UNIVERSIDAD
DE GRANADA

Departamento de Estadística e Investigación Operativa
Programa de Doctorado en Estadística Matemática y Aplicada

Modelo de conglomerado para el mapa de datos epidemiológicos

Tesis Doctoral

Dalila Camêlo Aguiar
Granada. 2021

UNIVERSIDAD
DE GRANADA

Departamento de Estadística e Investigación Operativa
Programa de Doctorado en Estadística Matemática y Aplicada

MODELO DE CONGLOMERADO PARA EL MAPA DE DATOS EPIDEMIOLÓGICOS

Tesis Doctoral

Dalila Camêlo Aguiar
Granada. 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Dalila Camêlo Aguiar
ISBN: 978-84-1306-967-8
URI: http://hdl.handle.net/10481/69864

Agradecimientos

En primer lugar agradezco al Eterno, Creador del cielo y de la tierra, que me dio vida para completar este trabajo.

A mi tutor y director de doctorado, Prof. Ramón Gutiérrez Sánchez, por darme la oportunidad de realizar esta tesis, y por todo el apoyo que me ha dado a lo largo de mi tiempo de estudiante, aportarme la confianza y seguridad necesarias para llevarla a cabo.

También estoy muy agradecida de haber sido estudiante en la Universidad de Granada. Mi gratitud a todos los profesores que he tenido, en especial al Dr. D. Andrés González Carmona, a mis examinadores, y a los miembros de este tribunal de tesis por tomarse el tiempo para brindarme comentarios tan útiles.

Especial agradecimiento a mi primo y esposo Edwirde, por su paciencia, comprensión y incentivo. A mis padres, José y Iolanda. A mis hermanos.

Contenido

Índice de figuras

Índice de tablas

Modelo de conglomerado para el mapa de datos epidemiológicos

Dalila Camêlo Aguiar

Resumen

En esta tesis doctoral, se presenta una solución basada en variables socioepidemiológicas a partir de casos notificados de tuberculosis, considerando un agrupamiento jerárquico similar al de Ward, llamado Ward-like, con inclusión de restricciones espaciales/geográficas; donde se ingresan dos matrices de disimilitud (D_0 y D_1). La primera en el espacio de característica (D_0) son las variables socioepidemiológicas y la segunda matriz de disimilitud en el espacio de restricciones son las distancias geográficas (D_1) entre los municipios y los pesos w son un nuevo conjunto de observaciones en los municipios, junto con un parámetro de mezcla α [0; 1] que permite al usuario establecer la importancia de cada matriz de disimilitud en el procedimiento de agrupamiento controlando el peso de la restricción en la calidad de las soluciones, a través de un valor de alfa que aumente la contigüidad espacial sin deteriorar significativamente la calidad de la solución en función de las variables de interés, es decir, las del espacio característico. Los datos analizados en los tres estudios son casos notificados de tuberculosis en los 223 municipios del Estado de Paraíba/Brasil en el período comprendido entre 2001 y 2018. Las variables son proporciones y se dividen en epidemiológicas y variables sociales. Los tres estudios respectivamente tomaron w diferentes: peso uniforme (índice de desigualdad colectiva del producto interno bruto) y pesos no uniformes (coeficiente de diversificación de la tuberculosis y la razón de incidencia estandarizada de tuberculosis). Los resultados contribuyeron significativamente al aumento de la claridad, tanto desde el punto de vista espacial como socioepidemiológico. El método se muestra viable en estudios epidemiológicos en la comprensión conjunta de factores de diferentes dimensiones, agregados desde una perspectiva espacial.

Por tanto es una herramienta de análisis que permite conocer mejor la realidad socioepidemiológica de los municipios.

Modelo de conglomerado para el mapa de datos epidemiológicos

Dalila Camêlo Aguiar

Resumo

Nesta tese de doutorado, apresenta-se uma solução baseada em variáveis socioepidemiológicas a partir dos casos notificados de tuberculose, considerando um agrupamento hierárquico semelhante ao de Ward, denominado Ward-like, incluindo restrições espaciais/geográficas; onde são introduzidas duas matrizes dissimilares (D_0 e D_1). A primeira no espaço característico (D_0) sendo elas variáveis socioepidemiológicas e a segunda matriz de dissimilaridade no espaço de restrições são as distâncias geográficas (D_1) entre os municípios e os pesos w são um novo conjunto de observações nos municípios, junto com um parâmetro de mistura α [0; 1] que permite ao usuário estabelecer a importância de cada matriz de dissimilaridade no procedimento de agrupamento, controlando o peso da restrição na qualidade das soluções, através de um valor alfa que aumenta a contiguidade espacial sem deteriorar significativamente a qualidade da solução em função das variáveis de interesse, ou seja, do espaço característico. Os dados analisados nos três estudos são casos notificados de tuberculose nos 223 municípios do Estado da Paraíba/Brasil no período de 2001 a 2018. As variáveis são proporções e se dividem em variáveis epidemiológicas e sociais. Os três estudos assumiram, respectivamente, w diferentes: peso uniforme (índice de desigualdade coletiva do produto interno bruto) e pesos não uniformes (coeficiente de diversificação da tuberculose e razão de incidência padronizada da tuberculose). Os resultados contribuíram significativamente para o aumento da clareza, tanto do ponto de vista espacial quanto socioepidemiológico. O método demonstrou ser viável em estudos epidemiológicos na compreensão conjunta de fatores de diferentes dimensões, agregados a partir de uma perspectiva espacial. Destarte, trata-se de uma ferramenta de aná-

lise que nos permite compreender melhor a realidade sócioepidemiológica dos municípios.

Capítulo 1

Introducción

1.1. Presentación

Pinker [1] indicó que un ser inteligente no puede tratar cada objeto que ve como una entidad única y diferente de cualquier otra cosa en el universo. Tiene que clasificar los objetos en categorías para poder aplicar al objeto en cuestión los conocimientos que ha adquirido con tanto esfuerzo sobre objetos similares encontrados en el pasado.

Además de ser una actividad conceptual humana básica, la clasificación de los fenómenos que se estudian es un componente importante de prácticamente toda la investigación científica. En las ciencias del comportamiento, por ejemplo, estos "fenómenos" pueden ser individuos o sociedades, o incluso patrones de comportamiento o percepción.

El investigador suele estar interesado en encontrar una clasificación en la que los elementos de interés se clasifiquen en un pequeño número de grupos o conglomerados homogéneos. Por lo general, la clasificación requerida es aquella en la que los grupos son mutuamente excluyentes (un elemento pertenece a un solo conglomerado) en lugar de superponerse (los elementos pueden ser miembros de más

de un conglomerado). Como mínimo, cualquier esquema de clasificación derivado debería proporcionar un método conveniente para organizar un conjunto grande y complejo de datos multivariados, con las etiquetas de clase proporcionando una forma parsimoniosa de describir los patrones de similitudes y diferencias en los datos (Everitt y Torsten, 2011 [2].

Las técnicas de conglomerados pueden lograr sus propósitos desde una perspectiva exploratoria. Varias son las situaciones en las cuales el análisis de conglomerados se hace presente y se aplica en muchos campos, como las Ciencias Naturales, la Medicina, etc.

Por ejemplo, en Psicología, donde se utiliza para clasificar a las personas según sus perfiles de personalidad (Speece et al., 1985 [3]); en la Investigación de Mercados, en la identificación del posicionamiento del producto (o servicio) en relación con los competidores del mercado y en la segmentación de los clientes según perfiles de consumo (Punj y Stewart, 1983 [4]).

También se aplica a los estudios socioepidemiologicos, donde la clasificación de enfermedades proporciona los fundamentos clave para la identificación y el tratamiento de enfermedades (Song, Merajver y Li, 2015 [5]).

El análisis de conglomerados, también, es muy útil para caracterizar las condiciones socioeconómicas de las poblaciones y avanzar hacia la profundización de las problemáticas que se funden en las condiciones de vida en las localidades y los territorios (Parra-Sánchez et al., 2017 [6]).

El análisis de conglomerados implica la categorización: dividir un grupo grande de observaciones en grupos más pequeños para que las observaciones dentro de cada un sean relativamente similares (es decir, para que en su mayoría tengan las mismas características) y las observaciones de grupos diferentes relativamente

distintos. El análisis de conglomerados se suele utilizar en la reducción de los datos para identificar un pequeño número de grupos (Helmuth, 1980 [7]; Everitt, 1992 [8]; Encinas, 2001 [9]).

