Geoinformatics 2004
Proc. 12th Int. Conf. on Geoinformatics − Geospatial Information Research: Bridging the Pacific and Atlantic
University of Gävle, Sweden, 7-9 June 2004

A HYBRID APPROACH TO DETECT SPATIAL-TEMPORAL OUTLIERS

Tao Cheng and Zhilin Li
Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom,
Kowloon, Hong Kong, {lstc; lszlli}@polyu.edu.hk

Abstract

A spatial outlier is a spatial referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood. It represents locations that are significantly different from their neighborhoods even though they may not be significantly different from the entire population. Here we adopt this definition to spatio-temporal domain and define a spatial-temporal outlier (STO) to be a spatial-temporal referenced object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatial or/and temporal neighborhood. Identification of STOs can lead to the discovery of unexpected, interesting, and implicit knowledge, such as local instability. 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 hybrid approach which integrates several data mining methods such as clustering, aggregation and comparisons to detect the STOs by evaluating the change between consecutive spatial and temporal scales.

INTRODUCTION

Outliers are data objects that appear inconsistent with respect to the remainder of the database (Barnett and Lewis, 1994). While in many cases these can be anomalies or noise, sometimes these represent rare or unusual events to be investigated further. In general, direct methods for outlier detection include *distribution-based*, *depth-based* and *distance-based* approaches. *Distribution*-based approaches use standard statistical distribution, *depth*-based technique map data objects into an m-dimensional information space (where m is the number of attribute) and *distance*-based approaches calculate the proportion of database objects that are a specified distance from a target object (Ng, 2001).

A spatial outlier is a spatial referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood. It represents locations that are significantly different from their neighborhoods even though they 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.

Many methods have been recently proposed to detect spatial outliers by the *distribution*-based approach. These methods can be broadly classified into two categories, namely 1-D (linear) outlier detection methods and multi-dimensional outlier detection methods (Shekhar et al, 2003). The 1-D outlier detection algorithms consider the statistical distribution of non-spatial attribute values, ignoring the spatial relationships between items.

The main idea is to fit the data set to a known standard distribution, and develop a test based on distribution properties (Barnett and Lewis, 1994; Johnson, 1992). 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 attribute dimensions and geo-spatial dimensions, and use all dimensions for defining neighborhood as well as for comparison. In the spatial methods, spatial attributes are used to characterize location, neighborhood, and distance, and non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. Among others, Shekhar et al. (2003) developed a unified modeling framework and identify efficient computational structure and strategies for detecting spatial outliers based on a single non-spatial attribute from a data set.

D*epth*-based techniques are also applied extensively as clustering for spatial outlier detection, i.e. identifying the neighborhood of an object based on spatial relationship, 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 ignoring the influence of some of the underlying spatial objects that might be different at different spatial locations despite the close proximity, i.e., the semantic relationship is not considerer in the clustering. An exception is that, Adam et al. (2004) identified spatial outliers by taking into account the spatial and semantic relationships among the objects.

Ng (2001) uses distance-based measures to detect unusual paths in two-dimensional space traced by individuals through a monitored environment. These 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, 2003).

In general, most existing methods only consider the non-spatial attributes of a data set, or only consider the spatial relations and ignore the semantic relations. Further, as all geographic phenomena evolve over time, temporal aspect should also considered (Yao, 2003). How to detect the temporal outliers or spatial-temporal outliers has been seldom discussed. Moreover, spatial and temporal relationships exist among spatial entities at various levels (scales, Yao, 2003). Such relationships should be considered and reveled in spatio-temporal detection. Our approach will build on existing approaches to evolve into a new methodology, which addresses the semantic aspects and dynamic aspects of spatio-temporal data in multi-scales.

ST OUTLIER DETECTION: PROBLEM DEFINITION AND PROPOSED ALGORITHMS

Here we adopt the definition of spatial outlier to spatio-temporal domain and define a spatial-temporal outlier (STO) to be a spatial-temporal referenced object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatial or/and temporal neighborhood.

