Change Analysis in Spatial Data by Combining Contouring Algorithms with Supervised Density Functions

Chun Sheng Chen[1] , Vadeerat Rinsurongkawong[1],
Christoph F. Eick[1], and Michael D. Twa[2]

[1] Department of Computer Science, University of Houston,
Houston, TX 77204
{lyons19, vadeerat, ceick}@cs.uh.edu
[2] Department of Optometry, University of Houston,
Houston, TX 77204
mtwa@optometry.uh.edu

Abstract. Detecting changes in spatial datasets is important for many fields. In this paper, we introduce a methodology for change analysis in spatial datasets that combines contouring algorithms with supervised density estimation techniques. The methodology allows users to define their own criteria for features of interest and to identify changes in those features between two datasets. Change analysis is performed by comparing interesting regions that have been derived using contour clustering. A novel clustering algorithm called DCONTOUR is introduced for this purpose that computes contour polygons that describe the boundary of a supervised density function at a given density threshold. Relationships between old and new data are analyzed relying on polygon operations. We evaluate our methodology in case studies that analyze changes in earthquake patterns.

Keywords: Change analysis, spatial data mining, region discovery, supervised density estimation technique, contour clustering algorithm, interestingness measures.

1 Introduction

Spatial datasets, containing geo-referenced data, are growing at a very high speed. Detecting changes in spatial datasets is important for many fields such as early warning systems that monitor environmental conditions or sudden disease outbreaks, epidemiology, crime monitoring, and automatic surveillance.

To address this need, this paper introduces novel methodologies and algorithms that discover patterns of change in spatial datasets. We are interested in finding what patterns emerged between two datasets, O_{old} and O_{new}, sampled at different time frames. Change analysis centers on identifying changes concerning interesting regions with respect to O_{old} and O_{new}. Moreover, two approaches to define interestingness perspectives are introduced. The first approach employs supervised density functions [8] that create density maps from spatial datasets. As we will explain later, regions (contiguous areas in the spatial subspace) where density functions take high (or low) values are considered interesting by this approach. A novel clustering algorithm DCONTOUR is introduced to identify interesting regions. For some applications, it is impossible to capture a user's interestingness perspective using a supervised density function. The second approach is proposed to overcome this limitation. It utilizes a preprocessing step that computes interesting regions that have been obtained by maximizing a plug-in reward-based interestingness function. DCONTOUR is then applied to the individual regions obtained, and change is analyzed by comparing the obtained contour polygons.

The contributions of this paper include:

1. A novel clustering algorithm called DCONTOUR is introduced. To the best of our knowledge, DCONTOUR is the first density-based clustering algorithm that uses contour lines to determine cluster boundaries that are described as polygons. Objects that are inside a contour polygon belong to a cluster. DCONTOUR operates on the top of supervised density functions that capture what is considered to be interesting places by a domain expert.
2. A framework for change analysis in spatial dataset is presented that compares interesting regions that have been derived using contour clustering. It analyzes change in interestingness by comparing contour polygons.

2 Change Analysis Using Supervised Density Estimation Approach

2.1 Supervised Density Estimation

We assume that objects $o \in O$ have the form ((x, y), z) where (x, y) is the location of object o and z—denoted as z(o) is the value of the variable of interest of object o. In the following, we will introduce supervised density estimation techniques. Density estimation is called supervised because in addition to the density based on the locations of objects, we take the variable of interest z(o) into consideration when measuring density. The density estimation techniques employ influence functions that measure the influence of a point $o \in O$ with respect to another point $v \in F$; in general, a point o's

influence on another point v's density decreases as the distance between o and v, denoted by d(o,v), increases. In contrast to past work in density estimation, our approach employs weighted influence functions to measure the density in datasets O: the influence of o on v is weighted by z(o) and measured as a product of z(o) and a Gaussian kernel function. In particular, the influence of object $o \in O$ on a point $v \in F$ is defined as:

$$f_{\text{inf luence}}(v,o) = z(o) * e^{-\frac{d(v,o)^2}{2*\sigma^2}} \quad (2\text{-}1)$$

