Distance-based LISA maps for multivariate lattice data

Pedro Delicado Sonia Broner
Universitat Politècnica de Catalunya

December 1, 2008

Abstract. In the context of areal data (a particular case of data with spatial dependence) we propose an algorithm to define spatial clusters. Our proposal is based on distance between the characteristics observed in different areas (individual). Thus it is able to be applied to any kind of observable characteristic on condition that an inter-individual distance can be defined. This way we provide a generalization of the well-known LISA maps that have been widely used for univariate data. We apply our proposals to the results of 2004 Spanish General Elections recorded at 248 neighborhoods in Barcelona.
Keywords: Clustering algorithm; compositional data; electoral data; spatial clustering.

1 Introduction

Cressie (1993) defines a *spatial process* as

$$\left\{ \boldsymbol{X}_s : \ s \in D \subseteq \mathbb{R}^d \right\}$$

where s is a generic data location in the d-dimensional Euclidean space (d is usually equal to 2), the set $D \subseteq \mathbb{R}^d$ could be fixed or random, and $\boldsymbol{X}_s$ are *random variables*, defined as random elements taking values in a metric space $\mathcal{X}$, where a distance d is defined. Typically $\mathcal{X}$ is $\mathbb{R}$ (the spatial process is then univariate) or $\mathbb{R}^p$ (corresponding to multivariate spatial process). The realization of a spatial process,

$$x_{s_1}, \ldots, x_{s_n}, \ s_i \in D, \ i = 1, \ldots, n,$$

constitutes a *set of spatial data*. The nature of the set D allows us to classify spatial data as:

- *Geostatistical data*, when D is a fixed subset of $\mathbb{R}^d$ with positive volume and n points $s_1, \ldots, s_n$ in D are chosen to observe the random elements $\boldsymbol{X}_{s_i}$, $i = 1, \ldots, n$.
- *Areal data*, when D is fixed and countable. Usually there is a bijection between D and a partition of a geographical area and, for any $s \in D$, $\boldsymbol{X}_s$ is a summary function of an event happened at the part of the area corresponding to s by this bijection.
- *Point process*, when we have a point process with locations given by the Nature. It could be possible to observe a set of characteristics (*marks*) at each observed point, in addition to its location.

In this paper we focus on data in lattice. In particular we propose a spatial clustering algorithm valid for a wide range of data types. Our proposal generalizes in some sense LISA maps, introduced by Anselin (1995) for univariate data in lattice. We base the algorithm on distances $d(x_{s_i}, x_{s_i})$ between the observed characteristics, thus making it valid for analyzing data in lattice of any kind as far as an appropriate distance function could be defined.

The rest of the work is organized as follows. Section 2 is a reminder of several known notions of local and global spatial dependence, that are useful to describe how LISA maps are done. Some drawbacks of LISA maps are given there. In Section 3 we introduce our proposal of spatial clustering, that we call *distance-based LISA maps*, overcoming the cited difficulties of usual LISA maps. An application to electoral data in Barcelona is developed in Section 4. Finally, some conclusion are listed in Section 5.

2 LISA maps

Let D be a discrete set and let $x_{s_1}, \ldots, x_{s_n}$, $s_i \in D$ for $i = 1, \ldots, n$, be a lattice data set. Usually locations s_i correspond to administrative areas defining a partition of the map of the region under study. For instance, if we are studying electoral results in the a big city, s_i can be electoral districts which the city is divided into (see Section 4 for such a real example referred to Barcelona).

The spatial information about location s_i is usually given by specifying the polygon defined by the boundary of the administrative area i. From those polygons it is possible to extract information about spatial characteristics of

each area. In particular, the geographical coordinates of the polygon centroid are sometimes used as spatial position of an area. More frequently, a $n \times n$ neighborhood matrix W is defined from the set of polygons, where element (i, j), w_{ij}, is 1 if polygons i and j have a common boundary and 0 otherwise ($w_{ii} = 0$ fora all i).

