Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2010/11

Spatial data mining for clustering: from the literature review to an application using RedCap

Federico Benassi, Chiara Bocci,
Alessandra Petrucci

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Spatial data mining for clustering: from the literature review to an application using RedCap

F. Benassi, C. Bocci, A. Petrucci
Department of Statistics "G. Parenti"
University of Florence

1. Introduction

The aim of the paper is both to review the scientific literature about spatial data mining methods - in particular spatial clustering methods developed in recent years - and to present an original application of the recently proposed RedCap[1] method (Guo, 2008) of spatial clustering and regionalization on Florentine Metropolitan Area (FMA). Demographic indicators computed on official data provided by the Italian Institute of Statistics (Istat), are the input of a spatial clustering and regionalization model in order to get a classification of the FMA municipalities into a number of homogeneous (with respect to demographic structure) and spatially contiguous zones.

In the optics of a progressive decentralization of the governance activities we believe that the FMA represents a very interesting case of study. This due to the fact that the individuation of new spatial areas built considering both the demographic characteristics of the resident population and the spatial dimension of the territory where this population insists could become a useful tool for local governance.

The paper is structured as follows. In section 2 we present some theoretical considerations about data mining and spatial data mining. In section 3 we analyze some of the most important methods of spatial clustering. In section 4 we briefly describe regionalization process and RedCap's major features. In section 5 we describe the FMA and present the results of the empirical application. In section 6 we propose some final conclusions.

[1]Regionalization with dynamically constrained agglomerative clustering and partitioning (Guo, 2008).

2. Data mining and spatial data mining: some theoretical considerations

For several years spatial data mining has been considered as the multi-dimensional equivalent of temporal data-mining (Roddick and Spiliopoulou, 1999) but today researchers agree to consider spatial data mining as an independent approach to analyzing data and measuring phenomena as confirmed by recent studies (Angayarkkani and Radhakrishnan, 2009; Behnisch and Ultsch, 2009, 2010; Jin and Guo, 2009; Moran and Bui, 2002; Szalay et al., 2000; Yu Pan and Faloutsos, 2002).

Before define spatial data mining it could be useful to explain what classic data mining is. Data mining is a relatively young discipline concentrating on the manipulation of extensive data bases. Lots of definitions have been elaborated by researchers belonging to different disciplines like mathematics, computer sciences, statistics and so on. One of the first points which should be define is the data mining's main goal. According to Miller and Han (2001), data mining searches for deeply hidden information that can be turned into knowledge of strategic decision making and answering fundamental research questions. Shekhar and Chawla (2003) give us more details about data mining process defining it as the process of discovering interesting and potentially useful patterns of information embedded in large data bases. Beginning from these basic considerations and following the theoretical formulation of Koperski et al. (1996), we can say that spatial data mining is a knowledge discovery process of extracting implicit interesting information, spatial relations, or other spatial pattern not explicitly stored in spatial databases.

Clarified these key points, the main differences between data mining and spatial data mining are straightforward. From a theoretical and conceptual point of view these are the same that exist between classic and spatial statistical analysis and this is due to the fact that data mining directly derives from the statistical science (Hoskin et al, 1997). It's well known that one of the most important assumptions of classical statistical analysis is that the data samples are independently generated; on the contrary, the spatial approach refuses this assumption and theorizes that the spatial location of the samples is an item that cannot be ignored (Tobler, 1970). Thus, it follows that data mining is connected to the concept of patterns while spatial data mining is connected to the concept of spatial patterns (Shekhar and Chawla, 2003). Obviously, these theoretical differences between classic and spatial data mining, have important repercussions in

operative terms. Apply a spatial (data mining) approach implies that the dimension of large databases become larger as spatially referenced objects also carry information concerning their representation in space by geometrical and topological properties (Koperski et al., 1996). This implies the following needs: more powerful techniques to manipulate data and extract knowledge; new kind of cartographic knowledge to represent the results obtained and make them readable to a non technical people (policy makers, local administrators etc.); finally, more flexible software in order to encourage the human interaction with the data.

Summarizing, the term "spatial data mining" refers to the search of spatial patterns into spatial (or geographical) databases. Necessarily this search must have some specific characteristics that allow us to define it as a knowledge discovery process. In particular, it should be non trivial and with the highest number of automated operations as possible in order to reduce to a minimum level the human efforts. Besides, the funded spatial patterns should be, with the regard to the research objectives, interesting, useful and unexpected (Shekhar and Chawla, 2003).

Nonetheless, there is no unique way to classifying data mining and spatial data mining methods and techniques. Han (1999) divides general (or classic) data mining into two main categories: descriptive data mining and predictive data mining. Descriptive data mining describes the behavior of data sets and presents interesting general properties of data while predictive data mining attempts to construct models in order to help predicting the behavior of the new datasets. Another important distinction that regards more specifically the statistical component of mining process is about the pattern recognition. In that case, researchers tend to distinguish between statistical pattern recognition – where everything is learnt from observations - and structural pattern recognition – where the most of the structure is imposed from a priori knowledge (Ripley, 1996). Others classify the mining process on the base of the analysis approach applied. Berry and Linoff (1997) define two approaches of analysis: top-down (confirmative) and bottom-up (explorative). The first approach, using mainly traditional statistics methods and techniques, tries to confirm or refuse some hypothesis by finding new aspects of a phenomenon not completely unknown. The second approach tries to find some unexpected information of a relatively unknown phenomenon. Another distinction is between unsupervised and supervised methods. In

the first case, there is no know grouping of the observations while in the second case the observations are known to be grouped in advance and the task is to group or to predict future observations (Ripley, 1996). Finally, we can find a great number of classifications with regards to spatial data mining techniques and methods. Ester et al. (1997) divide them into four general groups: spatial association rules, spatial clustering, spatial trend detection and spatial classification. Shekhar and Chawla (2003) define three non controversial techniques of spatial data mining: classification, clustering and association rules. Our attention is focused on spatial clustering techniques that we will analyze in the next paragraph.

3. Spatial clustering: an overview on methods and techniques

Spatial clustering is a process of grouping a set of spatial objects into meaningful subclasses (that is clusters) so that the members within a cluster are similar as much as possible whereas members of different clusters differ as much as possible from each other (Jiao and Liu, 2008). The spatial data mining role is to scale a spatial clustering algorithm to deal with the large geographical datasets (Shekhar and Chawla, 2003).

Spatial clustering algorithms and approaches can be separated into four general categories: partitioning method, hierarchical method, density-based method and grid-based method. This categorization is essentially based on the classification proposed Han et al. (2001, 2009). Following this, we propose a description of the most important spatial clustering methods.