If $\forall o \in O$ z(o)=1 holds, the above influence function becomes a Gaussian kernel function, commonly used for density estimation and by the density-based clustering algorithm DENCLUE [7]. The parameter σ determines how quickly the influence of o on v decreases as the distance between o and v increases. The overall influence of all data objects $o_i \in O$ for $1 \leq i \leq n$ on a point $v \in F$ is measured by the density function $\psi^O(v)$, which is defined as follows:

$$\psi^o(v) = \sum_{i=1}^{n} f_{\text{inf luence}}(v, o_i) \quad (2\text{-}2)$$

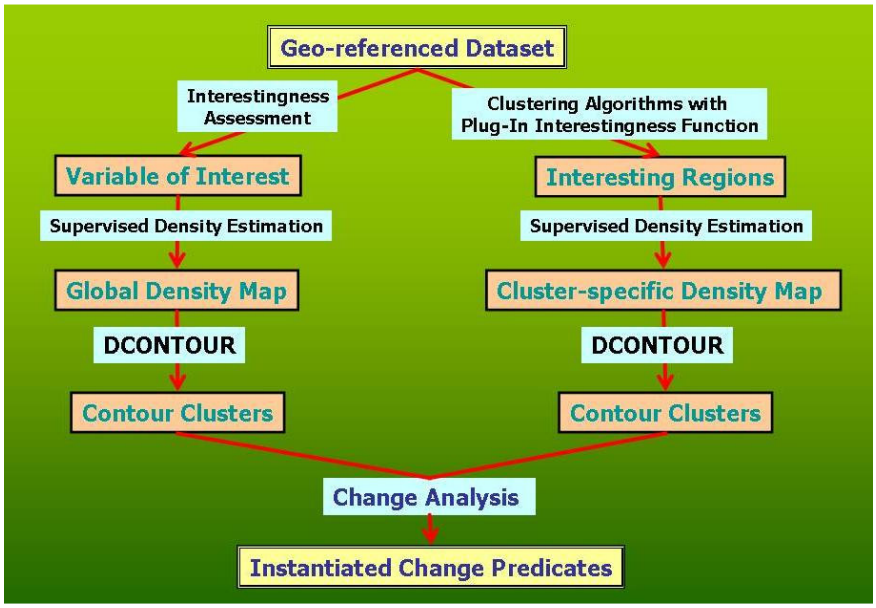


Fig. 2-1. Change analysis approaches investigated in this paper.

In summary, supervised density estimation does not only consider the frequency with which spatial events occur but also takes the value of the

variable of interest into consideration—in general, density increases as frequency and z(o) increase.

For instance, we might be interested in finding hotspot areas where the risk of earthquake is high; in this case z is defined as the severity of an earthquake. Using formula 2-2, a severity density map of earthquakes can be created and the hotspot locations of earthquakes can be directly identified as high density areas in the map. In section 5, more examples of supervised density functions will be given.

2.2 Change Analysis through Contour Clustering

We have developed a contour clustering algorithm named DCONTOUR that combines contouring algorithms and density estimation techniques. Fig. 2-2 gives the pseudo-code of the algorithm.

Input: Density function ψ^{ρ}, density threshold d.
Output: Density polygons for density threshold d.
1. Subdivide the space into D grid cells.
2. Compute densities at grid intersection points by using density function ψ^{ρ}.
3. Compute contour intersection points b on grid cell edges where $\psi^{\rho}(b) = d$ using binary search and interpolation.
4. Compute contour polygons from contour intersection points b.

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

Step 1 creates a grid structure for the space and step 2 computes the density for each grid intersection point. Step 2 will call the density function $O(D)$ times where D is the number of grid cells. In general, objects that are far away from a point contribute very little to the density of the point. Therefore, in order to speed up step 2, we implemented an "approximate" density function that only considers the influence of objects belonging to neighboring grid cells rather than all the objects in the dataset. Step 3 computes contour intersection points on grid edges. Since the density function is defined in the whole space and is nonlinear, binary serach on cell-edges is used in step 3 to limit the interpolation error. Fig. 2-3 gives an illustration of how contour intersection points for d=4.5 are constructed. As far as the right edge of the lower left cell is concerned, because 4.5 is between 4.4 and 5.5, a contour intersection point exists on this edge; interpolating between 4.1 and 5.5, a point on this edge is sampled and its density is computed which turns out to be 4.8. Because 4.8 is larger than d, we continue the binary search by sampling a point south of this point. The binary search terminates if the density difference between a sampled point and d is less than a threshold ε. Finally, in step 4, we connect contour intersection points b found on cell edges and continue this process on its neighboring cells until a closed polygon is formed or both ends of the

polyline reach the grid boundary. Step 4 uses an algorithm that was proposed by Cottafava, and Moli [3] to compute contour polygons.