La mayor parte del análisis de conglomerados se realiza con el objetivo de abordar la heterogeneidad de los datos. En lugar de tratar con un grupo de observaciones muy divergentes, dividimos explícitamente el grupo en subconjuntos más homogéneos. Sin embargo, separar los datos en subconjuntos más homogéneos no es lo mismo que encontrar agrupaciones de origen natural. Encontrar agrupaciones de origen natural requiere que haya grupos de observaciones con densidad local relativamente alta (es decir, hay muchas otras observaciones dentro de la misma área pequeña) separados por regiones de densidad local relativamente baja. En este caso, las agrupaciones en sí corresponden a una modalidad de datos y el número de agrupaciones corresponde a la cantidad de modas en una distribución de datos multimodal.

Mediante el uso de ejemplos y aplicaciones, esperamos hacer clara la distinción entre tratar la heterogeneidad (que siempre es posible, aunque no necesariamente deseable) y encontrar agrupaciones naturales, que solo es posible cuando la modalidad de datos (es decir, el número de modas en la distribución subyacente) es mayor que uno, como fue definido por Lattin et al., 2011 [10].

Se han desarrollado muchos enfoques diferentes para el análisis de conglomerados. En esta Tesis discutiremos los métodos de *árbol jerárquico*, en los que la solución de k conglomerados se forma mediante la unión de dos conglomerados $k+1$ y el método de partición (que separa las observaciones en un número determinado de subgrupos, y en el que la solución de la agrupación k y la solución de agrupación $k+1$ no están necesariamente alineadas.

En el análisis de conglomerados, las filas se agrupan, es decir, el objetivo es

analizar la estructura de un conjunto de objetos; su principal finalidad es que el análisis de conglomerados realiza una clasificación de objetos; cada objeto (encuestado) es un punto en el espacio de características y la tarea del análisis de conglomerados es la selección de la ”condensación” de puntos, la división de la población en subconjuntos homogéneos de objetos (segmentación).

En resumen, el agrupamiento es el proceso de dividir una muestra dada de objetos (observaciones) en subconjuntos (generalmente no superpuestos), llamados conglomerados, de modo que cada conglomerado consta de objetos similares y los objetos de diferentes conglomerados difieren significativamente.

Uno de los objetivos de la agrupación por conglomerados es identificar las relaciones internas entre los datos definiendo la estructura del conglomerado. La división de las observaciones en conglomerado de objetos similares permite simplificar el procesamiento de datos y la toma de decisiones mediante la aplicación de su propio método de análisis a cada conglomerado: *divide et impera* (“divide y vencerás”).

Una de las aplicaciones de la agrupación en conglomerados es resolver el problema de la compresión de datos. Si la muestra inicial es excesivamente grande, puede reducirla, dejando algunos de los representantes más característicos de cada conglomerado.

Para resolver problemas utilizando métodos de análisis de conglomerados, es necesario establecer el número de conglomerados por adelantado. En un lado, se intenta reducir el número de agrupaciones. En otro lado, es más importante garantizar un alto grado de similitud de objetos dentro de cada conglomerado, y puede haber tantos conglomerados como desee. En el tercer caso, son los objetos individuales que no encajan en ninguno de los conglomerados.

1.2. Motivación

La notable relación de incidencia que tiene la tuberculosis con las condiciones sociales exige conocer la dinámica de este agravamiento y su ocurrencia en el territorio (Santos Neto et al., 2017 [11]). Los enfoques de agrupamiento son una herramienta útil para detectar patrones en conjuntos de datos y generar hipótesis sobre posibles relaciones.

El método de conglomerado jerárquico Ward-like considera dos matrices de disimilitud con restricciones espaciales e geográficas. La primera matriz proporciona las disimilitudes en el espacio de característica calculadas con las variables socioepidemiológicas de la tuberculosis en el Estado de Paraíba/Brasil y la segunda matriz proporciona las disimilitudes en el espacio de restricción calculado a partir de las distancias geográficas entre los municipios. También determina un valor del parámetro de mezcla alfa que aumente la contigüidad espacial sin deteriorar significativamente la calidad de la solución en función de las variables de interés, es decir, las del espacio de característica. Por tanto, es fundamental considerar este enfoque en estudios socioepidemiológicos en la comprensión conjunta de factores de diferentes dimensiones, agregados desde una perspectiva espacial.

1.3. Objetivos

Esta Tesis se desarrolla en el contexto del método de conglomeredo jerárquico Ward-like. El algoritmo se utiliza de dos matrices de disimilitud D_0 y D_1 y un parámetro de mezcla $\alpha[0,1]$ a fin de econtrar soluções considerando restricciones espaciales y geográficas. Los principales objetivos que se buscan con la realización de esta Tesis son los siguientes:

- Estudiar el método de conglomerado Ward-like.
- Aplicar el algoritmo jerárquico Ward-like a ejemplos reales.

- Presentar una solución basada en variables socioepidemiológicas de la tuberculosis considerando un agrupamiento con restricciones espaciales/geográficas.
- Utilizar vectores de pesos uniforme y no uniformes.
- Determinar un valor del parámetro de mezcla alfa que aumente la contigüidad espacial sin deteriorar significativamente la calidad de la solución en función de las variables de interés, es decir, las del espacio de características, ya que alfa establece la importancia de la restricción en el procedimiento de agrupamiento y controla el peso de la restricción en la calidad de las soluciones.
- Obtener gráficos con resultados que se puedan visualizar e interpretar fácilmente en un mapa utilizando el paquete `ClustGeo`.
- Analizar los resultados y comprender la dinámica socioepidemiológica de la tuberculosis en el territorio del Estado de Paraíba.

1.4. Estructura de la Tesis.

La presente Tesis se divide en ocho capítulos. En esta sección, se presenta una breve introducción al análisis de conglomerados y objetivos que se pretenden alcanzar con la realización de esta Tesis (ver Capítulo 1). La metodologia de análisis de conglomerados y conglomerado jerárquico restringido (Capítulo 2), modelo de conglomerado jerárquico Ward-like incluye los capítulos (ver Capítulos 3 y 4).La detección de valores atípicos (ver Capítulo 5) y distancia geografica entre los municípios de Paraíba (ver Capítulo 6). Así, como los resultados de los estudios (ver Capítulo 7). La aplicación real del método de conglomerado jerárquico Ward-like que incluye restricciones espaciales/geográficas se dará a través de tres estudios utilizando casos notificados de tuberculosis en los 223 municipios del Estado de Paraíba/Brasil, donde se ingresan dos matrices de disimilitud, junto con un parámetro de mezcla $\alpha[0;1]$ que permite al usuario establecer la importancia de cada

matriz de disimilitud en el procedimiento de agrupamiento controlando el peso de la restricción en la calidad de las soluciones. La primera matriz de disimilitud en el espacio de característica (D_0) son variables socioepidemiológicas y la segunda matriz de disimilitud en el espacio de restricciones son las distancias geográficas (D_1) entre los municipios y los pesos w son un nuevo conjunto de observaciones en los municipios. Los tres estudios respectivamente tomaron w diferentes: índice de desigualdad colectiva de del producto interno bruto, coeficiente de diversificación de la tuberculosis y la razón de incidencia estandarizada de tuberculosis. Consideraciones finales y trabajo futuro (ver Capítulo 8). Las consideraciones finales también se pueden consultar en portugués (ver Capítulo 9). Al final de este documento se incluye el código fuente de la aplicación práctica (ver Anexo 1), artículos publicados en revistas científicas (ver Anexo 2) y una bibliografía.

Capítulo 2

Análisis de conglomerados

Casi todos los problemas de agrupación de cualquier tamaño apreciable requieren una solución heurística. Esto se debe a que a medida que aumenta el número de objetos en el conjunto de datos, el número de posibles soluciones de agrupación crece espectacularmente. El número de formas diferentes de dividir n objetos en k agrupaciones de tamaño n_1, n_2, n_3, ..., n_k dado por:

$$n!/[n_1!n_2!n_3!\ldots n_k!k!]$$

Donde n es el número de objetos.

Para tener una idea de la magnitud del resultado de cálculo, considerando un problema con tiene 12 objetos (municipios en esta Tesis). El número de maneras distintas de dividir esos objetos en cuatro conglomerados de igual tamaño ($n_1 = \ldots = n_5 = 3$) es mayor que 15400. De forma que se desea dividirlos en un número de conglomerados prefijado, de manera que (Peña, 2002 [13]):

- cada elemento pertenezca a uno, y sólo uno, de los conglomerados;
- todo elemento quede clasificado;
- cada grupo sea internamente homogéneo.

tal que se puedan estructurar los elementos de un conjunto de forma jerárquica por su similitud. Estrictamente, estos métodos no definen conglomerados, sino la estructura de asociación en cadena que pueda existir entre los elementos. La jerarquía construida permite obtener también una partición de los datos en conglomerados y también en un mapa temático de acuerdo con los conglomerados creados en el dendrograma.

