

# Analyzing Change in Spatial Data by Utilizing Polygon Models

Vadeerat Rinsurongkawong<sup>1</sup>, Chun Sheng Chen<sup>1</sup>, Christoph F. Eick<sup>1</sup>, and Michael D. Twa<sup>2</sup>

<sup>1</sup>Department of Computer Science,  
University of Houston  
Houston TX, 77204-3010  
{vadeerat, lyon19, ceick}@cs.uh.edu

<sup>2</sup>College of Optometry,  
University of Houston  
Houston TX, 77204-6052  
mtwa@optometry.uh.edu

## ABSTRACT

Analyzing change in spatial data is critical for many applications including developing early warning systems that monitor environmental conditions, epidemiology, crime monitoring, and automatic surveillance. In this paper, we present a framework for the detection and analysis of patterns of change; the framework analyzes change by comparing sets of polygons. A contour clustering algorithm is utilized to obtain polygon models from spatial datasets. A set of change predicates is introduced to analyze changes between different models which capture various types of changes, such as novel concepts, concept drift, and concept disappearance. We evaluate our framework in case studies that center on ozone pollution monitoring, and on diagnosing glaucoma from visual field analysis.

## Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *data mining, spatial databases and GIS.*

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *clustering.*

I.2.1 [Artificial Intelligence]: Application and Expert Systems – *Medicine and science.*

## General Terms

Algorithms, Measurement, Experimentation

## Keywords

Change Analysis, Polygon Models, Density-based Clustering, Concept Drift, Novelty Detection, Spatial Data Mining.

## 1. INTRODUCTION

With advances in data acquisition technologies, huge amount of spatial data are collected every day. Detecting and analyzing changes in such data is of particular importance and has many applications including developing early warning systems that monitor environmental or weather conditions, epidemiology, crime monitoring, automatic surveillance, emergency first responders' coordination for natural disasters. Addressing this need, we introduce a framework that includes methodologies and

tools to discover and analyze change patterns in spatial data. Change patterns capture how the most recent data differ from the data model established from the historical data.

The main challenges for developing systems that automatically detect and analyze change in spatial datasets include:

1. The development of a formal framework that characterizes different types of change patterns
2. The development of a methodology to detect change patterns in spatial datasets
3. The capability to find change patterns in regions of arbitrary shape and granularity
4. The development of scalable change pattern discovery algorithms that are able to cope with large data sizes and large numbers of patterns
5. Support for different perspectives with respect to which change is analyzed

In this paper, a framework for the detection and analysis of patterns of change is proposed. Our methodologies use spatial clustering and change analysis algorithms that operate on polygon models. In particular, clusters discovered by contour clustering algorithms serve as the models for the currently observed spatial data. These models have to be updated to reflect changes as new data arrive. Change analysis algorithms are responsible for detecting the change patterns of interest that capture discrepancies between new data and the current model. Finally, change reports are generated that describe newly emerging clusters, disappearing old clusters, and movement of the existing clusters.

Our contributions include the following:

1. The development of tools based upon spatial clustering and polygon operations to detect and analyze change patterns in spatial data
2. A definition of generalized change predicates that are utilized to detect and analyze a variety of specific change patterns of interest including concept drift and the emergence of novel concepts
3. A demonstration of these methods in two very different applications—ozone pollution monitoring, and the diagnosis of glaucoma progression from visual field analysis

The paper is organized as follows. Section 2 is an overview of our framework. Section 3 presents change analysis methodologies and tools. In section 4, our framework is evaluated in case studies.

Related work is discussed in section 5. Section 6 is our conclusion and future work.

## 2. OVERVIEW

In this paper, we address a problem of change analysis in spatial data; a framework that detects and analyzes the patterns of change is proposed. We introduce a change analysis method that operates on cluster models generated by density and interpolation functions and a contour clustering algorithm. A set of change predicates is introduced that is capable of defining change of concepts in spatial data such as concept drift, novel concept, concept disappearance, and local rearrangement.

A proposed change analysis is conducted as follows. Assume that a set of spatial datasets at different time frames,  $O_1, O_2, \dots, O_n$  are given, and we are interested in finding what patterns emerged in  $O_t$  compared to  $O_1, O_2, \dots, O_{t-1}$ . First, we generate a density map on spatial data in dataset  $O_t$  then a set of contour clusters  $\{c_1, c_2, \dots, c_n\}$  that represent interesting regions is created by running a contour clustering algorithm on the density map. Next, we create a cluster model  $M_t$  which is a set of polygons  $\{m_1, m_2, \dots, m_n\}$  that is generalized from dataset  $O_{t-1}, O_{t-2}, \dots, O_1$ . Finally, a knowledge base of change predicates is provided that allows analyzing various aspects of change in interesting regions based on changes in the set of contour clusters  $\{c_1, c_2, \dots, c_n\}$  from  $O_t$  with respect to a set of contour clusters  $\{m_1, m_2, \dots, m_n\}$  in the base polygon model  $M_t$ . The change analysis processes are repeated for dataset  $O_{t+1}$ .

### 2.1 System Architecture

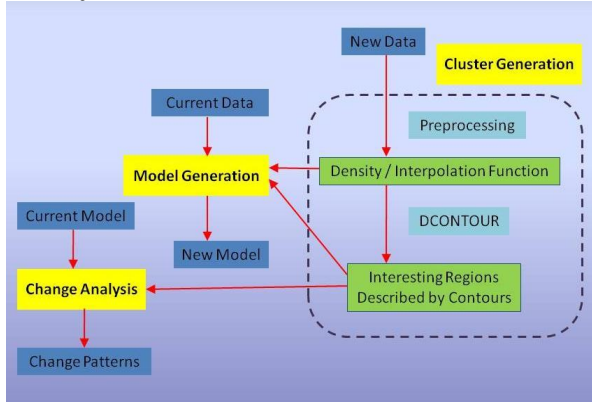


Figure 2-1. System architecture of the proposed framework

In this section, the system architecture is discussed. The overview of the system is illustrated in Figure 2-1. The system consists of three main modules: Cluster Generation, Model Generation, and Change Analysis.

