Research Article

Spatiotemporal Hotspots Analysis for Exploring the Evolution of Diseases: An Application to Oto-Laryngopharyngeal Diseases

Ferdinando Di Martino,1 Roberta Mele,1 Umberto E. S. Barillari,2 Maria Rosaria Barillari,2 IrinaPerfilieva,3 and Sabrina Senatore4

1 Università degli Studi di Napoli Federico II, Dipartimento di Architettura, Via Toledo 402, 80134 Napoli, Italy
2 Seconda Università degli Studi di Napoli, Dipartimento di Psichiatria, Neuropsichiatria Infantile, Audiofoniatria e Dermatovenereologia, L.go Madonna delle Grazie, 80138 Napoli, Italy
3 Centre of Excellence IT4 Innovations, Institute for Research and Applications of Fuzzy Modelling, University of Ostrava, 30. dubna 22, 70103 Ostrava, Czech Republic
4 Università degli Studi di Salerno, Dipartimento di Informatica, Via Ponte don Melillo, 80084 Fisciano, Salerno, Italy

Correspondence should be addressed to Ferdinando Di Martino; fdimarti@unina.it

Received 1 May 2013; Revised 29 July 2013; Accepted 29 July 2013

Academic Editor: Salvatore Sessa

Copyright © 2013 Ferdinando Di Martino et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents a spatiotemporal analysis of hotspot areas based on the Extended Fuzzy C-Means method implemented in a geographic information system. This method has been adapted for detecting spatial areas with high concentrations of events and tested to study their temporal evolution. The data consist of georeferenced patterns corresponding to the residence of patients in the district of Naples (Italy) to whom a surgical intervention to the oto-laryngopharyngeal apparatus was carried out between the years 2008 and 2012.

1. Introduction

In a GIS, the impact of phenomena in a specific area due to the proximity of the event (e.g., the study of the impact area of an earthquake, or the area constraint around a river basin) is performed using buffer area geoprocessing functions. Given a geospatial event topologically represented as a georeferenced punctual, linear, or areal element, an atomic buffer area is constituted by circular areas centered on the element. For example, if the event is the epicenter of an earthquake, georeferenced by a point, a set of buffer areas is formed by concentric circular areas around that point; the radius of each circular buffer area is defined a priori.

When it is not possible to define statically an area of impact and we need to determine what is the area affected by the presence of a consistent set of events, we are faced with the problem of detecting this area as a cluster on which the georeferenced events are thickened as well. These clusters are georeferenced, represented as polygons on the map, and called hotspot areas.

The study of hotspot areas is vital in many disciplines such as crime analysis [1–3], which studies the spread on the territory of criminal events, fire analysis [4], which analyzes the phenomenon of spread of fires on forested areas, and disease analysis [5–7], which studies the localization of focuses of diseases and their temporal evolution. The clustering methods mainly used for detecting hotspot areas are the algorithms based on density (see [8, 9]); they can detect the exact geometry of the hotspots, but are highly expensive in terms of computational complexity, and in the great majority of cases, it is not necessary to determine exactly the shape of the clusters. The clustering algorithm more used for its linear computational complexity is the Fuzzy C-Means algorithm (FCM) [10], a partitive fuzzy clustering method that uses the Euclidean distance to determine prototypes cluster as points.
Let $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$ be a dataset composed of $N$ pattern $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, where $x_{kj}$ is the $k$th component (feature) of the pattern $x_j$. The FCM algorithm minimizes the following objective function:

$$J(X, U, V) = \frac{1}{C} \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2$$  \hspace{1cm} (1)

where $C$ is the number of clusters, fixed a priori, $u_{ij}$ is the membership degree of the pattern $x_j$ to the $i$th cluster ($i = 1, \ldots, C$), $V = \{v_1, \ldots, v_C\} \subset \mathbb{R}^n$ is the set of points given by the centers of the $C$ clusters (prototypes), $m$ is the fuzzifier parameter and, $d_{ij}$ is the distance between the center $v_i = (v_{i1}, v_{i2}, \ldots, v_{in})^T$ of the $i$th cluster and the $j$th vector $x_j$, calculated as the Euclidean norm:

$$d_{ij} = \left\| (x_j - v_i) \right\| = \sqrt{\sum_{k=1}^{n} (x_{kj} - v_{ki})^2}. \hspace{1cm} (2)$$