Partitioning methods

The partitioning approach characterized early studies in clustering and still remain one of the most cited and used approach. A partitioning algorithm organizes the objects into clusters such that the total deviation of each object from its cluster center is minimized. The deviation of a point can be computed in different ways and is usually called similarity function (Han et al., 2001, 2009). At the beginning of the process each object is classified as a single cluster. In the following steps, all data points are iteratively reallocated to every clusters until a stopping criterion is met. Partitioning algorithms organize the objects into *k* clusters such that the similarity function of each object, with respect to its cluster representative, is minimized. A cluster representative

could be the cluster centre, the most centrally located object in the cluster or a probability distribution representing the cluster. Usually, similarity function correspond to Euclidean distance so minimize the similarity function between a given object and its cluster representative means to associate the object to the cluster having the closest representative (Varlaro, 2008). A great numbers of algorithms that belong to this method have been developed, nonetheless all of them can be considered as derivations of three basic algorithms: K-means (McQueen, 1976), K-medoid (Vinod, 1969; Kaufman and Rousseeuw, 1990) and EM–expectation maximization (Dempster et al., 1977; Yu et al., 1998; Bradley et al., 1998). K-means and k-medoid algorithms are very similar since in both methods the cluster is represented by a centrality measure that becomes the gravity center of the cluster (centroid or mean) in the first case and the most centrally located object in the cluster in the second (Han et al., 2001). On the contrary, EM clustering algorithms use a distribution consisting of a mean and a covariance matrix to represent each cluster so, instead of assigning each object to a dedicated cluster, these algorithms assign each object to a cluster according to a probability of membership which is computed from the distribution of each cluster (Han et al, 2001). From a more general point of view it's important to underline that every kind of partition methods is equivalent to a Voronoi diagram and each cluster is contained in one of the Voronoi polygons thus these methods tend to find clusters of spherical shape which is relatively restrictive for many applications (Sander et al., 1988; Shekhar and Chawla, 2003).

Like Han et al (2009) show, the objective criterion used in the K-Mean algorithm is typically the squared error function defined as:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2$$

where E is the sum of the square-error for all objects in the data set; x is the point representing a given object and m_i is the mean of cluster C_i (both x and m_i are multidimensional). The K-mean algorithm is relatively efficient in processing large data bases but presents important limitations. As mentioned above, it can't discover clusters with non convex shapes or clusters of very different size. It is also sensitive to noise and outlier data points (Han et al., 2001).

To reduce the limitations of clustering methods base on K-mean algorithms the K-medoid algorithms have been elaborated. In this case, instead of taking the mean value of the objects in a cluster as reference point, an actual object is picked to represent each cluster. As mentioned above this representative object is the medoid - the most centrally located object within the cluster - while each remaining object will be clustered with the representative object to which is the most similar (Han et al., 2009). The principle is to minimize the sum of dissimilarities between each object and its corresponding reference point in terms of an absolute error criterion defined as:

$$E = \sum_{j=1}^{k} \sum_{x \in C_j} |x - m_j|$$

where E is the sum of absolute error for all objects in the data set; x is the point representing a given object in cluster C_j and m_j is the representative object of C_j. The major limitation of this spatial clustering algorithms is that their computational dimension become very costly with large data sets (Han et al., 2009).

In the case of EM–algorithms each cluster can be represented mathematically by a parametric probability distribution. The data are clustered by using a finite mixture density model of M probabilistic distributions where each distribution represents a cluster. A mixture model can be formalized as:

$$P(x|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(x|\Theta_i)$$

where the parameters are $\Theta = (\alpha_1, \dots, \alpha_M, \Theta_1, \dots, \Theta_M)$ such that $\sum_{i=1}^{M} \alpha_i = 1$, and each p_i is a density function parameterized by Θ_i. That is, M component densities are mixed together with M mixing coefficients α_i (Han et al., 2009). The log-likelihood expression for this density from observed data x is given by:

$$log\big(\mathcal{L}(\Theta|X,Y)\big) = log\big(P(X,Y|\Theta)\big) = \sum_{i=1}^{N} log\left(P(x_i/y_i)P(x_i)\right)$$

$$= \sum_{i=1}^{N} \log\Big(\alpha_{yi} p_{yi}\big(x_i|\Theta_{yi}\big)\Big)$$

where we posit the existence of unobserved data $Y = \{y_i\}_{i=1}^{N}$ whose values inform us which component density "generated" each data item. For example, if the i-th element

was generated by the k-th mixture component, $y_i = k$ $(y_i \in \{1, \ldots, M\})$ (Han et al., 2009). The goal is to find Θ that maximize $\mathcal{L}$ and to achieve that the EM-algorithm is used. The EM-algorithm is in fact a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete. It can be viewed as an extension of the k-means paradigm because, instead of assigning each object to a dedicated cluster, it assigns each object to a cluster according to a weight representing the probability of membership (Han et al., 2009).

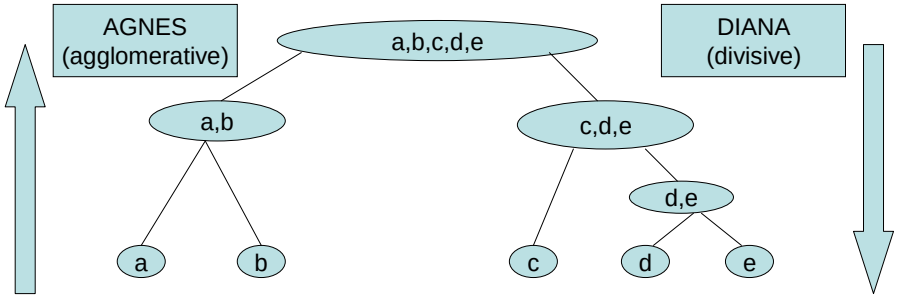
Some of the recent algorithms that are based on the partitioning method are: Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 1990), Clustering LARge Applications (CLARA) (Kaufman and Rousseeuw, 1990) and Clustergin LARge Applications based upon RANdomized Search (CLARANS) (Ng and Han, 1994) (Han et al., 2001, 2009; Varlaro, 2008).

Hierarchical methods

These clustering methods hierarchically decompose the spatial dataset by splitting or merging all clusters until a stopping criterion is met. The result of the decomposition is a dendrogram of spatial objects, that is a tree structure where each no-leaf node is composed by the same elements composing its children nodes. In this way, the result of a clustering task is not a partition of dataset but a hierarchy of clusters where each level describes a partition of source data (Varlaro, 2008). The dendrogram can either be created from leaves up to the root (agglomerative hierarchical clustering) or from the root down to the leaves (divisive hierarchical clustering) (Sander et all., 1998). The agglomerative hierarchical approach, also called bottom-up approach, starts with each object forming a separate group. At every interaction the two most similar clusters (according to the considered similarity function) are merged together into a new cluster until all of the objects are in a single cluster or until a termination condition holds (Han et al. 2001). The divisive hierarchical approach, also called top down approach, starts by considering each object belonging to the same general cluster and, at each iteration, one of the available clusters is split into smaller clusters according to some measure, until each object is in one cluster or a termination condition holds (Varlaro, 2008). AGNES (Agglomerative Nesting) is one of the first agglomerative

hierarchical algorithm while DIANA (Divisive analysis) is one of the first divisive algorithm (Han et al., 2001, 2009).

We propose now an example using AGNES and DIANA algorithms elaborated by Han et al. (2009).



