Chapter 10

SPATIAL OUTLIER DETECTION

"Time and space are modes by which we think and not conditions in which we live." – Albert Einstein

1. Introduction

Spatial outliers are objects which have *behavioral* attribute values that are distinct from those of their surrounding *spatial* neighbors. Thus, *spatial continuity* plays an important role in the identification of anomalies. This is an analogous principle to the concept of temporal continuity, which was discussed in the chapters on time series outlier detection. One of the most fundamental rules of spatial data is as follows [455]:
" Everything is related to everything else, but nearby objects are more related than distant objects."

Spatial data does not contain only spatial attributes, just as temporal data does not necessarily contain only temporal attributes. Instead, spatial locations form the *contextual points* at which other behavioral attributes of interest are measured. Thus, two kinds of attributes may be available:

- *Behavioral Attributes:* This is the attribute of interest which is measured for each object. For example, this could correspond to sea surface temperatures, wind speeds, car speeds, disease outbreak numbers, the color of an image pixel, etc. It is possible to have more than one behavioral attribute at a spatial location in given application.
- *Contextual Attributes (Spatial Location):* This is the location of interest at which the behavioral attribute is measured. Typically, this would contain two or three dimensions, when the data is expressed in terms of coordinates. In some cases, the contextual

C.C. Aggarwal, *Outlier Analysis*, DOI 10.1007/978-1-4614-6396-2_10,

attributes may be more complex, and may be expressed at the granularity of a *region of interest*, such as a county, zip-code etc. Alternatively, in an imaging application, the contextual attributes may correspond to individual pixels.

Spatial data shares a number of similarities with time-series data, in which one or more properties of interest (behavioral attributes) are measured at a given moment in time (contextual attribute). In fact, in *spatiotemporal* data, the contextual attributes may also contain a temporal component. This can be used to determine important spatiotemporal anomalies (or events) based on the underlying dynamics. For example, the dynamics of behavioral attributes such as humidity, wind speeds, sea surface temperatures and pressure can be used in order to identify and predict anomalous weather events. In such cases, both spatial *and* temporal continuity can play an important role in the prediction. It is also possible for the data to be *purely* spatiotemporal, in which no other behavioral attributes are present, and the trajectories of objects are measured over time. In such cases, no attribute needs to be treated as behavioral, since a joint analysis of both components provides the best insights in many applications. In some cases, it may be helpful to treat the temporal component as the contextual attribute, and the spatial components as the behaviorial attributes. For example, in a two-dimensional *real-time* trajectory mining application, this can be modeled as a bivariate time series, in which the evolving X-coordinate and Y-coordinate values are individual time series. In the *offline* trajectory shape analysis scenario, anomalies may correspond to unusual shapes, irrespective of their temporal provenance. The latter case is mostly a spatial analytics scenario, and the temporal aspects of the problem are limited. Therefore, trajectory-based applications can be modeled in multiple ways, depending upon the needs of the underlying application.

Spatial data is common in many real applications, such as the following:

- *Meteorological Data:* Numerous weather parameters are typically measured at different geographical locations, which may be used in order to predict anomalous weather patterns in the underlying data [510].

- *Traffic Data:* Moving objects may be associated with many parameters such as speed, direction etc. The location of an object is its contextual attribute. In many cases, such data is also spatiotemporal, since it has a temporal component. Finding anomalous behavior of moving objects [83] can provide numerous insights.

- *Earth Science Data:* The land cover types at different spatial locations may be the behavioral attributes. Anomalies in such patterns provide insights about anomalous trends in human activity such as de-forestation or other anomalous vegetation trends [287].

- *Disease Outbreak Data:* Data about disease outbreaks is often aggregated by spatial locations such as zip-code and county. Anomalous trends in such data [465] can provide information about the causality of the outbreaks.

- *Medical Diagnostics:* MRI and PET scans are spatial data in two or three dimensions. The detection of unusual localized regions in such data can help in detecting diseases such as brain tumors, the onset of alzheimer disease, and multiple sclerosis lesions [374, 206, 466, 418].

- *Demographic Data:* Demographic attributes such as age, sex, race, and salary can be used in order to identify demographic anomalies. Such information can be useful for target-marketing applications.

As in the case of temporal data, *abrupt changes in the behavioral attribute, which violate spatial continuity* provide useful information about the underlying contextual anomalies. For example, consider a meteorological application, in which sea surface temperatures and pressure are measured. Unusually high sea surface temperature in a very small localized region is a hot-spot which may be the result of volcanic activity under the surface. Similarly, unusually low or high pressure in a small localized region may suggest the formation of hurricanes or cyclones. In all these cases, spatial continuity is violated by the attribute of interest. Such attributes are often tracked in meteorological applications on a daily basis. In Figure 10.1, a color coded map of the sea surface temperatures on October 1, 2012 from the *NOAA Satellite and Information Service* is illustrated. Unusually high temperature anomalies are illustrated in red, whereas unusually low temperature anomalies are illustrated in blue.

In the context of *spatiotemporal* data, both spatial and temporal continuity is used for the purposes of outlier analysis. For example, a sudden change in the velocity of a few cars in a small localized region may suggest the occurrence of an accident or other anomalous event. Similarly, evolving events such as hurricanes and disease outbreaks are spatiotemporal in nature. Spatio-temporal methods for outlier detection [113, 114] are significantly more challenging because of the additional challenges of modeling the temporal and spatial components jointly.

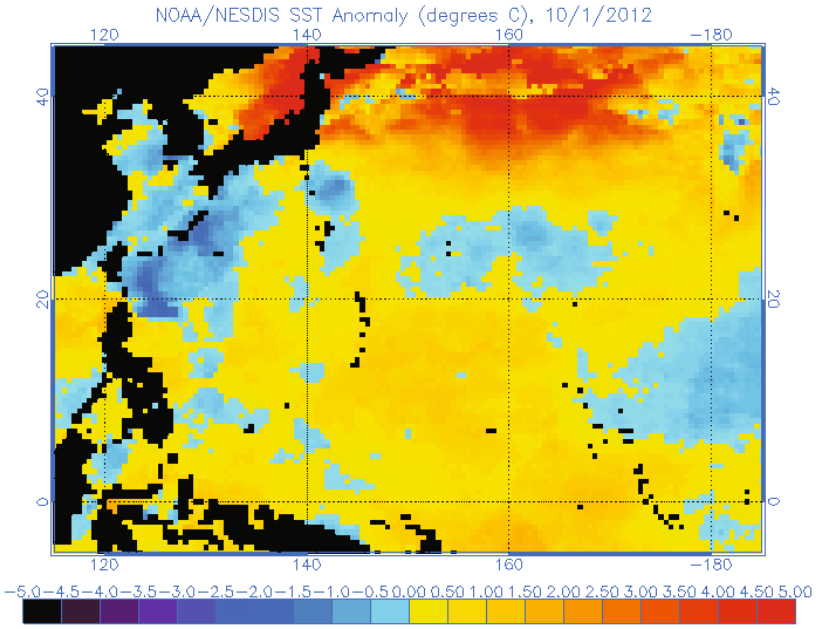


Figure 10.1. Sea surface temperature anomalies. Source: NOAA Satellite and Information Service

There are two main characteristics of spatial data, which are commonly used in outlier detection algorithms:

- *Spatial Autocorrelations:* This corresponds to the fact that behavioral attribute values in spatial neighborhoods are closely correlated with one another. However, unlike temporal data, where future values of the time-series are unknown, the values in all spatial directions of a data point can be used.