Using the Lagrange multipliers method for minimizing the objective function (1), we obtain the following solution for the center of each cluster prototype:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}, \hspace{1cm} (3)$$

where $i = 1, \ldots, C$ and for membership degrees $u_{ij}$,

$$u_{ij} = \frac{1}{\left( \sum_{h=1}^{C} \left( \frac{d_{ij}^2}{d_{ih}^2} \right) \right)^{2/(m-1)}} \hspace{1cm} (4)$$

subjected to the constraints:

$$\sum_{i=1}^{C} u_{ij} = 1, \quad \forall j \in \{1, \ldots, N\}, \hspace{1cm} (5)$$

$$0 < \sum_{j=1}^{N} u_{ij} < N, \quad \forall i \in \{1, \ldots, C\}.$$ 

Initially, the $u_{ij}$'s and the $v_i$ are assigned randomly and updated in each iteration. If $U^{(l)} = (u_{ij}^{(l)})$ is the matrix $U$ calculated at the $l$-th step, the iterative process stops when

$$\left\| U^{(l)} - U^{(l-1)} \right\| = \max_{i,j} \left| u_{ij}^{(l)} - u_{ij}^{(l-1)} \right| < \varepsilon, \hspace{1cm} (6)$$

where $\varepsilon > 0$ is a prefixed parameter.

This algorithm has a linear computational complexity; however, it is sensitive to the presence of noise and outliers; furthermore, the number of cluster $C$ is fixed a priori and needs to use a validity index for determining an optimal value for the parameter $C$.

In order to overcome these shortcomings, in [11, 12], the EFCM algorithm is proposed, where the cluster prototypes are hyperspheres in the case of the Euclidean metric. Like FCM, the EFCM algorithm is characterized by a linear computational complexity; furthermore, it is robust with respect to the presence of noise and outliers, and the final number of clusters is determined during the iterative process.

In [13, 14], the authors propose the use of the EFCM algorithm for detecting hotspot areas. The final hotspots are identified as the detected cluster prototypes and shown on the map as circular areas. In [4], the authors analyze the spatio-temporal evolution of the hotspots in the fire analysis. The pattern event dataset is partitioned according to the time of the event’s detection; so each subset is corresponding to a specific time interval. The authors compare the hotspots obtained in two consecutive years by studying their intersections on the map. In this way, it is possible to follow the evolution of a particular phenomenon.

The cluster prototypes detected from EFCM method are circular areas on the map that can approximate a hotspot area. Figure 1 shows an example of two circular hotspots, obtained as clusters.

Figure 1 shows three different regions.

(i) An area in which the hotspot $A$ is not intersected by the hotspot $B$ (corresponding to $A - (A \cap B) = A - B$): this region can be considered as a geographical area in which prematurely detected event disappears successively.

(ii) The region of intersection of the two hotspots $A \cap B$: this region can be considered a geographical area in which the event continues to persist.

(iii) An area in which the hotspot $B$ is not intersected by the hotspot $A$ (corresponding to $B - (A \cap B) = B - A$): this region can be considered as a geographical area in which the prematurely undetected event propagates successively.

We can study the spatio-temporal evolution of the hotspots by analyzing the interactions between the corresponding circular cluster prototypes obtained for consecutive periods,
and detecting the presence of new hotspots in regions previously not covered by hotspots and the absence of hotspots in regions previously spatially included in hotspot areas.

In this research, we present a method for studying the spatio-temporal evolution of hotspots areas in disease analysis; we apply the EFCM algorithm for comparing, in consecutive years, event datasets corresponding to otolaryngopharyngeal diseases diagnosis detected in the district of Naples (I). Each event corresponds to the residence of the patient who contracted the disease.

We study the spatio-temporal evolution of the hotspots analyzing the intersections of hotspots corresponding to two consecutive years, the displacement of the centroids, the increase or reduction of the hotspots areas, and the emergence of new hotspots.

In Section 2, we give an overview of the EFCM algorithm. In Section 3, we present our method for studying the spatio-temporal evolution of hotspots in disease analysis. In Section 4, we present the results of the spatio-temporal evolution of hotspots for the otolaryngologist-laryngopharyngeal diseases diagnosis events detected in the district of Naples (I). Our conclusions are in Section 5.

