Neighborhood based detection of anomalies in high dimensional spatio-temporal Sensor Datasets[1]

Nabil R. Adam
CIMIC,
Rutgers University
Newark, NJ 07102, USA
adam@adam.rutgers.edu

Vandana Pursnani Janeja
MSIS Department and CIMIC,
Rutgers University
Newark, NJ 07102, USA
vandana@cimic.rutgers.edu

Vijayalakshmi Atluri
MSIS Department and CIMIC,
Rutgers University
Newark, NJ 07102, USA
atluri@cimic.rutgers.edu

ABSTRACT
The behavior of spatial objects is under the influence of nearby spatial processes. Therefore in order to perform any type of spatial analysis we need to take into account not only the spatial relationships among objects but also the underlying spatial processes and other spatial features in the vicinity that influence the behavior of a given spatial object. In this paper, we address the outlier detection by refining the concept of a neighborhood of an object, which essentially characterizes similarly behaving objects into one neighborhood. This similarity is quantified in terms of the spatial relationships among the objects and other semantic relationships based on the spatial processes and spatial features in their vicinity. These spatial features could be natural such as a stream, and vegetation, or man-made such as a bridge, railroad, and chemical factory. The paper also addresses the identification of spatio-temporal outliers in high dimensions, in their neighborhood.

Keywords
Spatial neighborhood, Micro neighborhood, macro neighborhood, Sensors, outliers.

1. INTRODUCTION
Data mining, in general, deals with the discovery of non-trivial and interesting knowledge from different types of data. Traditional data mining [8] deals with numbers and categories, whereas spatial data mining deals with more complex data – spatial data. Specifically spatial data mining deals with identification of non-trivial and useful knowledge discovery in spatial data sets where spatial (point, lines, polygons, location) and non-spatial data, e.g., population count are stored. Spatial data has an important property that the nearest objects to a given spatial object are always linked by edges, which allows us to analyze proximity relationships among spatial objects [5]. Moreover spatial data mining deals with implicit spatial predicates like overlap, meet etc. Thus, in order to find similarity of spatial objects one needs to first identify which predicate to use.

Spatial dependency and heterogeneity are inherent properties of spatial objects. These make the treatment of spatial data mining different from traditional data mining techniques. Spatial dependency causes the attributes of some spatial object to be related. Spatial data analysis captures such dependencies in an important aspect called spatial autocorrelation [11,13].

Spatial heterogeneity causes the attribute values of spatial objects to vary greatly by a change in the spatial region where the object is located. As a result, small changes in a spatial region could result in changes in the attribute values of the spatial objects involved. For Example, in case of sensor readings do not only change with the change in the distance of the readings in the entire span of the river but also within a cross-section of the river, thus being affected by the change in depth of the river too.

Therefore, when considering a spatial object it is important to consider its spatial and non-spatial attributes, its implicit and explicit spatial relationships with other objects, as well as the region of influence of that object. For a given spatial object, the region of influence consists of the underlying spatial processes, which influence the behavior of this object and its neighboring objects. These spatial processes are not necessarily natural but could be man made for example: a chemical factory dumping toxic chemicals at the origin of the stream. A sensor placed in the stream will be under the influence of the spatial process of the dumping of the chemicals. The behavior of spatial objects is under the influence of nearby spatial processes. Therefore in order to perform any type of spatial analysis we need to take into account the spatial processes present in the vicinity of the spatial objects, which could influence the behavior of these objects. Once such a region of influence is identified, outliers and trends in the region can be discovered with a high level of relevance due to the implicit relationships in the region between the spatial objects and the spatial processes in the vicinity. This extends to the concept of neighborhood for a given spatial object. The neighborhood [3,8,11] cannot be implicitly identified just by the change in the spatial object, e.g., river, mountain, stream, city etc, nor it can be defined solely on the basis of spatial proximity (e.g., [2,4,5]). For effective discovery of outliers in a given spatial region it is important to take into account the spatial features in the vicinity of the objects as well as the underlying spatial processes in that region. Thus, being able to identify similarly behaving objects, would lead to the discovery of outliers within that region. In this

SAC'04, March14-17, 2004, Nicosia, Cyprus

[1] This work is partially supported by the Meadowlands Environmental Research Institute (MERI), which is a collaboration between the New Jersey Meadowlands Commission and CIMIC Rutgers Newark Campus.

paper we present a methodology for achieving this objective. Specifically, the paper addresses the issue of identifying outliers in high dimensional spatio-temporal data sets. Since spatial outlier detection deals with the identification of objects that behave very differently from their neighboring objects, it is therefore, critical to identify the right neighborhood for a given object. Such definition needs to go beyond a graph based neighborhood definition that has been addressed in [11]. Their definition of a neighborhood is similar to [3]. The neighborhood graph consists of the nodes, which correspond to the spatial objects from the spatial database and edges between the nodes, which 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 [3] to identify such relationships. However this process of selecting the spatial predicates and identifying the spatial relationship could be an intricate process. The definition of spatial neighborhood does not capture the semantic relationship between the attributes of the spatial objects and their respective areas of influence. A clustering technique uses Delaunay triangulation for spatial clustering [5] which connects the points by edges if they are within a certain threshold proximity. This approach also finds outliers as a by-product of clustering. This does not consider the semantic and implicit spatial relationships which could be useful in determining the cause of the outlierness. The disadvantage of using Delaunay triangulation 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 approximates to 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 delaunay triangulation. The algorithms need to account for these subtle changes as this might misrepresent the spatial relationships. Although it is computationally efficient to create the Delaunay triangulations than the Voronoi polygons [1] and subsequently derive the Voronoi diagrams, however Voronoi diagrams capture the proximity more completely than a Delaunay triangulation.

