



# A data mining framework for environmental and geo-spatial data analysis

Sujing Wang<sup>1</sup> · Christoph F. Eick<sup>2</sup>

Received: 31 January 2017 / Accepted: 16 September 2017 / Published online: 30 September 2017  
© Springer International Publishing AG 2017

**Abstract** Mining geo-spatial data is an important task in many application domains, such as environmental science, geographic information science, and social networks. In this paper, we introduce a data mining framework, which includes pre-processing of environmental and geo-spatial data, geo-spatial data mining techniques, and visual analysis of environmental and geo-spatial data. In particular, we propose new density-based clustering algorithms to identify interesting distribution patterns from geo-spatial data, a change pattern discovery technique to detect dynamic change patterns within spatial clusters, and a post-processing technique to extract interesting patterns and useful knowledge from geo-spatial data. Our density-based clustering algorithms are based on the well-established density-based shared nearest neighbor clustering algorithm, which can find clusters of different shape, size, and densities in high-dimensional data. The post-processing analysis technique allows automatic screening of interesting spatial clusters. The change pattern discovery algorithm is able to detect and analyze dynamic patterns of changes within spatial clusters. This paper focuses on developing a framework integrating a sequence of data mining process including clustering algorithm, analysis technique and pattern changing discovery algorithm. In contrast to previous works in this area, our approaches can cluster and analyze dynamically evolved complex objects, i.e., polygons. We evaluate the effectiveness of our techniques through a challenging real case study involving ozone pollution events in the Houston–

Galveston–Brazoria area. The experimental results show that our approaches can discover interesting patterns and useful information from geo-spatial air-quality data.

**Keywords** Geo-spatial data mining · Geographical information retrieval · Clustering · Change pattern discovery · Post-processing analysis · Polygons · Ozone pollution

## 1 Introduction

Due to technological advances, such as smartphones, general mobile devices, remote sensors, and sensor networks, different types of location-based data become increasingly available. Such data are called geo-spatial data. Geo-spatial data have explicit geographic positioning information included such as latitude and longitude. The geo-spatial data can also integrate multiple other types of data, such as temporal information, social information, textual data, multimedia data, and scientific measurements called enriched geo-spatial data. Such enriched geo-spatial data provide tremendous potentials and research challenges for discovering new useful knowledge. For example, the hourly ozone concentration is an ordinary data. If we add the monitor station location (latitude and longitude), we will get the geo-spatial data. We can also integrate the corresponding meteorological data such as outdoor temperature and wind speed from different resources to make it the enriched geo-spatial data for our analysis.

Meanwhile, mining geo-spatial data is an important task in many application domains, such as environmental science, geographic information systems, and social networks. However, traditional data mining techniques are inefficient in mining geo-spatial data because they do not incorporate the idiosyncrasies of the spatial domains such as spatial auto-correlation, spatial context, and spatial constraints. Spatial

✉ Sujing Wang  
swang3@lamar.edu ; sujing.wang@lamar.edu

<sup>1</sup> Department of Computer Science, Lamar University, Beaumont, TX 77710, USA

<sup>2</sup> Department of Computer Science, University of Houston, Houston, TX 77004, USA

auto-correlation in GIS helps understand the degree to which one object is similar to other nearby objects. Applying traditional data mining techniques to geo-spatial data can result in patterns that are biased or that do not fit the spatial data well [1]. Chawla et al. [2] highlight three reasons that geo-spatial data pose new challenges to data mining tasks: “First, classical data mining deals with numbers and categories. In contrast, spatial data are more complex and include extended objects such as points, lines, and polygons. Second, classical data mining works with explicit inputs, whereas spatial predicates (e.g., overlap) are often implicit. Third, classical data mining treats each input to be independent of other inputs, whereas spatial patterns often exhibit continuity and high auto-correlation among nearby features.” Chawla et al. [2] suggest that traditional data mining tasks be extended to deal with the unique characteristics intrinsic to geo-spatial data. Therefore, new data mining techniques are needed to address these challenges to provide effective solutions to search and mine the wealth of the geo-spatial data.

Clustering can help reveal interesting distribution patterns and serve as the foundation for other data mining and analysis techniques. It is one of the most commonly used data analysis techniques in many application domains. However, most of the existing clustering techniques are inefficient in dealing with enriched geo-spatial data. Moreover, several types of geo-spatial data, i.e., points, trajectories, and polygons, are available in real world applications. Clustering techniques for polygons are still rarely reported in the literature. Section 2 discussed related works in more details. Detail discussion of related works is that polygons serve an important role in the analysis of geo-spatial data as they provide a natural representation for certain types of objects, such as city blocks, city neighborhoods, and pollution hot-spots. We propose two new density-based clustering algorithms by extending the well-established density-based shared nearest neighbor (SNN) [3] clustering algorithm for polygons. Advantages of our clustering algorithms include its capability to find clusters with different shapes, sizes, and densities in high-dimensional data and its tolerance to noise. In addition, our algorithms do not require the number of clusters to be determined in advance.

Geographic dynamics refer to changes that occur across both spatial and temporal dimensions. Different change analysis techniques for geo-spatial data have been developed. However, most of them focus on points and trajectories. The current state-of-the-art is still lacking techniques for analyzing dynamic changes that may occur within the polygon-based spatial clusters across in both spatial and temporal dimensions, as well as formalizing their properties. New change detection algorithms are needed to automate the identification, representation, and computation of geographic dynamics for polygons. We propose a method which can incorporate spatial and temporal distance functions and thresholds to discover dynamic change patterns. The thresh-

olds can be set by the domain experts based on their need of analysis.

The changing discovery algorithm utilizing polygon models can automatically detect and analyze changes occurred across both the spatial and temporal dimensions within geo-spatial clusters. Change analysis is performed by comparing sets of polygon models which capture various types of changes such as expansion, dissipation, and disappear across both the spatial and temporal dimensions. For example, the pattern changing discovery algorithm can be used to study the change patterns of the ozone pollution events such as their formation and expansion. This type of study can help domain experts perform trend analysis and make prediction for future events.

Post-processing analysis techniques are also important tasks in data mining, which can help domain experts to explore and explain this data and the clustering results from a variety of viewpoints. Therefore, we also propose a post-processing analysis technique to extract interesting patterns and useful knowledge from geo-spatial clusters, and to detect dynamic change patterns within spatio-temporal clusters of polygons.

We demonstrate the effectiveness of our clustering and analysis techniques through a challenging real case study involving ozone pollution events in the Houston–Galveston–Brazoria (HGB) area. Ozone has been the major air-quality concern in the HGB area for many years. Due to this concern, an air-quality-monitoring network has been well established to continuously monitor the ground-level ozone concentration in this area. Various types of meteorological data related to ozone formation and transportation, such as outdoor temperature, solar radiation, wind speed, and wind direction, are also recorded by different monitoring stations in this area. Time series of  $NO_x$  concentrations measured in the HGB area are collected as well. All these multi-source geo-spatial data provide tremendous potential for discovering new useful knowledge about ozone pollution formations and transportation in HBG area. Our approaches can help domain experts identify interesting spatial patterns of ozone pollution events, examine important factors for controlling ozone concentrations, investigate the efficacy of emission control strategies, as well as make preliminary predictions for future ozone pollution events. Clustering algorithms can identify hourly patterns of high ozone concentrations occurred at similar areas. People with respiratory problems and people who are physically active outdoors can benefit from this analysis so that they can make plans in advance about spending time outdoors. For example, our analysis results could be used to suggest people avoid high ozone concentration time and location, such as 2: 00 pm to 4:00 pm near Highway I 45 daily. Decisions that areas along highway traffic and chemical plants could have high frequencies of ozone pollution events during summer time and people should try to avoid

planning outdoor activities at these areas and time can be implied by the analysis step findings.

Our research contributions are summarized below:

- A new framework for clustering and analyzing geo-spatial data is presented.
- Two new density-based clustering algorithms are developed.
- A post-processing analysis technique is implemented to identify interesting geo-spatial clusters.
- A change pattern discovery algorithm is introduced to detect and analyze dynamic patterns of changes within spatial clusters.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the framework. Section 4 presents two density-based clustering algorithms for geo-spatial data in detail. Section 5 introduces two analysis techniques for spatial clusters, e.g., anomalous clusters discovery and change pattern discovery techniques. Section 6 evaluates our work with case studies on geo-spatial ozone pollution data in the Houston–Galveston–Brazoria (HGB) area. Section 7 concludes our study and discusses potential future works.

## 2 Related work

Spatio-temporal clustering for point-wise objects and trajectories has been heavily studied in past work. Kulldorff [4] introduced basic spatial scan statistics to search spatio-temporal cylinders representing areas where the point objects occur consistently for a significant amount of time; spatio-temporal cylinders are circular regions occurring within a certain time interval. Iyengar [5] extended the basic spatial scan statistics [4] using flexible square pyramid shapes instead of cylinders for spatio-temporal clusters that can either grow or shrink over time and that can also move over time. Wang et al. [6] proposed two spatio-temporal clustering algorithms, i.e., ST-GRID and ST-DBSCAN. ST-DBSCAN is an extension of the DBSCAN algorithm to perform spatio-temporal clustering by introducing the second parameter of temporal neighborhood radius in addition to the spatial neighborhood radius. ST-GRID is a grid-based clustering approach which maps the spatial and temporal dimensions into cells. Birant et al. [7] also improved DBSCAN for spatio-temporal clustering and applied it to discover spatio-temporal distributions of physical seawater characteristics in Turkish seas. A density factor is assigned to each cluster for detecting some noise points when clusters of different densities exist. The density factor of a cluster captures the degree of the density of the cluster. ST-SNN and ST-SEP-SNN algorithms are based on the Shared Nearest Neighbor clustering algorithm

[3], which can automatically find clusters of different densities in high-dimensional data for complex geometry objects, i.e., polygons.

