A Spatiotemporal Data Mining Framework for Mining Ozone Pollution Data

Sujing Wang[1], Christoph F. Eick[2], Qiang Xu[3]

[1]Department of Computer Science, Lamar University, Beaumont, TX 77710, USA
Telephone: (010) 409 880-7798
Email: swang3@lamar.edu
[2]Department of Computer Science, University of Houston, Houston, TX 77204, USA
Telephone: (010) 713 734-3345
Email:ceick@cs.uh.edu
[3]Department of Chemical Engineering, Lamar University, Beaumont, TX 77710, USA
Telephone: (010) 409 880-7818
Email:qiang.xu@lamar.edu

1. Introduction

Due to advances in remote sensors and sensor networks, different types of dynamic and geographically distributed data become increasingly available. These data can also integrate multiple other types of data, such as temporal information, social information, textual data, multimedia data, and scientific measurements. Such enriched data provide a tremendous potential for discovering new useful knowledge, as well as research challenges.

Spatiotemporal data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from the spatiotemporal data. Conventional data mining techniques are inefficient in mining spatiotemporal data because they do not incorporate the idiosyncrasies of spatial and temporal domains. Therefore, new algorithm and technique are needed to provide effective solutions to mine spatiotemporal data. We introduce a spatiotemporal data mining framework, which includes a density-based spatiotemporal clustering algorithm and a post-processing analysis technique to extract interesting patterns and useful knowledge from spatiotemporal data. Our framework consists of the following four steps:

– Step 1: generate polygons from multi-source spatiotemporal point data;
– Step 2: apply spatiotemporal clustering algorithm to group polygons based on their spatial and temporal similarities to identify distribution patterns;
– Step 3: utilize post-processing analysis technique to identify interesting clusters;
– Step 4: visualize clusters.

We evaluate the effectiveness of our techniques through a challenging case study involving ozone pollution events in the Houston-Galveston-Brazoria (HGB) area.

2. Spatiotemporal Clustering Algorithm

The Shared Nearest Neighbour (SNN) (Ertoz 2003) clustering algorithm is a well-established density-based clustering algorithm. SNN defines the similarity between pairs of points in terms of how many nearest neighbours the two points share. We extend SSN by redefining the spatiotemporal similarity between polygons, the density-based concepts of core polygons and outliers as well.

Each polygon *p* in dataset *D* is associated with a time *t* when it occurs, and a set of other attributes. A cluster of polygons is a group of polygons that lie in close proximity in both space and time. Any function that can compute the spatial distances between a pair of polygons, such as Hausdoff distance (Hangouet 1995), can be used. Any temporal distance function that can compute the temporal distances between a pair of polygons can be adopted as well. The k-nearest spatial neighbour list and k-nearest temporal neighbour list for each polygon *p*, denoted by k-SN(*p*) and k-TN(*p*), are generated by keeping its k-nearest spatial neighbours and k-nearest temporal neighbours only. The nearest spatiotemporal neighbours of each polygon *p*, denoted by NN(*p*), is calculated as the union of the k-nearest spatial neighbour list and k-nearest temporal neighbour list of *p*:

$$NN(p) = \mathrm{k} - \mathrm{SN}(p) \cup \mathrm{k} - \mathrm{TN}(p) \tag{1}$$

The similarity between a pair of polygons *p* and *q*, denoted by similarity (*p*, *q*), is the number of the nearest spatiotemporal neighbours that they share:

$$similarity\ (p, q) = \left| NN(p) \cap NN(q) \right| \tag{2}$$

The SNN density of polygon *p* is defined as the number of polygons that share *Eps* or more nearest neighbours with *p*. *Eps* is a user specified density threshold:

$$density\ (p) = \left| \{ q \in D \,|\, similarity\ (p, q) \geq Eps \} \right| \tag{3}$$

The core polygons are identified by using a user specified parameter *MinP*; all polygons in dataset *D* that have SNN density no less than *MinP*:

$$CoreP\ (D) = \{ p \in D \mid density\ (p) \geq MinP \} \tag{4}$$

Clusters are then formed by computing the transitive closure of polygons that can be reached from an unprocessed core polygon using their respective nearest neighbours; this process continues until all core polygons have been assigned to a cluster. The remaining polygons that are not within a radius of *Eps* of any core polygon are classified as outliers, and are not included in any clusters.

