounts for both spatial proximity and affinity. We now further discuss the three types of outlier detection in such a neighborhood.

5.1 Spatio-temporal outliers

For each spatial object in a macro neighborhood we would like to identify any (time variant) spatio-temporal points, which are outliers with respect to the entire neighborhood. Thus, a spatio-temporal outlier is a point, which behaves differently not only as compared to other points of its own spatial object but also as compared to the spatio-temporal points of all spatial objects in its neighborhood. This leads us to the definition of spatio-temporal outlier.

Definition 7 (*Spatio-temporal outlier*) Given a macro neighborhood M_i and its corresponding micro neighborhoods $\{m_1, \ldots, m_n\}$ where each m_j has a set of spatio-temporal points $p_j = \{p_{j1}, \ldots, p_{jt}\}$, a point p_{jk} is said to be a spatio-temporal outlier if it lies at a greater distance than d from pt points in the macro neighborhood M_i, where d is a user defined distance threshold and pt is the threshold on the number of points, the point p_{jk} should be distant from.

The procedure is simple: once macro neighborhoods are generated, outlier detection technique can be applied to each neighborhood (all the points) to detect the outlying readings. This is expected to give more accurate results (less false positives) due to more accurate neighborhood generation. In this paper, we use Knorr and Ng's (1998) distance based outlier detection technique. This algorithm has been proposed in the context of traditional data mining and has the advantage of being simple and intuitive. Here we consider proximity of points in terms of distance threshold as the determining factor for outlierness, since we have already allocated the spatial objects to their

respective macro neighborhoods. In the outlier detection algorithm we set a threshold pt for the number of points (count) from which a certain point is at a greater than distance d. The threshold pt can be a user input or used as the number of points in the macro neighborhood/2 since a point cannot be at a greater distance than d from more than half of the points in the neighborhood. At the end of this process we have a set of spatio-temporal outliers $O = \{O_1, \ldots, O_n\}$ corresponding to each micro neighborhood in macro neighborhood $M_i = \{m_1, \ldots, m_n\}$ such that each $O_j = \{o_{j1}, \ldots, o_{jm}\}$ consists of individual spatio-temporal outliers for the micro neighborhood m_j. In the context of our water monitoring example these types of outliers correspond to the possibly temporary anomalous readings caused due to various reasons such as seasonal fluctuations, debris, high chemical concentration etc.

The spatio-temporal outlier set O is used for further analysis. Thus, if more than a certain number of points are spatio-temporal outliers for a spatial object then it can be further investigated if the object is an outlier in its entirety, i.e., a spatial outlier. This leads us to the discussion on spatial outlier detection.

5.2 Spatial outliers

There may be a situation that a large number of readings from one sensor are outliers. This would indicate that the sensor itself is malfunctioning, such that most or all of its readings are spatio-temporal outliers, deeming this sensor to be a spatial outlier. While it is possible to argue that an outlier object would not be grouped with other objects, our formation of the neighborhood considers not just the points (readings), but also the spatial features characterizing the region. This would group together objects expected to be similar.

Thus, once the spatio-temporal outliers within a neighborhood are determined, some function G (e.g., count, probability distribution, etc.) can be applied to determine if the entire spatial object is an anomaly. Here we consider the function G to be a count of number of spatio-temporal outliers for each spatial object in a micro neighborhood. We next define a spatial outlier.

Definition 8 (*Spatial outlier*) Given a macro neighborhood M_i and a set of spatio-temporal outliers O_p for each micro neighborhood $m_p \in M_i$, a spatial outlier is an object s_t associated with micro neighborhood $m_t \in M_i$, such that $G(s_t) > \theta$, where G is a count of the spatio-temporal outliers and θ is the user defined threshold value.

We outline an algorithm to detect spatial outliers in Algorithm 2. In the worst case scenario the complexity of the algorithm is $O(M.O)$ where M is the number of macro neighborhoods and O is the number of outliers in each macro neighborhood.

The algorithm takes as an input the macro neighborhood definitions and the spatio-temporal outliers identified for each micro neighborhood. If the number of spatio-temporal outliers ($\text{outlier}_{\text{count}}$) exceeds a user defined threshold δ, then the spatial object associated with it is marked as a spatial outlier. This threshold value can vary from half to all of the spatio-temporal points, since it seems obvious that if half or more of the points are anomalous, then this spatial object should be further investigated as a spatial outlier.

Algorithm 2 Identification of spatial outliers in macro neighborhoods

Require: k MacroNeighborhoods, $M_1, \ldots, M_k$
Require: Spatio-temporal outliers $O, O_1, \ldots, O_n$
for each MacroNeighborhood M_l **do**
  for all microNeighborhoods $m_i \in M_l$ **do**
    for all Outliers $O \in M_l$ **do**
      if $O_x \in m_i$ **then**
        {Increment outlier$_{\text{count}}$ for m_i }
      end if
    end for
    if outlier$_{\text{count}} \geq \theta$ **then**
      {Mark spatial object in microNeighborhood as Spatial Outlier}
    end if
  end for
end for

5.3 Spatio-temporally coalesced outlier

In case of spatio-temporal and spatial outliers it is difficult to identify a process, which could be causing the anomalous behavior. In some cases this may not be sufficient. For instance, in case of water monitoring sensors we not only want to identify anomalous readings or malfunctioning sensors but we also want to identify a situation where an anomaly has occurred due to a physical process such as the flow of a toxic chemical. Thus, it is important to distinguish this case from an anomalous reading or malfunctioning sensor so that additional action can be taken.

Since the points in a spatial neighborhood depict the underlying processes of similarly behaving spatial objects we can evaluate the scenario of a spatial process such as movement of an anomaly across the macro neighborhood. Such an anomaly would result in a subset of micro neighborhoods (in the macro neighborhood) to be linked by virtue of a time lag in the spatio-temporal outliers in them.

We now consider the time lag across the spatio-temporal outliers and define the concept of a spatio-temporally coalesced outlier. If a spatio-temporal outlier o_{1x} occurs in a micro neighborhood m_1 at time t_{1x} and another spatio-temporal outlier o_{2y} occurs in another micro neighborhood m_2 in the macro neighborhood M_i at time $t_{2y} = t_{1x} + \delta$, then a temporal outlier would indicate a link from m_1 to m_2 in terms of the traversal of the anomaly from one spatial object to another, separated by an interval of time δ. This leads to a definition of temporal outliers as follows:

Definition 9 (*Spatio-temporally coalesced outlier*) Given a macro neighborhood M_i and a set of spatio-temporal outliers O_p for each micro neighborhood $m_p \in M_i$, a spatio-temporally coalesced outlier consists of a pair of spatio-temporal outliers o_{ix}, o_{jy} such that the micro neighborhood $m_i \neq m_j$ and $((t_{ix} \geq (t_{jy} + \delta)$ or $(t_{jy} \geq (t_{ix} + \delta)))$.

In the context of our example, once the spatio-temporal outliers have been detected in a macro neighborhood we want to identify readings, which belong to different micro neighborhoods that have the temporal stamp differing by a threshold value δ, implying a possibility of a temporal anomaly propagating in the neighborhood.

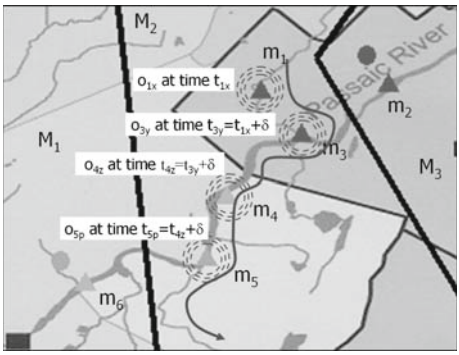

Fig. 6 Detecting a temporal anomaly as a link chain in the macro neighborhood

Figure 6 depicts sensors placed in the Passaic river in New Jersey. Let us assume we identified three macro neighborhoods M_1, M_2 and M_3. In the macro neighborhood M_2 there are 4 micro neighborhoods originating from spatial objects represented as triangles namely m_1, m_3, m_4 and m_5. The curved line across M_2 indicates the possible spatio-temporally coalesced anomaly traversing through the macro neighborhood, which is formed of a time lag between the pair of micro neighborhoods, such that there exists a temporal relationship between o_{1x} and o_{3y}, o_{3y} and o_{4z}, etc. Intuitively it can be seen that larger macro neighborhoods may possibly have longer chains, which implies that lower JC and SC would possibly lead to longer anomalies. We later attempt to demonstrate this with experimental results. We outline our approach to detect spatio-temporally coalesced outliers in Algorithm 3. The complexity of the algorithm is $O(n^2)$ where n is the number of spatio-temporal outliers.