There are extensive studies of spatio-temporal trajectory clustering techniques [8] as well. Gaffney et al. [9] proposed a clustering algorithm for continuous trajectories, which is based on a principled method for probabilistic modeling of a set of trajectories as individual sequences of points generated from a finite mixture model consisting of regression model components. Pelekis et al. [10] introduced different trajectory distance functions for trajectory clustering taking into account several spatio-temporal characteristics of the trajectories, such as direction, velocity, and co-location in space and time. Nanni and Pedreschi [11] presented an OPTICS-based temporal focusing approach for clustering moving object trajectories based on a simple notion of distance between trajectories. It clusters trajectories using all possible time intervals, evaluates the results and finds the best clustering. Rinzivillo et al. [12] proposed a progressive clustering approach to analyze the trajectories of moving objects supported by visualization and interaction techniques. It progressively applies different distance functions for spatio-temporal data in each step to optimize the outcome of the algorithm. Li et al. [13] introduced the concept of moving micro-cluster to catch some regularities of a moving object. The micro-clusters are kept geographically small at any time.

Joshi et al. [14] proposed a spatio-temporal polygonal clustering algorithm, STPC. STPC extends DBSCAN algorithm to cluster spatio-temporal polygons by redefining the neighborhood of a polygon as the union of its spatial neighborhood and temporal neighborhood. The temporal aspect is constant or reduced to a fixed interval or time instance when calculating spatial neighbors of a polygon. Moreover, the spatial dimension is instead held to a constant space when calculating temporal neighbors of a polygon. Therefore, STPC only clusters polygons that do not change their locations, sizes, and shapes over time. Only the non-spatial attributes or properties might change with time. Wang et al. [15] proposed a density-based clustering algorithm called DCONTOUR that uses contour lines to determine cluster boundaries, two distance functions for geo-spatial polygons to measure the distance between pair of overlapping polygons, and a clustering algorithm called POLY-SNN for geo-spatial polygons which does not consider the spatial and temporal domains during clustering. Therefore, POLY-SNN cannot identify spatio-temporal clusters. In contrast to STPC and POLY-SNN, ST-SNN and ST-SEP-SNN algorithms can cluster polygons that dynamically change their sizes and shapes through time and can move over time. Furthermore, ST-SNN and ST-SEP-SNN cope better with high-dimensional data and respond better to data with varying densities compared to DBSCAN-based clustering algorithms.

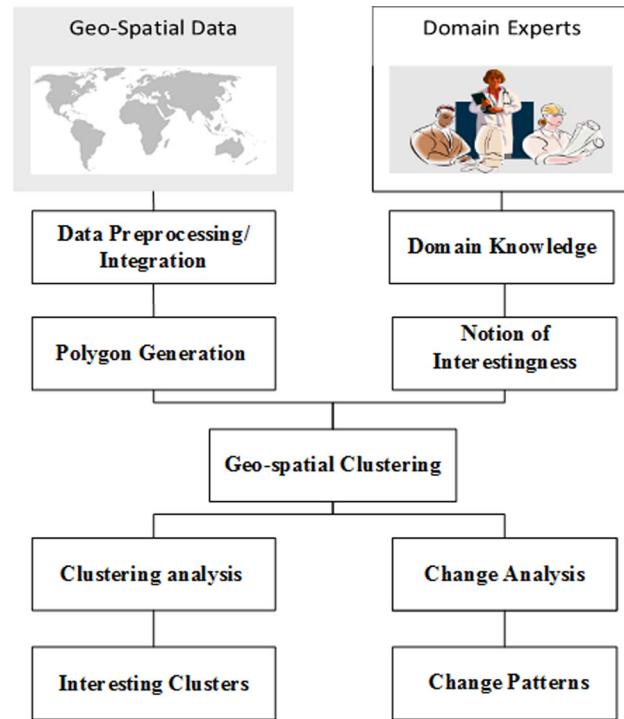
Research studies on spatio-temporal change discovery of point-based or trajectory data, such as moving clusters [16], flocks [17, 18], convoys [19], and trajectory clustering [20, 21], are increasingly common in the literature. Fewer methods are available for polygons. McIntosh and Yuan [22] introduced a methodology for analysis of polygon distributions, which focuses on the analysis of the internal values of polygon attributes, rather than changes in the spatial properties of the polygons. Rinsurongkawong et al. [23] proposed an approach to analyzing changes in spatial data by utilizing polygon models. Change patterns capture how the most recent data differ from the data model established from the historical data. Change analysis is performed by comparing sets of polygons which capture various types of changes. However, our change pattern discovery algorithm is for detecting geographic dynamics, which refer to the changes that occur across both the spatial and temporal dimensions simultaneously within spatio-temporal clusters of polygons. Stell et al. [24] proposed a novel approach to modeling the evolution of spatial entities over time by using bigraphs. The links in a bigraph are used to represent the sharing of a common ancestor and the places in a bigraph to represent spatial nesting. The bigraphical reaction rules provided are able to model situations such as two crowds of people merging together while still keeping track of the resulting crowd's historical links.

### 3 Framework architecture

Our framework is an integration of pre-processing techniques, two density-based clustering algorithms, a post-processing analysis technique, a change pattern discovery algorithm, and visualization techniques. The architecture of our framework is summarized in Fig. 1. It consists of the following five steps:

- Step 1 pre-processing of geo-spatial data.
- Step 2 generate polygons from geo-spatial point data.
- Step 3 integrate domain knowledge and apply clustering algorithms to group polygons based on their spatial and temporal similarities to identify interesting distribution patterns.
- Step 4.a utilize post-processing analysis technique to identify interesting clusters whose member variables have attribute values that deviate significantly from those of the entire population.
- Step 4.b detect and analyze dynamic patterns of change within geo-spatial clusters .
- Step 5 Visualize interesting clusters and change patterns.

We collected air-quality-related data from TCEQ's (Texas Commission on Environment) Reciprocal Area Final Clustering.



**Fig. 1** Framework architecture

Quality) Web site [25]. TCEQ uses a network of 44 monitoring stations in the HGB area which covers the geographical region within the longitude of  $[-95.8070, -94.7870]$  and the latitude of  $0.5\text{cm}$   $[29.0108, 30.7440]$ . Ozone has been the main air-quality concern in the HGB area for many years. Regional meteorological conditions combined with the variety of emissions from industries and transportation make the city a prime media for ground-level ozone formation [25]. The rapid growth of the city causes regional emissions to continue increasing.

We apply the standard Kriging interpolation method to compute the ozone hourly concentrations on  $20 \times 27$  grids that cover the HGB area and feed the interpolation function into the DCONTOUR algorithm [26] with a defined threshold, i.e., 80 ppb (parts per billion) to compute polygons. Such polygons describe ozone pollution hot-spots (areas whose hourly ozone concentrations are above the input threshold, i.e., 80 ppb). Next, clustering algorithms are adopted to cluster those polygons based on their spatial and temporal similarities to identify the ozone pollution spatio-temporal patterns in the HGB area. After that, we apply analysis techniques to identify interesting spatio-temporal clusters of polygons, and to detect and analyze the dynamic change patterns within spatio-temporal clusters of polygons. Our approaches can help analysts find interesting spatio-temporal patterns from ozone pollution events and make preliminary predictions for ozone pollution events in the future. For example, our algorithms can find hourly patterns of high ozone

concentrations that occurred in similar areas. Moreover, our post-processing analysis technique can identify clusters of polygons with attribute values that are much smaller or larger than others. Such anomalous clusters are exceptional in some sense and are often of unusual importance.

#### 4 Spatio-temporal clustering algorithms

Spatio-temporal clustering is an upcoming research topic which focuses on studying and implementing novel clustering techniques for spatio-temporal data. The Shared Nearest Neighbor (SNN) clustering algorithm [3] is a well-established density-based clustering algorithm. SNN defines the similarity between pairs of points in terms of how many nearest neighbors the two points share. The SNN clustering algorithm works well for ordinary data (non-spatial and non-temporal data). However, it is impractical to use for clustering spatio-temporal data. The clustering process for spatio-temporal data is more complex than for non-spatial and non-temporal data because spatio-temporal clustering algorithms have to consider the spatial and temporal neighbors of objects. Therefore, we improve SNN for spatio-temporal data. We extend SNN to cluster polygons generated from geo-spatial point data by redefining the spatio-temporal similarity between pairs of polygons taking into account both their spatial and temporal similarities. We also redefine the density-based concepts of core polygons. Each polygon  $p$  is associated with a time  $t$  when it occurs, and a set of non-spatial attributes. Henceforth, all polygons are spatio-temporal polygons unless specified otherwise. A spatio-temporal cluster of polygons is a group of polygons that lie in close proximity in both space and time domains. Given below are the definitions for the SNN-based concepts for polygons.

We propose two spatio-temporal clustering algorithms, ST-SNN and ST-SEP-SNN. The difference between ST-SNN and ST-SEP-SNN is based on the user-defined parameters given as input and how to compute the spatio-temporal distance between two objects. Any function that can compute the distance between a pair of polygons, such as Hausdorff distance [27], Fréchet distance [28], PDF [29], Overlay distance [30], Hybrid distance [15], etc., can be used in our clustering algorithms. Any function that can compute the temporal distance between a pair polygons can be adopted as well such as Eq. 1. Moreover, different temporal distance functions might be used for different analysis tasks. The following temporal distance function is an example used for our case study:

$$dist_t(p, q) = \begin{cases} abs(h(p) - h(q)), & abs(h(p) - h(q)) \leq 12 \\ 24 - abs(h(p) - h(q)), & \text{otherwise} \end{cases} \quad (1)$$

where  $p$  and  $q$  are a pair of polygons,  $h(p)$  function returns the hour information associated with polygon  $p$  ( $0 \leq h(p) \leq$

23), function  $abs()$  returns the absolute value. The following Hybrid distance [15] is applied to compute the spatial distances between polygons due to its capability of handling overlapping polygons:

$$dist_s(p, q) = w_s \times dist_{\text{Overlay}}(p, q) + (1 - w_s) \times dist_{\text{Hausdorff}}(p, q) \quad (2)$$

where  $w_s$  is the weight factor associated with the Overlay distance ( $0 \leq w_s \leq 1$ ). The  $k$ -nearest spatial neighbor list and  $k$ -nearest temporal neighbor list for each polygon  $p$ , denoted by  $k\text{-SPN-List}(p)$  and  $k\text{-TN-List}(p)$ , are generated by computing its  $k$ -nearest spatial neighbor and  $k$ -nearest temporal neighbor only. The nearest spatio-temporal neighbor list of a polygon  $p$ , denoted by  $NN(p)$ , is calculated as the intersection of the  $k$ -nearest spatial neighbor list and the  $k$ -nearest temporal neighbor list of polygon  $p$ :

$$NN(p) = k\text{-SPN-List}(p) \cap k\text{-TN-List}(p) \quad (3)$$

The similarity between a pair of polygons  $p$  and  $q$ , denoted by  $similarity(p, q)$ , is the number of the nearest spatio-temporal neighbors that they share:

$$similarity(p, q) = |NN(p) \cap NN(q)| \quad (4)$$

where  $NN(p)$  is the set of  $k$  nearest spatio-temporal neighbors of polygon  $p$ .