The Cluster Generation module employs density or interpolation functions and a contour clustering algorithm named DCONTOUR [3] to generate interesting regions from the new arriving data. (We discuss how DCONTOUR works in section 2.2.) The interesting regions resulted from the Cluster Generation module are presented as contour clusters. For some problems, generating density or interpolation functions from attributes of interest may be non-trivial. For example, in analyzing social network problems, domain experts may desire to find areas where people with high and low income live in close proximity to each other. In such case, a region discovery framework introduced in [6] can be

applied as a preprocessing step for discovering the regions which have high variance in income. Then the regions thus found can be inputs for the Cluster Generation module.

Next, the Model Generation module acquires the density or interpolation functions and cluster contours generated by the Cluster Generation module to create a new model.

Finally, the Change Analysis module analyzes change by applying polygon operations to the current model. A set of change predicates is introduced that identifies different types change, such as concept drift, novel concept, concept disappearance, and local rearrangement. Change reports that summarize such changes in the current data are automatically generated.

### 2.2 DCONTOUR

DCONTOUR [3] is a clustering algorithm that uses contouring algorithms to obtain spatial clusters and employs polygons as cluster models. Its input is a density function or interpolation function  $\psi^o$  that has been derived from a spatial dataset  $O$  and a threshold  $d$ . Its output is a set of polygons describing contiguous regions containing points  $v$  for which  $\psi^o(v) \geq d$ . Figure 2-2 gives a pseudo-code of the algorithm.

**Input:** Function  $\psi^o$ , threshold  $d$ .

**Output:** Polygons for threshold  $d$ .

1. Subdivide the space into  $D$  grid cells.
2. Compute  $\psi^o(g)$  for all grid intersection points  $g$ .
3. Compute contour intersection points  $b$  on grid cell edges where  $\psi^o(b) = d$  using binary search and interpolation.
4. Compute contour polygons from contour intersection points  $b$ .

Figure 2-2. Pseudo-code of the DCONTOUR algorithm

First, a grid structure for the space is created and then  $\psi^o(g)$  for each grid intersection point  $g$  is computed. Next, the contour intersection points are calculated using binary search and interpolation. Finally, contour intersection points found on cell edges are connected. The process is continued until a closed polygon is formed. DCONTOUR uses an algorithm proposed by Snyder [12] to derive contours from intersection points in the final step.

### 2.3 Polygon Models

In this paper, polygons are used as cluster models to discover change patterns. Basically, cluster boundaries in spatial space are represented by polygons. For prototype-based algorithms such as k-means, clusters can be modeled as Voronoi cells generated from a set of cluster representatives. In our work, cluster models for different time frames are generated by DCONTOUR as sets of polygons. Change analysis is then performed by comparing polygons by using basic polygon operations, such as union, intersection, area (or size) of polygons, and by analyzing distances between polygons.

### 2.4 Model Generation

To detect the changes of patterns in datasets at different time frames, we rely on cluster analysis to capture the arrival of new concepts, the disappearance of old concept, and concept drift. In this section, we describe how polygon models are generated and updated to reflect changes in current data. It is important for the polygon model to be responsive to relevant changes.

Assume dataset  $O_t$  is the dataset of time frame  $t$  ( $1 \leq t \leq T$ ) and polygon models are a set of polygons  $M_t = \{m_1, \dots, m_n\}$ . The initial polygon model  $M_1$  is generated by using the DCONTOUR clustering algorithm on dataset  $O_1$ . Each subsequent polygon model  $M_{t+1}$  is obtained from data in dataset  $O_{t+1}, O_t, \dots, O_1$ . It is important to note that each model captures the patterns of data in a series of time frames rather than from a snapshot of a single timeframe. Therefore, the model generating strategy needs to consider the impact of previous observations.

The goal of our proposed change analysis approach is to detect new patterns in the dataset  $O_{t+1}$  in the context of what happened in the past as captured in model  $M_t$  (based on datasets  $O_1, O_2, \dots, O_t$ ). One approach directly derives  $M_t$  from  $X = O_1 \cup \dots \cup O_t$ . However, this is not a feasible approach because of the number of data objects accumulated along the time will eventually become the performance bottleneck for running a clustering algorithm. Two possible strategies to deal with this issue include limiting the size of  $X$  by the use of random deletions, or using a sliding window approach that considers only the most recent  $k$  datasets. Additionally, data could be inversely weighted by age. This paper employs an un-weighted sliding window approach for model generation.  $M_t$  is derived by computing contour polygons from the instances of the last  $k$  datasets, where  $k$  is an input parameter.

### 3. A TOOL FOR CHANGE ANALYSIS

Change Analysis is conducted in the following steps:

1. A user selects a set of change predicates with respect to the changes to be analyzed.
2. The user selects parameter settings for the change predicates.
3. The parameterized change predicates are matched against a model  $M$  and a new dataset  $O$  to obtain sets, pairs, triples, ... of clusters that match the change predicates. We call the objects that match a change predicate instantiations of the change predicate.
4. Change reports are generated from the instantiations of the selected change predicates.

This section introduces basic change predicates to analyze changes between clusters in the dataset  $O$  and the model  $M$ .

