

Detecting Anomalous Geospatial Trajectories through Spatial Characterization and Spatio-Semantic Associations¹

Vandana P. Janeja, Vijayalakshmi Atluri and Nabil R. Adam

{vandana, atluri, adam}@cimic.rutgers.edu

CIMIC and MSIS Department

Rutgers University

Abstract

There are numerous applications that require one to identify anomalous geospatial trajectories in the domain of homeland security. Examples include: (i) a customs agent may want to discover anomalies among cargo routes to identify potentially dangerous shipments even before they cross the border, (ii) an FDA inspector may want to trace anomalous paths in the food supply chain to identify potential agro terrorism threats, etc. To accomplish this, in this paper, we propose an approach that relies on spatial characterization and spatio-semantic path associations. In particular, we consider atomic geospatial units and generate *micro neighborhoods* around them. We will then form similarly behaving regions, called *macro neighborhoods*, through selectively merging these micro neighborhoods by considering the spatial and semantic relationships among them. We will then identify associations among the geospatial trajectories and all the macro neighborhoods. Any strong association between a macro neighborhood and a part of the trajectory that does not reside in the macro neighborhood indicates a potential anomaly.

1. Introduction

Data mining techniques are used to discover interesting rules and patterns for useful analysis. Such analysis can be crucial in the decision-making process, as it is capable of extracting unknown and non-trivial patterns. During the recent years, data mining has undergone several developments for various application domains, including banking, credit cards, transportation, finance, etc. The focus of this paper is to develop a novel data mining technique to discover *anomalous geospatial trajectories*. A geospatial trajectory represents a connected path of a series of discrete points that have spatial coordinates, which could be a set of links connecting geocoded locations. The geospatial points forming the trajectories are associated with data that can be of both spatial and non-spatial attributes pertaining to different domains of interest. The process of geospatial characterization identifies regions of uniform distributions of these attributes such that it captures the underlying behavior of the spatial processes in the region in terms of the similarity of these attributes.

An anomalous geospatial trajectory is one that deviates from the normal. This is in terms of deviating from the expected path such that the points on the geospatial path have an association to some other spatial points that are not on the trajectory. There are numerous application domains that require one to identify anomalous geospatial trajectories. For example, if it is a transportation route, it could indicate stopovers at unexpected locations or deviation from a normal route. In the area of Border protection, a customs agent may want to identify anomalies among cargo routes to identify potentially dangerous shipments before they even cross the border, an FDA inspector may want to trace the anomalous paths in the food supply chain to identify agro terrorism acts, etc.

Specifically, in this paper, all our examples are focused in the domain of US Customs, where several cargo shipments cross the border and travel into various destinations within the United States. These cargo routes are essentially geospatial trajectories where data can be (i) spatial in terms of the spatial locations traversed and other attributes (such as roads, airports, bridges) associated with spatial features, and (ii) non spatial that qualify the path in terms of data about the path or from various partnering domains (such as agriculture, intelligence, coast guard and so on). Our approach takes into account all this semantic information into the knowledge discovery process. We address the behavior of geospatial trajectories, such as cargo routes, which traverse a region where information is attached to each part of the

¹ This work is supported in part by the National Science Foundation under grant IIS-0306838.

trajectory and to the region it traverses. This can be used for the identification of anomalous cargo routes. Anomalies in cargo routes may prop up in several ways - a cargo bound for a certain destination may make illegal stopovers, the source of origin could be different from what is documented in the manifest, among others. Anomalies among such geographical trajectories may potentially contain illegal shipments. In this paper, we propose an approach that relies on identifying the non-obvious associations among the trajectories and the regions identified using geospatial characterization, which can aid in identifying anomalous trajectories. In the following, we provide a motivating scenario that outlines the threats encountered in the border protection domain.

Example 1: *Illegal Transshipment* [USC03A] *Transshipment, in general, refers to the movement of goods through multiple stopovers en-route to its destination. While transshipment is legal and is commonly used by businesses, often it is exploited for the purpose of circumventing trade laws and restrictions applicable to the shipment, and to ship unlawful merchandise. Some of these items include contaminated or prohibited foods, counterfeit pharmaceuticals, ozone depleting Chlorofluorocarbons, illegal drugs, weapons, hazardous wastes, and substandard automotive and aircraft parts. As such, US Customs considers transshipment as a serious threat to consumers and/or to the environment within US in terms of the cargo itself, therefore it monitors the source of the cargo or the unauthorized stopovers a shipment would make, and uses its authority to combat the illegal transshipment [USC03B]. With a careful examination of certain attributes of the shipment, a seasoned customs officer can identify anomalies and can determine if a shipment is indeed a transshipment of restricted goods or via restricted ports. For instance, consider a shipment of fish from Spain: Although the bill of lading indicating that it has originated from North Africa, it could be overlooked with the current level of profiling. Assume that the container does not have the appropriate refrigeration for carrying fish, and the company has no prior history with the US Customs. In an ideal scenario, the customs officer would have at hand the countries' profile available from some intelligence agency indicating that Spain is a destination and minor transshipment point for Southwest Asian heroin and a key European gateway country for Latin American cocaine and North African hashish entering the European market [CIA03], or he/she simply would know it based on prior experience. However, if we were to use an automated tool, this cannot be identified merely on the basis of the attributes in the shipping manifest. As a result, even if the real origin of the shipment were North Africa, it would be identified as that from Spain and may evade inspection.*

In this paper, we propose an approach based on geo-spatial characterization, and demonstrate how an anomalous geospatial trajectory can be identified. Although in this paper we use examples and discussions pertaining to an anomalous cargo route and illegal transshipment, it is important to note that our approach is general and can be applied to detecting an anomalous geospatial trajectory in any application domain.