We have a data set of 5 objects $\{a, b, c, d, e\}$. Initially, AGNES places each object into a cluster of its own. At each stage, the algorithm joins the two clusters that are closest together (i.e., most similar). The cluster merging process repeats until all of the objects are eventually merged to form one cluster. DIANA does the reverse of AGNES. The measures for the distance between two objects or points are formulated by the following equations:

$$Minimum\ distance: d_{min}(C_i, C_j) = min_{p \in Ci, p' \in Cj} |p - p'|$$

$$Maximum\ distance: d_{max}(C_i, C_j) = max_{p \in Ci, p' \in Cj} |p - p'|$$

$$Mean\ distance: d_{mean}(C_i, C_j) = |m_i - m_j|$$

$$Average\ distance: d_{avg}(C_i, C_j) = min_{p \in Ci, p' \in Cj} |p - p'|$$

where $|p - p'|$ is the distance between two objects or points, p and p'; m_i is the mean for a cluster, C_i and n_i is the number of objects in C_i.When an algorithm uses the minimum distance to measure the distance between two clusters it is called single-linkage algorithm; on the contrary, when an algorithm uses the maximum distance it is called complete-linkage algorithm (Han et al., 2009). It is to note that, differently from the mean distance, minimum and maximum distances tend to be very sensitive to overlay or out noise data.

In contrast to partitioning algorithms, hierarchical algorithms do not need *k* as an input parameter. However a termination condition has to be define indicating when the merge or division process should be terminated. Alternatively, an appropriate level in the dendrogram has to be selected manually after the creation of the whole dendrogram.

A limitation of the hierarchical algorithms is that they are computationally intensive: as a pair wise confrontation of all the objects is required at each step, the computational dimension grows fast.

Some of the recently used hierarchical clustering algorithms are Balancend Iterative Reducing and Clustering using Hierarchies (BIRCH) (Zang et al., 1996), Clustering Using REpresentatives (CURE) (Guha et al., 1998), a Hierarchical Clustering Algorithm using Dynamic Modeling (CHAMELEON) (Karypis et al., 1999).

Density – based methods

These kind of methods has been developed to overcome the major limitations of clustering methods based on the distance between objects (i.e. partitioning and hierarchical methods). In fact, with these new family of methods based on the concept of density, clusters of arbitrary shapes can be discovered (Han et al., 2001). In the density-based methods clusters are regarded as dense regions, that is regions characterized by an high number of spatial objects. These dense regions are separated each other by regions of low density and constraining noise (Han et al., 2009). Despite their properties, density-based methods find difficulties when the number of dimension is high (Varlaro, 2008).

One of the most famous density-based clustering methods is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). The algorithm of this method grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise (Han et al, 2009). Another two important density-based methods are the Ordering Points to Identify the Clustering Structure (OPTICS) (Ankerst et al., 1999) and the Clustering Based on Density Distribution Functions (DENCLUE) (Hinneburg and Keim, 1998). OPTICS was proposed to overcome the limitation that DBSCAN algorithm presents in terms of wasted time due to its sensitivity to the input parameters. OPTICS, produces an ordering of the data points such that clustering result for any lower or similar value compare to the two input parameters can be visualized and computed easily. DENCLUE is a clustering method based on a set of density distribution functions. The method is based on some basic ideas: 1) the influence of each data point can be formally modeled using a mathematical function, called influence function, which describes the impact of a data

points on its neighborhood; 2) the overall density of the data space can be modeled analytically as the sum of the influence function of all data points; and 3) clusters can then be determined mathematically by identifying density attractors, where density attractors are local maxima of the overall density function (Han et al., 2001).

Grid-based methods

Using a grid based structure, the grid–based clustering methods overcomes the limitations of density based method mentioned above. In particular, in the grid-based methods data are quantized into a finite number of cells that represent the elements that will be clustered (Varlaro, 2008).

The main advantage of that approach is its fast processing time, since the time is independent on the number of data objects, but dependent on the number of cells (Han et al, 2001). Some of the most important grid-based algorithms are the STatistical Information Grid (STING) (Wang et al., 1997), which explores statistical information stored in the grid cells; Clustering In QUEst (CLIQUE) (Agrawal et al., 1998), which represents a grid and density-based approach for clustering in high dimensional data space; A Multi-Resolution Clustering Approach for Very Large Spatial Databases (WAVECLUSTER) (Sheikholeslami et al., 1988), which represents a grid-and density based approach for clustering in high-dimensional data space (Han et al., 2001).

4. Regionalization with dynamically constrained agglomerative clustering and partitioning (RedCap): a brief description

According to Guo (2008) we define regionalization as a process that divides a large set of spatial objects into a number of spatially contiguous regions while optimizing an objective function, which is normally a homogeneity (or heterogeneity) measure of derived regions. Therefore regionalization is a special kind of spatial clustering where the condition of spatial contiguity among spatial objects plays a priority role. As Guo (2008) suggests, existing regionalization methods can be classified in four groups: 1) non-spatial clustering followed by spatial processing; 2) non–spatial clustering with a spatially weighted dissimilarity measure; 3) trial-and-error search and optimization; 4) spatially constrained and partitioning. In the first case, the clusters are derived only regard to an attribute similarity an then the clusters are divided or merged

to regions in a geographic space (Guo, 2008). In the second case the algorithm modifies the similarity measure to incorporate spatial information explicitly using some measures like distance-weighted attribute similarity or treating geographic coordinates as additional variables (Guo, 2008). In the third case the algorithm takes a trial-and-error optimization approach and, finally, in the fourth case the algorithm considers spatial constraints with a non-spatial clustering method (Guo, 2008).

RedCap is a new method of spatial clustering and regionalization elaborated by Guo (2008). It is essentially based on a group of six methods for regionalization which are composed by the combination of three agglomerative clustering methods (Single Linkage clustering, SLK; Average Linkage clustering, AVG; Complete Linkage clustering, CLK) and two different spatial constraining strategies: First-Order constraining and Full-Order constraining (Guo, 2008). Referring to the work of Guo (2005, 2008) for technical and computational details about these six methods of regionalization we just briefly describe the theoretical context in which RedCap have to be collocated and how RedCap works in the case, applied in our analysis, of CLK-Full Order method.

Existing methods for multivariate spatial analysis and clustering span a continuum between computational and visual approaches. The first ones exploit the computational power and the formalism of statistical inference to search patterns while the second ones capitalize the ability of human vision to identify patterns and facilitate this process by presenting the data from different perspectives (Guo, 2005). Unfortunately, the historical development of these two methods for multivariate spatial analysis has proceeded independently as underlined by Guo (2005, 2006).

RedCap represents an integrated geographic discovery environment that is able to detect multivariate spatial patterns with high-dimensional geographic data; support human interactions to examine and explain the patterns; create new regionalization that minimize heterogeneity among clusters and at the same time satisfy the condition of spatial contiguity among them. The architecture of RedCap presents two fundamental steps: in the first step the method finds spatial clusters without imposing any spatial constraining strategy; in the second step the method completes the regionalization process. The results of these two steps are related and visualized on an interactive map.