The input to the algorithm is the macro neighborhood definitions and the spatio-temporal outliers identified. For each macro neighborhood if the spatio-temporal outliers belong to different micro neighborhoods and differ by a threshold value in time, then the two spatio-temporal outliers form a spatio-temporally coalesced outlier.

Properties of the proposed algorithms: For spatio-temporal outlier detection, once neighborhoods are identified, we simply apply Knorr and Ng's (1998) outlier detection procedure to identify outliers among all of the readings. Instead of Knorr and

Algorithm 3 Identification of temporal outliers in macro neighborhoods

Require: k MacroNeighborhoods, $M_1, \ldots, M_k$
Require: O Outliers, $O_1, \ldots, O_n$
for each MacroNeighborhood M_l **do**
  for every Outlier O_{ix} **do**
    for every other Outlier O_{jy} **do**
      if $m_i \neq m_j$ **then**
        if $t_{ix} \geq (t_{jy} + \delta)$ or $t_{jy} \geq (t_{ix} + \delta)$ **then**
          {Mark O_{ix} O_{jy} as temporal outlier}
        end if
      end if
    end for
  end for
end for

Ng's procedure, we could equally easily use any other outlier detection technique such as density based outlier detection (Ester et al. 1996). This gives us the same completeness and correctness guarantees provided by the specific outlier detection technique. Furthermore, assuming that the neighborhoods are optimally (correctly) identified, we can show that this allows us to have the highest accuracy possible in identifying outliers. This follows from the following reasoning: Given that outlierness is indicative of abnormality (i.e., outlier readings are uncommon) if a reading is a true outlier, even though it may locally look normal, a set of readings must exist among which it stands out. If the neighborhood were optimally identified, the set of readings in the neighborhood correspond to this set, and therefore the outlier reading should be appropriately identified. Similarly, if a reading is not an outlier in the global sense, again the appropriate neighborhood must exist among which it is normal, even if locally it shows up as an abnormal reading. Assuming that the neighborhood has been correctly chosen, the reading should now appear as normal. Our experiments bear this out, to the extent that the neighborhoods are chosen correctly. Typically a highly well defined neighborhood should show a higher degree of autocorrelation which may be measured using a neighborhood quality metric such as within neighborhood sum of squared errors (McGuire et al. 2008) or using a standard measure of spatial autocorrelation such as the Moran's I Statistic (Moran 1948). Thus, proper neighborhoods can be chosen, allowing for the most effective detection of spatio-temporal outliers.

Our algorithm for identifying spatial outliers simply checks to see whether more than a certain threshold of readings have been identified as outliers in the previous phase. Again, we can show that if the neighborhood has been appropriately identified, this correctly identifies all spatial outliers. The key is to note that a spatial outlier can be identified as long as all spatio-temporal outliers have been correctly identified. Since this is correctly done based on the neighborhoods, spatial outliers can also be identified, assuming that the threshold is set correctly.

Finally, assuming the macro neighborhood is optimally identified, spatio-temporally coalesced outliers can also be correctly identified, as long as the threshold value δ is correctly set.

We next discuss experimental results including datasets used, empirical results, setup for the three types of outlier detection techniques and prototype.

6 Experimental results

In this section we discuss the empirical performance of our proposed approach. The main questions we are interested in answering are: (1) *How well are we able to characterize similarly behaving objects into one neighborhood based on spatial and inferential relationships among objects?* (2) *How well are we able to identify outliers (spatio-temporal, spatial and temporal outliers) in such a spatial neighborhood?* We present various results to depict the answers based on our approach.

Datasets: We have performed experiments with datasets including the highway traffic monitoring dataset (Shekhar et al. 2001) and the water monitoring dataset (NASQAN 2002). Figure 7 gives a tabular overview of the datasets used along with their associated information. We now describe the datasets in detail.

Dataset	# of sensors	Pointset Number of readings **X** Attributes measured for each reading	Size of Feature Vector
Highway Sensors	60	189000 X 3	3
Water Monitoring – 7 Sensors	7	122 X 30	21
Water Monitoring- 40 Sensors	37	15000 X 123	10

Fig. 7 Datasets used for experimental analysis

Highway traffic monitoring dataset This is a large real world data set from the Minnesota Department of Transportation (Shekhar et al. 2001). The main attributes in the data are the time slots of 5 min during the day, volume and occupancy readings, for the sensor for that time slot. Each sensor is also associated with spatial location in the form of latitude and longitude. We created the feature vectors based on attributes such as highway name, direction of traffic flow and clustering of sensors. For our discussion in this paper we discuss the results obtained in data consisting of 60 sensors along the interstate I-35 W, North Bound and South Bound. For the purpose of validation, we also augmented the dataset with some known outliers, i.e., some points were made to exhibit an extreme behavior in terms of their values. This dataset is used for validation of outlier detection using our approach, as it is also used by graph based spatial outlier detection approach (Shekhar et al. 2001).

Water monitoring dataset: The USGS program "National Stream Quality Accounting Network" (NASQAN 2002) program is currently focused on monitoring the water quality of the nation's largest rivers—the Mississippi (including the Missouri and Ohio), the Columbia, the Colorado, and the Rio Grande rivers. NASQAN operates a network of ~679 water monitoring sensors where the concentration of chemicals, including pesticides and trace elements, is measured along with stream discharge. The EPA utilizes certain features [http://water.usgs.gov/-nasqan/progdocs/statables.html] to select the sensor locations for monitoring. We use these attributes to form our feature vectors, resulting in a feature vector of 21 features. Some of the features are: mean discharge (ft^3/s), incremental increase in drainage area (mi^2), incremental increase in stream flow (ft^3/s), drainage area (KM^2), contributing drainage area (mi^2), % urban, % forest, % cropland, % mixed crop and natural features, % grassland, population density per square mile.

The water monitoring sensor readings used for outlier detection consists of spatial attributes of latitude and longitude, which is useful in determining the spatial relationship. It consists of the temporal attributes of date and time of sampling and over 100 water-monitoring attributes. These attributes include: mean daily stream flow, temperature, specific conductance, dissolved oxygen, pH, alkalinity, suspended sediment, ammonia nitrogen, nitrite nitrogen, organic nitrogen plus ammonia nitrogen (filtered), organic nitrogen plus ammonia nitrogen (whole-water), nitrite plus nitrate, total phosphorus (whole-water), etc.

We discuss the results of our evaluation of two subsets of this data, one including 7 sensors and the other including 37 sensors, due to the size of the data we discuss in details the results obtained in 7 sensors and briefly outline results obtained in 37 sensors dataset.

We next discuss various aspects of our experimental results including the following:

A. Neighborhood formation: The first step of the macro neighborhood generation algorithm requires the identification of spatial relationships among the nodes (sensors). This is identified by applying the program TRIANGLE (Shewchuk 1996), which generates an edge for each of the two nodes that are adjacent to each other. This is the starting point of identifying neighbors based on spatial relationships. In addition we compute the JC, SC coefficients for the micro neighborhoods. Our aim is to show incrementally how these coefficients improve on the results obtained in the traditional approaches, which are discussed later. We study the effect of macro neighborhood formation by combining the spatial relationships with:(a) JC, (b) SC and (c) JC (AND/OR) SC
B. Outlier Discovery: We discuss the effect of the varying neighborhood formation, based on above criteria, on the discovery of spatio-temporal, spatial and spatio-temporally coalesced outliers. Specifically we study outlier detection in the neighborhoods based on spatial relationships and (a) JC, (b) SC and (c) JC (AND/OR) SC
C. Comparison with other approaches: In order to discuss the improvements that accounting for heterogeneity will bring about in our approach, we first lay out some results with traditional approaches which give us a context for the results we obtained in our approach. Specifically, we discuss the results for

- Comparison of our approach for neighborhood formation with traditional cardinality based approach (Ester et al. 1999; Shekhar et al. 2001).
- Comparison of our approach with Graph based spatial anomaly detection technique (Shekhar et al. 2001).
- Comparison with a Density based approach (Birant and Kut 2006).
- Comparison with traditional control charts (Chatfield 1983).