The paper is organized as follows. We next discuss a motivating example. In section 2 we outline the proposed approach in steps and then describe each step in detail. Section 3 includes a discussion of the data sets used to test our proposed approach as well as a discussion of the test results. Conclusions and future work is discussed in section 4.

1.1 A Motivating Example

The challenges posed by *spatial data mining* can be seen in an example scenario in the domain of water monitoring. A number of monitoring efforts are currently carried out in the New Jersey Meadowlands district by researchers at the Meadowlands Environmental Research Institute (MERI). Quarterly and continuous water sampling is carried out in and around the Hackensack River [15], here sensors are placed at different positions in the Hackensack River and its tributaries. Also the sensors are placed at various depths in the river stream. This monitoring effort is part of an overall effort to develop a decision support system for continuous monitoring of the water quality in the Hackensack River. Currently there is a network of sensors where each sensor covers a specific area within the river; Approximately 25 parameters are being monitored including, oxygen level, temperature, salinity, alkalinity, toxicity or presence of heavy metals, etc. These readings vary with the position of the sensors not only in terms of the expanse (length) of the river but also the depth of the river. Currently at MERI this data is validated based on threshold values determined by the domain expert. These validations are performed mainly to generate the corrected data to see behavior of the parameters at the various sensors. However this does not identify anomalous readings and the cause of such readings, whether it is malfunction or it is due to the change in toxicity level of the water. In the evaluation of our approach we are interested in identifying readings of sensors which are very different from other readings namely outliers. If a very large number of readings of a sensor are outliers it would indicate malfunctioning of the sensor. In this paper we limit our work to identifying the readings, which are outliers

Here the coverage area of a given sensor meets (is a neighbor of) the coverage area of one or more sensors. However this is further dependent on the spatial location of the sensors. Essentially the identification of the neighborhood can be a precursor to these knowledge discovery tasks namely spatial characterization, trend detection, outlier detection etc. In the case of sensor data, if we consider the spatial position of the sensors we would need to consider two factors: 1) the immediate region of influence of the sensor at that position and 2) the extended region of influence of the sensor. We assume that the proper region of influence for each sensor has been determined and is given. Our focus in this scenario is on discovering the sensor(s) whose reading is inconsistent with other sensors in neighboring regions and the time period when this anomaly occurred.

1.2 Preliminaries

Before we discuss our proposed approach it is essential to define certain preliminaries.

Jaccard Coefficient (JC): Jaccard coefficient is used to quantify similarity or dissimilarity of binary valued variables, i.e., having only two outcomes (0,1). Here the similarity or dissimilarity of two objects can be calculated using the contingency table as shown in table 1.

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

Table 1: Feature similarity for two Objects

For our approach quantifying the similarity match (1-1 match) than a non-similarity (0-0) match is more important. Unlike the matching or m-coefficient which considers both, JC gives more importance to a 1-1 match [6]. Therefore, we will use the JC to formalize the similarity in terms of links. An example follows based on table 1.

JC=positive match/(positive match+ mismatch) = 1/(1+1+2)=0.25

Here, the agreement of 1-1 is considered more important than the agreement of 0-0(negative match) therefore, the positive match is given more weight.

Silhouette Coefficient(SC): 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 [6]. Given a point x in a cluster A, then $a(x)$ is the average distance between the point x and the other points in A and $b(x)$ is the average distance between the point x and the points in the second closest cluster B. The Silhouette of x is then defined as[6]: $S(x) = b(x)-a(x)/\max\{a(x), b(x)\}$. Based on [6,8], the following are the evaluations of the SC of the point in the cluster. $S(x) = -1$

denotes highly overlapping structure, x that is on average closer to members of cluster B . $S(x) = 0$, in between A and B, x equally similar to cluster A and B. Hence, it is not clear whether x should be assigned to A or B. It can be considered as an "intermediate" case. $S(x) = 1$ good assignment of x to its cluster A

Silhouette coefficient, SC of cluster: is the average silhouette of all the points in the cluster. Based on [6], the following are the evaluations of the silhouette coefficient of the cluster. $0.7 < SC <= 1.0$ Strong structure; $0.5 < SC <= 0.7$ Medium Structure; $SC <= 0.25$ no structure (.25 represents the threshold level)

Outlier: An outlier is a point, which varies sufficiently from other points such that it appears to be generated by a different process from the one governing the other points. The current literature defines outlier and its difference from other points in terms of distance, density, etc. [4,7] The outlier is defined in terms of distance[7] as follows:

Definition 1: [Outlier] A object O in a dataset T is a DB (p, D) outlier if at least a fraction p of the objects in T are at a greater distance D from O.