The first step is based on the iterative algorithms of the Self Organizing Map (SOM) (Kohonen, 2001) that represents an intermediate approach between visual and computational approaches. The SOM projects high-dimensional data to a low-dimensional space with preserving nonlinear relationship by producing a similarity graph of the input data. The SOM result is visualized using two types of hexagons: 1) node hexagons, each of which contains a circle that is scaled to depict the number of data items in the node; 2) distance hexagons, each of which is shaded to represent the multivariate distance between two neighboring hexagons. This kind of graphic display of the SOM result is called U-matrix (unified distance matrix) (Kohonen, 2001). RedCap presents also a method of encoding patterns with colors utilizing systematic variation in both hue and lightness to construct a 2D array of logically ordered but discriminable colors (Guo, 2005).

The second step of analysis is based on a contiguity matrix and a set of constrained strategies that drive the agglomerative clustering method. The Complete Linkage clustering method (CLK) defines the distance between two clusters as the dissimilarity between the furthest pair of data points (Guo, 2008):

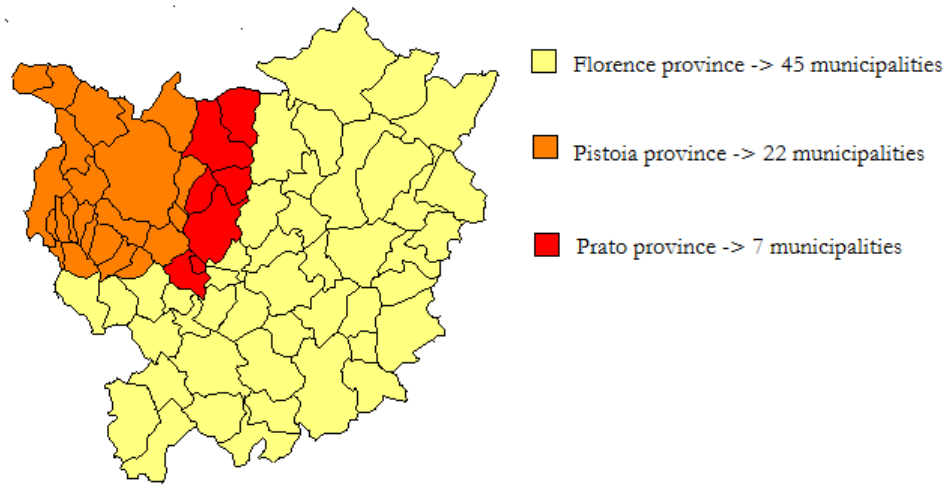
$$d_{CLK}(L, M) = max_{u \in L,\ v \in M}(d_{uv})$$

where L and M are two clusters, $u \in L$ and $v \in M$ are two data points and d is the dissimilarity between *u* and *v*. At the beginning of the process, each individual data is a cluster by itself then original data are updated with the information of the SOM algorithms and the most similar (given by the distance definition) pair of clusters are selected and merged into one. The merging process incorporates the contiguity constraints using the Full Order constraining strategy (Guo, 2008). Contiguity-constrained agglomerative clustering requires that two clusters cannot be merged if they are not spatially contiguous. This is the differential element between classic spatial clustering and regionalization. A Full-Order constraining strategy includes all edges in the clustering process, and the distance between two clusters is defined over all edges. This strategy is dynamic because it updates the contiguity matrix after each merge to track all edges which connect two different clusters (Guo, 2008).

5. Application and Results[2]

The Florentine Metropolitan Area (FMA) is defined by the deliberation of the regional council of Tuscany n.130 on 13/2/2000. This area is composed by three provinces (Firenze, Pistoia and Prato) divided in 73 municipalities (Fig.1). Due to its recent definition a few studies on its population structure and dynamic are available in the literature (Petrucci et al., 2008; Petrucci et al., 2006; Vignoli et al., 2007) and, in particular, there are no studies that consider directly the spatial dimension in the analysis. The FMA is a very heterogeneity area in terms of demographic structures and dynamics, settlements models, geo-morphological structures and economic specialization. In addiction the FMA is strongly affected by many phenomena of mobility: residential migrations, daily migrations (commuting), international migrations. Due to this strong heterogeneity and in the optics of a progressive decentralization of the activities of governance we believe that FMA represents a very interesting case of study.

Figure 1. Florentine Metropolitan Area



We select as input variables five demographic indexes (computed for each municipality of FMA) plus the spatial attributes of each municipality. The five demographic indexes, computed by using data on resident population produced by Italian National Institute of Statistics (Istat), are:

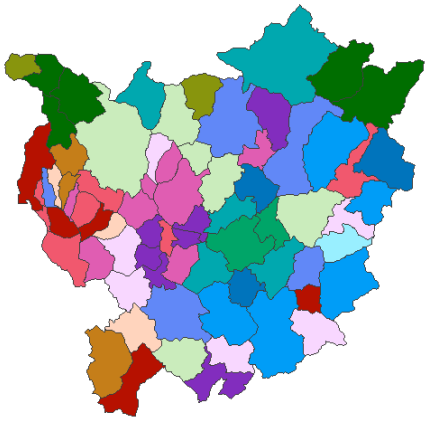
[2] These results were presented to the Joint Meeting GfKl – CLADAG 2010, held in Florence 8-10 September 2010.

Youth dependency index (IDG)	*Aging index* (IV)	*Elderly dependency index* (IDA)	*Population in active age substitution index* (IS)	*Population in active age replacement index* (IR)
$\frac{P_{0-14}}{P_{15-64}} * 100$	$\frac{P_{>64}}{P_{0-14}} * 100$	$\frac{P_{>64}}{P_{15-64}} * 100$	$\frac{P_{40-64}}{P_{15-39}} * 100$	$\frac{P_{60-64}}{P_{15-19}} * 100$

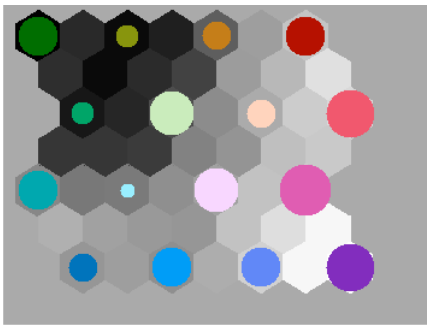
Firstly, an explorative analysis based on the results of the SOM algorithms is carried out. Applying a visual approach we build groups of similar clusters through the results of clustering process visualized on the unified distance matrix without taking into account the condition of spatial contiguity among them.

