

Regionalization with dynamically constrained agglomerative clustering and partitioning. An application on spatial segregation of foreign population in Italy at regional level

Federico Benassi and Raffaele Ferrara

Abstract In this paper we propose an application of a new method of regionalization and spatial clustering called RedCap. Firstly we describe this method and how it works and then we propose a possible application of it: finding homogenous and spatially contiguity areas in Italy with the regard of spatial segregation of foreign population. The results obtained offer first indications about the spatial dimension of foreign population in Italy and could be a start point to develop new and more deeply analysis in the future applying this kind of method.

Key words: Geographic Information System, Migration, Spatial distribution of foreign population, Regionalization, Spatial data analysis

1 Introduction

Regionalization aims at dividing a large set of spatial objects into a smaller number of spatially contiguous regions while optimizing an objective function, which is normally a measure of the attribute similarity in each region (Guo, 2008).

¹ Federico Benassi, Department of Statistics “G. Parenti”, University of Florence; email: benassi@ds.unifi.it
Raffaele Ferrara, Department of Statistical Sciences, University of Naples “Federico II”; email: rf.ferrara@virgilio.it

In this paper we test a new method of regionalization through finding homogeneous areas in Italy, with regard to the phenomenon of spatial segregation of foreign population.

The applied method (RedCap¹) is relatively innovative because it uses spatial-data mining techniques joining computational and visual approaches to regionalization and spatial clustering. In particular, RedCap uses the Som (Self Organizing Map – Kohonen, 2001) to identify and visualize clusters and then combines a specific agglomerative clustering method (complete linkage clustering) and a constraining spatial contiguity strategy (full order clustering) among clusters to identify new regionalization.

In our work, we computed some indexes of spatial segregation of foreign population in Italy at regional level. For each region the indexes are computed using the distribution of foreign resident population at municipality level in 2009 (Istat). Then, we used these indexes as input variables into RedCap. We finally obtained homogeneous areas. This result could be a start point to develop further analysis especially applying the RedCap method.

The paper is organized as follows: section 2 presents a brief description of the methodology; section 3 describes the indexes used as input variables into RedCap; section 4 shows the results; section 5 presents concluding remarks and possible future development of this research.

2 A brief description of RedCap and the full order constrained complete linkage clustering method

2.1 *Regionalization and spatial clustering*

Regionalization is a special form of spatial clustering where spatial units are grouped together, based on a set of defined criteria and a set of contiguity or adjacency constraints (Guo, 2008). From an operative point of view, regionalization, and in general spatial clustering, is essentially a combinatorial problem. Existing methods for multivariate spatial analysis and clustering span a continuum between computational and visual approaches (Guo, 2005). 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. RedCap represents an integrated geographic

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

Regionalization with dynamically constrained agglomerative clustering and partitioning. An application on spatial segregation of foreign population in Italy at regional level 3
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.

2.2 Self organizing map as a tool to identify spatial clusters

The Self organizing map (Som) represents an intermediate approach between visual and automatic computational methods. The Som projects high-dimensional data to a low-dimensional space with preserving nonlinear relationship by producing a similarity graph of the input data (Kohonen, 2001; Guo, 2005). In this work, the input data for the Som component are the attribute space x and the variables that we introduce into the software (see section 3 and 4). The construction of a Som is an iterative process that depends on the size of the Som and the complexity of the data (the number of iterations must be at least 500 times the number of Som nodes (Kohonen, 2001; Guo, 2005)). Each Som node is associated with a vector which represents the position of this node in the input attribute space. The Som first initializes each node by assigning its codebook vector randomly (Kohonen, 2001; Guo, 2005). During the iterative learning process, each codebook vector is adjusted according to the data items falling inside and the codebook vectors of its neighboring nodes are adjusted accordingly (Guo, 2005). This iterative process allows to identify a (first) winner node that will be the vector with most similar output vector (x) related to the input one (y). The measure of similarity adopted is the Euclidean distance:

$$DIS(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

the neighboring nodes are the nodes that have a distance from the winner node less than R , where R is the σ of a Laplacian of Gaussian distribution:

$$\nabla^2 G = \frac{4}{\sigma^2} \left(\frac{x^2 + y^2}{\sigma^2} - 1 \right) e^{-\frac{x^2 + y^2}{\sigma^2}} \quad (2)$$

this iterative process continues since all the nodes (winner node plus neighbor nodes) are individuated and their number is the same of the dimension of the Som. After this learning process is complete, each node (winner node plus neighbor nodes) has a new position in the input attribute space. With their new positions and topological relationships in the 2D layout, the Som nodes form a nonlinear, smooth surface in the input attribute space, which can be regarded as the result of a non linear regression (Guo, 2005). 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 (Guo, 2005). This kind of graphic display of the Som result is called U-matrix (unified distance matrix) (Kohonen, 2001). In Redcap Guo, adopted 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).

2.3 *Identify similar clusters and contiguity clusters trough CLK-Full order clustering*

To understand an agglomerative clustering method is important to identify how this method defines distance between clusters. The complete linkage clustering method (CLK) defines the distance between two clusters as the dissimilarity between the furthest pair of data points (Guo, 2008):

$$d_{CLK}(L, M) = \max_{u \in L, v \in M} (d_{uv}) \quad (3)$$

where L and M are two clusters, $u \in L$ and $v \in M$ are two data points and d_{uv} is the dissimilarity between u and v . At the beginning of the process, each individual data is a cluster by itself. The original data are up dated with the information of the Som algorithms then the most similar (given by the distance definition) pair of clusters are selected and merged into one cluster. The merging process repeats until all data points are in the same cluster. The strategy used to incorporate contiguity constraints in this hierarchical clustering methods is called full ordering constraining (Guo, 2008). In general terms, 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. As defined by Guo, a full-order constraining strategy includes all edges in the clustering process, and the distance between two clusters throughout all considered 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). The algorithm of this method produces different trees. Different trees define different searches for the partitioning algorithm. Subsequently, trough the partition of the spatial contiguous tree, we obtain a number of subtrees, each of them corresponding to a spatial contiguous region. As indicated above, the tree is based on distances between clusters while the partition is carried out to optimize an objective function, which is to minimize the total heterogeneity value of all regions (Guo, 2008). The partitioning algorithm iteratively partitions a spatial contiguous tree into K regions by cutting a subtree (i.e. a region or a macro region) into two ones at each step. The best subtree to cut at each step is the one with the largest homogeneity gain which is computed for each tree (or subtree) with the following equation:

$$h_g^*(R) = \max(H(R) - H(R_a) - H(R_b)) \quad (4)$$

where R_a and R_b are the two subtrees from a possible cut of R , and h_g^* is the homogeneity gain (i.e. heterogeneity reduction) for the tree after the best cut. Given a spatial contiguous tree, two subtrees (and thus two regions) are generated by removing any edge from the tree. To obtain K regions, $K-1$ edges have to be removed (Guo, 2008).

3 Three indexes to measure the spatial segregation of foreign population in Italy

Spatial segregation of foreign population is a very complex multidimensional phenomenon. Anyway, Massey and Denton (1988) defined five conceptual dimensions of it: *evenness* (concerns the distribution of foreigners in the territory); *exposure* (concerns the potential level of contact and interaction among individuals belonging to different groups); *concentration* (space in relative terms occupied by a minority group); *centralization* (level of spatial localization of the group respect to the core of an urban area) and *clustering* (how the areas where minority groups live are contiguous to each others) (Massey, Denton, 1988). The first index is the relative index of dissimilarity which is one of the most famous measures of spatial segregation (Wong, 1993). It is an homogeneity measure and it is a function of the population distribution of two groups (X , Y) in the sub areas i that compose the territory of analysis:

$$D = \frac{1}{2} \sum_i \left| \frac{x_i}{X} - \frac{y_i}{Y} \right| \quad (5)$$

The value of D index is included between 0 (the two groups are in the same proportion in all the areas, this means absence of segregation) and 1 (one group takes up an area in exclusive terms, this means perfect segregation) (Duncan, Duncan, 1955; Massey, Denton, 1988). It is important to underline the limits of this index. First of all, it is a non spatial index (it doesn't consider the spatial distribution of the groups in the territory); it is also affected by a scale effect: its value tends to change in function of the spatial level of the data (Strozza, Ferrara, Labadia, 2006). In our study, we applied this index to the foreign population and the control group is the Italian population.