Voronoi Diagrams: Voronoi diagrams [10] is a technique from computational geometry which divides the plane into polygons with certain properties. Voronoi diagrams for a set of objects is defined as follows:

Definition 2: [Voronoi Diagrams] The Voronoi diagram of a set of objects O is the subdivision of the plane into n polygons, with the property that a point q lies in the polygon corresponding to an object o_i iff $dist(q,o_i) \leq dist(q,o_j)$ for each $o_j \in O$ with $j \neq i$. That is, any point in a Voronoi polygon $V(o_i) = \{q \mid \|q-o_i\| \leq \|q-o_j\|$ for $i \neq j\}$.[10].

2. THE PROPOSED APPROACH

A number of approaches for spatial outlier detection have been proposed in the literature (e.g., [4,5,11]), some are by-products of clustering. A common limitation of these approaches is: identifying the neighborhood of an object based only on spatial relationships, and considering the proximity factor as the main basis for deciding if an object is an outlier with respect to neighboring objects or to a cluster and ignoring the influence of some of the underlying spatial processes that might be different at different spatial locations despite the close proximity of the two objects. We believe that an effective approach for identifying spatial outliers must take into account the spatial and semantic relationships among the objects considering the underlying spatial processes as well as the features of these spatial objects that might be different for different objects despite the close proximity of the objects. Thus, each object has an immediate neighborhood or a region of influence, which we call as Micro Neighborhood.This can be extended or merged with other adjacent regions based on semantic and spatial relationships between these neighborhoods or regions, we call this extended neighborhood as Macro Neighborhood. This identification of the neighborhood is a precursor to the outlier detection in spatio temporal datasets.

Our proposed approach builds on existing approaches to evolve into a new methodology, which addresses some limitations of existing outlier detection in spatial and spatio-temporal datasets. A summary of the overall approach is presented below, followed by a detailed discussion of each of the steps. In the context of this approach we refer to the sensor as spatial objects and the readings for each sensor as points.

1. Generation of Micro Neighborhood. This involves the generation of Voronoi polygons [13] around each spatial object. Here the *input* is the set of objects characterized by their spatial locations. The *output* is the Voronoi diagram defining an immediate spatial neighborhood around each object in the form of a Voronoi polygon around the object (referred to as Micro Neighborhood). The performance of algorithms for generating Voronoi diagrams suffers as the number of dimensions increases [1,10]. Our approach, however, is based on voronoi diagrams in 2 dimensional space only. Once the micro neighborhoods are identified the next step is to find the relationships between the polygons so that they can be merged to form an extended neighborhood.

2. Identification of Spatial Relationships. Here the *input* is the Voronoi diagram supplemented by an edge list indicating an edge shared between two polygons. The Voronoi diagram and the edge list are generated using a Triangle: a 2D mesh generator [12]. The *output* is the identification of spatial relationships between two micro neighborhoods (voronoi polygons). The output evolves from the edge list to generate the adjacencies in the form of a neighborhood matrix indicating if a micro neighborhood is a neighbor of any of the other micro neighborhoods. Currently we are considering a spatial relationship of adjacency as represented by the adjacency matrix. This is further discussed in section 2.2.

3. Identification of semantic relationships. We use JC and SC [6] to capture the semantic relationships among neighborhoods. We discuss, in section 2.3 the rationale for using both coefficients. Here the *input* is a set of micro neighborhoods each characterized by a feature vector representing the spatial processes in that micro neighborhood. The *output* is JC indicating the similarity of features between two micro neighborhoods. Another part in this step is calculating SC [6]. Here the *input* is the set of micro neighborhoods each characterized by a set of points (readings over a period of time) in that micro neighborhood. The *output* is SC indicating the level of overlap or similarity of the two spatial micro neighborhoods, in terms of readings over a period of time.

4. Generation of Macro Neighborhood. Here the *input* is the Neighborhood (adjacency) matrix, JC and SC. The *output* is the Macro Neighborhood, which is generated based on evaluation of spatial relationship, JC and SC. If two objects have a spatial relationship, and if the JC is greater than a certain threshold and SC is less than a certain threshold then the two polygons are merged. The outlier detection will be performed on the various points.

5. Detecting Outliers. Outliers are identified based on the distance values among various points (readings). An outlier is a reading that is at a greater distance than a threshold value from a certain number of points. Here we build on an existing technique for outlier detection, specifically the Distance based outlier detection [7]. Currently we use the Euclidean distance. Following is a discussion of each of the above steps in detail.

2.1 Generation of Micro Neighborhood

Our neighborhood definition is based on the concept of Voronoi diagrams [10], as defined in section 1.2, which divides the plane according to the nearest neighbor rule where each object is associated with a given region of the plane that is closest to it. Several algorithms for identifying Voronoi polygons have been extensively studied in the literature [1, 10]. We next discuss how

we make use of the Voronoi diagrams to define the micro neighborhood of a spatial object.

Let us assume that we have a finite set of n distinct spatial objects in the plane S $=\{s_1,s_2,...,s_n\}$. In the context of our domain example, a spatial object is a stationary sensor around which we generate the periphery in the form of a Voronoi polygon. The Voronoi diagram of S is the subdivision of the plane into n polygons, with the property that a feature *q (such as location of a chemical factory)* lies in the polygon corresponding to an object s_i iff $dist(q,s_i) \leq dist(q,s_j)$ for each $s_j \in S$ with $j \neq i$. That is, any feature in a Voronoi polygon $V(s_i) = \{q \mid ||q-s_i|| \leq ||q-s_j||$ for $i \neq j\}$.[10]. Thus, each feature in a Voronoi polygon is associated with the object in that polygon implicitly as its neighborhood.