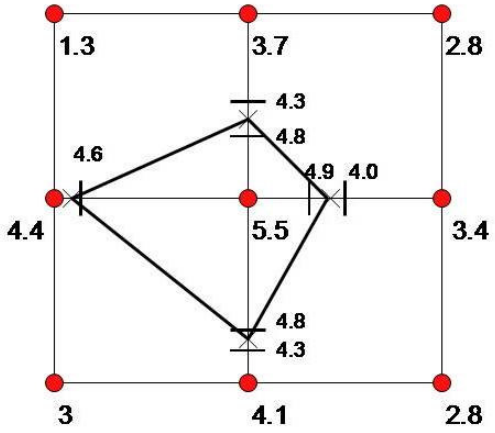


Fig. 2-3. Contour construction for d=4.5

Traditional contouring algorithms operate on datasets of the form ((x,y),u) where u is a measurement of an attribute of interest at the location (x,y), and use interpolation to infer values of u in locations that were not sampled. DCONTOUR, on the other hand, creates contour polygons for a given density intensity using supervised density maps as its input. A contour polygon acts as a boundary of interesting regions that are above (or below) a specific density threshold; objects surrounded by each individual polygon are defined as a cluster.

The proposed change analysis approach that is depicted in Fig. 2.1, applies DCONTOUR to O_{old} and O_{new}, and analyzes change with respect to the obtained contour polygons. How this is exactly done will be discussed in Sections 4 and 5.

3 Change Analysis by Contouring Interesting Regions

The supervised density estimation approach assesses interestingness by assigning an interesting value z(o) to the objects in a dataset. Unfortunately, for some applications it is impossible to capture a user's interestingness perspective using a supervised density function. We will give an example to illustrate this point. Let us assume we perform change analysis with respect to two earthquake datasets, and we are only interested in comparing regions for which the variance of earthquake depth regions is high—basically we are

interested in changes in places were deep earthquakes are next to shallow earthquakes. In this case, interestingness of a region depends on the variance with respect to the depth of the earthquakes that belong to this region; that is, the squared difference of the depth of an earthquake and the mean-depth of a region can be viewed as a proxy for interestingness. Unfortunately, the mean value itself depends on the scope of the region and is therefore not known in advance. Consequently, it is impossible to capture this notion of interestingness using a supervised density function, because z(o) cannot be measured prior to knowing which region an object o belongs to. Basically, we need an approach that computes interesting regions first prior to computing z(o). Such an approach will be introduced in the following.

The approach (see also Fig. 2-1) relies on contouring individual interesting regions that have been determined in a preprocessing step. First, we run a region discovery algorithm to identify interesting regions. Next, we run DCONTOUR on the clustering results of the region discovery algorithm to determine contour clusters. Finally, the derived contour polygons are compared.

A region discovery framework, that has been introduced in [4], to find scientifically interesting places in spatial datasets, is used for the preprocessing step. The goal of region discovery is to find a set of regions that maximize an externally given interestingness function. To illustrate this approach, we introduce an interestingness function for the earthquake depth variance example discussed earlier. The interestingness of a region r, i(r), is defined as follows:

$$i(r) = \begin{cases} 0 & \frac{Var(r,depth)}{Var(O,depth)} \leq th \\ \left(\frac{Var(r,depth)}{Var(O,depth)} - th\right) * |r|^{\beta} & otherwise \end{cases} \qquad (3\text{-}1)$$

where $$Var(r,depth) = \frac{1}{|r|-1} \sum_{o \in r} (depth(o) - \mu_{depth}(r))^2 \qquad (3\text{-}2)$$

In the formula depth(o) denotes depth of an earthquake o. The interestingness function computes the ratio of the region's variance with respect to depth and the dataset's variance. Regions whose ratio is above a given threshold th receive rewards. |r| is the number of objects in a region r. β is a parameter that determines a premium associated with the number of objects in a cluster—choosing higher values for β usually leads to the discovery of larger regions. $\mu_{depth}(r)$ is an average earthquake depth in a region r. The interestingness function parameters β and th are determined in close collaboration with a domain expert. Region discovery algorithm identifies interesting regions by maximizing i(r). After interesting regions have been identified, a supervised

density function is created for each region. Finally, contour polygons are created for each region using DCONTOUR and compared.

