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

IJDR

International Journal of Development Research
Vol. 10, Issue, 04, pp. 35478-35483, April, 2020
https://doi.org/10.37118/ijdr.18753.04.2020

ISSN: 2230-9926

RESEARCH ARTICLE

OPEN ACCESS

WARD-LIKE HIERARCHICAL CLUSTERING WITH DISSIMILARITIES AND NON-UNIFORM WEIGHTS IN CASES OF TUBERCULOSIS IN PARAÍBA, BRAZIL

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

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

ARTICLE INFO

Article History:
Received 03rd January, 2020
Received in revised form
14th February, 2020
Accepted 11th March, 2020
Published online 30th April, 2020

Key Words:

Ward-like algorithm, Spatial constraints, Tuberculosis, State of Paraíba.

**Corresponding author: Dalila Camêlo Aguiar*

ABSTRACT

In this article, we propose to present a solution based on socio-epidemiological variables of TB, considering a clustering with spatial/geographical constraints for the State of Paraíba, Brazil. The Ward-Like hierarchical clustering method uses two dissimilarity matrices, 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 constraint space calculated from the geographical distances (D_1) together with an α mixing parameter and the non-uniform weight *w* assigned to the calculation of the dissimilarity matrix defined by the diversification coefficient (DC) of TB. Statistical analyses were undertaken in R. According to DC, most micro-regions are diversified, indicating that the epidemiological situation of TB does not depend on any specific variable. In D_0, the clusters are dispersed and are not strictly contiguous. Geographically more compact clusters are obtained after the introduction of D_1 and $\alpha = 0.2$, slightly favoring socioepidemiological homogeneity (26.11%) versus geographic homogeneity (17.58%), mainly influenced by cluster 2. Clusters 3 and 5 were separated based on the proportion of TB patients of working age. Cluster 4 had the lowest cure proportion of all clusters. The Ward-Like algorithm is shown to be viable in socio-epidemiological studies in understanding the behavior of TB from a spatial perspective.

Citation: *Dalila Camêlo Aguiar, Ramón Gutiérrez Sánchez and Edwirde Luiz Silva Camêlo et al.* "Ward-like hierarchical clustering with dissimilarities and non-uniform weights in cases of tuberculosis in paraíba, Brazil*", International Journal of Development Research*, 10, (04), 35478-35483.

INTRODUCTION

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 choose the clustering method and the optimal number of clusters. In TB epidemiology, for example, this challenge is great for being a data-driven approach involving many subjective decisions. However, in some clustering problems, it is relevant to impose constraints on the set of allowed solutions. Tuberculosis (TB) still poses a huge global health threat, with some 10 million new cases per year. In Brazil, it is estimated that the incidence of TB is increasing after many years of decline, owing to an upward trend between 2016 and 2018 (WHO, 2019). TB incidence is disproportionately high among people in poverty (Reis-Santos, 2019). 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 (Brazil, 2019). In State of Paraíba, the situation is even more critical, Aguiar *et al* (2019) identified a cure rate of 55% in the studied period (2007–2016). The State of Paraíba is composed of 223 municipalities it has 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). The remarkable relation that TB has with social conditions demands anunderstanding of the dynamics of this aggravation and its occurrence in the territory through geospatial analyses (Santos Neto *et al.*, 2017). The aim of this study is present a solution based on socio-epidemiological variables considering the Ward-like clustering with non-Euclidean dissimilarities and non-uniform weights attributed to the diversification coefficient of TB in the 23 microregions of the State of Paraíba in defining the importance of the constraint in the clustering procedure through the mixing parameter α.