Let  $c, c_1, c_2$ , be clusters in  $O$  and  $X_O$  the set of all clusters in  $O$ ; let  $m, m_1, m_2$  be a clusters in  $M$  and  $X_M$  the set all clusters in  $M$ . The operators ' $\cap$ ' and ' $\cup$ ' denote polygon intersection and union;  $|c|$  computes the size (area) of a polygon  $c$ . In this case, agreement between  $c$  and  $m$  can be computed as follows:

$$Agreement(c, m) = \frac{|c \cap m|}{|c \cup m|}$$

Agreement measures how similar two polygons  $c$  and  $m$  are. In addition to agreement, containment between two clusters is defined as follows:

$$Containment(c, m) = \frac{|c \cap m|}{|c|}$$

Basically, containment measures the degree to which a cluster  $c$  is contained in another cluster  $m$ .

Many change predicates involve distances between clusters. There are many different ways to measure distances between clusters. Our current work uses Average Link or Group Average [15] as a

metric to measure the distance between two clusters  $c$  and  $m$ . Average Link is defined as follows:

$$Distance(c, m) = \frac{1}{|c||m|} \sum_{o \in c} \sum_{v \in m} distance(o, v)$$

where  $o$  is an object belonging to cluster  $c$  and  $v$  denotes an object that lies in the scope of  $m$ .

Agreement, containment and average link distance are utilized to define more complex change predicates; below we list several popular change predicates:

- 1) Stable Concept  $(c, m) \leftrightarrow Agreement(c, m) \geq \text{very high} \wedge Distance(c, m) \leq distance_{min}$
- 2) Concept Drift  $(c, m)$ 
  - 2.1) Moving  $(c, m) \leftrightarrow (distance_{min} < Distance(c, m) < distance_{max}) \wedge Agreement(shift(c, m), m) \geq \text{medium}$
  - 2.2) Growing  $(c, m) \leftrightarrow (Containment(c, m) < Containment(m, c)) \wedge Agreement(c, m) \approx Containment(c, m) \wedge (distance_{min} < Distance(c, m) < distance_{max})$
  - 2.3) Shrinking  $(c, m) \leftrightarrow (Containment(c, m) > Containment(m, c)) \wedge Agreement(c, m) \approx Containment(m, c) \wedge (distance_{min} < Distance(c, m) < distance_{max})$
- 3) Local Rearrangement  $(c_1, c_2, m_1, m_2)$ 
  - 3.1) Merging  $(c_1, m_1, m_2) \leftrightarrow (Containment(c_1, m_1) < Containment(m_1, c_1)) \wedge (Containment(c_1, m_2) < Containment(m_1, c_2)) \wedge (Distance(c_1, m_1) \leq distance_{max}) \wedge (Distance(c_1, m_2) \leq distance_{max}) \wedge (Containment(c_1, m_1) + Containment(c_1, m_2) \geq \text{medium})$
  - 3.2) Splitting  $(c_1, c_2, m_1) \leftrightarrow (Containment(c_1, m_1) > Containment(m_1, c_1)) \wedge (Containment(c_2, m_1) > Containment(m_1, c_2)) \wedge (Distance(c_1, m_1) \leq distance_{max}) \wedge (Distance(c_2, m_1) \leq distance_{max}) \wedge (Containment(m_1, c_1) + Containment(m_1, c_2) \geq \text{medium})$
- 4) Novel Concept  $(c) \leftrightarrow \forall m \in X_M (Agreement(c, m) \approx 0 \wedge Distance(c, m) \geq distance_{max})$
- 5) Disappearing Concept  $(m) \leftrightarrow \forall c \in X_O (Agreement(c, m) \approx 0 \wedge Distance(c, m) \geq distance_{max})$

In the above definitions, *very high*, *high*, *medium*, *low*,  $distance_{max}$  and  $distance_{min}$  are parameters whose values are selected based on application specific needs.  $x \approx y$  is a predicate that returns true if "x and y have approximately the same value".

In the above definition, concept drift captures cases when a cluster moves or changes in size. Local rearrangement occurs when a cluster splits into two or more clusters, or when two or more clusters are merged into a single cluster. In the case of concept drift with moving type,  $(Agreement(shift(c, m), m))$  measures the agreement of  $c$  and  $m$  assuming  $c$  is shifted back to the position of  $m$ .

In general, the polygon clusters for  $O$  and  $M$  are matched against all change predicates, and pairs of clusters are reported for each change predicate obtaining instantiated change predicates. For example, in the case of the "Stable Concept" change predicate, the following set of pairs of clusters is computed:

$$\{(c,m) \mid c \in X_O \wedge m \in X_M \wedge \text{Agreement}(c,m) \geq \text{very high} \wedge \text{distance}(c,m) \leq \text{distance}_{min}\}$$

Next, the pairs of clusters that match the change predicate, namely the instantiations of the ‘‘Stabile Concept’’ change predicate are sent to the report generator and a summary is generated. For example,  $\{(c1,m12), (c3,m17)\}$  indicates that  $c1$  corresponds to  $m12$ , and  $c3$  corresponds to  $m17$  and that the two pairs of clusters are highly stabile between  $X$  and  $M$ . In general,  $X_M$  and  $X_O$  are matched against all change predicates and the instantiations of each change predicate are reported as the results of change analysis.

Moreover, change predicates can be easily written as SQL queries, assuming the query language supports polygon operations. For example, one could use ORACLE SPATIAL as an implementation platform for change predicates. In general, a knowledge base of change predicates can be created for a particular application domain, and our change analysis framework can be directly applied to automate change analysis. For a different application domain, some change predicates may be different and some parameters in change predicates may have to be modified, but—most importantly—all other components of our change analysis framework can be reused.

