Transactions in GIS, 2006, 10(2): 253–263

Research Article

A Multiscale Approach for Spatio-Temporal Outlier Detection

Tao Cheng
Department of Land Surveying and Geo-Informatics
The Hong Kong Polytechnic University

Zhilin Li
Department of Land Surveying and Geo-Informatics
The Hong Kong Polytechnic University

Abstract
A spatial outlier is a spatially referenced object whose thematic attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood. It represents an object that is significantly different from its neighbourhoods even though it may not be significantly different from the entire population. Here we extend this concept to the spatio-temporal domain and define a spatial-temporal outlier (ST-outlier) to be a spatial-temporal object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatial or/and temporal neighbourhoods. Identification of ST-outliers can lead to the discovery of unexpected, interesting, and implicit knowledge, such as local instability or deformation. Many methods have been recently proposed to detect spatial outliers, but how to detect the temporal outliers or spatial-temporal outliers has been seldom discussed. In this paper we propose a multiscale approach to detect ST-outliers by evaluating the change between consecutive spatial and temporal scales. A four-step procedure consisting of classification, aggregation, comparison and verification is put forward to address the semantic and dynamic properties of geographic phenomena for ST-outlier detection. The effectiveness of the approach is illustrated by a practical coastal geomorphic study.

1 Introduction

In the database domain, outliers refer to data that appear inconsistent with respect to the remainder of the database (Barnett and Lewis 1994). While in many cases outliers can be anomalies or noise, sometimes they represent rare or unusual events to be investigated

Address for correspondence: Tao Cheng, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: lstc@polyu.edu.hk

further. In general, there are three direct approaches for outlier detection: *distribution-based*, *depth-based* and *distance-based*.

Distribution-based approaches make use of standard statistical distributions and tests to detect the outliers. A key drawback of this approach is that most of the distributions used are univariate. Though some tests are multivariate, fitting the data to the standard distribution is costly, and may not produce satisfactory results (Breunig et al. 1999). The *depth*-based approaches map data objects into a m-dimensional information space (where m is the number of attributes) and assign a depth to each data value. The depth-based approaches become inefficient for large data sets with high dimensionality such as images and geographic data (Ng 2001). The *distance*-based approaches calculate the proportion of data that are of a specified distance from a target data value. The *distance*-based approaches have better computational efficiency than the depth-based approaches for large datasets. However, they require the existence of an appropriate (meaningful) distance function (Ng 2001), which is sometimes difficult to define.

A spatial outlier is a spatially referenced object whose thematic attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood. It represents the object that is significantly different from its neighbours even though it may not be significantly different from the entire population (Shekhar et al. 2003). Identification of spatial outliers can lead to the discovery of unexpected, interesting, and implicit knowledge, such as local instability and deformation.

Many methods have been recently proposed to detect spatial outliers by *distribution*-based approaches. These methods can be broadly classified into two categories, namely one-dimensional (linear) outlier detection and multi-dimensional outlier detection (Shekhar et al. 2003). The one-dimensional outlier detection methods consider the statistical distribution of thematic attribute values, ignoring the spatial relationships between the objects. The main idea is to fit the data set to a known standard distribution, and develop a test based on distribution properties (Johnson 1992, Barnett and Lewis 1994). Multi-dimensional outlier methods can be further grouped into two categories, namely homogeneous multi-dimensional metric based methods and spatial methods. The homogeneous multi-dimensional metric based methods do not distinguish between thematic attribute dimensions and spatial dimensions, and use all dimensions for defining the neighborhood as well as for comparison. In the spatial methods, spatial attributes are used to characterize location, neighborhood and distance, and thematic attribute dimensions are used to compare a spatially referenced object to its neighbors. For example, Shekhar et al. (2003) developed a unified modeling framework to detect spatial outliers based on a single thematic attribute.