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 data are reported cases of TB in the State of Paraíba, 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 diversification coefficient of TB in the State of Paraíba. Data collection took place during February 2020. As units of analysis, municipalities and microregions were used. 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: Usually the researcher is faced with the difficulty of clustering a set of n objects into k disjoint clusters. 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. The most common type is the contiguity constraints (in space or in time). Such constraints 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 (Chavent, 2017b). An adjacency matrix is used to find a connection between the borders of each city in the State of Paraíba. So, two clusters are regarded as contiguous if there are two objects, one from each cluster, which are linked in the contiguity matrix. Several authors in different areas of knowledge have implemented of constrained clustering procedures (Duque *et al.* 2011, Bécue-Bertaut *et al.* 2017, Dehman *et al.* 2015, Legendre 2014, and Ambroise *et al.* (1997, 1998)).

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_1provides the differences in "constraint space". 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. 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. 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 function *hclustgeo* of the *ClustGeo* package performs the hierarchical clustering of *Ward.D*, using a 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) and the weights $w = (w_1, \dots, w_n)$ of observations as arguments. Here the diversification coefficient (DC) socio-epidemiological of the microregions of the State of Paraíba will be applied as non-uniform weights. The sum of the heights in the dendrogram is equal to the total pseudo-inertia of the data set. The formula for pseudo-inertia is:

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

Where $\mu_k = \sum_{i \in C_k} w_i$ is the weight of C_k. The lower the pseudo-inertia I (C_k), the more homogeneous are the observations belonging to the cluster C_k. The function *hclustgeo* is a wrapper of the usual *hclust* function with the following arguments: a) Distance: D_o (Manhattan distance). The socio-epidemiological distances; b) Distance: D_1. The geographic distances between the municipalities; calculating a distance matrix for geographic points using R through packages: *sgeostat* (Majure and Gebhardt, 2016), *geosphere* (Hijmans, 2019) and Imap (Wallace, 2012). These functions calculate distance matrix for geographic for latitude and longitude points of the center of gravity of the municipalities; c) Methods: "Ward.D" and d) Members: $w = DC_i^* = \frac{LDC_i - 1}{L - 1}$ (diversification coefficient). 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.

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

Diversification coefficient: The diversification coefficient tries to measure the degree to which the value of a TB notification in a microregion comes from a variety more or less accused of different variables (new cases TB or relapse for instance), or if on the contrary, it comes from a relatively low number of variables. If a microregion has a high coefficient of specialization, it is because its occurrence is more influenced by a specific variable, in which case, diversification is minimal. On the other hand, if a microregion is classified as diversified, it means that its TB epidemiological situation does not depend much on any specific variable, that is, they are all equally influenced by the set of variables, in whose case diversification is maximum. The diversification coefficient of microregion i is defined as follows (González and Céspedes, 2004):

$$DC_i = \frac{\left(\sum_{j=1}^{L} Y_{ij}\right)^2}{L\sum_{j=1}^{L} Y_{ij}^2} \qquad (3)$$

Where DC is the magnitude of socio-epidemiological variables, whose data are in the form of a matrix, where Y_{ii} is the value that takes the socio-epidemiological variable j (j=1,…,4) in microregion i *(i=1,…,23)*. The DC is a quantity between ${}^{1}/_{L}$ and 1, $\frac{1}{L} \leq DC_i \leq 1$, being${}^{1}/_{L}$when the diversification is minimal and 1 when it is maximum. The following formula is used to normalize this coefficient between zero and one: $DC_i^* = L/L - 1(Di - 1/L)$ or, of equivalent form: $DC_i^* = \frac{LDC_i - 1}{L-1}$.