The SNN density of polygon  $p$  is defined as the number of polygons that share  $Eps$  or more nearest neighbors with polygon  $p$ :

$$density(p) = |\{q \in D | similarity(p, q) \geq Eps\}| \quad (5)$$

The core polygons are identified by using a user-specified parameter  $MinPs$  and all polygons in the dataset  $D$  that have the SNN density of at least  $MinPs$ :

$$CoreP(D) = \{p \in D | density(p) \geq MinPs\} \quad (6)$$

Clusters are then formed by computing the transitive closure of the polygons that can be reached from an unprocessed core polygon using their respective nearest neighbor lists; 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 polygons are classified as outliers and not included in any clusters.

We use two polygons  $p$  and  $q$  as an example to illustrate the computation. For example, polygon  $p$  represented an ozone hot-spot occurred at 4:00 pm, polygon  $q$  occurred at 2:00 pm, the temporal distance between  $p$  and  $q$  is 2 based on Eq. 1. Assume the overlay distance between polygons  $p$  and  $q$  is 0.5; the Hausdorff distance between  $p$  and  $q$  is 0.3; the

hybrid distance between  $p$  and  $q$  is 0.4 based on Eq. 2 with weight  $w$  equal to 0.5. If we choose ( $k=5$ ), assume the K-nearest temporal neighbor list for polygon  $p$  5-TN-List( $p$ ) is  $\{5,10,12,15,20\}$ ; the 5-nearest spatial neighbor list k-SPN-List( $p$ ) is  $\{10,12,15,25,27\}$ , the nearest spatio-temporal neighbor list of polygon  $p$ , NN( $p$ ) is  $\{10, 12, 15\}$  based on Eq. 3. The 5-nearest temporal neighbor list for polygon  $q$ , 5-TN-List( $p$ ), is  $\{3,10,12,16,25\}$ ; the 5-nearest spatial neighbor list of polygon  $q$ , 5-SPN-List( $q$ ), is  $\{10,12,16,25,29\}$ ; the nearest spatio-temporal neighbor list of polygon  $q$ , NN( $q$ ), is  $\{10, 12, 16, 25\}$ . The similarity between  $p$  and  $q$  is 3 according to Eq. 4 because the intersection of NN( $p$ ) and NN( $q$ ) is  $\{10, 12, 16\}$ . The rest of algorithm is performed as SNN.

ST-SEP-SNN does not integrate spatial distance and temporal distance as a single spatio-temporal distance. In contrast to ST-SEP-SNN, ST-SNN uses a weighted sum of the spatial distance and the temporal distance between two polygons  $p$  and  $q$  to calculate the spatio-temporal distance between polygons  $p$  and  $q$ , denoted by  $dist_{st}(p, q)$ :

$$dist_{st}(p, q) = w_{sp} \times dist_s(p, q) + (1-w_{sp}) \times dist_t(p, q) \quad (7)$$

where  $w_{sp}$  is the weight factor associated with the spatial distance ( $0 \leq w_{sp} \leq 1$ );  $dist_s$  is any function that can compute the normalized spatial distance between two polygons  $p$  and  $q$ ;  $dist_t$  is any functions that can compute the normalized temporal distance between two polygons  $p$  and  $q$ . Then, we rank the obtained spatio-temporal distance matrix to get the  $k$ -nearest spatio-temporal neighbor list for each polygon. In this case, the sizes of the nearest neighbor lists for all polygons in the data are the same. ST-SNN form clusters the same way as ST-SEP-SNN. The pseudo-code of ST-SNN and ST-SEP-SNN is given in Algorithm 1.

In general, ST-SEP-SNN uses separate  $k$ -nearest spatial neighbor list and  $k$ -nearest temporal neighbor list and does not try to integrate the spatial distance and the temporal distance into a single spatio-temporal distance. The nearest neighbor lists of all polygons may have different cardinalities  $m$  ( $m \leq k$ ). This property distinguishes ST-SEP-SNN from ST-SNN.

Both ST-SNN and ST-SEP-SNN require several user-defined parameters that have significant impacts on clustering results. These user-defined parameters need to be changed and adapted according to the data being clustered:

- $k$  the size of the nearest neighbor list. It is the most important parameter as it determines the granularity of the clusters. In general, if  $k$  is too small, both ST-SNN and ST-SEP-SNN will tend to find many small clusters and a lot of outliers. On the other hand, if  $k$  is too large, both ST-SNN and ST-SEP-SNN will tend to find only a few large clusters.

**Input:** polygon dataset  $D = \{p_1, p_2, \dots, p_n\}$ , size of nearest neighbor list  $k$ , the core polygon threshold  $MinPs$ , the similarity threshold  $Eps$

**Output:** set of clusters  $C_i$  of polygons  
**Algorithm:**

```

1. for every polygon  $p$  in  $D$ 
2.   compute  $k$ -nearest neighbor list:  $NN(p)$ 
3.   mark  $p$  ‘unprocessed’.
4. end for
5. for every pair of polygons  $p$  and  $q$  in  $D$ 
6.   compute  $similarity(p, q)$ ;
7. end for
8. for every polygon  $p$  in  $D$ 
9.   compute  $density(p)$ 
10.  if  $density(p) \geq MinPs$  then
11.    mark  $p$  as ‘core polygon’;
12.  end if
13. end for
14. for every core polygon  $p$  in  $D$ 
15.  if  $p$  is marked ‘unprocessed’ then
16.    form a cluster  $C_i$  that can be reached from  $p$  following the entries of the respective  $NN$ -lists of core polygons;
17.  end if
18.  mark all polygons in  $C_i$  ‘processed’;
19. end for
20. return set of generated clusters  $C_i$  of polygons;
```

**Algorithm 1:** Pseudo-code of ST-SNN and ST-SEP-SNN

- $MinPs$  the core polygon threshold. It is the minimum number of the shared neighbors required for the core polygons. It allows users to control how many polygons are needed to qualify a polygon as a core polygon.  $MinPs$  should be smaller than  $k$ .
- $Eps$  the density threshold. It is used as the criteria to define the SNN density of each polygon.  $Eps$  should be smaller than  $k$  as well.

Both ST-SNN and ST-SEP-SNN follow the structure of SNN. The time complexity of ST-SNN and ST-SEP-SNN are the same as SNN which is  $O(n^2)$  without the use of an indexing structure, where  $n$  is the number of polygons in the dataset. If an indexing structure such as a  $k-d$  tree or an  $R^*$  is used, the time complexity will be reduced to  $O(n \times \log(n))$ . The space complexity is  $O(k \times n)$  since only the  $k$ -nearest neighbor need to be stored, while the  $k$ -nearest neighbor can be computed once and used repeatedly for different runs of the algorithms with different parameter values.

## 5 Analysis techniques for spatio-temporal clusters

### 5.1 Post-processing analysis technique

Our post-processing analysis technique allows automatic screening of the obtained clusters to identify interesting ones,

whose members have attribute values that deviate significantly from those of the entire population. For example, we try to find clusters of polygons with attribute values that are much smaller or larger than others, e.g., cluster of polygons with extremely high ozone concentrations. Such anomalous clusters are exceptional in some sense and are often of unusual importance. The domain experts could use our post-processing analysis technique to automatically identify such clusters. Therefore, it is desirable to have some assessment of the degree to which the attribute values of a cluster are anomalous. Since box plots is a commonly used method for showing the distribution of values of a single numerical attribute, and for comparing how attribute values vary among different clusters of objects, our post-processing analysis technique is developed based on box plots. We assume a dataset  $D$  with  $n$  attributes and a set of clusters in  $D$  identified by spatio-temporal clustering algorithm; e.g., ST-SEP-SNN and ST-SNN. Let  $(a_{i,j}, b_{i,j})$  be the interquartile range ( $IQR$ ) for attribute  $j$  of a cluster  $C_i$ , with  $a_{i,j} > b_{i,j}$ , and  $(a'_j, b'_j)$  be the  $IQR$  for attribute  $j$  of dataset  $D$  with  $a'_j > b'_j$ , we compute the degree of deviation for attribute  $j$  in cluster  $C_i$  compared with the dataset  $D$  as follows:

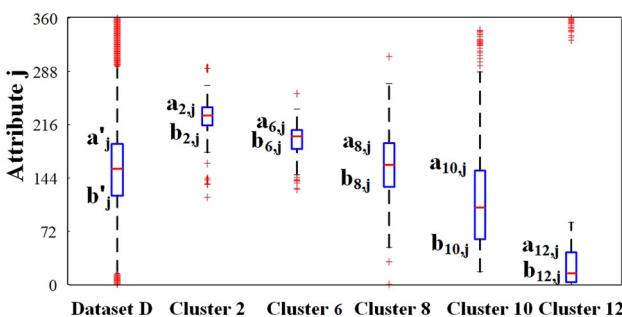
$$R_{i,j} = 1 + \frac{\max(a'_j - a_{i,j}, 0) + \max(b'_j - a_{i,j}, 0) - \max(a'_j - b_{i,j}, 0) - \max(b'_j - b_{i,j}, 0)}{a_{i,j} - b_{i,j}} \quad (8)$$

$R_{i,j}$  could be any numbers between  $-1$  and  $1$ . We explain the equation described above with the help of several examples shown in Fig. 2, which displays the box plots for wind direction of several selected clusters, e.g., clusters 2,6,8,10, and the dataset  $D$ .

For example, the deviation degree of wind direction for cluster 2,  $R_{2,j}$  can be calculated as:

$$R_{2,j} = 1 + \frac{0 + 0 - 0 - 0}{a_{2,j} - b_{2,j}} = 1 \quad (9)$$

where  $a_{2,j} > b_{2,j} > a'_j > b'_j$  holds as shown in Fig. 2.  $R_{2,j}$  is equal to 1, which means that cluster  $C_2$  has significant



**Fig. 2** Box plots for attribute  $j$

different values (much larger) of wind directions compared to the entire dataset  $D$ , as shown in Fig. 2.