A *spatial cluster* is defined as a set of areas that are close to each other having similar observed values for the variable of interest $\boldsymbol{X}$. This kind of clusters would exist when variable $\boldsymbol{X}$ presents spatial dependence at local level. Anselin (1995) introduced Local Indicators of Spatial Association (LISA) to quantify local dependence of a random variable $\boldsymbol{X}$. For the case of continuous univariate $\boldsymbol{X}$, one of the most used LISA is the Local Moran's index, defined for location s_i as

$$I_i = \frac{(x_{s_i} - \overline{x}_n)}{\frac{1}{n}\sum_{j=1}^{n}(x_{s_j} - \overline{x}_n)^2} \sum_{j=1}^{n} w_{ij}^*(x_{s_j} - \overline{x}_n),$$

where $\overline{x}_n$ is the mean value of the variable $\boldsymbol{X}$ in the region under study and $w_{ij}^* = w_{ij}/(\sum_{j=1}^{n} w_{ij})$ are the elements of the neighborhood matrix W normalized to add up to 1 row-wise. Spatial dependence around location s_i is significant when the observed value of I_i is much grater than values corresponding to random allocation of values variable $\boldsymbol{X}$ to neighbors of s_i. Therefore a permutation test is a valid tool to identify locations with high spatial dependence, that could be positive (if I_i is significantly high) or negative (if I_i is significantly low), implying respectively spatial clustering around s_i or that s_i an spatial outlier.

The sum of all Local Moran's indexes form the (Global) Moran's I index (Moran 1950) of spatial dependence,

$$I = \sum_{i=1}^{n} I_i = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_{s_i} - \overline{x}_n)(x_{s_j} - \overline{x}_n)}{\sum_{i=1}^{n}(x_{s_i} - \overline{x}_n)^2},$$

a global measure of spatial autocorrelation. Again a permutation test is useful to determine if the observed Moran's I is significantly high (positive spatial correlation) or low (negative spatial correlation).

The information contained in local Moran's Indexes I_i is usually summarized in a map, known as LISA map, drawn as follows (see Anselin 1995, and the software GeoDA, Anselin 2003). Only tracts with significant spatial dependence are colored and four colors are used for these areas. Red color is used to mark tracts with high values of $\boldsymbol{X}$ surrounded by other tracts with high values (usually labeled as HH). Blue color is used for areas with low values of $\boldsymbol{X}$ with neighbors also having low values of $\boldsymbol{X}$ (usually labeled as

LL). These two cases correspond to values of I_i is significantly high. Other two different colors are used to represents areas where I_i is significantly low, corresponding to spatial outliers: one of them represents areas whit high values of $\boldsymbol{X}$ surrounded by tracts with low values (labeled HL), an the other color corresponds to the opposite situation (labeled LH).

LISA maps are an excellent tool for exploratory spatial data analysis when the characteristic of interest is a one-dimensional continuous variable. They offer an easy way to visualize two main spatial clusters (HH and LL) and where the spatial outliers are p'laced (HL and LH). Nevertheless several drawbacks of LISA maps can be pointed out:

1. They are not able to detect more than two spatial clusters. Assume for instance that we study a triangular region and that the variable of interest is three-modal, with each mode corresponding to one of the three vertexes of the region (assume, for instance that they are around values -1, 0, and 1, and that they have a similar weight). A LISA map for this situation will probably detect two spatial clusters corresponding to vertexes where the two extreme mode have been allocated, but will not mark the other vertex as a spatial cluster. Or even worse: the third cluster could be detected, but the tracts belonging to it would be colored with one of the four available colors almost at random, because the observed value of $\boldsymbol{X}$ at one of these tracts, as well as the average values at their neighbors, will be close to 0 and their sign will be positive or negative with equal probability.

2. The cluster detected by LISA maps (those marked as HH or as LL) are not guarantee to be spatially connected. For instance, in a squared region the vertexes could allocated the tracts whit highest values of variable $\boldsymbol{X}$, while the lowest values correspond to the remainder tracts. In this situation the cluster HH would be not connected, with four connected components located at the vertexes.

3. They are not valid for random variables $\boldsymbol{X}$ with dimension greater than one, because the definition of local Moran's Indexes is valid only for one-dimensional data.

In the next section we introduce distance-based LISA maps, which overcome these difficulties maintaining the easy interpretability of LISA maps.