- *Spatial Heteroscedasticity:* This corresponds to the fact the variances of the behavioral attribute depend on spatial location [433].

While the first property is the primary criterion for outlier analysis, the second has also proven to be useful in many scenarios. This is because when certain regions are likely to have greater variance as a matter of expectation, then abrupt changes in those regions are less likely to be significant. Such insights have lead to local methods [433], which are based on ideas derived from local density-based methods (LOF) [78].

Numerous methods have been proposed in the literature for detecting spatial outliers. The primary ones among them use *variations of the behavioral attribute* within a neighborhood in order to define outliers. Such

outliers use either multidimensional analysis methods or graph-based methods. In addition, many of the temporal auto-correlation methods discussed in the previous chapter can also be generalized to the spatial domain, when the data is completely specified over the various dimensions.

Most of the work on spatial outliers is about finding *abrupt changes* which violate spatial auto-correlations. Such outliers are *contextual* outliers. While the standard statistical tests for deviation detection are useful in this case, it is sometimes useful to intuitively visualize the key outlier points. The spatial nature of the data also lends itself to more intuitive visualization methodologies such as visualization. Two examples of such methodologies are *variogram clouds* and *pocket plots* [203, 354]. The former will be described in detail in this chapter.

As in the case of time-series databases, it is also useful to find unusual *shapes of patterns implied by the distribution of the behavioral attribute* in a database of multiple spatial distributions. For example, the color distribution in an image or MRI scan may correspond to an unusual shape, when compared to other images in the database. Such an image may be of interest for further analysis. Such outliers are *collective* outliers in the context of spatial data.

Supervised methods are also very useful in the spatial domain, where it is desirable to determine unusual shapes from multiple spatial patterns. For example, while many conditions such as weather patterns of interest, or brain tumors in MRI scans may be rare on a *relative* basis, a significant amount of training data may be available on an *absolute* basis for modeling purposes. In medical applications, large numbers of pathological examples are sometimes available for modeling purposes. Similarly, many examples of pathological patterns of unusual shapes may be available in meteorological and earth science applications. In such cases, it is useful to utilize supervision for the purposes of outlier detection. Supervised methods are particularly useful in the context of outlier detection in such cases, because of the unusually high complexity of a database containing multiple spatial patterns. Such methods are closely related to topics such as image classification. The topic of image classification is a large area of interest in its own right. While this is beyond the scope of this book, some discussion of related work will be provided in this chapter.

A close relationship exists between temporal and spatial outlier detection, because both methods use concepts of *behavioral attribute continuity with respect to one or more contextual attributes*. The main difference lies in the fact that spatial contextual attributes are often multidimensional, whereas time is a single attribute. Furthermore, time is uni-

directional, where only values in the past are known, whereas spatial attributes are known in the different directions of all axes. Nevertheless, in many applications, these differences are not significant enough to invalidate the applicability of temporal methods. While recent work has adapted temporal techniques to some spatial applications such as anomalous image shape detection [469], many other temporal techniques have the potential for use in the spatial domain. This chapter will point out the different temporal techniques, which are also applicable to the spatial domain. It should be noted that in some cases, these temporal methods are indeed not applicable, especially when the spatial contextual attribute cannot be expressed in terms of a comprehensive set of coordinates in a multidimensional plane. For example, the spatial attribute may be specified with a rough granularity, such as a county or zip-code, or may be available only for a small subset of points in the spatial plane.

This chapter is organized as follows. In the next section, neighborhood-based algorithms for outlier analysis will be studied. Both multidimensional and graph-based methods will be studied in this section. Autoregressive models for anomaly detection are presented in section 3. Visual methods for detecting spatial outliers with variogram clouds are addressed in section 4. Unusual shape discovery in multidimensional spatial data will be addressed in section 5. Methods for spatiotemporal outlier detection are presented in section 6. The use of supervision for anomaly detection in spatial data is studied in section 7. The conclusions and summary are presented in section 8.

2. Neighborhood-based Algorithms

Neighborhood-based algorithms can be very useful in the context of a wide variety of tasks. In these algorithms, abrupt changes in the spatial neighborhood of a data point are used in order to diagnose outliers. Such algorithms depend upon the exact way in which the spatial neighborhood is defined, the function used to combine these neighborhood values into an expected value, and the computation of the deviations from the expected values. The neighborhood may be defined in many different ways [3, 268, 317, 401–404], depending upon the nature of the underlying data.

- *Multidimensional Neighborhoods:* In this case, the neighborhoods are defined on the basis of multidimensional distances between data points.
- *Graph-based Neighborhoods:* In this case, the neighborhoods are defined by linkage relationships between spatial objects. Such

neighborhoods may be more useful in cases, where the location of the spatial objects may not correspond to exact coordinates (eg. county or zip code), and graph-representations provide a more general modeling tool.

This section will study method for neighborhood-based outlier detection with the use of multidimensional and graph-based methods.

2.1 Multidimensional Methods

While traditional multidimensional methods can also be used to detect outliers in spatial data, such methods do not distinguish between the contextual attributes and the behavioral attribute. Therefore, such methods are not optimized for outlier detection in spatial data, especially in cases where the outliers are defined on the basis of the behavioral attribute.

Numerous methods have been defined, which use the spatial neighborhood of the data with the use of multidimensional distances on the spatial (contextual) attributes. Thus, the contextual attributes are used for determining the k nearest neighbors, and the deviations on the behavioral attribute values are used in order to predict outliers. A variety of distance functions can be used on the multidimensional spatial data for determination of proximity. The choice of the distance function is important, because it defines the choice of the neighborhood which is used for comparison with the true value. For a given spatial object o, with behavioral attribute value $f(o)$, let $o_1 \ldots o_k$ be its k-nearest neighbors. Then, a variety of methods may be used to compute the predicted value $g(o)$ of the object o. The most straightforward method is the mean:

$$g(o) = \sum_{i=1}^{k} f(o_i)/k$$

Alternatively, $g(o)$ may be computed as the median of the surrounding values of $f(o_i)$, in order to reduce the impact of extreme values. Then, for each data object o, the value of $f(o) - g(o)$ represents a deviation from predicted values. The extreme values among these deviations may be computed using a variety of methods discussed in Chapter 2. These are reported as outliers.

2.1.1 Local Outliers. An observation in [433] is that all local deviations are not equally important from the perspective of outlier analysis. For example, consider the case where the sea-surface temperatures are being measured at different spatial locations. In some spatial

regions, the changes in temperatures may naturally show larger variations than others. Therefore, the same variation cannot be treated with equal importance in all regions. Specifically, the outlier scores in high variance regions need to be suppressed. In such cases, it may be useful to quantify the changes around a data point in a local way. For example, instead of using the value of $f(o) - g(o)$ as discussed above, it is possible to use a normalized value of $\frac{f(o)-g(o)}{L(o)}$, where $L(o)$ represents a *spatially local* quantification of the deviations around o. For example, $L(o)$ could represent the standard deviations of the behavioral attribute values in the spatial neighbors of o.

In practice, a variety of different methods could be used in order to characterize the local deviations around the spatial object o. The work in [433] has also defined a deviation measure *SLOM* which is based on the LOF methods for defining local spatial outliers. This approach is sensitive to the spatial heteroscedasticity of the data, in which the behavior of the spatial locality is carefully accounted for in constructing the outlier score.