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. It is important to know whether the TB situation in the State of Paraíba is diversified or not. Based on the socio-epidemiological variables, we will calculate the diversification coefficient. The values of the diversification coefficient for 223 microregions are shown in Figure 1

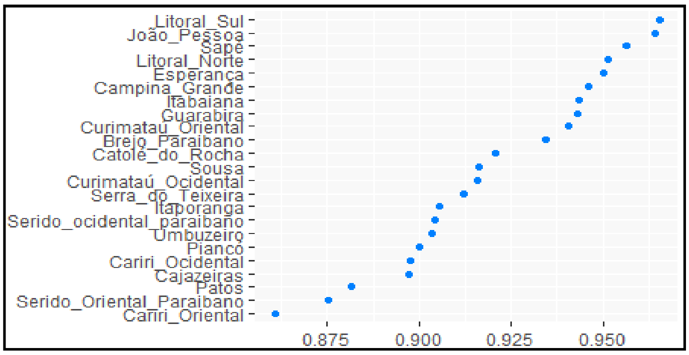

Figure 1. Values of the diversification coefficient (DC) of socio-epidemiological variables of microregions, Paraíba, Brazil

If a microregion is classified as diversified, it means that its TB epidemiological situation does not depend much on any specific variable, that is, they are all equally influenced by the set of variables. It can be seen in Figure 1 that most microregions have a diversification measure close to 1, with minimum value of approximately 0.68 and a maximum of 0.967, Cariri Oriental and Litoral Sul respectively. Diversification in the Cariri Oriental microregion is diminished by the existence of inequality between the variables epidemiological (new cases and cure) and social variables (less than 10 years' formal education and working age (20-49)), focusing more on one of them. This diversification coefficient is the weight of the constraint on the quality of the solutions and is controlled by α which defines the importance of the constraint in the cluster procedure. 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 2 shows the dendrogram of the dissimilarity matrix D_0, that is, the differences in the feature space of socio-epidemiological variables and map of the partition corresponding to the five clusters.

Figure 2 (a) shows the dendrogram according to Ward-Like method criterion using the distance matrix of the 223 municipalities using only the four socio-epidemiological variables according to diversification measures. The visual inspection of the dendrogram in Figure 2a suggests to retain K = 5 clusters. We can use the map provided in the estuary data to visualize the corresponding partition in five clusters Figure 2 (b). Geographically, we perceive clusters well dispersed according with socio-epidemiological variables, that is, the clusters are not strictly contiguous. It is observed that the 5 clusters are well spread out within the State of Paraíba. An important feature is the city of Maturéia, Mogeiro, Belém, Lucena e Riacho de Santo Antônio. All cities were well distributed according to groupings. Through the *choicealpha function* of the package *ClustGeo* find an alpha value for relative importance between the D_0 and D_1 dissimilarity matrices. An alpha value of 0.3 was considered shown the partition taking into account the geographical constraints in Figure 3. Obtaining the partition taking into account the geographic constraints in Figure 3, shows the value α which aims to increase the spatial contiguity, seen in detail in Table 1.

Table 1. Normalized proportion of explained pseudo-inertias

Alpha values	Q_0norm	Q_1norm
$\boldsymbol{\alpha = 0.17}$	0.80773244	0.68104151
$\boldsymbol{\alpha = 0.18}$	0.71786338	0.76987949
$\boldsymbol{\alpha = 0.19}$	0.75936603	0.74331210
$\boldsymbol{\alpha = 0.20}$	0.73858351	0.82422833
$\boldsymbol{\alpha = 0.21}$	0.75132496	0.80288570

When $\alpha = 0$ the geographical dissimilarities are not taken into account and when α=1 it is the socio-epidemiologic distances which are not taken into account and the clusters are obtained with the geographical distances only. The plot in Figure 3 (left) would appear to suggest choosing $\alpha = 0.2$ which corresponds to a loss of only (1-0.7385 = 26,11%) of socio-epidemiologic with diversification coefficient of each city, and 17,58% increase in geographical homogeneity. The increased geographical cohesion of this partition can be seen in Figure 4.