Segundo Peres et al (2012) [15], el termino conglomerado debe ser usado cuando no existe cualquier información sobre como es la organización de los datos. Así, normalmente, se denomina conglomerado el proceso por el cual se estudian las relaciones de similaridad entre los objetos. La tarea de agrupamiento puede ser descrita como la busca por una función H, que tenga capacidad de mapear un conjunto de X de vectores de entrada $\vec{x} \in E^d \times W \rightarrow C$, donde que d es la dimensión del espacio E, o sea, el número de coordenadas del vector $\vec{x}$, W es un espacio de parámetros ajustables por medio de un algoritmo de inducción no supervisado e $W = arg\_\text{mín}_w\, dist(\vec{x}p, \vec{x}q)$, siendo p e q los índices de dos objectos cualesquiera y distintos asociados a un mismo conglomerado.

También, tradicionalmente, las siguientes condiciones se deben ser consideradas en la resolución del conglomerado, en que c es el número de grupos en el modelo de grupos resultantes (Leandro et al., 2016 [17]):

$$C_k \neq, k = 1, \ldots, c \cup_{k=1}^{c} C_k = X; C_k \cap C_l, k, l = 1, \cdots, cek \neq l.$$

El procedimiento para llevar a cabo un Análisis de Conglomerados son dividido en cinco pasos (Encinas, 2001 [9]):

Paso 1 - *Selección*: esta primera fase consiste en seleccionar a los individuos objeto del estudio.
Paso 2 - *Determinación de la matriz de disimilitudes*: se trata de determinar las distancias, similitudes o disimilitudes de los individuos.
Paso 3 - *Ejecución del algoritmo*: una vez determinadas las disimilitudes de los individuos, se procede a ejecutar el algoritmo que formará las diferentes agrupaciones o conglomerados de individuos.
Paso 4 - *Representación gráfica*: determinada la clasificación, el paso siguiente consiste en obtener una representación gráfica (dendrograma) de los conglomerados obtenidos, de modo que se puedan visualizar los resultados obtenidos.
Paso 5 - *Interpretación*: conseguido el propósito de la clasificación, el último paso a llevar a cabo es la interpretación de los resultados alcanzados.

2.1. Métodos jerárquicos

2.1.1. Distancias y similaridades

Una de las técnicas matemáticas más probadas para comparar relaciones de manera más sistemática es medir la distancia entre dos puntos. Por el contrario, podríamos intentar medir la similitud entre dos puntos (distancias pequeñas que corresponden a grandes similitudes y grandes distancias que corresponden a pequeñas similitudes). También, podríamos intentar medir la similitud entre dos puntos (distancias pequeñas que corresponden a grandes similitudes y grandes distancias que corresponden a pequeñas similitudes).

La mayoría de los esfuerzos para producir una estructura de conglomerado bastante simple a partir de un conjunto de datos complejo requieren una medida de "proximidad" o "similitud". A menudo hay una gran subjetividad involucrada en

la elección de una medida de similitud. Las consideraciones importantes incluyen la naturaleza de las variables (discretas, continuas, binarias), escalas de medición (nominal, ordinal, intervalo, razón) y conocimiento de la materia. Cuando los elementos (unidades, casos) se agrupan, la proximidad suele indicarse mediante algún tipo de distancia. Por el contrario, las variables suelen agruparse en función de los coeficientes de correlación o medidas de asociación similares.

Las medidas de distancia más estándar en Matemáticas se llaman métricas, que deben satisfacer ciertas condiciones o axiomas (como ser simétrico).

Definición 2.1.1 *Generalmente una medida de distancia $d(R, S)$ entre dos puntos R y S satisface las siguientes propiedades, donde T es cualquier otro punto intermedio:*

Formalmente, una distancia d definida en un conjunto F es una aplicación entre el producto cartesiano $F \times F$ y los números reales no negativos $\Re^+ \cup \{0\} = [0, \infty)$, de modo que a cada par de elementos $(x, y) \in F \times F$ se le asigna un número real no negativo r que define la distancia entre los puntos x y b de F (López, 2005 [16]).

$$\begin{aligned} d : F \times F &\to \Re^+ \cup \{0\} = [0, \infty) \\ (x, y) &\to r \Leftrightarrow d(x, y) = r \end{aligned}$$

La distancia d verifica las siguientes condiciones:

- $d(x, y) \geq 0$. Toda distancia es definida positiva, es decir, la distancia entre dos elementos cualesquiera es mayor o igual que cero, y sólo es cero si $y = x$.
- $d(x, y) = d(y, x)$. Se trata de la propriedad de simetría, lo que equivale a decir que la distancia de x y y es la misma que la de y y x.
- $d(x, y) \leq d(x, z) + d(z, y)$. Se trata de la desigualdad triangular, es decir, la distancia entre dos puntos cualesquiera, x y y, es menor o igual que la suma de la distancia de x a un tercer punto z, más la distancia de z a y.

- Si $i \neq j$, entonces $d(i,j) > 0$.

Ahora definimos las medidas de distancia d que se utilizan a menudo en la agrupación. Sea $x' = (x_1, \ldots, x_r)$ y $y' = (y_1, \ldots, y_r)$. Luego

- *Distancia euclídea al cuadrado*: $d(i,j)^2 = \sum_k (x_{ik} - y_{jk})^2$. La distancia euclídea al cuadrado entre dos individuos se define como la suma de los cuadrados de las diferencias de todas las coordenadas de los dos puntos.
- *Distancia euclídea*: $d(x,y) = \sqrt{(x-y)'(x-y)}$. La distancia euclídea se define como la raíz cuadrada positiva de la distancia anterior.
- *Distancia estadística*: $d(x,y) = \sqrt{((x-y)'A(x-y)}$. Por lo general, $A = T^{-1}$, donde T contiene las varianzas y covarianzas muestrales. Sin embargo, sin un conocimiento previo de los distintos conglomerados, estas cantidades de muestra no se pueden calcular. Por esta razón, a menudo se prefiere la distancia euclidiana para la agrupación.
- *Métrica de Minkowski*: $d(x,y) = (\sum_{i=1}^{p} |x_i - y_i|^m)^{1/m}$. Para $m = 1, d(x,y)$ mide la distancia de la çiudad-bloque.[en]tre dos puntos en p dimensiones. Para $m = 2, d(x,y)$ se convierte en la distancia euclidiana. En general, la variación de m cambia el peso dado a diferencias más grandes y más pequeñas.
- *Distancia City-Block o de Manhattan*: calcula las diferencias absolutas entre las coordenadas de par de objetos (Karthikeyan y Gomathi, 2014 [12]). Matemáticamente, se puede representar como:

$$d_1(i,j) = \sum_k |x_{ik} - y_{jk}| \tag{2.1}$$

La distancia de Manhattan es un caso particular de la distancia o medida de Minkowski cuando $q = 1$ y resulta ser la suma de las diferencias, en valor absoluto,

de todas las coordenadas de los dos individuos cuya distancia se calcula.

En esta Tesis optamos por la distancia de Manhattan porque el método de Ward ya se ha generalizado para su uso en distancias no euclidianas. Según Strauss y Maltitz [14], el algoritmo de agrupación de Ward puede usarlo junto con las distancias de Manhattan.

Definición 2.1.2 *El coeficiente de similaridad según la variable $j = 1, \ldots, p$ entre dos elementos muestrales (e, f), se define como una función, s_{jef}, no negativa y simétrica:*

- $s_{jii} = 1$
- $0 \leq s_{jef} \leq 1$
- $s_{jef} = s_{jfe}$

Si obtenemos las similaridades entre dos elementos para cada variable podemos combinarlas en un coeficiente de similaridad global entre los elementos (Peña, 2002 [13]). El coeficiente propuesto por Gower es definido por

$$s_{ef} = \frac{\sum_{j=1}^{p} v_{jef} s_{jef}}{\sum_{j=1}^{p} v_{jef}}$$

donde v_{jef} es una variable ficticia que es igual a uno si la comparación de estos dos elementos la variable j tiene sentido, y será cero si no queremos incluir esa variable en la comparación.

2.1.2. Formación de los grupos: análisis jerárquico de conglomerados

Dado que, al calcular la matriz de distancias, se conoce qué observaciones están más próximas y más alejadas entre sí, es necesario formar los grupos, lo que implica tomar dos decisiones: seleccionar el algoritmo de agrupación que se elige

y determinar un número de grupos razonable (Hair et al, 1999 [18] y Jiménez y Manzano, 2005 [19]).

Tomar estas decisiones no es fácil, ya que existen docenas de algoritmos de agrupación. La mayoría de los autores, de hecho, recomiendan utilizar diversos procedimientos y comparar sus resultados (Sharma, 1996 [20]; Johnson, 1998 [21]). Si distintos métodos aportan agrupaciones similares, será razonable suponer que existe una agrupación natural objetiva. Si no fuera así, habría que examinar las distintas agrupaciones a la luz de un marco teórico o de trabajos precedentes para elegir el resultado más razonable.