D. Additional results: In addition we also discuss some other results of our approach including:

- Comparison of Precision
- Selection of SC and JC thresholds
- Summary of results with larger datasets
- Prototype

6.1 Neighborhood formation

We discuss the various results in the *Water Monitoring dataset*. For better understanding we first consider a smaller subset of 7 sensors. The adjacency graph of these sensors is shown in Fig. 8a indicating a purely spatial relationship based neighborhood.

6.1.1 Combining the spatial relationships with JC

Figure 8 shows the JC threshold and the corresponding macro neighborhood formed consisting of the micro neighborhood for sensors labeled with identifiers 1–7. The

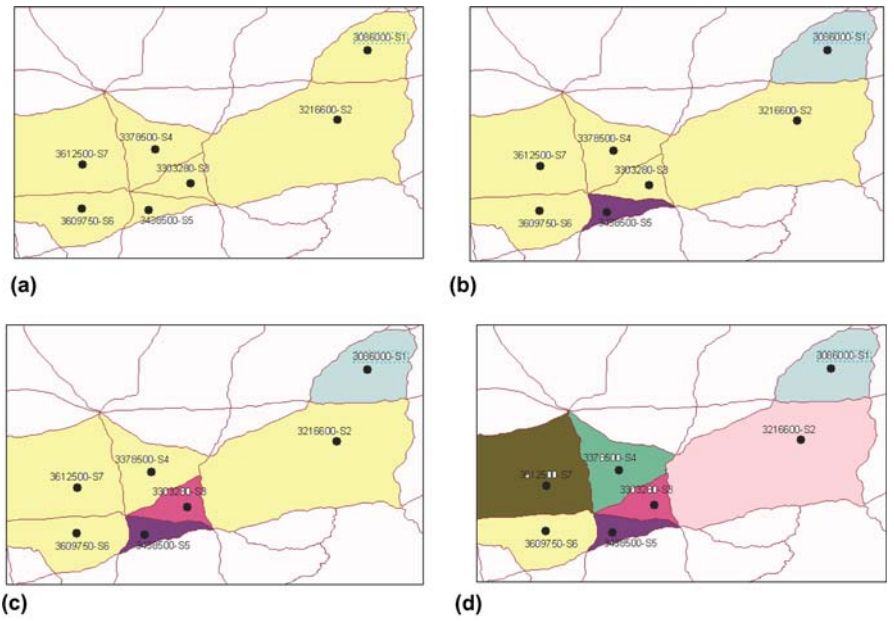

Fig. 8 Neighborhood formation with varying JC thresholds. (**a**) Spatial relationship and $JC_{th} = 0$, (**b**) spatial relationship and $JC_{th} = 0.1-0.3$, (**c**) spatial relationship and $JC_{th} = 0.5$, (**d**) spatial relationship and $JC_{th} = 0.6-1.0$

varying shades in the figure indicate a demarcation of macro neighborhoods within the entire neighborhood formation. As shown in the Fig. 8b, c, d by adding the condition on JC along with the spatial relationships the macro neighborhood is not clumped into one big region (as shown in Fig. 8a) and we end up with multiple neighborhoods.

The results indicate that the neighborhood generated with a smaller value of JC threshold builds on the neighborhood generated with a more restrictive, larger value of, JC threshold and vice versa. For example, one of the macro neighborhoods generated at JC threshold of 0.5 (Fig. 8c) consists of the micro neighborhoods for sensors 2, 4, 6, 7 and the macro neighborhood generated at JC threshold of 0.1–0.3 (Fig. 8b) consists of the micro neighborhoods for sensors 2, 3, 4, 6, 7. With the change in the threshold value at from 0.2 to 0.5, 2, 4, 6, 7 are still grouped together and 3 is separated out. Thus, the neighborhood shows the incremental merging on the basis of less restrictive threshold value of JC threshold and incremental distilling on the basis of more restrictive value of JC.

6.1.2 Combining the spatial relationships with SC

We next investigated the sensitivity of the formation of macro neighborhood to the threshold value of SC. As shown in Fig. 9b, c, d by adding SC along with the spatial relationships the neighborhood is not clumped into one big region. As shown in Fig. 9, as the threshold value of SC increases the neighborhood gets more refined. With a SC threshold value of 0.91 the neighborhood is more granular than a threshold value of

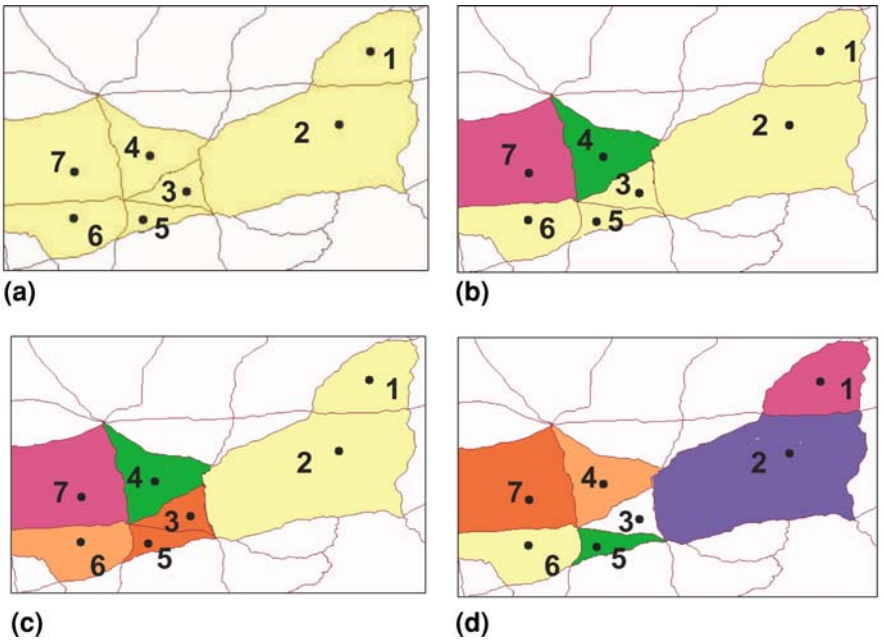

Fig. 9 Neighborhood formation with varying SC thresholds. (**a**) Spatial relationship and $SC_{th} = 0-0.7$, (**b**) spatial relationship and $SC_{th} = 0.9$, (**c**) spatial relationship and $SC_{th} = 0.91$, (**d**) spatial relationship and $SC_{th} = 1.0$

0.7. This is mainly because fewer micro neighborhoods will have such a high level of overlap and thus they are not merged into one macro neighborhood. In general it can be seen that the neighborhood is clumped together if we do not restrict the overlap (reflected in low SC) as shown in Fig. 9a. Basically we allow neighborhoods to be merged if overlap is low. Similarly as we begin restricting the overlap (SC threshold of 1 indicating the highest degree of overlap) the merging of neighborhood is reduced and we start seeing more segregated neighborhoods as shown in Fig. 9b, c, d. However, as compared to JC, SC needs more fine tuning as we are dependent on the wide ranging readings of each sensor.

6.1.3 Combining the spatial relationships with JC AND/OR SC

When JC, SC are used in combination the results are further refined into smaller neighborhoods as can be seen in Figs. 10 and 11. We discuss the effect of varying both JC and SC thresholds for neighborhood formation using the logical AND, OR operators.

A very low JC AND SC, leads to micro neighborhoods clumped together into one macro neighborhood, as SC AND JC are increased the neighborhood becomes more refined as shown in Figs. 10a and 11a.