Consider cluster 6, its box plot overlaps with the box plot for  $D$  and  $a'_{6,j} > a'_{j,D} > b'_{6,j} > b'_{j,D}$ .  $R_{6,j}$  can be calculated as:

$$R_{6,j} = 1 + \frac{0 + 0 - (a'_j - b'_{6,j}) - 0}{a'_{6,j} - b'_{6,j}} = \frac{b'_{6,j} - a'_j}{a'_{6,j} - b'_{6,j}} \quad (10)$$

Apparently,  $0 < R_{6,j} < 1$ . As shown in Fig. 2, the box plots of cluster 6 overlap with the box plot for dataset  $D$  above its 50th percentile line. Therefore,  $R_{6,j}$  is greater than 0 but less than 1 which means the wind directions of cluster 6 do not deviate significantly from those of the entire population.

For cluster 8, its box plot is covered completely by the box plot for  $D$ , and  $a'_{8,j} > a'_{j,D} > b'_{8,j} > b'_{j,D}$  which means that the wind directions of cluster 8 have similar values with the entire dataset  $D$ . Therefore, the wind directions of cluster 8 are not interesting attributes.

$$R_{8,j} = 1 + \frac{(a'_j - a'_{8,j}) + 0 - (a'_j - b'_{8,j}) - 0}{a'_{8,j} - b'_{8,j}} = 0 \quad (11)$$

Consider cluster 10, its box plot overlaps with the box plot for  $D$  below its 50th percentile line, and  $a'_{10,j} > a'_{j,D} > b'_{10,j} > b'_{j,D}$ :

$$\begin{aligned} R_{10,j} &= 1 + \frac{(a'_j - a'_{10,j}) + 0 - (a'_j - b'_{10,j}) - (b'_{j,D} - b'_{10,j})}{a'_{10,j} - b'_{10,j}} \\ &= \frac{b'_{10,j} - b'_{j,D}}{a'_{10,j} - b'_{10,j}} \end{aligned} \quad (12)$$

Obviously,  $-1 < R_{10,j} < 0$ . As shown in Fig. 2, the wind directions of cluster 10 overlap with dataset  $D$  below its 50th percentile line which means that the wind directions of cluster 10 have smaller values compared with dataset  $D$ . For cluster 12, its box plot is displayed below the box plot for  $D$ , and  $a'_{12,j} > b'_{j,D} > a'_{12,j} > b'_{12,j}$  which means the wind directions of cluster 12 have much smaller values compared with dataset  $D$ . Therefore, cluster 12 is an interesting cluster based on wind directions.

$$R_{12,j} = \frac{-a'_{12,j} + b'_{12,j}}{a'_{12,j} - b'_{12,j}} = -1 \quad (13)$$

In general,  $R_{i,j}$  is interesting if  $R_{i,j}$  is equal to 1 or  $-1$ , which means that cluster  $C_i$  has significant different values of attribute  $j$  compared to the entire dataset  $D$ . The interestingness score of cluster  $C_i$  is calculated based on the values

of all  $R_{i,j}$  associated with  $C_i$ . Let  $O_i = \{r_{i,1}, r_{i,2}, \dots, r_{i,n}\}$  be the set of deviation degrees of  $n$  attributes in  $C_i$ ; in general, the interestingness score of cluster  $C_i$  is a function of  $O_i$ :

$$I(C_i) = f(O_i) \quad (14)$$

Different interestingness functions may be adopted for different analysis tasks. Moreover, domain knowledge is crucial in determining interestingness functions. Therefore, the following interestingness functions are proposed based on domain experts' notion of interestingness. For a cluster  $C_i$ , we calculate the Ozone Formation Potential Index (*OFPI*) defined by the following equation:

$$OFPI_i = \frac{1}{3}R_{i,NO_x} + \frac{1}{3}R_{i,T} + \frac{1}{3}R_{i,SR} \quad (15)$$

where  $R_{i,NO_x}$  is the degree of deviation of  $NO_x$  concentrations in  $C_i$ ,  $R_{i,T}$  is the degree of deviation of the outdoor temperatures in  $C_i$ ,  $R_{i,SR}$  is the degree of deviation of solar radiations in  $C_i$ . Note that *OFPI* function is a linear function of these three variables, i.e.,  $R_{i,NO_x}$ ,  $R_{i,T}$ , and  $R_{i,SR}$ , because  $NO_x$  concentration, outdoor temperature and solar radiation are key control factors in ozone pollution events. The negative  $R_{i,j}$  values will contribute negatively to *OFPI* because lower temperatures, lower  $NO_x$  concentrations, and lower amounts of solar radiation will slow down the ozone formation process. However, the other two attributes, i.e., wind speed and wind direction, contribute to the ozone pollution dispersion process. Hence, the Ozone Dispersion Index (*ODI*) for cluster  $C_i$  is calculated as follows:

$$ODI_i = \exp(1 - 0.4 \times |R_{i,WD}| + 0.6 \times R_{i,WS}) \quad (16)$$

where  $R_{i,WD}$  is the degree of deviation of wind directions in  $C_i$ ,  $R_{i,WS}$  is the degree of deviation of wind speeds in  $C_i$ . Note that *ODI* is an exponential function due to the fact that dispersion distribution satisfies Gaussian air pollutant dispersion equation. We use the absolute value of  $R_{i,WD}$  because the lower degree of wind direction and high degree of wind direction contribute equally to the impact scope of the ozone dispersion process. The interestingness score (*IS*) for cluster  $C_i$  is computed as follows:

$$IS_i = OFPI_i \times ODI_i \quad (17)$$

The proposed interestingness function is only an example for identifying interesting clusters related to ozone pollution impact scope. Our goal is to find clusters with either relative high or low interestingness scores, i.e., clusters with either relative large or small ozone pollution impact scopes under unusual environmental conditions and  $NO_x$  concentrations. Those are the anomalous clusters that domain experts are

interested in and want to further analyze. Other interestingness functions for different analysis tasks can be developed as well.

## 5.2 Change pattern discovery algorithm

In this section, we introduce a methodology for discovery and analysis of changes that may occur within a spatio-temporal cluster. The study of the change patterns within clusters can help us classify areas experiencing similar phenomenon across a period of time and to perform trend analysis and make predictions about the future occurrence of ozone pollution events. For example, in order to perform trend analysis and make predictions, it will be very helpful if we can study the change patterns of the ozone pollution events such as their formation and expansion. Therefore, we categorize individual changes according to the spatial relationship among polygons within the same cluster into the following four primitive patterns:

- Formation when the number of polygons at time  $t_i$  is increased from zero at time  $t_{i-1}$ .
- Expansion the overall areas covered by all polygons occurred at time  $t_i$  is increased compared to time  $t_{i-1}$ .
- Dissipation the overall areas covered by all polygons occurred at time  $t_i$  is decreased compared to time  $t_{i-1}$ .
- Disappear the number of polygons at time  $t_i$  is changes to zero from nonzero at time  $t_{i-1}$ .

These individual changes can also be linked into sequences to describe the evolution of a spatio-temporal cluster over time. Duration represents the amount of time that the ozone pollution events exist within a cluster. The pseudo-code of polygon-based change pattern discovery algorithm (Poly-CD) is given in Algorithm 2.

Poly-CD also measures the spatial properties of multiple polygons within a cluster, such as overlap, distance, and directional relation. Hausdorff distance [31] is applied to compute the distance between a pair of polygons. Overlay distance [30] is utilized to calculate the overlap between a pair of polygons. In order to track the movement of the spatio-temporal clusters with respect to space and time, Poly-CD also computes the centroid of each polygon. Tracking the movement of the centroids of polygons in a cluster will enable us to identify the directions of the clusters' movement. The change pattern vector can be used to find similar change patterns at different locations and may in turn help in predicting the future change patterns of ozone pollution events. The change patterns discovered from the spatio-temporal clusters can be used in order to make preliminary predictions for the future movement of the cluster. For example, if a cluster is more likely to expand in the near future, more resources

**Input:** set of spatio-temporal clusters  $C_i$  in  $D$   
**Output:** a vector of change pattern of cluster  $C_i$ , duration of  $C_i$   
**Algorithm:**

1. **for** every spatio-temporal cluster  $C_i$  in  $D$
2.    initialize  $change_{C_i} = \text{null}$ ;  $duration_{C_i} = 0$ ;
3. **end for**
4. **for** each time  $t$  in  $C_i$
5.    compute the number of polygons  $N_t$  occurred at time  $t$ .
6. **end for**
7. **for** every two continues time  $t_i$  and  $t_{i-1}$  occurred in cluster  $C_i$
8.    compute the change patterns:
9.    **if**  $N_t \geq 1$  and  $N_{t-1} = 0$  **then**
10.       $change_{C_i}.\text{add(formation)}$ ;  $duration_{C_i}++$ ;
11.    **if**  $N_t \geq 1$  and  $N_{t-1} \neq 0$  and  $area_t > area_{t-1}$  **then**
12.       $change_{C_i}.\text{add(expansion)}$ ;  $duration_{C_i}++$ ;
13.    **if**  $N_t \geq 1$  and  $N_{t-1} \neq 0$  and  $area_t < area_{t-1}$  **then**
14.       $change_{C_i}.\text{add(dissipation)}$ ;  $duration_{C_i}++$ ;
15.    **if**  $N_t = 0$  and  $N_{t-1} \geq 1$  **then**
16.       $change_{C_i}.\text{add(disappear)}$ ;
17. **end for**
18. **return** the change pattern vector of cluster  $C_i$ ,  $change_{C_i}$ , and the duration of cluster  $C_i$ ,  $duration_{C_i}$

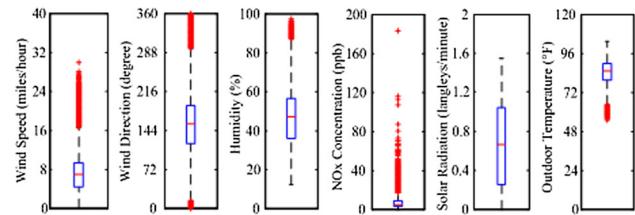
#### Algorithm 2: Pseudo-code of Poly-CD

should be assigned to this region in order to prevent the cluster from expanding.