Los métodos más utilizados de agrupación se dividen en dos grandes grupos: jerárquicos y no jerárquicos.

1. Métodos jerárquicos: los métodos jerárquicos o agrupamientos jerárquicos van generando grupos en cada una de las fases del proceso buscando el número de cluster hasta hacer una agrupación óptima, es decir, inicialmente, cada individuo es un grupo en sí mismo. Sucesivamente se van formando grupos de mayor tamaño fusionando grupos cercanos entre sí. Finalmente, todos los individuos se unen en un solo grupo.

2. Métodos no jerárquicos: en este método los grupos no se forman en un proceso secuencial de fusión de grupos de menor tamaño, es necesario establecer inicialmente un número de grupos *a priori* y los individuos se van clasificando en cada uno de esos grupos.

En esta Tesis, consideraremos los métodos jerárquicos, detallando principalmente el método de Ward. A continuación se detallan los principales métodos jerárquicos.

Método del vecino más próximo o enlace simple

Este método considera que la distancia entre dos conglomerados es la distancia más corta desde un miembro de un conglomerado a otro miembro de otro conglomerado, si los datos consisten en similaridades entre dos conglomerados, entonces se considera la mayor similaridad desde cualquier miembro de un conglomerado a otro miembro de otro conglomerado (Sneath y Sokal, 1973 [22]).

De manera que, si después de haber realizado la k-ésima etapa, tenemos formados $(n-k)$ conglomerados, la distancia entre el conglomerado C_i que posee n_i elementos, con el conglomerado C_j que posee n_j elementos, y dado que x_i pertenece al conjunto C_i y que x_j pertenecen al conjunto C_j seria entonces:

$$d(C_i, C_j) = Min\{d(x_l, x_m)\} \quad \text{para} \quad l = 1, 2, \ldots, n_i \quad \text{y} \quad m = 1, 2, \ldots, n_j$$

Al emplear medidas de distancia. Si tuviésemos similaridades, en ese caso tendríamos:

$$s(C_i, C_j) = Max\{s(x_l, x_m)\} \quad \text{para} \quad l = 1, 2, \ldots, n_i \quad \text{y} \quad m = 1, 2, \ldots, n_j.$$

Así que, el proceso de unir las dos agrupaciones en el proceso sería minimizar las distancias o maximizar las similitudes, como se muestra a continuación:

$*$ En caso de utilizar distancias, tenemos:

$$\begin{aligned} &d(C_i, C_j) = Min_{(i_1, j_1)=1,2,\ldots,(n-k)|i_1 \neq j_1}\{d(C_i, C_j)\} = \\ &Min_{(i_1, j_1)=1,2,\ldots,(n-k)|i_1 \neq j_1}\{Min_{x_l \in C_{i1}} \text{ y } x_m \in C_{j1} \text{ y } d(x_l, x_m)\} \\ &\text{para} \quad l = 1, 2, \ldots, n_{i1} \quad \text{y} \quad m = 1, 2, \ldots, n_{j1}. \end{aligned}$$

$*$ En caso de utilizar similaridades, tenemos:

$$s(C_i, C_j) = Max_{(i_1,j_1)=1,2,\ldots,(n-k)|i_1 \neq j_1}\{s(C_i, C_j)\} =$$
$$Max_{(i_1,j_1)=1,2,\ldots,(n-k)|i_1 \neq j_1}\{Max_{x_l \in C_{i1}} \text{ y } _{x_m \in C_{j1}} \text{ y } s(x_l, x_m)\}$$
$$\text{para} \quad l = 1, 2, \ldots, n_{i1} \quad \text{y} \quad m = 1, 2, \ldots, n_{j1}.$$

Método del vecino más lejano o enlace completo

Según King, 1967 [23] este método considera la mayor distancia entre el par de observaciones más lejano en dos conglomerados. Si la medida es la distancia, se toma la máxima distancia de los individuos del grupo al nuevo individuo. Si se toma la similitud o semejanza entre el grupo formado y el nuevo individuo, entonces se recoge el mínimo de los individuos del grupo al nuevo individuo.

Este método suele producir agrupaciones más estrechas que el método del vecino más próximo, pero estas agrupaciones estrechas pueden terminar muy juntas y también tiende a minimizar las distancias dentro de los conglomerados.

De manera que, si después de haber realizado la k-ésima etapa, tenemos formados $(n-k)$ conglomerados, la distancia entre el conglomerado C_i que posee n_i elementos, con el conglomerado C_j que posee n_j elementos, y dado que x_i pertenece al conjunto C_i y que x_j pertenece al conjunto C_j seria entonces:

$$d(C_i, C_j) = Max\{d(x_l, x_m)\} \quad \text{para} \quad l = 1, 2, \ldots, n_i \quad \text{y} \quad m = 1, 2, \ldots, n_j.$$

Al emplear medida de ese tipo en base al tipo de variables de la matriz de datos y tuviésemos similaridades, en ese caso tendríamos el siguiente:

$$s(C_i, C_j) = Min\{s(x_l, x_m)\} \quad \text{para} \quad l = 1, 2, \ldots, n_i \quad \text{y} \quad m = 1, 2, \ldots, n_j.$$

Así que, el proceso de unir las dos agrupaciones en el proceso sería minimizar las distancias o maximizar las similitudes, como se muestra a continuación:

∗ En caso de utilizar distancias, tenemos:

$$d(C_i, C_j) = Min_{(i_1,j_1)=1,2,\ldots,(n-k)|i_1\neq j_1}\{d(C_i, C_j)\} =$$
$$Min_{(i_1,j_1)=1,2,\ldots,(n-k)|i_1\neq j_1}\{Min_{x_l\in C_{i1}} \text{ y } _{x_m\in C_{j1}} \text{ y } d(x_l, x_m)\}$$
$$\text{con} \quad l = 1, 2, \ldots, n_{i1} \quad \text{y} \quad m = 1, 2, \ldots, n_{j1}.$$

∗ En caso de utilizar similaridades, tenemos:

$$s(C_i, C_j) = Max_{(i_1,j_1)=1,2,\ldots,(n-k)|i_1\neq j_1}\{s(C_i, C_j)\} =$$
$$Max_{(i_1,j_1)=1,2,\ldots,(n-k)|i_1\neq j_1}\{Max_{x_l\in C_{i1}} \text{ y } _{x_m\in C_{j1}} \text{ y } s(x_l, x_m)\}$$
$$\text{con} \quad l = 1, 2, \ldots, n_{i1} \quad \text{y} \quad m = 1, 2, \ldots, n_{j1}.$$

Método de agrupación de vinculación promedio o vinculación inter-grupo

El método de agrupación de vinculación promedio es aquel donde la distancia entre cada par de observaciones en cada grupo se suma y se divide por el número de pares para obtener una distancia promedio entre grupos (Ward, 1963 [24]; Murtgah, 1984 [25]).

Consideramos el conglomerado C_i que tiene n_i elementos y está compuesto por dos conglomerados, C_{i1} y C_{i2} con n_{i1} y n_{i2} elementos respectivamente cada uno, y el conglomerado C_j que posee n_j elementos, de manera que la distancia o similitud será la siguiente,

$$d(C_i, C_j) = \frac{d(C_{i1}, C_j) + d(C_{i2}, C_j)}{2}.$$

Método de la mediana

Se define la distancia entre 2 conglomerados como la distancia entre las medianas de cada conglomerado, donde la mediana es localizada en el valor mediano de

cada variable sobre todos los miembros del conglomerado.

Si el tamaño de los conglomerados es muy diferente, entonces el centroide del nuevo conglomerado estará situado muy cerca del más grande, e incluso podría estar dentro de él, por esta motivo Gower (1985) [26] sugiere una estrategia alternativa, llamada "método de la mediana", y puede ser adecuado tanto para las medidas de distancia como para las de similitud (Mucha y Sofyan, 2000 [27]), siendo este método es similar al del centroide.

En este método no se tiene en cuenta el tamaño de los conglomerados a la hora de efectuar los cálculos, ya que si los tamaños de los dos conglomerados n_{i1} y n_{i2} del conglomerado C_i son muy diferentes entre sí, puede ocurrir que el centroide de dicho conglomerado m^i esté excesivamente influenciado por el componente mayor y, luego, no se tenga en cuenta las cualidades del menor.

La estrategia a seguir en este método para calcular la distancia mediana, considerando que $n_{i1} = n_{i2}$, hace que el centroide del conglomerado C_i se sitúe entre los conglomerados C_{i1} e C_{i2}, luego el centroide del conglomerado en estudio (C_i, C_j) se sitúa en el punto central o mediana del triángulo formado por los conglomerados $C_{i1}, C_{i2}C_j$ (Gutiérrez et al, 1994 [30]).