When we combine the two coefficients, we need to fine tune them much more as to compared to when used alone. Also it was seen that SC had much more sensitivity as compared to JC. In general we observed that AND produces better results in identify-

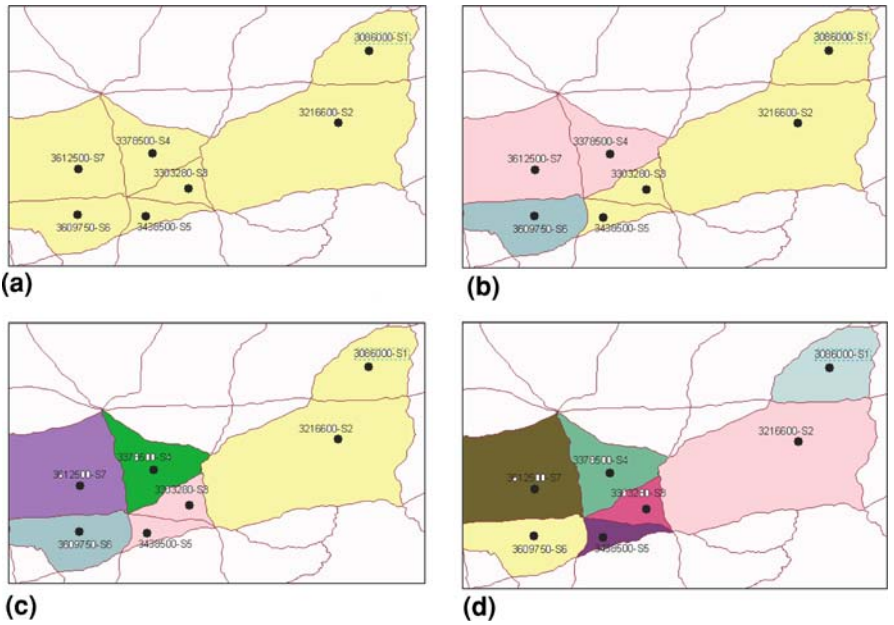

Fig. 10 Neighborhood formation with SC OR JC. (**a**) Spatial relationship $JC_{th} = 0-0.5$, $SC_{th} = 0-0.5$, (**b**) spatial relationship, $JC_{th} = 0.555$, $SC_{th} = 0.9073$, (**c**) spatial relationship $JCt_{th} = 0.6$, $SC_{th} = 0.908$, (**d**) spatial relationship $JC_{th} = 0.8-1.0$, $SC_{th} = 0.92-1$

ing the neighborhoods as will be illustrated with results in outlier discovery in these neighborhoods.

We next discuss the outlier discovery in each of these neighborhoods.

6.2 Outlier discovery

We now follow with a discussion of some observations pertaining to the results obtained for *spatio-temporal outliers* (*ST*), *Spatial Outliers* (*SO*) *and Spatio-Temporally Coalesced Outliers* (*STC*) in neighborhood formed using the two coefficients.

6.2.1 In neighborhood based on spatial relationships and JC

- At low threshold value of JC, the number of ST outliers is high and at high threshold value of JC, the number of ST outliers is less. This is mainly because the neighborhood becomes more refined at higher values of JC threshold.
- ST Outliers detected at high threshold value of JC are a subset of those detected at low threshold value of JC. The process systematically eliminates outliers, which do not conform to the neighborhood. Figure 12a shows the total true ST outliers, false positives identified and the true ST outliers identified at various setting of JC. The distance is set to the value d*s* obtained from the outcomes of spatial relationships

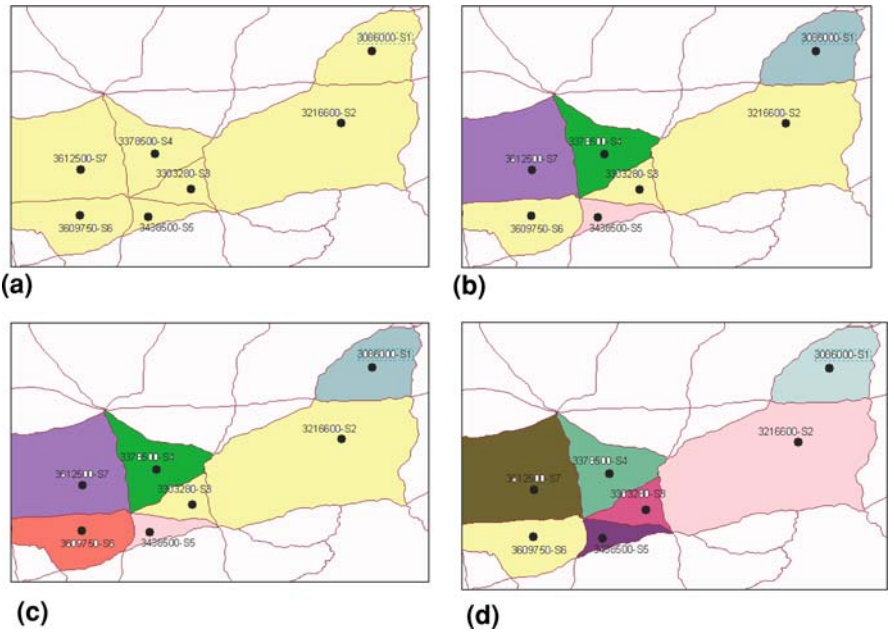

Fig. 11 Neighborhood formation with SC AND JC. (**a**) Spatial relationship $JC_{th} = 0$, $SC_{th} = 0-0.5$, (**b**) spatial relationship, $JC_{th} = 0.1$, $SC_{th} = 0.89$, (**c**) spatial relationship $JC_{th} = 0.5$, $SC_{th} = 0.9$, (**d**) spatial relationship $JC_{th} = 0.8-1.0$, $SC_{th} = 0.92-1$

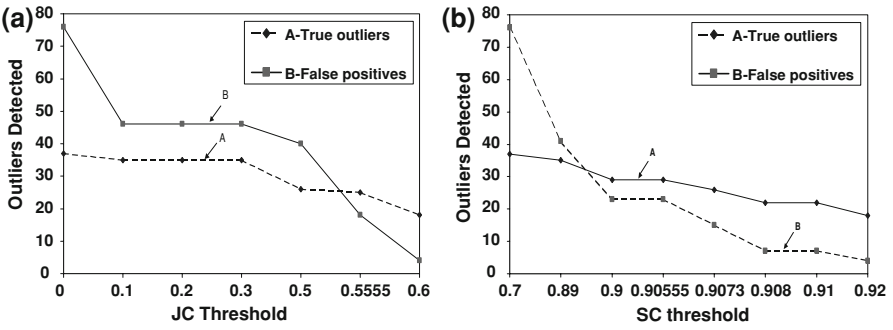

Fig. 12 Results in spatio-temporal outlier detection with varying SC, JC thresholds. (**a**) Number of spatio-temporal outliers, false positives detected with spatial relationship and JC. (**b**) Number of spatio-temporal outliers, false positives detected with spatial relationship and SC

based neighborhood definition as will be discussed in later sections. We can see that as JC becomes more refined the false positives identified become less and the true ST outliers detected become constant after a certain setting of JC.

– It is observed that if any micro neighborhood has no ST outliers at 0.2 threshold of JC, the same will be true, for this macro neighborhood, at other threshold values such as 0.5 or 0.8. Thus, consistency is not compromised in the process of outlier detection. Overall it can be seen that at various distance thresholds the

false positives are reduced in a neighborhood defined using JC along with spatial relationships.
- Spatial outlier detection deals with identifying sensors, that have over a certain threshold, which we call as spatial threshold S_{th}, of their readings as outliers. This threshold could be varied between 30% of the readings to over 80% of the readings. It was found that with JC = 0.5 and $S_{\text{th}} = 50\%$ two sensors 2 and 7 are labeled as spatial outliers. Moreover with JC = 0.8 and $S_{\text{th}} = 50\%$ one sensor 7 is labeled as a spatial outlier. Indeed if we further evaluate the outlier detection we can see the two sensors in the case of JC = 0.5 behaving anomalously.
- For lower values of JC where a bigger macro neighborhood is formed, STC outliers are identified with a consistency. For instance STC anomaly is discovered as follows for JC = 0.2. These conform to the neighborhood formation for JC = 0.21. Sensor 3 on 12/11/2001 at 1200 is coalesced with 4 on 12/11/2001 at 1240 2. Sensor 3 on 3/6/2002 at 1330 is coalesced with 7 on 3/6/2002 at 1400.