3 Distance-based LISA maps

Assume now that the random variable of interest $\boldsymbol{X}$ takes values in a metric space $\mathcal{X}$, where the distance function d is defined. Let $d_{ij} = d(x_{s_i}, x_{s_i})$ be the distance between the observed values of $\boldsymbol{X}$ at tracts s_i and s_j, for $i, j = 1, \ldots, n$. Let D be the $n \times n$ distance matrix with element (i, j) equal to d_{ij}. Let W be the neighborhood matrix. Fix a significance level $\alpha \in [0, 1]$ and a large number of permutations P to be used in a permutations test (we use $\alpha = 0.05$ and $P = 999$ in our examples). The algorithm we propose for spatial clustering is as follows.

Algorithm: Distance-based LISA maps

Step 0: Remove the isolated areas (those having no connection with any other tract).

Step 1: Detecting global outliers.

1.1 For $i = 1, \ldots, n$ compute $d_{\min}(i) = \min j \neq i d_{ij}$, draw a boxplot of these quantities and mark as global outliers areas s_i such that $d_{\min}(i)$ is larger than the upper whisker.

1.2 Remove global outliers from the data-set and recheck for the existence of new isolated areas (those connected only to tracts marked as global outliers), that might also be removed. Let $A \subseteq \{1, \ldots, n\}$ be the set of remaining tracts.

Step 2: Marking tracts significantly similar to (and significantly different from) their neighbors. For all $i \in A$, do the following:

2.1 Compute $\bar{d}_i$: the average of values d_{ij}, where j are the elements in A such that s_i and s_j are neighbors, that is, $w_{ij} = 1$. (Note: other location measures could also be used instead of the average).

2.2 Repeat for $p = 1, \ldots, P$:

· Obtain a permutation π of $j \in A - \{i\}$.

· Compute $\bar{d}_i^p$, the average of values $d_{i\pi(j)}$, $j \in A$ such that $w_{ij} = 1$.

2.3 Compute q_1 and q_2, the quantiles $\alpha/2$ and $(1 - \alpha/2)$ of $\bar{d}_i^p$, $p = 1, \ldots, P$.

2.4 If $\bar{d}_i < q_1$ mark tract s_i as member of a spatial cluster. If $\bar{d}_i > q_2$ mark tract s_i as spatial outlier.

Step 3: Including non-marked tracts that are similar to a neighbor marked tract.

For all i such that s_i has been marked as member of a spatial cluster:

3.1 Let d_i^* be the maximum distance d_{ij}, for s_i to its neighbors s_j also marked as members of a spatial cluster. If there are no such a neighbor, let $d_i^* = q_1$.

3.2 Visit all their neighbors j such that s_j has not been marked jet. If $d_{ij} \leq d_i^*$ then mar also S_j as member of a spatial cluster.

Step 4: Identifying spatial clusters.

Consider the data-set $C = \{x_{s_i} : s_i \text{ is marked as member of a spatial cluster}\}$ and use any standard algorithm to performs a cluster analysis on C. Let $C_1, \ldots, C_k$ the resulting partition of C.

Step 5: Drawing the map.

Draw a map of the region and color each area in accordance with their mark (for global and spatial outliers) or with the cluster C_j it belongs. Non-marked areas are left in white.

4 A real example: Electoral data in Barcelona

We consider results of the 2004 General Election to the Spanish Parliament in the city of Barcelona. In 248 Zones of Study (ZRP, from the initials in Catalan), which are groupings of neighborhoods defined for statistical purposes, seven variables are recorded: the participation (proportion of people having right to vote that really votes), the proportion of votes obtained by the main five political parties and the proportion corresponding to other minor parties. The five main political parties are CIU (center-right nationalist party), PSC (Catalan Socialist Party), ICV (left ecologist party), PP (Popular Party), ERC (left nationalist party). The data are obtained from the Department of Statistics of the Barcelona Municipal Council web page (`http://www.bcn.cat/estadistica/angles/index.htm`).

First we study participation, a univariate quantity. Thus it is possible to draw a standard LISA maps for this variable (see Figure1, upper panel). The HH cluster (red color) corresponds to areas where people has high educational and economical level. The LL cluster (blue color) contains areas with high proportion of working class. There area three tracts whit high participation surrounded by other with low values (HL, pink color), and one in the opposite case (LH, light blue).