Fig. 2 General results (without constraining strategy)

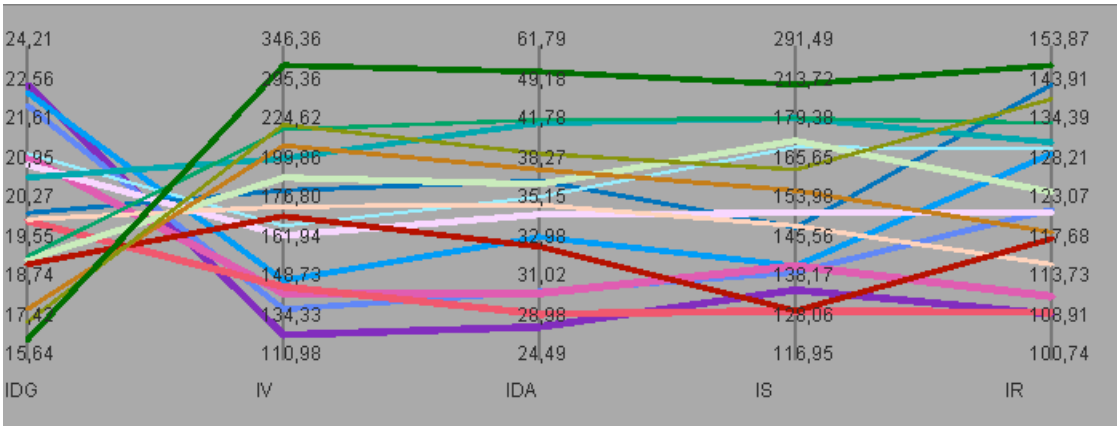
a. Multivariate Mapping



b. Clustering with SOM



c. Multivariate visualization of clusters (Parallel Coordinate Plot)

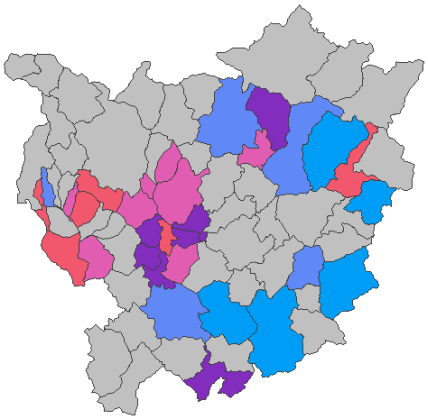


Starting from 73 municipalities we identify 16 clusters that are defined by the node hexagons on the SOM (Fig.2b). Territorial units with the same color belong to the same cluster and clusters with similar colors present a low level of dissimilarity. The level of similarity is measured and visualized by the Parallel Coordinate Plot (PCP) (Fig.2c). In this plot we can observe the profile of each cluster. More in detail, we have five parallel axes (one for each index) scaled by nested means method (Guo, 2005). What is important to compare is the profile of each segment (that is to say each cluster) comparing to the central value of each axe that is the value of that index computed for the total area.

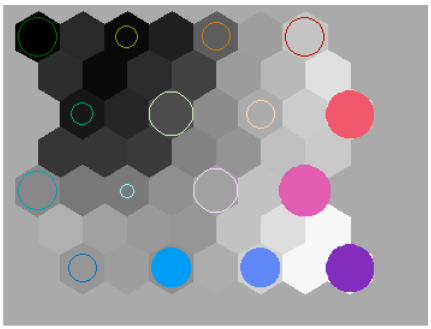
On the base of the results showed in Fig.2, we classify the 16 clusters in three main groups. The first group, that we define "young", is composed by 5 clusters and 29 municipalities (Fig.3). This group has a relatively young structure and a high level of

Fig. 3 Group 1: "Young"

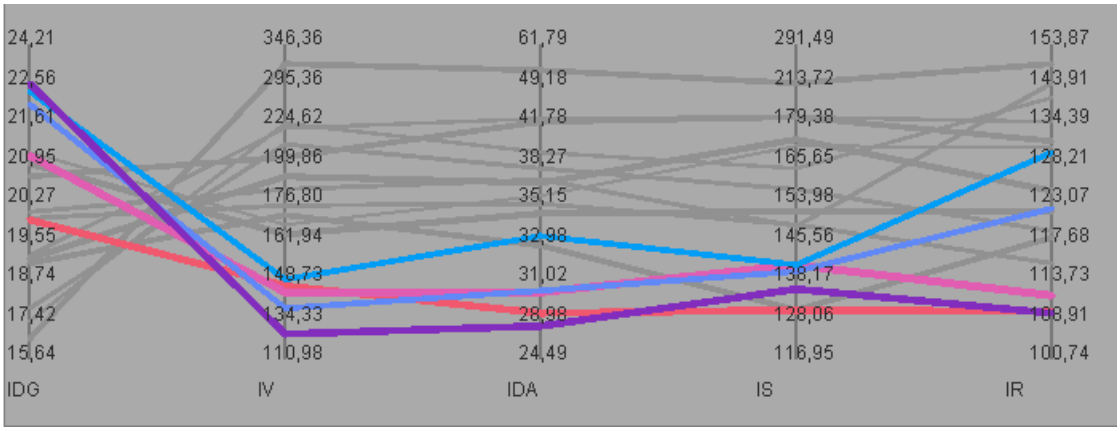
a. Multivariate Mapping



b. Clustering with SOM



c. Multivariate visualization of clusters (Parallel Coordinate Plot)

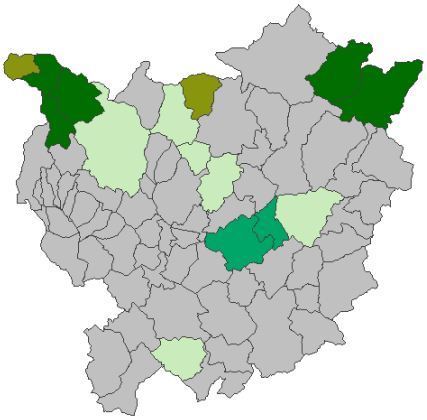


inner homogeneity as the colors of the node hexagons and the PCP show. The 5 clusters present a low level of the IV, IDA and ISPA indexes, a high/medium level of IDG and, finally, a low/medium level of IRPA index (Fig.3c).

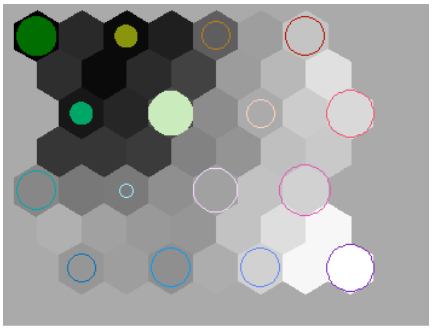
The second group, that we define "old", is composed by 4 clusters and 15 municipalities. This second groups of clusters is characterized by a population with a very old structure as the PCP shows clearly (Fig.4c). In fact, the level of IDG is low while the level of the others indexes (IV, IDA, IS, IR) is medium/high in one case and very high in the others three cases. This group of clusters presents a great level of inner homogeneity as the colors of node hexagons in SOM and the profiles of the PCP show (Fig.4).