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 divided it into two half planes. Thus, a feature located on one side of the bisector is closer to that half plane than the other. As we keep adding new spatial objects, more half planes are formed and the region of influence of the object is the intersection of the half planes. The Voronoi polygons form a polygonal partition of the plane -- called 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 [10]. We identify the immediate area of influence, which we call as the micro neighborhood of a spatial object, using the Voronoi diagrams. However, we will also need to consider other attributes of the object itself for example, in case of a sensor we would need to consider the range of the station as well. Below is the definition of the micro neighborhood.

Definition 3: [micro neighborhood M_i] : The micro neighborhood can be defined in terms of the region of influence of a spatial object or dominance [13] of one object over the other**.**

Given a set S of spatial objects $s_1, s_2, ...,s_n$, In the context of the above example, S is a set of sensors. The dominance of s_1 over s_2 is defined as the subset of the plane being at least as close to s_1 as to s_2 i.e., $Dom(s_1, s_2)=\{x \in$ features set $\mid d(x,s_1) \leq d(x,s_2)\}$ Where d is the Euclidean distance function. Here x belongs to the feature set which includes spatial features that can have their own spatial processes such as a chemical factory, a river, a rail track, etc. Thus a chemical factory will be considered part of micro neighborhood of s_1 since the distance $d(x,s_1) \leq d(x,s_2)$. This distance relationship is implicitly identified by the formation of the Voronoi polygons.

This results in identifying the feature vector for each micro neighborhood, which is further used to identify similarities between various micro neighborhoods.

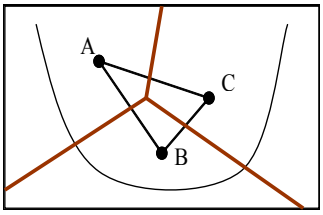

Figure 1: Sensors in the cross section of a river

In the context of our above example, the cross section of the river if we have three sensors A, B, C, will implicitly fall into their own micro neighborhood as shown in figure 1. The identification of the micro neighborhood based on Voronoi diagrams is used as an initial step to neighborhood merging, although this will identify sites based on spatial proximity only. In order to capture the semantic relationships among micro neighborhoods, we need to identify the similarities among them in terms of the features and behavior of objects over a period of time as well as the spatial relationships among them. This issue is addressed in sections 2.2 and 2.3.

2.2 Identification of Spatial Relationships

Given a set of spatial objects and their micro neighborhood we need to determine any relationship that exists among these micro neighborhoods. We first identify the spatial relationships among them.

Spatial relationships are binary relations between pairs of objects; these spatial objects can be points or spatially extended objects such as lines, polygons and polyhedrons [5]. These spatial relationships include topological, distance and direction. 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, which s determined by the property of Voronoi polygons [10] such that if two Voronoi polygons share an edge then they are adjacent to each other. This process consists of utilizing the edge list generated by Triangle: 2D mesh generator [12] for the Delaunay triangulation, which is a precursor to the Voronoi polygons. An edge in the edge list is of the format edge# <from spatial object> <to spatial object>, where the spatial objects have spatial coordinates. Thus, if there is an edge between 2 spatial objects (sensors) it implies that the corresponding micro neighborhoods (polygons) are adjacent. This is because in order to construct a Voronoi diagram we first connect the spatial objects by edges to form a triangulation and then bisect these edges to generate Voronoi polygons as explained in section 2.1. Thus we can see that a triangulation edge evolves into a common edge for two Voronoi polygons, in our case between two micro neighborhoods. Currently we consider adjacency as a representative spatial relationship and we form an adjacency matrix where a 1 indicates the existence of an adjacency relationship and a 0 represents the absence of such relationship. This becomes the neighborhood adjacency matrix. The adjacency spatial relationship can be extended to consider other complex spatial relationships, such as a combination of adjacency and direction (north, south, north east etc.) relationship or distance and direction.

2.3 Identification of Semantic Relationships

The existing approaches limit the neighborhood definition to be based only on the spatial relationship. We would like to extend the definition to include the semantic relationships as well. Once the polygons are formed around each spatial object it creates a periphery of the immediate neighborhood in the form of micro neighborhood. This micro neighborhood can now be characterized by the presence or absence of spatial features or other spatial processes. For example: the presence of a factory in the micro neighborhood, the presence or absence of a bridge, railroad, stream, the presence or absence of certain type of vegetation, etc. Such information can be accumulated with the help of domain experts for example. In many cases such studies are done before placing sensors. In case of the NASQAN [14] data, the process of determining the placing of the sensors is described in terms of "qualities of the region". The result of this analysis is a feature vector of 0's and 1's showing the absence or presence of a feature in the micro neighborhood. In our above example, a spatial object is a water-monitoring sensor, which will also have an associated

set of readings over a period of time (with each reading being considered a point in the neighborhood). Thus, we can make use of the features in the micro neighborhood and also the points (readings) in the neighborhood to identify semantic relationships among them. Moreover, if the features in the micro neighborhood are not sufficient for the identification of similarity coefficient or are do not exist then the readings can be used to identify the semantic relationship.