The *depth*-based approaches are also applied extensively as clustering for spatial outlier detection, i.e. identifying the neighborhood of an object based on spatial relationships, and considering the proximity factor as the main basis for deciding if an object is an outlier with respect to neighboring objects or to a cluster. The limitation of these approaches is that the influence of some of the underlying spatial objects is ignored, which might be different at different spatial locations even though the close proximity is ignored, i.e. the semantic relationship is not considered in the clustering. An exception is the work of Adam et al. (2004) who identified spatial outliers by taking into account the spatial and semantic relationships among the objects.

Ng (2001) used the distance-based approach to detect unusual paths in two-dimensional space traced by individuals through a monitored environment. The distance-based measures allow the identification of unusual trajectories based on entry/exit points, speed and

geometry; these trajectories may correspond to unwanted behaviors such as theft. Other methods used in data mining such as classification and aggregation, are also applied in spatial outlier detection (Miller 2006).

In general, most existing methods only consider the thematic attributes of a data set, or only consider the spatial relations but ignore the semantic relations. As all geographic phenomena evolve over time, temporal aspects should also be considered (Yao 2003). Moreover, spatial and temporal relationships exist among spatial objects at various levels (scales). Such relationships should be considered and revealed in ST-outlier detection. However, detection of temporal outliers or spatio-temporal outliers (ST-outliers) has been seldom discussed. Our approach will build on existing approaches to propose a new methodology, which addresses the semantic and dynamic aspects of spatio-temporal data at multiple scales.

In the following sections, we first give a formal definition of ST-outliers, and then present a four-step multiscale approach to detect ST-outliers, which is followed by a case study. The last section summarizes the major findings and proposes directions for further research.

2 ST-Outlier Detection: Problem Definition

Here we extend the definition of spatial outlier given by Shekhar et al. (2003) to the spatio-temporal domain.

Usually time is included by dimensioning-up the spatial dimensions in order to accommodate spatio-temporal data. However, the nature of time requires the coding of the system and structuring of the data according to the relevant process aspect of time in order to make them useful (Cheng 2005). In most systems, time is generally considered to be uni-directional and linear. Conversely, space is commonly perceived as bi-directional and nonlinear (Roddick and Lees 2001).

Therefore, when we define ST-outliers, we should think about the process involved during the period that the spatio-temporal data are obtained. It implies that even if an object is a spatial outlier at a particular time; it may not be a ST-outlier in terms of the process involved. Figure 1a, b and c represents an area (with changing attributes represented by grids) at time T_1, T_2 and T_3, respectively. If we consider the attributes at each epoch separately, we may think that the points indicated by "*O*" are outliers. However, if we consider the attributes changing in time sequences, we may see the continuous change (or the trend) of the points indicated by "*O*". The appearance of "*O*" is continuous from T_1 to T_2 and T_3 with more points around "*O*" at T_1. Therefore, the points indicated by "*O*" are not outliers in a spatio-temporal framework. However, the point indicated by "*T*" at T_2 is a spatial outlier at T_2. It is also a ST-outlier since it is also obviously different from its temporal neighbourhood.

Hence we can define a spatio-temporal outlier (ST-outlier) to be a spatio-temporal object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatio-temporal neighbourhood. Let's assume that S_T is a spatio-temporal framework; f is an attribute function; *NB* is the neighbourhood relationship; F^N_{aggr} is a neighbourhood aggregation function; and F_{diff} is a difference function; *STT* is a statistical test procedure for determining statistical significance. A ST-outlier can be formally defined as follows:

(1) **Spatial neighbours –** Spatial neighbours refer to the objects that appear at the same time *T* and are spatially adjacent. The formal representation is as follows:

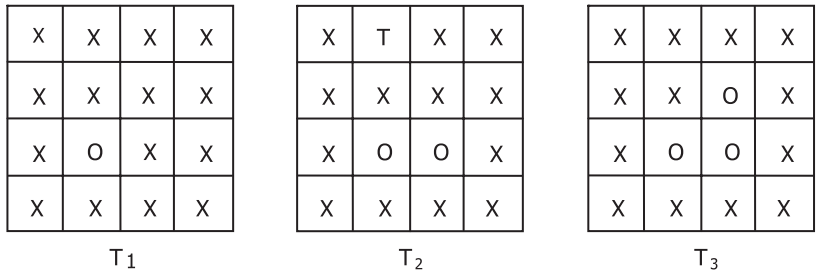

Figure 1 An area represented by grids at three epochs

$$If \quad Adjacent(S_{T,a}, S_{T,b}) = True \quad Then \quad NB(S_{T,a}, S_{T,b}) = True \tag{1}$$