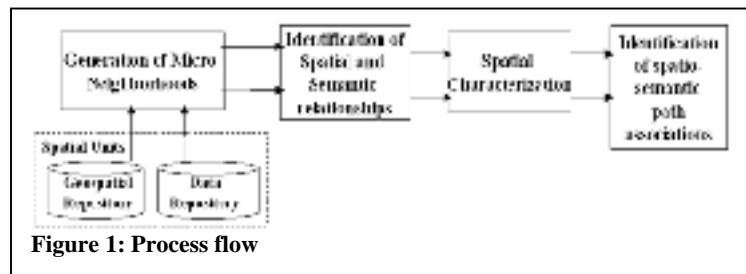
In the following, we outline the process of identifying illegal transshipments with respect to the specific example above. Since a cargo shipment follows a specific route, one may exploit the spatial and non-spatial attributes associated with each geographic location in the route. Since several geo-spatial associations are possible, our analysis will use various spatial layers for the spatial distribution of imports in the region under consideration, coastal zones, terrorist activity zones, drug zones, transshipment hubs, etc. These layers represent knowledge from multiple domains. A particular cargo can be strongly associated with some part of the route than other parts. Alternatively, it could be strongly associated with some geographic region that is not on the route, which would indicate the possibility of transshipment prompting further investigation. Current spatial data mining approaches are based on spatial autocorrelation represented by the spatial relationships, and therefore cannot discover the associations described above. This is because, in addition to the spatial autocorrelation, they require spatial heterogeneity in terms of the behavior of the non-spatial attributes across the region. In other words, to determine the accurate distribution and associations of such attributes to the spatial location, it is essential to identify the correct neighborhood in terms of both spatial autocorrelation and spatial heterogeneity. More specifically, our data mining process involves dividing the geo spatial region under consideration into similarly behaving regions. This leads to the characterization of the region based on spatial and semantic relationships. Each such region is associated with a composite feature vector capturing the

underlying spatial and semantic information. Given a cargo route, it is first divided into segments where each segment will be associated with its feature vector. The composite feature vector of the characterized region and the feature vector of the segments are used to identify associations, which could be either expected or unexpected. If the segment is most strongly associated with its own region, then it validates the shipment route. However, any strong association between a characterized region and a segment that *does not reside* in the region indicates a potential transshipment. For example, if one of the attributes of the cargo route segment is related to drug activity, then all the regions that are characterized as active drug zones should be taken into consideration for further investigation to identify potential relations.

The rest of the paper is organized as follows. Section 2 provides an overview of our approach. Section 3 presents the detailed steps involved in identifying anomalous geospatial trajectories along with the experimental results. Section 4 provides an overview of related work and Section 5 the conclusions and future work.

2. Overview of Our Approach

We begin our approach by considering the smallest geospatial entities, called *spatial units*. As an



example, a spatial unit could be a city associated with spatial coordinates and other spatial and non-spatial information including economic data, agricultural data, boundary data (demarcations of borders or a country), etc., whose sources are from various domains. For example, the Central Intelligence Agency (CIA) maintains a fact book for

each country where different attributes are gathered including social, economic, geographic factors in categories such as Geography, People, Government, Economy, Communications, Transportation, Military and Transnational Issues [CIA03].

We generate a *micro neighborhood* around each spatial unit based on the concept of Voronoi polygons. Using the similarity among the non-spatial attributes represented by semantic relationships, and spatial relationships, we merge the micro neighborhoods to form a larger region, called the *macro neighborhood*. Essentially, we form a larger spatial entity from the spatial units that are spatially related and are similarly behaving with respect to their attributes. The goal of constructing a macro neighborhood is to characterize a region. We employ the Jaccard coefficient to quantify the similarity. Since the data from multiple domains is used, a macro neighborhood represents the spatial characterization of a region in multiple domains. The characterized region is the input for the detection of anomalous geospatial trajectories. A composite feature vector for each macro neighborhood is generated and the associations (called path associations) are identified between the macro neighborhood and a micro neighborhood on a geospatial trajectory. This continues in a drill down manner to identify geospatial trajectory that has a path association with micro neighborhoods that are not part of the trajectory. We illustrate these various steps involved in our approach in figure 1.

3. Detection of Anomalous Geospatial trajectories

In this section, we present the details of each of the processing steps outlined in section 2, namely, generation of micro neighborhood, identification of spatial and semantic relationships, spatial characterization through the generation of macro neighborhood and finally identification of path associations to detect anomalous geographic trajectories.