2.2 Graph-based Methods

In graph-based methods, spatial proximity is modeled with the use of links between nodes. Thus, nodes are associated with behavioral attributes, and strong variations in the behavioral attribute across neighboring nodes are recognized as outliers. Graph-based methods are particularly useful when the individual nodes are not associated with point-specific coordinates, but may correspond to regions of arbitrary shape. In such cases, the links between nodes can be modeled on the basis of the neighborhood relationships between the different regions. Graph-based methods define spatial relationships in a more general way, since semantic relationships can also be used to define neighborhoods. For example, two objects could be connected by an edge, if they are in the same *semantic* location such as a building, restaurant, or office. In many applications, the links may be weighted on the basis of the strength of the proximity relationship. For example, consider a disease outbreak application in which the spatial objects correspond to county regions. In such a case, the strength of the links could correspond to the length of the boundary between two regions.

Let S be the set of neighbors of a given node i. Then, the concept of spatial continuity can be used in order to create a *predicted* value of the behavioral attribute based on those of its neighbors. The strength of the links between i and its neighbors can also be used in order to compute the predicted values as either the weighted mean or median on

the behavioral attribute of the k nearest spatial neighbors. For a given spatial object o, with behavioral attribute value $f(o)$, let $o_1 \ldots o_k$ be its k linked neighbors based on the relationship graph. Let the weight of the link (o, o_i) be $w(o, o_i)$. Then, the linkage-based weighted mean may be used to compute the predicted value $g(o)$ of the object o.

$$g(o) = \frac{\sum_{i=1}^{k} w(o, o_i) \cdot f(o_i)}{\sum_{i=1}^{k} w(o, o_i)}$$

Alternatively, the weighted median of the neighbor values may be used for predictive purposes. Since the true value of the behavioral attribute is known, this can be used in order to model the deviations of the behavioral attributes from their predicted values. As in the previous case, the value of $f(o) - g(o)$ represents a deviation from the predicted values. Extreme value analysis can be used on these deviations in order to determine the spatial outliers. This process is identical to what was discussed before for the multidimensional case. As in all outlier analysis algorithms, a variety of extreme-value analysis methods of Chapter 2 can be used on these deviations in order to determine the outliers. The nodes with high values of the normalized deviation may be reported as outliers.

2.3 Handling Multiple Behavioral Attributes

In many cases, multiple behavioral attributes may be associated with the contextual attributes. For example, in a meteorological application, both temperature and pressure values may be available with the spatial attributes. In these cases, the deviations may be computed on each behavioral-attribute, and then these values need to be combined into a single deviation value, which provides the final outlier score. For this purpose, any of the multivariate extreme value analysis methods in section 3 of Chapter 2 may be used. In particular, the work in [112] has proposed the use of the Mahalanobis distance-based method of Chapter 2 for extreme value analysis. However, it is also possible to use other depth-based, or angle-based methods discussed in that chapter in order to determine the underlying outliers.

3. Autoregressive Models

Spatial data shares a number of similarities with temporal data. Both kinds of data measure a behavioral attribute (eg. temperature) with respect to a contextual attribute (eg. space or time). In many scenarios, spatial data is available in the form of coordinates, and the values of the behavioral attribute may be available at *each possible spatial reference*

point in the grid. Such data arises commonly in weather contour maps, images, MRI scans etc. In cases, *where the data is completely specified at most points in the grid*, it is possible to use auto-regressive models in order to determine unusually large deviations in the data, in a way which is completely analogous to the temporal scenario.

Let X_{t_1,t_2} be the value of the behavioral attribute at the spatial location (t_1, t_2). In the temporal auto-regressive model, the predicted value of the behavioral attribute is based on a 1-dimensional window of *past* history of length p (see section 2.1 of Chapter 8). In the 2-dimensional spatial scenario, this can be generalized to a square window of size $(2 \cdot p + 1) \times (2 \cdot p + 1)$, with p coordinates in either direction. More generally, in the case of 3-dimensional spatial data, one can use a cube of size $(2 \cdot p + 1) \times (2 \cdot p + 1) \times (2 \cdot p + 1)$. As in the case of the 1-dimensional auto-regression for temporal data in section 2.1 of Chapter 8, a 2-dimensional model can be defined as follows.

$$X_{t_1,t_2} = \sum_{i=-p}^{p} \sum_{j=-p}^{p} a_{ij} \cdot X_{t_1-i,t_2-j} + c + \epsilon_{t_1,t_2}$$

The value of a_{00} is always set to 0, and is missing from the above summation, since a spatial value cannot be used to predict itself. The values of a_{ij} need to be learned from the underlying training data. Thus, such an equation can be created for each value of (t_1, t_2). When the number of spatial-coordinates available is much larger than $(2 \cdot p+1) \times (2 \cdot p+1)$, this is an over-determined system of equations, and can be solved in a similar way with *least-squares regression*, as discussed in the methods of section 2.1 of Chapter 3. Thus, the process of determining the regression coefficients is very similar to the case of temporal data.

In the above system of equations, the value of c is a constant, and the value of ϵ_{t_1,t_2} represents the noise, or the *deviation* from the expected values. Large absolute values of this deviation represent the anomalies in the underlying data. Therefore, the extreme value analysis techniques of Chapter 2 can be used in order to determine those deviations which vary significantly from the norm. These values are assumed to be independent identically distributed random variables, which are drawn from a normal distribution. Thus, the extreme value analysis methods of Chapter 2 can be used in order to detect the anomalies.

The afore-mentioned discussion provides a generalization of Autoregressive (AR) models from temporal to spatial data for illustrative purposes. In practice, it is possible to generalize *all* the regression models (ARMA, ARIMA, PCA) to the spatial scenario, by using the appropriate slice of values from the spatial data. As in the temporal case, it is even possible to create multivariate spatial regression models, where

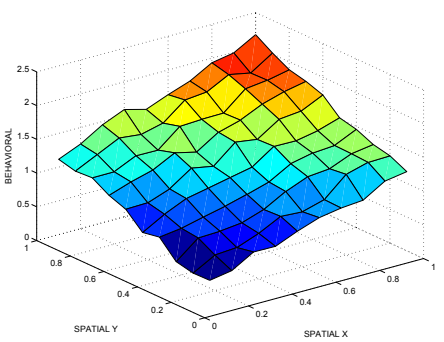
multiple behavioral attributes are available. Typically such behavioral attributes may be correlated with one another (eg. temperature and humidity), and it is desirable to determine unusually large local deviations with the help of multivariate correlations. Some of these generalizations are presented as exercises for the reader at the end of this chapter.

While the autoregressive nature of spatial data is very widely recognized, such models have rarely been used for anomaly detection in the spatial literature. This is partially a result of the high computational complexity of auto-regressive models with an increasing number of coefficients. Such models also cannot easily handle spatial data which is incompletely specified by spatial location, region-based locations or semantic locations. Nevertheless, such models can be very useful in many scenarios such as image analysis or weather patterns, where large amounts of reasonably complete data are available for analysis. In such cases, the statistical robustness of these methods is likely to be higher than simpler neighborhood-based models.