(2) **Spatial outliers –** A spatial outlier is the object that is obviously different from its spatial neighbours, i.e.

$$If \ \ (NB(S_{T,a}, S_{T,b_1,b_2,\ldots,b_n}) = True \ \&\& \ STT\{F_{diff}[f(S_{T,a}) - F_{aggr}(S_{T,b_1,b_{2,\ldots,b_3}})]\} = True)$$
$$Then \quad S_{T,a} \text{ is a spatial outlier at time } T. \tag{2}$$

(3) **Spatio-temporal neighbours –** The spatio-temporal neighbours are the objects that are spatial neighbours and appear in consecutive periods, i.e.

$$If \quad Overlay(S_{T,a}, S_{T+1,b}) \neq \phi$$
$$Then \quad NB_T(S_{T,a}, S_{T+1,b}) = True \tag{3}$$

Here NB_T means that $S_{T,a}$ and $S_{T+1,b}$ are spatio-temporal neighbours at time T.

(4) **ST-outliers –** A ST-outlier is an object that is obviously different from its ST-neighbours, i.e.

$$If \ \ (NB_T(S_{T,a}, S_{T+1,b_1,b_2,\ldots,b_n}) = True \ \&\& \ STT\{F_{diff}[f(S_{T,a}) - F_{aggr}(S_{T+1,b_1,b_{2,\ldots,b_3}\cap a})]\} = True)$$

and

$$If \ \ (NB_T(S_{T,a}, S_{T-1,b_1,b_2,\ldots,b_m}) = True \ \&\& \ STT\{F_{diff}[f(S_{T,a}) - F_{aggr}(S_{T-1,b_1,b_{2,\ldots,b_m}\cap a})]\} = True)$$
$$Then \quad S_{T,a} \text{ is a ST-outlier.} \tag{4}$$

where $(b_1, b_2, \ldots, b_n)$ and $(b_1, b_2, \ldots, b_m)$ are the spatial-temporal neighbours of $S_{T,a}$, $(b_1, b_2, \ldots, b_n)$ *n a* and $(b_1, b_2, \ldots, b_m)$ *n a* represents the spatial overlay between them and $S_{T,a}$, respectively.

In order to detect ST-outliers, we need to identify the ST-objects and compare them with their spatio-temporal neighbours. If the differences between the attribute of an object located at x at time T and the statistical attribute values of x's neighbours at time $T-1$ and $T+1$ are large, the object is likely a ST-outlier. This means that an object cannot be a ST-outlier if it has the same or similar attribute values at different times, or its spatial neighbours change into the same or similar attribute values with the change of time. The next section will discuss the algorithm for the detection of ST-outliers.

3 Multiscale Approach for ST-Outlier Detection

In order to detect ST-outliers from a dataset, the existing methods for spatial outlier detection can be modified to address the semantic and dynamic aspects of spatio-temporal

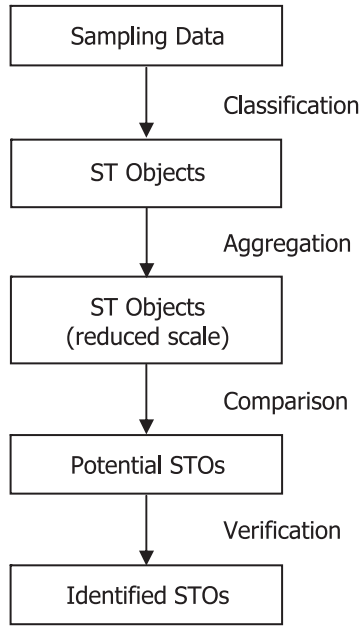

Figure 2 Four steps to detect the spatio-temporal outliers

data in multiscales. Since clustering is a basic method for outlier detection, we adopt it by including the semantic knowledge in the process of ST-object identification. Then, the multi-scale property of geographic phenomena is considered. If a spatial object (which is created from clustering) disappears after aggregation, it might be a suspected ST-outlier. Since spatial objects are dynamic, the verification of ST-outliers will consider the temporal continuity in addition to spatial continuity. Therefore, a four-step approach is proposed to identify the spatio-temporal outliers, i.e. classification, aggregation, comparison and verification (Figure 2). Here we call it a multiscale approach since the aggregation and verification compare the change between two consecutive scales in space and time.

