

SPATIAL HIERARCHICAL CLUSTERING

Alexandre Xavier Ywata CARVALHO¹
Pedro Henrique Melo ALBUQUERQUE¹
Gilberto Rezende de ALMEIDA JUNIOR¹
Rafael Dantas GUIMARÃES¹

- **ABSTRACT:** *This paper studies a methodology for hierarchical spatial clustering of contiguous polygons, based on a geographic coordinate system. The studied algorithm is built upon a modification of traditional hierarchical clustering algorithms, commonly used in the multivariate analysis literature. According to the method studied in this paper, at each step of the sequential process of collapsing clusters, only neighboring clusters (groups of original polygons, i.e. municipalities, census tracts, states) are allowed to be collapsed to form a bigger cluster. Two types of neighborhood are used: polygons with one edge in common (rook neighborhood) or polygons with at least one point in common (queen neighborhood). In this paper, the methodology is employed to create clusters of Brazilian municipalities, for the year 2000, based on a group of socio-economic variables. Several clustering methods are investigated, as well as several types of vector distances. The studied methods were: centroid method, single linkage, complete linkage, average linkage, average linkage weighted, Ward minimum variance and median method. The studied distances were: L_p norm (particularly, L_1 and L_2 norms), Mahalanobis distance and variance corrected Euclidian distance. Finally, a discussion on selection of the number of clusters is presented.*
- **KEYWORDS:** *Cluster analysis, regionalization methods, hierarchical algorithms.*

1 Introduction

The last few decades have witnessed a considerable evolution of spatial data processing techniques. Many of which have been important in helping, for instance, the analysis of territorial heterogeneities, which may be related to social-demographic, cultural, behavioral, climatic or geographic diversities. These spatial differences may arise from great territorial dimensions (the case of some countries) or particularities on the political and economic history of each region. This geographical heterogeneity makes the implementation of public policies aiming regional development a complex task, requiring

¹ Quantitative Methods, Department of Regional and Urban Studies of the Institute for Applied Economics Research (IPEA), SBS – Quadra 1 - Bloco J - Ed. BNDES ; 70076-900 - Brasília - DF – Brasil.. E-mail: alexandre.ywata@ipea.gov.br

techniques, both quantitative and qualitative, that provide adequate spatial disaggregation, identifying, within the national territory, areas with similar characteristics. Once these areas are established, specific policies can be applied.

When it comes to intra-urban development policies, even though territorial dimensions are much smaller in comparison to those of the whole nation, the problem of heterogeneity is present in all big Brazilian cities. Crime-combating policies, for instance, require different strategies for different city areas, given the characteristics of the crimes might differ considerably in each area (see Hirschfield and Bowers, 2001). Similarly, housing policies must also take into consideration social-demographic particularities of each area within a metropolitan region. Therefore, the problem of adequate data processing of spatial heterogeneity, from the public policy viewpoint, is present in the case of intra-urban policies as well.

Besides the use of data processing techniques for treating spatial heterogeneity in studies of regional and intra-urban development, there is a wide array of uses in other areas of knowledge. In treating deforestation data, the researcher might be interested in identifying areas where deforestation has been more intense in the last few decades. Studies in epidemiology require that the researcher be able to detect, with some precision, areas with greater incidence of certain disease, thus leading to the development of more efficient combat strategies. Marketing strategists might be interested in dividing a metropolitan area in smaller, homogeneous areas, according to social-economic characteristics that influence consumption patterns. In this sense, it is possible to make sounder marketing campaigns.

One of the most popular and efficient data processing techniques for identifying homogeneous aggregates inside a heterogeneous group is the cluster analysis. Treating heterogeneous data through clustering them in homogeneous parts is an old technique and is present in most books on multivariate analysis. The idea of grouping data that have similar characteristics goes from the use of one-dimensional clusters, where data on a numeric line are grouped, to more recent developments in the field of spatial clustering. Berkhin (2002) provides a timeline explaining the various clustering techniques. For a general description of clustering algorithms, see Hastie, Tibshirani and Friedman (2001), Khattree and Naik (2000), Berry and Linoff (1997), Alpaydin (2004).

Specifically concerning geographical data, spatial clustering is a powerful technique that can adapt to the most varied cases and thus has developed quickly and become increasingly popular. Considerable part of these clustering techniques is based on probabilistic models, on which Bayesian or Maximum Likelihood procedures are employed. Li, Ramachandran, Movva Graves, Plale and Vijayakumar (2006), for example, have used clustering techniques to improve weather forecast. In other example, Gangnon and Clayton (2003) studied leukemia cases in New York in order to enhance spatial clustering methods by using Monte-Carlo simulations and Markov Chains. Lawson and Denison (2002) provide a collection of articles on spatial clustering applied to various areas. Generally, techniques that are based on probabilistic models try to identify homogeneous groups based on the distribution of events in the space, or just based on a variable of interest (e.g. air temperature, number of leukemia cases per population, etc.)

Concerning probabilistic models aimed at identifying areas with greater intensity of occurrence of an event, the hotspot analysis is very popular in treating crime data (see Eck et al., 2005, Hirschfield and Bowers, 2001). In this case, it is common to consider geographic coordinates data (latitude and longitude where the crimes have occurred).

Techniques to estimate non-parametric density in two dimensions are employed. Hotspots correspond to areas in the space where estimated density is higher. The goal of hotspot analysis is not to divide a certain city, for example, in a partition of homogeneous sub-areas. The goal resides in identifying places where the occurrence of certain event is more pronounced.

Another commonly employed technique for identifying critical areas is the Scan technique (see Kulldorff, 1997, Glaz and Balakrishnan, 1999, Glaz et al., 2001). Instead of using geographical location data (Cartesian coordinates), the Scan method is used in situations in which the available data correspond to geographical polygons (cities, sectors, etc.). The technique consists in moving a fixed-size window through space, in search of areas where the dot density surpasses certain limit, compared to some appropriate distribution to the data in study (e.g. Bernoulli for binary data). It is a computational-intensive technique and has been largely used in the detection of epidemics.

In this paper, we study a methodology for analyzing spatial data that is conceptually different from the hotspot and Scan techniques. Instead of detecting areas where the occurrence of certain event is significantly more pronounced, the technique we study aims at partitioning the region of study (e.g. the national territory) in sub-regions with similar characteristics. The analyzed information are: (i) polygons in a geographic coordinate system (cities, states, sectors etc.); (ii) characteristic variables of each polygon in the geographic coordinate system. We might be interested in analyzing spatial heterogeneity within Brazilian cities (polygons), according to per capita income, longevity, average schooling, house conditions (characteristic variables). We denote the presented method Spatial Hierarchical Clustering and it consists in a modification to traditional hierarchical clustering methods (see Hastie, Tibshirani and Friedman, 2001, Khattree and Naik, 2000, Berry and Linoff, 1997).

Traditional hierarchical clustering methods consist in identifying homogeneous clusters progressively, through metamorphosis (merging or separation) of previous clusters in the sample (see Gower, 1967, Jain and Dubes, 1988, Jain et al., 1999). Hierarchical clustering can be made by agglomeration (starting with as many clusters as objects and merging those to new clusters) or division (starting with only one big cluster and dividing it into new clusters). Cluster metamorphosis is decided considering the distance between objects, which differentiates the clustering methods. The core of the process consists in building a proximity matrix (Jain and Dubes, 1988, Jain et al., 1999) between pair of objects. In the case of hierarchical clustering by agglomeration, close objects are aggregated in clusters and the proximity matrix is updated. The process interacts until the algorithm reaches the determined number of clusters.

According to hierarchical spatial clustering algorithms studied in this paper, traditional hierarchical clustering algorithms are modified in order to limit the identification of strictly contiguous geographical regions with similar social-economic characteristics (or any other group of variables). Forcing contiguity brings some advantages:

- 1) The main goal of clustering analysis is to build homogeneous groups from a geographical area according to certain variables (e.g. social-economic variables). The implicit hypothesis is that the used variables will be enough to describe the characteristics of the studied cities. However, it may be the case that many other relevant variables are not included in the database, which may lead to some loss in

the clustering analysis. On the other hand, it may be expected that many absent variables feature strong spatial correlations, which means that contiguous cities have similar characteristics (see Anselin, 1988, Anselin and Florax, 2000, Pace and Berry, 1997). In this case, using clustering algorithms that impose contiguity might reduce the loss of information.

- 2) Specifically considering works in the branches of regional and intra-urban development, for example, the main goal is identifying homogeneous regions within a country or a sector, where different development policies could be applied. In this sense, contiguity is crucial since the intention is to come up with public policies that focus in geographical areas that have some degree of neighborhood.

The three challenges of the analyst are to choose the distance between data vectors (attributes), to define the final number of clusters, and to choose the clustering method. The clustering methods employed in this paper were: average linkage, centroid, single linkage, complete linkage (unweighted), complete linkage (weighted), Ward minimum variance and median. These clustering methods can be based upon different vector distances: Euclidean (L_2), Manhattan (L_1), L_p (more general case), Mahalanobis and Variance Corrected (see Khattree and Naik, 2000). Finally, we bring the discussion concerning the definition of the number of clusters, using the following criteria: CCC , pseudo- F , pseudo- t^2 , R^2 and semipartial- R^2 (see Sarle, 1983, Khattree and Naik, 2000).

Alternative algorithms to building spatial clusters – clusters with contiguous units – are described, for example, in Maravalle and Simeone (1995) and Maravalle, Simeone and Naldini (1997). These authors suggest algorithms based on the transformation of a map into a graph, and the following reduction of the graph into a generating tree. Applications of clustering algorithms through graphs in Brazil are presented in Assunção, Lage and Reis (2002) and Chein, Lemos and Assunção (2005). Despite the fact that the algorithms presented in Maravalle and Simeone (1995) and Maravalle, Simeone and Naldini (1997) have the same analytical objectives as the hierarchical spatial clustering algorithms presented in this paper, hierarchical algorithms seem to be a more intuitive approach, in which all steps of the process are clear both to users of the new methodology and to readers. Also, the spatial hierarchical clustering methodology allows for the immediate incorporation of various clustering methods (single linkage, complete linkage etc) and different vector distances. Furthermore, traditional criteria for deciding the number of clusters (CCC , pseudo- F , pseudo- t^2 , R^2 and semipartial- R^2) may also be easily incorporated.

Surveys on constrained classification can be found in Batagelj and Ferligoj (1998), Gordon (1996), Duque, Ramos and Surinach (2007). A clustering algorithm under spatial contiguity constraint is provided in Luo (2001); in that paper, the author suggests a K -means method, in which the objective function contains a term to penalize the absence of contiguity between clustered units. The idea of restricting all merges in hierarchical clustering algorithms was implemented in Wiperman (1999); in that manuscript, the author considers two clustering methods: single linkage and complete linkage, and applies the algorithm to the creation of territories for automobile insurance in the province of British Columbia, Canada. The application presented in Wiperman (1999) considered 11 geographic areas, and a detailed analysis of merging steps is provided.

Even though the methodology of constrained hierarchical clustering to ensure geographic contiguity is not new (see Wiperman, 1999, and Duque et al., 2007, for example), this paper complements the literature in the following ways:

- 1) Many types of clustering methods as well as different types of vector distances are addressed, and an exhaustive comparison between them is provided;
- 2) A comparison between different definitions of neighboring polygons is discussed;
- 3) The behavior of usual measures for number of cluster selection is studied;
- 4) The real data application studies the problem of clustering the 5,507 Brazilian municipalities (according to the political map in 2000), and that has been an issue recurrently faced by researchers and policy makers. We hope the exercises presented here will help future applied studies.

The paper is divided in five sections, including this introduction. The second part discusses the methodology used in the formation of homogeneous groups of cities (clusters), the discussion on clustering methods, the calculations of different vector distances, and the discussion on the criteria for selecting the number of clusters. The third section features a comparison between the various clustering methods and the different distances used, based on a case study using social-economic data from Brazilian cities on year 2000. The fourth section presents some conclusions.

2 Methodology

In this section, we describe the algorithm for formation of homogeneous groups of cities, census tracts, or any other geographic polygons. As we shall see in more detail later, the algorithm studied in this paper consists in a modification of widely-recognized hierarchical clustering methods.

2.1 Algorithm for forming homogeneous spatial groups

In traditional clustering algorithms (hierarchical or not), when geographical units are grouped, the homogeneous groups created are not necessarily formed by strictly neighboring cities. It may occur that within the same cluster we find geographically separated cities. The formation of homogeneous groups of cities with members that are not necessarily contiguous may not be a problem in many applications. Indeed, it may be the case that the analyst or researcher is interested in identifying regions on the periphery of São Paulo, for example, that are similar in social-economic characteristics to the regions in the city center. In the following topic, we discuss modifications to the traditional clustering method, in order to incorporate the restriction of contiguous geographical units (e.g. cities, sectors, states).

2.2 Spatial hierarchical clustering algorithms

Given the great applicability of clustering methods, the literature in the area has evolved remarkably, so as more algorithms have been created, providing more efficiency

and more suitability for different situations. Hastie, Tibishirani and Friedman (2001), Jain and Dubes (1988), Jain et al. (1999), discuss these algorithms. The clustering algorithms can be divided in three main categories: 1) combinatorial algorithms; 2) mixture models; 3) mode seeking. The latest two categories are based in some form of probabilistic models for the data generating process. Combinatorial algorithms, on the other hand, are based on heuristic rules to search for best groupings, trying to minimize some overall variability criteria.

The algorithm employed in this work can be classified as a combinatory algorithm and has a hierarchical cluster-formation structure. The following steps describe the modification on the hierarchical algorithm, in order to satisfy the neighborhood restriction between the units of each homogeneous group. In order to make the presentation easier, the described steps refer to clustering of cities; however, the discussion is immediately applicable to any other kind of geographic unit.

- 1) Let C be an initial database of N geographical units (the empirical study presented in this paper uses municipalities as these units). Initially, each of these N observations is an isolated cluster and has a set of attributes (variables) $[x_{i,1}, x_{i,2}, \dots, x_{i,m}]$. For each of these N units it is necessary to determine the list of neighboring objects, according to some spatial criterion. Two definitions of neighborhood were investigated. On the first case, cities are considered neighbors if they share at least one edge (considering a geographic coordinate system) – this type of neighborhood is known in spatial statistics literature as rook neighborhood. On the second case, cities are considered neighbors if they share at least one point – this type of neighborhood is known as queen neighborhood. Obviously, queen neighborhood is less restrictive than rook neighborhood.
- 2) Calculate the vector distance between all pairs formed by strictly neighboring elements from the list of N units, and build a (symmetric) proximity matrix.
- 3) Let I and J be the two geographical units featuring the smallest vector distance between them, with the restriction that I and J are neighbors. We group the pair I and J in one single cluster. The number of clusters is now $N-1$.
- 4) On the definition of the new cluster, formed by units I and J , it is necessary to combine the lists of neighbors. Therefore, a new list of neighbor cities² will be formed from the union of the neighbor list of city I and the neighbor list of city J .
- 5) For the new $N-1$ clusters, after the junction described on step 3, the proximity matrix has to be updated. Updating the proximity (or distance) matrix will depend on the clustering method being employed. For the single linkage method, for example, the distance between two clusters I and J is the minimum of the vector distances between all pairs of variable vectors in the two clusters (Jain et al., 1999). For the complete linkage method, on the other hand, the distance between two clusters is the maximum of the vector distances between all pairs of vectors. Section 2.3 presents a detailed discussion on the clustering methods.

² Two clusters A and B are considered neighbors when at least one city from A is neighbor of one city of B .

- 6) Repeat steps 3 to 5 until there is only one cluster left, which will contain all N original geographical units.

Just as it happens in the case of traditional hierarchical clustering, at the end of the process one has a tree that characterizes the groupings that take place at each step of the algorithm. Again, the researcher can use some of the traditional indicators (for example, CCC , pseudo- F and pseudo- t^2) for the choice of the most appropriate number of groups. However, since the algorithm used herein uses substantial modifications on the traditional hierarchical clustering algorithms, the properties of these statistical indicators not necessarily match those from traditional clustering (not spatial), thus indicating the need for posterior studies on the behavior of these indicators. On the other hand, the direct use of statistical criteria not necessarily leads to a number of clusters that makes sense according to the objectives of each study. One can opt to choose the number of groups whose economic interpretation makes more sense. The choice of the number of clusters via subjective criteria was used, for instance, in Chein, Lemos and Assunção (2005), where there were selected 100 clusters for all Brazilian territory. Anyway, Section 3.3 offers a discussion on the behavior of some criteria for selecting the number of clusters commonly found in the literature.

2.3 Clustering methods

In this section, we present some of the clustering methods commonly found in the hierarchical clustering literature. The list of methods below is not exhaustive and the reader can use the references provided in this paper in order to be familiar with other algorithms. Besides the clustering methods presented below, this section presents a list of vector distances. By combining the clustering methods to the vector distances, it is possible to come up with a great variety of alternatives. These many combinations are studied in Section 3.2.

Average linkage (unweighted)

According to the average linkage method, also known as McQuitty method, the distance between two clusters is the average distance between all pairs of variable vectors drawn from the two clusters. For non-spatial agglomeration clustering, this method tends to join clusters with low variance and it is slightly biased towards producing clusters with the same variance. The expression for the distance measure between clusters K and L is given by

$$D_{K,L} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j),$$

where $d(x_i, x_j)$ is a distance between vectors x_i and x_j (containing the characteristics of polygons i and j), as discussed in Section 2.4. Another way of implementing the hierarchical clustering algorithm is through the update of the proximity (or distance) matrix between clusters. Every time a new cluster C_M is created through the joint of clusters C_L and C_K existent from the previous step, the proximity matrix is updated in order to consider distances to the new cluster. This update can be made directly through

the distances from the previous matrix, utilizing combinatorial formulas. Consider a cluster C_j . For the average linkage method, the distance between any cluster C_j and the new cluster C_M can be obtained from the previous distances, using the following expression

$$D_{J,M} = \frac{1}{2}D_{J,K} + \frac{1}{2}D_{J,L}$$

The average linkage method considers an average of all members from the clusters whose distance is being calculated. Therefore, this method is less influenced by extreme values, as is the case of single linkage and complete linkage methods.

Single linkage

According to the single linkage method (or closest neighbor method), the distance between two clusters is the minimum of the distances between all pairs of variable vectors drawn from the two clusters. For non-spatial agglomeration clustering, it has many desirable technical properties, but performs badly in Monte-Carlo experiments. The distance measure between clusters is defined as

$$D_{K,L} = \min_{i \in C_K, j \in C_L} d(x_i, x_j)$$

The distance measure between a cluster CJ and the new cluster CM can be updated by the combinatorial formula

$$D_{J,M} = \frac{1}{2}D_{J,K} + \frac{1}{2}D_{J,L} - \frac{1}{2}|D_{J,K} - D_{J,L}|$$

Since it does not impose restrictions on the shape of the clusters, this method sacrifices the possibility of obtaining compact clusters, with the advantage of allowing for irregular or elongated clusters. In non-spatial agglomeration clustering, the single linkage method also tends to cut the tails of the distributions before separating the main clusters.

Complete linkage method

According to the complete linkage method, the distance between two clusters is the maximum of the distances between all pairs of variable vectors drawn from the two clusters:

$$D_{K,L} = \max_{i \in C_K, j \in C_L} d(x_i, x_j)$$

The distance measure between a cluster CJ and the new cluster CM can be obtained by the combinatorial formula

$$D_{J,M} = \frac{1}{2}D_{J,K} + \frac{1}{2}D_{J,L} + \frac{1}{2}|D_{J,K} - D_{J,L}|$$

It is strongly biased towards producing compact clusters with similar diameters, and can be severely distorted by moderate outliers. It is a method that assures that all items in a cluster are at a minimal distance from one another.

Ward's minimal variance method

Ward's minimal variance method is biased towards generating clusters of the same size. The method is based on the sum of squares of errors (SSE) of each cluster (sum of squares of deviations from the cluster centroid). We add the SSEs of all G clusters, generating the TSSE. The method consists in analyzing all possible pairs of joined clusters, identifying which joint produces the smallest increase in SSE. In this method, the distance between two clusters is the ANOVA sum of squares between two clusters for all variables. At each generation, it minimizes the intra-cluster sum of squares obtainable by the joint of two clusters. Frequently, it is advised to use the ratio SQE/SQET over the absolute SQE. The distance measure between clusters CK and CL measure is defined as

$$D_{K,L} = \frac{d(\bar{x}_K, \bar{x}_L)^2}{\left(\frac{1}{N_K} + \frac{1}{N_L}\right)},$$

where \bar{x}_K and \bar{x}_L are the mean vectors within clusters CK and CL respectively. It is a method that aims at maximizing the likelihood at each level of hierarchy under the hypotheses of mixture of multivariate normal distributions, equal covariance spherical matrixes and equal sample probabilities. In non-spatial agglomeration clustering, it tends to join clusters with small number of observations and is strongly biased towards producing clusters of the same shape and number of observations. It is also influenced by outliers.

Centroid method

Developed by Sokal and Michener in 1958, centroid linkage method considers the distance measure between two clusters as the square of the distance between the centroids of the clusters:

$$D_{K,L} = d(\bar{x}_K, \bar{x}_L)^2$$

being it a comparison of averages, outliers have little influence. In other aspects, it can be less efficient than other methods such as average linkage and Ward. The bigger of two joined clusters tends to dominate the new cluster.

Average linkage weigthed

Weighted average linkage method differs from the original average linkage method due to different weights inserted in the combinatorial formula. The new formula is

$$D_{J,M} = \frac{n_K}{n_L + n_K} D_{J,K} + \frac{n_L}{n_L + n_K} D_{J,L},$$

where n_K and n_L are the numbers of observations on clusters CK and CL respectively.

Median method

The median method has the combinatorial formula for updating the distance matrix given by

$$D_{J,M} = \frac{1}{2} D_{J,K} + \frac{1}{2} D_{J,L} - \frac{1}{4} D_{K,L}.$$

This method was developed by Gower (1967).

2.4 Types of vector distances

This section presents the types of topological distances that were used in this paper. Let x and y be vectors containing the characteristic variables for two polygons. In the case of municipality data, for instance, x and y may be vectors containing the variables of per capita income, average longevity, average schooling, Gini index and so on. Let x_i denote the i -th element (scalar) of vector x .

Lp norm

Lp norm is the most commonly used type of metric, where parameter p is a penalization factor for outliers. According to the Lp norm, the distance between vectors x and y is defined as

$$d(x, y) = \sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p}$$

Euclidean distance (L2 norm)

The Euclidean norm or distance is a particular case of the Lp norm for $p = 2$. The Euclidean distance between vectors x and y is defined as

$$d(x, y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}$$

L1 Norm (Manhattan, taxicab or city block distance)

The Manhattan distance is a particular case of the Lp norm for $p = 1$. The Manhattan distance between vectors x and y is defined as

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Mahalanobis distance

The Mahalanobis distance can be seen as a generalization of the quadratic form of the Euclidean distance (see Mahalanobis, 1936). It differs from the Euclidean distance since it is invariant to scale, that is, it does not depend on the scale of the measures in study. Let Σ be the variance-covariance matrix for the variables characterizing the polygons to be grouped. For data vectors with v variables, the matrix Σ is symmetric, with dimension $v \times v$. In real applications, the matrix Σ is calculated directly from the database. Once the matrix Σ is calculated, the Mahalanobis distance between x and y is defined as

$$d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}.$$

In the particular case where the variance-covariance matrix is the identity matrix, the Mahalanobis distance coincides with the Euclidean distance.

Variance corrected Euclidean distance or normalized Euclidean distance

The variance corrected distance corresponds to a variation of the Mahalanobis distance, in which the matrix Σ is diagonal (elements outside the main diagonal are null) and the elements of the main diagonal are the variances of the v variables of the data vector, characterizing each polygon. In this case, the use of the variance corrected distance corresponds to the use of the Euclidean distance, where all variables on the original database are divided by their respective standard deviations. One advantage of using the variance corrected distance is the correction for scale problems between variables. Depending on the clustering method used, variables with larger scale may end up having an artificially bigger weight when forming the clusters. On the other hand, as discussed in Hastie, Tibshirani and Friedman (2001), bringing all variables to the same scale by dividing them by their standard deviations does not necessarily bring advantages in identifying homogeneous groups. See Milligan and Cooper (1987) for a discussion on the normalization of variables for clustering.

2.5 Criteria for selecting the number of clusters

In Section 2.3, we presented the sequential algorithm for forming new clusters from a set of clusters from the previous step. This algorithm continues until there is only one cluster left (or a minimum number of clusters, respecting the neighborhood restriction). In this process, we can build various measures to help on the selection of the number of clusters to be used on the study of interest of each researcher. One of the most popular indexes is Sarle's CCC (Cubic Clustering Criterion) (see Sarle, 1983), which, in non-spatial hierarchical clustering algorithms, tests the H_0 hypothesis that the data were sampled from a uniform distribution, against the H_1 hypothesis that the data were sampled from a mixture of spherical multivariate normal distributions with equal variances

and sample probabilities. Positive values for CCC cause the rejection of H_0 . We then plot the values of CCC against the number of clusters and look for peaks where CCC exceeds 3. The formula for the CCC is given by

$$CCC = \log \left[\frac{1 - E[R^2]}{1 - R^2} \right] \times v,$$

where v is the number of variables in the database, R^2 is the R^2 criterion, $E[R^2]$ is the expected R^2 and $\log[.]$ is the natural logarithm. The formulas for R^2 and expected R^2 criteria are presented on Sarle (1983).

Other very popular criteria are pseudo-t2, semipartial-R2 and pseudo-F. The latter measures the separation between clusters at the current hierarchy level. High values for pseudo- F indicate that the average vectors for each cluster are different, that is, each cluster is significant on that configuration. Therefore, one way of using pseudo- F is to look for peak values on the chart of the pseudo- F versus the number of clusters; the chosen number of clusters corresponds to the peak on the pseudo- F . On the other hand, if the pseudo- t^2 statistic is high at a certain level of the process of joint of two clusters, then these clusters should not be joined, since their average vectors can be considered different. Therefore, the existing literature recommends that we search for peak values on the sequence of pseudo- t^2 statistics and use the number of clusters immediately above the number of clusters corresponding to the peak. Finally, the semipartial- R^2 criterion calculates the proportional reduction on the variance due to the joint of two clusters (C_k and C_l). Small values indicate that the two clusters can be considered only one, whereas high values indicate that the clusters are probably different. For more detail on the criteria for selection the number of clusters, see Khattree and Naik (2000).

3 Case study

This section presents a case study to investigate the properties of the clustering methods presented on Section 2.3 and the vector distances discussed on Section 2.4. The used database is one that features social-economic information on Brazilian cities, based on the municipality data from year 2000. There is a total of 5507 municipalities. The choice of this database was motivated by the need for studies and public policies focused on homogeneous and contiguous areas within Brazilian territory. Therefore, the case study is useful not only to present some general ideas on the various clustering methods, but also to point out which method is most appropriate on specific regional development studies. Section 3.1 presents a discussion on the used database. Section 3.2 presents the results of the spatial clustering exercise, comparing the different clustering methods. Section 3.3 presents the results on the investigation of the criteria for selecting the number of clusters.

3.1 Database

The data utilized in this paper were extracted from the Brazilian Census 2000 (IBGE, 2002) and the “Atlas do Desenvolvimento Humano” (IPEA, PNUD and FJP, 2003). For a better discussion on the used data, see Carvalho, Albuquerque, Mota and

Piancastelli (2008). The selected social-economic characteristics for the Brazilian municipalities were:

- a) Unemployment rate of the municipality (employed population divided by total population).
- b) Percentage of the municipality's population in urban areas.
- c) Demographic Variables: longevity and birth rate of the municipality in year 2000.
- d) Urban infrastructure and condition of the residences: percentage of residences with public lighting, identification (ZIP code), sewage, piped water, paved streets, power and waste management.
- e) Academic performance: percentage of children from 5 to 6 years old at school; percentage of children from 7 to 14 years old with access to elementary school; percentage of teenagers from 15 to 17 years old with access to high school; percentage of people from 18 to 24 years old with access to higher education; percentage of children from 7 to 14 years old with more than one year of delay on education; percentage of elementary school teachers with higher education; average of years of education for 25 year old people or older; percentage of illiterate people 15 year old or older; percentage of illiterate people from 10 to 14 years old.
- f) Per capita income of the residence, percentage of per capita income derived from work earnings, percentage of per capita income derived from government help, in 2000.
- g) Percentage of poor people in the municipality in 2000.
- h) Income inequality (Gini index) in 2000.
- i) Variables related to public health: mortality for children up to 1 year old and up to 5 years old in 2000 and probability of survival up to 60 years in 2000.
- j) Murder rate in the municipality in 2002.

Besides the used data described above, we have used the municipality map of 2000 in Brazil, containing information from a geographic coordinate system. This information was used to build the neighborhood structure between cities. Due to the presence of islands in the national territory, the process of sequential aggregation of clusters went on until the minimum number of three clusters (when there is no more neighborhood, be it rook or queen).

3.2 Effects of the clustering algorithms and the types of distance between vectors

Figures 1 presents the clusters formed using the Euclidian distance and the Ward clustering method. Figure 2 presents the clusters for the Mahalanobis distance, with the average linkage algorithm. The reader can refer to Carvalho et al. (2009) for all the other cluster maps (considering different clustering methods, different vector distances and different neighborhood definitions). In order to allow for the comparison between methods, it was used 100 clusters in all maps – which was the number used in Chein et al. (2005). Carvalho et al. (2008) used 91 groupings. Appendix 1 presents the Box-plots evaluating the size of the formed clusters.

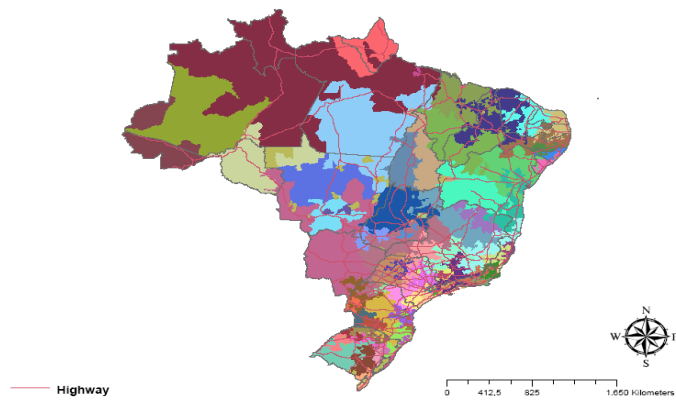


Figure 1 - Ward's method, using Euclidian distance, and rook neighborhood (100 clusters).

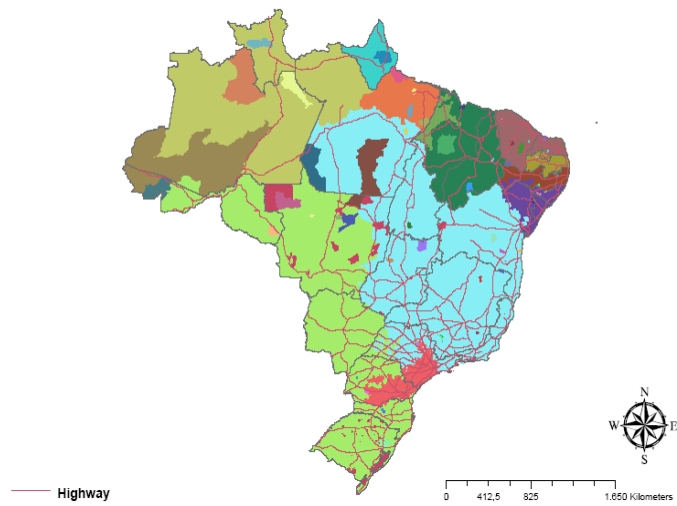


Figure 2 - Average linkage method, using Mahalanobis distance, and queen neighborhood (100 clusters).

From the constructed maps, we noticed the single linkage, average linkage, average linkage weighted, centroid and median methods tend to form clusters very different in terms of numbers of geographical units. This conclusion is supported by the Box-plots on Appendix 1. The methods of complete linkage and especially Ward are the ones that tend to form clusters with the most homogeneous sizes – this fact was somehow expected, given the discussion on Section 2.4.1. On the other hand, among the five distances studied, the Manhattan distance (L_1) is the one that provides more homogeneous clusters in terms of number of geographic units. It is noticed that for the general distance L_p , the closer to $p=1$, the more homogeneous are the sizes of the clusters. This might be related to the fact that L_1 distance gives less weight to outliers than the Euclidean distance, for example.

3.3 Behavior of the criteria for selecting the number of clusters

Appendix 2 presents the charts for the selection criteria of the number of homogeneous groups. The charts refer to the seven methods, considering only the Euclidean distance and only the Rook neighborhood. For other distances and the Queen neighborhood, the conclusions are similar. The criteria presented on the charts are the R^2 , the *CCC*, the pseudo- t^2 statistic and the pseudo- F statistic. The semipartial- R^2 criterion presented similar behavior to the pseudo- t^2 and was not shown.

On the chart for the R^2 statistic, we also show the expected R^2 . This value is the same for all clustering methods. It is noticed that the expected R^2 is considerably higher than the R^2 for all seven clustering methods. This was already expected: expected R^2 is calculated (see Sarle, 1983) under the condition of no restriction of contiguity between clusters, while the R^2 for spatial clustering tend to be smaller than the R^2 for traditional clustering algorithms. This fact causes values for the *CCC* to be very negative (see formula for the *CCC*); therefore, the rule of choosing the number of clusters when the *CCC* has peaks above three is no longer valid. This fact opens a question for future research: how to calculate the R^2 more appropriately for the case of spatial hierarchical clustering. A more appropriate calculation for R^2 would result in a correction for the *CCC*, which may still be conservative for spatial clustering methods. One alternative is to employ a Monte Carlo approach, as described in Jain and Dubes (1988).

The pseudo- t^2 and pseudo- F criteria present a chart behavior that is similar to the case of non-spatial hierarchical algorithms. Notice that the pseudo- F criterion presents a sequence of not very abrupt variations, containing local peaks, which may indicate the number of clusters to be chosen. On the other hand, the sequence for pseudo- t^2 criterion presents some points of very high peaks. Due to this fact, we have used the logarithmic scale for the vertical axis on the charts for this criterion. Literature suggests the choice of $u + 1$ clusters, where u is the number of clusters that corresponds to a peak. Therefore, the pseudo- t^2 seems to provide clearer evidence for the choice of the number of clusters.

3.4 Effects of the type of neighborhood (Rook versus Queen)

Observing the graphs in Appendix 1, we notice that the results for both rook and queen neighborhoods are very similar considering the number of municipalities in each cluster. The similarity is also present for the criteria for selection of the number of clusters (the graphs for the Queen neighborhood are very similar to the graphs for the Rook

neighborhood – see Carvalho et al., 2009). On the other hand, the constructed maps (see Carvalho et al., 2009) show that the groups formed with the two neighborhoods can be substantially different, even when one uses the same clustering method and the same vector distance.

At a first glance, the use of queen neighborhood, by requiring only one shared point to characterize neighborhood, can lead to the formation of more irregular groups than those formed with rook neighborhood. On the other hand, since it is less restrictive, queen neighborhood is expected to provide clusters that feature smaller variability intra-clusters. In the exercise presented on the next section, however, this smaller variability for the queen neighborhood is not verified.

3.5 Comparison with political groupings of Brazilian municipalities

In this section, we present a comparison between the groups obtained via spatial hierarchical clustering and existing political divisions in Brazil. The used groups, for comparison purposes, are: micro-regions, meso-regions and federation units (states). There are a total of 27 states, 558 micro-regions and 137 meso-regions. Therefore, in order to compare the results to the political divisions, we have used configurations with 27, 558 and 137 groups. For each of these three numbers of groups, we calculate the sum of squares of the deviations to the mean of each cluster (*TCSS*). This was used as a performance index for the group configuration – it provides an idea of the intra-cluster variability for each method³. The expression for *TCSS* is given by

$$TCSS = \sum_{k=1}^G \sum_{i \in C_k} \sum_{l=1}^v [x_{k,i,l} - \bar{x}_{k,l}]^2,$$

where v is the total number of variables on the database, G is the number of clusters ($G = 27, 137$ or 558), C_k is the set of municipalities in cluster k , $x_{k,i,l}$ is the variable l in city i , and $\bar{x}_{k,l}$ is the average of variable l , within cluster k . In order to compare the various clustering methods and the political divisions, the reported values are the relative variability of each method versus the variability for the corresponding political division. In this case, the relative variability $\Delta TCSS_{Method}$ is given by

$$\Delta TCSS_{Method} = 100 \times TCSS_{Method} / TCSS_{Political\ Division}.$$

Table 1 below presents the performance measure for the different clustering methods and the different vector distances. Note that the clusters obtained by spatial hierarchical clustering with Ward's method have smaller total variabilities than the political divisions. This is valid for micro-regions (558 clusters), meso-regions (137 Clusters) and states (27 clusters). In the case of micro-regions, the Ward method generates a reduction of up to

³ Another measure to evaluate the performance of the different clustering methods and the different vector distances is to use a variation of the *TCSS*, where the distance between groups is normalized by the number of municipalities in each group, so as to provide a comparison less dependent on the cluster sizes. Nonetheless, we decided to use the *TCSS* as defined above because, from a policy maker standpoint, obtaining clusters with very unequal sizes may be undesirable: large clusters may imply in less efficient public policies – see discussion on Section 4.

almost 50% in variability. For the complete linkage method, the measure of intra-cluster variability is also smaller, generally, than that of the political division; the exception is when we use Mahalanobis distance. For the other five methods, the results are less encouraging; the resulting variability with single linkage, average linkage, average linkage weighted, median and centroid is bigger than that given by the political divisions in the case of states and meso-regions. For micro-regions, however average linkage and average linkage weighted present, for most of the distances, smaller variability than that of the political divisions.

Table 1 - Percentage of total variability of the different clustering methods, in comparison to the political divisions of Brazilian municipalities

Spatial Hierarchical Clustering Method		Number of Groups					
Clustering method	Distance between vectors	27		137		558	
		Rook	Queen	Rook	Queen	Rook	Queen
Single Linkage	Euclidean	214	214	259	259	297	297
	L_1 - Manhattan	216	216	265	265	305	305
	L_p (p=1.5)	215	215	263	263	302	302
	Euclidean (normalized)	217	217	265	265	301	301
	Mahalanobis	216	216	265	265	296	296
Complete Linkage	Euclidean	91	96	77	80	56	56
	L_1 - Manhattan	93	91	77	77	59	58
	L_p (p=1.5)	97	92	74	77	56	56
	Euclidean (normalized)	101	97	84	83	66	66
	Mahalanobis	156	187	100	97	85	84
Average Linkage (unweighted)	Euclidean	215	218	133	148	97	106
	L_1 - Manhattan	195	215	175	127	88	84
	L_p (p=1.5)	189	127	136	118	90	84
	Euclidean (normalized)	118	147	137	139	100	96
	Mahalanobis	164	137	117	116	98	98
Average Linkage (weighted)	Euclidean	204	203	110	106	71	71
	L_1 - Manhattan	110	112	112	115	73	75
	L_p (p=1.5)	205	205	114	114	74	71
	Euclidean (normalized)	209	209	125	125	93	91
	Mahalanobis	212	212	251	249	150	153

Continua

Table 1 - Percentage of total variability of the different clustering methods, in comparison to the political divisions of Brazilian municipalities (continuação)

Spatial Hierarchical Clustering Method		Number of Groups					
Clustering method	Distance between vectors	27		137		558	
		Rook	Queen	Rook	Rook	Rook	Rook
Median	Euclidean	199	219	244	259	223	150
	L_I - Manhattan	219	220	251	154	201	158
	L_p (p=1.5)	218	198	175	240	183	255
	Euclidean (normalized)	217	217	193	170	191	176
	Mahalanobis	218	215	160	162	161	173
Centroid	Euclidean	204	205	119	109	101	96
	L_I - Manhattan	212	212	252	251	132	132
	L_p (p=1.5)	204	204	121	121	118	118
	Euclidean (normalized)	211	211	133	134	132	133
	Mahalanobis	212	212	257	257	283	284
Ward	Euclidean	77	76	66	66	53	52
	L_I - Manhattan	80	79	72	71	57	56
	L_p (p=1.5)	78	78	66	66	53	53
	Euclidean (normalized)	85	85	76	76	63	63
	Mahalanobis	90	89	86	86	77	76
Political divisions of cities	States	100		---		---	
	Meso-regions	---		100		---	
	Micro-regions	---		---		100	

Elaborated by the authors.

The fact that the complete linkage and specially the Ward methods have generated clusters with smaller variability than that of the political divisions does not mean the numerically obtained clusters are better than the existing political divisions. However, the idea is that spatial hierarchical clustering can be more adequate than existing political divisions when the researcher is interested in studying groups that are as homogeneous as possible. Besides that, the groups can be generated according to a set of specific variables on the interest of the researcher.

For most methods, the spatial hierarchical clustering algorithms do not produce more homogeneous groups, in terms of the used intra-cluster variability measure, than the political divisions. This can be explained by the fact that these methods tend to form very unequal groups in terms of the number of cities (see Section 3.2). Due to this, some of the

clusters contain more than half of the Brazilian municipalities, for which aggregate variability is very high. At the aggregate level, the total variability for all groups turns out to be much higher than that for political groups (micro-regions, meso-regions or states).

The hierarchical clustering method that has presented the smallest variability measure was Ward, using the Euclidean distance. This fact was already expected, because, as stated on traditional (non-hierarchical) clustering literature, Ward's algorithm tends to form clusters by minimizing aggregate variance. Besides, the used variability measure (*TCSS*) was built intrinsically using the Euclidean norm. Possibly, had we used other variability measures, more linked to other types of distance, the resulting variability could be smaller in the case of clustering methods based on the Manhattan distance, for instance, or the Mahalanobis distance. Anyway, the present exercise reinforces the use of spatial hierarchical clustering algorithms, when possible; with some preference to Ward's method.

It is worth noting that, even allowing for more flexibility on cluster formation, the queen neighborhood does not necessarily imply groups with smaller variability than those obtained using rook neighborhood. This can be explained by the hierarchical nature of the clustering algorithms used herein. Had we used an algorithm that minimizes explicitly total variability, possibly the queen neighborhood would incur in a smaller variability than that of the rook neighborhood. However, the hierarchical algorithm is not an algorithm of explicit minimization, which means the formed clusters do not necessarily correspond to the clusters formed by an algorithm of maximization of some objective function.

Conclusions

This paper studied a methodology for the hierarchical formation of spatial groups. The studied algorithm is basically a modification of the traditional hierarchical clustering algorithm: at each step of the process of merging two clusters to form a new one, it is imposed that the joint can only be made between spatially neighbor clusters (according to some geographic coordinate system). In this case, we have considered two types of neighborhood: rook (polygons with one edge in common) and queen (polygons with one point in common). Due to the fact that the spatial hierarchical clustering algorithm studied in this paper is an extension of the traditional spatial hierarchical clustering algorithm, we can incorporate various clustering methods from the known literature. The clustering methods employed in this paper are: Ward minimum variance, centroid, median, single linkage, complete linkage, average linkage, average linkage weighted. Besides, we have employed different definitions for the distance between vectors. The distances used are: L_p , Euclidean, L_1 , Mahalanobis and variance corrected Euclidean.

By performing a detailed study using social-economic variables for Brazilian cities in year 2000, this paper addresses various properties of the clustering algorithms, the distances between vectors and the neighborhood types. The results show that the Ward and complete linkage methods tend to provide clusters with similar sizes. On the other hand, the other methods tend to produce clusters with very different sizes. This suggests that the use of Ward and complete linkage can be more advisable when trying to identify homogeneous and geographically close areas.

The fact that single linkage, average linkage average linkage weighted, centroid and median methods result in the formation of groups that are very different in terms of size

does not necessarily exclude them from practical applicability. The formation of a few very big groups together with many smaller others can be used as a tool to identify groups of few municipalities that are significantly different from the others, in relation to some variables of interest. Therefore, these five methods can be used as alternatives for the scan methods (see Kulldorff, 1997, Glaz and Balakrishnan, 1999, Glaz et al., 2001), for instance.

For the distribution of the number of cities in each geographical group and for the behavior of the cluster number selection criteria, the obtained results were very similar, using both rook and queen neighborhoods. In terms of geographical visualization, though, the queen neighborhood allows for more flexibility on the formation of the groups, since it requires only one shared point to characterize neighborhood between two geographic units. The rook neighborhood, by requiring a shared edge in order to characterize neighborhood, allows for the formation of groups with less irregular shape. Due to the hierarchical nature of the clustering algorithms studied in this paper, clusters formed with queen neighborhood do not necessarily present smaller variability than those formed with rook neighborhood.

The paper also presents a comparison between the groups obtained through spatial hierarchical clustering and the existing political division in Brazil. The used political divisions were: micro-regions, meso-regions and states. There are a total 27 states, 558 micro-regions and 137 meso-regions. Therefore, in order to compare the results for the clusters and for the political divisions, configurations with 27, 558 and 137 clusters were used. For each of these three numbers of groups, the sum of the squares of the deviations to the mean of each cluster was calculated, as a measure of total intra-cluster variability. The results show the capacity of the complete linkage and mainly Ward's method to generate groups with smaller variability than that of the political divisions. In the case of micro-regions, for instance, Ward's method allowed for the formation of homogeneous groups with half the variability of the political divisions. For the other five methods, due to the tendency of forming clusters of very unequal sizes, the obtained total variability resulted, in many cases, higher than that of the political divisions.

Many questions are still open for investigation, leading to the improvement of the method studied herein. First, it would be interesting to extend the distances between vectors in order to incorporate other kinds of variables. The distances studied in this paper are more appropriate for continuous variables. There could be modifications on the procedures of spatial hierarchical clustering algorithms in order to deal with binary data or categorical data in general. Besides, it would be interesting to develop specific methodologies to deal with the combination of different types of variables (categorical and continuous, for example) within the same database (Hiu et al., 2001, present a clustering algorithm for different types of data; their proposed algorithm is not based on hierarchical procedures though).

The algorithms studied in this paper are purely heuristic and are not based on intrinsic probabilistic models, from which Bayesian estimation procedures or maximum likelihood methods can be employed. Even the heuristic procedures presented herein do not necessarily follow the same theoretical behavior studied on papers on non-spatial hierarchical clustering. There emerges, then, a study branch for new spatial clustering methods based on probabilistic models. A possible branch of probabilistic models that can be adapted to account for spatial contiguity are the semi-supervised methods, with restrictions between pairs of observations (these methods are described in Basu et al.,

2008, and Chapelle et al., 2006). Another topic for further investigation is the formal evaluation of the properties of the heuristic procedures studied in this paper.

Finally, the selection of the number of clusters is another question for further investigation. The behavior of traditional methods for selecting the number of clusters, in non-spatial algorithms, was briefly addressed. The properties for these usual criteria in the case of non-spatial algorithms probably do not apply to the heuristic methods of spatial clustering. Indeed, an important point is that very often, in non-spatial clustering, the interest is in reaching a reasonably small number of groups (e.g. 10); with few groups, it becomes easier to interpret them and to generate typologies that may eventually become popular. On the other hand, in spatial clustering, reaching a small number of groups is not necessarily the goal. The objective of spatial clustering might be to identify close and homogeneous cities in order to make more efficient public policies. For this reason, it is interesting to have clusters that are not so big in order to avoid long distances between cities within the same cluster. Therefore, policy makers might be interested in identifying many homogeneous groups with a small number of geographical units inside them.

CARVALHO, A. X. Y.; ALMEIDA JUNIOR, G. R.; ALBUQUERQUE, P. H. M.; GUIMARÃES, R. D. Clusterização espacial hierárquica. *Rev. Bras. Biom.*, São Paulo, v.27, n.3, p.412-443, 2009.

- RESUMO: Este artigo analisa uma metodologia para clusterização hierárquica espacial de polígonos contíguos, com base em um sistema de coordenadas georeferenciadas. O algoritmo estudado é construído a partir de uma modificação do algoritmo de clusterização hierárquica tradicional, comumente utilizado na literatura de análise multivariada. A cada passo do processo sequencial de junção de *clusters*, impõem-se que somente conglomerados (grupos de polígonos originais, como municípios, estados ou setores censitários) vizinhos possam ser unidos para formar um novo cluster maior. Foram definidos como vizinhos polígonos que possuem um vértice em comum (vizinhança do tipo Queen) ou uma aresta em comum (vizinhança do tipo Rook). Este artigo apresenta aplicações da metodologia para clusterização dos municípios brasileiros, no ano de 2000, com base em um conjunto de variáveis sócio-econômicas. Diversos métodos de clusterização são estudados, assim como diferentes tipos de distâncias entre vetores. Os métodos estudados foram: *centroid*, *single linkage*, *complete linkage*, *average linkage* e *average linkage weighted*, *Ward minimum variance* e método da mediana. As distâncias utilizadas foram: norma L_p (em particular as normas L_1 e L_2), Mahalanobis e distância Euclidiana corrigida pela variância (*variance corrected*) – uma variação da distância de Mahalanobis. Finalmente, apresenta-se uma discussão sobre alguns indicadores comumente utilizados para seleção do número de clusters.
- PALAVRAS-CHAVE: Análise de agrupamentos, métodos de regionalização, algoritmos hierárquicos.

References

- ALPAYDIN, E. *Introduction to machine learning*. The MIT Press, 2004. 460p.
- ANSELIN, L. *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic, 1988.

- ANSELIN, L.; FLORAX, R. *Advances in spatial econometrics*. Heidelberg: Springer-Verlag, 2000. 513p.
- ASSUNÇÃO, R.; LAGE, J.; REIS, E. Análise de conglomerados espaciais via árvore geradora mínima. *Rev. Bras. Estat.*, Rio de Janeiro, v.63, n.220, p.7-24. 2002.
- BASU, S.; DAVIDSON, I.; WAGSTAFF, K. *Constrained clustering: advances in algorithms, theory, and applications*. London: Chapman & Hall/CRC Press, 2008. (Data Mining and Knowledge Discovery Series).
- BATAGELJ, V.; FERLIGOJ, A. Constrained clustering problems. In: *Adv. data Sci. Classif.*, Amsterdam: Springer verlag, 1998, p.137-144.
- BERRY, M. J. A.; LINOFF, G. *Data mining techniques*. Hoboken: John Wiley and Sons, 1997.
- CARVALHO, A. X. Y.; ALBUQUERQUE, C. W.; MOTA, J. A.; PIANCASTELLI, M. (Org.). *Dinâmica dos Municípios*. Brasília: IPEA, 2008. 326p.
- CARVALHO, A. X. Y.; ALMEIDA, G. R.; ALBUQUERQUE, P. H. M.; GUIMARÃES, R. D. *Clusterização hierárquica espacial: Texto de Discussão*. Brasília: IPEA, 2009.
- CHAPELLE, O.; SCHAKOPF, B.; ZIEN, A. *Semi-supervised learning*. Cambridge: MIT Press, 2006. 528p.
- CHEIN, F.; LEMOS, M. B.; ASSUNÇÃO, J. J. Desenvolvimento desigual: evidências para o Brasil. In: ANAIS DO ENCONTRO NACIONAL DE ECONÔMICA, 33., 2005, Niterói. *Anais ...* Niterói, 2005. p.1-20.
- CHOMITZ, K. M.; Da MATA, D.; CARVALHO, A.; MAGALHAES, J. C. R. *Spatial dynamics of labor markets in Brazil*. Brasília, 2005. (World Bank Policy Research Working Paper 3752).
- Da MATA, D.; DEICHMANN, HENDERSON, J. V.; LALL, S.; WANG, H. *Determinants of city growth in Brazil*, Cambridge: National Bireaux of Economic Research, 2005. 49p. (NBER Working Paper, series 11585).
- Da MATA, D.; PIN, C.; RESENDE, G. Composição e consolidação da infra-estrutura domiciliar nos municípios brasileiros. In: *Diferenças regionais no Brasil: caracterização e evolução nos últimos anos*. Brasília: IPEA. (no prelo).
- DUQUE, J. C.; RAMOS, R.; SURIÑACH, J. Supervised regionalization methods: *Survey Intern. Reg. Sci. Rev.*, v.30, n.3, p 195-220, 2007.
- ECK, J. E.; CHAINEY, S.; CAMERON, J. G.; LEITNER, M.; WILSON R. E. *Mapping crime: understanding hot spots*. Washington: U.S. Department of Justice, 2005.
- GLAZ, J.; BALAKRISHNAN, N. *Scan statistics and applications*. Birkhäuser, 1999.
- GLAZ, J.; NAUS, J.; WALLENSTEIN, S. *Scan statistics*. Springer, 2001.
- GORDON, A. D. A survey of constrained classification. *Comput. Stat. Data Anal.*, Amsterdam, v.21, p.17-29, 1999.
- GOWER, J. C. A comparison of some methods of cluster analysis, *Biometrics*, Washington, v.23, n.4, p.623-637, 1967.

- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning. data mining, inference and prediction*. Amsterdam: Springer, 2001.
- HIRSCHFIELD, A.; BOWERS, K. (Ed.). *Mapping and analysing crime data. lessons from research and practice*. Taylor and Francis, 2001.
- HIU, T. D.; FANG, J.; CHEN, Y.; WANG, JERIS C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 7., 2001, San Francisco. *Proceedings ...* San Francisco: ACM SIGKDD, 2001.
- IBGE - Instituto Brasileiro de Geografia e Estatística. *Censo demográfico 2000: Documentação dos microdados da amostra*. Brasília, 2002.
- IPEA, PNUD, FJP. *Atlas do desenvolvimento humano no Brasil*. Brasília, 2003.
- JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. New Jersey: Prentice Hall, 1988.
- JAIN, A. K.; MURTY, M. N.; FLYNN, M. N. Data clustering: a review. *ACM Comput. Surveys*, New York, v.31, n.3, p.264-323, 1999.
- KHATTREE, R.; NAIK, D. N. *Multivariate data reduction and discrimination with SAS software*. Boca Raton: Wiley Inter-Science, 2000.
- LAWSON, A. B.; DENISON, D. G. T. (Ed.). *Spatial cluster modelling*. London: Chapman and Hall/CRC., 2002. 310p.
- LUO, M.; MA, Y.; ZHANG, H.; A spatial constrained K-means approach to image segmentation. In: INFORMATION, COMMUNICATIONS AND SIGNAL; PACIFIC RIM CONFERENCE ON MULTIMEDIA, 4., 2003, Singapore. *Proceedings ...* Singapore, 2003, v.2, p.738-742.
- LUO, Z. *Clustering under Spatial contiguity constraint: a penalized K-means method*. Penn: Department of Statistics - Penn State University, 2001. (Technical Report).
- MAHALANOBIS, P. C. On the generalised distance in statistics. *Proc. Nat. Inst. Sci. India*, Calcutta, v.2, n.1, p.49-55, 1936.
- MARAVALLE, M.; SIMEONE, B. A spanning three heuristic for regional clustering. *Commun. Stat. Theor. Methods*. New York, v.24, p. 629-63, 1995.
- MARAVALLE, M.; SIMEONE, B.; NALDINI, R. Clustering on trees. *Comput. Stat. Data Anal.* Amsterdam, v.24, p.217-234, 1997.
- MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, Williamsburg, v.50, p.159-179, 1985.
- MILLIGAN, G. W.; COOPER, M. C. *A study of variable standardization*. Columbus, OH: The Ohio State University, 1987. p.87-63. (*College of Administrative Science Working Paper Series*)
- MURTAGH, F.; A survey of algorithms for contiguity-constrained clustering and related problems. *Comp. J.*, London, v.28, n.1, p.82-88, 1985.
- PACE, K.; BARRY, R. Sparse spatial autoregressions. *Stat. Probab. Lett.*, Amsterdam, v.33, p.291-297, 1997.

SARLE, W.S. *Cubic clustering criterion*, SAS Technical Report A-108, Cary, NC: SAS Institute, 1983.

WIPPERMAN, B. *Hierarchical agglomerative cluster analysis with a contiguity constraint* Simon Fraser University. 1999.

Received 21.04.2009.

Approved after revision 08.10.2009.

Appendix 1. Box-plots to evaluate the number of municipalities within each cluster.

Rook Neighborhood

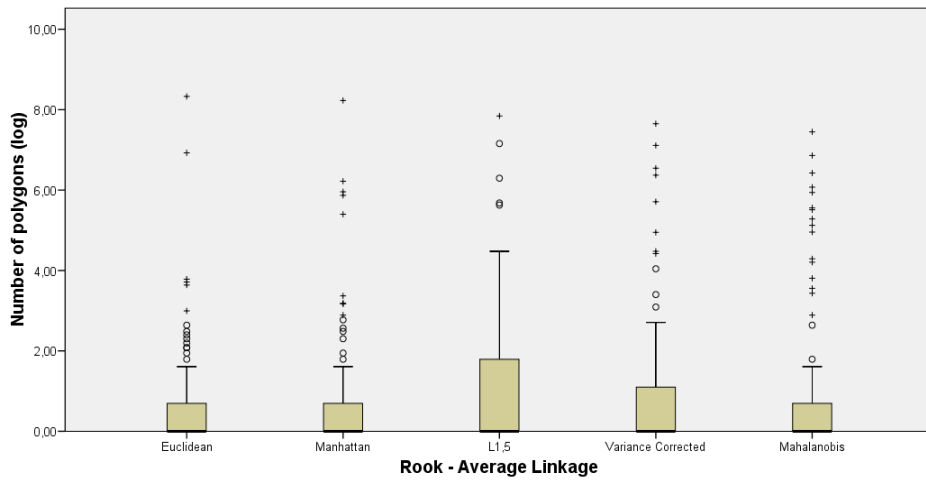


Figure A2-1 - Distribution of the number of municipalities - average linkage.

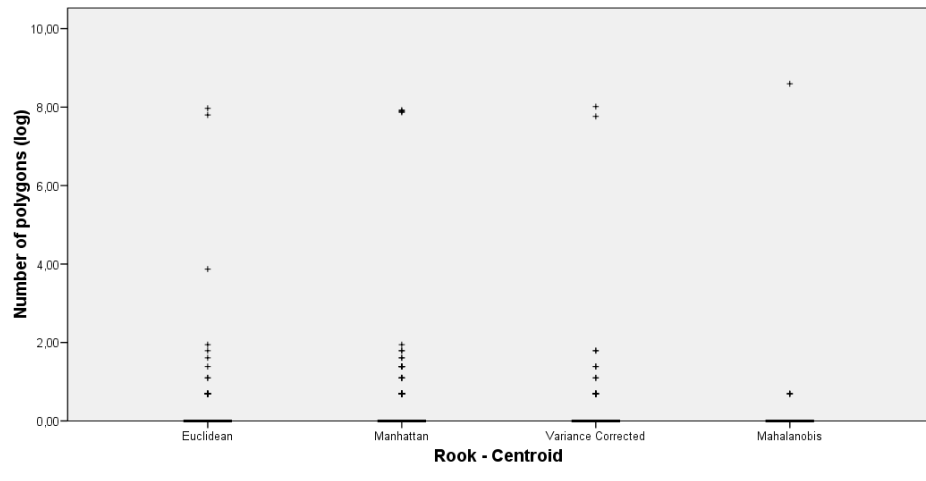


Figure A2-2 - Distribution of the number of municipalities – centroid.

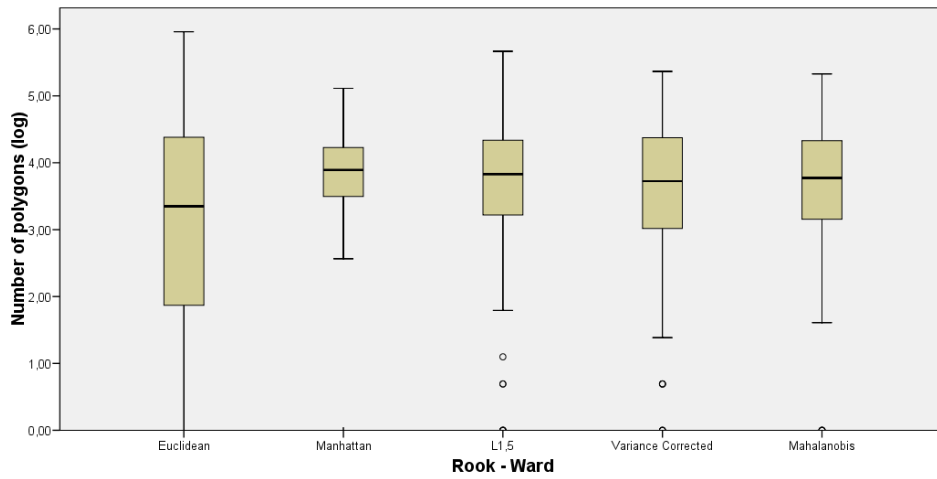


Figure A2-3 - Distribution of the number of municipalities – Ward.

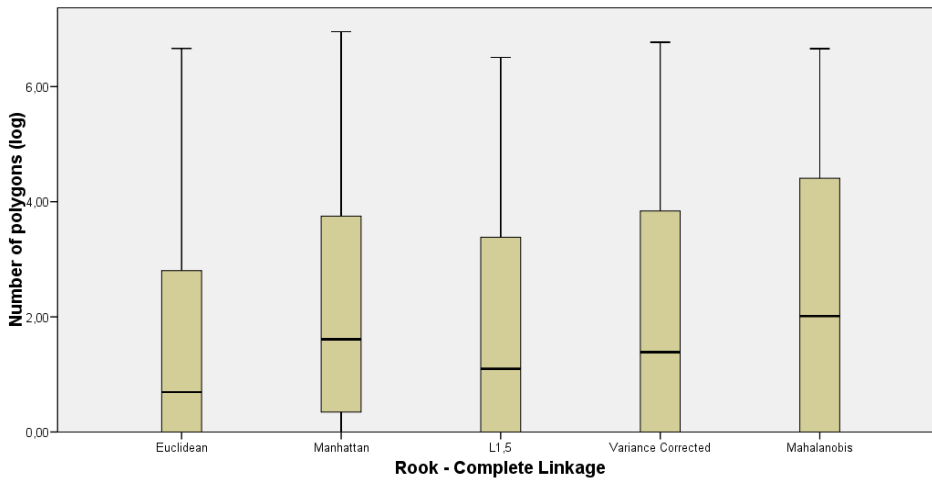


Figure A2-4 - Distribution of the number of municipalities – complete linkage.

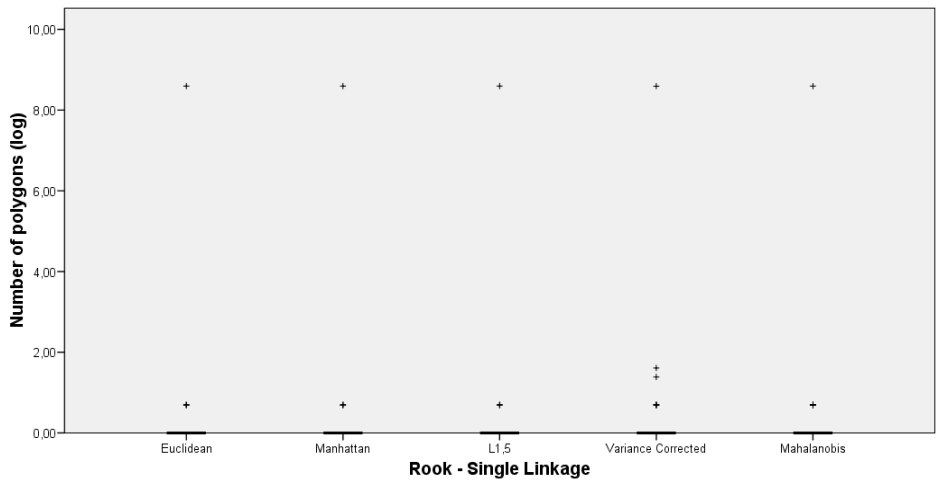


Figure A2-5 - Distribution of the number of municipalities – single linkage.

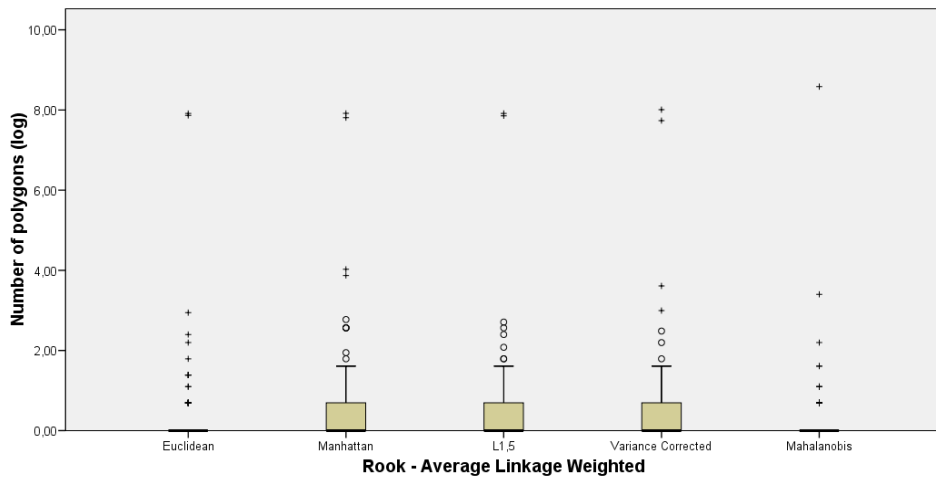


Figure A2-6 - Distribution of the number of municipalities – average linkage weighted

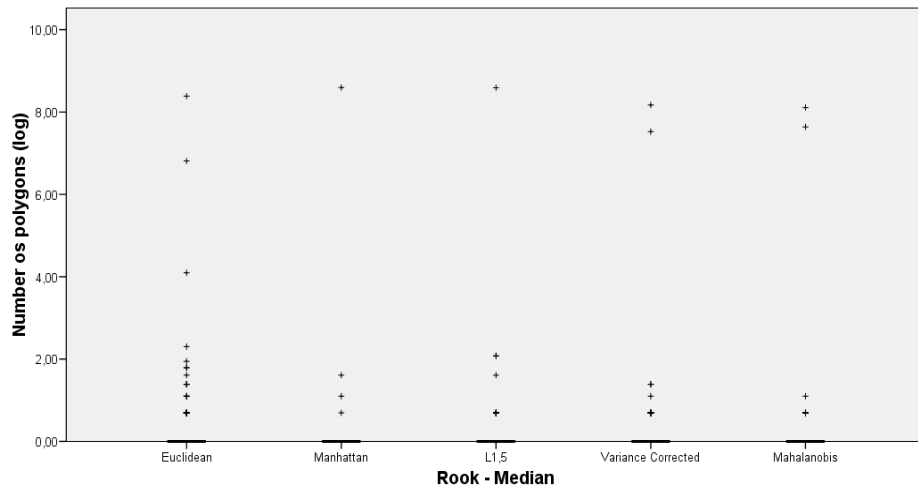


Figure A2-7 - Distribution of the number of municipalities – median.

Queen Neighborhood

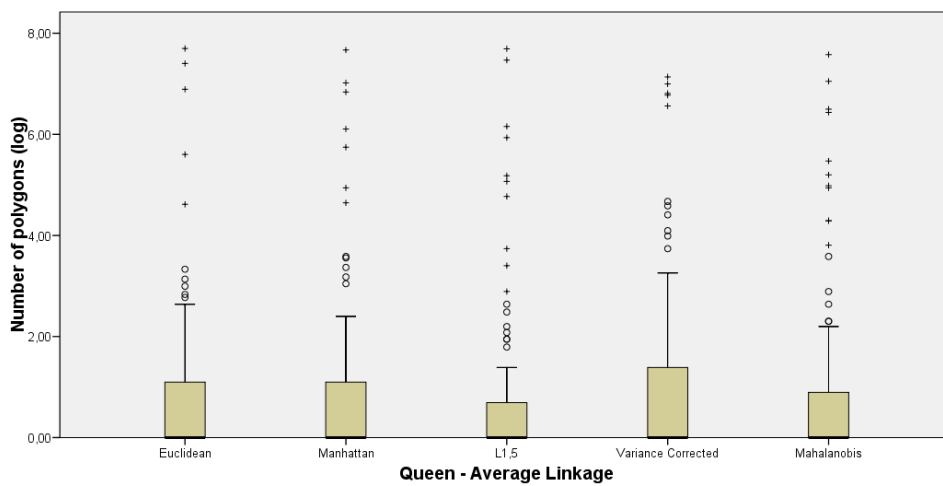


Figure A2-8 - Distribution of the number of municipalities – average linkage.

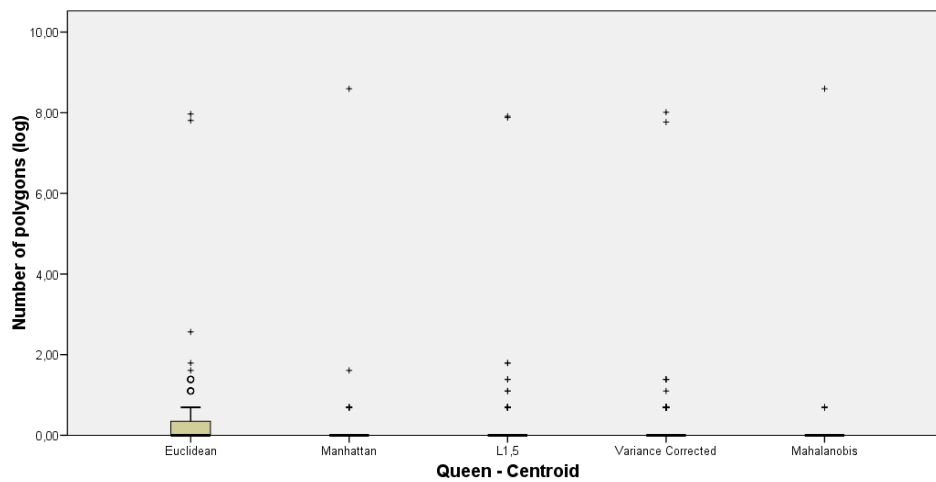


Figure A2-9 - Distribution of the number of municipalities – centroid

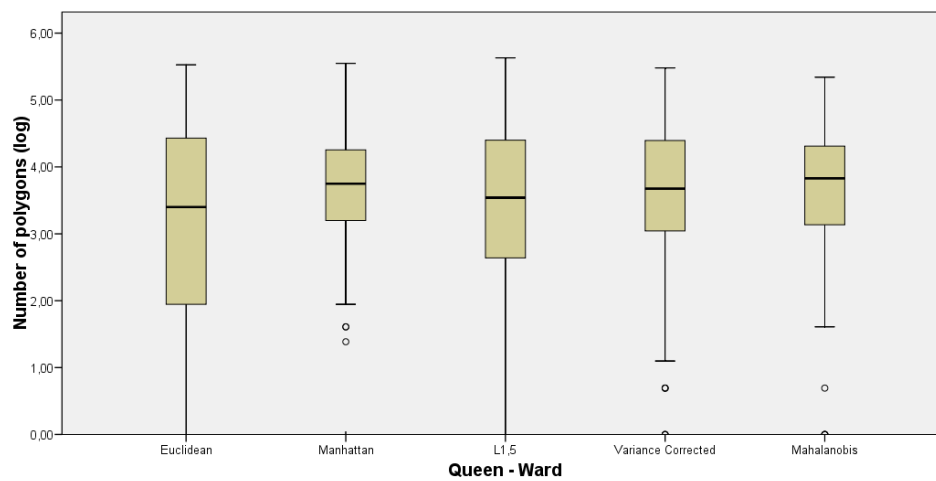


Figure A2-10 - Distribution of the number of municipalities – Ward.

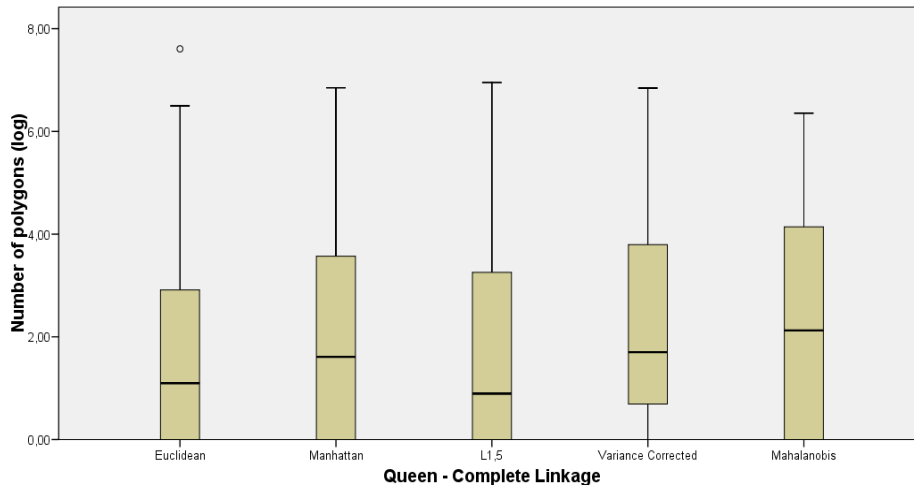


Figure A2-11 - Distribution of the number of municipalities – complete linkage.

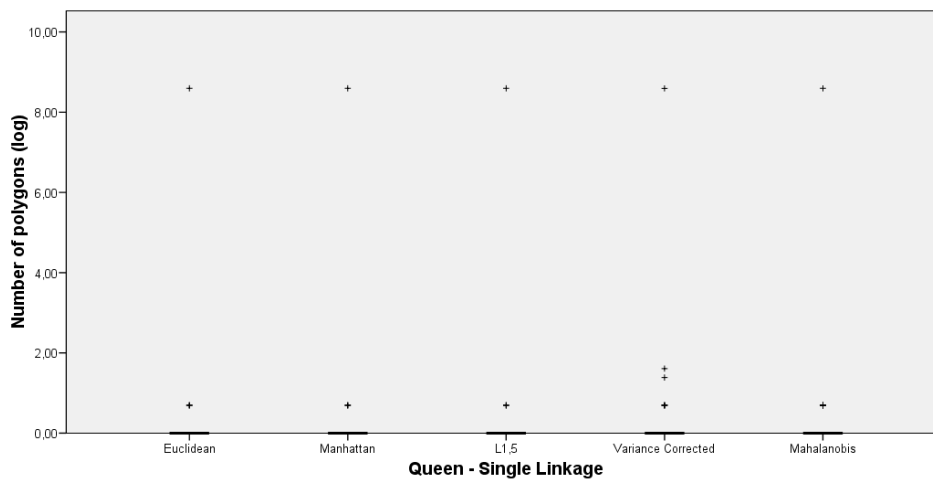


Figure A2-12 - Distribution of the number of municipalities – single linkage.

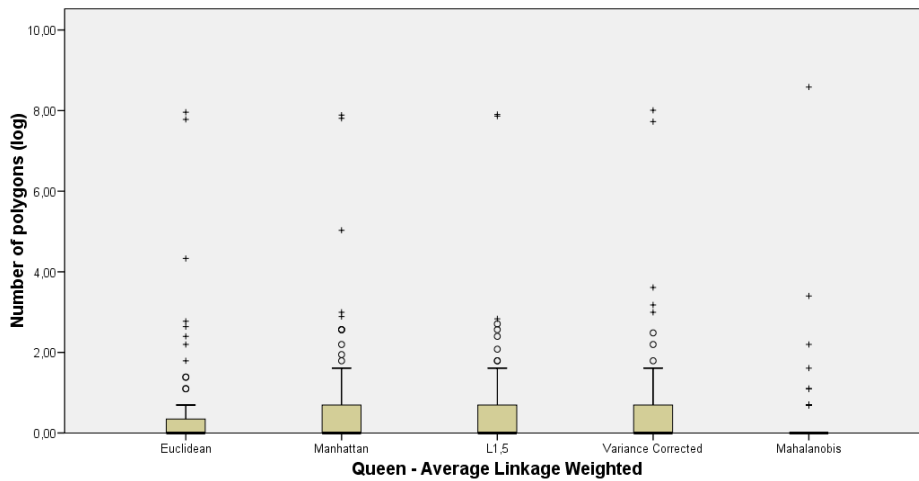


Figure A2-13 - Distribution of the number of municipalities – average linkage weighted.

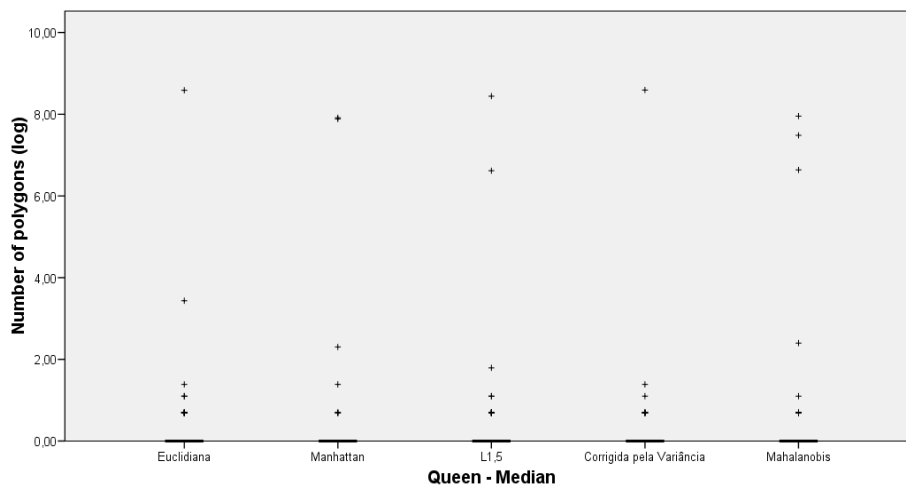


Figure A2-14 - Distribution of the number of municipalities – median.

Appendix 2. Plots of the criteria for selecting the number of clusters.

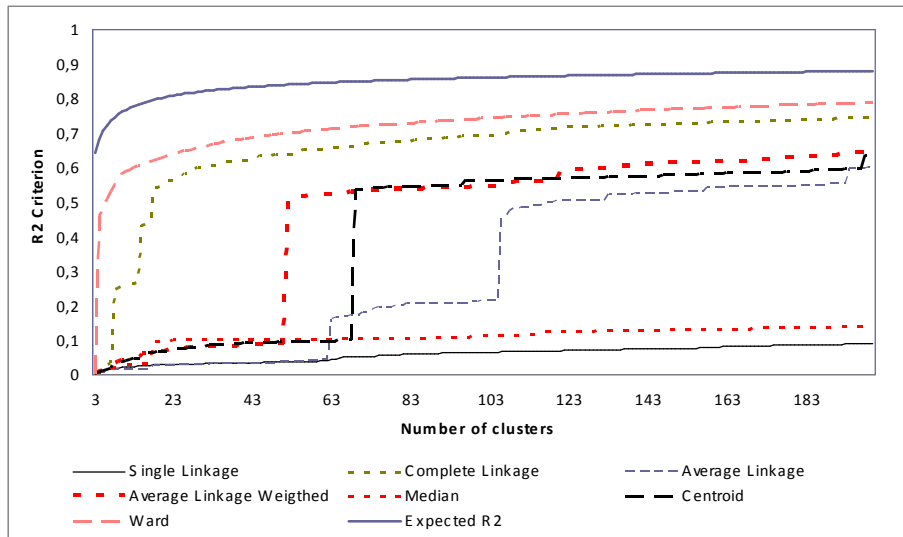


Figure A2-1 - R² criterion with Rook neighborhood.

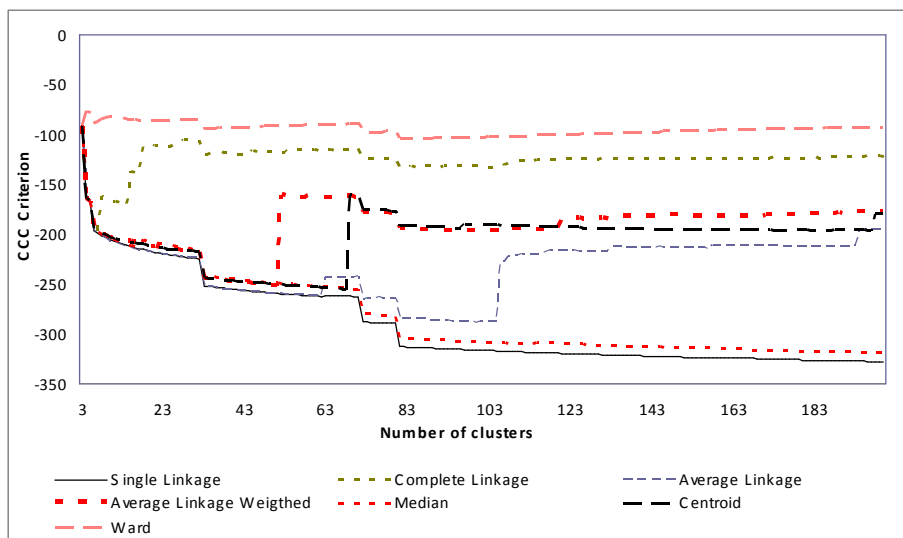


Figure A2-2 - CCC criterion with Rook neighborhood.

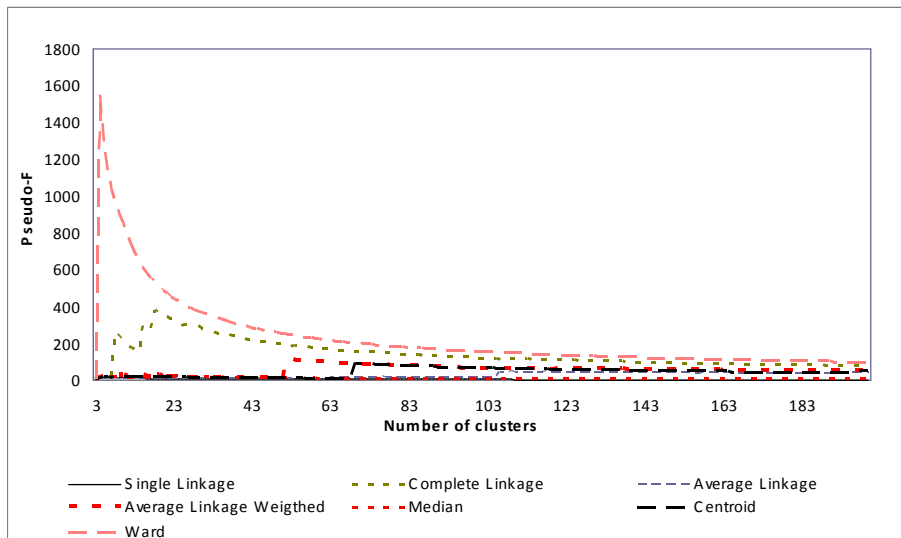


Figure A2-3 - Pseudo-F criterion with Rook neighborhood

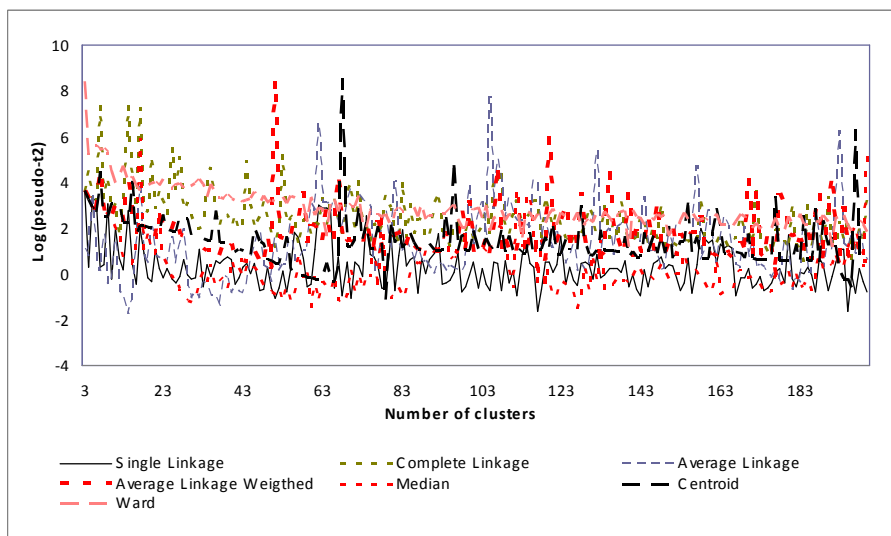


Figure A2-4 - Pseudo-t² criterion with Rook neighborhood.