We now discuss how we make use of existing coefficients of measures of similarities to capture the semantic relationships among micro neighborhoods. The topic of similarity and overlap has been very well studied in clustering techniques and a number of similarity coefficients, e.g., the JC, have been identified in the literature (see for example, [6]). In case of *binary valued attributes, namely the feature vectors of the micro neighborhoods,* we use the JC For *other types of attributes (non-binary valued*), such as the readings of sensors*,* we use the SC as shown in the clustering structure evaluation [6].Both these coefficients are discussed in section 1.2 SC quantifies the overlap between two clusters, we exploit this characteristic to measure the overlap in the micro neighborhoods based on the readings over a period of time at each of the sensors.

This measure of overlap shows the relationship between two micro neighborhoods. In our case of neighborhood merging, if we discover that the points of a micro neighborhood share similarities with some other micro neighborhood's points such that SC is between 0 and 0.25 then we can address the merging of these two neighborhoods. Essentially each micro neighborhood is characterized by its own points, they can be considered similar to each other, based on higher value of JC and lower value of SC. Based on the above discussion we now define a Semantic relationship as follows:

Definition 4: [Semantic Relationship]: Given Micro Neighborhood m_i, m_{i+1} the semantic relationship is identified by JC (m_i, m_{i+1}) and SC (m_i, m_{i+1}); such that the higher the JC and the lower the SC, the stronger is the semantic relationship between m_i and m_{i+1}.

2.4 Generation of Macro Neighborhood

In order to capture the relationships among the micro Neighborhoods we need to identify the similarity among them. This can be done using spatial relationships as identified in [3] and also using semantic relationship as identified in section 2.3.

Each micro neighborhood identified using Voronoi diagrams can be considered as an implicit sub cluster or grouping (e.g., a set of readings of a sensor in that spatial region). The idea is to identify which sub clusters share similarities with each other and merge them to form a larger cluster. This large cluster would form the macro neighborhood. The macro neighborhood can be defined in terms of spatial relationships between the micro-neighborhood polygons and the semantic relationship in terms of similarity of spatial and non-spatial attributes across the micro-neighborhoods as well as the overlap of data points in them. This leads us to the definition of a Macro Neighborhood.

Definition 5: [macro neighborhood MaN] Macro Neighborhood is a graph with the outer Edges E' of the micro neighborhood polygons M_i and Links Li where a link $l = (m_i, m_{i+1})$ holds iff spatial neighbor (m_i, m_{i+1}) and semantic neighbor (m_i, m_{i+1}).

Here spatial neighbor (mi, mi+1) refers to the spatial relation between polygons and the semantic neighbor refers to the semantic relation based on the test for JC and S such that $JC \geq \delta_1$ or $SC \leq \delta_2$, where δ_1 is a threshold value for measuring the JC and δ_2 is the threshold value of SC. Figure 2 shows an algorithm for generating macro neighborhoods. The algorithm takes as input the polygon set generated from the Voronoi diagram, which gives the micro neighborhoods. For each polygon we identify if there is a spatial relationship between two micro neighborhoods. We then identify if there is a certain level of similarity between two polygons based on the corresponding values of JC and SC then these two micro neighborhoods can be merged.

The spatial and non-spatial data regarding the spatial objects is in files, which is extracted into matrices. The Polygons (micro neighborhoods) from the Voronoi diagrams are initialized into the poly matrix, similarly, the data matrix contains data points about each spatial object so for example a water monitoring sensor transforms into a polygon or micro neighborhood, the data matrix will contain the readings for that sensor. Once the various matrices are initialized and populated (not represented here), we then look for spatial relationship of adjacency between two micro neighborhoods. If there exists a spatial relationship we further examine semantic relationship in terms of the SC and the JC. Once these relationships are determined to hold, we merge the micro neighborhoods, which essentially group together all the data points of the two micro neighborhoods maintaining both the old micro neighborhood id and the new one, which it receives from the merging. Further implementation details are discussed in section 3.1

2.5 Outlier detection

As we discussed earlier several techniques have addressed for spatial outlier detection [5,11,13]. In the case of graph based spatial outlier detection [11], emphasis has clearly been given to the identification of the neighborhood. This neighborhood, however, considers the linkage between spatial objects based on the spatial relationships such as distance, direction shown in the connectivity graph. Thus, the subtle cases where the spatial relationship may not be the only determining factor could be left out. Some other techniques also discover outliers as a by-product of clustering [5]. However, the focus here is clustering and the outliers are the objects, which are left out of the cluster, in this process there is no concept of relationships but distance or proximity between the objects. That is, proximity is considered as the driving factor. It is possible that a given point is an outlier with respect to more than one cluster. It is therefore, important not only to identify the outliers but also to identify the neighborhood. This can lead to further analysis of identification of trends and the causes of such trends. Therefore, we not only want to consider the objects based on spatial relationships but also semantic relationships. This leads us to our definition of a spatial outlier.

Definition 4: [Spatio-temporal Outlier] A point x_i is said to be a spatio-temporal outlier iff it differs sufficiently from other points in the macro neighborhood. Here the Macro neighborhood consists of all the micro neighborhood merged into it under the spatial and the semantic relationship restrictions.

Algorithms for identifying spatio-temporal outliers: The spatial outlier detection is the final step in our approach. Once we identified, the macro neighborhood of a set of spatial objects, we can employ any of the available outlier detection techniques. In

this paper we utilize a technique based on the distance based outlier detection technique proposed in [7]. This algorithm has been proposed in the context of traditional data mining and has the advantage of being simple and intuitive. At this juncture we can consider proximity in terms of distance threshold as one of the determining factor since we have allocated the spatial objects to their respective neighborhoods. In the outlier detection algorithm we set a threshold for the number of points (count) from which a certain point is at a greater distance than d. The threshold count 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.