3. Post-processing Analysis Technique

Our post-processing analysis technique allows automatic screening of all clusters for interesting ones whose polygons have attribute values that deviate significantly from the typical attribute values. For example, we try to find clusters of polygons with attribute values that are much smaller or larger than all others. Such anomalous clusters are exceptional in some sense, and are often of unusual importance. We assume a dataset *D* with *n* attributes, and a set of clusters in *D* identified by spatiotemporal clustering algorithm. Let (*a*, *b*) be the interquartile range (IQR) of a cluster, C_i, for attribute *j* with *a* > *b*, and (*a'*, *b'*) be the IQR of the dataset *D* for attribute *j* with *a'* > *b'*, we compute the degree of deviation for each attribute *j* in a cluster C_i compared with the dataset *D*:

$$R_{i,j} = \frac{\max\left(a'_j - a_{i,j}, 0\right) + \max\left(b'_j - a_{i,j}, 0\right) - \max\left(a'_j - b_{i,j}, 0\right) - \max\left(b'_j - b_{i,j}, 0\right) + a_{i,j} - b_{i,j}}{a_{i,j} - b_{i,j}} \tag{5}$$

$R_{i,j}$ could be any number between -1 and 1. If $R_{i,j}$ is 1, it indicates the box plot for attribute *j* in C_i is above the upper end of the box plot (75%) for attribute *j* in *D*. If $R_{i,j}$ is -1, it indicates that the box plot for attribute *j* in C_i is below the lower end of the box plot for attribute *j* in *D*. The interestingness score of a cluster, C_i, is calculated based on all $R_{i,j}$ associated with C_i. Let $o_i = \{r_1,\ldots r_n\}$ be the set of deviation degrees of *n* attributes in C_i. In general, the interestingness score of each cluster C_i is a function of O_i:

$$I(c_i) = f(O_i) \tag{6}$$

Different interestingness functions may be adopted for different analysis tasks in different application domains.

3. Case Study

We collected multiple air quality data from TCEQ’s (Texas Commission on Environmental Quality) website. TCEQ uses a network of 44 monitoring stations in the HGB area which covers the longitude of [-95.8070, -94.7870] and the latitude of [29.0108, 30.7440]. It collects the ground-level ozone concentration and various meteorological conditions, such as temperature, solar radiation, wind speed, and wind direction at each monitoring station. NOx concentration is collected as well. We apply a standard Kriging interpolation method to compute the ozone hourly concentrations on 20 x 27 grids that cover the HGB area, and feed the interpolation function into DCONTOUR algorithm (Chen 2009) with a user defined threshold ozone value, i.e. 80 ppb (parts per billion) to generate polygons. Such polygons describe ozone pollution hotspots - areas with hourly ozone concentrations above the threshold value. A spatiotemporal clustering algorithm is developed to cluster those polygons based on their spatial and temporal similarities. Our algorithm can help domain experts find interesting spatiotemporal patterns from ozone pollution events and make preliminary predictions for ozone events in the future. For example, our algorithms can find hourly patterns of the high ozone concentrations occurred in similar areas. For people having active outdoor activities or respiratory problems, this can help them make appropriate plans in advance. Figure 1 visualizes one such cluster, i.e. cluster 16. All six polygons in cluster 16 occurred at 3 p.m. along highway interstate 45 north. Clusters 16 are formed due to the emissions from highway traffic vehicles and strong solar radiations which usually happen between 14:00 pm and 15:00 pm each day. If different temporal distance function is adopted, our spatiotemporal clustering algorithm can group polygons that lie at similar locations and occur within certain time intervals, such as three or four hours, as well. Figure 2 visualizes one such cluster, i.e. cluster 9.