$*$ En caso de utilizar distancias, tenemos:

$$d(C_i, C_j) = \frac{1}{2}[d(C_{i1}, C_j) + d(C_{i2}, C_j)] - \frac{1}{4}d(C_{i1}, C_{i2})$$

$*$ En caso de utilizar similaridades, tenemos:

$$s(C_i, C_j) = \frac{1}{2}[s(C_{i1}, C_j) + s(C_{i2}, C_j)] + \frac{1}{4}[1 - s(C_{i1}, C_{i2})]$$

Método del Centroide

El método del centroide calcula la distancia entre los centroides de dos grupos. A medida que los centroides se mueven con nuevas observaciones, es posible que los grupos más pequeños sean más similares al nuevo grupo más grande que a sus conglomerados individuales, lo que provoca una inversión en el dendrograma (Silva Camêlo et al., 2020 [28]). Este problema no surge en los otros métodos de vinculación porque los conglomerados que se fusionan siempre serán más similares a ellos mismos que al nuevo conglomerado más grande. En consecuencia, la distancia entre los conglomerados se puede reducir entre distintos pasos y, por tanto, puede hacer que existan problemas en los resultados de análisis (Wolfson et al., 2004 [29]), por eso la semejanza entre dos conglomerados viene dada por la semejanza entre sus centroides.

Al efectuar los cálculos en este método, se tiene en cuenta el tamaño de los conglomerados. Entonces, suponiendo que pretendemos medir la distancia entre el conglomerado C_i que tiene n_i elementos y está compuesto por dos conglomerados, a saber C_{i1} y C_{i2} con n_{i1} y n_{i2} elementos respectivamente cada uno, y el conglomerado C_j que tiene n_j elementos, sabiendo que los centroides de los conglomerados anteriormente descritos son m^j, m^{i1} y m^{i2} y siendo vectores n dimensionales (Gallardo, 1994 [30]).

Las componentes del centroide del conglomerado C_i vienen dada por:

$$m_l^{i1} = \frac{n_{i1}m_l^{i1} + n_{i2}m_l^{i2}}{n_{i1} + n_{i2}} \quad \text{con} \quad l = 1, 2, \ldots, n.$$

Luego la distancia euclídea al cuadrado entre los dos conglomerados se unen los de menos distancia es obtenida por

$$d^2(C_i, C_j) = \sum_{l=1}^{n} (m_l^j - m_l^i)^2$$
$$d^2(C_i, C_j) = \frac{n_{i1}}{n_{i1}+n_{i2}} d^2(C_i, C_j) + \frac{n_{i2}}{n_{i1}+n_{i2}} - \frac{n_{i1}}{(n_{i1}+n_{i2})^2} d^2(C_i, C_j).$$

Método de Ward o varianza mínima

El método de Ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos conglomerados los cuales tengan el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada conglomerado, de cada individuo al centroide del conglomerado. El objetivo del método de Ward es encontrar en cada etapa aquellos dos conglomerados cuya unión proporcione el menor incremento en la suma total de errores, o sea, que los nuevos conglomerados formados sean más homogéneos (Mucha y Sofyan, 2000 [27]; Gallardo, 2011 [31]).

El fundamento del método de Ward es relativamente distinto al de los presentados hasta ahora y se utiliza con algunas variantes a la hora de llevar a cabo su implementación informática (Encinas, 2001 [9]). Para este método se considera la distancia euclídea al cuadrado como medida de disimilitud. Llamando

$$d(x_i, x_j)^2 = ||x_i - x_j||^2$$

a dicha distancia entre los puntos x_1 y x_j, la varianza total (o inercia) del conjunto de puntos es la cantidad dada por la siguiente expression:

$$I = \sum_i m_i ||x_i - G||^2,$$

siendo G el centro de gravedad de los puntos dados, con masas respectivas m_i.

Existiendo una partición del conjunto de individuos en q conglomerados, el q-ésimo conglomerado tiene como centro de gravedad a G_q y masa m_q. Entonces la

inercia se puede descomponer como la suma de la varianza que existe dentro de los conglomerados y la que hay entre unos conglomerados y otros, del siguiente modo:

$$I = \sum_q m_q ||G_q - G||^2 + \sum_q \sum_{i \in q} m_i ||x_i - G_q||^2.$$

Si x_i y x_j son dos elementos de masas m_i y m_j, respectivamente, que se unen en un elemento x de masa $m = m_i + m_j$, con

$$x + (m_i x_i + m_j x_j)/(m_i + m_j),$$

entonces, podemos descomponer la varianza I_{ij} de x_i y x_j con respecto a G a través de la ecuación:

$$I_{ij} = m_i ||x_i - x||^2 + m_j ||x_j - x||^2 + m||x - G||^2.$$

El último término es el único que permanece constante si se cambian x_i y x_j por su centro de gravedad x. La reducción en la varianza es dada por:

$$\Delta I_{ij} = m_i ||x_i - x||^2 + m_j ||x_j - x||^2.$$

Sustituyendo x por su valor como función de x_i y x_j, tenemos:

$$\Delta I_{ij} = \left(\frac{m_i m_j}{m_i + m_j} \right) ||x_i - x||^2 = \left(\frac{m_i m_j}{m_i + m_j} \right) d(x_j - x_j)^2.$$

De este modo, la estrategia que se sigue para hacer conglomerados con este método consiste en encontrar los individuos x_i y x_j con la condición de hagan mínima ΔI_{ij}, en lugar de ser los individuos más cercanos. Por tanto, puede considerarse a ΔI_{ij} como un nuevo índice de disimilitud. Por medio de esta estrategia, los individuos con menor peso son los que más pronto se unen. El cuadrado de

la distancia de un punto z a un centro de conglomerado x, se puede escribir en función de las distancias a los puntos x_i y x_j:

$$d(x-z)^2 = \left(\frac{1}{m_i+m_j}\right)\left(m_i d(x_i,z)^2 + m_j(x_j,z)^2) - \left(\frac{m_i m_j}{m_i+m_j}\right) d(x_i,x_j)^2\right),$$

por medio de la descomposición de la varianza de (x_i y x_j) esta fórmula se establece con respecto a z, en la varianza con respecto a x y en la varianza de x con respecto a z:

$$m_i||x_i-z||^2 + m_j||x_i-z||^2 = (m_+m_j)||x_i-z||^2 + \left(\frac{m_i m_j}{m_i+m_j}\right)||x_i-z||^2.$$

2.2. Conglomerado jerárquico restringido

Por lo general, el investigador tiene dificultades en agrupar un conjunto de n objetos en k grupos disjuntos. Pronto, muchos métodos propusieron encontrar la mejor partición según un criterio de homogeneidad basado en diferencias, o para un modelo de mezcla de funciones de distribución multivariante. El tipo más común son las restricciones de contigüidad.

Dichas restricciones ocurren cuando se requiere que los objetos en un grupo no solo sean similares entre sí, sino que también comprendan un conjunto contiguo de objetos (municipio), es decir, la contigüidad entre cada par de objetos viene dada por una matriz $C = (c_{ij})_{n\times n}$, donde $c_{ij} = 1$ si los objetos i-ésima y j-ésima son contiguos, y 0 si no lo son.

Entonces, un grupo C se considera contiguo si hay una ruta entre cada par de objetos en C, es decir, el subgrafo está conectado (Chavent et al., 2017b [32]). Se han modificado varios algoritmos de agrupamiento clásico para tener en cuenta este tipo de restricción, según destaca Murtagh, 1985a [33], Legendre, 2011 [34] y Bécue-

Bertaut et al., 2014 [35]. Por ejemplo, el procedimiento jerárquico estándar basado en la fórmula de Lance y Williams (1967) [36] se puede restringir fusionando solo conglomerados contiguos en cada etapa. Pero, ¿qué define a los conglomerados "contiguos"? Por lo general, dos conglomerados se consideran contiguos si hay dos objetos, uno de cada conglomerado, que están vinculados en la matriz de contigüidad.

Una matriz de adyacencia es una de las formas de representar un gráfico, se trata de una matriz cuadrada de n filas $\times n$ columnas, siendo n el número de vértices del grafo, donde cada elemento a_{ij} vale 1 cuando haya una arista (relación) que una los vértices i y j, caso contrario el elemento a_{ij} vale 0.

Las propiedades de la matriz de adyacencia son:

- Para un grafo no dirigido la matriz de adyacencia es simétrica a lo largo de la diagonal principal, es decir, la entrada a_{ij} es igual a la entrada a_{ji};
- El número de caminos $C_{i,j}(k)$, atravesando k aristas desde el vértice i al vértice j, viene dado por un elemento de la potencia k-ésima de la matriz de adyacencia:

$$C_{i,j}(k) = [A^k]_{ij}$$

En los gráficos no dirigidos, las matrices de adyacencia son simétricas a lo largo de la diagonal principal, es decir, la entrada a_{ij} es igual a la entrada a_{ji}.