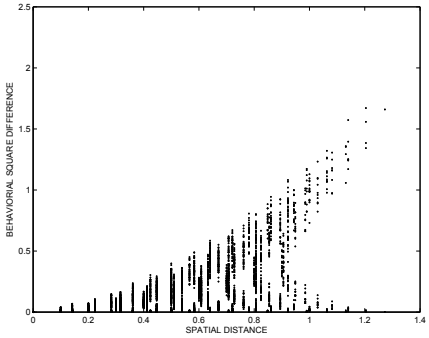
4. Visualization with Variogram Clouds

A number of visualization techniques such as pocket plots and variogram clouds are used in order to visualize spatial outliers. The former will be discussed here in detail, because of their relative popularity. Since spatial outliers are based on *disagreement* in the continuity of the behavioral attribute *in relation* to the spatial attribute, a natural method to visualize this would be to create a scatter plot between the pairwise spatial distances and the pairwise behavioral attribute (square) deviation. The spatial distance is simply the euclidian distance between a pair of points. The behavioral attribute deviation is defined as the half the square distance between the behavioral attribute values. A scatter plot is created between the spatial distances on the X-axis, and the behavioral square deviations on the Y-axis, for every pair of points in the data set. The idea is that smaller spatial distances will likely correspond to smaller behavioral attribute variances and vice-versa. In particular, large variations of the behavioral attribute for smaller spatial distances should be considered deviants. Such points on the variogram cloud can be traced back to the original data to determine pairs of points which are spatially close, but behaviorally different.

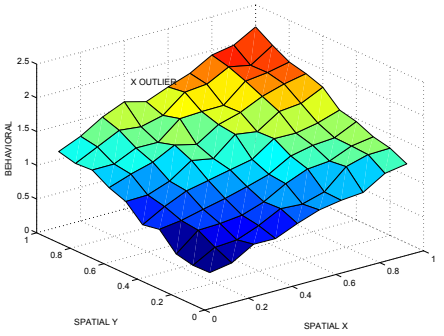
In order to illustrate the impact of outliers on variogram clouds, an example will be used. First, the data set for the variogram clouds of Figure 10.2 will be described. In this case, a grid of 100 points on the spatial plane are used with coordinates drawn from $X, Y, = 0.1, 0.2 \ldots 1.0$. The



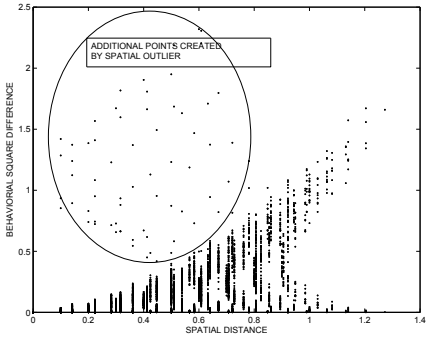
(a) Smooth spatial variation with no outlier



(b) Variogram Cloud with no outlier



(c) Smooth spatial variation with added outlier



(d) Variogram Cloud with added outlier

Figure 10.2. Effect of adding spatial outlier to variogram cloud

value of the behavioral attribute Z was generated as follows:

$$Z = X + Y + \epsilon$$

Here ϵ is a small amount of noise, which was randomly generated from the uniform distribution in $[0, 0.2]$. This spatial variation of the attribute is quite smooth, since the noise is small relative to the global variation in values of the behavioral attribute. The spatial profile of the generated data is illustrated in Figure 10.2(a), and the corresponding variogram cloud is illustrated in Figure 10.2(b). It is evident that low values of the spatial distance always corresponding to low deviations of the behavioral attribute. While it is possible for high spatial deviations to be related to low behavioral deviations, the converse is not true.

Subsequently, a single outlier is added to the data by distorting the behavioral attribute of one of the spatial values in the grid of Figure 10.2(a). The corresponding outlier is shown in Figure 10.2(c), and is marked explicitly. Note that the spatial data sets in Figures 10.2(a) and 10.2(c) are virtually identical, with the only difference between them being the outlier created by a distorted behavioral attribute value. The corresponding variogram cloud is illustrated in Figure 10.2(d). It is evident that in this case, a new set of points have been added to the variogram cloud in which significant behavioral deviations exist even at low spatial distances. Multiple such deviant points are created corresponding to the different data points in the immediate spatial locality of the added outlier. Such points can easily be isolated visually and linked back to the original points in the data. Thus, this approach provides an easy visual and intuitive way to isolate the spatial outliers in the data set.

One challenge of creating a variogram cloud is the high computational complexity. Note that a single point exists in the variogram cloud for each pair of data points in the original data. Therefore, the number of points in the variogram cloud scales quadratically with the number of points in the spatial data. This can make the approach rather slow, when the number of data points is large. In practice, it is difficult to create a variogram plot for situations in which the data contains a few hundred thousand spatial data points. This can be a significant problem, since spatial data sets are often quite large in practice.

One observation about the variogram cloud is that it is not always necessary to represent *every pair* of points on the plot. Data points which are spatially very far away add little insights about the outlier behavior. Therefore, each spatial dimension can be discretized into ranges, and this creates a 2-dimensional grid in the data. The pairwise relationships between all spatial points *within* this grid can be used in order to create

the variogram cloud. This significantly reduces the computational complexity of creating the variogram cloud. For example, consider the case, where the original data set contains N points, which are discretized into a $t \times t$ grid with approximately[1] N/t^2 data points in each. Then, the computational complexity of creating a variogram cloud for each grid is $O(N^2/t^4)$. Of course, since there are a total of t^2 grids, the aggregate computational complexity is $O(N^2/t^2) < O(N^2)$. This provides a speedup factor of $O(t^2)$. Of course, in this case an optimistic scenario was assumed where the data points were uniformly distributed into the grid structure. It can be shown theoretically that a speedup factor of at least t can be obtained with this approach. This is because the speed up achieved with a grid partitioning into $t \times t$ ranges will always be better than the discretization along only one dimension into t ranges with an equal number of data points. A significant speedup may be obtained even for modest values of t, without significant reduction in the quality of the visual discrimination between the outliers and the normal points.

5. Finding Abnormal Shapes in Spatial Data

The problem of finding unusual shapes in spatial data finds numerous applications such as image analysis. For example, the detection of unusual shapes from brain PET scans or MRI scans can help detect conditions such as tumors, alzheimer and sclerosis [374, 466], or can help identify anomalous conditions such as hurricanes from weather maps. For example, consider the satellite image illustrated in Figure 10.3. The anomalous shape in the image corresponds to hurricane *Fran*, which was a large destructive hurricane, which hit Cape Fear in North Carolina on September 1996. The hurricane can easily be identified by its characteristic shape in the satellite image. However, such a shape may not appear in other similar satellite images on normal days, and is therefore an unusual event. Another example from the medical domain is illustrated in Figure 10.4, where the PET scans from a normal person and an alzheimer patient are presented. The colored regions correspond to the uptake of the radioactive tracer administered in a PET scan (behavioral attribute). It is evident that this behavioral attribute shows very different spatial behavior for normal and diseased individuals.

In their simplest form, shapes can be modeled by the contours (or boundaries) of regions with particular ranges of behavioral attribute val-

[1] In practice, the different grid regions may contain a different number of data points because of spatial correlations. However, in many applications such as image data, pixels may be available for every spatial coordinate. Therefore, the division into grids will create a uniform division of the data points.