If more than a certain number of points are outliers for a spatial object, then it can be further investigated if the object is an outlier in its entirety (Spatial outlier). In essence for the example of water monitoring sensors, if more than a set of readings are outliers then the sensor can be investigated to be a spatial outlying object. Once the outliers have been detected in a macro neighborhood we want to identify readings, which belong to different micro neighborhoods in the bigger macro neighborhood but have the same temporal id implying a possibility of a temporal anomaly

```
MacroNbdGen()
Initializing polygon matrix for all spatial objects
poly[number_Of_Polygons][number_of_Attributes]
Initializing data matrix for all points in the micro neighborhood of the spatial objects
data[number_Of_Points][ number_Of_Dimensions]
Initializing feature matrix for all spatial objects
featureSet[objects][features]
Initializing adjacency edge matrix for all spatial objects
edgeset[number_Of_Edges][ number_Of_EdgeAttributes]
Processing all edges to identify neighbors
for(int q=0;q< number_Of_Edges;q++)
{
for(int j=1;j<=2;j++)Edge is of the form <edge#><spatial object><spatial object>, so j starts at 1
and goes till 2
{
x=edges[q][j]            Using edge matrix values as subscripts to neighbor adjacency matrix
y=edges[q][j+1]
Setting x as neighbor of y
        neighbor[x][y]=1;neighbor[y][x]=1;
}

for each polygon p_i in poly
    for each polygon p_j in poly
        Identifying the Spatial Relationship
        if spatial_relation(p_i,p_j)  i.e., if(neighbor[i][j]==1)
        Extract feature vector for each spatial object and send to calculate JC
        JCij=Jaccard(p_i,p_j);
        Extract data for each polygon and send to calculate SC
        SC_ij=silhouetteCoefficient(p_i,p_j)
        if (JC_ij > δ1)&&(SC_ij < δ_2)
        p_i_Neighborhood_id=p_i_id
    end for
end for
```

Figure 2: Algorithm for Spatial Neighborhood generation

propagating in the neighborhood. We defer the discussion of this work to the future research.

3. EXPERIMENTAL RESULTS

In this section we discuss the empirical performance of our proposed approach. The main question we are interested in answering is "How well are we able to characterize similarly behaving objects into one neighborhood" taking into account not only the spatial relationships among objects, but also semantic relationships based on the spatial processes and spatial features in their vicinity. A consequence of identifying the neighborhoods is the outlier detection.

3.1 Date Sets and Implementation Details

We used two datasets in our study: the highway traffic monitoring dataset [11] and the water monitoring dataset [14]. The highway traffic monitoring dataset was used for validation purposes. The aim is to ensure that our approach is able to capture the example outliers described in [11]. The water monitoring dataset is the closest to the illustrative domain as described in the motivating example. A discussion of the datasets is included below.

Highway Traffic Monitoring Dataset: The Graph based spatial outlier detection technique [11] was evaluated on a large real world data set from the Minnesota Department of Transportation. The paper and the final project report discussed some known examples of outlier detection. The dataset includes the traffic readings for the stations on I 35 W North Bound and South Bound. We experimented with a subset of this dataset that consisted of 60 stations along I 35 W NB and SB. The main attributes in the data are the time slots of 5 minutes during the day, volume and occupancy readings for the station for that temporal reading. Each station is also associated with spatial location in the form of latitude and longitude. The feature matrix was created based on such attributes as highway name, direction of traffic flow and clustering of stations. For the purpose of validation, we augmented the dataset with some "known" outliers.

Water Monitoring Dataset: This dataset is taken from the USGS program "National Stream Quality Accounting Network" (NASQAN) [http://water.usgs.gov/nasqan/progdocs/index.html]. This 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 approximately 41 stations where the concentration of chemicals, including pesticides and trace elements, is measured along with stream discharge. We experimented with a subset of the full dataset, specifically we included data related to 7 stations only. Since our algorithms require an input of a feature matrix, which shows the presence/absence of characteristics in the micro neighborhood. The feature matrix for the water monitoring data utilizes the features [http://water.usgs.gov/nasqan/progdocs/statables.html] used by the EPA to select the sensor locations for monitoring. Sensors in the NASQAN program are chosen at major nodes within the river basin network to provide characterization of large sub basins of these rivers. For creating the feature matrix we considered 21 features. Some of the features are: Mean discharge (ft3/s), Incremental increase in drainage area(mi2), Drainage areaKM2, Percent urban , Percent forest, Percent mixed crop and natural features, Population density per square mile .

The similarity matrix captures, in an approximate way, the similarity of the various stations in terms of these features. This, however, can be further refined by including domain experts' input as to which features are more critical than others and perhaps assigning higher weight to each of these features. We defer the work on feature selection to future research. The water monitoring data used for outlier detection consists of spatial attributes of latitude and longitude, which is useful in determining the spatial relationship, e.g., adjacencies. It consists of the

temporal attributes of date and time of sampling. The data also consists of over 100 water-monitoring attributes. These include: Mean daily streamflow, Temperature, Specific conductance, Dissolved oxygen, pH, Alkalinity, Suspended sediment, Suspended sediment, Ammonia nitrogen, Nitrite nitrogen, Organic nitrogen plus ammonia nitrogen (filtered) etc.