3.1 Step 1: Classification (Clustering)

This involves the classification or clustering of the input data based upon the background knowledge of the data. The clustering method is designed based upon the prior knowledge and characteristics of the data. If the data are raster-based images, supervised classification might be applied or a classifier might be built based upon prior knowledge (i.e. a semantics-based approach). Other methods such as neural networks can also be applied if no prior knowledge about the data available. The purpose of this step is to form some regions that have significant semantic meanings, i.e. the spatio-temporal objects of interest.

3.2 Step 2: Aggregation (Filtering)

The results obtained in the previous step might contain outliers that are different from their spatial neighbours. Usually people use an aggregation function to check the difference between the objects and their neighbours. The aggregation function can be grouped

into three categories, namely, distributive, algebraic, and holistic (Gray et al. 1995, Shekhar et al. 2001). Examples of distributive aggregation functions include count, max, and sum. Average, variance, standard deviation, *maxN*, *minN* are all algebraic aggregation functions. The holistic aggregation functions include the median for example.

Spatial relationships between the objects have to be defined for the aggregation function. The computation complexity is high. Therefore, a new approach will be defined here. The main idea of this approach is to change the spatial scale of the data. If there are spatial outliers, they usually disappear if the scale of processing is reduced, similar to what happens with the multi-resolution clustering algorithm being used for outlier detection (Han et al. 2001). It can be realized by changing the resolution of measurement so that the clustered results are different with different scales. With a decrease in scale, the difference between the objects and their spatial neighbours will decrease and the small regions which contain outliers will be removed (Cheng et al. 2004).

Therefore, aggregation here refers to the changing (decreasing) of spatial resolution (spatial scale) of the data for clustering. It is also called filtering since the outliers (noises) will be filtered after the spatial scale change.

3.3 Step 3: Comparison (Detecting)

In this step, the results obtained at two spatial scales are compared in order to detect the potential spatial outliers. Therefore, the results derived from Step 1 will be compared with the results derived from Step 2. The objects that are found in step 1 and are missed (or filtered) in Step 2 are potential ST-outliers, just like the regions circled in Figure 6.

The comparison can be realized either by EVA (exploratory visualization analysis) or VDM (visual data mining). EVA methods tend to be highly interactive, providing interlinked and dynamic tools to explore the data, but without rigid control over the format that the scene might take. By contrast, VDM generally uses very specific algorithms so that uncovered objects or patterns will have a pre-defined visual appearance. In addition, EVA tends to be human-led and draws heavily from the perception literature, whereas VDM takes its cue from the numeric properties of the data and relies on statistical theory, pattern recognition and machine learning. The VDM methods are more structured and rigorous, but less flexible and perhaps less geographically intuitive (Gahegan 2001).

3.4 Step 4: Verification (Checking)

The outliers detected in the previous step can be suspected as ST-outliers. According to the definition of ST-outliers, they need to be verified since some of them might be noise and some are not, representing the real change of spatial objects. Since the natural change is usually smooth and continuous, what we see previously will be seen consecutively. Therefore, the verification checks the temporal neighbours of the suspected ST-outliers detected in the previous step. If the attribute value of such a ST-outlier is not significantly different from its temporal neighbours, this is not a ST-outlier. Otherwise, it is confirmed as a ST-outlier.

For the suspected outliers appearing at a particular time, we compare them with their temporal neighbours, i.e. the clustered results obtained before and after, to see if there is a continuous pattern. The attribute function of these suspected outliers will be compared with the aggregation functions of their temporal neighbours. If the difference is larger than a statistical criterion, it is a ST-outlier. Otherwise, it is a ST-object (see Equation 4).