The second index is the interaction index which measures the potential level of contact between the minority group X (foreigners) and the majority group Y (Italians):

$$xPy = \sum_{i=1}^n \left[\frac{x_i}{X} \right] \left[\frac{y_i}{Y} \right] \quad (6)$$

The value of this index is included between 0 and 1. This reflects the probability that a person belonging to the minority group could share a spatial

area with a person belonging to the majority group. This index is obtained as the mean value of the proportions of the persons belonging to the majority group weighted with the proportion of persons belonging to the minority group (Massey, Denton, 1988; Strozza, Ferrara, Labadia, 2006). The major limit of this index is that this index is asymmetrical. The third index, Delta, is a measure of concentration and it is referred to the relative proportion of physical space occupied by a minority group. This index is calculated as follows:

$$DEL = \frac{1}{2} \sum_{i=1}^n \left| \frac{x_i}{X} - \frac{a_i}{A} \right| \quad (7)$$

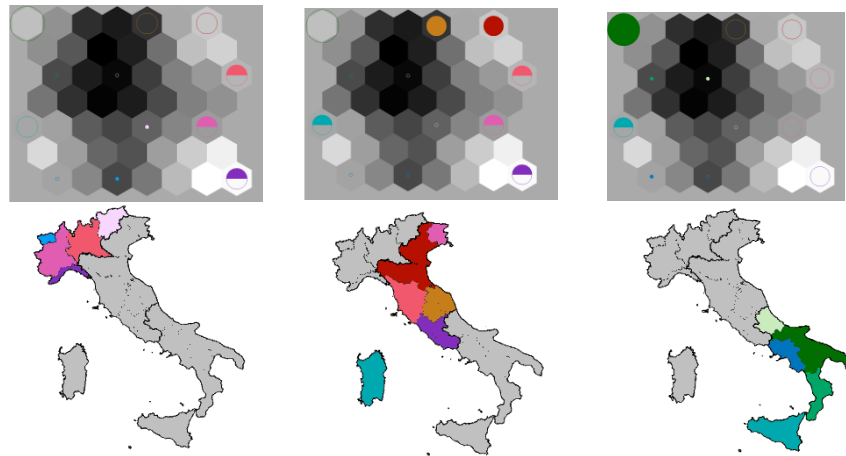
where a_i is the area of the territorial unit i and A is the area of the total of the territorial unit. This index describes the proportion of the members belonging to the minority group who should change their residential area in order that the density of their group becomes equal in all the territory in analysis (Massey, Denton, 1988; Strozza, Ferrara, Labadia, 2006).

4 Results

We computed 5 simple indexes for each of the 20 Italian regions using the distribution of foreign resident population at municipality level in 2009 (Istat). Besides the three indexes described above, we computed for each region two more simple indexes: 1) the proportion of resident foreign population on the total resident population; 2) the proportion of foreign population resident in capital city on the total of foreign population resident in the region. These 5 indexes are the input variables of our regionalization process. Before starting the process, each indicator is normalized to Z –score distribution and weighted (in this case we assigned the same weight to the 5 indexes). The control variable of the regionalization process is the resident population in each region in 2009. We fixed a minimum population for each new region of 15 million and a maximum of 4 new regions. Before starting the second step of the process and completing the regionalization, the software creates the clusters by using the method described in the section 2. The result of this first step is elaborated by a Som (Self Organizing Map) (Kohonen, 2001) and visualized through an interactive map. This process identifies some natural clusters and in particular four groups. The first one is composed by Lombardia, Veneto, Toscana, Marche, Umbria and Sardegna; the second one by Liguria, Piemonte, Trentino Alto-Adige, Friuli Venezia Giulia and Lazio; the third one by Valle d’Aosta, Campania, Sicilia and Sardegna and the fourth one by Abruzzo, Molise, Puglia, Basilicata and Calabria. These 4 groups of clusters present a degree of heterogeneity both in terms of non complete similarity and also in terms of spatial contiguity. The second step of the regionalization method (as