Figure 10.3. NASA Satellite Image of Hurricane *Fran*: The anomalous shape is characteristic of a hurricane

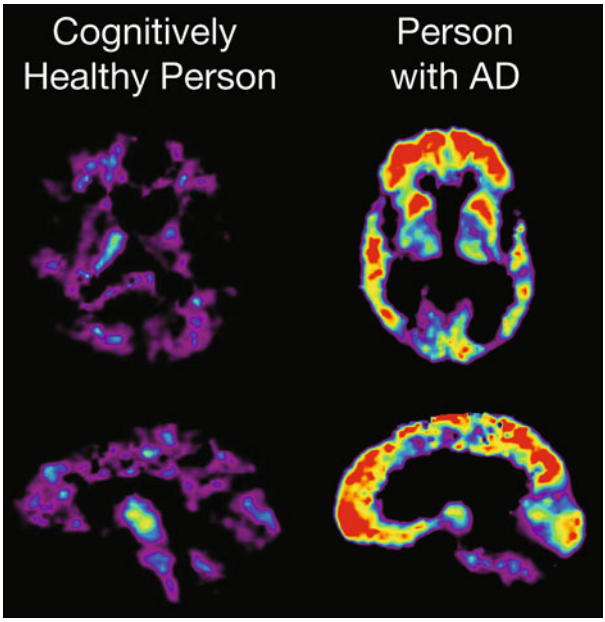
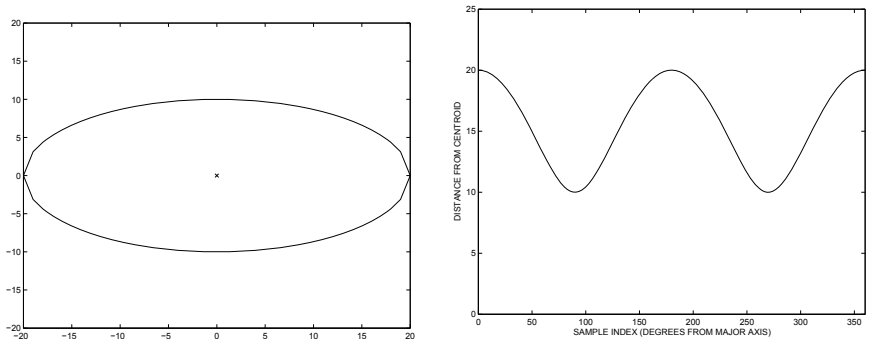


Figure 10.4. PET Scans of brain for cognitively healthy person versus an alzheimer patient: Image courtesy of the National Institute on Aging/National Institutes of Health

ues in the data. For example, in the case of Figure 10.3, the boundaries of such regions can be extracted by direct analysis of sensor and satellite readings such as pressure, cloud cover, temperature, wind speed, and humidity or from the (already processed) color histogram of the corresponding image.

A key simplification for shape analysis is that the contours of an object can be represented as a synthetic time-series. One possible way to achieve this is to use the distance from the centroid of the object to the boundary of the object, and compute a sequence of real numbers derived in a clockwise sweep of the boundary [504]. This yields a time series of real numbers, and is referred to as the *centroid distance signature*. This transformation can be used to map the problem of mining shapes to that of mining time-series, a domain which is much more easier to address from an analytical perspective. For example, consider the elliptical shape illustrated in Figure 10.5(a) with centroid denoted by X. Then, the time-series representing the distance from the centroid, by using 360 different equally spaced angular samples, is illustrated in Figure 10.5(b). In this case, the sample points are started at one of the major axes of the ellipse. If the sample point starts at a different place, or if the shape is rotated (with the same angular starting point), then this causes a cyclic translation of the time-series. The resulting time-series may be normalized in different ways depending upon the needs of the application:

- If no normalization is performed, then the outlier analysis approach is sensitive to the absolute sizes of the underlying objects. This may be the case in many medical images such as MRI scans, in which all spatial objects are drawn to the same scale.
- If all time series values are multiplicatively scaled down by the same factor to unit mean, then such an approach will allow the matching of shapes of different sizes, but will discriminate between different levels of relative variations in the shapes. For example, two ellipses with very different ratios of the major and minor axes will be discriminated well.
- If all time series are translated to zero mean and multiplicatively scaled to unit variance (as is normally done for time-series analysis), then such an approach will match shapes where *relative* local variations in the shape are similar, but the overall shape may be quite different. For example, such an approach will not discriminate very well between two ellipses with very different ratios of the major and minor axes, but will discriminate between two such shapes with different relative local deviations in the boundaries.



(a) Elliptical Shape (b) Distance from Centroid

Figure 10.5. Conversion from shapes to time-series

The only exception is a circular shape, which appears as a straight line. Furthermore, noise effects in the contour will be differentially enhanced in shapes which are less elongated. For example, for two ellipses with similar noisy deviations at the boundaries, but different levels of elongation (major to minor axis ratio), the overall shape of the time-series will be similar, but the local noisy deviations in the extracted time series will be *differentially* suppressed in the elongated shape. This can sometimes provide a distorted picture from the perspective of shape analysis. A perfectly circular shape may show unstable and large noisy deviations in the extracted time-series because of image rasterization effects. The solution proposed in [469] is to treat circular shapes specially, though the unintended effects of such normalization may have unusually complex effects across a broader spectrum of shapes.

In general, it may be advisable to pick the normalization method in an application-specific way.

The problem of shape analysis is further complicated by the effect that transformations such as rotations can have on the underlying data. For example, consider the images illustrated in Figure 10.6. All images correspond to the same object, but two of them are rotated with respect to the original shape, and the last is a mirror image of the original shape. It is clear that the rotation makes it much more difficult to match the two images, if the time-series representation does not account for the rotation or the mirror image effects of the representation. Errors in matching the two shapes also lead to errors in outlier detection, especially when the outlier detection process uses a proximity-based method. It is important to note that all applications do not necessarily require the accounting of rotations. For example, in an MRI scan, where the correct orientation of

the scan is known, such rotational transformations may not be needed. However, in the following, the most general case, which accounts for rotations will be discussed.

An immediate observation is that *a rotation of the shape leads to a linear cyclic shifting of the time series generated by using the distances of the centroid of the shape to the contours of the shape.* For a time series of length n denoted by $a_1 a_2 \ldots a_n$, a cyclic translation by i units leads to the time series $a_{i+1} a_{i+2} \ldots a_n a_1 a_2 \ldots a_i$. Then, the *rotation invariant euclidian distance* $RIDist(T_1, T_2)$ between two time series $T_1 = a_1 \ldots a_n$ and $T_2 = b_1 \ldots b_n$ is given by the minimum distance between T_1 and all possible rotational translations of T_2 (or vice-versa). Therefore, the following is true:

$$RIDist(T_1, T_2) = \min_{i=1}^{n} \sum_{j=1}^{n} (a_j - b_{1+(j+i) \bmod n})^2$$

Note that the reversal of a time-series corresponds to the mirror-image of the underlying shape. Therefore, mirror images can also be addressed by using this approach.

The shape discords can then be determined by computing the series whose kth nearest neighbor distance to its closest neighbor is as large as possible. The top n such shapes need to be found. As in all distance-based algorithms, a brute-force approach on a database with N shapes would require $O(N^2)$ distance computations, unless pruning methods are used.