4 Experimental Results

4.1 Data sets

Ameland, a barrier island in the north of the Netherlands, was chosen as a case study area. The process of coastal change involves the erosion and accumulation of sediments along the coast, which is scale-dependent in space and time. It can be monitored through the observation of annual changes of landscape units such as foreshore, beach and foredune (Cheng and Molenaar 1999).

The landscape units are defined based upon water lines. The foreshore is the area above the closure depth and beneath the low water line, the beach is the area above the low water line and beneath the dune foot, the foredune is the first row of the dunes inland from this dune foot. Based on height observations, it is possible to derive a measure of foreshore, beach and duneness. Height observations have been made by laser scanning of the beach and dune area and by echo sounding on the foreshore. These data have been interpolated to form a full height grid of the test area. In the following analysis, the error of the height grid, which was used as the original fine resolution DEM, is ignored.

The data set used in this study covers part of the island. The DEMs of six consecutive years (1989–95) are displayed in Figure 3. The images are of size 54×60 with the spatial resolution set at 60 m by 60 m. It is hard to identify the outliers in these images. The purpose of our experiment is to use the multiscale approach to detect the outliers in these annual DEMs.

4.2 Implementation Details

We applied the four steps discussed in the previous section to detect the ST-outliers in the DEM data.

First, we classified the DEMs into three landscape classes according to the approximate definition given by Dutch geomorphologists. For example, the areas with height between