6.2.2 *In neighborhood based on spatial relationships and SC*

- Higher the threshold value of SC, lesser is the number of outlier detected. Figure 12b shows the false positives identified and true ST outliers identified by considering SC along with spatial relationship at fixed ds. Here we can see that the true outliers, false positives detected are lower at a more refined level of SC (varied from 1.0 to 0). At SC < 0.7 the true ST outliers detected becomes constant. This indicates least overlap and thus each sensor is kept in its own macro neighborhood and has the most refined ST outlier detection.
- For all variations of SC the results for SO are consistent with ones obtained with JC. Here also the sensor 7 is identified as a spatial outlier. However, the results with SC are more robust such that even for smaller thresholds of SC the same sensor 7 is identified as a spatial outlier. This is consistent for all variations of S_{th} and SC together.
- For lower values of SC threshold where a bigger macro neighborhood is formed, STC outliers are identified with a consistency. For instance the STC anomaly is discovered as follows for SC of 0.5. 1. Sensor 2 on 3/7/2002 at 1300 coalesced with 6 on 3/7/2002 at 1330.

6.2.3 *In neighborhood based on spatial relationships and SC AND/OR SC*

- At a fixed distance threshold ds, the false positives for ST outlier detection with JC AND SC can be seen in the graph in Fig. 13b and similarly the false positives for JC OR SC can be seen in the graph in Fig. 13a. A very low JC AND SC, leads to micro neighborhoods clumped together into one macro neighborhood and number of ST outliers is high, as SC AND JC is increased the neighborhood becomes more refined and number of ST outliers are also refined or reduced. This refinement in outlier detection is also reflected in the true outliers and false positives detected. We saw that the logical AND operator performs slightly better in terms of reducing false positives as compared to OR.

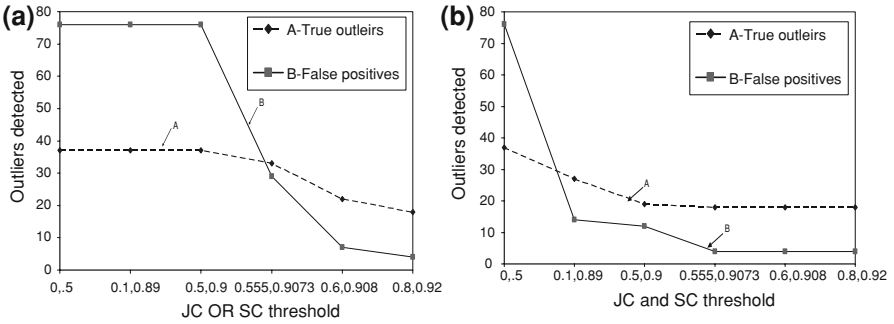

Fig. 13 Results in spatio-temporal outlier detection with varying thresholds of SC, JC versus spatial relationship based neighborhood. (**a**) Number of spatio-temporal outliers, false positives detected with spatial relationship versus SC OR JC. (**b**) Number of spatio-temporal outliers, false positives detected with spatial relationship versus SC AND JC

– The refinement in ST outliers does not compromise the consistency in the results, such that if a certain micro neighborhood does not contribute to the outlier set in one case, it does not do so in another case as well. The ST outliers in the refined case (e.g., SC = 0.92, JC = 0.6) are a subset of ST outliers in the broader case (e.g., SC = .9, JC = 0.5). On using both the coefficients with AND operator the resulting SO is an intersection of the individual values of JC, SC thresholds. Essentially the variation in JC is captured with the results of SC. It was observed that the sensors discovered while varying the coefficients individually are seen here as well, thus it is an additive result where sensors 2 and 7 are identified as spatial outliers.
– STC outliers discovered by varying JC, SC together are similar to the other two cases discussed above. For instance 1. Sensor 3 on 12/11/2001 at 1200 is coalesced with 4 on 12/11/2001 at 1240. Similarly 2. Sensor 3 on 3/6/2002 at 1330 coalesced with 7 on 3/6/2002 at 1400.

6.3 Comparison with other approaches

We next discuss some results with traditional approaches which give us a better understanding of how we perform in comparison to these approaches.

6.3.1 Comparison of neighborhood formation with traditional cardinality based approach

The neighborhood formed with pure spatial relationship is shown in Figs. 8a and 9a when JC = 0 and SC = 0. As is evident from this neighborhood graph, there is a high level of connectivity among the nodes simply based on adjacency. Thus, identifying the neighborhood based only on the spatial relationships would result in one big neighborhood. If we consider cardinality based neighborhood (Ester et al. 1999; Shekhar et al. 2001) we can have several possible neighborhood formations, since it is not order invariant. Moreover determination of cardinality is based on user's choice similar to our coefficient threshold choice. Let us say, for example, we consider 10 sensors whose

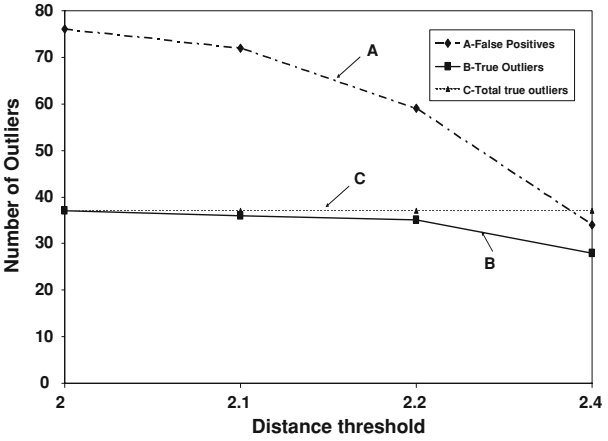

Fig. 14 Spatio-temporal outlier detection: false positives and true outliers detected with neighborhood based on purely spatial relationships

data has two distributions α and β. Now if our cardinality is 6 one of the sensors from the alternate distribution may show up as an outlier. Similarly if cardinality is 5 or <5 the sensors from alternate distributions may be grouped together, causing one or more of them to be detected as an outlier. This behavior is reflected in spatio temporal outlier detection as shown in Fig. 14. Here the false positives are shown as compared to the true outliers in the data.

We also compared the false positives obtained using a traditional neighborhood vs the neighborhood created with JC, SC and both of them in combination with the AND, OR operators.

It can be seen that at various distance thresholds the false positives are reduced in a neighborhood defined using JC along with spatial relationships versus in a neighborhood defined based on purely spatial relationships as shown in Fig. 15a. Similarly, it was observed that at various distance thresholds the false positives are reduced in a neighborhood defined using SC along with spatial relationships versus in a neighborhood defined based on purely spatial relationships as shown in Fig. 15b.

It can be observed in Figs. 15 and 16a, that the false positives are reduced as compared to just using spatial relationships in forming the neighborhood. However, the logical operator AND performs better in terms of false positives as compared to OR. The false positives with varying distance thresholds and keeping JC threshold set at 0.1 and SC threshold set at 0.89 are shown in Fig. 16a. Here also we can see that JC AND SC produces the lowest number of false positives.

6.3.2 Comparison of our approach with graph based spatial anomaly detection technique (Shekhar et al. 2001)

6.3.2.1 Highway traffic monitoring dataset We illustrate a comparison of our approach with the graph based spatial outlier detection approach (Shekhar et al. 2001). This comparison is in terms of the detection of the spatio-temporal outliers. For any

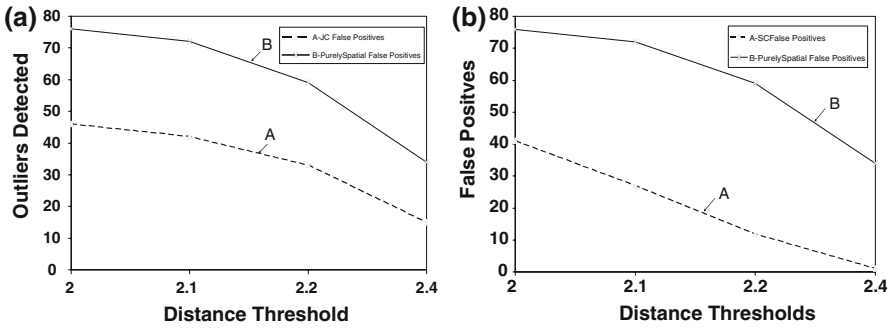

Fig. 15 Comparison of false positives. (**a**) Comparison of false positives detected with spatial relationships and JC versus purely spatial relationship. (**b**) Comparison of false positives detected with spatial relationships and SC versus purely spatial relationship