3.1 Generation of Micro neighborhoods

We consider the smallest geographic entity as a spatial unit. We assume S be the set of spatial units, $S = \{s_1, s_2, \dots, s_n\}$. A spatial unit can be formally defined as follows.

Definition 1: [Spatial Unit]: Each spatial unit, $s_i \in S$ 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 $a_i = \{a_{i1}, a_{i2}, a_{i3}, \dots, a_{im}\}$.

The attributes $\{a_{i1}, a_{i2}, a_{i3}, \dots, a_{im}\}$ of each s_i are transformed into categorical values (binary) to form a feature vector $f_i = \{f_{i1}, f_{i2}, f_{i3}, \dots, f_{im}\}$. For example, a spatial unit can be a city characterized by certain attributes that capture the underlying natural or man made spatial processes such as agricultural produce, the area covered by water, number of importers located in the area, and so on. Thus, if an airport is present in that city, its corresponding feature value is set to 1, otherwise it is set to 0. In the following, we describe the procedure to determine the micro neighborhood, originally proposed in [AJA04].

The micro neighborhood is constructed based on the concept of Voronoi tessellations [OBS00], where each spatial unit is allotted its region of influence in terms of an intersecting half plane. This directly follows from the process of creating a Voronoi diagram. To construct a Voronoi diagram, first

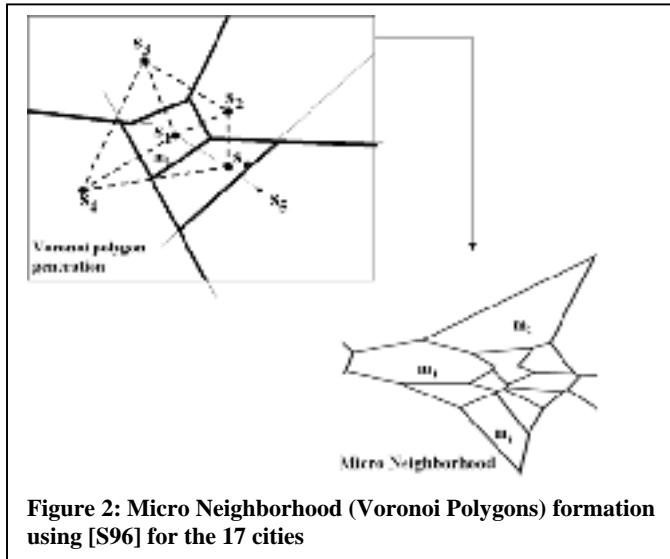


Figure 2: Micro Neighborhood (Voronoi Polygons) formation using [S96] for the 17 cities

two spatial units are connected by a line segment, which is then bisected into two half planes. For example, consider figure 2 where s_1 to s_6 are six spatial units. The dotted lines indicate line segments connecting these special units. The solid lines represent the bisecting lines, which form a Voronoi polygon surrounding s_1 . We denote the Voronoi polygon surrounding a spatial unit s_i with $V(s_i)$. The Voronoi polygons form a polygonal partition of the plane, called the Voronoi diagram of the entire set of spatial units S , denoted by $V(S)$. It is composed of Voronoi edges and Voronoi vertices forming a polygonal cell around each spatial unit. As new spatial units are added, more half planes are formed, and the region of influence of the spatial unit is the intersection of the half

planes. Thus, $V(S)$ is comprised of the entire proximity information about S in an explicit and computationally useful manner. Thus:

Definition 2 [micro neighborhood]: Given a set of spatial units $S = \{s_1, s_2, \dots, s_n\}$, a micro neighborhood m_i is the polygon bounded by a Voronoi polygon $V(s_i)$.

The micro neighborhood thus represents the region of influence of a spatial unit or its dominance over another spatial unit for an attribute [OBS00]. In other words, a feature located on one side of the bisector is closer to that half plane than the other.

Definition 3 [Region of Dominance]: Given two spatial units s_i and s_j , and their feature vectors f_i and f_j such that $f_{ik} \in f_i$ and $f_{jk} \in f_j$, we define the *dominance* of s_i over s_j as follows: $\text{dominance}(s_i, s_j)$ iff $d(f_{ik}, s_{ik}) \leq d(f_{jk}, s_{jk})$, where d is a distance function.

Thus, a micro neighborhood is a bounded polygon m_i around each spatial unit s_i , encompassing this knowledge of dominance for each spatial unit. The formation of micro neighborhood is performed using an application, the Two-Dimensional Quality Mesh Generator and Delaunay Triangulator [S96], to generate the Voronoi polygons. For the generation of micro neighborhoods, for the application to transshipment in cargo routes, we consider the spatial locations of various cities as the spatial units, and use them to generate the micro neighborhoods.