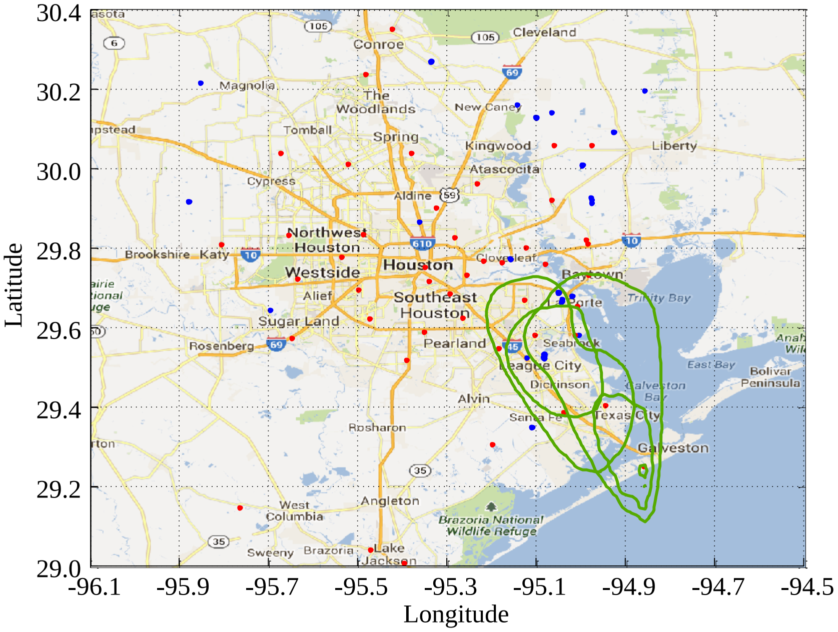

Figure 1. Visualization of cluster 16.

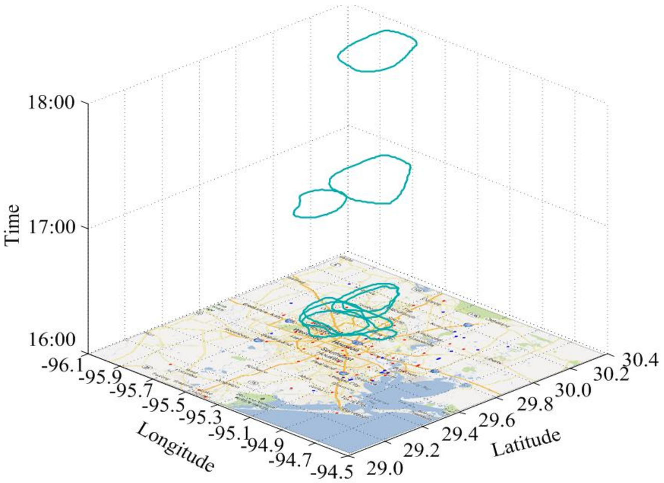

Figure 2. Visualization of cluster 9.

We apply post-processing analysis technique to identify interesting spatiotemporal clusters that are unusual compared to all other clusters. Figure 3 visualizes one such cluster, cluster 26. Cluster 26 is selected due to relative high values of solar radiation (deviation degree 1), and relative high values of temperatures (deviation degree 1). Therefore, cluster 26 has larger interestingness value.

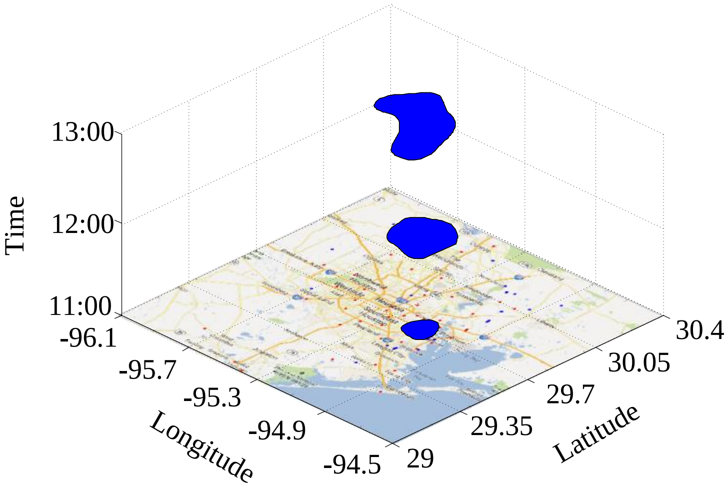

Figure 3. Visualization of cluster 16.

4. References

Chen C, Rinsurongkawong V, Eick C, and Twa M, 2009, Change analysis in spatial data by combining contouring algorithms with supervised density functions, *Proceedings of the 13th Asia-Pacific Conference on Knowledge Discovery and Data Mining*, Bangkok, Thailand, April 2009.

Ertoz L, Steinback M, and Kumar V, 2003, Finding clusters of different sizes, shapes, and density in noisy high dimensional data, *Proceedings of the 3rd SIAM International Conference on Data Mining*, San Francisco, CA, USA, May 01-03, 2003.

Hangouet J, 1995, Computing of the Hausdorff distance between plane vector polylines, *Proceedings of the 8th International Symposium on Computer-Assisted Cartography*, Charlotte, North Carolina, USA, February 27-29, 1995.