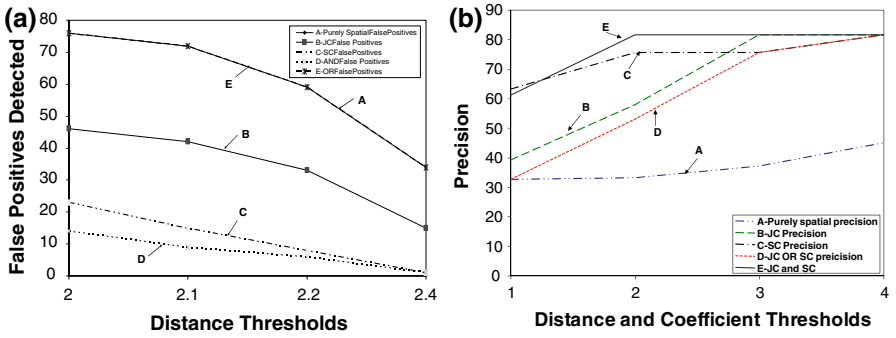

Fig. 16 Comparison of various cases in spatio-temporal outlier detection. (**a**) Comparison of false positives detected with spatial relationships and SC, JC versus purely spatial relationship. (**b**) Comparison of precision for various cases

given neighborhood formation we evaluate the outlier detection. It is observed that as JC is reduced, i.e., the threshold for merging similar micro neighborhood is low, the number of different types of outliers increase and more number of micro neighborhoods are merged, and vice versa. Also consistency across various JC thresholds is maintained for example: sensor 60 forms a neighborhood of itself and has 153 outliers. This is true across the different threshold values of JC.

In addition, Our results for outlier detection confirmed results obtained in Shekhar et al. (2001) as follows. The results show a temporal outlier in sensors 29 and 30 from the time 9:30 to 10:15 a.m. Moreover the anomaly is also observed for 2:30 p.m. for the two sensors. Further if we see sensors 31–34, it shows an anomaly with a gap of about 10 min. This could probably lead to detection of the progression of the anomaly, which in this case may be due to an accident, which leads to congestion along the highway.

Lastly, for the purpose of validating our approach, 13 spatio-temporal outlier cases were randomly dispersed throughout the data, which consisted of 108,900 readings. Each one of the 13 spatio-temporal outliers was detected in the outlier set along with the other spatio-temporal outliers. Although the approach (Shekhar et al. 2001)

discusses the cardinality of the neighborhood and the depth of the neighborhood, it does not explicitly discuss the membership of the highway monitoring sensors in the respective neighborhoods, thus further validation for the neighborhood identification was not possible for this dataset.

6.3.3 Comparison with a density based approach (Birant and Kut 2006)

We consider a representation of this approach since we were unable to acquire the exact code from the authors. the approach essentially discovers the spatial neighborhoods using a spatial distance threshold eps_1 and then applies a density based method such as DBSCAN (Ester et al. 1996) to find outliers. the density based method in turn requires two parameters namely the radius eps_2 and minimum points within the radius MinPts. to approximate this method we consider the spatial neighborhood based on adjacency of the sensors and then perform density based outlier detection using DBSCAN (Ester et al. 1996). The approach (Birant and Kut 2006) is obviously sensitive to eps_1, eps_2, minpts, however, we remove the need for eps_1 since we consider adjacency based spatial relationship rather than a distance based spatial relationship for our experiment. In this regard we should see improvements in Birant and Kut (2006) since the sensitivity to the eps_1 is eliminated.

We performed experiments in the water monitoring dataset in such a neighborhood which comprised of sensors 3,086,000(1), 3,216,600 (2). This neighborhood is equivalent to a macro neighborhood found with $JC = 0.6$ and $SC = 0.9$. In our experiments we found that with a varying eps_2, MinPts for density based approach, both approaches perform comparably in terms of accuracy. Our accuracy was 0.89 and the density based approach varied from 0.86 to 0.91. However, in terms of precision density based approach varied from 0.5 to 0.66 whereas our approach precision was 0.44. In terms of recall the density based approach had a recall around 0.66 whereas our recall was 1. Thus, in terms of False positives our approach performed worse than the density based approach.

6.3.4 Comparison with traditional control charts (Chatfield 1983)

Most of the time Series approaches mainly consider temporal variation and temporal autocorrelation. However, often times the temporal behavior is governed by the spatial behavior especially where the dataset is spatio temporal and not purely temporal. Therefore we cannot only rely on pure temporal analysis. However, for completion we compare our results with a simple time series technique called as control charts discussed below. We explore two such types of charts for a comparison with the results from our approach specifically (a) Multivariate CUmulative SUM (MCUSUM) chart and Exponentially Weighted Moving Average (MEWMA) chart (Chatfield 1983). In case of MCUSUM chart various charts were generated for each sensor individually which corresponds to a case of $JC = 0.8$. In this case the anomalies discovered are overlapping. Similarly for MEWMA chart.

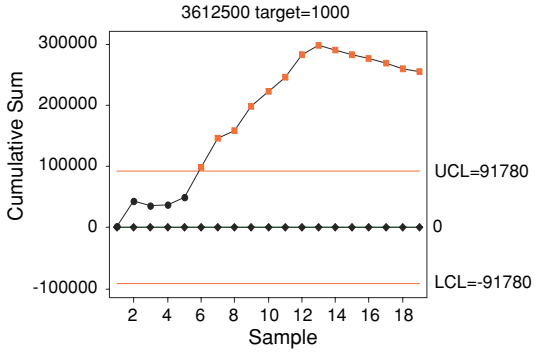

Fig. 17 MCUSUM chart for sensor 7

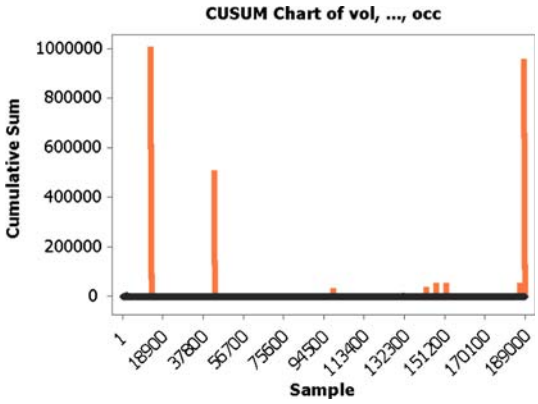

Fig. 18 MCUSUM for the traffic dataset

As shown in Fig. 17, the sensor 7 (id: 3612500) is consistently shown as having several anomalies strengthening the results of spatial outlierness for the same sensor using our approach. However, the key difference is identified when some known outliers are inserted into the data. In a step for validation, we discussed above that, 13 known outliers were inserted into the traffic dataset and all of them are discovered by our neighborhood based outlier detection technique. However, MCUSUM, Fig. 18, can only discover 9 of them and MEWMA, Fig. 19, can discover 10 of them. Our approach was able to identify all 13 of the known outliers.

6.4 Additional results

6.4.1 Comparison of precision

In Fig. 16b we compare the precision acquired by considering a neighborhood based on (1) purely spatial relationships, (2) spatial relationships and JC, (3) spatial relationships and SC, (4) spatial relationships with SC AND JC and (5) spatial relationships with SC OR JC. It can be seen that neighborhood created with SC AND JC has the best

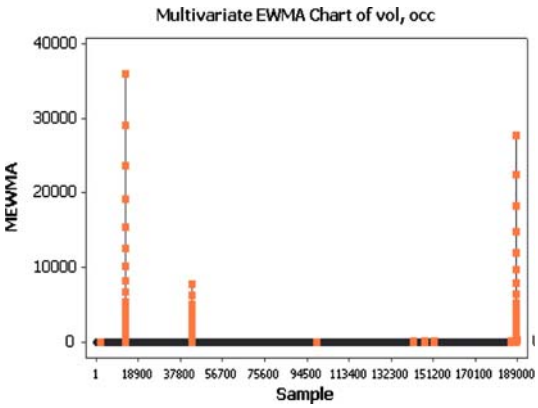

Fig. 19 MEWMA for the traffic dataset

precision of all. Moreover with the data comprising of 7 sensors from the water monitoring dataset, using SC with spatial relationship provides better results than when using JC in combination with spatial relationship. This is mainly because SC gives a true representation of a neighborhood in terms of the readings over a period of time.