2. The EFCM Algorithm

In the EFCM algorithm, we consider clustering prototypes given by hyperspheres in the n-dimensional feature's space. The ith hypersphere is characterized by a centroid \( v_i \) and a radius \( r_i \).

Indeed, if \( r_i \) is the radius of \( V_i \), we say that \( x_j \) belongs to \( V_i \) if \( d_{ij} \leq r_i \).

The radius \( r_i \) is obtained considering the covariance matrix \( P_i \) associated with the ith cluster, defined as

\[
P_i = \frac{\sum_{j=1}^{N} u_{ij}^m (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{N} u_{ij}^m}
\]

whose determinant gives the volume of the ith cluster. Since \( P_i \) is symmetric and positive, it can be decomposed in the following form:

\[
P_i = Q_i \Lambda_i Q_i^T,
\]

where \( Q_i \) is an orthonormal matrix and \( \Lambda_i = (\lambda_{ij}) \) is a diagonal matrix. The radius \( r_i \) is given by the following formula (see [12]):

\[
r_i = \frac{1}{n} \sqrt{\prod_{k=1}^{n} \lambda_{ik}^{1/n}} = \sqrt{\det(P_i)^{1/n}}.
\]

The objective function to be minimized is the following:

\[
J(X, U, V) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m (d_{ij}^2 - r_i^2),
\]

where the membership degrees \( u_{ij} \) are updated as

\[
u_{ij} = 1 \times \left( \sum_{h=1}^{C} \left( \frac{(d_{ij}^2 \max (0, 1 - r_i^2/d_{ij}^2))}{(d_{ij}^2 \max (0, 1 - r_i^2/d_{ij}^2)))^{1/(m-1)}} \right) \right)^{-1}
\]

\[
= \frac{1}{\sum_{h=1}^{C} \left( d_{ij}^2 u_{ij}/d_{ij}^2 \right)^{2/(m-1)}}
\]

We set \( d_{ij}^2 = d_{ij}^2 w_{k} \) and define the number \( \varphi_j = \text{card}[k \in \{1, \ldots, C\} : d_{ij}^2 = 0] \) for any \( j = 1, \ldots, N \); thus, we obtain

\[
u_{ij} = \begin{cases} 
0 & \text{if } d_{ij}^2 > 0, \\
1 & \text{if } d_{ij}^2 = 0, \\
\varphi_j & \text{if } \varphi_j > 0.
\end{cases}
\]

However, the usage of (12) produces the negative effect of diminishing the objective function (10) when a meaningful number of features are placed in a cluster and this fact can prevent the separation of the clusters. Then a solution to this problem consists in the assumption of a small starting value of \( \epsilon \) and then it is increased gradually with the factor \( \epsilon^{(l)} / \epsilon^{(l-1)} \), where \( \epsilon^{(l)} \) is the number of clusters at the lth iteration and \( \epsilon^{(l-1)} \) is defined recursively as \( \epsilon^{(0)} = 1, \epsilon^{(l)} = \min(\epsilon^{(l-1)}, \epsilon^{(0)}) \), by setting

\[
I_{ij} = \frac{\sum_{j=1}^{N} \text{min} (u_{ij}, u_{ik})}{\sum_{j=1}^{N} u_{ij}}
\]

and the symmetric matrix \( S = (S_{ik}) \), where \( S_{ik} = \max[I_{ik}, I_{kj}] \) is defined as well. If \( S_{ik} \) is the matrix S at the lth iteration (\( l > 1 \)) and the threshold \( \alpha^{(l)} = 1/(\epsilon^{(l-1)} - 1) \) is introduced as limit, then two indexes \( \alpha^* \) and \( \alpha^* \) are determined such that \( S_{ik}^{(l)} \geq \alpha^{(l)} \) and thus \( \alpha^* \) and \( \alpha^* \) are merged by setting

\[
u_{ij}^{(l)} = u_{ij}^{(l)} + u_{kj}^{(l)}, \quad \forall j \in \{1, \ldots, N\},
\]

\[
C^{(l)} = C^{(l-1)} + 1.
\]

The \( k^* \) th row can be removed from the matrix \( U^{(l)} \). In other words, the EFCM algorithm can be summarized in the following steps:

1. The user assigns the initial number of clusters \( C^{(0)} \), \( m > 1 \) (usually \( m = 2 \)), \( \epsilon > 0 \), the initial value \( S_{ik}^{(0)} = 0 \), and \( \epsilon^{(0)} = 1 \).