In order to detect STOs from a data set, the existing methods for spatial outlier detection can be modified to addresses the semantic aspects and dynamic aspects of spatio-temporal data in multi-scales. Since clustering is a basic method for outlier detection, we start it by including the semantic knowledge in the process. Then, the multi-scale property of natural phenomena is considered. If a spatial object (which is created from clustering) disappears after aggregation, it might be a potential STO. Since spatial objects are dynamic, the verification of STOs will consider the temporal consecutive in addition to spatial continuity. Therefore a four-step approach is proposed to identify the spatio-temporal outliers (see Figure 1). Here we call it a hybrid approach since it adopts methods of clustering, aggregation and comparison.

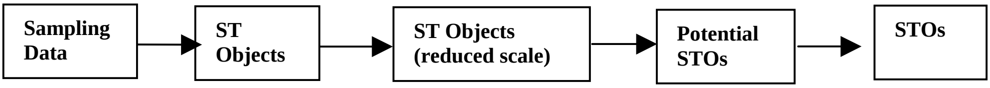

Figure 1: Four steps to detect the Spatio-temporal Outliers (STOs).

1) classification (clustering)

This involves the classification or segmentation or clustering the input data based upon the background knowledge of the data. There are several ways to do so which depends on the characteristics of data. If the data is raster-based images, supervised classification might be applied or a classifier might be built based upon prior-knowledge (semantics-based approach). The purpose of this step is to form some regions that have significant semantic meanings.

2) aggregation (filtering)

This step involves the multi-scale aggregation of clustered result to check the stability of the clustering. In this step, outliers might be merged out.

3) comparison (identify the merged objects)

This step compare the clustered results derived from step 2 with the results derived from step 3 and identify the objects that are missed in step 3, which are potential STOs. In these step, the comparison is implemented at two consecutive spatial scales.

4) verification (checking spatial and temporal neighbors)

This step checks the spatial and temporal neighbors of the potential STOs identified in the previous step. If the semantic value of such a STOs does not have significant differences with its spatial neighbor, it may not be a STO; then check its temporal neighbors. If the difference with the temporal neighbors is not large, this checking is not a STO. Otherwise, it is confirmed as a STO.

EXPERIMENTAL RESULTS

Data sets

Ameland, a barrier island in the north of the Netherlands, was chosen as a case study area. The process of coast 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 changes of landscape units such as foreshore, beach and foredune.

The landscape units are defined based upon water lines. The foreshore is the area above the closure depth and beneath the low water line, 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 dune foot. Based on height observation, 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 raster of the test area. Experiments show that the uncertainty of the interpolated heights of the raster can be expressed by standard deviation (σ = 0.15 m). However, in the following analysis, the error of the height raster, which was used as the original fine resolution DEM, is ignored.

The data set we used covers part of the island. The DEMs of six consecutive years is displayed in Figure 2. The purpose of our experiment is to detect the outliers in these six year DEMs.

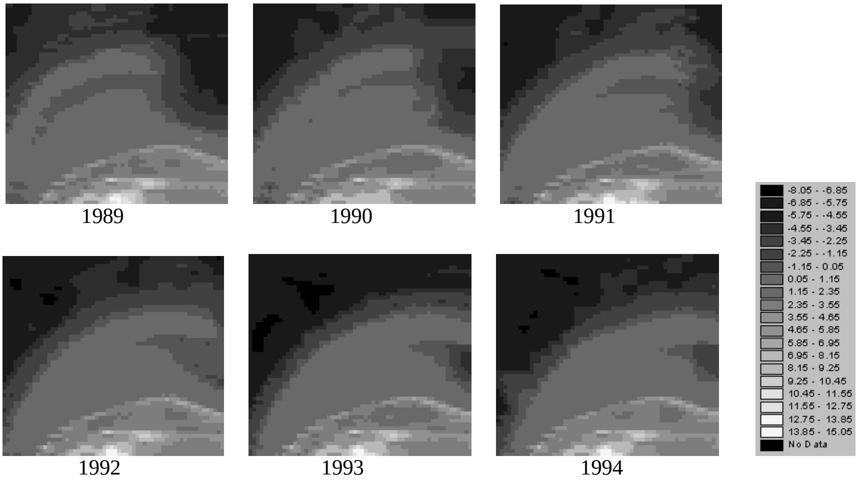

Figure 2: DEMs of Ameland in six consecutive years.