4 Change Analysis Predicates

This section introduces basic predicates that capture different relationships for change analysis. Given two clusterings X and X' for O_{new} and O_{old}, respectively, relationships between the regions that belong to X and X' can be analyzed. Let r be a region in X and r' be a region in X'. In this case, agreement between r and r' can be computed as follows:

- Agreement(r,r')= $|r \cap r'|/|r \cup r'|$

In general, the most similar region r' in X' with respect to r in X is the region r' for which Agreement(r,r') has the highest value. In addition to agreement, we also define predicates novelty, relative-novelty, disappearance and relative-disappearance below.

Let r, r_1, r_2,..., r_k be regions discovered at time t, and r', r_1', r_2',…, r_k' be regions that have been obtained for time t+1.

- Novelty (r') = $(r' — (r_1 \cup \ldots \cup r_k))$
- Relative-Novelty(r') = $|r' — (r_1 \cup \ldots \cup r_k)|/|r'|$
- Disappearance(r) = $(r — (r'_1 \cup \ldots \cup r'_k))$
- Relative-Disappearance(r) = $|r — (r'_1 \cup \ldots \cup r'_k)|/|r|$

Novelty measure captures regions that have not been interesting in the past. On the other hand, disappearance is used to discover regions where those characteristics are disappearing. Relative-novelty and relative-disappearance measure percentages of novelty and disappearance. We claim that the above and similar measurements are useful to identify *what is new* in a changing environment. Moreover, the predicates we introduced so far can be used as building blocks to define more complex predicates.

It is also important to note that the above predicates are generic in the sense that they can be used to analyze changes between the old and new data based on different types of clustering. The change analysis approach that we introduced in sections 2 and 3 uses polygons as cluster models. Consequently, in our particular approach the operators '$\cap$', '$\cup$', and '—' denote polygon intersection, union and difference and |r| computes the size of a polygon r. For example, agreement between two polygons r and r' is computed as the ratio of the size of the intersection between r and r' over the size the union of r and r'.

5 Demonstration

We demonstrate our methods on an earthquake dataset which is available on website of the U.S. Geological Survey Earthquake Hazards Program http://earthquake.usgs.gov/. Information recorded includes the location (longitude, latitude), the time, the severity (Richter magnitude) and the depth

(kilometers) of earthquakes. We uniformly sampled earthquakes dated from January 1986 to November 1991 as dataset O_{old} and earthquakes between December 1991 and January 1996 as dataset O_{new}. Each dataset contains 4132 earthquakes.

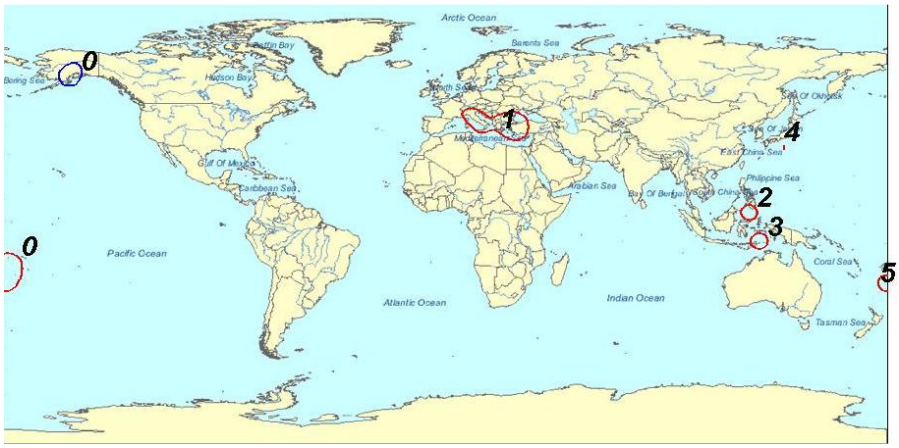
The change analysis framework is tested in two case studies: 1) discovering regional changes in strong correlations with respect to earthquake depth and severity in section 5.1; 2) detecting changes of regions in which deep and shallow earthquakes are in close proximity in section 5.2.