## 4. CASE STUDIES

We evaluate our framework in two case studies centering on diagnosing glaucoma from visual field analysis and on ozone pollution monitoring in Houston, Texas.

### 4.1 Case Study in Diagnosing Glaucoma from Visual Field Analysis

The data used in this case study come from a longitudinal study of vision loss in glaucoma, a leading cause of blindness worldwide. One of the key tests that clinicians use to diagnose glaucoma and monitor its progression is a test of central and peripheral vision sensitivity. A typical visual field test evaluates a patient’s ability to detect faint spots of light at 54 discrete points distributed over the central 30-degrees of field of vision. At each point, higher value corresponds with better vision. In glaucoma, patients experience gradual and progressive damage to the optic nerve, and this corresponds with several characteristic patterns of visual field loss [9]. A total of 2,848 records from 232 patients comprise the dataset. The mean age was  $61 \pm 11$  years with an average of 6 yearly examinations. The visual field data were preprocessed by adjusting raw values to a common age corrected value [8]. A spatial map of the total deviation from the average age corrected value for each point were then classified into 6 levels: -9 dB (worse vision), -6 dB, -3 dB, 3 dB, 6 dB, and 9 dB (better vision), which are represented in visualizations by clusters in red, orange, yellow, green, blue-green, and green color respectively.

This case study represents a special case of our more general framework. In the cluster generation processes, an interpolation function is used instead of a density estimation function because the points in the visual field dataset are evenly distributed in a Cartesian grid structure. We have applied a bicubic spline interpolation function [10] to increase the number of sample points in the grid structure by a factor of 6.

The bicubic spline – an extension of cubic spline on two dimensional grid points – is defined as follows:

$$f(x, y) = \sum_{i=1}^4 \sum_{j=1}^4 c_{ij} t^{i-1} u^{j-1}$$

$$f_x(x, y) = \sum_{i=1}^4 \sum_{j=1}^4 (i-1)c_{ij} t^{i-2} u^{j-1}$$

$$f_y(x, y) = \sum_{i=1}^4 \sum_{j=1}^4 (j-1)c_{ij} t^{i-1} u^{j-2}$$

$$f_{xy} = \sum_{i=1}^4 \sum_{j=1}^4 (i-1)(j-1)c_{ij} t^{i-2} u^{j-2}$$

where  $c_{ij}$  are constant values, and  $0 \leq u, t \leq 1$ .

We have generated cluster models by using a sliding window technique with a window size 2. For data at time  $t$ , a model  $M_t$  is created by averaging data at time  $t-1$  and  $t-2$ . Parameter settings of change predicates are shown in Table 5-1a.

**Table 5-1a. Parameter settings of change predicates for visual field dataset.**

Agreement	Low=0.2	Medium=0.5	High=0.8
Containment	Low=0.2	Medium=0.5	High=0.8
Distance	Min=5	Max=15	

Figure 5-1a depicts the visual field of the right eye of Case 1. The images from top to bottom indicate progressive visual field loss due to glaucoma. On the 5<sup>th</sup> visit at -9 dB level (red), cluster id 0 and 1 in the model are merged and become data cluster id 0. Containment( $c0,m0$ ) is 0.56 and is less than Containment( $m0,c0$ ) which is 0.92. Containment( $c0,m1$ ) is 0.058 and is less than Containment( $m1,c0$ ) which is 0.82. Distance( $c0,m0$ ) and Distance( $c0,m1$ ) are 5.1 and 10.3 which are less than  $\text{distance}_{max}$ . On the 7<sup>th</sup> visit at the same level, data cluster id 1 is obviously a novel concept for this level. Its agreements with every red cluster in its associated model are all 0s and its distance to the closest cluster in the model is more than 23.7.

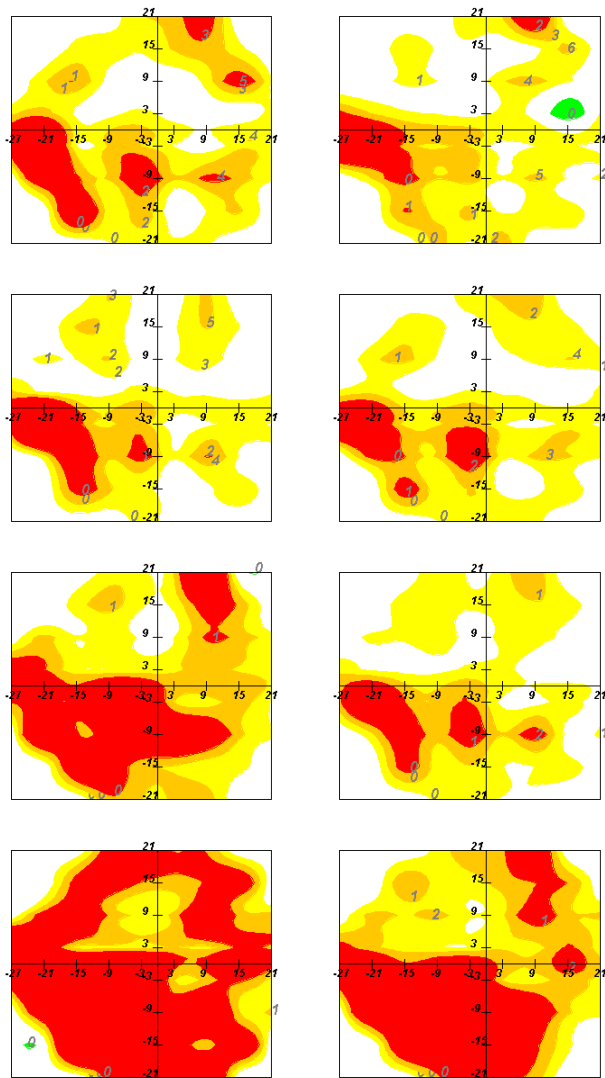
From Figure 5-1b. and Table 5-1b., high agreements and low distances of data cluster id 0 at -6 dB level (orange) on all visits indicate that it is a stabile concept with respect to its associated clusters. It should be noted that the models are kept on updating so the changes in the data are handled.