Esta matriz de adyacencia es utilizada para encontrar una conexión entre los límites de cada elementos. En consecuencia, dos grupos se consideran contiguos si hay dos elementos, uno de cada grupo, que está vinculado en la matriz de contigüidad. Varios autores, en diferentes áreas del conocimiento, han implementado procedimientos de conglomerado restringido: Ambroise et al., 1997 [37]; Ambroise

y Govaert, 1998; [38]; Legendre, 2011 [34]; Duque et al., 2011 [39]; Dehman, 2015 [40]; Bécue-Bertaut et al., 2017 [35]; Aguiar, et al. 2020a, 2020b [41, 42].

Miele et al. 2014 [47], por ejemplo, propuso un método restringido espacialmente basado en modelos que integran la información geográfica dentro de un marco de regularización del algoritmo de maximización esperada (EM) agregando algunas restricciones a la estimación de máxima verosimilitud de los parámetros. Es un método de partición con restricciones de vecindad, mientras que el método tipo Ward-like formulado por Chavent et al., 2017a [43] es un método de conglomerado jerárquico (y no de partición), que incluye restricciones espaciales/geográficas (no necesariamente restricciones de vecindad) (Chavent et al., 2017b [32]).

En esta tesis, estudiaremos el método de conglomerado jerárquico Ward-like (y no de partición) que incluye restricciones espaciales (no necesariamente restricciones de vecindad) que utiliza dos matrices de disimilitud D_0 (calculada a partir de las variables socioepidemiológicas) y D_1 (calculada a partir de las distancias geográficas) y un parámetro de mezcla $\alpha \in [0, 1]$.

Capítulo 3

Conglomerado jerárquico Ward-like

Con un algoritmo similar a Ward, el método Ward-like es un algoritmo de conglomerado jerárquico restringido que optimiza la combinación convexa $D_\alpha = (1-\alpha)D_0 + \alpha D_1$ de este criterio calculado con dos matrices de disimilitud, D_0 y D_1 y un parámetro de mezcla $\alpha \in [0;1]$.

La primera matriz de disimilitud $D_0 = [d_{0,ij}]$ se construye a partir de la matriz de distancias de Manhattan entre los 223 municipios realizada con p = 5 variables socioepidemiológicas, es decir, la matriz de las diferencias en el espacio de características, y la matriz de disimilitud $D_1 = [d_{1,ij}]$ que se construye a partir de la distancia geográfica entre los 223 municipios, es decir, la matriz D_1 de las diferencias en la restricción-espacio.

El criterio minimizado en cada etapa es una combinación convexa del criterio de homogeneidad calculado con D_0 y el criterio de homogeneidad calculado con D_1.

El parámetro α (el peso de esta combinación convexa) concede la importancia relativa de D_0 en comparación con D_1. Este parámetro controla el peso de la restricción sobre la calidad de las soluciones, es decir, para un valor dado de $\alpha[0;1]$, el

parámetro de mezcla α controla claramente la parte de pseudo-inercia debida a D_0 y D_1. Teniendo en cuenta que cuando α aumenta, la homogeneidad calculada con D_0 disminuye mientras que la homogeneidad calculada con D_1 aumenta.

La propuesta es determinar un valor de alfa que aumente la contigüidad espacial sin deteriorar demasiado la calidad de la solución sobre las variables de interés. Por lo tanto, la intención es determinar un valor de α que aumente la homogeneidad espacial y geográfica sin deteriorar demasiado la calidad de la solución sobre las variables de interés.

3.1. Conglomerado jerárquico Ward-Like con disimilitudes y pesos no uniformes

Consideremos un conjunto de n observaciones, siendo w_i el peso de la i-ésima observación para $i = 1, \ldots, n$ y $D = [d_{ij}]$ una matriz de disimilitud $n \times n$ asociada con las n observaciones, donde d_{ij} es la medida de disimilitud entre las observaciones i y j.

Recordemos que cuando la matriz de disimilitud D no es una matriz de distancias euclidias, el criterio de inercia habitual (también denominado criterio de varianza) utilizado en el enfoque de conglomerado jerárquico de Ward (1963) [24] no tiene sentido y el algoritmo de Ward implementado con la fórmula de Lance y Williams (1967) [36] tiene que ser reinterpretado (Chavent, 2017b) [32]. Sin embargo, el método de Ward ya se ha generalizado para su uso con distancias no euclidianas, para la distancia $l1$ o distancias de Manhattan, conforme vemos en Strauss y von Maltitz (2017) [14].

3.1.1. Método Ward-like

Pseudo-inercia - Para entender el funcionamento del método de Ward-like és importante comprender la pseudo-inercia. Entonces, consideremos una partición $P_K = (C_1, \ldots, C_K)$ en K grupos. La pseudo inercia de un grupo C_k generaliza la inercia al caso de datos de disimilitud (euclidianos o no) de la siguiente manera:

$$I_\alpha(C_k) = \sum_{i \in C_k} \sum_{j \in C_k} \frac{w_i w_j}{2\mu_k} d_{ij}^2 \tag{3.1}$$

en que $\mu_k = \sum_{i \in C_k} w_i$ es el peso de C_k. Cuanto menor es la pseudo-inercia $I(C_k)$, más homogéneas son las observaciones que pertenecen al grupo C_k.

La pseudo-inercia dentro del clúster de la partición P_K es definida por:

$$W(P_K) = \sum_{k=1}^{K} I(C_k).$$

Cuanto más pequeña es esta pseudo-inercia $W(P_K)$, más homogénea es la partición en los K grupos.

Espíritu del conglomerado jerárquico Ward. Para obtener una nueva partición P_K en K grupos de una partición determinada $P_K + 1$ en $K + 1$ grupos, la idea es agregar los dos grupos A y B de $P_K + 1$ de manera que la nueva partición tenga una inercia mínima dentro del grupo (heterogeneidad, varianza), es decir:

$$arg \min_{A,B \in P_{k+1}} W(P_K), \tag{3.2}$$

donde $P_K = P_{K+1} \setminus \{A, B\} \cup \{A \cup B\}$ y $W(P_K) = W(P_K + 1) - I(A) - I(B) +$

$I(A \cup B)$.

Dado que $W(P_{K+1})$ está fijo para una partición determinada P_{K+1}, el problema de optimización de la Equación (3.2) es equivalente a:

$$\min_{A,B \in P_{k+1}} I(A \cup B) - I(A) - I(B). \tag{3.3}$$

Por tanto, el problema de optimización se logra definiendo

$$\delta(A, B) = I(A \cup B) - I(A) - I(B)$$

como la medida de agregación entre dos conglomerados que se minimiza en cada paso del algoritmo de conglomerado jerárquico. Conviene observar que $\delta(A, B) = W(P_K) - W(P_K + 1)$ puede verse como el aumento de la inercia dentro del grupo, o sea, pérdida de homogeneidad.

Processo conglomerado jerárquico Ward-like para disimilitudes no euclidianas De acuerdo con Chavent et. al, (2017b) [32] en el caso de datos de disimilitud el procedimiento de conglomerado jerárquico Ward-like es interpretado de la siguiente manera:

Paso $K = n$: Inicialización.

La partición inicial P_n en n grupos, es decir, cada grupo solo contiene una única observación.

Paso $K = n - 1, \ldots, 2$: obteniendo la partición en K grupos de la partición en $K + 1$ grupos.

En cada paso K, el algoritmo agrega los dos grupos A y B de $P_K + 1$ de acuerdo con el problema de optimización de la Ecuación (3.3) de manera que el aumento

de la pseudo-inercia dentro del grupo es mínimo para la partición seleccionada sobre las otras en K grupos.

> Paso $K = 1$: detener. Se obtiene la partición P_1 en un grupo (que contiene las n observaciones).

El conjunto jerárquicamente anidado de dichas particiones $\{P_n, \ldots, P_K, \ldots, P_1\}$ está representado gráficamente por un árbol (también llamado dendrograma) donde la altura de un grupo $C = A \cup B$ es $h(C) := \delta(A, B)$.

Debido a la ecuación de Lance y Williams (1967) [36], las medidas de agregación entre el nuevo conglomerado $A \cup B$ y cualquier conglomerado D de $P_K + 1$ se calculan en cada etapa:

$$\begin{aligned} \delta(A \cup B, D) &= \frac{\mu_A + \mu_D}{\mu_A + \mu_B + \mu_D}\delta(A, D) + \frac{\mu_B + \mu_D}{\mu_A + \mu_B + \mu_D}\delta(B, D) \\ &= -\frac{\mu_D}{\mu_A + \mu_B + \mu_D}\delta(A, B). \end{aligned} \tag{3.4}$$

En el primer paso, la partición es P_n y las medidas de agregación entre los únicos se calculan con

$$\delta_{ij} := \delta(\{i\}, \{j\}) = \frac{w_j}{w_i + w_j} d_{ij}^2,$$

y almacenado en la matriz $n \times n$, $\Delta = [\delta_{ij}]$. Para cada paso K posterior, se utiliza la fórmula de la Equación (3.4) de Lance e Williams para construir la correspondiente matriz de agregación $K \times K$.