## 6 Case study

### 6.1 Multi-source geo-spatial data

An area with specified air-quality index violating the NAAQS (National Ambient Air Quality Standards) is defined as the air-quality non-attainment area. The HGB area is currently classified as an ozone non-attainment area [25]. To improve the air quality in this area, an air-quality-monitoring network has been well established to continuously monitor the ground level of ozone concentration, various meteorological conditions, and  $NO_x$  concentration in this area, which provide large amounts of spatio-temporal data associated with ozone pollution events. Time series measurements for such data constitute enriched geo-spatial data. Data mining techniques are needed to facilitate the information extraction and knowledge discovery from such enriched geo-spatial data. In particular, we collected the raw data from the time-frame of 1 am on April 1, 2010, through 11 pm on November 30, 2010, from TCEQ's Web site [25]. These data include hourly measurements of ground-level ozone concentration, solar radiation, outdoor temperature, wind direction, wind speed, and  $NO_x$  concentration. We then apply the DCONTOUR algorithm to generate polygons. These polygons are represented by closed contour lines. The area within each polygon has an hourly ozone concentration higher than the user input threshold, e.g., 80 ppb. The shape and area of a



**Fig. 3** Statistical distribution of ozone pollution data

polygon reflect the impact location and scope of an ozone pollution event. The ozone pollution events are affected by multiple factors of ozone precursors and meteorological conditions. A total of 460 polygons have been identified for the density threshold of 80 ppb. Figure 3 provides the box plots for the statistical distributions of four corresponding meteorological attributes and  $NO_x$  concentration. It can be observed that the distribution of wind directions cover from  $0^\circ$  to  $360^\circ$ , and the majority are between  $120^\circ$  and  $200^\circ$ , which means that the wind can come from any direction, e.g., the wind can blow from the Gulf of Mexico as the result of a typical sea breeze encompassing east-southeast to south on shore winds in the HGB area [32]. This winds can help transport the emissions from point sources, area sources, and on-road mobile sources from upwind to different downwind regions. The solar radiation is between 0.3 and 1.05 langleys/min, which is relatively strong and in favor of ozone formation. The higher wind speed between 4 and 9 miles/h promotes the precursor transportation. The average  $NO_x$  concentration is below 10 ppb and the peak concentration is 183 ppb. The outdoor temperatures are between 75 and 90°F. In summary, the integrated conditions of high outdoor temperature, high  $NO_x$  concentration, and high solar radiation create suitable conditions of high ozone concentration events in the HGB area. The spatio-temporal clusters identified by our clustering algorithms satisfy all of these conditions.

### 6.2 Spatio-temporal clustering evaluations

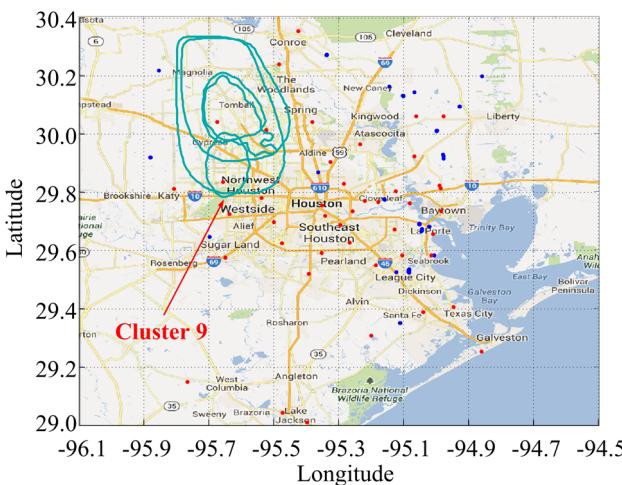
In the following section, we focus on demonstrating the effectiveness of both ST-SNN and ST-SEP-SNN clustering algorithms in finding interesting spatio-temporal patterns from ozone pollution events in HGB area. A spatio-temporal cluster of polygons is a group of polygons representing high ozone concentration hot-spots that are in close proximity in both space and time, and possibly share other attributes. Statistical analyses are also presented to interpret the discovered patterns. Hybrid distance [15] is applied to compute the spatial distance between a pair of polygons due to its capability of handling overlapping polygons.

### 6.2.1 ST-SNN clustering and analysis

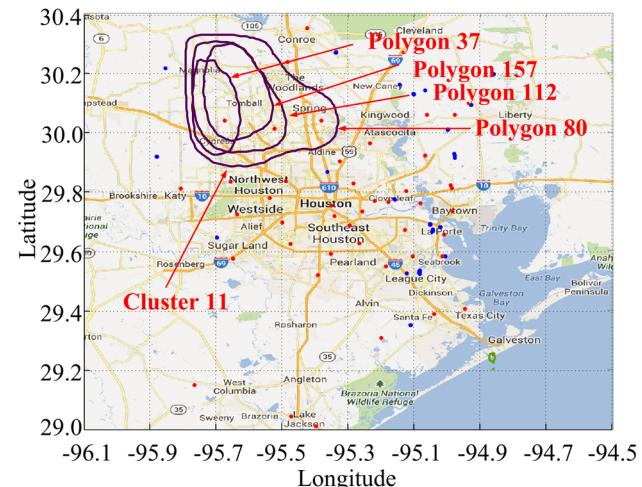
The task of this case study is to apply ST-SNN to find interesting spatio-temporal patterns from ozone pollution data in HGB area. In particular, we are interested in finding hourly patterns of ozone pollution events that occurred in similar areas. This type of study can help domain experts identify not only similar impact scopes of the ozone pollution events in space but also their corresponding time instants or time intervals. This can help domain experts gain knowledge from the past. Unlike the time slicing approaches, which perform a snapshot clustering at each time stamp or time interval, ST-SNN takes into account both spatial and temporal distances between polygons and is able to detect clusters of polygons that are similar in both spatial and temporal dimensions simultaneously.

The input parameters for ST-SNN are  $k = 5$ ,  $MinPs = 3$ ,  $Eps = 2$ ,  $w = 0.5$ . We use a relatively smaller value for  $k$ , i.e., 5, to find smaller compact clusters. There are 17 clusters found by ST-SNN. Figures 4, 5, 6 and 7 visualize the clusters 9, 11, 14, and 16, respectively.

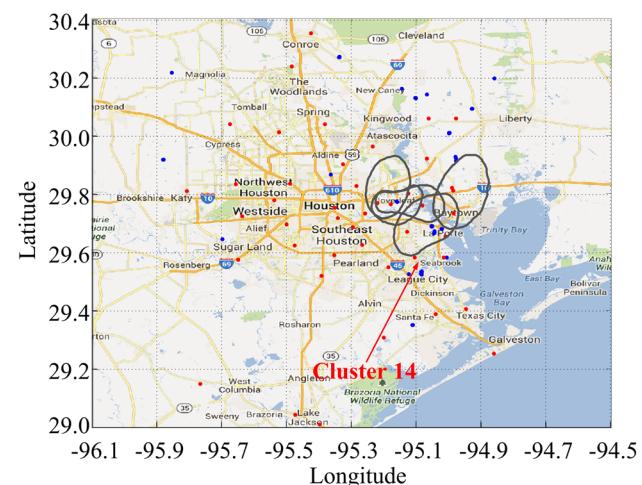
As expected, ST-SNN is able to find clusters of polygons that are very similar in space and time; for example, ST-SNN could successfully identify 17 clusters of polygons that are very close spatially and occur at the exact same time on different dates. We also identify the locations of the emission sources and monitor stations represented by the blue points and red points, respectively, on each figure. Clusters 9 and 11 lie closely in space, but are identified as two different clusters due to different temporal similarities, i.e., all five polygons in cluster 9 occurred at 1 pm, whereas all four polygons in cluster 11 occurred at 3 pm. Clusters 14 and 16 are also detected due to the high spatial and temporal similarities. All five polygons in cluster 14 occurred at 2 pm along highway interstate 10 east. All six polygons in cluster 16 occurred at 3



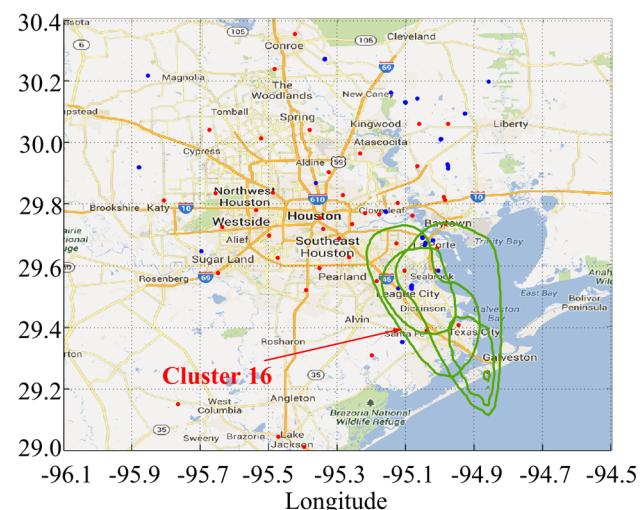
**Fig. 4** Visualization of cluster 9 identified by ST-SNN



**Fig. 5** Visualization of cluster 11 identified by ST-SNN

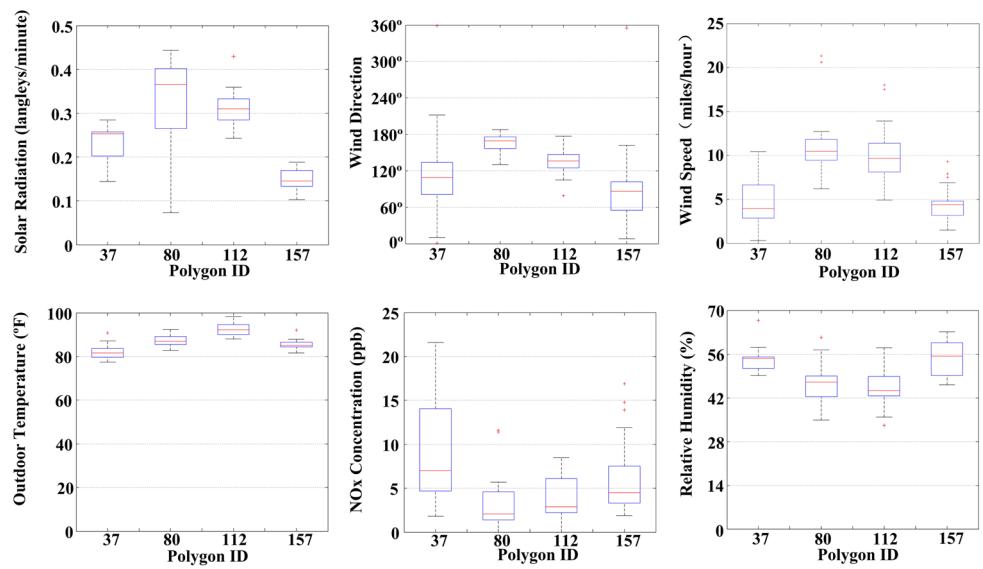


**Fig. 6** Visualization of cluster 14 identified by ST-SNN



**Fig. 7** Visualization of cluster 16 identified by ST-SNN

**Fig. 8** Box plots for cluster 11 identified by ST-SNN



pm along highway interstate 45 north. Clusters 6 and 14 are formed due to the emissions from highway traffic vehicles and strong solar radiation which usually happen between 2 pm and 3 pm each day.