Fig. 4 Group 2: "Old"

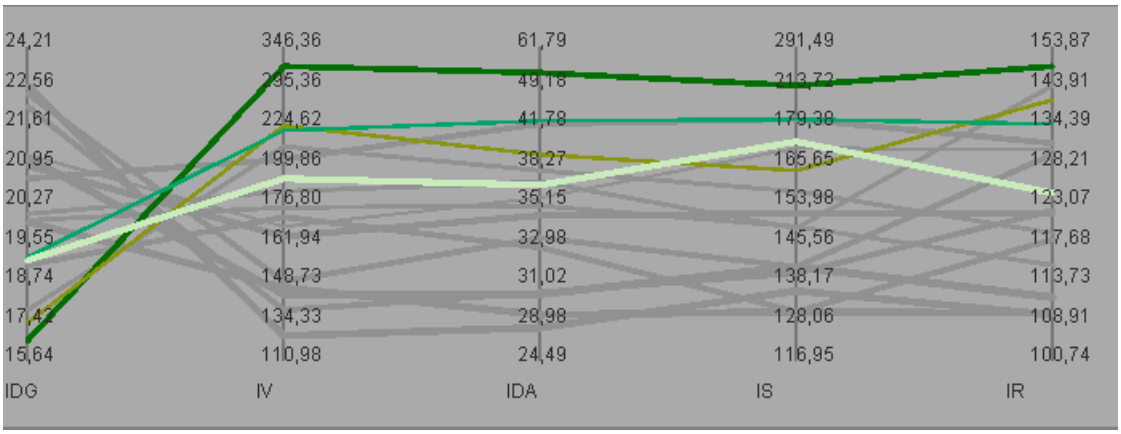
a. Multivariate Mapping



b. Clustering with SOM



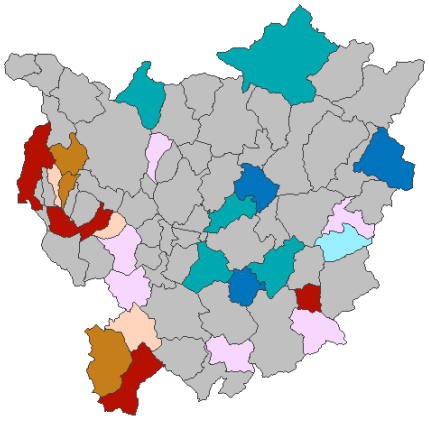
c. Multivariate visualization of clusters (Parallel Coordinate Plot)



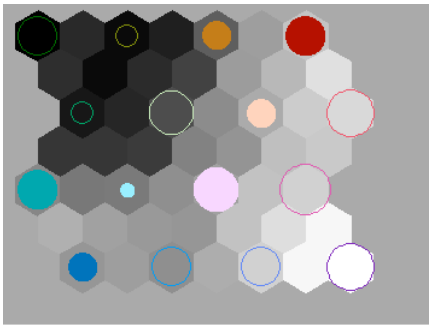
The third group of clusters, that we define “medium”, is composed by 6 clusters and 26 municipalities. As we can see in the Fig.5 this third group is composed by clusters that present a medium level of all indexes. The inner homogeneity of this group is relatively low as we can see by the colors of the node hexagons of the SOM (Fig.5b) and in the profile of the clusters represented in the PCP (Fig.5c). Actually, this group could be divided in two sub-groups: “medium high” (individuated by the blue nodes) and “medium low” (individuated by the pink-brown nodes).

Fig. 5 Group 3: “Medium”

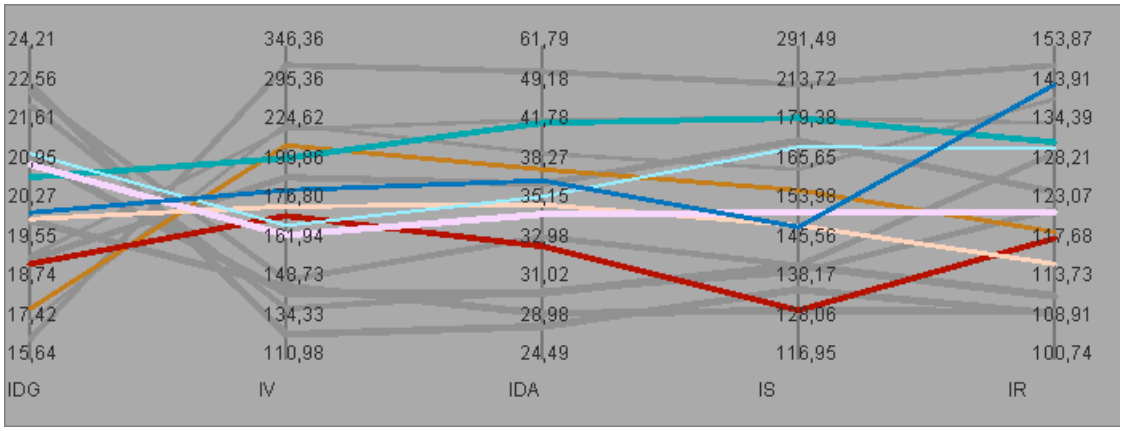
a. Multivariate Mapping



b. Clustering with SOM



c. Multivariate visualization of clusters (Parallel Coordinate Plot)



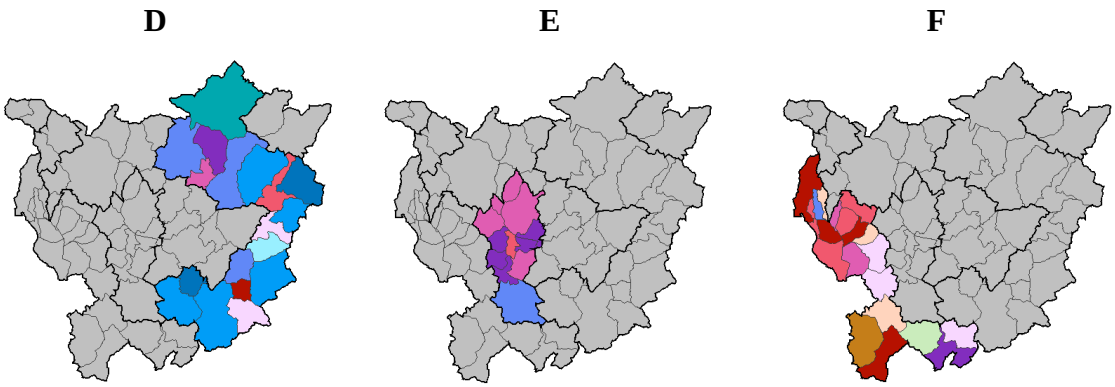
After the explorative analysis we apply the Complete Linkage Clustering method together with a Full Order constraining strategy (CLK-Full Order) in order to obtain *n* area that, given the condition of spatial contiguity, minimize the inner heterogeneity regarding to the demographic structure of the population. We individuate

six areas (Fig.6 and Fig.7). Areas A and B, are very similar to each other and both are characterized by a high level of inner homogeneity. In fact, by a visual analysis we can clearly see that the municipalities that belong to these areas have basically the same colors (Fig.2a). Therefore, areas A and B present a very old structure of their resident population as we can verify by observing the PCP of the "old" group of clusters (Fig.4c). The similarity between areas A and B is not only in terms of demographic structure: these areas are in fact mountain areas characterized by rural settlements and both are localized on the boundary of the FMA. The fact that these areas present a very old demographic structure is probably connected to their recent demographic history. As known, these areas (like the majority of the mountain areas of Italy) were interested by a depopulation process caused by a strong internal rural-urban migration flows. The actors of these internal migration movements were mainly young people and young couples in search of a more modern and dynamic society (typically urban) where the opportunity of study and employment were higher. For similar reasons, these areas are not become destination areas for international migrants (especially for international labor migrants) that usually prefer areas where the labor market is dynamic and characterized by a high incidence of informal sectors (typically the urban and peri-urban areas). Most probably this double migratory mechanism in addition to the aging process that involve the whole Italian population, have determined the extremely old structure of the population of these two areas.