3.2 Identification of Spatial and Semantic relationships

Spatial Relationships: Spatial units are governed by the spatial relationships such as topological (adjacent, inside, disjoint, ..), direction (north, south, east, west, north-east ...), and distance. A spatial relationship can be formally defined as follows [EKS97]:

Definition 4: [Spatial relationship]: Given a set of spatial units $S = \{s_1, s_2, \dots, s_n\}$ and their corresponding micro-neighborhoods $\{m_1, m_2, \dots, m_n\}$, we say that there exists a spatial relationship between m_i and m_j , denoted by $\text{sp}(m_i, m_j)$, iff there exists either a topological relationship between s_i and s_j , direction relationship between s_i and s_j , or a distance relationship between s_i and s_j .

In our approach, we utilize the inherent adjacency relationships, which can be extracted from the Voronoi polygons. Recalling from section 3.1, the Voronoi polygons are formed by creating an edge connecting the two spatial coordinates for the spatial units, and then bisecting the edge to divide the region into two half planes. The connecting edge essentially leads to the formation of two adjacent polygons, sharing a common edge. In other words, if there is a common edge for any two spatial units, then there exists an adjacency relationship between them. This can be utilized to identify if the two spatial units are neighbors based on their spatial relationship. Although other topological and distance relationships can be identified based on this knowledge, however, in this paper, we limit our focus to adjacency relationships only.

Semantic Relationships: Identifying the existence of semantic relationships among the spatial units can be accomplished by first determining the similarity between these spatial units with respect to their feature vectors. This can be performed using similarity coefficients such as the m-coefficient, Jaccard coefficient, etc. [KR90]. In this paper, we use the Jaccard coefficient (JC) since we are primarily interested in the similarity among the features, but not the dissimilarity among them. A comparative analysis of the feature vectors is deferred to future work. JC is used to quantify similarity or dissimilarity of binary valued variables. Here the similarity or dissimilarity of two spatial units can be calculated using the contingency table as shown in table 1.

| | | |
|---|---|---|
| | 1 | 0 |
| 1 | 1 | 1 |
| 0 | 2 | 0 |

Table 1: Feature similarity for two Objects

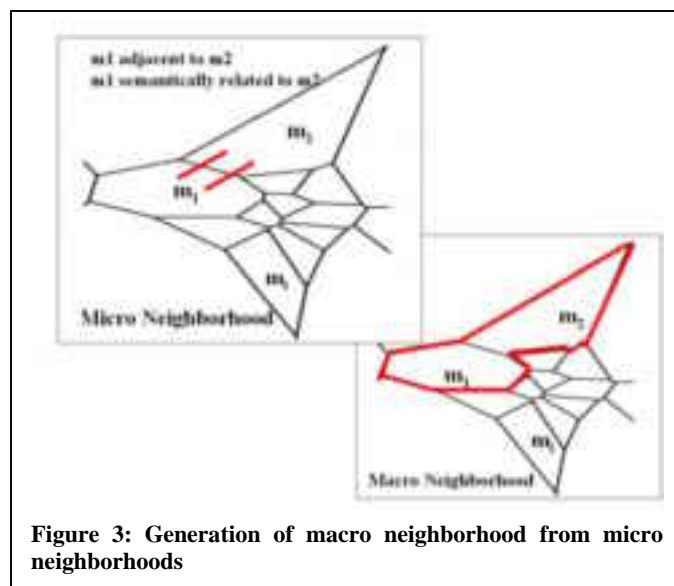
For our approach, similarity is quantified in terms of the positive match (1-1 match) and the positive mismatch (1-0 or 0-1 mismatch); it does not give importance to the negative match (0-0 match). We utilize the JC for this approach that gives more importance to a 1-1 match and 0-1 or 1-0 mismatch unlike the matching or m-coefficient, which considers all [KR90]. An example follows based on table 1. $JC = \text{positive match} / (\text{positive match} + \text{positive mismatch}) = 1 / (1 + 1 + 2) = 0.25$. Here, the agreement of 1-1 is considered more important than the agreement of 0-0

(negative match). That is, the positive match is given more weight. Next, we formally define the semantic relationship in terms of the similarity of feature vectors for the micro neighborhoods.

Definition 5 [Semantic Relationship]: Given two micro neighborhoods m_i and m_j , and the corresponding feature vectors f_i and f_j , the degree of the semantic relationship between m_i and m_j , denoted by $sm(m_i, m_j) = JC(f_i, f_j)$.

Note that the higher the value of JC, the stronger the semantic relationship between m_i and m_j .

3.3 Spatial Characterization



Our approach to the identification of anomalous geospatial trajectories is based on spatial characterization of data from multiple domains. Spatial characterization essentially aggregates similarly behaving micro neighborhoods with spatial and semantic relationships into one region called macro neighborhood [AJA04]. Recall that a micro neighborhood leads to the identification of a region of influence of the underlying spatial attributes in the spatial unit (in our example, a city). However, if the spatial units that have a spatial relationship with each other and behave similarly in terms of their features, they can be merged to form a larger region of influence of the underlying spatial processes. The formation of macro neighborhood is an aggregation step, where the micro neighborhoods are merged based on the spatial and semantic relationships. Here the semantic relationship

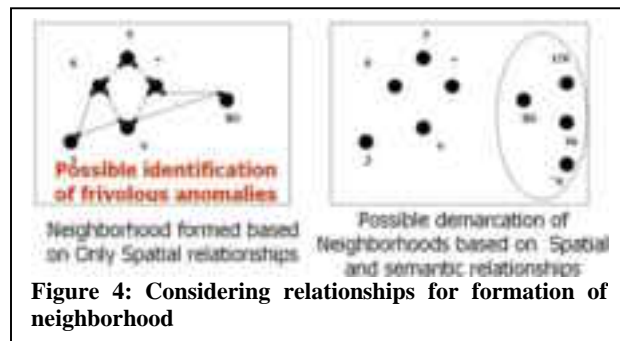
Here the semantic relationship

is quantified in the form of the similarity between them using the JC. In the following, we define the macro neighborhood, adapted from [AJA04].