Further, we closely inspect cluster 11. Polygons 37, 157, 112, and 80 in cluster 11 are in the ascending order according to their impact scopes. Their extending directions are to the east. The box plots shown in Fig. 8 provide the statistical comparisons of  $NO_x$  concentration, solar radiation, outdoor temperature, wind speed, wind direction, and relative humidity of these polygons, respectively. It can be observed that the majority of wind directions associated with these polygons distribute in the range between  $60^\circ$  and  $180^\circ$ , which means that the wind comes from the Gulf of Mexico as the result of a typical sea breeze encompassing east-southeast to south on the shore winds in the HGB area [32]. The wind transports the emissions from both point sources and highway traffic vehicles that locate at the upwind direction to this area, creating a suitable condition for ozone formation in this area. Polygon 80 has relatively stronger solar radiation which facilitates the ozone formation and higher wind speed which speeds up the emission transportation. Therefore, polygon 80 has the largest impact scope in cluster 11. Polygon 37 has the lowest wind speed, the largest distribution of wind directions, and low solar radiation, which cause polygon 37 to have the smallest impact scope in cluster 11, even though it has higher  $NO_x$  concentrations compared with other polygons in cluster 11. All the information suggests that the wind speed, wind direction, and solar radiation may be major contributors for the impact scope of ozone pollutant events in this area, while  $NO_x$  concentration, relative humidity, and outdoor temperature may not. It also demonstrates that the land and sea breeze can affect ozone formation by transferring the emissions from the emission sources.

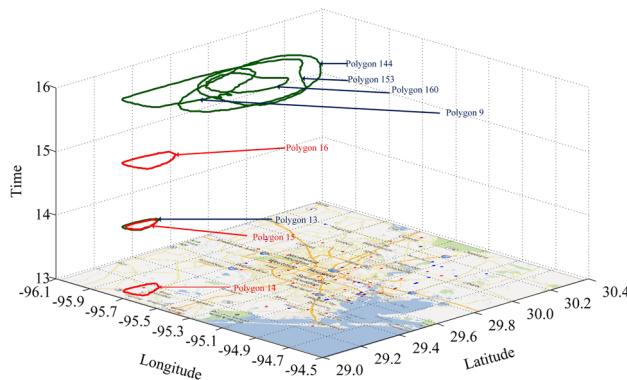
If the value of  $k$  is large, such as 8, or 10, ST-SNN finds fewer clusters. If we increase the weight associated with spatial distance  $w$ , clusters that are closer in space and less close in time will be identified, such as clusters of polygons that occur within several hour intervals but lying very closely in space. If we use  $w$  equal to 1.0, ST-SNN becomes a spatial clustering algorithm. We can also adopt different temporal distance functions, for example, in order for ST-SNN to find groups of polygons that lie closely in space and occur within certain time intervals instead of a particular time instant; the following temporal distance function could be adopted:

$$dist_t(p, q) = \begin{cases} 1, & abs(h(p) - h(q)) < t \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

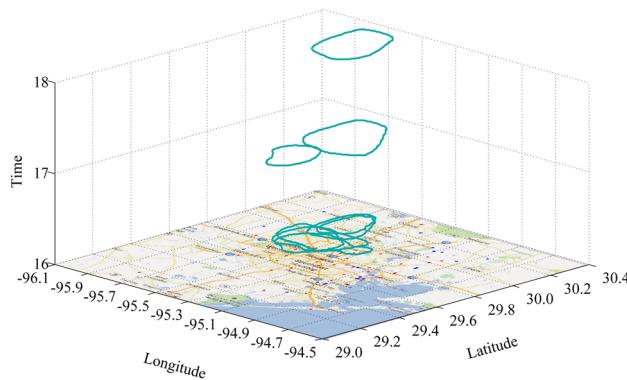
where  $p$  and  $q$  are two polygons,  $h(p)$  returns the value of hour information associated with polygon  $p$ , function  $abs()$  return the absolute value, and  $t$  is the time interval threshold input by the users.

#### 6.2.2 ST-SEP-SNN clustering and analysis

In this case study, we apply ST-SEP-SNN and the same temporal distance function for the purpose of comparison. The input parameters for ST-SEP-SNN are  $k = 5$  and  $MinPs = 3$ . There are 12 clusters identified by ST-SEP-SNN. Six of them contain polygons that occurred at the same time on different dates at similar locations, which are very similar to the clusters identified by ST-SNN. However, the other six clusters contain polygons that occurred within certain time intervals, such as 3 or 4 h, instead of a particular time. Unlike ST-SNN, ST-SEP-SNN intends to group polygons that lie at similar locations but occur within certain



**Fig. 9** Visualization of cluster 4 identified by ST-SEP-SNN



**Fig. 10** Visualization of cluster 9 identified by ST-SEP-SNN

time intervals into one large cluster instead of dividing them into two or more smaller compact clusters. This is because ST-SNN uses the weighted sum of the spatial distance and temporal distance to compute the spatio-temporal distance between a pair of polygons; then, it ranks the spatio-temporal distance matrix to get the  $k$ -nearest neighbor list for each polygon, whereas the ST-SEP-SNN first ranks the spatial distance matrix and temporal distance matrix separately; next, it computes the nearest neighbor list using the intersection of the  $k$ -nearest spatial neighbor list and the  $k$ -nearest temporal neighbor list. Thus, the  $k$ -nearest temporal neighbor list may include some polygons that have larger temporal distances if they were ranked in the top  $k$ , which may be excluded in the  $k$ -nearest neighbor list by ST-SNN due to their large temporal distances. Figures 9 and 10 visualize two such clusters, i.e., clusters 4 and 9.

Cluster 4 is an interesting cluster. It includes eight polygons, i.e., polygons 9, 13, 14, 15, 16, 144, 153, and 160. Polygons 14, 15 and 16, shown in red in Fig. 9, occurred in three continuous hours, i.e., 1 pm, 2 pm, and 3 pm on the same day. They are in the ascending order according to the area each polygon covered, and the scope extending direction is to the northeast. The box plots in Fig. 11 provide the statistical comparisons of  $NO_x$  concentration and

related meteorological data. Based on Fig. 11, from 1 pm to 3 pm, the mean of the wind direction rotated clockwise and the wind speed increased gradually, which enhanced the dispersion effect. Therefore, the areas covered by polygons 14, 15, and 16 were enlarged. However, these enhancement factors were counterbalanced by the reduction in the solar radiation from 1 pm to 3 pm, which resulted in the slow change of the areas covered by polygons 14, 15, and 16 over time. Compared with other polygons, these three polygons have large variances of solar radiation, wind direction, and wind speed, and lower mean values of relative humidity, while their  $NO_x$  concentrations and outdoor temperatures are quite similar to those of the other polygons. Further investigations of such clusters may help domain experts detect dynamic evolution of the ozone pollution events in this area.

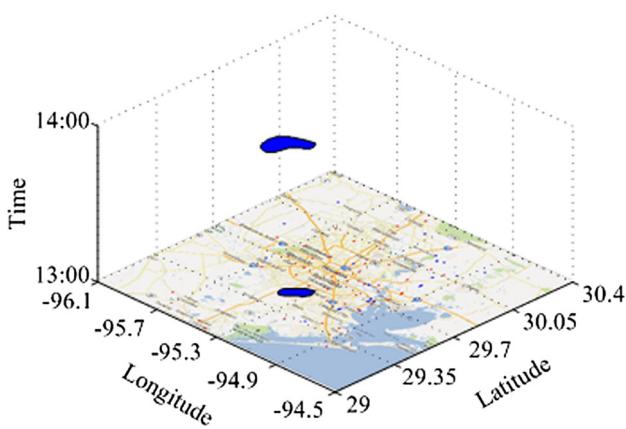
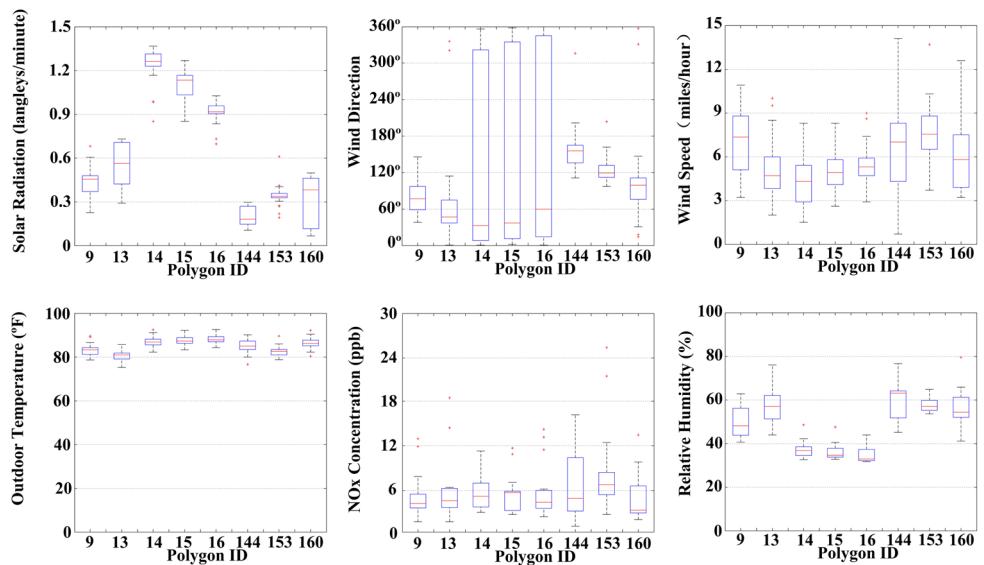
### 6.3 Post-processing analysis technique evaluation