Conglomerado jerárquico Ward-like cuando las disimilitudes son distancias euclidianas. Cuando las diferencias son distancias euclidianas calculadas a partir de una matriz de datos numéricos X de dimensión $n \times p$, por ejemplo, la pseudo-

inercia de un conglomerado C_k definido en la Ecuación (3.1) es ahora igual a la inercia de las observaciones en C_k:

$$I(C_k) = \sum_{i \in C_k} w_i d^2(x_i, g_k)$$

donde $xi \in \Re^p$ es la i-ésima fila de X asociada con la i-ésima observación, y $g_k = \frac{1}{\mu_k} \sum_{i \in C_k} w_i x_i \in \Re^p$ es el centro de gravedad de C_k. Por tanto, la medida de agregación $\delta(A, B)$ entre dos conglomerados se escribe como:

$$\delta(A, B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} d^2(g_A, g_B).$$

3.2. Conglomerado jerárquico Ward-Like con dos matrices de disimilitudes

Consideremos un conjunto de n observaciones, sea w_i el peso de la i-ésima observación para $i = 1, \ldots, n$ y tomando como entrada dos matrices de disimilitud $D_0 = [d_{0,ij}]$ y $D_1 = [d_{1,ij}]$ $n \times n$. Suponiendo que las n observaciones son municipios, D_0 puede basarse en una matriz de datos numéricos de p_0 variables cuantitativas medidas en las n observaciones y D_1 puede ser una matriz que contiene las distancias geográficas entre las n observaciones.

3.2.1. Algoritmo de conglomerado jerárquico con dos matrices de disimilitud

Para un cierto valor de $\alpha \in [0, 1]$, el algoritmo funciona de la siguiente manera. Hay que tener en cuenta que la partición en K conglomerados se indexará a partir de ahora por α de la siguiente manera: P_α^K.

Definición 3.2.1 *La pseudo-inercia mixta del grupo P_α^K (de aquí en adelante llamada inercia mixta) se define como:*

$$I_\alpha(C_k^\alpha) = (1-\alpha)\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{0,ij}^2 + \alpha\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{1,ij}^2, \tag{3.5}$$

siendo $\mu_k^\alpha = \sum_{i\in C_k^\alpha} w_i$ el peso de $C - k^\alpha$ y $d_{0,ij}$, respectivamente $d_{1,ij}$, es la disimilitud normalizada entre las observaciones i y j en D_0 y también en D_1.

La pseudo inercia mixta dentro del conglomerado (de aquí en adelante llamada inercia mixta dentro del conglomerado) de una partición $P_K^\alpha = (C_1^\alpha, \ldots, C_K^\alpha)$ es la suma de la inercia mixta de sus conglomerados:

$$W_\alpha(P_K^\alpha) = \sum_{k=1}^{K} I_\alpha(C_k^\alpha). \tag{3.6}$$

Espíritu del conglomerado jerárquico Ward-like. Como se ha visto anteriormente, para obtener una nueva partición P_K^α en K conglomerados de una partición dada $P_K^\alpha + 1$ en $K+1$ grupos, la idea es agregar los dos conglomerados A y B de P_{K+1} de manera que la nueva partición tenga un mínimo de mezcla inercia dentro del conglomerado. El problema de optimización ahora se puede expresar de la siguiente manera:

$$\arg \min_{A,B\in P_{K+1}^\alpha} I_\alpha(A\cup B) - I_\alpha(A) - I_\alpha(B). \tag{3.7}$$

Proceso de agrupamiento jerárquico Ward-like.

Paso $K = n$: Inicialización.

Las diferencias se pueden reescalar entre 0 y 1 para obtener el mismo orden de magnitud, de este modo,

$$D_0 \leftarrow \frac{D_0}{\text{máx}(D_0)} \quad \text{y} \quad D_1 \leftarrow \frac{D_1}{\text{máx}(D_1)}.$$

La partición inicial $P_n^\alpha =: Pn$ en n grupos (es decir, cada grupo solo contiene

una observación) es única y, por lo tanto, no depende de α.

Paso $K = n-1, \ldots, 2$: obtención de la partición en K conglomerados a partir de la partición en $K+1$ grupos.

En cada paso K, el algoritmo agrega los dos grupos A y B de $P_\alpha K+1$ de acuerdo con el problema de optimización de la Equación 3.7 de manera que el aumento de la inercia mixta dentro del conglomerado sea mínimo para la partición seleccionada sobre las otras en K conglomerados.

Más precisamente, en el paso K, el algoritmo agrega los dos conglomerados A y B de manera que la medida de agregación correspondiente sea mínima

$$\delta_\alpha(A,B) := W_\alpha(P^\alpha_{K+1}) - W_\alpha(P^\alpha_K) = I_\alpha(A \cup B) - I_\alpha(A) - I_\alpha(B)$$

Paso $K = 1$: detener. Se obtiene la partición $P^\alpha_1 =: P_1$ en un conglomerado. Hay que tener en cuenta que esta partición es única y, por lo tanto, no depende de α.

El valor (altura) de un conglomerado $A \cup B$ en el dendrograma de la jerarquía correspondiente, viene dado por el valor de criterio del conglomerado aglomerativo $\delta_\alpha(A,B)$.

La ecuación de Lance y Williams (3.4) en la práctica, sigue siendo cierta en este contexto (donde δ debe reemplazarse por δ_α). La medida de agregación entre dos *singletons* (conjunto que contiene solo un elemento) se escribe ahora:

$$\delta_\alpha(\{i\},\{j\}) = (1-\alpha)\frac{w_i w_j}{w_i + w_j} d^2_{0,ij} + \frac{w_i w_j}{w_i + w_j} d^2_{1,ij}.$$

Posteriormente se aplica la ecuación de Lance y Williams a la matriz

$$\Delta_\alpha = (1-\alpha)\Delta_0 + \alpha\Delta_1,$$

dónde Δ_0 y Δ_1 es la matriz $n \times n$ de los valores $\delta_{0,ij} = \frac{w_i w_j}{w_i+w_j} d^2_{0,ij}$ y $\delta_{1,ij} = \frac{w_i w_j}{w_i+w_j} d^2_{1,ij}$, respectivamente.

Es importante destacar algunas observaciones al respecto del procedimiento propuesto del algoritmo Ward-like. Este es diferente de aplicar directamente el algoritmo de Ward a la matriz de "disimilitud" obtenida mediante la combinación convexa $D_\alpha = (1-\alpha)D_0 + \alpha D_1$. La principal ventaja del algoritmo Ward-like, es que el parámetro de mezcla α controla claramente la parte de pseudo-inercia debida a D_0 y D_1 en (3.5). Este no es el caso cuando se aplica directamente el algoritmo de Ward a D_α, ya que se basa en una sola pseudo-inercia.

Cuando $\alpha = 0$, el conglomerado jerárquico se basa solo en la matriz de disimilitud D_0, del mismo modo, cuando $\alpha = 1$ el conglomerado jerárquico se basa solo en la matriz de disimilitud D_1. A continuación se propone un procedimiento para determinar un valor adecuado para el parámetro de mezcla α.

3.3. Un procedimiento para determinar un valor adecuado para el parámetro de mezcla α

El parámetro de mezcla α establece la importancia de D_0 y D_1 en el proceso de conglomerado. Un punto esencial es la elección de un valor adecuado para el parámetro de mezcla $\alpha \in [0,1]$. Una solución práctica es condicionar K y elegir un valor α que mejor se comprometa entre la pérdida de la homogeneidad socio-epidemiologica y la pérdida de la homogeneidad geográfica o mismo determinar un valor de α que aumente la homogeneidad geográfica de una partición en K conglomerados sin afectar negativamente la homogeneidad socioeconómica.

Como idea de esta ideal del procedimiento propuesto en la determinación de un valor adecuado para el parámetro de mezcla α, Chavent et al. (2017 [43]) propusieron que la matriz de disimilitud D_1 contuviera distancias geográficas entre n municipios, mientras que la matriz de disimilitud D_0 fueran las distancias basadas en una matriz de datos X_0 de $n \times p_0$ de p_0 variables socioeconomica medidas en estos n municipios. En esta Tesis, utilizaremos una matriz de disimilaridad D_0 que contiene distancias basadas en una matriz de datos de variables socioepidemiologica. Estas homogeneidades pueden medirse usando las apropiadas pseudo inercias dentro del conglomerado.