3.2 Discussion of Results

Highway Traffic Monitoring Dataset: JC and SC are varied to determine the impact on the neighborhood identification in the various test cases while keeping the distance threshold constant for the outlier detection. It is observed that as JC is reduced, the number of outliers increases and more number of polygons are merged, and vice versa. Also consistency across thresholds is also maintained for example: station 60 forms a neighborhood of itself and has 153 outliers, this is consistent across the different values of JC.

Our results for outlier detection confirmed results obtained in [11] as follows. The results show a spatio-temporal anomaly in stations 29 and 30 from the time 9:30 to 10:15 am. Moreover the anomaly is also observed for 2:30 PM for the two stations. Further if we see station 31-34, it shows an anomaly with a gap of about 10 minutes. This could probably lead to detection of the progression of the anomaly. For the purpose of validating our approach, 13 outlier cases were randomly dispersed throughout the data, which consisted of 108,900. Each one of the 13 outliers was detected in the outlier set along with the other outliers. Although the approach [11] 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.

The Water Monitoring Dataset: 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 [12], which generates an edge for each two nodes that are judged adjacent to each other. The adjacency here is considered as a type of spatial relationship.

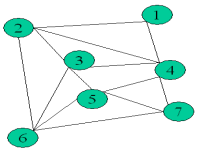

Figure 3: Adjacency graph for the spatial objects

This is the starting point of identifying neighbors based on spatial relationships. The spatial adjacencies are then expressed into an adjacency matrix and the corresponding graph is shown in figure 3. As is evident from this graph, there is a high level of connectivity among the nodes simply based on adjacency. Identifying the neighborhood based only on the spatial relationships would result in one big one big neighborhood, i.e., the entire graph would collapse into one big node. Here the spatial objects initially lie inside the micro neighborhood (voronoi polygon), after applying the spatial relationship we get the macro neighborhood. Next we describe three sets of experiments by varying the JC and SC

Case 1: Neighborhood identification and outlier detection based on the JC along with the spatial relationships

By adding the JC along with the spatial relationships the macro neighborhood is not clumped into one big region and we end up with three neighborhoods. We next investigated the sensitivity of the resulted neighborhood and number of detected outliers, to the threshold value of JC. Related results are shown in figure 4. Below is a discussion of some observations pertaining to these results.

Observation 1: Incremental building of Macro Neighborhood

The results indicate that the neighborhood generated with a more expansive JC threshold builds on that generated with a more restrictive JC threshold and vice versa. For example the macro neighborhood generated at JC threshold of 0.5 builds on neighborhood generated with JC threshold of 0.8. Here macro neighborhood generated at JC threshold of 0.5 consists of the polygons 2,4,6,7 and the macro neighborhood generated at JC threshold of 0.2 consists of the polygons 2, 3, 4, 6, 7. And despite change in the threshold value, 6 and 7 are grouped together. Thus, the neighborhood shows the incremental merging on the basis of less restrictive threshold value of JC.

Observation 2: Refinement in outliers detected

Outlier detection will produce different results as the neighborhood changes. Results obtained when identifying neighborhood only on the basis of spatial relationships are different as compared to when taking the JC into account. Moreover, as the threshold for JC varies the number of outliers is more refined. As shown in figure 4, the higher the threshold value of JC, the less number of outlier detected. This is mainly because the neighborhood becomes more refined. Thus, if there is a selective set of points in a neighborhood then the outliers are based on distance from the proper set of objects in the neighborhood and not all random points, which could lead to more outliers than actually present.

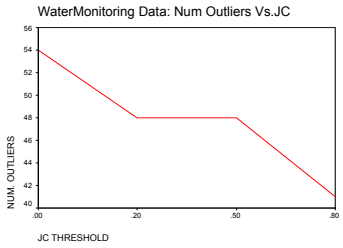

Figure 4: JC threshold vs. Number of outliers detected

Observation 3: Systematic elimination of outliers

At low threshold value of JC, the number of outliers is high and at high threshold value of JC, the number of outliers is less. However, outliers detected at high threshold value of JC are a subset of those detected at low threshold value of JC. For example, at 0.5 threshold value of JC the outlier detected are 2,3,4,8 as compared to the outliers 2,4 detected at 0.8 threshold value of JC. Thus, the process systematically eliminates outliers which do not conform to the neighborhood.

Observation 4: Consistency in Outlier detection

It is seen that sensor id 1 has no outliers at 0.2 threshold of JC. This is consistently at other threshold values of JC 0.5 and 0.8 thus, consistency is not compromised in the process of outlier detection.

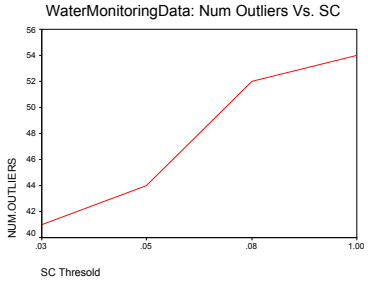

Figure 5: SC threshold vs. Number of outliers detected

Case 2: Neighborhood identification and outlier detection based on the SC along with the spatial relationships

We next investigate the

sensitivity of the resulted neighborhood and number of detected outliers, to the threshold value of SC. Related results are shown in figure 5. Below is a discussion of some observations pertaining to these results.