The goal of our post-processing analysis technique is to help domain experts identify interesting clusters generated by spatio-temporal clustering algorithms that are unusual compared to other clusters. We apply the post-processing analysis techniques discussed in Sect. 5.1 to identify those two clusters. We considered Ozone Formation Potential Index (OFPI), Ozone Dispersion Index (ODI), and Interestingness Scores (IS) for the purpose of post-processing analysis. Figures 12 and 13 visualize two such clusters identified by our post-processing analysis technique, i.e., cluster 23, and 26, respectively. It can be observed that cluster 23 has a very small impact scope within the 2 h duration, whereas cluster 26 has relatively larger impact scopes within the 3 h duration. Table 1 summarizes the corresponding results of Ozone Formation Potential Index (OFPI), Ozone Dispersion Index (ODI), and Interestingness Scores (IS) for clusters 23 and 26. The deviation degrees of all attributes in clusters 23 and 26 are summarized in Table 2. The OFPI value of cluster 23 is 0.13 because cluster 23 has relatively low values of the solar radiation (deviation degree 0.39) and low  $NO_x$  concentration (deviation degree -1) compared with the entire dataset. Furthermore, cluster 23 has the lowest ODI value (1.00) among all clusters because it has the lowest wind speed and the smallest range of wind direction as revealed by deviation degrees are -1 for both attributes shown in Table 2. On the contrary, cluster 26 has a larger value of OFPI (0.70) and ODI (1.78) due to the high values of solar radiation (deviation degree 1) and high values of temperatures (deviation degree 1). Therefore, cluster 26 has a larger IS value (1.25).

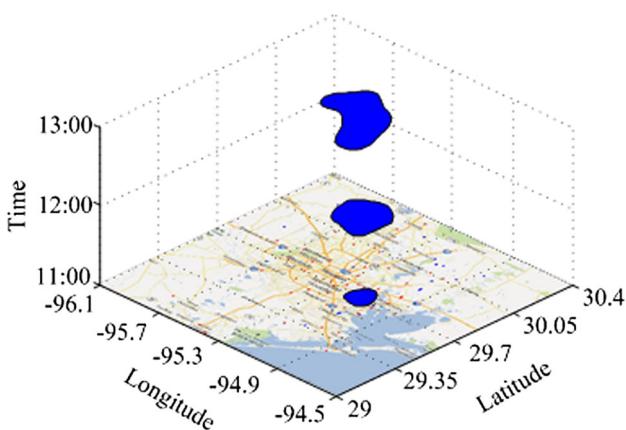
### 6.4 Change pattern discovery evaluation

In this case study, we use Hybrid distance [15] to compute the spatial distance between polygons. The spatial distance

**Fig. 11** Box plots for cluster 4 identified by ST-SEP-SNN



**Fig. 12** Visualization of cluster 23



**Fig. 13** Visualization of cluster 26

threshold  $\theta_s$  is set to the average distance of the data; temporal distance threshold  $\theta_t$  is set to 24 h because domain experts are interested in daily ozone pollution patterns. Note that

**Table 1** The interesting scores of clusters 23 and 26

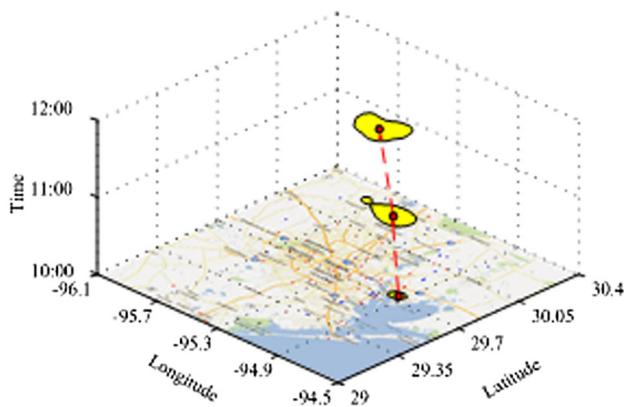
Cluster ID	OFPI	ODI	IS
23	0.13	1.00	0.13
26	0.70	1.78	1.25

**Table 2** The deviation degrees of clusters 23 and 26

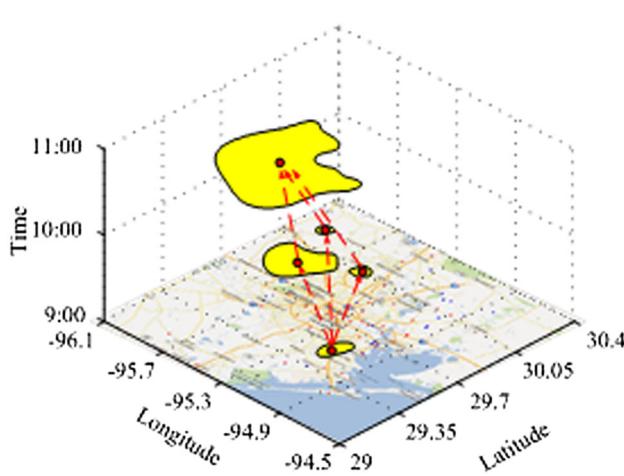
Cluster no.	Temperature	Wind direction	Wind speed	NO <sub>x</sub>	Solar radiation
23	1.00	-1.00	-1.00	-1.00	0.39
26	1.00	-0.19	-0.58	0.11	1.00

different temporal distance functions and thresholds can be adopted for different analysis tasks. We use a relatively small number as core polygon threshold  $MinPs$ , i.e., 2. There are 45 clusters identified by ST-SNN. Figures 14, 15 and 16 visualize three clusters, i.e., cluster 10, 12, and 35, respectively. The centroid of each polygon is marked as a red dot. As shown in Fig. 15, there are three ozone pollution hot-spots occurred at 10:00 am at different locations in HGB area. It is possible that multiple ozone pollution events occurred concurrently at different locations in HGB area. Therefore, there are three centroids at the same time, i.e., 10:00 am.

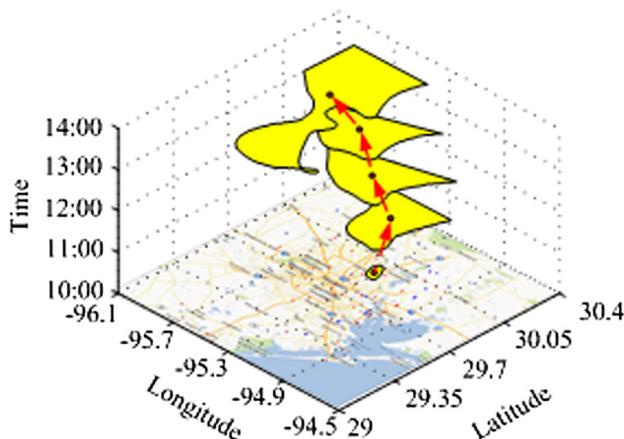
The dynamic change patterns of daily ozone pollution events in a specific region are very complex. The main concern is the expansion and duration of the impact region during ozone pollution events. A long duration time means that the corresponding ozone pollution event is very serious and may be caused by abnormal emissions from local industrial plants. For example, the duration of cluster 35 is 5 h, and the impact regions of the ozone pollution events in cluster 35 have been



**Fig. 14** Visualization of cluster 10

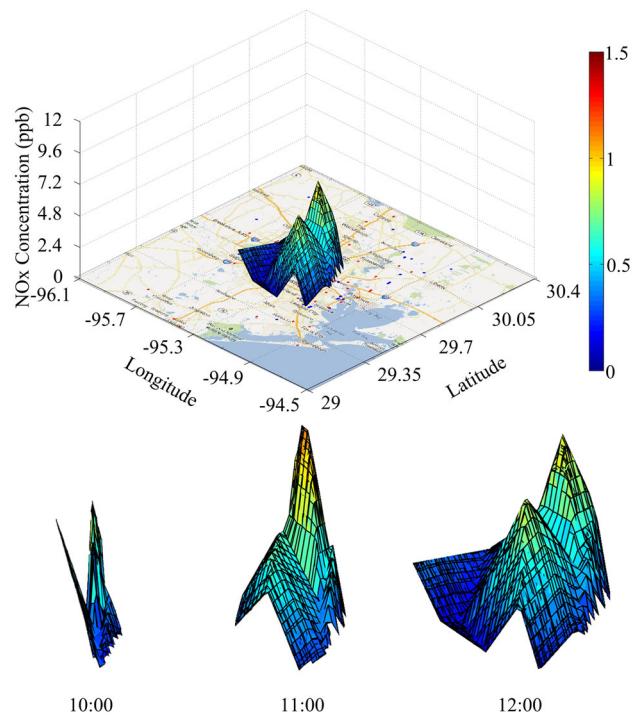


**Fig. 15** Visualization of cluster 12

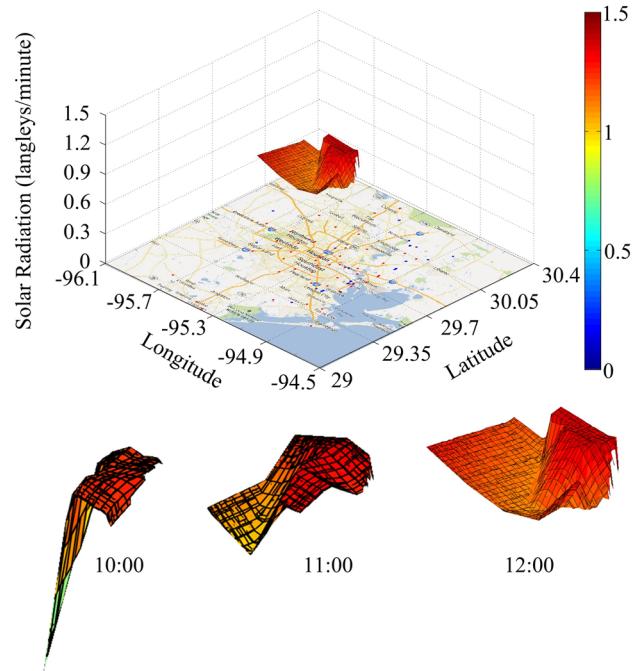


**Fig. 16** Visualization of cluster 35

largely expanded because the wind is not fast enough to effectively transport the ozone pollution to the downwind direction. The change pattern for cluster 35 is formation → expansion → expansion → expansion → expansion → disappear.