2. The membership degrees \( u_{ij}^{(0)} \) (\( j = 1, \ldots, N \) and \( i = 1, \ldots, C(0) \)) are assigned randomly.
3. Hotspots Detection and Evolution in Disease Analysis

Each pattern is given by the event corresponding to the residence of the patient to whom a specific disease has been detected. The two features of the pattern are the geographic coordinates of the residence.

The first step of our process is a geocoding activity necessary for obtaining the event dataset starting by the street address of the patients.

To ensure an accurate matching for the geopositioning of the event, we need the topologically correct road network and the corresponding complete toponymic data.

The starting data include the name of the street and the house number of the patient's residence. After the matching process, each data is converted in an event point georeferenced on the map.

In Figure 2, the road network of the district of Naples is shown; the name of the street is labeled on the map; the events are georeferenced as points on the map.

Figure 3 shows the data corresponding to an event selected on the map.

After geo-referencing each event, the event dataset can be split, partitioning them by time interval. For example, the event in Figure 3 can be split by the field "Year."

For each subset of events, we apply the EFCM algorithm to detect the final cluster prototypes.

In this research, we point out the analysis of the temporal evolution and spread of oto-laryngo-pharyngeal diseases detected within the district of Naples. The datasets, divided by time sequences corresponding to periods of one year, are made up of patterns for different events georeferenced corresponding to ailments encountered in patients for which an intervention and the subsequent histological examination were pointed out as well. The event refers to the geopositioning of the location of the patient.

The data have been further divided by the type of the disease for analyzing the distribution and evolution of each specific disease on the area of the study.

The EFCM algorithm has been encapsulated in the GIS platform ESRI ArcGIS. Figure 4 shows the mask created for setting the parameters and running the EFCM algorithm.

We can set other numerical fields for adding other features to the geographical coordinates.

Initially, we set the initial number of clusters, the fuzzifier m, and the error threshold for stopping the iterations. After running EFCM, the number of iterations, the final number of clusters, and the error calculated at the last iteration are reported. The resultant clusters are shown as circular areas on the map and can be saved in a new geographic layer.

The final process concerns the comparative analysis of the hotspots obtained by the clusters corresponding to each subsets of events.
Figure 5 shows an example of display on the map of hotspots obtained as final clusters for two consecutive subset of events.

In order to assess the expansion and the displacement of a hotspot, we measure the radius of the hotspot and the distance between the centroids of two intersecting hotspots.

In the next section, we present the results obtained by applying this method for the data corresponding to surgical interventions to the oto-laryngo-pharyngeal apparatus in patients residents in the district of Naples between the years 2008 and 2012.

We divide the dataset per year and analyze various types of diseases.

Among the types of the most frequent diseases, the following were analyzed:

(i) carcinoma,
(ii) edema of bilateral Reinke,
(iii) hypertrophy of the inferior turbinate,
(iv) nasal polyposis,
(v) bilateral vocal fold prolapse.

In the next section, we show the most significant results obtained by applying this method to the each partitioned dataset of events.

### Table 1: Edema of bilateral Reinke disease—final number of clusters and final error for year.

| Year | Initial number of clusters | Final number of clusters | $|U^{(l)} - U^{(l-1)}|$ | $\epsilon$ |
|------|-----------------------------|--------------------------|------------------------|------------|
| 2008 | 15                          | 4                        | $0.48 \times 10^{-2}$  | $1 \times 10^{-2}$ |
| 2009 | 15                          | 4                        | $0.55 \times 10^{-2}$  | $1 \times 10^{-2}$ |
| 2010 | 15                          | 4                        | $0.71 \times 10^{-2}$  | $1 \times 10^{-2}$ |
| 2011 | 15                          | 4                        | $0.67 \times 10^{-2}$  | $1 \times 10^{-2}$ |
| 2012 | 15                          | 3                        | $0.53 \times 10^{-2}$  | $1 \times 10^{-2}$ |

### 4. Test Results

We present the results obtained on the event dataset described above in the period between the years 2008 and 2012.

We consider first the subset of data corresponding to the edema of bilateral Reinke disease.

We fix the fuzzifier parameter to 0.1, the initial number of clusters to 15, and the final iteration error to $1 \times 10^{-2}$.

Table 1 shows the results obtained for each year.