Definition 6: [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 an $sp(m_i, m_j)$ and $sp(m_i, m_j) \geq \delta$, where δ is the threshold.

In the above definition, $sp(m_i, m_j)$ refers to the spatial relation between the micro neighborhoods m_i and m_j , $sp(m_i, m_j)$ refers to the degree of the semantic relationship between them, and δ is a user defined threshold value. Essentially, the contour of the macro neighborhood is formed by eliminating the common edge(s) between the adjacent micro neighborhood polygons 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 figure 3.

Most approaches limit the neighborhood identification to be based on spatial relationships only [SLZ01, EFKS98]. This is only a representation of spatial autocorrelation, which states that nearby things are closely related in terms of distance and other parameters. However, this can overlook the combined effect of the inherent spatial heterogeneity along with spatial autocorrelation. Due to this, several important processes can be misrepresented in the neighborhood.

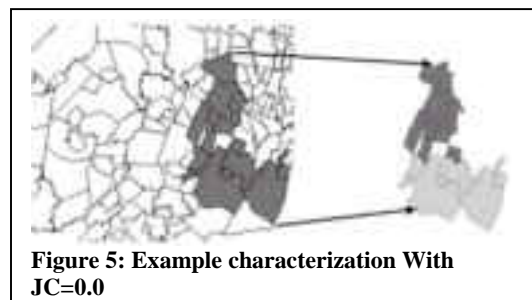


For example, if we are considering anomaly detection for a certain parameter, in a neighborhood as shown in the figure 4, then some extreme values may be considered to be part of the neighborhood. This is mainly based on spatial autocorrelation even though they may be part of some other semantically related neighborhood. To account for this, we consider semantic relationships to take into account the values for various spatial features, which can be natural or

manmade. This will take into consideration the heterogeneity in the region with respect to similarity or dissimilarity of these features. Thus we consider both spatial and semantic relationships so that both spatial autocorrelation and heterogeneity are taken into account.

Formation of micro neighborhoods leads us to the characterization of the spatial region in terms of the features associated with each spatial unit. In some cases, it is possible that the feature vector be refined. For example, if two spatial units, in our case two cities, lie across state or country borders, the feature of border will have more importance, or have more significance to other features associated with a change of the country or state, such as features pertaining to the government, economic factors, etc. Thus, the key here is to attach as much semantic knowledge as possible to the spatial characterization.

In the domain of US Customs, it is also possible to exploit the knowledge of existing routes in the spatial characterization such that the region is characterized in terms of the features of the spatial unit and the routes that pass through the unit. For example, a region can be selectively characterized in terms of features associated with import cargo routes and import related features such as produce in the region, carriers in the region, etc. This can be done by the analysis based on an additional coefficient that can capture the overlap with the routes passing through the neighborhood. We next discuss results of some experimentation performed for spatial characterization.

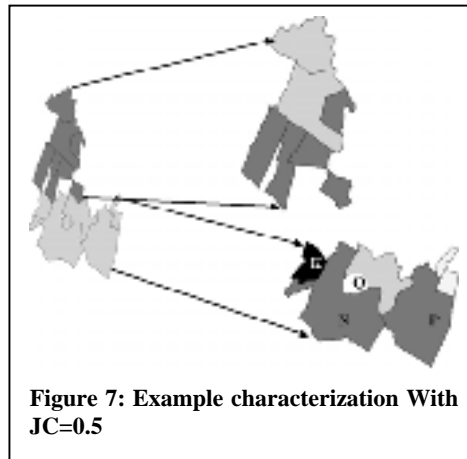
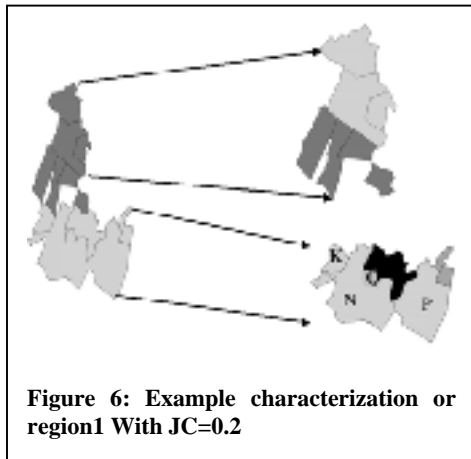