The major difference between this problem and the unusual time-series shape discovery problem discussed in section 3 of Chapter 8 is that the rotational invariant distances are used instead of the euclidian distances. Furthermore, the distances are computed on whole time-series instead of on subsequences. While it may be possible in theory to use the method of Chapter 8, by making some modifications to address rotational invariance, longer lengths of whole sequences (compared to subsequences), may cause greater challenges in pruning. For example, rotational variations can be addressed by explicitly incorporating rotational variations of the time-series into the database, just as subsequences of a time-series are incorporated into the database for subsequence discord discovery in section 3 of Chapter 8. Care needs to be taken in avoiding self-similarity from the same shape during the distance computations, just as self-similarity is avoided in time series discord discovery. Therefore, the techniques in section 3 of Chapter 8 can be used in theory in order to find discords. Of course, the addition of multiple rotational variations of the shapes to the database is likely to slow down the discovery process. It also leads to some redundancy in the representation,

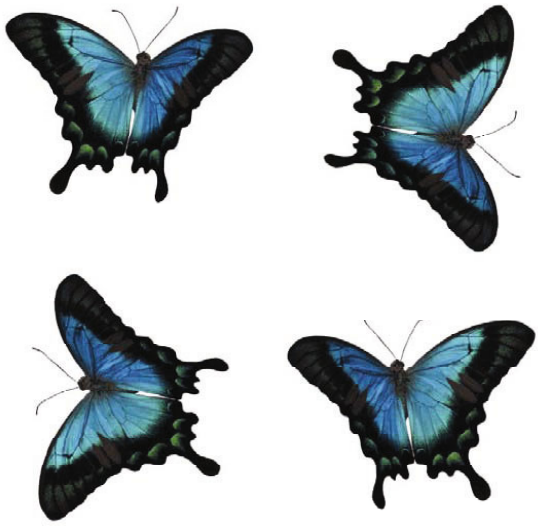


Figure 10.6. Rotation and mirror-image effects on shape matching for outlier analysis

because all rotational variations of the same object will have the same outlier score.

The work in [469] uses a different pruning method based on LSH-approximations [230] of the symbolic aggregate approximations of the time-series. The overall organization of the approach is similar to the algorithm discussed in section 3 of Chapter 8. Both methods first sort the objects by approximate outlier tendency in order to perform the outlier search in an ordered way, which optimizes the pruning behavior. For each object, pruning is performed with approximate nearest neighbor distances. However, the specific technique used for pruning is different in the two scenarios.

A nested loop approach is used to implement the method. The algorithm examines the candidate shapes iteratively in an outer loop, and progressively improves the estimate of each candidate's k-nearest neighbor distance in an inner loop. The inner loop essentially computes the distances of the other shapes to the candidate. At the end of the execution of a candidate-specific inner loop, the approach then either includes the candidate in the current set of top-n outlier score estimates, or discards the candidate at some point during the computation of its k-nearest neighbor in the inner loop. This is referred to as early inner loop termination. This inner loop can be terminated early, when the currently approximated k-nearest neighbor distance for that candidate shape is less than the score for the nth best outlier found so far. Clearly,

such a shape cannot be an outlier. In order to obtain the best pruning results, the candidate shapes in the outer loop need to be heuristically ordered, so that the earliest shapes examined have the greatest tendency to be outliers. The pruning performance is also best, when the points are ordered in the inner loop, such that the k-nearest neighbors of the candidate shape are found early. It remains to explain how the heuristic orderings required for good pruning are achieved.

As in the case of time-series subsequences, each time series is mapped onto an LSH word with the use of Symbolic Aggregate Approximation. Assume that the resulting SAX words have length m. Locality sensitive hashing [230] randomly samples $r < m$ distinct positions in the SAX representation. Therefore, two SAX words which are more similar are more likely to map to the same string. This is also referred to as the *locality sensitivity property* of the LSH-hashing approach, and the similarity can be robustly quantified by examining the mapping behavior over multiple hash functions. However, this does not account for the rotational invariance of the matching process. In order to account for the possible rotations, a rotational invariant LSH function is defined. This function first picks $r < m$ position indices randomly, and then samples these r position indices from all possible m rotations of the SAX word. Clearly, similar shapes will lead to LSH-based collisions, even in the presence of rotations. The LSH-hashing process is repeated with multiple hash functions in order to provide greater robustness to the collision-based counts.

For each SAX word, a count is maintained of its number of LSH-based collisions. This provides approximate information about its outlier score. Shapes with smaller counts need to be processed first as candidates in the outlier loop, since they have greater likelihood of being outliers. Furthermore, shapes which collide with one another frequently in LSH-based hashing are more likely to be nearest neighbors. Therefore, shapes which have the largest number of collisions with the current outer loop candidate are examined first in the inner loop for distance computations. This provides the heuristic order of processing in the inner loop. The reader is referred to [469] for a detailed description of the algorithm.

6. Spatio-temporal Outliers

Spatio-temporal data is very common in many real applications in which behavioral attribute values are continuously tracked at different spatial locations. For example, consider a chemical factory dumping chemicals in a river. In such cases, the concentrations of chemicals in the water cannot be described by using either only spatial or temporal

contextual attributes. Thus, the contextual attributes need to contain *both* spatial and temporal components. Spatiotemporal data is extremely common in all forms of sensor data, in which behavioral attribute readings are continuously transmitted by sensors at different spatial locations. An example is provided in [472], where precipitation data from different spatial locations and times is aggregated. It is desirable to determine localized spatial regions which are also close together in time, whose precipitation values are significantly different from their "neighboring" values. So how should neighboring values be defined in the case of spatiotemporal data?

Virtually all the spatial methods discussed in earlier sections of this chapter can be generalized to spatiotemporal data, as long as the concept of neighborhood is properly defined in order to make it relevant for the spatiotemporal scenario:

- Spatial methods can be used on temporal snapshots of the data in order to determine the relevant outliers at different instants. However, such an approach is incomplete, because it fails to identify violations of temporal continuity.

- Some algorithms have been proposed in order to separately identify spatial outliers and temporal outliers, and then combining the results in order to provide the spatiotemporal outliers [71]. However, the decoupling of spatial and temporal aspects of the problem at an earlier stage is obviously a sub-optimal solution.

- Spatio-temporal neighborhoods of data points may be used in order to determine predicted values. Thus, the only difference from purely spatial methods, is that the expanded set of contextual attributes are now used in order to define the neighborhoods for analysis and prediction. As in the previous case, deviations from the predicted values can be used in order to determine outliers. In some techniques such as neighborhood methods, the challenge is to combine the (contextual) distances along the spatial and temporal dimensions in a meaningful way. One simple way of achieving this would be to normalize the standard deviation across each of the contextual attributes to one unit before computation of distances. If desired, weights can be used in order to provide more importance to one or more of the contextual attributes.

The last of the above methods is the most general, because it can detect significant changes *both* across spatial and temporal attributes in an integrated and meaningful way. It is also important to note that spatial and temporal continuity may not be equally important, depending upon

the underlying application. For example, in an application where precipitation level is the behavioral attribute [472], spatial continuity may be slightly more important than temporal continuity. In such cases, appropriate scaling can be performed on the different dimensions, in order to define neighborhoods in a way which provides greater importance to one or more contextual attributes.