By adding SC along with the spatial relationships the neighborhood is not clumped into one big region and we end up with three neighborhoods.

Observation 1: Incremental building of Macro Neighborhood

As the threshold value of SC decreases the neighborhood gets more refined. With the SC threshold value of 0.03 the neighborhood is more granular than a threshold value of 0.8. This is mainly because fewer polygons will have such a high level of overlap and thus they are not merged into one neighborhood.

Observation 2: Refinement in outliers detected

Similar to Case 1, results obtained when identifying neighborhood only on the basis of spatial relationships are different as compared to obtained by into account SC. As shown in figure 5, the lower the threshold value of SC, the less number of outlier detected. This is in conformance to observation 1 for identification of the neighborhood. Observation 3 and 4 are found similar to those under Case 1.

Case 3: Neighborhood identification and outlier detection based on both JC and SC along with the spatial relationships

Observation 1: Refinement in outliers detected

When JC and SC are used in combination the results from JC are further refined into smaller neighborhood.

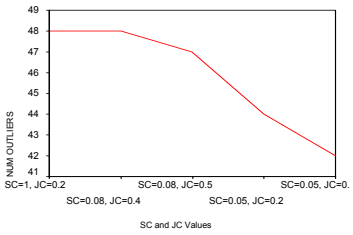

Figure 6: SC and JC threshold vs. Number of outliers detected

Here with using only JC 2,4,6,7 are grouped together, however when we refine the condition with SC the neighborhood is broken into two parts of 2,6 and 4,7.

The various observations of JC and SC together can be seen in the graph in figure 6. Mainly if we see a very low JC and a very high SC leads to polygons lumped together into one neighborhood and number of outliers as high, as SC is reduced and JC is increased the neighborhood becomes more refined and number of outliers are also refined or reduced. This refinement in Outlier numbers does not compromise the consistency in the results, such that if a certain polygon does not contribute to the outlier set in one case, it does not do so in another case as well. The outliers in the refined case (e.g.: SC=0.05, JC=0.5) are a subset of outliers in the broader case (e.g.: SC=1.0, JC=0.2)

4. CONCLUSIONS AND FUTURE WORK

In this paper we discussed spatial neighborhood based outlier detection and the definition of the neighborhood based on spatial and semantic relationships. The semantic relationship is identified using features in the vicinity or micro neighborhood of a spatial object. The feature vector for a micro neighborhood can consist of many features, however only a few can be critical for the neighborhood identification. These can be selected with the help of a domain expert or by devising a semi automatic technique. In the current approach we use two separate coefficients for identifying the semantic relationships, for this a composite index needs to be devised which facilitates the identification of the semantic relationships. Moreover once we identify the outliers with respect to the neighborhood, it would be intuitive to see which point is the point of origin of the anomaly. Further trends of variation in a certain attribute value like toxicity can be explored.

The paper also motivates an explicit comparison of the proposed methodology with certain other methods described in the paper. The aspects of the integration of the proposed methodology with a spatial DBMS (e.g. Oracle Spatial) could be also analyzed.

5. REFERENCES

[1] F. Aurenhammer. Voronoi Diagrams: A Survey of a Fundamental Geometric Data Structure. ACM Computing Surveys, Vol 23(3), 345-405, 1991

[2] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander. Algorithms for characterization and trend detection in spatial databases. In Proceedings of 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD), 1998.

[3] M. Ester, H. P. Kriegel, and J. Sander. Spatial Data Mining: A Database Approach. In Proceedings of the International Symposium on Large Spatial Databases, Berlin, Germany, July 1997, pp. 47-66.

[4] M. Ester, H. -P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In Proceedings of 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD), 1996.

[5] I. Kang, T. Kim, and K. Li. A Spatial Data Mining Method by Delaunay Triangulation. In Proceedings of the 5th International Workshop on Advances in Geographic Information Systems (GIS-97), pages 35-39, 1997.

[6] L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.

[7] E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In Proceedings of 24th Int. Conf. Very Large Data Bases, VLDB, 1998

[8] H. J. Miller and J. Han, Geographic Data Mining & Knowledge Discovery, Publisher: Taylor & Francis; 1st edition

[9] Minnesota Highway traffic dataset: http://www.cs.umn.edu/research/shashi-group/TrafficData/

[10] A. Okabe, B. Boots, K. Sugihara, S. Chiu. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. pp. 291-410. John Wiley, 2000.

[11] S. Shekhar, C. Lu, and P. Zhang. Detecting Graph-Based Spatial Outlier: Algorithms and Applications(A Summary of Results). In Computer Science & Engineering Department, UMN, Technical Report 01-014, 2001.

[12] J. R. Shewchuk, *Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator*. First Workshop on Applied Computational Geometry (Philadelphia, Pennsylvania), pages 124-133, ACM, May 1996

[13] D. Unwin, Introductory Spatial analysis, Publisher: Routledge Kegan & Paul. January 1982

[14] USGS, National Stream Water Quality Network (NASQAN), Published Data: http://water.usgs.gov/nasqan/progdocs/index.html

[15] Water Monitoring, the Meadowlands Environmental Research Institute, and the New Jersey Meadowlands Commision : http://cimic.rutgers.edu/hmdc_public/monitoring/