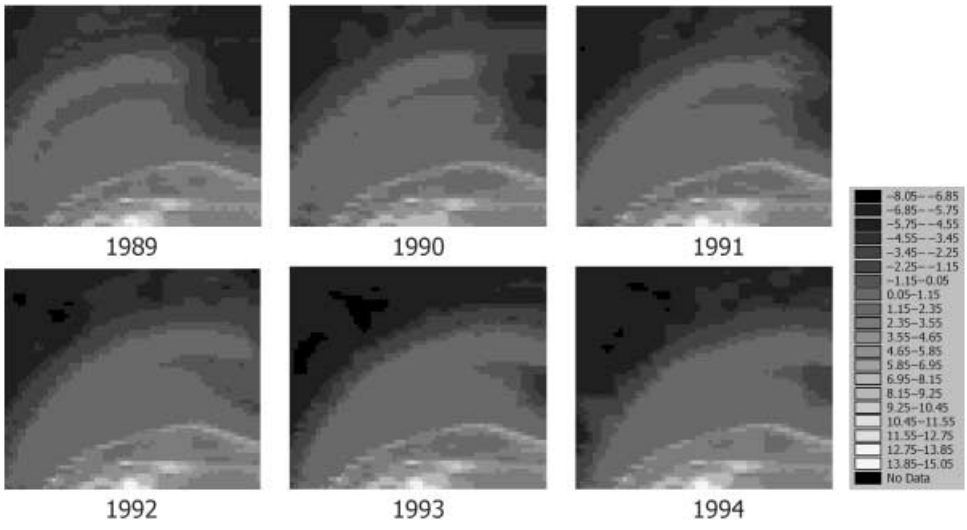

Figure 3 DEMs of Ameland in six consecutive years

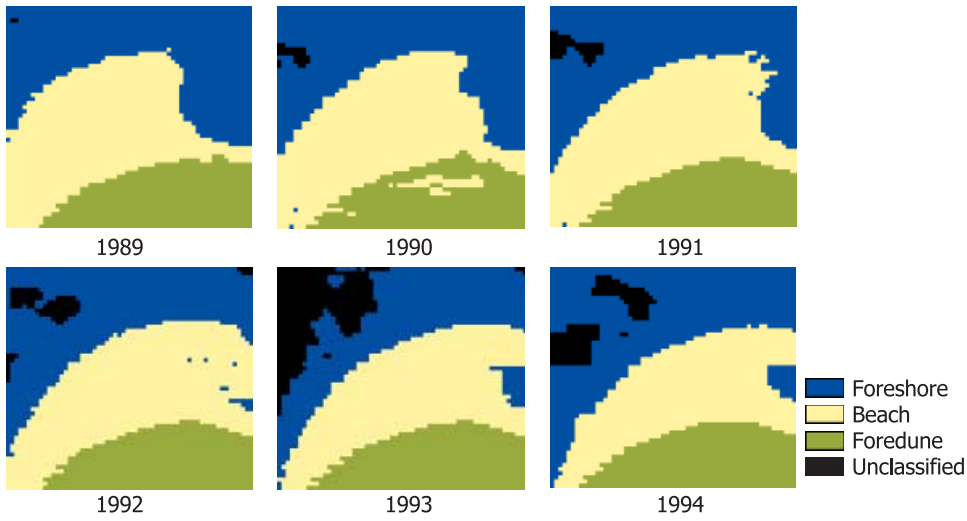

Figure 4 Step 1 – Clustering of the DEMs (1*1). This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

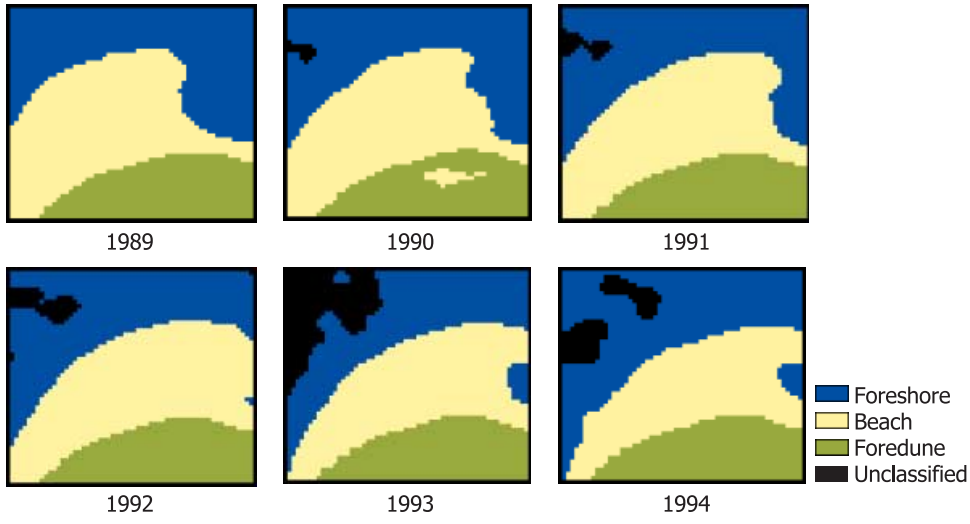

Figure 5 Step 2 – Clustering of aggregated DEMs (3*3). This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

–6 m ~ –1.1 m are the foreshore; the areas with height between –1.1 m ~ 2 m are the beach; and the areas with height between 2 m ~ 25 m are the foredune. The classification results are shown in Figure 4.

Next, we changed the spatial scale of the DEMs by averaging the height value using a 3 by 3 moving window. We classified the averaged (aggregated) results again into