Figure 6 – Regions A, B and C

Area C presents a medium level of inner homogeneity as indicated by the different shades of colors of the municipalities that belong to this area (Fig.6). The demographic structure of the population of this area is medium old: some municipalities (the green ones) have an old structure; others (the pink and the brown) have younger demographic structure. Area C is composed by two important urban centers, Firenze and Pistoia, and by a peri-urban area around these two cities. Therefore, we can say that area C presents an urban settlements and a medium old demographic structure but with an internal spatial structure that can be divided in two sub-structures: a more properly urban area – the core of FMA - with a relatively old demographic structure and a peri-urban area with a relatively younger structure. These results, especially with regards to Firenze and Pistoia, find confirmation in the evidences of Petrucci et al. (2008). By a theoretical point of view this demographic situation can be explained by the following considerations: a) in recent years Firenze and Pistoia (like the majority of medium and large size cities of Italy) were involved in the suburbanization process. Typically this process involves mainly people with a relatively old age structure that move themselves in search of areas less urbanized and with a higher quality of life (Benassi et al., 2009). In the case of Pistoia and, especially, Firenze this process probably involves also young people and young couples that decide to leave parental home and, due to the extremely asymmetric structure of the house market that characterize these cities, they probably migrated to less central areas; b) the high cost of life and houses in Firenze and Pistoia affects also the residential choice of international migrants; c) the urban way of life is typically associated with relatively low level of Total Fertility Rate.

Figure 7 – Regions D, E and F



According to the core-ring model for the study of urban development (Van den Berg, 1992) areas D and F (Fig.7) can be defined as a ring around the more properly urban area. Area D presents a medium/old demographic structure and a medium inner homogeneity as the different colors of the municipalities of this area underline. This area is characterized by a semi-urban and rural settlements. Some specific sub-areas of area D (the Chianti area for example) are destination for retirement migrants from north Europe (especially U.K. and Germany) and north America (Benassi and Porciani, 2010). Due to this spatial proximity to the core of the metropolitan system this area is also probably the destination area for internal migrants involved in the suburbanization process of area C. Area F is the second part of the ring around the core of the FMA system. The inner homogeneity of this area is relatively low but presents a younger demographic structure compare to the area D. This is probably due to the fact that this area is not a destination area for retirement migrants but, on the contrary, is certainly a destination area for international labor migrants and also for young people and young couples. A is in fact more influenced by the dynamics of area E, that represents the *alter ego* of area C. Like this area, in fact, is an urban area characterized by an urban settlements but, differently from area C, it presents a very young demographic structure. The reason of this dual situation is that area E is strongly involved in international migration movements. As known, it is an attractive area for international migrants (especially the Chinese community) that are very concentrated in this area and, in particular, in the municipality of Prato. On the other hand the suburbanization process mainly driven by young people and young couples probably interests this area as a destination area for the people that migrate from area C.

6. Concluding remarks

The RedCap method has some advantages but also some limitations. It is very ductile, user friendly, free, allows to interact directly with the data, and takes into account directly the spatial dimension. On the other hand, it is not a probabilistic clustering method.

The spatial analysis of the demographic structure of the resident population of the Florentine Metropolitan Areas has produced some important results that clearly

show how the spatial attributes influence the demographic structure of the population. The FMA is a demographic complex spatial system where coexist: a) mountain areas with a very old demographic structure (areas A and B); b) dual core metropolitan areas composed by a relatively young area (E) and a relatively old area (C); c) two ring areas that are basically the spatial extension of the FMA core (areas D and F).

Starting from these empirical evidences we want to underline that ignoring spatial dimension can lead to misleading inference. The use of appropriate methods for the detection of spatial clusters can improved the measurement and interpretation of urban socio-economic phenomena and provide a useful information to local authorities and policy makers for regional and urban planning.

Bibliography

Agrawal R., Gerhke J., Gunopulos P. and Raghavan P., *Automatic subspace clustering of high dimensional data for data mining applications*, in proceedings of the International Conference on Management of Data, pp. 95-105, 1998.

Angayarkkani, K. and Radhakrishnan N., *Efficient forest fire detection system: a spatial data mining and image processing based approach*, in «International Journal of Computer Science and Network Security», Vol. 9, No.3, pp. 100-107, 2009.

Ankerst M., Breunig M., Kriegel H.-P. and Sander J., *OPTICS: Ordering points to identify the clustering structure*, in proceeding of the International Conference on Management of Data (SIGMOD), Philadelphia, pp.49-60, 1999.

Behnisch M. and Ultsch A., *Are there Cluster of Communities with the same dynamic behavior?,* in *Classification as a tool for research*, Part.3, Springer, Berlin, pp. 445-453, 2010.

Behnisch M. and Ultsch A., *Urban data-mining: spatiotemporal exploration of multidimensional data*, in «Building Research & Information», 37(5-6), pp. 520-532, 2009.

Benassi F., Bottai M., Giuliani G., *Migrazioni e processi di urbanizzazione in Italia. Spunti interpretativi in un'ottica biografica*, in Macchi M.J. (a cura di), *Geografie del popolamento: metodi, casi e teorie*, Edizioni dell'Università di Siena, Siena, pp. 71-78, 2009.

Benassi F. and Porciani L., *The dual demographic profile of migration in Tuscany*, in Salzmann T., Edmonston B. and Raymer J. (eds), *Demographic Aspects of Migration*, VS-VERLAG Springer, pp. 209-226, 2010.

Berry M. and Linoff G., *Data mining techniques for marketing, sales, and customer support*, Wiley, New York, 1997.

Bradley P.S., Fayyad U.M. and Reina C.A., *Scaling EM (Expectation Maximization) Clustering to Large Databases*, Microsoft Technical Report, 1998.

Dempster A.P., Laird N.M. and Rubin D.B., *Maximum likelihood form incomplete data via EM algorithm*, in «Journal of the Royal Statistical Society», Series B, 39, pp.1-38, 1977.

Ester M., Kriegel H.-P. and Sander J., *Spatial Data Mining: A Database Approach*, in proceedings of 5th International Symposium on Advances in Spatial Databases, pp.47-66, 1997.

Ester M., Kriegel H.P., Sander J. and Xu X., *A density-based algorithm for discovering clusters in large spatial databases*, in proceedings of International Conference on Knowledge Discovery and Data mining, pp. 226-231, 1996.