Regionalization with dynamically constrained agglomerative clustering and partitioning. An application on spatial segregation of foreign population in Italy at regional level described in section 2) is to achieve spatial contiguity clusters minimizing the heterogeneity inside of them. The results are shown in the following charts:

Figure 1 - *Som and Map of the new regionalization obtained (with constraining strategy)*



we have to underline that a degree of heterogeneity inside the new groups of regions still remains, nevertheless the new groups satisfy the spatial contiguity condition. The first group is composed by Piemonte, Trentino, Lombardia, Valle d'Aosta and Liguria. In this group the proportion of foreign population on the total resident population is high as well as the level of spatial concentration of foreign population (not only in the main urban areas but also in the industrial districts). On the contrary, the level of potential interaction with the Italian population is low. The second group is composed by Friuli, Veneto, Emilia Romagna, Toscana, Umbria, Marche, Sardegna and Lazio. In this second group the foreign population has low level of settlement dissimilarity in comparison to the Italian ones, medium level of spatial concentration (also if in some cases the proportion of foreigners resident in regional capital cities is relevant) and a medium level of potential interaction with the Italian population. The last group is composed by the regions of Southern Italy. In these regions the proportion of foreign population on the total resident population is very low; generally, the foreigners are distributed in all the territory and not present high level of concentration in the major urban areas. At the same time, the probability of sharing the same spatial area with the Italian population is high. In this group, Sicilia and Campania represent particular study cases: the level of spatial concentration of the foreign population is very high (similar to the ones of North Italy). Is not a case that without applying the constraining strategy these two regions (with Sardegna and Valle d'Aosta) constitute a single cluster. Nevertheless the high level of concentration of foreign population, the potential interaction between foreign and Italian population in these four groups is high. This happens mainly because also the Italian population in these areas are really concentrated. In

conclusion, nevertheless the nature essentially descriptive of our analysis and the limits of the indexes used, we can underline a first important information. In Northern and Central Italy, where the proportion of foreign population is high and where this population is concentrated in specific areas of the territory, the policy on migration taken at municipality level should have a major probability to be efficiency than policy taken at regional level. On the contrary, in Southern Italy, where the proportion of foreign presence is low and its spatial distribution involves all the territorial units, the immigration phenomenon should be manage at regional level.

5 Concluding remarks and future developments

This paper represents the first step of a wider research which is still in progress. At the moment we try to describe a relatively new regionalization method, analyze how it works, and show one of the possible applications. Also the proposed application, spatial distribution/segregation of foreign population in Italy at regional level, is to be considered in progress. In the second step of the work we would like to modify the indexes applied (in particular it could be interesting transform these indexes in “spatial indexes”, meaning indexes which reflect also the morphological aspects of the territory). In addition, it will be useful to introduce new indexes belonging to conceptual dimensions of spatial segregation not taken into account in this analysis (centralization and clustering) and new indexes referred to some important dimensions of foreign presences in Italy (i.e. index of economic and labor integration, index of migratory history and so on) and also to refer the analysis on more fine territorial dimension (i.e. Italian province).

References

1. Duncan, O.D., Duncan, B.: A methodological analysis of segregation indexes. In: *American Sociological Review*, Vol.20, No.2, pp.210-217 (1955)
2. 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)
3. Guo, D.: 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)
4. Massey, D.S., Denton N.A.: The dimensions of residential segregation. In: *Social Force*, Vol.67, No.2, pp.281-315 (1988)
5. Strozza. S., Ferrara, R., Labadia, C.: Alcune misure della distribuzione territoriale delle popolazioni: dall'esame della letteratura a qualche applicazione. In: *Rivista Italiana di Economia, Demografia e Statistica*, Vol. LX, No.3-4, pp.441-449 (2006)
6. Wong, D.W.S.: Spatial indices of segregation. In: *Urban Studies*, Vol.30, pp.559-572 (1993)
7. Kohonen, T., *Self-organizing maps*, Springer, Berlin, New York (2001)