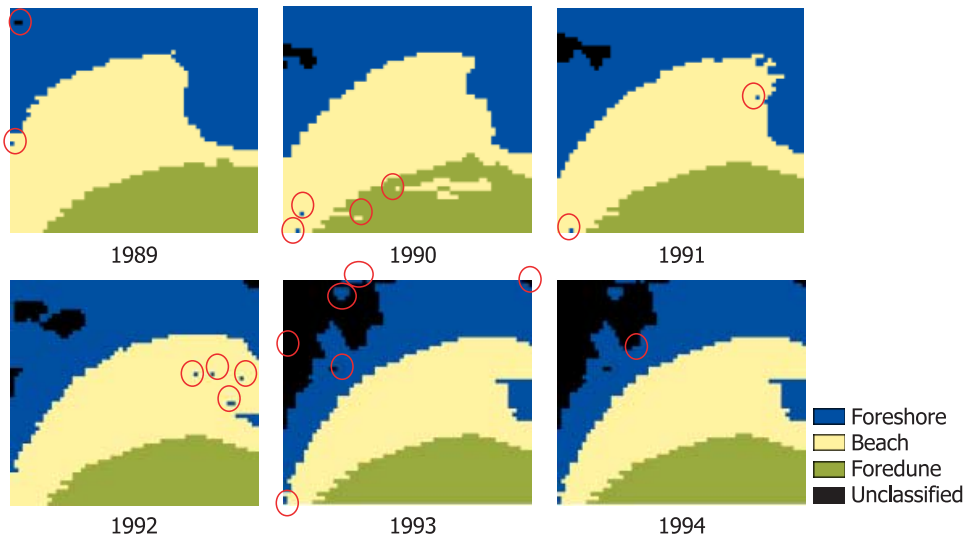

Figure 6 Step 3 – Suspected spatio-temporal outliers. This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

three landscape classes, according to the class definition in the previous step. The aggregated classifications are shown in Figure 5.

Later, we compared the images in Figures 4 and 5 of the same year and found regions that were identifiable in Figure 4 but disappeared in Figure 5. These regions are suspected ST-outliers (which are circled in Figure 6).

Last, we compared the height values of these suspected ST-outliers in consecutive years. If the change in height is continuous then the suspected ST-outlier is not a ST-outlier. For example, the ST-outlier that appeared in the upper-left corner in 1990 became part of a large dark area in later years. It means the change is continuous and this is not an outlier from a temporal perspective.

In this step, we first calculate the statistics of the circled areas in each year; then overlay the data with its consecutive year and calculate the statistics for the corresponding regions. If the difference is small, then they are not ST-outliers, i.e. those cases circled with dashed lines in Figure 7. If the differences are large, they are ST-outliers, i.e. those cases with continuous circled lines in Figure 7. We use the algebraic function for the statistical testing, i.e. average, variance, standard deviation of the height values for the grids belonging to the small regions are compared with their temporal neighbours.

4.3 Discussion

The threshold value of the statistical test is quite critical. Based upon the geomorphic situation of the area in our case, 0.5 m is set to be the threshold value for the difference. In future applications, the uncertainty of the verification can be evaluated since the statistical test can have different levels of confidence.

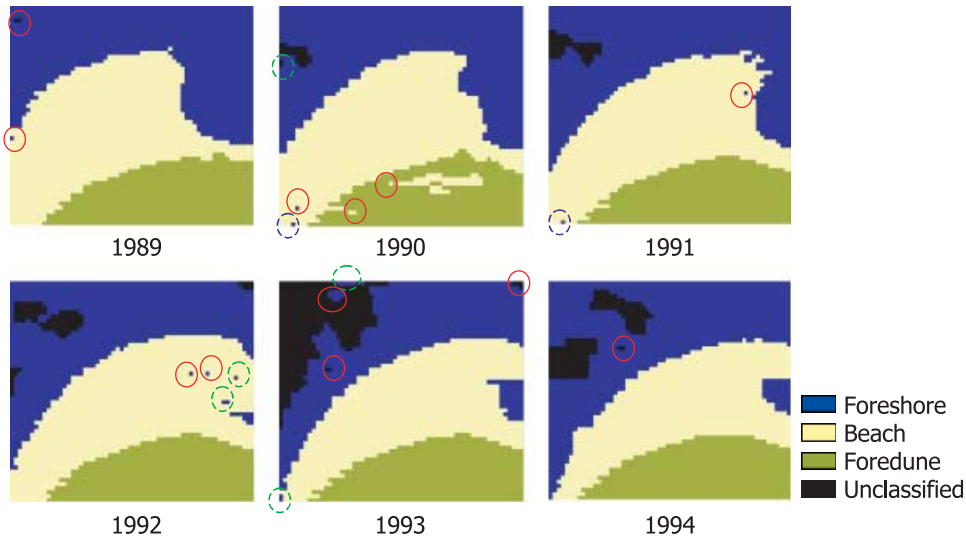

Figure 7 Step 4 – Verified ST-outliers (those circled with dashed lines are not ST-outliers). This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

5 Conclusions and Future Work

In this paper we discussed spatial-temporal outlier detection. We defined a spatial-temporal outlier (ST-outlier) as being a spatially-temporally referenced object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatio-temporal neighborhood.