In Figure 5-1c., two green clusters with id 0 and 1 in the model of the 2<sup>nd</sup> visit are disappearing since their agreements with clusters of the 4<sup>th</sup> visit are zeroes. From Table 5-1c., data cluster 0 and 1 at -9 dB level (red) in the 4<sup>th</sup> visit are growing clusters as indicated by their Agreement( $c,m$ ) values which are close to the Containment( $c,m$ ) values. Moreover, their distances from their associated model clusters are less than  $\text{distance}_{min}$ , pointing towards concept drift. Their sizes also show that they are growing; however, without agreement, containment, and distance values, we cannot know which clusters they are associated with in the model.

In summary, the proposed approach for polygonal change analysis allows one to quantify changes and to associate the obtained statistics with disease progression. Consequently, our work provides a valuable tool and a methodology for the development of automatic, computerized tools for glaucoma disease staging.

**Table 5-1b. Change predicate values for cluster id 0 at -6 dB level (orange) of visual field test of Case 2.**

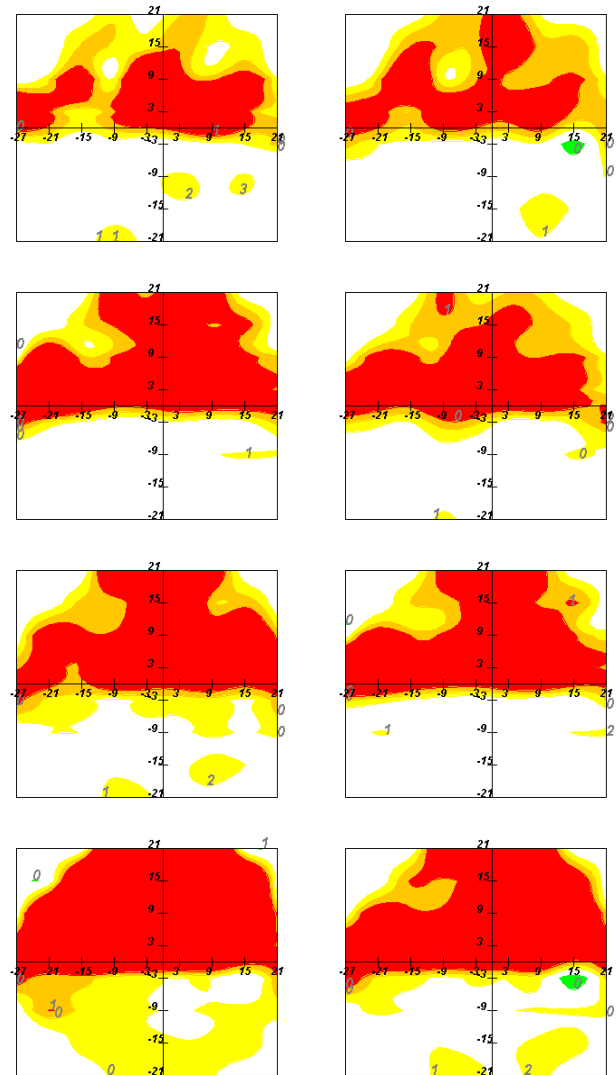
Visit No.	Size c0	Size m0	Agreement (c0,m0)	Containment		Distance (c0,m0)
				(c0,m0)	(m0,c0)	
3 <sup>rd</sup>	619	793	0.66	0.92	0.72	1.6
5 <sup>th</sup>	902	876	0.84	0.90	0.93	3.1
7 <sup>th</sup>	953	868	0.88	0.90	0.98	2.7
9 <sup>th</sup>	953	868	0.88	0.89	0.98	2.6



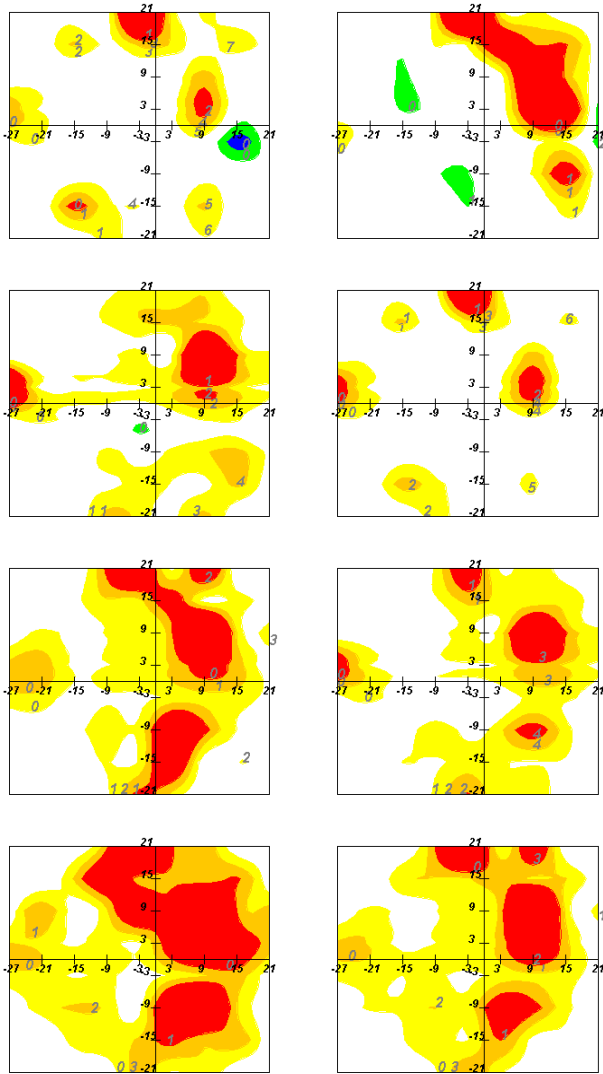
**Figure 5-1a. Visualizations of visual field test on right eye of Case 1; images on the left column are test results on the 3<sup>rd</sup>, 5<sup>th</sup>, 7<sup>th</sup>, and 9<sup>th</sup> visits respectively; image on the right column are images of the models for the data in the left column.**