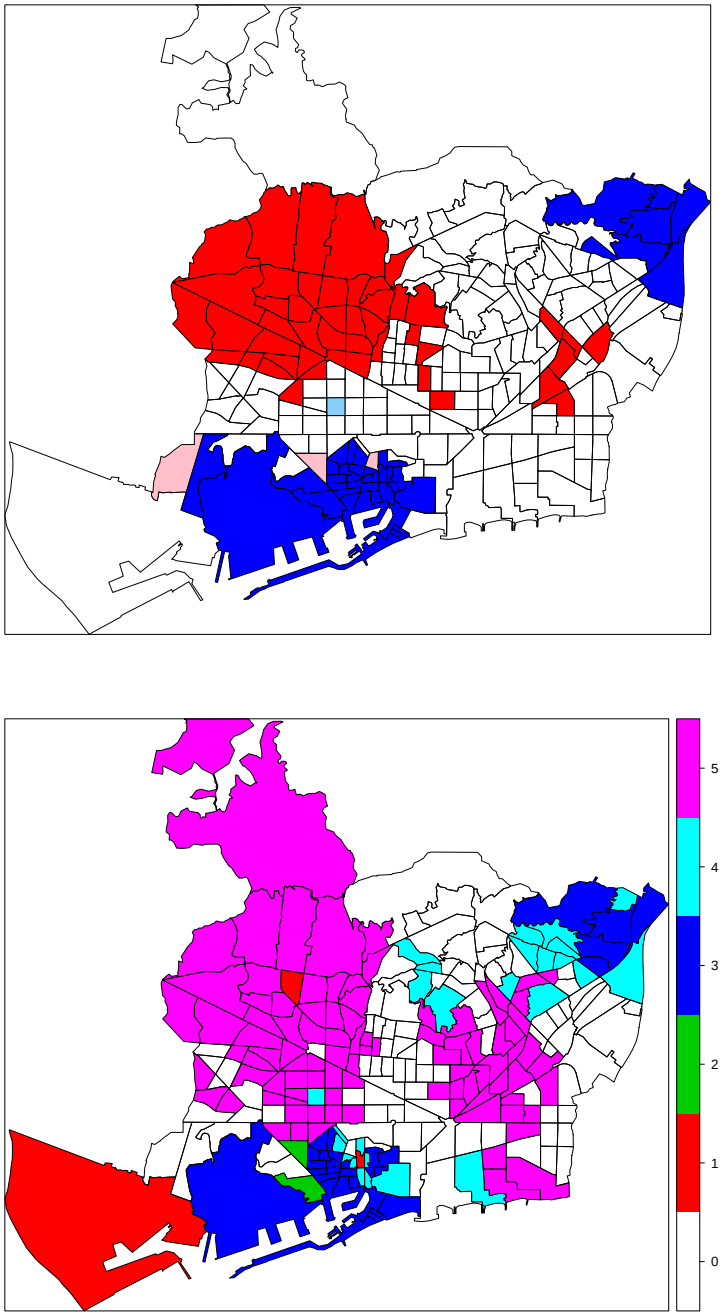

Figure 1: Participation in 2004 General Elections at ZRPs of Barcelona. *Upper panel:* Standard LISA map. Color code: HH, red; LL, blue; HL, pink; LH, light blue. *Lower panel:* Distance-based LISA map.

We apply now the algorithm presented ion the previous section to obtain the distance-based LISA map corresponding to participation. The resulting map is in the lower panel of Figure1. Areas colored in red (marked with a number 1) are the global outliers: there are 3 of them in this example. Observe that one of the areas marked as HL in the standard LISA map is a neighbor of the global outlier located in the middle of the old town (the small area in the center, close to harbor). Spatial outliers are marked with number 2 and colored in green: there are 2. One of them was marked as HL in the standard LISA map (the triangular shaped one). Areas with no color (white areas marked with a 0) are no marked in any way (they are no outliers and no clustered). There are three spatial clusters in this example (with labels 3, 4, and 5, and colors blue, magenta and light blue, respectively). The blue cluster here approximately corresponds to the cluster LL in the standard LISA map (also colored in blue). The magenta cluster includes the red HH cluster in the standard LISA map. The third cluster (light blue) is a transition from areas with low participation (blue cluster) to those with high values (magenta cluster).