Guo D., *Regionalization with dynamically constrained agglomerative clustering and partitioning*, in «International Journal of Geographical Information Sciences», Vol.22, No.7, pp.801-823, 2008.

Guo, D., Chen J., MacEachren A. M. and Liao K., *A Visualization System for Spatio-Temporal and Multivariate Patterns (VIS-STAMP)*, in «IEEE Transactions on Visualization and Computer Graphics» 12(6), pp. 1461-1474, 2006.

Guo, D., Gahegan M., MacEachren A.M. and Zhou B., *Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach*, in «Cartography and Geographic Information Science», vol. 32, No. 2, pp. 113-132, 2005.

Guha S., Rastogi R. and Sim K., *CURE: an efficient clustering algorithm for large databases*, in proceedings of International Conference on Management of Data, pp. 73-84, 1998.

Jin H. and Guo D., *Understanding Climate Change Patterns with Multivariate Geovisualization*, in proceedings of the International Conference on Data Mining Workshops, IEEE Press, pp. 217-222, 2009.

Han J., *Data Mining*, in Urban J. and Dasgupta P. (eds.), *Encyclopedia of Distributed Computing*, Kluwer Academic Publishers, 1999.

Han J., Kamber M. and Tung A.K.H, *Spatial clustering methods in data mining: a survey*, in Miller H.J. and Han J. (eds.), *Geographic data mining and knowledge discovery*, Taylor & Francis, pp. 1-29, 2001.

Han J., Lee J.G. and Kamber M., *An overview of clustering methods in geographic data analysis*, in Miller H. J. and Han J. (eds.), *Geographic data mining and knowledge discovery*, 2nd edition, Taylor & Francis, pp. 149-170, 2009.

Hinneburg A. and Keim D., *An efficient approach to clustering in large multimedia databases with noise*, in proceedings KDD, 1998

Hosking J.R.M, Pednault E.P.D. and Madhu S., *A statistical perspectives on data mining*, in «Future Generation Computer System», 133, pp.17-134, 1997.

Jiao L., Liu Y., *Knowledge discovery by spatial clustering based on self-organizing feature map and composite distance measure*, in «The international archives of the photogrammetry, remote sensing and spatial information sciences», Vol.XXXVII, Part. B2, pp.219-224, 2008.

Kaufman L. and Rousseeuw P.J., *Finding Groups in Data: an introduction to cluster analysis*, John Wiley & Sons, 1990.

Karypis G., Han E.H. and Kumar V., *CHAMALEON: A hierarchical clustering algorithm using dynamic modeling*, in «Computer», 32, pp. 68-75, 1999.

Koperski K., Adhikany J. and Han J., *Knowledge discovery in spatial database: progress and challenges*, in proceedings of the workshop on research issues on data mining and knowledge discovery, Montreal, pp.55-70, 1996.

Ng R. and Han J., *Efficient and Effective Clustering method for spatial Data mining*, in proceedings of International Conference on Very Large Data Bases, pp. 144-155, 1994.

Miller H.J. and Hann J., *Geographic data mining and knowledge discovery*, Taylor & Francis, 2001.

MacQueen J., *Some methods for classification and analysis of multivariate observation*, in proceedings of the 5th Berkley Symposium on Mathematics, Statistics and Probabilities, edit by L. Le Cam and J. Neyman, Vol.1, pp. 281-297, 1967.

Moran C.J., and Bui E.N., *Spatial data mining for enhanced soil map modeling*, «International Journal of Geographical Information Science», Vol.16, Issue 6, pp. 533-549, 2002.

Petrucci A., Salvati N., Salvini S. and Vignoli D., *Invecchiamento e mobilità nell'area metropolitana fiorentina*, «Rivista di Economia e Statistica del Territorio», (2), pp.81-103, 2008.

Petrucci A., Salvini S. and Vignoli D., *Vieillesement et évolution du peuplement dans l'aire fiorentine*, in G.F. Dumont (eds.), *Les territoires face au vieillissement en France et en Europe. Géographie – Politique – Prospective*, Carrefours Les Dossiers, ellipses, Paris, 2006.

Ripley B. D., *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press, 1996.

Roddick J.F. and Spiliopoulou M., *A bibliography of temporal, spatio and spatio-temporal data mining research*, in SIGKDD EXPLORATIONS; 1999 ACM SIGKDD, Vol. 1, Issue I: pp. 34-38, 1999.

Sander J., Ester M., Kriegel H.P. and Xu X., *Density-based clustering in spatial data bases: a new alghorithm and its applications*, in «Data mining and Knowledge Discovery, an International Journal» Kluwer Academic Publisher, Vol.2, No.2.

Sang H., Gelfald A.E., Lennard C., Hegerl G., Hewitson B., *Interpreting Self-Organizing Maps trough space time models*, in «The Annals of Applied Statistics, Institute of Mathematical Statistics», Vol. 2, No. 4, pp. 1194-1216, 2008.

Sheikholeslami G., Chatterjee S. and Zhang A., *Wave Cluster: A Multi-Resolution Clustering Approach for Very Large Spatial DataBases*, proceedings on International Conference on Very Large Data Bases (VLDB, '98), pp.428-439, 1998.

Shekhar S., Chawla S., *Spatial Data Base: a tour*, Prentice Hall, 2003.

Szalay A., Kunszt P., Thakar A. and Gray J., *Designing and mining multi-terabyte astronomy archives: The sloan digital sky survey*, in proceedings of SIGMOD, 2000.

Tobler W., *A computer movie simulating urban growth in the Detroit region*, in «Economic Geography», Vol. 46 No. 2, pp. 234-240, 1970.

Van den Berg, *Urban Europe: a study of growth and decline*, Oxford University Press, 1992.

Varlaro A., *Spatial clustering of structured objects*, Phd thesis in Computer Science, University of Bari, 2008.

Vignoli D., Dugheri G., Ferro I., Salvini S. and Secondi L., *L'area fiorentina: quanti siamo e quanti saremo*, Ufficio Statistica del Comune di Firenze, Serie *La statistica per la città*, 2007.

Vinod H., *Integer programming and the theory of grouping*, in «Journal of American Statistical Association», No. 64, pp. 506-517, 1969.

Wang W., Yang J. and Muntz R., *STING: a statistical information grid approach to spatial data mining*, in proceedings of International Conference on Very Large Data Base, pp. 186-195, 1997.

Yu D., Chatterjee S., Sheikholeslami G. and Zhang A., *Efficiency detecting arbitrary shaped clusters in very large datasets with high dimensions*, SUNY, Buffalo, Computer Science Technical Report, pp. 98-08, 1998.

Yu Pan J. and Faloutsos C., *"GeoPlot": spatial data mining on video libraries*, in proceedings of the eleven international conference on Information and knowledge management, pp. 405 - 412, 2002

Zhang T., Rarnakrishnan R. and Livny M., *BIRCH: an efficient data clustering method for very large databases*, in proceedings of International Conference on Management of Data, pp.103-114, 1996.

Copyright © 2010
Federico Benassi, Chiara Bocci,
Alessandra Petrucci