**Table 5-1c. Change predicate values at -9 dB level (red) of visual field test of Case 3 in the 4<sup>th</sup> visit.**

Size c0	Size m0	Agreement (c0,m0)	Containment		Distance (c0,m0)
			(c0,m0)	(m0,c0)	
105	12	0.07	0.07	0.63	9.2
Size c1	Size m2	Agreement (c1,m2)	Containment		Distance (c1,m2)
84	23	0.14	0.16	0.57	5.6



**Figure 5-1b. Visualizations of visual field test on right eye of Case 2; images on the left column are test results on the 3<sup>rd</sup>, 5<sup>th</sup>, 7<sup>th</sup>, and 9<sup>th</sup> visit respectively; image on the right column are images of the models for the data in the left column.**



**Figure 5-1c. Visualizations of visual field test on the left eye of Case 3; images on the left column are test results on the 2<sup>nd</sup>, 4<sup>th</sup>, 6<sup>th</sup>, and 8<sup>th</sup> visit respectively; image on the right column are images of the models for the data in the left column.**

## 4.2 Case Study in Ozone Pollution Monitoring

The Texas Commission on Environmental Quality (TCEQ) is a state agency responsible for environmental issues including the monitoring of environmental pollution in the Texas. As seen on its website [16], the agency collects hourly ozone concentration data for metropolitan areas across the state. TCEQ uses a network of 27 ozone-monitoring stations in the Houston-Galveston area. The area covers the geographical region within [-95.8061, -94.8561] longitude and [29.0436, 30.3436] latitude. High ozone concentrations are normally observed in a day that has high UV radiation and low wind speed. TCEQ issues ozone pollution warnings once the 1-hour ozone concentration exceeds 75 parts per billion (ppb) based on the EPA’s ozone standard.

In this case study, we apply our methodology to analyze the progression of high ozone concentrations during high-level ozone days. We divide the Houston metropolitan area into a 20×27 grid

and use ordinary Kriging interpolation method [4] to estimate the hourly ozone concentration on the 20×27=540 grid intersection points. Using Kriging is motivated by the fact that the number of ozone-monitoring stations are far less than the number of the intersection points of the grid structure.

Kriging interpolation is a common method used by scientists in the environmental research. Particularly, Kriging interpolation deals with the uneven sampling issue in the sample space. The objective of the Kriging interpolation is to estimate the value of an unknown function,  $f$ , at a point  $\tilde{x}$ , given the values of the function at some observed points,  $x_1, \dots, x_n$ , and weights  $w_1, \dots, w_n$ .