Figure 4 a gain in spatial homogeneity is perceived, mainly in cluster 2, next appears cluster 1. The municipalities in yellow circles went well located in the microregions of Cariri Ocidental, Piancó, Cajazeiras and Seridó Oriental for cluster 2 and the municipalities of the Litoral Norte in cluster 1. Significant changes occurred mainly in cluster 3. Figure 5 shows the boxplots of the variables for each cluster of the partition Figure 4. Cluster 1 presented a behavior similar to cluster 2. It seems that groups 3 and 5 were separated based on the proportion of TB patients in working age, because the municipalities in cluster 3 have lower proportions of TB patients in working age and with less than 10 years of study,

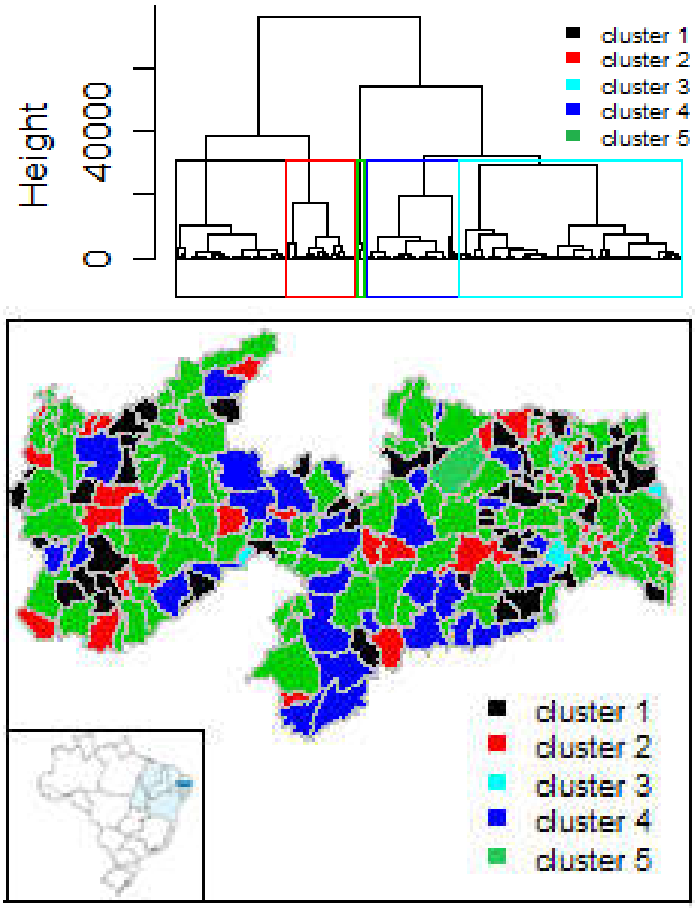

Figura 2. a) Dendrogram of the *n* = 223 municipalities based on the *4* socio-epidemiologic variables (that is using D_0 only). b) Map of the partition with 5 clusters only based on the socio-epidemiological variables for diversification coefficient

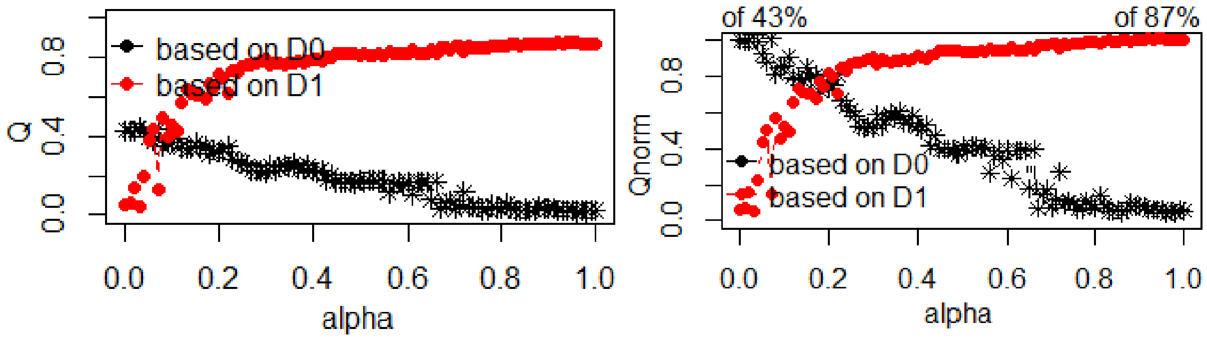

Figure 3. Choice of α for a partition in $K = 5$ clusters when D_1 is the geographical distances between municipalities. Left: proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in dashed line). Right: normalized proportion of explained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

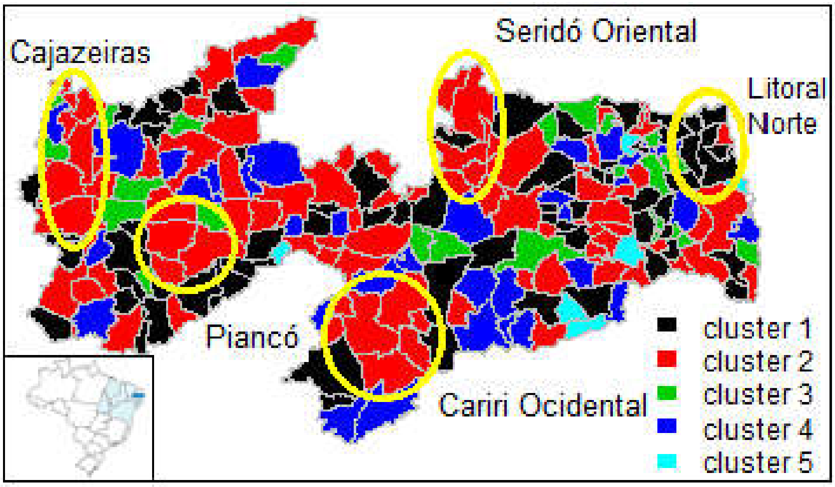

Figure 4. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the geographical distances between the municipalities D_1 with $\alpha = 0.2$.