We have experimented with data for 17 US cities (shown in figure 5). For the purposes of validation and better understanding of the process, we have considered cities where real data is readily available. Specifically, we have considered cities in northern New Jersey with populations of 10,000 or greater and for April 2000. However, for all practical purposes, these cities can be from anywhere in the world. Each city is qualified by their geographical coordinates and non-spatial attributes, which include area under water, area of the land, number of

females, number of males, number of housing units, number of vacant units, etc. [USGS03]. We have evaluated the spatial relationships using the Voronoi polygons and the data about edges between the spatial units generated by the Triangle program [S96]. The similarity between the polygons is evaluated on the basis of a vector comprised of 20 features. We describe the results of the characterization in different cases:

- I. No similarity coefficient is used ($JC = 0$), i.e., macro neighborhood formed based on spatial relationships only.
- II. JC with different threshold values (0.2 and 0.5), i.e., macro neighborhood formed based on spatial and semantic relationships.

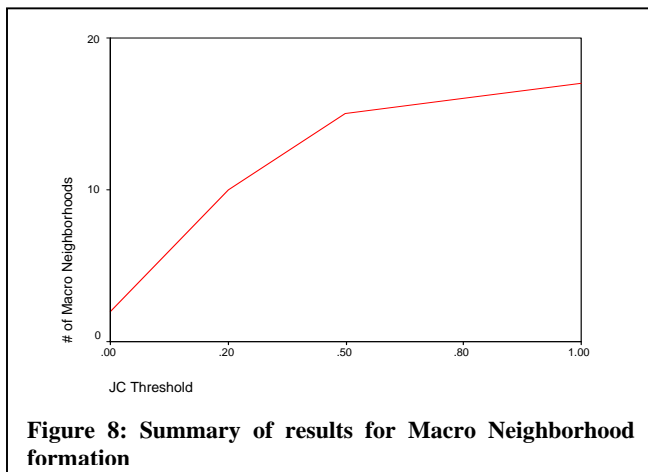
These results are pictorially depicted in figures 6 and 7, which are composed using Arc View [AV04]. The process of creation of shape files is not automatic at this point, however, this will be addressed in



future work. As it can be seen from figures 6 and 7, when we consider only the spatial relationships, the region under consideration has been characterized as two macro neighborhoods only. However, when we start considering the spatial as well as the semantic relationships, the

region starts to be characterized into more refined macro neighborhoods. More granular macro neighborhoods start emerging with the refining of the semantic relationships in terms of the JC . It is important to note here that the macro neighborhoods at a more refined level of the JC are evolved from the macro neighborhoods at the lower level of JC . For example, the group of CITY_N, CITY_P at $JC=0.5$ is the macro neighborhood evolved from the macro neighborhood of CITY_K, CITY_N, CITY_P, CITY_Q at $JC=0.2$.

Similarly, this macro neighborhood at $JC=0.2$ is formed from the macro neighborhood at $JC=0.0$,



which is shown in figures 6 and 7 where the cities are labeled. As can be seen in figure 5, when $JC=0.0$, the region is characterized into two regions only. When a smaller number of cities is considered, it is possible that the characterization will be in the form of one large neighborhood. Thus, one may conclude that simply considering spatial relationships does not give a well refined characterization. But, when we consider semantic relationships in the form of similarity of features quantified by JC , we observe that a more refined characterization starts to emerge. It is evident in figures 6 and 7 where JCs are 0.2 and 0.5, respectively. We can see more refinement in each of the macro

neighborhoods at the higher levels of JC . This new macro neighborhoods at $JC=0.2$ are from the macro neighborhood formed at $JC=0.0$. Similarly, the macro neighborhoods formed at $JC=0.5$ are from the macro neighborhood formed at $JC=0.2$. Thus, the neighborhood formation is consistent across the

refinement of JC. As we can see from figure 8, the number of macro neighborhoods increases with the increase in the threshold of the JC. This is because, the macro neighborhoods are formed based on a higher threshold of similarity between the micro neighborhoods. At JC threshold of 1.0, we are trying to identify spatial units that are similar in all features, which is a rare or non-existent case. Thus, at this threshold, the number of macro neighborhoods equals the number of micro neighborhoods i.e. 17, as none of them are merged. However, at a JC threshold of 0.2, many spatial units will be similar, thus the number of macro neighborhoods formed is 10 with more macro neighborhoods. For JC threshold of 0.0, we are only considering the spatial relationships, thus two big macro neighborhoods are formed. We next discuss the detection of anomalous geospatial trajectories.

3.4 Identification of spatio-semantic path associations

The spatial characterization can be utilized to analyze a geospatial trajectory to identify non-obvious associations between points on the geospatial trajectory and the other spatial units that are not obvious. For example, in case of our example of the cargo route in the transshipment, there is a possibility of a hidden association apart from the predefined points such as the origin and destination of the route. In figure 9, the region is characterized based on drug activity and economic features. Although the route shown on the manifest is from B to C, it is possible to find if the shipment being carried has any association with other locations such as A, which can then help in identifying potential transshipments.