$$f(\tilde{x}) = \sum_{i=1}^n w_i f(x_i)$$

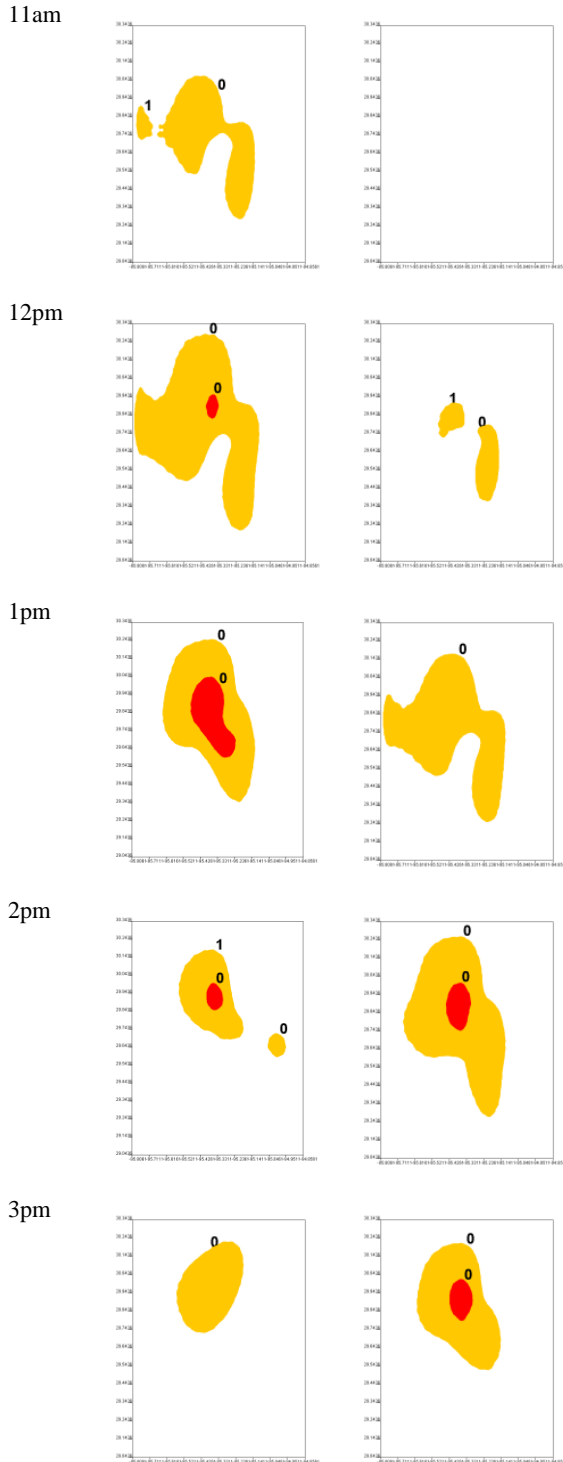
$$\sum_{i=1}^n w_i = 1$$

We are interested in analyzing the progression of ozone pollution over time. Moreover, we are interested in general patterns of ozone concentration progression, but not in identifying suddenly occurring ozone concentrations that disappear quickly. Consequently, our approach creates polygon models from sequences of ozone concentration snapshots using a sliding window approach; in this case study ozone a 2 hour sliding window is used. Figure 5-2a visualizes the progression of ozone hotspots from 11am to 3pm on August 26<sup>th</sup>, 2008. The orange polygons capture areas having ozone level above 75 ppb. The red polygons represent areas having 1-hour ozone concentration above 100 ppb. Table 5-2a summarizes parameter settings of change predicates for the dataset.

**Table 5-2a. Parameter settings of change predicates for ozone concentration dataset**

Agreement	Low=0.2	Medium=0.5	High=0.8
Containment	Low=0.2	Medium=0.5	High=0.8
Distance	Min=0.05	Max=0.5	

From Figure 5-2a, at 11am, two new orange hotspots are visible. At 12pm, the two orange clusters (orange polygons id 0 and 1) in the model merged into a larger data cluster (orange polygon id 0),  $\text{Containment}(c0,m0)$  is 0.137 and is less than  $\text{Containment}(m0,c0)$  which is 1.  $\text{Containment}(c0,m1)$  is 0.06 and is less than  $\text{Containment}(m1,c0)$  which is 1.  $\text{Distance}(c0,m0)$  and  $\text{Distance}(c0,m1)$  are 0.332 and 0.263 which are less than the  $\text{distance}_{max}$ . A higher ozone level hotspot represented by red polygon id 0 also becomes visible at 12pm, the  $\text{Agreement}(c0,m0)$  is 0 and  $\text{Distance}(c0,m0)$  is greater than  $\text{distance}_{max}$ . The red polygon id 0 in the data grows to its largest size of the day at 1pm. Although this red polygon seems to be a polygon that grows from the red polygon id 0 at 12pm, but we declare that it is still a novel concept according to the novel concept predicate (see its agreement and distance in Table 5-2b). This is because the model does not immediately recognize the small, red polygon id 0 as a pattern but does so at 1pm after another red polygon occurs in the same area. In general, creating a model based on multiple ozone readings makes the model more fault tolerant, and reduces the probability of false alarms.



**Figure 5-2a. Visualizations of hourly Ozone data on August 26<sup>th</sup>, 2008; images on the left column are Ozone hotspots discovered by DCONTOUR from 11:00am to 03:00pm; image on the right column are images of the models for the data in the left column.**

**Table 5-2b. Change predicate values for polygon (red) id 0 of high ozone concentration (above 100 ppb) a from 11:00 to 15:00 on August 26<sup>th</sup>,2008 (NaN is not a number caused by a number divided by 0)**

Time	Size c0	Size m0	Agreement (c0,m0)	Containment		Distance (c0,m0)
				(c0,m0)	(m0,c0)	
11:00	0.0	0.0	NaN	NaN	NaN	0.0
12:00	0.006	0.0	0.0	0.0	NaN	Inf.
13:00	0.061	0.0	0.0	0.0	NaN	Inf.
14:00	0.01	0.025	0.337	0.913	0.348	0.043
15:00	0.0	0.02	0.0	NaN	0.0	Inf.

The size of the orange polygon starts to shrink at 1pm with the Containment(c0,m0) which is 0.7406 is less than the Containment(m0,c0) which is 0.7181 and Agreement(c0,m0) which is 0.5738 is close to Containment(m0,c0) which is 0.7181. The Distance(c,m) which is 0.092 is between  $distance_{max}$  and  $distance_{min}$  and points towards concept drift. The red polygon also starts to shrink at 2pm and completely disappears at 3pm.

## 5. RELATED WORK

Recently, several techniques have been proposed to detect and analyze change patterns in different types of data. In work by Asur et al. [1], a technique for mining evolutionary behavior of interaction graphs is proposed. Yang et al. [17] proposed a technique to discover the evolution of spatial objects. Association analysis is used to detect changes. In [7], Jiang and Worboys propose a technique to detect topological changes in sensor networks. The topological changes that can be discovered include region appearance and disappearance, region merging and splitting, and hole formation and elimination. A framework for tracking cluster transition in data streams is proposed by Spiliopoulou et al in [13]. The proposed framework can detect external and internal transitions of clusters. Kifer et al. [11] present a method for the detection and estimation of change in data streams. The proposed algorithm reduces the problem of analyzing continuously streaming data to analyzing two static sample sets: a reference window and a sliding window containing current points in the data streams. The algorithm compares discrepancies between two windows by means of analyzing probability distributions.

In summary, the work in [17] and [7] is similar to our work in that they focus on spatial data. Unlike [17] which uses association analysis, our work utilizes cluster analysis. Similar to our technique, the techniques in [1] [17], [7], [13], and [6] analyze changes in data but require that the identity of objects must be known or restrict analysis to objects that are characterized by nominal attributes. In contrast, our work is applicable for data with unknown object identity and for datasets that contain numerical attributes.