We present the details relating to the comparison of the hotspots obtained by considering the event data for the years 2011 and 2012.

Figures 6 and 7 show, respectively, the hotspots obtained by using the pattern subset of events that occurred in the years 2011 and 2012.

Figure 8 shows the overlap of the hotspots obtained for the two years: in red, the hotspots corresponding to the year 2011; in blue, the ones corresponding to the year 2012.

Table 2 shows in the first two columns the labels of the hotspots in 2011 and 2012, in third (resp., fourth) column the radius obtained in 2011 (resp., 2012), and the distance between the centroids is given in the fifth column.

The results show that only hotspot 3 obtained for the year 2011 remains almost unchanged in the year 2012. Instead, hotspots 1 and 2 seem to merge into a single larger hotspot (the hotspot 1 obtained for the year 2012), and hotspot 4, that shifts about 1 km, is expanded; the radius of this hotspot in 2012 is about 6.5 km (hotspot 3 obtained for the year 2012 in Figure 8).
Table 2: Edema of bilateral Reinke disease—comparison results of the hotspots obtained for the years 2011 and 2012.

<table>
<thead>
<tr>
<th></th>
<th>Intersecting hotspot</th>
<th>Radius 2011 hotspot (km)</th>
<th>Radius 2012 hotspot (km)</th>
<th>Centroid’s distance (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.724</td>
<td>3.848</td>
<td>1.759</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.943</td>
<td>3.848</td>
<td>1.507</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.434</td>
<td>3.453</td>
<td>0.074</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.591</td>
<td>6.519</td>
<td>1.115</td>
</tr>
</tbody>
</table>

Table 3: Nasal polyposis—comparison results of the hotspots obtained for the years 2011 and 2012.

<table>
<thead>
<tr>
<th></th>
<th>Intersecting hotspot</th>
<th>Radius 2011 hotspot (km)</th>
<th>Radius 2012 hotspot (km)</th>
<th>Centroid’s distance (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3.087</td>
<td>4.951</td>
<td>2.656</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4.915</td>
<td>7.103</td>
<td>1.052</td>
</tr>
</tbody>
</table>

Now we show the results obtained for the disease nasal polyposis.

Figure 9 shows the overlap of the hotspots obtained for the two years, 2011 and 2012.

In Table 3, the comparison’s results are reported.

The results in Figure 9 show that in 2011 and 2012 there are two hotspots: the one covering an area of the city of Naples and the other covering many Vesuvian towns. The two hotspots, which in 2011 covered a circular area with a radii of about 3 and 5 km, respectively, in 2012 cover a circular area with radii of about 5 and 7 km, respectively.

The histogram in Figure 10 shows the trend of the radii of the two hotspots in the course of time.

It is relevant the spread in recent years of the hotspot that surrounds the Vesuvian towns (the radius of this hotspot, from about 2 km in the year 2008, is about 7 km in the year 2012).

Another significant trend concerns the hotspots obtained for the carcinoma disease.

Also, in this case, the two main hotspots cover the city of Naples and many Vesuvian towns. In in this case, we have a very high spread of the hotspot covering the city of Naples (cfr. Figure 11); in recent years, the radius of this hotspot is increased up to 9.5 km.

5. Conclusions

The hyperspheres obtained as clusters (circles in case of two dimensions) by using EFCM can represent hotspots in hotspot analysis; this method has a linear computational complexity and is robust to noises and outliers. In hotspots analysis, the patterns are bidimensional and the features are formed by geographic coordinates; the cluster prototypes are circles that can represent a good approximation of hotspot areas and can be displayed as circular areas on the map.

In this paper, we present a new method that uses the EFCM algorithm for studying the spatio-temporal evolution of hotspots in disease analysis.
We consider the residence's information of patients in the district of Naples (Italy) to whom a surgical intervention to the oto-laryngo-pharyngeal apparatus was carried out between the years 2008 and 2012. A geocoding process is used for geo-referencing the data; then, the georeferenced dataset is partitioned per year and type of disease; we compare the hotspots obtained for each pair of consecutive years and analyze the trend of each hotspot over time measuring the variation of the radius and the distance between intersecting cluster centroids concerning two consecutive years.

The results show a consistent spread in the last years of the nasal polyposis disease hotspot covering some Vesuvian towns and of the carcinoma disease hotspot covering the city of Naples.

References