To start with, the region is characterized into 5 macro neighborhoods, M1, M2, M3, M4 and M5. The cargo manifest shows that the point of origin and destination are spatial units B and C, respectively. Note that A, B, and C may belong to different countries. When we analyze the associations of the micro

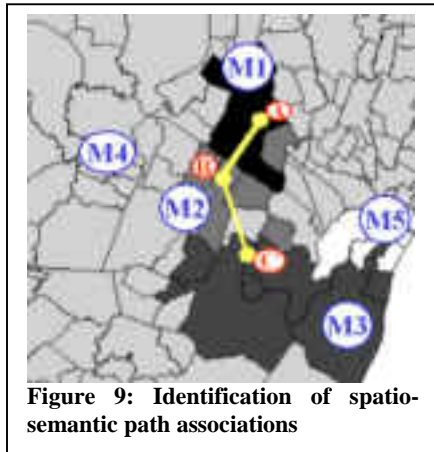


Figure 9: Identification of spatio-semantic path associations

neighborhood for B, which we call as the Origin micro neighborhood, it is found that apart from being strongly related to its own macro neighborhood M2, B is also related strongly with the macro neighborhood M1. This strong association can be quantified between the origin micro neighborhood such as for city B and Macro neighborhood such as M1.

Earlier, we have defined a spatial and semantic relationship between various micro neighborhoods to form a macro neighborhood in the generation of spatial characterization. However, there can be many micro neighborhoods. Now that we have the macro neighborhood, we can first identify the association between the origin micro neighborhood and the complete set of macro neighborhoods defining the region. Once the strongest association is found, the association is further refined to find the micro neighborhood, which are most strongly associated with the

origin micro neighborhood. We take the following approach to identify the above associations.

Since the macro neighborhood essentially is a polygon of similarly behaving micro neighborhoods, it can be represented by a composite vector. This composite vector will take into consideration the values for all of the micro neighborhoods within this macro neighborhood. It is formed by computing the average of the attributes associated with the spatial unit of each micro neighborhood. Thus, we have an attribute vector associated with each macro neighborhood. Now we can transform this into a binary valued feature vector representing each macro neighborhood. We define the composite feature vector below.

Definition 7 [composite feature vector] Let M be a macro neighborhood comprising of a set of micro neighborhoods $\{m_1, m_2, \dots, m_n\}$, and a set of attributes $\{a_{i1}, a_{i2}, a_{i3}, \dots, a_{ip}\}$ for the spatial unit s_j of each $m_i \in M$. A composite vector $cf = \{cf_1, cf_2, cf_3, \dots, cf_p\}$ of M , where each cf is the average of $(a_{i1}, a_{i2}, a_{i3}, \dots, a_{ip})$ transformed into a binary valued feature.

Once we have the composite feature vector for each macro neighborhood, we identify the semantic relationships between the feature vector of the origin micro neighborhood and each composite feature vector of the set of macro neighborhoods. We quantify this using a path association. We define the path association as follows:

Definition 8: [**Path Association**] Given a set of macro neighborhoods $M = \{M_1, M_2, \dots, M_m\}$ and their corresponding composite feature vectors $cf_i = \{cf_{i1}, cf_{i2}, cf_{i3}, \dots, cf_{ip}\}$, and given the origin micro neighborhood m_o such that $m_o \notin M_i$, the degree of the path association, $pa(M_i, m_o) = JC(cf_i, f_o)$.

Note that the higher the value of $pr(M_i, m_o)$, the stronger the path association between M_i and m_o .

The idea here is to find the association of the locations on the routes i.e. B to C with other locations. Specific to this example, the goal is to find the association of the cargo being shipped to other locations apart from B to C. This would lead to the identification of possible transshipment sources. We next outline an algorithm (Algorithm 1) for detection of anomalous geospatial trajectories. Essentially, we

Algorithm 1: [Detection of anomalous geospatial trajectories]

Input: A set of geospatial trajectories, m_i , the set of micro neighborhoods, δ

Output: Anomalous geospatial trajectory and associated external micro neighborhood m_e

Procedure:

Compute the macro neighborhoods from m_i

Identify the origin micro neighborhood m_o

Compute the $cf_p = \{cf_{p1}, cf_{p2}, cf_{p3}, \dots, cf_{pm}\}$ for each macro neighborhood

M_p

Where some $s_{ij} \in m_q$ and some $m_q \in M_p$

for all m_o

for all M_p

{
if $pa(m_o, M_p) \geq pa(m_o, M_k)$, where $m_o \in M_k$
for all $m_i \in M_p$

{
if $sm(m_o, m_i) \geq \delta$

flag as anomalous geospatial trajectory
 $m_e = m_i$

}
}

its own macro neighborhood. Once a macro neighborhood apart from its own is identified, we can further drill down and find an association to a micro neighborhood. We indicate such a micro neighborhood as an external micro neighborhood m_e . Here we also find an association that this origin micro neighborhood has with its own macro neighborhood. Once such a micro neighborhood is identified, we can mark the route as possible transshipment or anomalous route.