6.1 Spatiotemporal Data: Trajectories

A special case of spatiotemporal data is one in which no behavioral attributes are present, and the data comprises a set of moving object trajectories. Such data can be treated as a form of bivariate temporal data, by treating the X-coordinates and Y-coordinates of each object as the behavioral attributes, and time as the only contextual attribute. This results in two related time series at the same instants. Thus, the methods for temporal data analysis can be applied very effectively to such cases. Such analysis, when applied to single time-series, can identify sudden *changes* in trajectory directions and velocity. This can be very useful in detecting information about significant changes in cyclone or hurricane trajectories [94]. In other cases, a database of multiple trajectories may be available, and it is desirable to determine unusual shapes of trajectories. The temporal component is less important in this case, since the trajectories may have been created at different times. In such cases, it is possible to use subsequence analysis on these time-series in order to determine those trajectories which behave very differently from the remaining series by determining time-series of unusual shapes [304]. However, unlike the univariate scenario [304], spatial time-series are at least bivariate, and it is much harder to find unusual shapes in terms of the *combination behavior of the two time series*.

For the first case of real-time change analysis, the prediction-based outlier detection methods discussed in section 2 of Chapter 9 can be applied separately on each of the X-coordinate and Y-coordinate time series. This results in a residual value along each of the two coordinates. If each of these residuals is modeled as a normal distribution, then the sum of the squares of the Z-values of these residuals is a χ^2 distribution with two degrees of freedom. This can provide an outlier score, along with a corresponding probability value.

While real-time *change* analysis of such scenarios can be performed more effectively by using temporal modeling, *unusual shape* detection of trajectories can be best performed by abstracting out the temporal component, and performing the spatial analysis directly on the trajectories. In such cases, each spatial object has a shape, and the difference of this

shape to its nearest neighbor trajectories are used in order to determine outliers. Since such trajectories may contain a large number of time-stamps, it may often be difficult to determine outliers on the entire sets of trajectories. In such cases, unusual subsequences of the trajectories may be used in order to identify outliers. This case is similar to that of identifying unusual shapes in images, which is discussed in section 5 of this chapter.

Some specific methods such as TROAD have also been proposed in the literature [292] for unusual shape detection in trajectories. In particular, the partition-and-detect framework [292] first partitions the trajectories into a set of sub-trajectories. Note that this is somewhat analogous to the concept of partitioning time series into subsequences (or finding outliers in subspaces of numerical data), since outliers cannot easily be determined on the full series (with high implicit dimensionality). The sub-trajectories are created with a two-level partitioning which is allowed to be coarse-grained at the higher levels, and fine-grained at the lower level. Subsequently, those sub-trajectories, which are not similar to other ones in the data are reported as the outliers. The similarity is measured with the use of both distance-based and density-based methods. Note that the choice of the distance function is critical, and can regulate the nature of the outlier found. For example, a distance function which is sensitive to the *location* of the trajectory is likely to find an outlier based on location of the trajectories. On the other hand, a distance function which is sensitive to the angle between trajectory segments is likely to be sensitive to directions of movement. The precise definition of the distance function is application dependent, though a variety of such functions can be used in conjunction with the partitioned set of sub-trajectories.

The work in [292] defines a t-partition as a line segment from the trajectory. Intuitively, this can be considered analogous to comparison-unit schemes discussed in Chapter 9, which are used in the context of sequence data. A t-partition is said to be outlying using the variation[2] of the k-nearest neighbor distance definition, first proposed in [261]. Intuitively, a t-partition is considered an outlier, if a sufficient number of trajectories in the database are not close to it. The definition of closeness is based on measuring the portion of the trajectory, which is close to the t-partition. As in comparison-unit schemes for discrete sequences, the results from the different "units" (or partitions) are combined together to declare a trajectory as an outlier, if a sufficient number of its

[2]That variation fixes the nearest neighbor distance, and computes the required value of k rather than the other way around.

partitions are also outlying. Furthermore, the locality sensitive density-based approach of [78] has also been generalized to this case, by creating a density-sensitive outlier score for the trajectories.

6.2 Anomalous Shape Change Detection

In spatial data such as weather data, PET scans, and MRI scans, unusual changes in the contours of the shapes may be used in order to predict anomalous events. For example, the formation of a hurricane or a tumor over multiple time stamps will show up as an unusual change in the shapes of the corresponding image representations of the weather data or the MRI scan. The determination of such changes is more complex than those of detecting unusual *point changes* in the data. However, the detection of unusual point changes can be a first step towards detecting regions of anomalous change in the data, by clustering the change points in the spatial data. Not all regions of change may necessarily correspond to anomalies. For example, increasing age may create certain characteristic change contours in an PET scan, which should be considered normal. In practice, this problem is not very different from finding unusual shapes in the original data, with the main difference being that the contours of the shapes are constructed on the basis of the changes in the behavioral attributes between two snapshots. The normally occurring changes in the data over time will usually be quite different from the anomalous changes. Therefore, a differencing operation on two temporal snapshots of the data may be required as a pre-processing step, before applying outlier analysis algorithms. A detailed description of many such change analysis methods may be found in [92].

7. Supervised Outlier Detection

In many applications, a significant amount of training data may be available in order to determine anomalies. Such supervision could occur in either spatial data (with contextual attributes and behavioral attributes), or spatiotemporal data such as trajectory data. In all cases, supervision can be used in order to greatly enhance the effectiveness of the outlier analysis process.

7.1 Supervised Shape Discovery

Spatial data is particularly common in many forms of image data such as weather maps, PET scans or MRI scans. For example, consider the case of MRI scans, where 3-dimensional images of the brain may be available for analysis. The anomalies in the data such as tumors and lesions may show up as characteristic regions in the data, which are

rare but are nevertheless indicative of specific kinds of abnormalities. In such cases, previous examples of anomalous and normal scans may be available for the purposes of training. While unsupervised anomaly detection can help outlier analysis up to a point, the use of supervision can increase the sophistication of the analysis by revealing specific *kinds* of abnormalities. In most applications, at least semi-supervision is used, where examples of normal spatial profiles are available for analysis. The collection of normal examples is typically not very difficult in most application-specific scenarios, since copious examples of normal instances are usually available.

For all forms of shape classification, the actual *representation* of the shape is the most important step. For example, the *centroid distance signature* discussed in this chapter [504] is one possible way of representing the shapes, but by no means the only one. A thorough review of shape representation techniques may be found in [504]. The shape to time-series transformation discussed in section 5 of this chapter can be used in order to transform the shape classification problem to the time-series classification problem. Any of a number of methods (such as subsequence-based k-nearest neighbor methods) can be used for time-series classification in this case. Numerous methods for time-series classification may be found in the literature [343, 490]. These methods typically try to determine discriminative shapes of the series (or shapelets) which distinguish the normal and abnormal series. In the context of spatial data, such abnormal series are typically derived from abnormal shapes from a spatial perspective. In the *semi-supervised* case, the distances of the test series to examples of normal profiles can be used in order to create outlier scores for the underlying series. The only distinction from the available methods for time-series analysis is that care must be taken in order to account for different rotational variants of the shape in particular application-specific scenarios.

The problem of supervised classification of unusual shapes is also closely related to the problem of detecting and recognizing specific shapes in images. This problem has been studied extensively in the field of computer vision and image analysis. The problem of supervised shape recognition is an important area of research in its own right, and is beyond the scope of this book. The reader is referred to [54, 92, 316, 504] for a detailed description of such methods for image classification, analysis and change detection in the image domain. The major modification to these methods is the incorporation of rare class detection and cost-sensitive methods into these algorithms, using the methods of Chapter 6. Since many of the algorithms discussed in Chapter 6 are meta-algorithms, they

can be used in conjunction with any of the classification techniques in the literature.