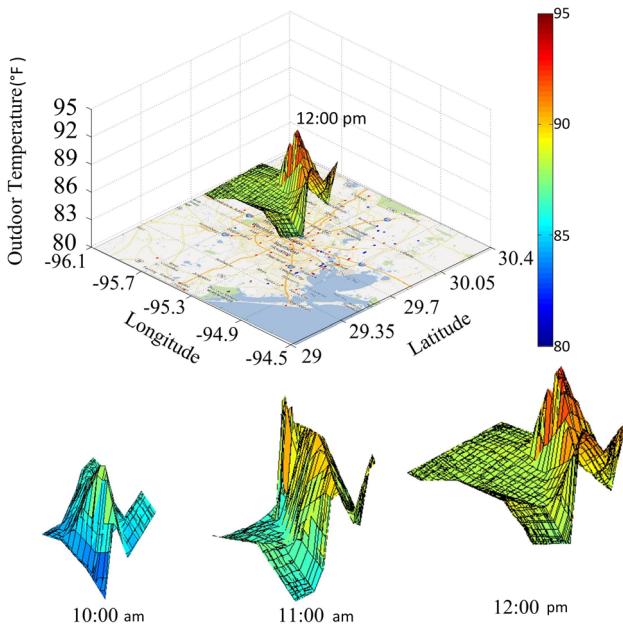


**Fig. 17** Dynamic changes of 3D NO<sub>x</sub> concentration profiles for cluster 10



**Fig. 18** Dynamic changes of 3D solar radiation profiles for cluster 10

Further, we closely inspect cluster 10, which is a typical change pattern of daily ozone pollution events. The dynamic profiles of  $NO_x$  concentration, solar radiation, and outdoor temperature for these three polygons in cluster 10 are shown in Figs. 17, 18 and 19, respectively. Figure 17 displays the



**Fig. 19** Dynamic changes of 3D outdoor temperature profiles for cluster 10

dynamic changes of 3D  $NO_x$  concentration profiles disclosed from 10 am to 12 pm. The peak  $NO_x$  concentration occurred at 11 am. For solar radiation shown in Fig. 18, the maximum value occurred at 12 pm. Note that an ozone pollution event was started at 10 am due to the impact of the high  $NO_x$  concentration, solar radiation, and outdoor temperature; in the following continuous 2 h, the pollution hot-spots were expanded because the major impacts were enhanced; meanwhile, the ozone pollution hot-spots were continuously moving toward the northwest direction partially due to the impact of the wind flow.

## 7 Conclusion and future work

The main goal of our research is to develop novel spatio-temporal clustering and analyze framework for spatio-temporal data. Polygons are very useful as they provide a natural representation for particular types of spatial objects and provide a useful tool to analyze discrepancies, progression, change, and emergent events. Two density-based spatio-temporal clustering algorithms, called ST-SNN and ST-SEP-SNN, are developed by extending the generic shared nearest neighbor clustering algorithm. We redefine the nearest spatio-temporal neighborhood of a polygon and the density-based concepts for polygons. Both ST-SNN and ST-SEP-SNN can find clusters of varying shapes, sizes, and densities in high-dimensional data, even in the presence of outliers. We also propose a change pattern discovery algorithm to atomically detect and analyze dynamic changes within

spatio-temporal clusters of polygons and a post-processing analysis technique to identify interesting spatio-temporal clusters of polygons for domain experts. Experiments on spatio-temporal data involving ozone pollution events in the HGB area demonstrate that our methodology is effective and can discover interesting spatio-temporal patterns and change patterns of ozone pollution events. Moreover, statistic and post-processing analysis techniques can help domain experts to identify interesting patterns, to learn from the past, and be better prepared for the future.

In terms of future work, we plan to design and develop efficient algorithmic solutions and computing infrastructures for integrating large-scale spatio-temporal analysis with modern computing frameworks (e.g., cluster), and taking advantage of existing the powerful computational resources such as Hadoop-GIS to support efficient big spatio-temporal data analysis.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Han, J., Kamber, M., Tung, A.: Spatial Clustering Methods in Data Mining: A Survey, Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS. Taylor and Francis, Abingdon (2001)
2. Chawla, S., Shekhar, S., Wu, W., Ozesmi, U.: Modeling spatial dependencies for mining geospatial data. In: Proceedings of the 2001 SIAM International Conference on Data Mining (2001)
3. Ertoz, L., Steinback, M., Kumar, V.: Finding clusters of different sizes, shapes, and density in noisy high dimensional data. In: Proceedings of the 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA, May (2003)
4. Kulldorff, M.: A spatial scan statistic. Commun. Stat. Theory Methods **26**, 1481–1496 (1997)
5. Iyengar, S.: On detecting space-time clusters. In: Proceedings of the 10th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August (2004)
6. Wang, M., Wang, A., Li, A.: Mining spatial-temporal clusters from geodatabases. Lect. Notes Comput. Sci. **4093**, 263–270 (2006)
7. Birant, D., Kut, A.: ST-DBSCAN: an algorithm for clustering spatial-temporal data. Data Knowl. Eng. **60**, 208–221 (2007)
8. Kisilevich, S., Mansmann, F., Rinzivillo, S., Nanni, M.: Spatio-temporal clustering: a survey. In: Data Mining and knowledge Discovery Handbook, pp. 269–298 (2010)
9. Gaffney, S., Smyth, P.: Trajectory clustering with mixtures of regression models. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Minin, San Diego, CA, USA, August (1999)
10. Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G., Theodoridis, Y.: Similarity search in trajectory databases. In: Proceedings of the 14th International Symposium on Temporal Representation and Reasoning, Alicante, Spain, June (2007)
11. Nanni, M., Pedreschi, D.: Time-focused clustering of trajectories of moving objects. J. Intell. Inf. Syst. **27**, 267–289 (2006)

12. Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., Andrienko, G.: Visually driven analysis of movement data by progressive clustering. *Inf. Vis.* **7**, 225–239 (2008)
13. Li, Y., Han, J., Yang, J.: Clustering moving objects. In: Proceedings of the 10th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August (2004)
14. Joshi, D., Samal, A., Soh, L.: Spatio-temporal polygonal clustering with space and time as first-class citizens. *GeoInformatica* **17**, 387–412 (2013). doi:[10.1007/s10707-012-0157-8](https://doi.org/10.1007/s10707-012-0157-8)
15. Wang, S., Eick, C.: A polygon-based clustering and analysis framework for mining spatial datasets. *GeoInformatica* **3**, 569–594 (2014). doi:[10.1007/s10707-013-0190-2](https://doi.org/10.1007/s10707-013-0190-2)
16. Li, Z., Ding, B., Han, J., Kays, R.: Swarm: Mining relaxed temporal moving object clusters. In: PVLDB, vol. 3, pp. 723–734 (2010)
17. Benkert, M., Gudmundsson, J., Hubner, F., Wolle, T.: Reporting flock patterns. In: COMGEO (2008)
18. Gudmundsson, J., van Krevel M.: Computing longest duration flocks in trajectory data. In: GIS (2006)
19. Jeung, H., Yiu, M.L., Zhou, X., Jensen C.S., Shen, H.T.: Discovery of convoys in trajectory databases. In: PVLDB (2008)
20. Chen, L., Ozsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: SIGMOD (2005)
21. Vlachos, M., Gunopulos, D., Kollios, G.: Discovering similar multidimensional trajectories. In: ICDE (2002)
22. MaIntosh, J., Yuan, M.: A framework to enhance semantic flexibility for analysis of distributed phenomena. *Int. J. Geogr. Inf. Sci.* **19**, 999–1018 (2005)
23. Rinsuengkawong, V., Chen, C.-S., Eick, C.F., Twa, M.: Analyzing change in spatial data by utilizing polygon models. In: Proceedings of International Conference on Computing for Geospatial Research and Application, Washington DC, USA, June (2010)
24. Stell, J., Mondo, G.D., Thibaud, R., Claramunt, C.: Spatio-temporal evolution as bigraph dynamics. In: COSIT 2011: Spatial Information Theory, pp. 148–167 (2011)
25. Texas commission on environmental quality. <http://www.tceq.state.tx.us>. Accessed May 2011
26. Chen, C., Rinsuengkawong, V., Eick, C., Twa, M.: Change analysis in spatial data by combining contouring algorithms with supervised density functions. In: Proceedings of the 13th Asia-Pacific Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand, April (2009)
27. Hangouet, J.: Computing of the hausdorff distance between plane vector polylines. In: Proceedings of the 8th International Symposium on Computer-Assisted Cartography, Charlotte, North Carolina, USA, February (1995)
28. Buchin, K., Buchin, M., C. W.: Computing the frchet distance between simple polygons in polynomial time. In: Proceedings of the 22nd ACM Symposium on Computational Geometry, Sedona, Arizona, USA, June (2006)
29. Joshi, D., Samal, A., Soh, L.: A dissimilarity function for clustering geospatial polygons. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS), Seattle, Washington, USA, November (2009)
30. Wang, S., Chen, C., Rinsuengkawong, V., Akdag, F., Eick, C.: A polygon-based methodology for mining related spatial datasets. In: Proceedings of the 18th ACM SIGSPATIAL Conference on Advances in Geographic Information Systems Workshop on Data Mining for Geoinformatics (DMGI), San Jose, CA, USA, November (2010)
31. Atallah, M., Ribeiro, C., Lifschitz, S.: Computing some distance functions between polygons. *Pattern Recogn.* **24**(8), 775–781 (1991)
32. Lu, R., Turco, R.: Air pollutant transport in a coastal environment. part i: Two-dimensional simulations of sea-breeze and mountain effects. *J. Atmos. Sci.* **51**, 2285–2308 (1994)