We analyze now proportion of votes obtained by the main five political parties (and the remainder), taken as a 6-dimensional variable. This vector is an example of *compositional data* (see Pawlowsky-Glahn and Egozcue 2001, for instance). Therefore the appropriate distance measure between observations is no longer Euclidean distances but other that take into account the compositional character of the data. In Pawlowsky-Glahn and Egozcue (2001) the authors recommend to use the following distance:

$$d^2(\mathbf{x},\mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left(\log\frac{x_i}{x_j} - \log\frac{y_i}{y_j}\right)^2,$$

where $\mathbf{x} = (x_1,\ldots,x_k)$ and $\mathbf{y} = (y_1,\ldots,y_k)$ are two compositional data with k components.

The corresponding distance-based LISA map is shown in Figure 2. In this case there are 15 areas marked as global outliers (in red in the map) and 26 that are not marked (in white). There are no areas marked as spatial outliers (in green). The remainder 207 areas are clustered in 4 clusters, numbered as 3 (76 areas, in blue), 4 (47 areas, in light blue), 5 (75 areas, in magenta) and 6 (9 areas, in yellow).

In order to understand this classification we draw boxplot of the proportion of votes (transformed by the logit function: $f(p) = \log(p/(1-p))$) of each party separated by clusters. The results are in Figure 3. At cluster 3 (blue) the two main parties (PSC and CIU) are very close to the global mean, whereas smaller left parties (ICV and ERC) present slightly better

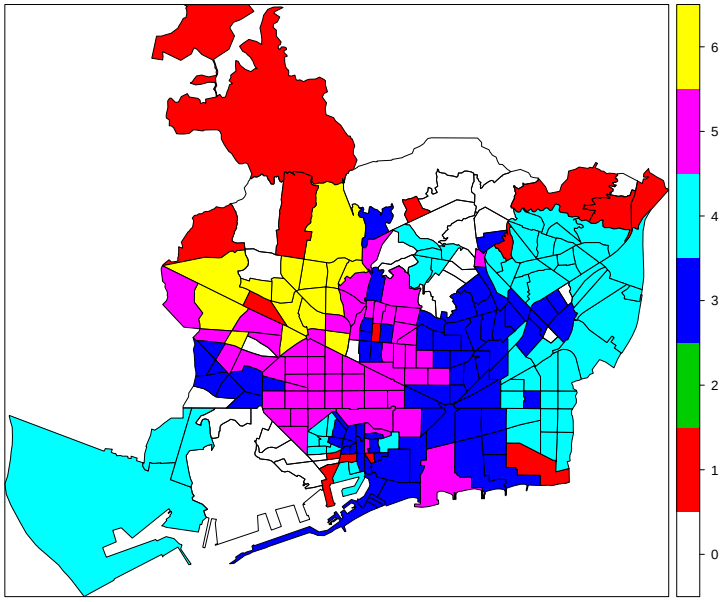

Figure 2: Distance-based LISA map for the proportion of votes obtained by each political party in 2004 General Elections at ZRPs of Barcelona.

results than in the whole city, and the opposite happens for the smaller right party (PP). At cluster 4 (light blue) the nationalist parties (CIU and ERC) have worst results than the global ones and the opposite happens for the no-nationalist parties (PSC and PP). The fifth party (ICV) has no a clear position on nationalism. Cluster 5 (magenta) is characterized by the position of CIU and PSC: here CIU has better results than the overall result, and the opposite happens for PSC. Finally, in cluster 6 (yellow) the two right parties (CIU and PP) obtain their best results and the left parties (PSC, ICV and ERC) has lower proportion of votes than in the overall result.

5 Conclusions

We have presented an alternative way to define spatial clusters that also offer the possibility of drawing a kind of LISA map. Our proposal is based on distances between observations. Therefore it is able to be used for data of a very wide list of types, including multivariate or compositional data. The last step of the proposed algorithm consist on doing a standard clustering. Thus all known techniques for selecting the number of clusters can be applied here (see, for instance, Tibshirani, Walther, and Hastie 2001).