7.2 Supervised Trajectory Discovery

In many cases, supervision may be available in the form of labels associated with trajectories. For example, consider a case where the trajectories of a large number of ships are available, and it is desirable to identify the suspicious ones based on their trajectory patterns. In some cases, previous examples of anomalous trajectories may be available. These can be used in order to detect significant anomalous patterns in the underlying data. This is a homogeneous attribute scenario, since the unusual shapes are based purely on the spatial and temporal attributes, rather than on a behavioral attribute.

The ROAM method [300] uses a discrete *symbolic* approximation of the trajectories, which converts the numerical coordinate sequence into a symbolic sequence based on the directions of movement and significant changes in this direction. For example, motifs could correspond to *right-turn*, *u-turn* or *loop*. Every movement pattern can be described as a sequence of these primitive movement patterns. The important motifs can be mined directly from the data by using a clustering approach. If desired, additional meta-attributes may be associated with the symbols corresponding to characteristics of the movement such as the speed. This is however different from the concept of behavioral attributes, since these attributes do not play the behavioral role in the learning process.

Once the discrete representation has been created, the sequences together with their labels can be fed to any sequence-based classifier, which identifies how different sequences are related to the class labels. While the ROAM method was applied in the context of supervised models, it is important to note that the feature transformation used in this work can also be used in the context of unsupervised scenarios.

8. Conclusions and Summary

The problem of spatial outlier detection arises in many domains such as demographic analysis, disease outbreaks, image analysis, and medical diagnostics. Spatial outlier detection shares significant resemblance with temporal outlier detection in terms of the effects of contextual attributes on the continuity of the behavioral attributes. Therefore, a number of methods in the temporal domain can be used for outlier detection in the spatial domain. Spatio-temporal outlier detection is even more complex and challenging, since it combines spatial and temporal characteristics effectively for outlier analysis.

Spatial data can often be treated as an abstraction of image data, when the spatial data is specified in a complete way. In such cases, numerous methods for image analysis can be used for outlier detection. In fact, in many applications such as MRI scans and weather maps, such data are indeed expressed as images. The analysis of such data involves the determination of unusual shapes from the distribution of the spatial attributes. Such analysis can be performed both in the unsupervised and supervised scenarios.

9. Bibliographic Survey

The problem of finding spatial outliers is different from that in multidimensional data because of the different kinds of attributes which are present in spatial data. The most common kinds of methods for finding spatial outliers use changes in the spatial proximity in order to determine outliers [3, 268, 317, 401–404]. Spatial proximity can be defined either with the use of multidimensional distances, or graph-based distances. Spatial distances are more relevant when the contextual attributes are expressed in terms of coordinates. On the other hand, when the reference attributes correspond to spatial regions or semantic locations, graph-based methods are more relevant, since distances and proximity can be expressed as general functions across links. A random walk approach to determine free form spatial scan windows is discussed in [234]. The application of outlier detection to heterogeneous neighborhoods is discussed in [235]. The work in [473] introduces a spatial likelihood ratio test in order to determine local grid regions in which the variation of the behavioral attribute is different from the remaining data in a statistically significant way. Furthermore, such methods can also be used in the context of multiple behavioral attributes [112]. Spatial data also shows local heterogeneity because of different levels of variance in different parts of the data. Therefore, a local method for spatial outlier detection was proposed in [433].

The standard auto-regressive models for temporal data [387] can be extended to spatial data, when the behavioral attribute values are completely specified over all the different reference values. This is often the case with many forms of image data. The problem of unusual shape detection in images is an important one from the perspective of outlier analysis. Some recent work [469] has been performed on finding unusual shapes in images in an efficient way. Supervised methods for shape detection and change analysis are also widely available in the literature [54, 92, 316, 504]. The work in [206] uses Multivariate Gaussian Markov Random Fields in order to find unusual shapes in medical image data.

Spatial data is closely related to temporal data in the context of the continuity shown by the behavioral attributes. Numerous methods for auto-regressive modeling [387] can also be generalized to the case of spatial data. A significant amount of data in spatial domains also has a temporal component, when the attributes are tracked at multiple timestamps. This requires methods for spatiotemporal outlier detection [113, 114]. An application of spatiotemporal outlier detection to precipitation data is discussed in [472]. A method for detecting flow anomalies in the context of sensors which located upstream or downstream from one another is discussed in [251]. When the differences in the values of the sensors exceeds a given threshold, it is flagged as a spatiotemporal anomaly. A method for explicitly quantifying the level of local change in a spatiotemporal data stream is proposed in [16]. This method also has the ability to perform online processing, and is discussed in detail in Chapter 8. Methods for detecting anomalies in vegetation data with the use of Principal Component Analysis (PCA) are discussed in [287].

The detection of outliers in trajectories can be modeled either spatially or temporally. Therefore, both spatial and temporal methods are relevant to this case. Significant *changes* in trajectory directions is useful for many applications such as hurricane tracking [94]. In such cases, the trajectory can be treated as bivariate temporal data, and change analysis can be applied to this representation. For this purpose, the prediction-based deviation detection techniques of the previous chapter can be helpful. The works in [83, 181] determine anomalies in moving object streams in real time, by examining patterns of evolution. On the other hand, the detection of anomalous trajectory *shapes* is a very different problem. The earliest methods for trajectory shape outlier detection were proposed in [263]. However, this method transforms the trajectories into point data by using a set of features describing meta-information about the trajectories. Unsupervised methods for trajectory outlier detection, which actually use the sequence information explicitly were first investigated in [344, 292]. The work in [344] uses the fourier transform in order to represent the trajectories in terms of the leading coefficients, and find anomalies. In the second method [292], trajectories are divided into different line segments and anomalous patterns are identified in order to determine outliers. Supervised methods for anomaly detection in trajectory data may be found in [300]. These methods transform the data into discrete sequences, and a classifier is learned in order to relate the trajectories to the class labels. Another method proposed in [302] proposes methods for finding outliers in vehicle traffic data. However, these methods are not designed for determining outliers on individual

objects, but are designed for finding anomalous traffic regions (or road segments) on the basis of aggregate spatial traffic characteristics.

10. Exercises

1. Construct the closed form solution to the AR regression model proposed in this chapter. Use the methods proposed in Chapter 3 for this purpose.

2. Construct PCA models for relating multiple behavioral attributes at the same spatial location. Use analogous models to those discussed in Chapter 8 for this purpose.

3. Construct PCA models for relating multiple behavioral attribute values over spatially local slices of size $p \times p$. Use analogous spatial models to the time-series models proposed in Chapter 8 for this purpose.

4. What is the time complexity of the methods proposed in Exercises 2 and 3.

5. Create a generalization of the time-series shape detection algorithm discussed in section 3 of Chapter 8 [258] to the spatial shape detection scenario. Refer to the details in [258] for specific details of pruning based on Symbolic Aggregate Approximation.

6. Implement the algorithm developed in Exercise 6 using a C++ implementation. Test it over benchmark data sets discussed in [469].

7. Implement the algorithm discussed in this chapter for unusual shape detection. Refer to [469] for specific details of LSH-based pruning. Test it over benchmark data sets discussed in [469]. How does the speed compare to the algorithm developed in Exercise 7.