4. Related Work

Several studies in spatial data mining have addressed the issue of identifying spatial neighborhood. [EKS97] addresses several database primitives for spatial data mining. They identify neighborhood graph of spatial objects, based on spatial relationships such as topological, direction and distance relationships. However, this process of identifying the spatial relationship predicate could be an intricate process in itself. [EFKS98] utilizes the spatial database primitives in spatial characterization and identification of spatial trends starting from a spatial object. However, it does not account for spatial auto correlation and spatial heterogeneity. Many approaches assume that observations are independent, yet if spatial autocorrelation exists, then the assumption is no longer true and could lead to inferential errors. Similarly, if spatial autocorrelation is accounted for, then spatial heterogeneity also needs to be quantified.

[KKL97] is a clustering technique, which uses Delaunay triangulation technique that connects the points by edges if they are within a certain threshold proximity. This approach also finds outliers as a by-product of clustering. However, it does not consider the semantic and implicit spatial relationships to identify the clustering structure. Further, 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 [OBS00]. 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

have the set of macro neighborhoods formed during spatial characterization, where each macro neighborhood is comprised of micro neighborhoods. Given that a cargo route is comprised of some spatial units corresponding to some origin micro neighborhoods, each of these have their feature vectors associated with them. Therefore, we can find the associations between the composite feature vectors for each macro neighborhood and the origin micro neighborhood. We identify associations that are equal to or stronger than the association that this micro neighborhood has with

spatial relationships. Although it is computationally efficient to create the Delaunay triangulations than the Voronoi polygons [A91] and subsequently derive the Voronoi diagrams, Voronoi diagrams capture the proximity more completely than a Delaunay triangulation.

[AJA04] utilizes the spatial relationships defined in [EKS97] and augments the neighborhood definition by taking into consideration the semantic relationships in terms of the various attributes, which essentially performs geospatial characterization. This takes into consideration both spatial autocorrelation and spatial heterogeneity. However, it does not identify anomalous geospatial trajectories.

5. Conclusions and Future Work

In this paper, we have proposed a novel approach for detecting anomalous geospatial trajectories using spatial characterization and spatio-semantic path associations. This process is comprised of several steps: First, we have considered atomic geospatial units and generated *micro neighborhoods* around them. Second, we have merged these micro neighborhoods into similarly behaving regions called *macro neighborhoods*, by considering the spatial and semantic relationships among them. Each macro neighborhood is then associated with a composite feature vector. As a last step, we divide the geospatial trajectory into segments and associate each segment with its feature vector. The composite feature vector of the characterized region and the feature vector of the segments are used to identify associations. Any strong association between a macro neighborhood and a part of the trajectory that does not reside in it is identified as a potential anomaly.

The feature vector for a micro neighborhood can consist of many features, however, only a few can be critical for the neighborhood identification. Moreover, there may be some features that have more importance in certain situations. For example, if two spatial units such as cities are across state or country borders then one of the features would need to be added or weighted differently to reflect this, as the spatial behavior would be different. As part of our future work, we intend to incorporate selectivity criteria to more accurately describe the behavior of the micro neighborhoods. Further, we have used Jaccard coefficient for similarity matching, however we would like to explore other coefficients as well such as Tanimoto, simple matching (M), Russel and Rao (RR), Dice and Gower, etc.

References

- [AJA04] N.R. Adam, V.P. Janeja, V. Atluri, "Neighborhood Based Detection of Anomalies in High Dimensional Spatio-temporal Sensor Datasets" ACM Symposium on Applied Computing, March 2004.
- [AV04] ArcView 8.3, <http://www.esri.com/>
- [CIA03] Central Intelligence Agency(CIA), The world fact book, <http://www.cia.gov/cia/publications/factbook/geos/sp.html>
- [EKS98]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.
- [KR90] Kauffman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- [EKS97] M. Ester, H. P. Kriegel, and J. Sander. Spatial Data Mining: A Database Approach. In Proceedings of the International Symposium on Large Spatial Databases, Germany, July 1997, pp. 47-66.
- [S96]J. R. Shewchuk, Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. First Workshop on Applied Computational Geometry, Pennsylvania , pages 124-133, ACM, May 1996
- [SLZ01]S. Shekhar, C. Lu, and P. Zhang. Detecting Graph-Based Spatial Outlier: Algorithms and Applications (A Summary of Results). In CSE Department, UMN, Technical Report 01-014, 2001
- [OBS00] A. Okabe, B. Boots, K. Sugihara, S. Chiu. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. pp. 291-410. John Wiley, 2000.
- [USC03A]TECHNICAL INFORMATION FOR PRE-ASSESSMENT SURVEY (TIPS): http://www.cbp.gov/ImageCache/cgov/content/import/regulatory_5faudit_5fprogram/focused_5fassessment/fap_5fdocuments_5f10_5f2003/exh5l_2epdf/v2/exh5l.pdf
- [USC03B] Public Health and Safety http://www.cbp.gov/xp/cgov/enforcement/ice/investigations/public_health.xml
- [USGS03] U.S. National Atlas Cities<http://pubs.usgs.gov/of/2003/of03-001/data/basemaps/usa/cities/>