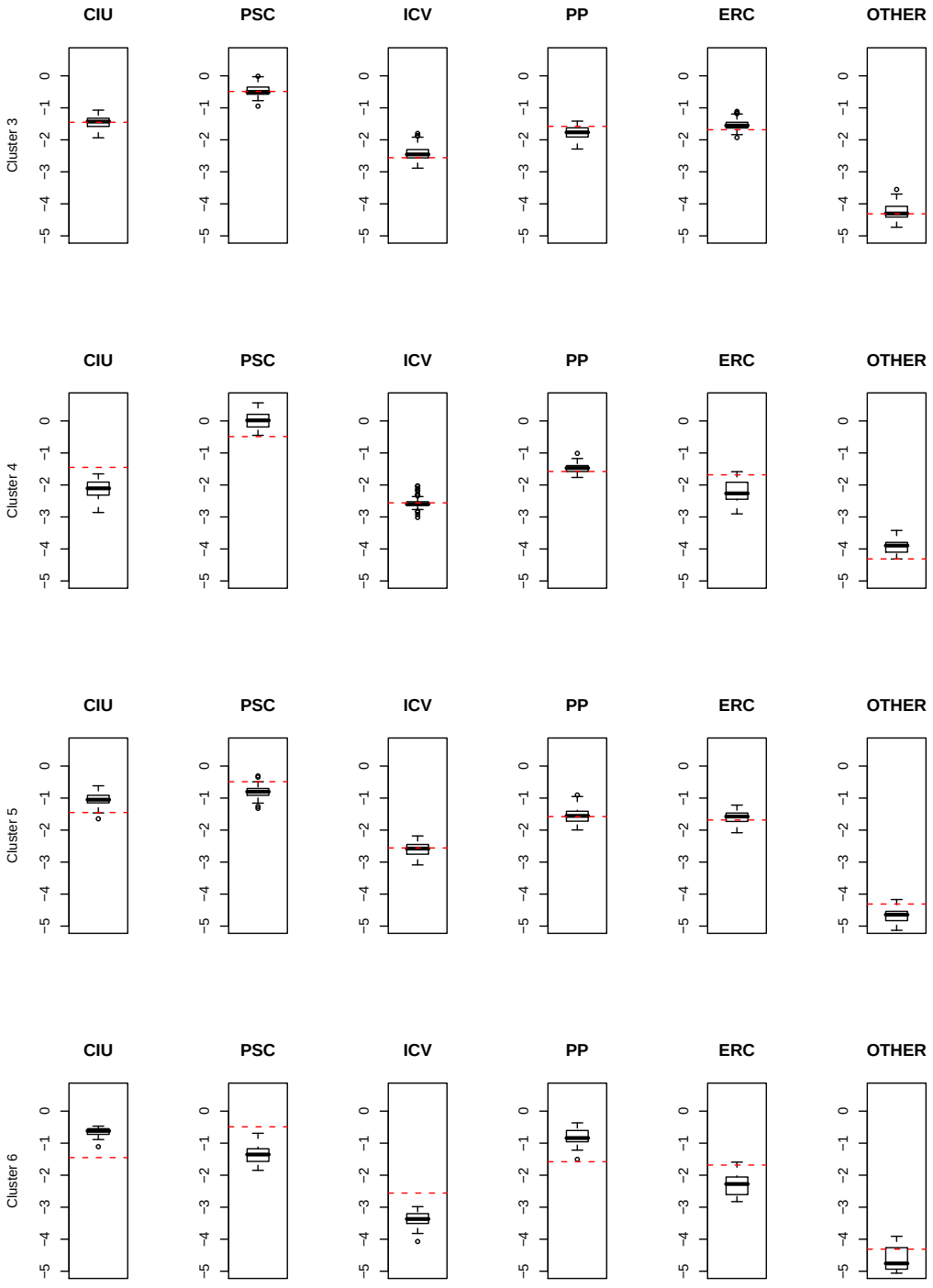

Figure 3: Guide to interpret the distance-based LISA map for the proportion of votes obtained by each political party in 2004 General Elections at ZRPs of Barcelona.

Acknowledgments

Work supported by the Spanish Ministerio de Educación y Ciencia and FEDER grant MTM2006-09920.

References

Anselin, L. (1995). Local indicators of spatial association-lisa. *Geographical Analysis 2*, 93–115.

Anselin, L. (2003). *GeoDA^{TM} 0.9 User's Guide*. Spatial Analysis Laboratory. Department of Agricultural and Consumer Economics. University of Illinois, Urbana-Champaign. Center for Spatially Integrated Social Science.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*, 17–33.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 15*(5), 384–398.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society-B 63*, 411–423.