5.1 Changes in Earthquake Severity/Depth Correlation

In this section, we discuss a case study that analyzes changes in strong positive or negative correlations between the depth of the earthquake and the severity of the earthquake. Accordingly, the variable of interest, z(o) is defined as follows:

$$z(o) = \frac{(severity(o) - \mu_{severity})}{\sigma_{severity}} * \frac{(depth(o) - \mu_{depth})}{\sigma_{depth}} \qquad (5\text{-}1)$$

where $\mu_{severity}$ and μ_{depth} are the mean values of the severity and depth of the dataset and $\sigma_{severity}$ and σ_{depth} are the standard deviation of the severity and depth, respectively. It should be noted that that z(o) takes positive and negative values, and that the constructed density function now contains hotspots (areas high positive density) and cool spots (areas of high negative density). Figures 5-1.a shows the results of running DCONTOUR twice to generate the red and blue contour polygons in the figure; once with a negative density threshold and once with a positive threshold.



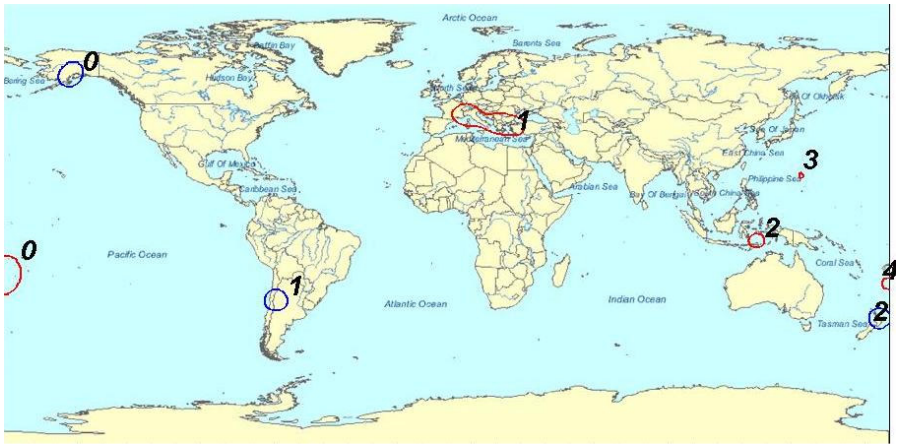


Fig. 5-1.a. Contour polygons generated by DCONTOUR for O_{old} (upper figure) and O_{new} (lower figure). The blue polygons indicate areas with significant negative correlations (deep earthquakes are always less severe and shallow earthquakes tend to be strong). Red polygons are areas having positive correlations between the two variables.