There are many papers that address the problems of concept drift or novelty detection but the Spinoza et al. [14] are the first that tackles the problem of novelty detection in the presence of concept drift. The paper introduces a cluster-based technique that can detect novel concepts as well as deal with concept drift in data streams; cluster models are created by k-means clustering algorithm. Our framework, on the other hand, relies on contour clustering algorithm to create polygon models; moreover, polygons in our approach do not need to be convex. An advantage of our work over [14] is that our change predicates allow

discovery of not only novelty and concept drift but also other types of change patterns.

## 6. CONCLUSION AND FUTURE WORK

We introduce a framework for change analysis that utilizes polygons as cluster models to capture changes in spatial datasets between different time frames. The proposed framework has been applied to two real-world datasets to analyze the changes in the visual fields and the ozone concentrations. The framework consists of three modules: Cluster Generation, Model Generation, and Change Analysis. The Cluster Generation module creates contour clusters from newly arriving data using density or interpolation functions and a contour clustering algorithm that operates on those functions. The polygon models are updated in the Model Generation module to reflect changes in data that occur over time. The Change Analysis module uses sets of change predicates, which can be viewed as queries that detect and analyze changes through polygon operations. In general, the change analysis tool is highly generic and supports arbitrary sets of change predicates as long as they operate on sets of polygons. Our prototype implementation uses the Java-based Cougar<sup>2</sup> framework [2], [5]; however, we are beginning to re-implement the change analysis tool as a subcomponent of a spatial database system.

The case studies show that our framework can capture various kinds of changes in spatial datasets. We plan to implement a tool for optometrists to study various change patterns that are associated with different stages in the progression of glaucoma. The tool aims to assist the optometrists to better understand the development and progression of the disease. Our ultimate vision of this project is to use the tool to help diagnose glaucoma and to provide an expert system to assist optometry students in learning about the different stages of the disease.

Moreover, we plan to integrate our change analysis framework and change report generators into early warning systems so that alarms are raised and change reports are generated automatically when critical events are discovered. To accomplish this goal, efficient data-driven algorithms that integrate change analysis tools into early warning systems have to be developed.

## 7. REFERENCES

- [1] Asur, S., Parthasarathy, S., and Ucar, D. 2007. An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining .
- [2] Bagherjeiran, A., Celepcikay, O. U., Jiamthapthaksin, R., Chen, C.-S., Rinsurongkawong, V., Lee, S., Thomas, J. and Eick, C. F. 2009. Cougar<sup>2</sup>: An Open Source Machine Learning and Data Mining Development Framework. In Proceedings of Open Source Data Mining Workshop (2009)
- [3] Chen, C.-S., Rinsurongkawong, V., Eick, C.F., and Twa, M.D. 2009. Change Analysis in Spatial Data by Combining Contouring Algorithms with Supervised Density Functions. In Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining
- [4] Cressie, N. 1993. Statistics for spatial data. New York: Wiley.
- [5] Cougar<sup>2</sup>, <https://cougarsquared.dev.java.net>
- [6] Eick, C.F., Vaezian, B., Jiang, D., and Wang, J. 2006. Discovery of Interesting Regions in Spatial Datasets Using Supervised Clustering. In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases .
- [7] Jiang, J., and Worboys, M. 2008. Detecting Basic Topological Changes in Sensor Networks by Local Aggregation. In Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems
- [8] Johnson, C.A., Sample, P.A., Cioffi G.A., Liebmann, J.R., and Weinreb, R.N. 2002. Structure and Function Evaluation (SAFE): I. Criteria for Glaucomatous Visual Field Loss using Standard Automated Perimetry (SAP) and Short Wavelength Automated Perimetry (SWAP). *Am J Ophthalmol*; 134(2): 177-185.
- [9] Keltner, J.L., Johnson, C.A., Cello K.E., Edwards, M.A., Bandermann, S.E., Kass, M.A., and Gordon, M.O. 2003. Classification of Visual Field Abnormalities in the Ocular Hypertension Treatment Study. *Arch Ophthalmol*; 121(5): 643-650.
- [10] Keys, R. 1981. Cubic Convolution Interpolation for Digital Image Processing. *IEEE Transactions on Signal Processing, Acoustics, Speech, and Signal Processing* 29: 1153
- [11] Kifer, D., Ben-David, S., and Gehrke, J. 2004. Detecting Change in Data Streams. In Proceedings of the 30th International Conference on Very Large Data Bases.
- [12] Snyder, William V. 1978. Algorithm 531: Contour Plotting [J6]. *ACM Transactions on Mathematical Software (TOMS) (ACM)* 4, no. 3, 290 – 294
- [13] Spiliopoulou, M., Ntoutsis, I., Theodoridis, Y., and Schult, R. 2006. Monic – Modeling and Monitoring Cluster Transitions. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining .
- [14] Spinosa, E.J., Carvalho, A.P.L.F., and Gama, J. 2007. OLINDDA: A Cluster-based Approach for Detecting Novelty and Concept Drift in Data Streams. In Proceedings of the 22nd Annual ACM Symposium on Applied Computing .
- [15] Tan, P.-N., Steinbach, M., Kumar, V. 2006. Cluster Evaluation. *Introduction to Data Mining*, Pearson Education Inc., Pages 532-555.
- [16] The Texas Commission on Environmental Quality, [http://www.tceq.state.tx.us/nav/data/ozone\\_data.html](http://www.tceq.state.tx.us/nav/data/ozone_data.html)
- [17] Yang, H., Parthasarathy, S., and Mehta, S. 2005. A Generalized Framework for Mining Spatio-temporal Patterns in Scientific Data. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining .