ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 10, Issue, 04, pp. 35374-35380, April, 2020
https://doi.org/10.37118/ijdr.18706.04.2020

RESEARCH ARTICLE **OPEN ACCESS**

HIERARCHICAL CLUSTERING WITH SPATIAL CONSTRAINTS IN TUBERCULOSIS DATA

***[1]Dalila Camêlo Aguiar, [2]Ramón Gutiérrez Sánchez and [3]Edwirde Luiz Silva Camêlo**

[1]PhD student of the Doctoral Programme in Mathematical and Applied Statistics, University of Granada, Granada, Spain
[2]PhD in Statistics, Professor at the University of Granada, Granada, Spain
[3]PhD in Statistics, Professor at the State University of Paraíba, Campus Campina Grande, Paraíba, Brazil

ARTICLE INFO

Article History:
Received 08th January, 2020
Received in revised form
14th February, 2020
Accepted 20th March, 2020
Published online 30th April, 2020

Key Words:
Ward-like Hierarchical Clustering,
Spatial Constraints, Tuberculosis,
State of Paraíba.

Corresponding author: Dalila Camêlo Aguiar

ABSTRACT

Study on socio-epidemiological variables of TB, considering a clustering with spatial/ geographical restrictions for the State of Paraíba, Brazil. For the application of Ward's hierarchical clustering method, two dissimilarity matrices were calculated, the first provides the dissimilarities in the feature space calculated from the socio-epidemiological variables (D_0) and the second provides the dissimilarities in the calculated restriction space from the geographical distances (D_1) together with an alpha mixing parameter and the weight w attributed to calculation of the dissimilarity matrix as being collective inequality index. Statistical analyses were undertaken in R. In D_0 the clusters are dispersed and are not strictly contiguous, the five clusters are marked mainly by the high proportion of new cases. Geographically more compact clusters are obtained after the introduction of D_1 and $\alpha = 0.1$, slightly favoring socioeconomic homogeneity (24%) versus geographical homogeneity (64%) mainly influenced by clusters 1 and 3. With $\alpha = 0.2$ the socio-epidemiological and geographic homogeneity are favored although they are more compact, this partition is slightly worse than the previous one because it gives more importance to the neighborhoods. The method is shown to be feasible in epidemiological studies in the joint understanding of factors of different dimensions, aggregated from a spatial perspective.

Citation: ***Dalila Camêlo Aguiar, Ramón Gutiérrez Sánchez and Edwirde Luiz Silva Camêlo.*** **2020.** "Hierarchical clustering with spatial constraints in tuberculosis data*", International Journal of Development Research*, 10, (04), 35374-35380.

INTRODUCTION

One of the main concerns in Public Health surveillance is detectation and track of clusters of diseases i.e., the presence of high incidence rates around a particular location, which usually means a higher risk of suffering from the disease of study. Cluster analysis consists in distinguishing, in the set of analysed data, the groups, called clusters. These groups are disjoint subsets of the data set, having such a property that data belonging to different clusters differ among themselves much more than the data, belonging to the same cluster (Wierzchoń and Kłopotek, 2018). It is known how difficult it is for researchers to group a set of n objects into k separate sets. However, in some clustering problems, it is relevant to impose restrictions on the set of allowed solutions. This article makes use of a recent hierarchical method of clustering (and not partitioning), including spatial restrictions (not necessarily neighborhood restrictions) (Chavent *et al*., 2018a) in the imposition of contiguity restrictions on the set of permitted solutions for the local mapping of tuberculosis (TB) socio-epidemiological data in State of Paraíba. It is one of the to 27 federative units in Brazil and is divided into 223 municipalities. It has the twenty-first largest territorial area (0.66%) in the country and is the fourteenth contingent population among the states of Brazil with more than 4.018 million inhabitants (1.91%) according to 2019 estimates by the Brazilian Institute of Geography and Statistics (IBGE, 2019). Aguiar *et al* (2019) in their study observed that in the period from 2007 to 2016, 13,413 cases of TB were reported in the State of Paraíba, with an annual average of 1,336.6 and cure rates lower than those recommended by the World Health Organization (WHO) and Ministry of Health/Brazil. The remarkable relation that TB has with social conditions demands an understanding of the dynamics of this aggravation and its occurrence in the territory through geospatial analyses (Santos Neto *et al*., 2017). Therefore, the objective is to present a solution based on socio-epidemiological variables considering a clustering with spatial/geographical restrictions for the State of Paraíba.

MATERIAL AND METHODS

Study design and data sources: The data analyzed in this study are notified cases of TB in the 223 municipalities in the State of Paraíba in the period between 2001 and 2018, using a secondary source, through the database, registered in the Notifiable Diseases Information System (SINAN, 2020) and made availables on the website of the Informatics Department of the Unified Health System (DATASUS). The variables are ratios and are divided into epidemiological (new cases and cure) and social variables such as years of study (less than 10 years' formal education) and working age (20-49). A matrix was also calculated with the geographic distances between the municipalities and the weight w attributed to the calculation of the dissimilarity matrix D as being the collective inequality index of the GDP in the State of Paraíba. As units of analysis, municipalities and microregions were used. Data collection took place during February 2020. As units of analysis, municipalities and microregions were used. Data collection took place during February 2020. For data analysis, the program was used R version 3.6.2 (R Core Team, 2019). As this is a secondary data survey and does not directly involve human beings, this study was not submitted to the Research Ethics Committee's evaluation.

Constrained hierarchical clustering: The hierarchical cluster or hierarchical cluster analysis (HCA) as it is also known, is a popular method for cluster analysis in big data research and data mining in order to establish a hierarchy of clusters. HCA tries to group to individuals with similar characteristics into clusters (Murtagh, 2014; Petushkova *et al*., 2014; Zhang, 2017). Usually the researcher is faced with the difficulty of clustering a set of n objects into k separate sets. Soon, many methods were proposed to find the best partition according to a homogeneity criterion based on differences, or for a multivariate distribution function mix model. Soon, many methods were proposed to find the best partition according to a homogeneity criterion based on differences, or for a multivariate distribution function mix model. However, in some clustering problems, it is relevant to impose constraints on the set of allowable solutions. The most common type are the contiguity constraints (in space or in time). Such restrictions occur when the objects in a cluster are required not only to be similar to one other, but also to comprise a contiguous set of objects (municipality), i.e. the contiguity between each pair of objects is given by a matrix $C = (c_{ij})_{n\times n}$, where $c_{ij} = 1$ if the *i*th and the *j*th objects are regarded as contiguous, and 0 if they are not. A cluster C is then considered to be contiguous if there is a path between every pair of objects (municipality) in C (the subgraph is connected), that is, a cluster C is then considered connected if there is a path between each pair of municipality in C. Several classical clustering algorithms have been modified to take this type of constraint into account (see e.g., Murtagh 1985a; Legendre and Legendre 2012; Bécue-Bertaut *et al*. 2014b). So, two clusters are regarded as contiguous if there are two objects, one from each cluster, which are linked in the contiguity matrix. Although this can lead to reversals (i.e. inversions, upward branching in the tree) in the hierarchical classification it has been proven that only the complete link algorithm is guaranteed to produce no reversals when relational constraints are introduced in the ordinary hierarchical clustering procedure (Ferligoj and Batagelj 1982). Several authors in different areas of knowledge have implemented of constrained clustering procedures (Duque *et al*. 2011, Bécue-Bertaut *et al*. 2017a, Dehman *et al*. 2015, Legendre 2014, and Ambroise *et al*. (1997b, 1998a)). The previous procedures which impose strict contiguity may separate objects (municipality) which are very similar into different clusters, if they are spatially apart. Other non-strict constrained procedures have then been developed, including those referred to as soft contiguity or spatial constraints Other non-strict constrained procedures have then been developed, including those referred to as soft contiguity or spatial constraints. Oliver and Webster (1989) and Bourgault *et al*. (1992) suggest running clustering algorithms on a modified dissimilarity matrix. This dissimilarity matrix is a combination of the matrix of geographical distances and the dissimilarity matrix computed from non-geographical variables (that here will be socio-epidemiological variables). According to the weights given to the geographical dissimilarities in this combination, the solution will have more or less spatially contiguous clusters.

Ward-like hierarchical clustering: The Ward-like hierarchical clustering method (not partitioning) including spatial/geographic constraints (not necessarily neighborhood constraints) was proposed by Chavent *et al* (2018a). With an algorithm similar to Ward, Ward-like is a constrained hierarchical clustering algorithm which optimizes a convex combination of this criterion calculated with two dissimilarity matrices,D_0 and D_1 beyond a mixing parameter $\alpha \in [0; 1]$. The first dissimilarity matrix D_0 is constructed from the distances between socio-epidemiological variables, this is, the matrix presents the differences in the 'feature space' and the dissimilarity matrix D_1 is built with the geographic matrix, i.e., the matrix D_1 provides the differences in "constraint space". The procedure for the choice the mixing parameter is assigned by α, which defines the importance of the constraint in the grouping procedure. The minimized criterion at each stage is a convex combination of the homogeneity criterion calculated with D_0 and the homogeneity criterion calculated with D_1. The parameterα (the weight of this convex combination) controls the weight of the constraint on the quality of the solutions. When α increases, the homogeneity calculated withD_0 decreases, conversely, the homogeneity calculated increases with D_1. Therefore, idea is to determine a value of α which increases the spatial-contiguity without deteriorating too much the quality of the solution on the variables of interest.

Considering a set of n observations. Let w_i be the weight of the *i*th observation for $i = 1, \dots, n$. Let $D = [d_{ij}]$ be a $n \times n$ dissimilarity matrix associated with the n observations, where d_{ij} is the dissimilarity measure between observations i and j. The Ward-like method considers a partition $P_K = (C_1, \dots, C_K)$ in K clusters. The pseudo inertia of a cluster C_K generalizes the inertia to the case of dissimilarity data (Euclidean or not) in the following way:

$$I(C_K) = \sum_{i\in C_k} \sum_{j\in C_k} \frac{w_i w_j}{2\mu} d_{ij}^2 \qquad (1)$$

where $\mu_k = \Sigma_{i\in C_k} w_i$ is the weight of C_k. The smaller the pseudo-inertia $I(C_K)$ is, the more homogenous are the observations belonging to the cluster C_k. The pseudo within-cluster inertia of the partition P_K is therefore, $W(P_K) = \sum_{k=1}^{K} I(C_k)$. The smaller this pseudo within-inertia $W(P_K)$ is, the more homogenous is the partition in K clusters. The quality criterion Q_0 and Q_1 of the partitions P_K^{α} obtained with different

values of $\alpha \in [0,1]$ and choose the value of alpha which is a trade-off between the lost of socio-epidemiological homogeneity and the gain of geographic cohesion. With *ClustGeo* (R Package) developed by Chavent *et al*, (2017b) it is possible to implement this hierarchical clustering algorithm and the procedure for choosing alpha α. The function *hclustgeo* of the *ClustGeo* package performs the Ward-like hierarchical clustering using the dissimilarity matrix D (which is an object of the *dist* class, that is, an object obtained with the *dist* function or a dissimilarity matrix transformed into an object of the *dist* class with the *as.dist* function) of observations as arguments. We opted for the uniform weight defined by the collective inequality index. The function *hclustgeo* is a wrapper of the usual *hclust* function with the following arguments: methods (Ward.D), $d = \Delta$(Mahattan distance) and members $w = MD = \sum_{i=1}^{h} d_i f_i$ (collective inequality index). The sum of the heights in the dendrogram is equal to the total pseudo-inertia of the data set Eq. (1).

Manhattan distance: We opted for the Manhattan distance because the Ward method has already been generalized for use with non-Euclidean distances, according Strauss and Maltitz (2017) concluded in their study that Ward's clustering algorithm can be used in conjunction with Manhattan distances, without the characteristic of minimising within-cluster variation and maximising between-cluster variation being violated, and that for this specific case it produced better results than using Euclidean distances. Manhattan distance it is also known as City block distance, and absolute value distance or L1 distance. Manhattan distances a distance that follows a route along the non-hypotenuse sides of a triangle. This metric is less affected by outliers than the Euclidean and squared Euclidean metrics:

$$d(i,j) = \sum_{k=1}^{n} |X_{ik} - X_{jk}| \qquad (2)$$

Collective inequality index: Its about a regional statistical indicator, the collective inequality index is a decomposable measure and will be defined as the weight w attributed to the calculation of the dissimilarity matrix D. For the calculation of the collective inequality index, will be used GDP of the 23 micro regions of the State of Paraíba (H) is used, which takes values $h_1, h_2, \ldots, h_{23}$ with absolute frequencies $n_i, n_2, \ldots, n_{23}$ over a finite population of size $N = 23$. According to the characteristic proposed by Zaiger (1983), a measure of decomposable inequality is given by:

$$I_{\beta(H)} = \sum_{i=1}^{23} \Gamma_\beta \left(\frac{h_i}{\bar{h}}\right) f_i$$

Being $f_i = n_i/N$ the relative frequency and $\Gamma_\beta(h)$ a defined function, for the value of $\beta < 0$, whose function will be: $\Gamma_\beta(h) = h^\beta - 1$. González and Céspedes (2004) establishes the collective inequality index (CII) as being:

$$CII = I_{-1}(H) = \sum_{i=1}^{23} \Gamma_{-1}\left(\frac{h_i}{\bar{h}}\right) f_i = \sum_{i=1}^{23} \left[\left(\frac{h_i}{\bar{h}}\right)^{-1} - 1\right] f_i = \sum_{i=1}^{23} \left(\frac{\bar{h}}{h_i} - 1\right) f_i = \sum_{i=1}^{23} d_i f_i$$

RESULTS AND DISCUSSION

In the period of 2001-2018, 24.258 cases of TB were reported in the State of Paraíba, among which 80% were new cases, 65% were cured of the disease, 46.8 had less than ten years of schooling, 63.2% were between the ages of 20 and 49-years-old and 67% were male. The goal set by the WHO is to cure 85% of new bacilliferous TB cases by 2020 (WHO, 2017), however, as observed in the 2018 data, Brazil (71.4%) it falls short of reaching this goal and the situation is even more critical for the State of Paraíba (55.5%) (Brasil, 2019). In a study conducted in 2016, the authors concluded that in Brazil the lower the patients' level of education (less than 9 years' formal education), the higher the numbers of new cases of TB and the higher the rates of healing and treatment abandonment, throughout the country (Camêlo *et al*, 2016). We know socio-economic determinants have a substantial impact on infectious disease control, For this reason, we have included the collective inequality index (CII), although it has influenced the increase in heterogeneity among the municipalities due to the economic inequalities between them.

Clustering approaches are a useful tool to detect patterns in data sets and generate hypothesis regarding potential relationships. The role of cluster analysis is, therefore, to uncover a certain kind of natural structure in the data set (Wierzchoń and Kłopotek, 2018). Figure 1 shows the dendrogram of the dissimilarity matrix D_0, that is, the differences in the feature space of socio-epidemiological variables.The visual inspection of the dendrogram in Figure 1 suggests to retain $K = 5$ clusters. The 223 municipalities were grouped into their respective clusters according to their socio-epidemiological similarity, namely, cluster 1 (68), cluster 2 (58), cluster 3 (52), cluster 4 (41) and cluster 5 only 4 municipalities. The partition corresponding to the five clusters can be seen on the map in Figure 2. Geographically, we perceive clusters well dispersed according to socio-epidemiological variables, that is, the clusters are not strictly contiguous. The interpretation of clusters according to the initial socio-epidemiological variables is interesting. Figure 7 shows the variable boxplots for each cluster (top row). Cluster 1 has the lowest proportion of years of study in TB patients in the study area, contrary has higher incidences of new cases. Cluster 2 shows a high proportion of new cases and a low proportion of TB patients of working age. Cluster 3 has high rate of new cases, a low rate of schooling (below the average value of the study area) and the lowest rate of patients with active-age TB in all clusters. Cluster 4 has lower rates of cure and a high proportion of new cases (although its median proportion is lower when compared to other clusters). Cluster 5, has high rate of people of working age, with low schooling and median rate of cure rate slightly higher than that of new cases.

To obtain geographically more compact clusters, we will introduce the matrix D_1 of geographical distances into *hclustgeo*. For this, it is necessary that a mixing parameter be selected α to improve the geographical cohesion of the 5 groups without adversely affecting the socio-epidemiological cohesion. In Figure 3, we have the mixing parameter $\alpha \in [0,1]$ defines the importance of D_0 and D_1 in the clustering process with separate calculations for socio-economic homogeneity and the geographic cohesion of the partitions obtained for a range of different values of α and the 5 clusters. Obtaining the partition taking into account the geographic restrictions in Figure 3, shows the value α which aims to increase the spatial contiguity. When $\alpha = 0$ the geographical dissimilarities are not taken into account and when $\alpha = 1$ it is the socio-epidemiologic distances which are not taken into account and the clusters are obtained with the geographical distances only.

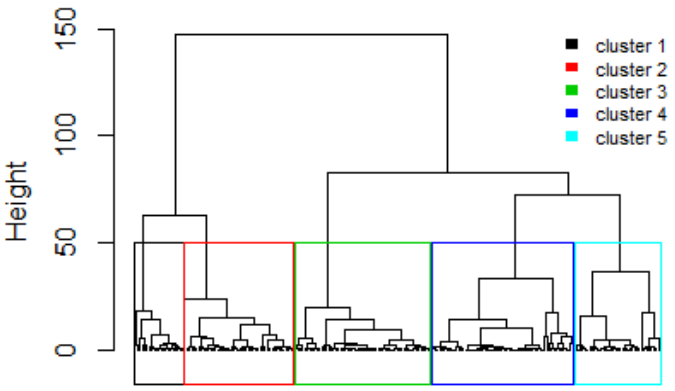

Figure 1. Dendrogram of the n = 223 municipalities based on the *4* socio-epidemiologic variables (that is using D_0 only).

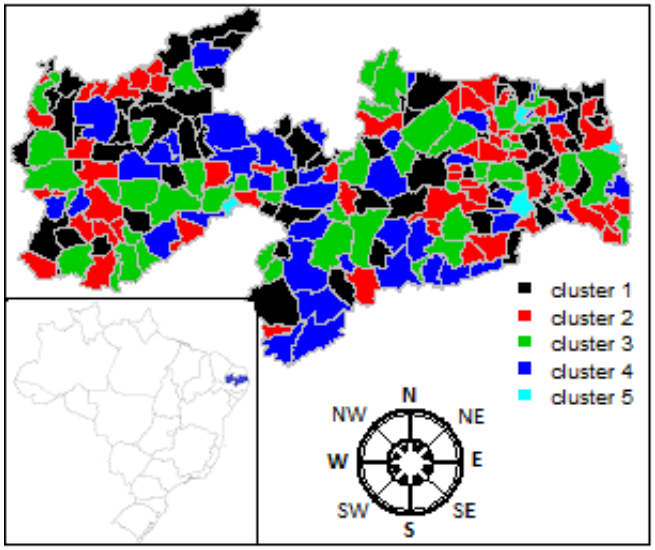

Figure 2. Map of the partition with K= 5 clusters only based on the socio-epidemiological variables (that is using D_0 only)

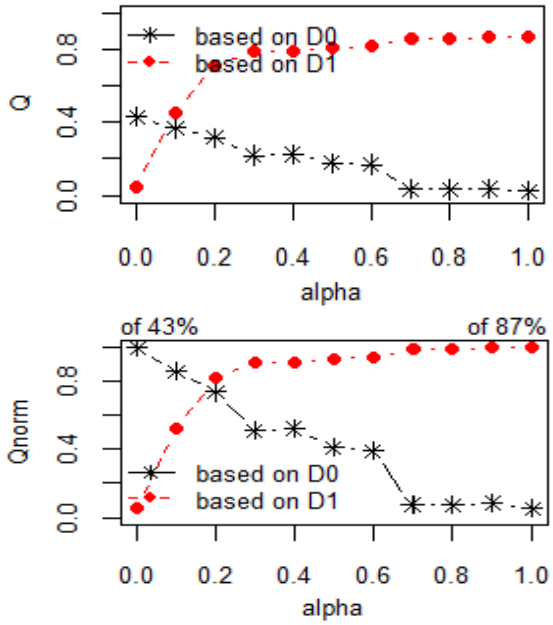

Figure 3. Choice of α for a partition in $K = 5$ clusters when D_1 is the geographical distances between municipalities. Left: proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in dashed line). Right: normalized proportion of explained pseudo-inertias $Q_0^*(P_K^\alpha)$ versus α (in black solid line) and $Q_1^*(P_K^\alpha)$ versus α (in dashed line)

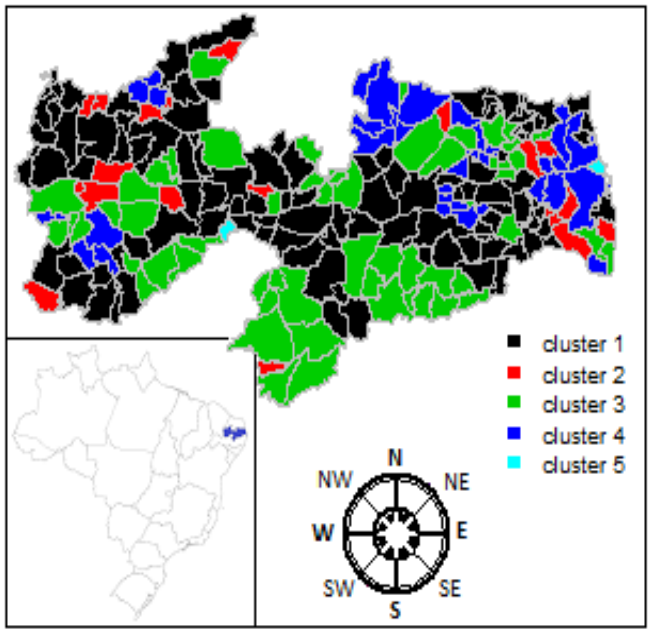

Figure 4. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the geographical distances between the municipalities D_1 with $\alpha = 0.1$.

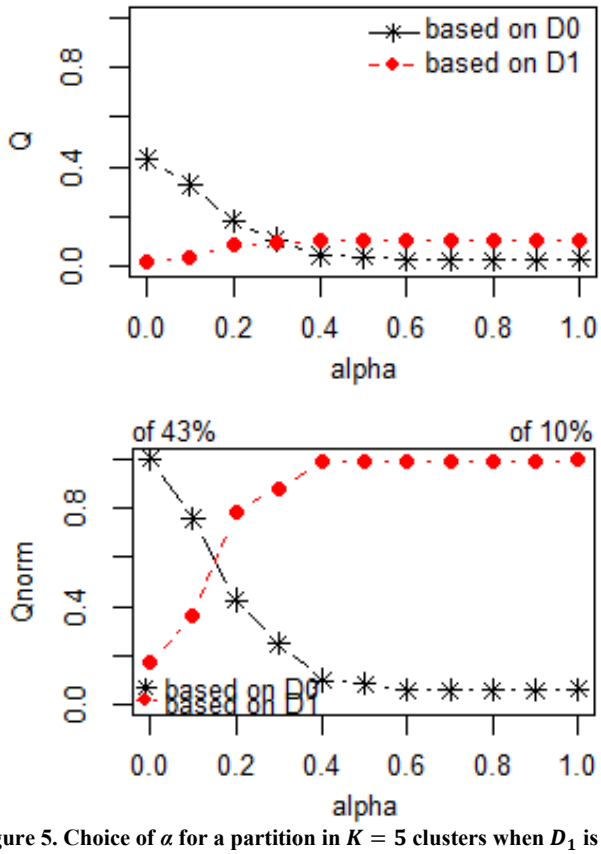

Figure 5. Choice of α for a partition in $K = 5$ clusters when D_1 is the neighborhood dissimilarity matrix between municipalities. Left: proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in dashed line). Right: normalized proportion of explained pseudo-inertias $Q_0^*(P_K^\alpha)$ versus α (in black solid line) and $Q_1^*(P_K^\alpha)$) versus α (in dashed line).

Figure 3 gives the plot of the proportion of explained pseudo-inertia calculated with D_0 (the socio-epidemiological distances) which is equal to 0.43 when $\alpha = 0$ and decreases when α increases (black solid line). On the contrary, the proportion of explained pseudo-inertia calculated

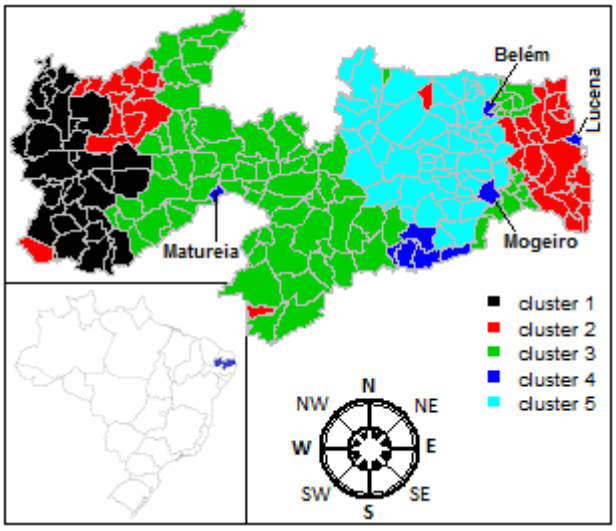

Figure 6. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the "neighborhood" distances of the municipalities D_1 with $\alpha = 0.2$.

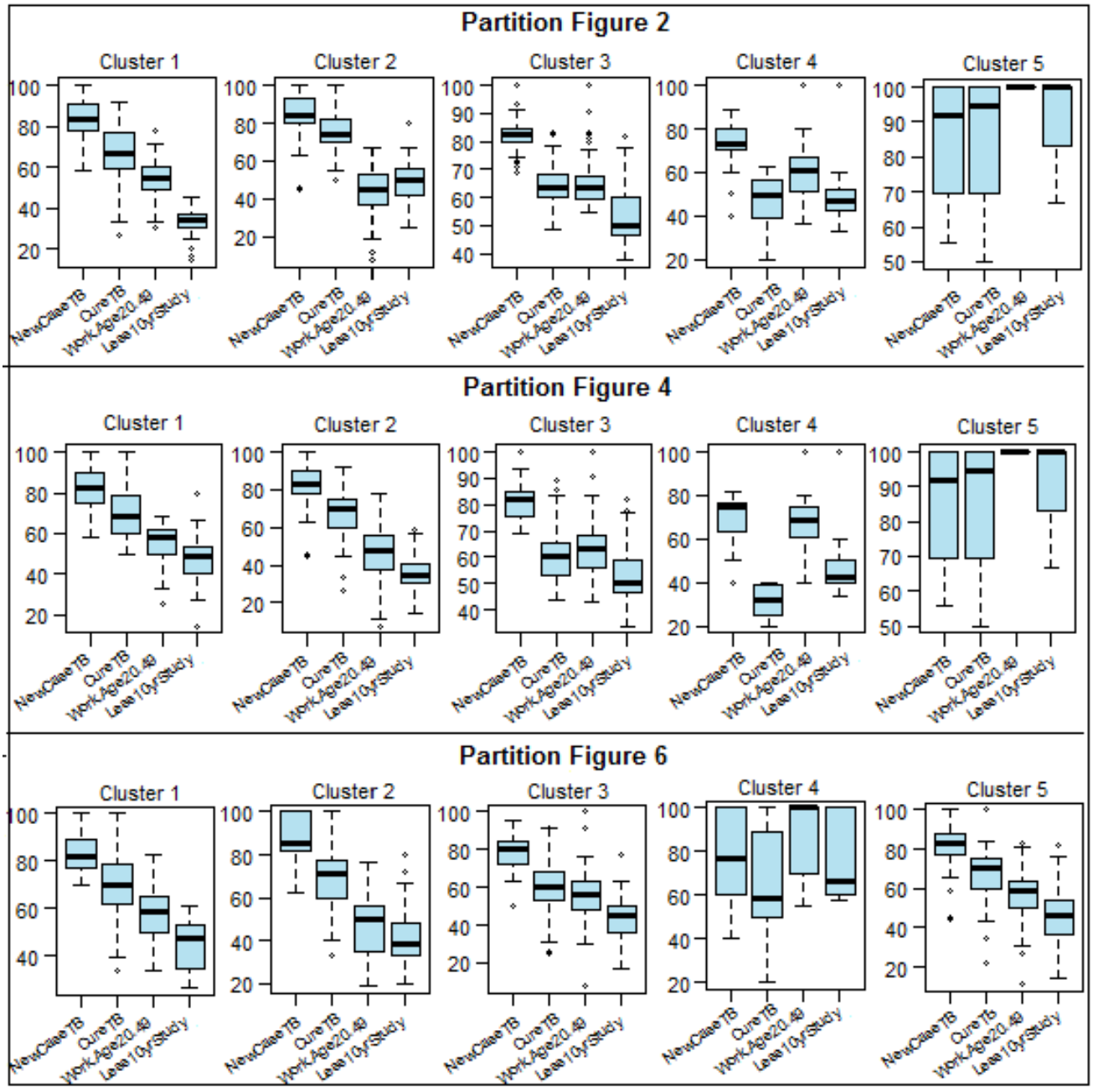

Figure 7. Comparison of the final partitions Figure 2, Figure 4 and Figure 5 in terms of variables.

with D_1 (the geographical distances) is equal to 0.87 when $\alpha = 1$ and decreases when α decreases (dashed line). Here, the plot of the normalized proportion of explained inertias suggests to retain $\alpha = 0.1$ or 0.2. The value alpha = 0.1 slightly favors the socio-economic homogeneity versus the geographical homogeneity. According to the priority given in this application to the socio-epidemiological aspects, the final partition obtained with $\alpha = 0.1$, which corresponds to a loss of only (1-0.76) 24% of socio-epidemiological homogeneity, and a (1-0.36) 64% increase in geographical homogeneity. The increased geographical cohesion of this partition with D_0 and D_1 and $\alpha = 0.1$ can be seen in Figure 4. Figure 4 a gain in spatial homogeneity is perceived, mainly in cluster 1 and 3. Clusters 2 and 4 were significantly altered. Figure 7 shows the boxplots of the variables for each cluster of the partition (middle line). The change in cluster 4 (Partition Figure 4) relative to cluster 4 (Partition Figure 2) it was primarily due to the cure variable, with the lowest rate in the study area. Cluster 2 (Partition Figure 4) has a higher median proportion of TB patient with working age and lower schooling rate, the opposite occurs in cluster 2 (Partition Figure 2), higher schooling rate and lower median proportion of working age. Cluster 5 (Partition Figure 4) is identical to cluster 5 (Partition Figure 2). The next plot, Figure 5, shows the choice of alpha for partition.

At the right of Figure 5, the plot of the normalized proportion of explained inertias (that is $Q_0(P_K^\alpha)$ and $Q_1(P_K^\alpha)$) suggests to retain $\alpha = 0.2$ slightly favoring socio-epidemiological homogeneity versus geographical homogeneity. It remains only to determine this final partition for $K = 5$ clusters and $\alpha = 0.2$. The corresponding map is given in Figure 6. Figure 6 shows that the clusters are spatially more compact than those in Figure 5. However, it is known that this approach creates divergences in the adjacency matrix, which gives more importance to the neighborhoods. However, as the approach is based on soft contiguity restrictions, municipalities that are not neighbors may be in the same clustering according occurs with the municipalities of Lucena, Belém, Matureia and Mogeiro in cluster 4. The quality of the partition in Figure 6 is slightly worse than that of the partition in Figure 4, according to the Q_0 criterion (32.61% versus 36.98%).

Concluding Remarks

The application of the Ward-like hierarchical clustering method proves to be feasible in epidemiological studies since it allows two matrices to be considered concurrently, the first with differences in the feature space (socio-epidemiological variables) and the second with differences in the constraint space (geographical distance) with an alpha mixing parameter in order to improve the geographical cohesion of the clusters without adversely affecting the socio-epidemiological cohesion. Thus, when considering spatial constraints, the hierarchical clustering becomes even more complete, once it will detect patterns in data sets of different dimensions. Therefore, its application becomes indispensable for a better understanding of the socio-epidemiological and economic reality of the municipality, as it is an analysis tool that allows to make more accurate decisions in the elaboration of public policies and more effective health actions in coping with TB, given that such a disease is directly related to the socioeconomic gradient in the level of poverty and social context. The difficulties of the State of Paraíba and Brazil itself with TB, especially with the cure of new bacilliferous cases, are worrisome and the scenario could be even worse since the financing of TB in Brazil has been decreasing significantly since 2018; in 2019, the national TB budget was only 38 (US$ millions), in addition to changes in the regulation of federally funded investment in strategic areas of health and strict limits imposed on the growth of public spending until 2036.

Acknowledgments: None.

REFERENCES

_____. Ministério da Saúde. *Sistema de Informação de Agravos de Notificação Tuberculose* – casos confirmados no Sistema de Informação de Agravos de Notificação – SINAN. Brasília, DF; 2017. Available in: <http://www2.datasus.gov.br/>. Accessed February 8, 2020.

Aguiar DC, Silva Camelo EL and Carneiro RO. 2019. Análise estatística de indicadores da tuberculose no Estado da Paraíba. doi: 10.13037/ras.vol17n61.5577. ISSN 2359-4330 Rev. Aten. Saúde, São Caetano do Sul, v. 17, n. 61, p. 05-12, jul./set.

Ambroise C, Govaert G. 1998a. Convergence of an EM-type algorithm for spatial clustering. Pattern Recognition Letters 19(10): 919-927.

Ambroise C., Dang M., Govaert G. 1997b. Clustering of Spatial Data by the EM Algorithm. In: A. Soares et al. (eds), geoENV I-Geostatistics for Environmental Applicattions, Kluwer, Dordrecht, pp. 493-504.

Applying of hierarchical clustering to analysis of protein patterns in the human cancer-associated liver. *PLoS One* 2014;9:e103950. 3.

Bécue-Bertaut M, Alvarez-Esteban R, S_anchez-Espigares JA. 2017a. *Xplortext*: Statistical Analysis of Textual Data R package. <https://cran.r-project.org/package=Xplortext>. R package version 1.0.

Bécue-Bertaut M, Kostov B, Morin A , Naro G 2014b. Rhetorical strategy in forensic speeches: multidimensional statistics-based methodology. *Journal of Classication* 31(1): 85-106.

Bourgault G, Marcotte D, Legendre P. 1992. The Multivariate (co) Variogram as a Spatial Weighting Function in Classification Methods. Mathematical Geology 24(5): 463-478.

BRASIL. Ministério da Saúde. Boletim Epidemiológico. Secretaria de Vigilância em Saúde. Brasil Livre da Tuberculose: evolução dos cenários epidemiológicos e operacionais da doença. Ministério da Saúde 3 Volume 50, Nº 09, Mar. 2019.

Camêlo E, Aguiar D, Silva R, Figueiredo TMRM, Gonzaléz Carmona A and Sánchez RG. 2016. Tuberculosis in Brazil: New Cases, Healing and Abandonment in Relation to level of Education. *International Archives Of Medicine,* 9**.** doi:10.3823/1939.

Chavent M, Kuentz-Simonet V, Labenne A and Saracco J. 2017b. ClustGeo: Hierarchical Clustering with Spatial Constraints. R package version 2.0. <https://CRAN.R-project.org/package=ClustGeo>.

Chavent M, Kuentz-Simonet V, Labenne A and Saracco J. 2018a. Clust Geo: an R package for hierarchical clustering with spatial constraints. *Comput Stat* 33, 1799–1822. <https://doi.org/10.1007/s00180-018-0791-1>.

Core Team R. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

Dehman A, Ambroise C, Neuvial P. 2015. Performance of a blockwise approach in variable selection using linkage disequilibrium information. BMC Bioinformatics 16:148.

Duque JC, Dev B, Betancourt A, Franco JL. 2011. ClusterPy: Library of spatially constrained clustering algorithms, RiSE-group (Research in Spatial Economics). EAFIT University. <http://www.rise-group.org/risem/clusterpy/>. Version 0.9.9.

Ferligoj A, Batagelj V. 1982. Clustering with relational constraint. *Psychometrika* 47(4):413-426.

González FP and Céspedes JC. Técnicas cuatitativas para el análisis regional. España: Editorial Universidad de Granada, 2004.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Paraíba* - Panorama. Cidades. 2019. Available in: <https://cidades.ibge.gov.br>. Accessed February 8, 2020.

Legendre P, Legendre L 2012. Numerical Ecology, vol. 24. Elsevier.

Legendre P. 2014 *const.clust*: Space-and Time-Constrained Clustering Package. <http://adn.biol.umontreal.ca/ numericalecology/Rcode/>.

Murtagh F. 1985. Multidimensional clustering algorithms. Compstat Lectures, Vienna: Physika Verlag.

Murtagh F. Hierarchical Clustering. In: Lovric M. editor. International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer; 2014:633-5. [Google Scholar].

Oliver M, Webster R. 1989. A Geostatistical Basis for Spatial Weighting in Multivariate Classication. Mathematical Geology 21(1):15-35.

Petushkova NA, Pyatnitskiy MA, Rudenko VA, et al. 2014.

Santos Neto M, *et al.,* Spatial distribution of tuberculosis cases in a priority Brazilian northeast municipality for control of the disease. *International Journal of Development Research*. Volume: 7, Article ID: 10611, 6 pages.

Strauss T, von Maltitz MJ 2017. Generalising Ward's Method for Use with Manhattan Distances. PLoS ONE 12(1): e0168288. doi:10.1371/journal.pone.0168288.

Wierzchoń S.T., Kłopotek M.A. 2018. Cluster Analysis. In: Modern Algorithms of Cluster Analysis. Studies in Big Data, vol 34. Springer, Cham.

World Health Organization-WHO. Global Tuberculosis Report 2017 [Internet]. Geneva: WHO; 2020 [cited 2020 Feb 7]. Available from: <http://apps.who.int/iris/bitstream/ 10665/259366/1/9789241565516-eng.pdf?ua=1>.

Zaiger, D. 1983. "Inequalities for the Gini coefficient of composite populations", *Journal of Mathematical Economics*, 12.

Zhang Z, Murtagh F, Van Poucke, S Lin, S and Lan P. 2017. Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. *Annals Of Translational Medicine*, 5(4), 9. doi:10.21037/13789.

****