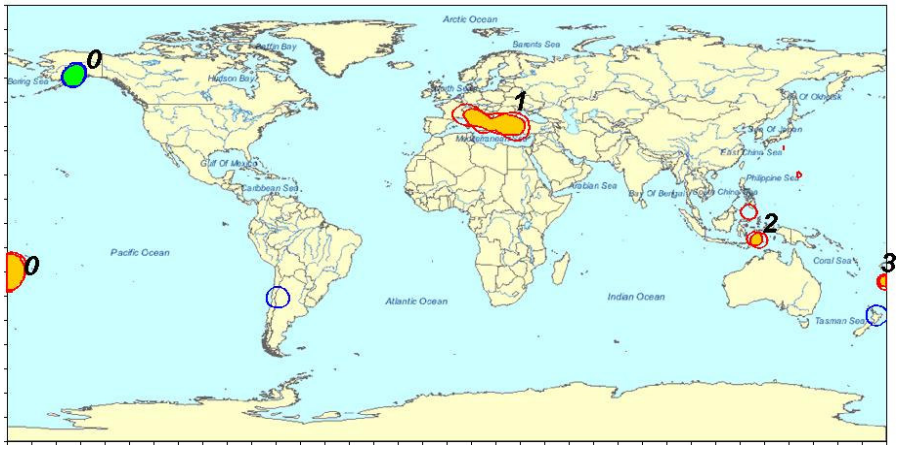


Fig. 5-1.b. Overlap of contour polygons of O_{old} and O_{new} dataset

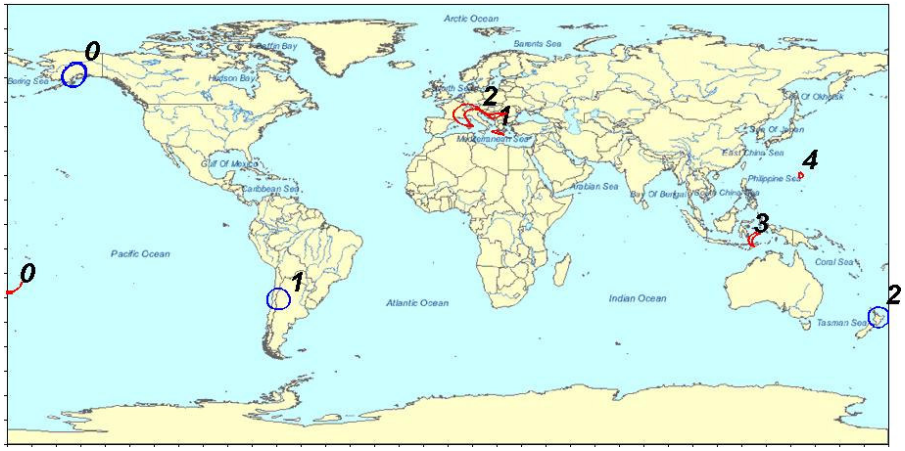


Fig. 5-1.c. Novelty areas of dataset O_{new} compared to dataset O_{old}

Table 5-1.a. Agreement of contour polygons in Figure 5-1.a.

Agreement				
Area#	0	1	2	3
Red area	0.84	0.6	0.48	0.56
Blue area	0.87			

Table 5-1.b. Novelty and relative-novelty for polygons in Figure 5-1.a.

Red area#	0	1	2	3	4
Novelty	3.6	3.06	53.3	9.00	2.43
Relative-novelty	0.04	0.01	0.25	0.31	1
Blue area#	0	1	2		
Novelty	10.3	62.66	54.74		
Relative-novelty	0.13	1	1		

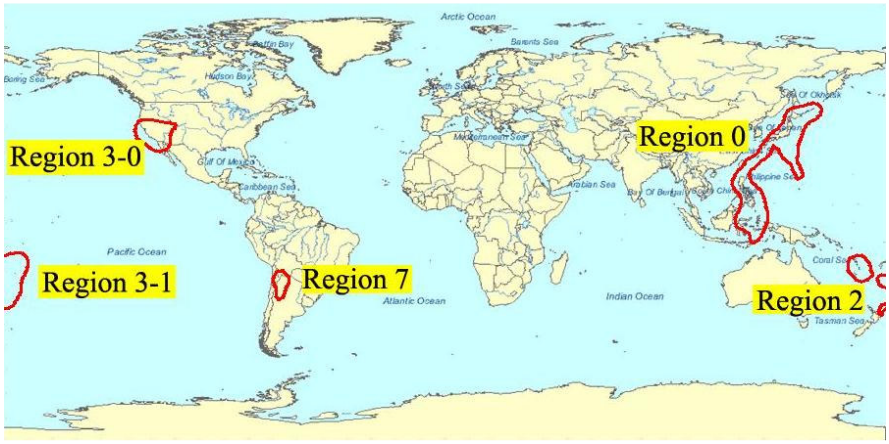
In Fig. 5-1.b, intersection areas of the two datasets filled by orange are positive-correlated areas and filled by green are negative-correlated areas. In Table 5-1.a, the high agreement of area red#0 and blue#0 indicates that the high positive and negative correlation of severity and depth of earthquakes in these areas have not changed from the old data to the new data. Fig. 5-1.c shows the novel areas in dataset O_{new} comparing to dataset O_{old} for both positive-correlated and negative-correlated areas. The novelty and relative-novelty are listed in Table 5-1.b. We can see that area red#4, blue#1 and blue#2 have the relative-novelty equal to 100% indicating that these areas are new clusters only exist in dataset O_{new}. This can be verified by comparing the contour polygons of dataset O_{new} with the ones of dataset O_{old} in Fig. 5-1.a.