6.4.2 Selection of SC and JC thresholds

In general it was observed that SC AND JC produces the best precision. We present some observations for the selection for the right coefficient. Firstly, the coefficient can be selected based on the quality of the data. Overall it was observed that SC is a more suitable coefficient when feature data is sparse as compared to the point set data used to compute the SC. In certain other cases where the readings are noisy or have high degree of overlap, then JC is more appropriate coefficient. Secondly the coefficient to be used can be selected based on the availability of data. Thus, if the readings are readily available then SC can be used in other cases where the feature set is more comprehensive JC can be used. In most cases a combination of both SC AND JC provides the best results. Here we would like to note that there is a high degree of tuning involved for selecting the threshold values of the coefficients. However, we have certain advantages. First our approach is order invariant unlike the traditional cardinality based approaches. Second our approach provides a framework for accounting for both autocorrelation and heterogeneity. In addition a different coefficient may be used for better flexibility in selection of threshold values and less dependence on tuning. For instance we could use Moran's I and Local Moran's metric (Haining 2003). We defer this to future work.

6.4.3 Summary of results with larger datasets

Another set of experimentation was performed with a larger dataset consisting of 37 sensors and 15000 points. Each sensor was qualified by 11 features. The overall results depicted a refinement over purely spatial relationship based neighborhood formation.

However, a marked improvement was seen with JC rather than SC. Our understanding is that the sensor reading data, exhibited a high degree of overlap due to which the neighborhoods were merged and showed a similar behavior as spatial relationship based neighborhoods in most cases. Thus, we can see that in certain cases the feature data, which is primarily used for JC computation, may be more appropriate than the actual point data or vice versa (as in 7 sensor data), for macro neighborhood formation. The difference can be judged with the neighborhood formation. If the neighborhood formation is markedly different with the JC, SC thresholds as compared to using purely spatial relationships then the following knowledge discovery will be more refined with the coefficient based macro neighborhood formation.

6.4.4 Prototype

The prototype can be found at—http://165.230.127.120/Website/proto/frame.htm. A snapshot showing the prototype can be seen in Fig. 20. It has been implemented in ARC IMS and ARC VIEW (2002). The implementation is in Java so that it can be integrated with other monitoring modules (such as displaying to PDA). The prototype consists of several GIS layers, which can be activated on the right panel. These include the layers for all the sensors. The website is initiated with the layer for all the sensors as the active layer and the layers generated for various stages in neighborhood formation at various threshold levels of JC, SC. Each layer can be queried for more information about each of the sensors. A script is run in the background, which initiates the identification of the anomalies. This script sends an input to this interface and the anomaly layer is activated. Here the anomaly is the sensor, which has been detected as a spatial outlier. Further, this can be modified to also show the spatio-temporal outliers and the progression of the temporal outlier across the macro neighborhood. The geospatial layers can be further queried using the query builder with various data fields for the sensors. The data for the sensors can be attached to the layers for further analysis by

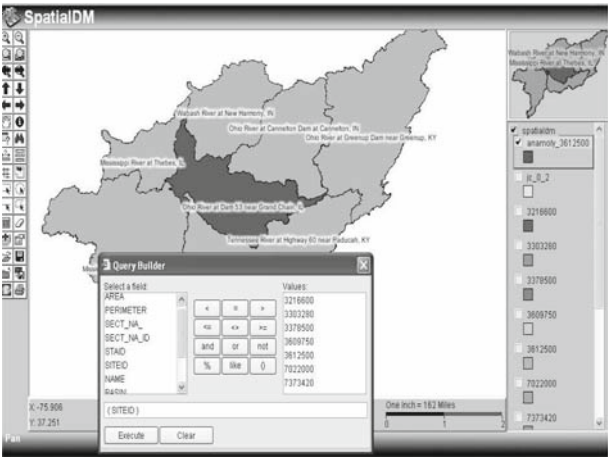

Fig. 20 Geospatial prototype interface

the scientists. Moreover, this interface can interact with various modules to generate alerts for the investigators such as scientists, field workers etc. and shown via mobile devices.

7 Related work

Spatial neighborhood formation is a key aspect to any spatial data mining technique (Ester et al. 1997; Lu et al. 2003; Shekhar et al. 2001; Huang et al. 2004; Sun and Chawla 2004; Huang et al. 2006 etc.), especially anomaly detection. Firstly, we evaluate different spatial data mining techniques for the neighborhood formation they have used. Secondly we evaluate different spatial anomaly detection approaches.

The issue of graph based spatial outlier detection using a single attribute has been addressed in Shekhar et al. (2001). Their definition of a neighborhood is similar to the definition of neighborhood graph as in Ester et al. (1997). The neighborhood graph consists of the nodes, which correspond to the spatial objects from the spatial database and edges between the nodes. These edges are present, if and only if there exists a spatial relation between the two nodes such as topological, direction and distance relationships. Various database primitives are proposed (Ester et al. 1997) to identify such relationships. However, this process of selecting the spatial predicates and identifying the spatial relationship could be an intricate process in itself. Moreover, definition of spatial neighborhood does not capture the inferential relationship between the attributes of the spatial objects and their respective areas of influence. Shekhar et al. (2003) provides a general definition of a spatial outlier and shows that the various tests for finding outliers are the special cases of this generic definition. However, they focus on outliers in a single attribute and do not address outliers in multidimensional space using multiple attributes. In addition the spatial neighborhood definition is based purely on the location of the spatial object using relationships such as adjacency or distance.

Sun and Chawla (2004) proposes a Spatial Local Outlier Measure (SLOM) which captures the local behavior in a spatial neighborhood and identifies local outliers. However, this technique considers the traditional notion of neighborhood defined based on distance or spatial relationships. Moreover it suppresses the reporting of outliers in highly unstable areas, where data is too heterogeneous. Lu et al. (2003) propose an approach to discover spatial outliers with multiple attributes. The neighborhood is based on spatial relationship of adjacency. They use Mahalanobis distance to analyze spatial data with multiple attributes, considering the average and median of the neighborhood values to determine the deviation of a spatial object from the normal behavior in the neighborhood.

Huang et al. (2004, 2006) propose approaches to identify collocation patterns in a proximity based spatial neighborhood. This approach considers spatial neighbors defined based on spatial relationships such as adjacency and metric relationships using Euclidean distance. Subsequently co-location of spatial features is detected in this neighborhood definition. Essentially in all of the above approaches the spatial neighborhood does not account for spatial autocorrelation and heterogeneity in combination.

Lu et al. (2007) detect and track region outliers in meteorological data. Here region outliers are a group of adjacent points whose features are inconsistent with those of

their surrounding neighbors. The Neighborhood connectivity is primarily based on adjacency of spatial points. Kou et al. (2007) identifies point and region outliers by generating a graph based on k-nearest neighbor relationship. It subsequently allocates weights to edges based on differences of non-spatial attributes and removes edges with high weight edges to identify outliers. The KNN graph again is based on basic spatial relationships to form the neighborhood.

Ng and Han (1994) proposes a spatial clustering technique for grouping of similar spatial objects. This approach depends on DBLearn for selection of tuples. Another clustering technique uses Delaunay triangulation (DT) for spatial clustering (Kang et al. 1997), finds outliers as a by-product of clustering. It connects the points by edges if they are within a certain threshold proximity. However, both these approaches do not consider the inferential and implicit spatial relationships, which could be useful in determining the cause of the outlierness. Autoclust (Estivill-Castro and Lee 2000) also uses DT for identifying clusters of spatial points mainly considering the length of the edge as the main factor for clustering. The disadvantage of using DT is that, we need to assume non-collinearity among objects, however, many times we need to analyze collinear points. Moreover at least 3 points are required to create the triangulation. In some cases the triangulation is not complete because the points might not be sharing a common edge of the Voronoi polygons due to this it forms something called as a Delaunay pretriangulation therefore in order to create the complete triangulation in the quadrangle with more than 4 points, the points are joined to create the DT. The algorithms need to account for these subtle changes as this might misrepresent the spatial relationships. Although it is computationally efficient to create the DT than the Voronoi polygons (Aurenhammer 1991) and subsequently derive the Voronoi diagrams, however, Voronoi diagrams capture the proximity more completely than a DT.