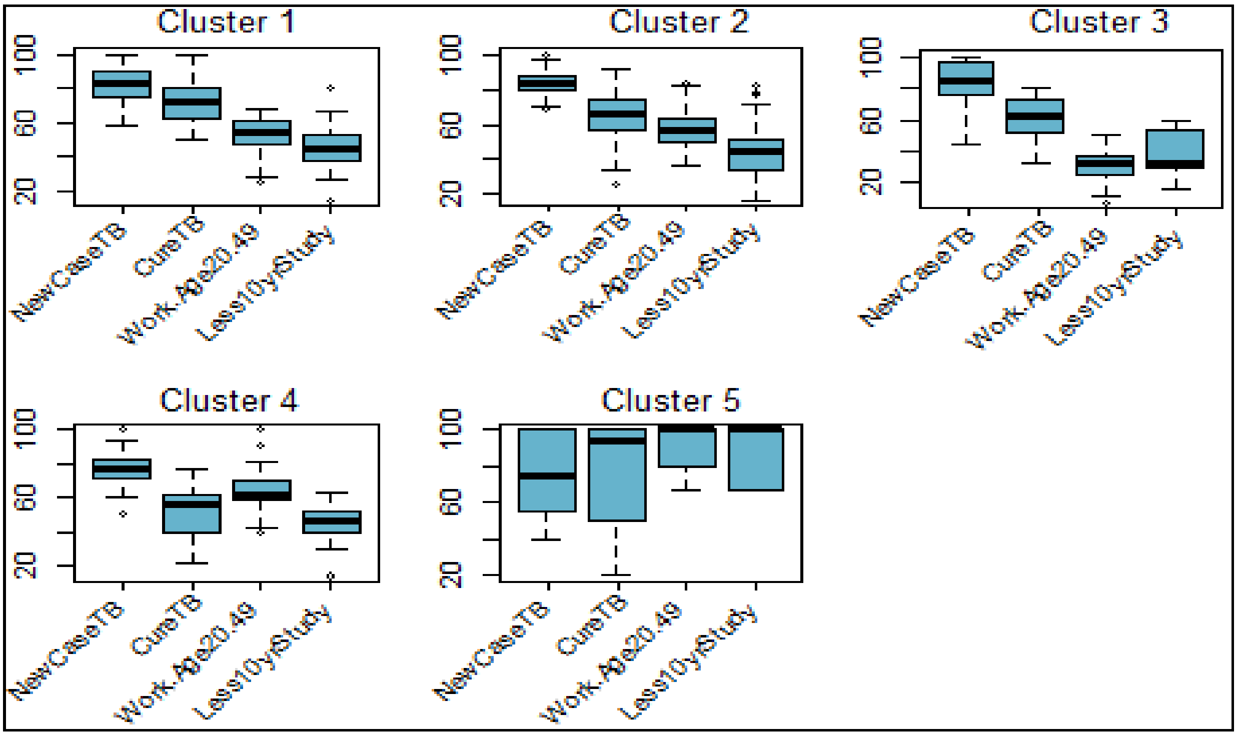

Figure 5. Comparison of clusters in the partition of Figure 5 in terms of variables

the opposite occurs in cluster 5, with higher proportions of people with less than 10 years of study at working age. Clusters 1, 2, 3 and 4 are characterized by the high proportion of new cases with greater variation in cluster 5. Cluster 4 has the lowest cure rate of all clusters. Although it has the lowest median proportion of new cases, cluster 5 has high rates of cure, higher proportions of people of working age and with less than 10 years of schooling, in 6 municipalities, Maturéia, Gado Bravo, Mogeiro, Belém , Lucena and Umbuzeiro.

Conclusion

When considering spatial/geographical constraints, the hierarchical clustering becomes even more complete, as it detects patterns in data sets of different dimensions. Therefore, the application of the Ward-Like method becomes indispensable for a better understanding of the socio-epidemiological reality of the State of Paraíba from a spatial perspective.

Acknowledgment: None.

REFERENCES

Aguiar DC, Silva Camelo EL and Carneiro RO. 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., 2019.

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), geo ENV I-Geostatistics for Environmental Applicattions, Kluwer, Dordrecht, pp. 493-504.

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.

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. ClustGeo: an R package for hierarchical clustering with spatial constraints. *Comput Stat* 33, 1799–1822 (2018a). <https://doi.org/10.1007/s00180-018-0791-1>.

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.

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

Hijmans RJ. 2019. geosphere: Spherical Trigonometry. R package version 1.5-10. <https://CRAN.R-project.org/package=geosphere>.

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

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

Majure JJ, Gebhardt A. 2016). sgeostat: An Object-Oriented Framework for Geostatistical Modeling in S+. R package version 1.0-27. <https://CRAN.R-project.org/package=sgeostat>.

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

Reis-Santos B, Shete P, Bertolde A, Sales CM, Sanchez MN, *et al.* (2019) Tuberculosis in Brazil and cash transfer programs: A longitudinal database study of the effect of cash transfer on cure rates. PLOS ONE 14(2): e0212617. https://doi.org/10.1371/journal.pone.0212617.

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.

SINAN - Sistema de Informação de Agravos de Notificação. Tuberculose – casos confirmados no Sistema de Informação de Agravos de Notificação. Ministério da Saúde, Brazil: Brasília, DF; 2020 [citado em 2020 fevereiro 7]. Disponível em: http://www2.datasus.gov.br/.

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.

Wallace JR (2012). Imap: Interactive Mapping. R package version 1.32. <https://CRAN.R-project.org/package=Imap>.

WHO - World Health Organization. Global tuberculosis report 2019. Geneva: World Health Organization; 2019. Licence: CC BY-NC-SA 3.0 IGO.

Wierzchoń ST, Kłopotek MA. (2018). Cluster Analysis. In: Modern Algorithms of Cluster Analysis. Studies in Big Data, vol 34. Springer, Cham.

****