5.2 Changes in High Variance of Earthquake Depth

In this section, we analyze changes in high variance regions with respect to earthquake depth. As discussed earlier only the second change analysis approach is suitable for this problem. We ran the region discovery framework using CLEVER clustering algorithm [5] with the interestingness function that was introduced in Section 3. The CLEVER output characterizes regions by the set of earthquakes that belong to the region. Next, supervised density functions were created for each clusters separately, defining the variable of interest z(o) as follows:

$$z(o)=|depth(o)-\mu_{depth}(r)| \quad (5\text{-}2)$$

where r is the region to which object o belongs to. Earthquakes belonging to a region are weighted by the absolute difference between earthquake depth and the regions average earthquake depth. Basically, earthquakes whose depth deviates significantly from the average depth contribute more to density. Next, DCONTOUR is applied to each region's supervised density function, and contour polygons are derived, as shown in Fig. 5-2.a. The polygon areas indicated by red lines show the boundary of the regions that received rewards. From Table 5-2.a, the high value of agreement of 0.74 of region 3 in O_{old} and region 5 in O_{new} indicates that the variance of earthquake depth did not change significantly; what changed can be observed in Fig. 5-2.a. The novelty of regions in O_{new} and the disappearance of regions in O_{old} are illustrated in Fig. 5-2.b. The relative disappearance value of region 7 is 100% indicate that the high variance in this region completely disappeared in the new data.



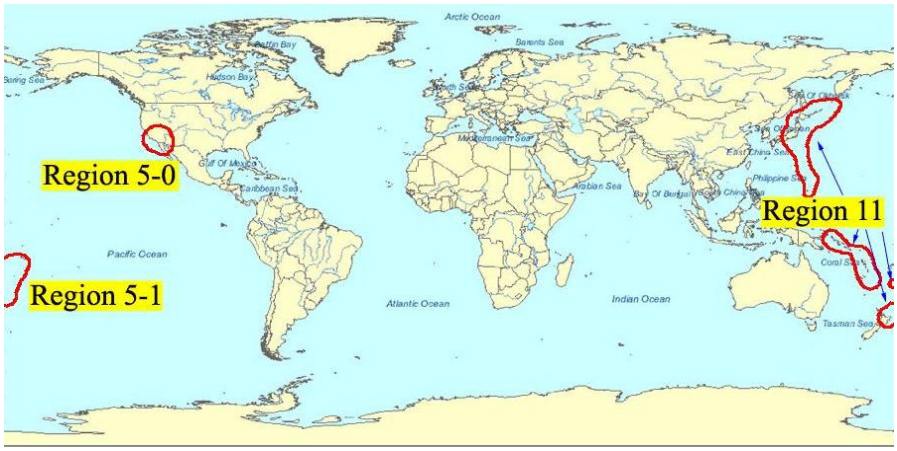


Fig. 5-2.a. Results showing regions where variance of earthquake depth is high in O_{old} data (upper figure), and O_{new} data (lower figure).

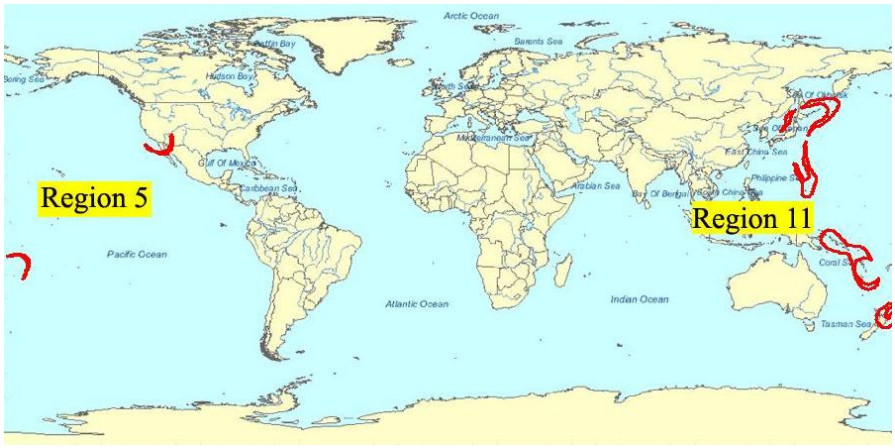


Fig. 5-2.b. Disappearance areas of regions in O_{old} data (upper figure) and novelty areas of regions in O_{new} data (lower figure).