We proposed a multiscale approach to detect the ST-outliers by evaluating the change between consecutive spatial and temporal scales. A four-step procedure consisting of classification, aggregation, comparison and verification is put forward to address not only the spatial relationships but also the semantic and dynamic properties of geographic phenomena for the ST-outlier detection. The effectiveness of the approach is illustrated by a practical coastal geomorphological study.

As for further research, the effect of the granularity of the spatial and temporal scales should be investigated. Further, quantitative calibration of the difference between two consecutive spatial and temporal scales should also be established, and the uncertainty of the ST-outlier detection should be evaluated.

Acknowledgements

The research is supported by the Hong Kong Polytechnic University (no. G-YW92) and the Major State Basic Research Development Program of China (973 Program, no. 2006CB701306). Thanks are due to the anonymous referees for their comments on an earlier version of this paper.

References

Adam N R, Janeja V P, and Atluri V 2004 Neighbourhood based detection of anomalies in high dimension spatio-temporal sensor datasets. In *Proceedings of the ACM Symposium on Applied Computing*, Nicosia, Cyprus: 576–83

Barnett V and Lewis T 1994 *Outliers in Statistical Data*. New York, John Wiley and Sons

Breunig M M, Kriegel H P, Ng R T and Sander J 1999 OPTICS-OF: Identifying local outliers. In Zytkow I M and Rauch J (eds) *Principles of Data Mining and Knowledge Discovery*. Berlin, Springer-Verlag Lecture Notes in Computer Science No 1704: 262–70

Cheng T and Molenaar M 1999 Diachronic analysis of fuzzy objects. *GeoInformatica* 3: 337–56

Cheng T, Fisher P, and Li Z 2004 Double vagueness: Effect of scale on the modeling of fuzzy spatial objects. In Fisher P F (ed) *Developments in Spatial Data Handling*, *Proceedings of the Eleventh International Symposium on Spatial Data Handling (SDH2004)*, Leicester, UK: 299–313

Cheng T 2005 Modelling and visualising linear and cyclic changes. In Fisher P and Unwin D J (eds) *Re-presenting GIS*. London, John Wiley and Sons: 205–13

Gahegen M 2001 Visual exploration in geography. In Miller H J and Han J (eds) *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis: 261–87

Gray J, Bosworth A, Layman A, and Pirahesh H 1995 Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In *Proceedings of the Twelfth IEEE International Conference on Data Engineering*, New Orleans, Louisiana: 152–9

Miller H J 2006 Geographic data mining and knowledge discovery. In Wilson J P and Fotheringham A S (eds) *Handbook of Geographic Information Science*. Oxford, Blackwell Publishing: forthcoming

Ng R 2001 Detecting outliers from large datasets. In Miller H J and Han J (eds) *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis: 218–35

Han J, Kamber M, and Tung A K H 2001 Spatial clustering methods in data mining. In Miller H J and Han J (eds) *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis: 188–217

Johnson R 1992 *Applied Multilevel Statistical Analysis*. Englewood Cliffs, NJ, Prentice Hall.

Roddick J F and Lees B G 2001 Paradigms for spatial and spatio-temporal data mining. In Miller H J and Han J (eds) *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis: 33–49

Shekhar S, Lu C T, and Zhang P 2003 A unified approach to detection of spatial outliers. *GeoInformatica* 7: 139–66

Shekhar S, Huang Y, Wu W L, and Lu C T 2001 What's special about Spatial Data Mining: Three Case Studies. In Grossman R L, Kamath C, Kegelmeyer P, Kumar V, and Namburu R R (eds) *Data Mining for Scientific and Engineering Applications*. Berlin, Kluwer: 405–19

Yao X 2003 Research issues in spatio-temporal data mining. In *Proceedings of the University Consortium for Geographic Information Science Workshop on Geospatial Visualization and Knowledge Discovery*, Lansdowne, Virginia