Implementation details

We applied the four steps discussed in the previous section. First, we classified the DEMs into three landscape classes. The classification results are shown in Figure 3.

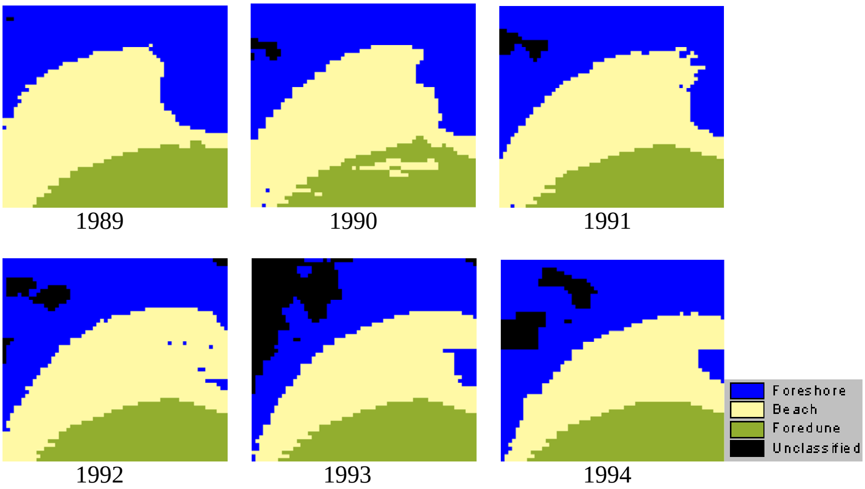

Figure 3: Step 1 – Clustering of the Data.

Then, we changed the spatial scale of the DEMs by averaging the height value by a 3*3 window. Then we classified them again into three landscape classes. The results are shown in Figure 4.

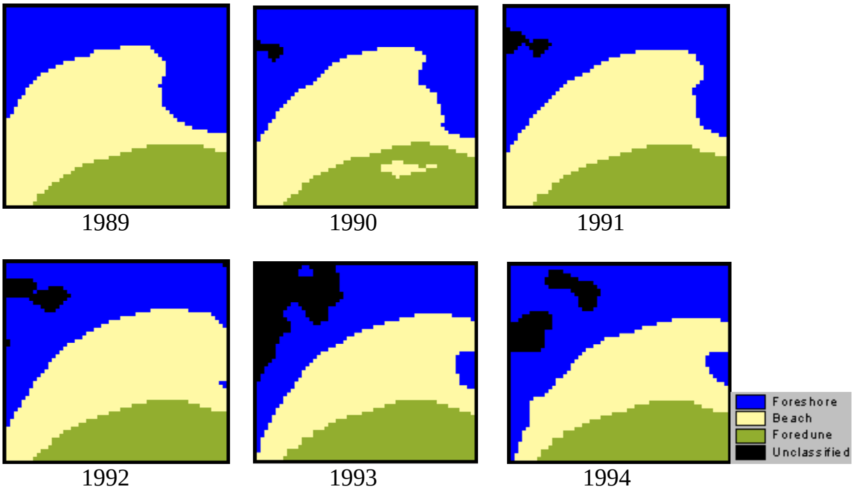

Figure 4: Step 2 – Multi-scale change and potential STOs.

Later, we compared Figure 3 and Figure 4 and found the potential STOs (which are circled in red in Figure 5). They are the regions (objects) that are different/disappeared in Figure 4.