Scan statistics (Naus 1965) is another spatial anomaly detection technique which deals with identifying anomalous windows, where a group of objects is anomalous with respect to the entire dataset. A variation of the simple scan statistic is the spatial and spatio-temporal scan statistic (Kulldorff 1997; Kulldorff et al. 1998), which detects unusual space or space-time windows. It uses a circular window for spatial and cylindrical window for spatio-temporal processes. However, this approach does not account for the combination of both autocorrelation and heterogeneity. Moreover, the window shape is fixed such as circular or cylindrical.

Time series analysis, has evolved into a very well studied area in statistics and computer science literature. Several techniques have been studied (Chatfield 1983; Dasgupta and Forrest 1999; Shahabi. et al. 2000; Keogh et al. 2002) such as novelty detection for the detection of any divergence from normality. Most of these approaches mainly take into consideration temporal variation and temporal autocorrelation. However, our work encompasses both spatial and spatio-temporal datasets, where often times the temporal behavior is governed by the spatial behavior. Therefore we cannot only rely on pure temporal analysis. Recently Kang et al. (2008) introduced the promising problem of finding dominant persistent Flow Anomalies similar to our spatio-temporally coalesced outliers. However, their focus is more on the time series data generated by two sensors at a time and does not look at the dynamic combination of spatial and temporal properties together.

8 Conclusions and future work

In this paper we discussed spatial neighborhood based anomaly detection in sensor datasets. We first generated micro neighborhoods around spatial objects. We then identified spatial and inferential relationships between them to generate macro neighborhoods. The spatial relationship was identified using the spatial attributes such as the spatial coordinates and the inferential relationship was identified using features and point sets in the micro neighborhood of a spatial object. Thus, our neighborhood definition captured spatial autocorrelation and heterogeneity both. Subsequently we identified spatio-temporal, spatial and spatio-temporally coalesced outliers in this neighborhood.

In the future we would like to explore a technique for feature selection to identify the critical features that can be used to determine the inferential relationships. Currently we use two separate coefficients for identifying the inferential relationships, for this a composite coefficient needs to be devised which facilitates the identification of the inferential relationships. Moreover, we would like to improve the efficiency of the algorithms we have proposed in this paper. The aspects of the integration of the proposed methodology with a spatial DBMS (e.g., Oracle Spatial) also need to be analyzed.

Acknowledgments The authors would like to thank Dr. Kirk Barrett and Ms. Vasundhara Chaudhuri for early input on the domain knowledge about the water monitoring process at MERI. The authors would also like to thank Ms. Vani Sheshadri for help in testing the programming code.

References

ARC (2002) ARC IMS 4.0, ArcView 8.3. http://www.esri.com/

Aurenhammer F (1991) Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput Surv 23(3):345–405

Birant D, Kut A (2006) Spatio-temporal outlier detection in large databases. J Comput Inf Technol 14(4): 291–297

Chatfield C (1983) Statistics for technology, a course in applied statistics. Science Paperbacks. Chapman & Hall/CRC, Boca Raton, FL

Dasgupta D, Forrest S (1999) Novelty detection in time series data using ideas from immunology. In: International conference on intelligent systems

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases. In: KDD, AAAI Press, USA, pp 44–49

Ester M, Kriegel H, Sander J (1997) Spatial data mining: a database approach. In: 5th International symposium on advances in spatial databases, Springer, London, pp 47–66

Ester M, Frommelt A, Kriegel HP, Sander J (1998) Algorithms for characterization and trend detection in spatial databases. In: 4th International conference on KDD

Ester M, Kriegel HP, Sander J (1999) Knowledge discovery in spatial databases. In: KI '99: proceedings of the 23rd annual German conference on artificial intelligence, Springer, London, pp 61–74

Estivill-Castro V, Lee I (2000) Autoclust: automatic clustering via boundary extraction for mining massive point—data sets. In: 5th International conference on geocomputation

Griffith D (1987) Spatial autocorrelation: a primer. Assoc Am Geogr

Haining R (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge

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

Huang Y, Pei J, Xiong H (2006) Mining co-location patterns with rare events from spatial data sets. Geo-Informatica 10(3):239–260

Kang I, Kim T, Li K (1997) A spatial data mining method by delaunay triangulation. In: 5th ACM international workshop on advances in geographic information systems, pp 35–39. doi:10.1145/267825.267836

Kang JM, Shekhar S, Wennen C, Novak P (2008) Discovering flow anomalies: a sweet approach. In: ICDM, IEEE computer society, pp 851–856

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons Inc., Hoboken, NJ

Keogh E, Lonardi S, Chiu BY (2002) Finding surprising patterns in a time series database in linear time and space. In: 8th ACM international conference on knowledge discovery and data mining, ACM Press, New York, NY, pp 550–556. doi:10.1145/775047.775128

Knorr EM, Ng RT (1998) Algorithms for mining distance-based outliers in large datasets. In: 24th International conference on very large data bases, NY, USA, pp 392–403. http://www.vldb.org/conf/1998/p392.pdf

Kou Y, Lu CT, Santos RFD (2007) Spatial outlier detection: a graph-based approach. In: ICTAI '07: proceedings of the 19th IEEE international conference on tools with artificial intelligence, vol 1 (ICTAI 2007), IEEE Computer Society, Washington, DC, pp 281–288. doi:10.1109/ICTAI.2007.169

Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26(6):1481–1496

Kulldorff M, Athas WF, Feurer EJ, AMiller B, Key CR (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in los alamos, new mexico. Am J Public Health 88(9):1377–1380

Lu C, Chen D, Kou Y (2003) Detecting spatial outliers with multiple attributes. In: 15th IEEE international conference on tools with artificial intelligence, p 122

Lu CT, Kou Y, Zhao J, Chen L (2007) Detecting and tracking regional outliers in meteorological data. Inf Sci 177(7):1609–1632

McGuire MP, Janeja V, Gangopadhyay A (2008) Spatiotemporal neighborhood discovery for sensor data. In: Proceedings of the 2nd international workshop on knowledge discovery from sensor data (Sensor-KDD 2007), held in conjunction with the 14th international conference on knowledge discovery and data mining (ACM SIG-KDD 2008)

Miller HJ, Han J (2001) Geographic data mining and knowledge discovery. Taylor & Francis Inc., New York, NY

Moran P (1948) The interpretation of statistical maps. J R Stat Soc B 10(243):51

NASQAN (2002) USGS, National stream water quality network (NASQAN), published data. http://pubs.usgs.gov/dds/wqn96cd/html/wqn/wq/region05.htm. Accessed 25 Aug 2009

Naus J (1965) The distribution of the size of the maximum cluster of points on the line. J Am Stat Assoc 60:532–538

Ng RT, Han J (1994) Efficient and effective clustering methods for spatial data mining. In: 20th International conference on very large data bases, Morgan Kaufmann, Los Altos, CA, pp 144–155

Okabe A, Boots B, Sugihara K, Chiu S (2000) Spatial tessellations: concepts and applications of Voronoi diagrams. John Wiley & Sons Ltd., West Sussex, England

Shahabi C, Tian X, Zhao W (2000) TSA-tree: a wavelet-based approach to improve the efficiency of multi-level surprise and trend queries on time-series data. In: 12th International conference on scientific and statistical database management

Shekhar S, Lu C, Zhang P (2001) Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: 7th ACM international conference on knowledge discovery and data mining, pp 371–376. doi:10.1145/502512.502567

Shekhar S, Schrater P, Vatsavai R, Wu W, Chawla S (2002) Spatial contextual classification and prediction models for mining geospatial data. In: IEEE transaction on multimedia

Shekhar S, Lu CT, Zhang P, Shekhar S, Lu CT, Zhang P (2003) A unified approach to spatial outliers detection. GeoInformatica 7:139–166

Shewchuk JR (1996) Triangle: engineering a 2d quality mesh generator and delaunay triangulator. In: Selected papers from the workshop on applied computational geormetry, towards geometric engineering, Springer, London, pp 203–222

Sun P, Chawla S (2004) On local spatial outliers. In: 4th IEEE international conference on data mining, pp 209–216

Unwin D (1982) Introductory spatial analysis. Methuen, London