Table 5-2.a. Agreement of contour polygons in Figure 5-2.a.

Agreement				
Region Old # / Region New #	0	2	3	7
5	0	0	0.74	0
11	0.25	0.08	0	0

Table 5-2.b. Novelty and relative-novelty for regions in O_{new} and disappearance and relative-disappearance for regions in O_{old} in Figure 5-2.a.

Region New #	Novelty	Relative-Novelty	Region Old #	Disappearance	Relative-Disappearance
5	16.98	0.06	0	356.29	0.59
11	297.20	0.47	2	13.77	0.14
			3	73.31	0.23
			7	51.29	1

6 Related Work

Our change analysis is closely related to clustering analysis. Recently, scan statistic algorithms [10, 9] were introduced to detect significant clusters newly emerged in geographical space. This differs from our approach since the algorithms are limited to hotspot discovery, and are not capable of detecting other types of change.

In 2006, a framework for change description that perceived changes of clusters as changes of states in a state space was proposed by Fleder et al. [6]. A framework for tracking external and internal cluster transitions in a data stream was introduced by Spiliopoulou et al [11] in the same year. In 2007, a

technique for mining evolutionary behavior of interaction graphs was proposed by Asur et al. [1]. In general, these methods [1, 6, 11] can detect many types of change patterns but they require that the identity of objects must be known or objects must be characterized by nominal attributes. The advantage of our approaches is that we can detect various types of changes in data with continuous attributes and unknown object identity.

Existing contour plotting algorithms can be seen as variations of two basic approaches: level curve tracing [12] and recursive subdivision [2]. Level curve tracing algorithms scan a grid and mark grid-cell boundaries that are passed by the level curve. Contour polygons are created by connecting the marked edges. Recursive subdivision algorithms start with a coarse initial grid and recursively divide grid cells that are passed by the level curve. Our algorithm, DCONTOUR, uses level curve tracing.

7 Summary

Developing techniques for discovering change in spatial datasets is important and providing methods to detect change for continuous attributes and for objects that are not identified apriori are advantages of the techniques we describe here. In this paper, change analysis techniques that rely on comparing clusters for the old and new data based on a set of predicates are proposed. A contour clustering algorithm named DCONTOUR that combines supervised density functions with contouring algorithms is introduced to automate this task.

In general, our work is a first step towards analyzing complex change patterns. The novel contributions of this paper includes: 1) using density functions in contouring algorithm; 2) change analysis is conducted by interestingness comparison; 3) degrees of change are computed relying on polygon operations; 4) two novel change analysis approaches were introduced: one directly uses supervised density functions; the second approach derives density functions from interesting regions that have been obtained using a region discovery algorithm that relies on reward-based interestingness functions.

References

1. Asur, S., Parthasarathy, S., Ucar, D.: 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 (2007)
2. Bruss, A. R.: Extracting Topographic Features from Elevation Data using Contour Lines. Working Paper 146, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA (1977) http://hdl.handle.net/1721.1/41980

3. Cottafava, G. and Le Moli, G.: Automatic Contour Plotting. Communication of the ACM, vol. 12, no. 7 (1969)
4. Withheld
5. Withheld
6. Fleder, D. and Padmanabhan, B.: Cluster Evolution and Interpretation via Penalties. In Proceedings of the 6th IEEE International Conference on Data Mining – Workshops (2006)
7. Hinneburg, A. and Keim, D. A.: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (1998)
8. Jiang, D., Eick, C. F. and Chen, C.-S.: On Supervised Density Estimation Techniques and Their Application to Clustering. In Proceeding of the 15th ACM International Symposium on Advances in Geographic Information Systems (2007)
9. Kulldorff, M.: Prospective Time-Periodic Geographical Disease Surveillance using a Scan Statistic. Journal of the Royal Statistical Society A, vol. 164, 61-72 (2001)
10. Neil, D. B., Moore, A. W., Sabhnani, M., and Daniel K.: Detection of Emerging Space-Time Clusters. In Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (2005)
11. Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., and Schult, R.: Monic – Modeling and Monitoring Cluster Transitions. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)
12. Watson, D.: Contouring — A Guide to the Analysis and Display of Spatial Data. Computer Methods in the Geosciences, vol. 10, Pergamon Press, Oxford, (1992)