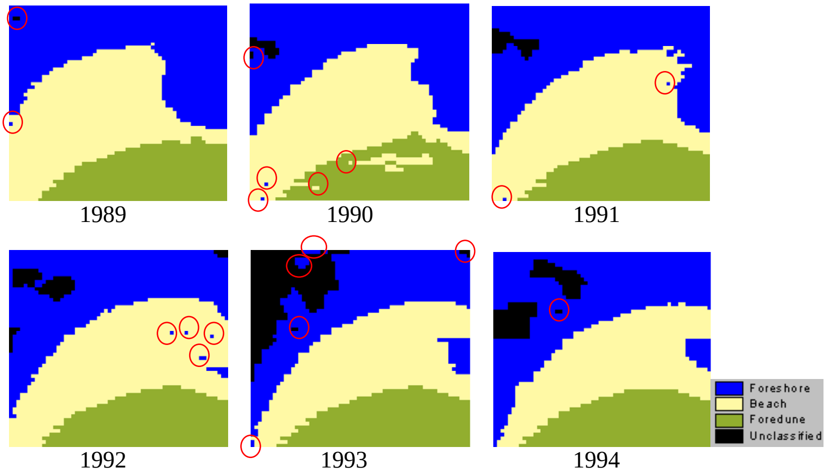

Figure 5: Step 3 – Comparison and potential STOs.

For further verification, we compared the height values of these potential STOs in the consecutive years. If the change of height is continuous then the potential STO is not a STO. For example, the STO appeared in 1991 (in upper-left corner) became part of a big dark area in 1991. It means the change is continuous and this is not an outlier in temporal perspective. Finally, we identified the STOs, which are circled in red in Figure 6. For those circled in dashed line in Figure 6, they are not STOs.

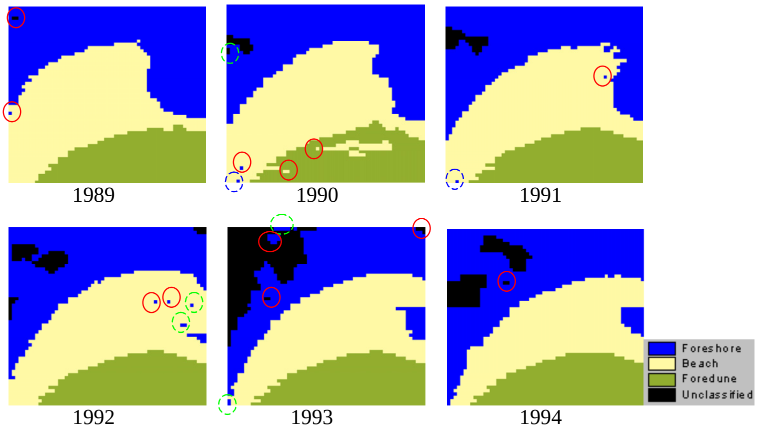

Figure 6: Step 4 – Verified STOs.

CONCLUSIONS AND FUTURE WORK

In this paper we discussed spatial and temporal based outlier detection. We defined a spatial-temporal outlier (STO) to be a spatial-temporal referenced object whose thematic attribute values are significantly different from those of other spatially and temporally referenced objects in its spatial or/and temporal neighborhood. We propose a hybrid approach that integrates several data mining methods such as clustering, aggregation and comparisons to detect the STOs by evaluating the change between consecutive spatial and temporal scales. As for further research, the effect of granulites of spatial and temporal scales should be investigated. Further, quantitative calibration of the difference between two consecutive spatial and temporal scales should also be established.

ACKNOWLEDGEMENTS

Support for the first author from The Hong Kong Polytechnic University through the Postdoctoral Research Fellowship (no. G-YW92) and the project G-T29E is gratefully acknowledged.

REFERENCES

Adam, N.R., Janeja, V.P. and Atluri, V., 2004: Neighbourhood based detection of anomalies in high dimension spatio-temporal sensor datasets. *2004 ACM symposium on Applied Computing*, March 14 – 17, Nicosia, Cyprus, pp. 576 – 583.

Barnett, V. and Lewis, T., 1994: Outliers in Statistical Data, John Wiley.

Johnson, R., 1992: *Applied Multilevel Statistical Analysis*, Prentice Hall.

Miller, H., 2003: Geographic data mining and knowledge discovery. In:. Wilson, J.P. and Fotheringham, A.S. (Eds) *Handbook of Geographic Information Science*, in press.

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, 218-235.

Shekhar, S., Lu, C.T. and Zhang, P., 2003: A unified approach to detection spatial outliers, *GeoInformatica* 7, 139-166.

Yao, X., 2003, Research issues in spatio-temporal data mining, a white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, Nov. 18-20.