Sea $\beta \in [0, 1]$, la noción de proporción del total mixto (pseudo) inercia explicada por la partición P_K^α en K conglomerados viene dado por:

$$Q_\beta(P_k^\alpha) = 1 - \frac{W_\beta(P_k^\alpha)}{W_\beta(P_1)} \in [0, 1].$$

- Cuando $\beta = 0$, la situación es:
  - El denominador $W_0(P_1)$ es la (pseudo) inercia total basada en la matriz de disimilitud D_0 y el numerador es la (pseudo) inercia dentro del conglomerado $W_0(P_K^\alpha)$ basada en la matriz de disimilitud D_0, es decir, solo desde el punto de vista socioepidemiológico en nuestro estudio, Capítulo 7.
  - Luego, cuanto mayor sea el valor del criterio $Q_0(P_K^\alpha)$, más homogénea será la partición P_K^α desde el punto de vista socioepidemiológico (esto es, cada conglomerado $C_k^\alpha, k = 1, \ldots,$ tiene una inercia baja $I_0(C_k^\alpha)$, lo que significa que los individuos dentro del conglomerado son similares).
  - Pero cuando la partición considerada P_K^α se ha obtenido con $\alpha = 0$, el criterio $Q_0(P_K^\alpha)$ es obviamente máximo (ya que la partición P_K^0 se obtuvo usando solo la matriz de disimilitud D_0), y este criterio naturalmente

tenderá a disminuir a medida que α aumente de 0 a 1.

- Cuando $\beta = 1$, tenemos:
  - Del mismo modo, cuando $\beta = 1$, el denominador $W_1(P_1)$ es la (pseudo) inercia total basada en la matriz de disimilitud D_1 y el numerador es la (pseudo) inercia dentro del conglomerado $W_1(P_K^\alpha)$ basada en la matriz de disimilitud D_1, es decir, solo desde un punto de vista geográfico en nuestra aplicación, Capítulo 7.
  - Por esta razón, cuanto mayor sea el valor del criterio $Q_1(P_K^\alpha)$, más homogénea será la partición P_K^α desde el punto de vista geográfico.
  - Aunque cuando la partición considerada P_K^α se ha obtenido con $\alpha = 1$, el criterio $Q_1(P_K^\alpha)$ es obviamente máximo (esto porque la partición P_K^1 se obtuvo utilizando solo la matriz de disimilitud D_1), y este criterio naturalmente tenderá a disminuir a medida que α disminuya de 1 a 0.
- Cuando $\beta \in]0, 1[$, tenemos:
  - Para un determinado valor de $\beta \in]0, 1[$, el denominador $W_\beta(P_1)$ es una (pseudo) inercia mixta total que no se puede interpretar fácilmente en la práctica, y el numerador $W_\beta(P_K^\alpha)$ es la inercia mixta (pseudo) dentro del conglomerado. Aunque, es importante tener en consideración que cuando la partición considerada P_K^α se ha obtenido con $\alpha = \beta$, el criterio $Q_\beta(P_K^\alpha)$ es obviamente máximo por construcción, y tenderá a disminuir a medida que α se aleja de β.
  - Por último, observe que este criterio $Q_\beta(P_K^\alpha)$ es decreciente en K. Además, $\forall \beta \in [0, 1]$, inclusive es fácil ver que $Q_\beta(P_n) = 1$ y $Q_\beta(P_1) = 0$. Cuantos más conglomerados haya en una partición, más homogéneos son estos conglomerados (esto es, con una inercia baja). Así pues, este criterio no puede utilizarse para seleccionar un número adecuado K de conglomerados.

3.4. Selección del parámetro de mezcla α

Con el interés en determinar un valor de α que aumente la homogeneidad geográfica de una partición en K conglomerados sin deteriorar demasiado la homogeneidad socioepidemiologica, podemos utilizar este criterio para seleccionar el parámetro de mezcla alfa determinando un número de K conglomerados, la idea es la siguiente.

Consideremos una cuadrícula dada de valores J para $\alpha \in [0, 1]$:

$$G = \alpha_1 = 0, \alpha_1, \ldots, \alpha_J = 1.$$

Dado que para cada valor $\alpha_j \in G$, a una partición correspondiente a $P_K^{\alpha_j}$ en K conglomerados se obtiene utilizando el algoritmo propuesto de conglomerado hierárquico de Ward-like.

Para las J particiones $P_K^{\alpha_j}, j = 1, \ldots, J$, se evalúa el criterio $Q_0(P_K^{\alpha_j})$. La gráfica de los puntos $(\alpha_j, Q_0(P_K^{\alpha_j})), j = 1, \ldots, J$ proporciona una forma visual de observar la pérdida de homogeneidad socioeconómica de la partición $P_K^{\alpha_j}$ (de la partición socioepidemiologica "pura" P_K^0) a medida que α_j aumenta de 0 a 1.

Del mismo modo, para las J particiones $P_K^{\alpha_j}, j = 1, \ldots, J$, se evalúa el criterio $Q_1(P_K^{\alpha_j})$. La gráfica de los puntos $(\alpha_j, Q_1(P_K^{\alpha_j})), j = 1, \ldots, J$ proporciona una forma visual de observar la pérdida de homogeneidad geográfica de la partición $P_K^{\alpha_j}$ (de la partición geográfica "pura" P_K^1) cuando α_j disminuye de 1 a 0.

Estos dos gráficos (superpuestos en la misma figura) permiten al usuario elegir un valor adecuado para $\alpha \in G$, que es una compensación entre la pérdida de homogeneidad socioepidemiologica y una mayor cohesión geográfica (cuando se ve a través de valores crecientes de α).

Otra situación es cuando las dos (pseudo) inercias totales $W_0(P_1)$ y $W_1(P_1)$ utilizadas en $Q_0(P_K^\alpha)$ y $Q_1(P_K^\alpha)$ son muy diferentes.

Para comprenderlo mejor, consideremos, por ejemplo, que la matriz de disimilitud D_1 es una matriz de disimilitud de "vecindad", construida a partir de la correspondiente matriz de adyacencia A: es decir, $D_1 = 1_n - A$ con $1_{n,ij=1} \forall (i,j), a_{ij}$ igual a 1 si las observaciones i y j son vecinas y 0 en caso contrario, y $a_{ii} = 1$ por convención. Con este tipo de matriz de disimilitud local D_1, la cohesión geográfica para algunos conglomerados suele ser pequeña: de hecho, $W_1(P_1)$ podría ser muy pequeño y, por lo tanto, el criterio $Q_1(P_K^\alpha)$ toma valores generalmente mucho más pequeños que los obtenidos por $Q_0(P_K^\alpha)$.

Por consiguiente, no es tan sencillo seleccionar un valor adecuado para el parámetro de mezcla α ya que los dos gráficos están en dos escalas muy diferentes, pero una forma de remediar este problema es considerar una renormalización de las dos parcelas.

En lugar de razonar en términos de valores absolutos del criterio $Q_0(P_K^\alpha)$ y también $Q_1(P_K^\alpha))$ que es máximo en $\alpha = 0$ (respectivamente, $A = 1$), renormalizaremos $Q_0(P_K^\alpha)$ y $Q_1(P_K^\alpha)$ de la siguiente manera:

$$Q_0^*(P_K^\alpha) = \frac{Q_0(P_K^\alpha)}{Q_0(P_K^0)} \quad \text{y} \quad Q_1^*(P_K^\alpha) = \frac{Q1(P_K^\alpha)}{Q1(P_K^1)}$$

luego razonamos en términos de proporciones de estos criterios. Por lo tanto, la gráfica correspondiente $(\alpha_j, Q_0^*(P_K^{\alpha_j})), j = 1, \ldots, J$ (resp. $(\alpha_j, Q_1^*(P_K^{\alpha_j})), j = 1, \ldots, J$) comienza desde el 100 % y disminuye a medida que α_j aumenta de 0 a 1 (respectivamente cuando α_j disminuye de 1 a 0).

3.5. Elección del número K conglomerados

El procedimiento propuesto para seleccionar un valor adecuado para el parámetro de mezcla α funciona para un número determinado de conglomerados K.

Primero es necesario seleccionar K y centrase en el dendrograma del conjunto jerárquicamente anidado de tales particiones $P_n^0 = P_n, \ldots, P_K^0, \ldots, , P_1^0 = P_1$ solo con base en la matriz de disimilitud D_0 (es decir, para $\alpha = 0$, esto es, considerando solo el punto de vista socioepidemiologico en nuestro estudio). Según el dendrograma, el usuario puede seleccionar un número apropiado K de conglomerados de acuerdo con su regla favorita.

Resulta interesante la interpretación de los conglomerados según las variables socioepidemiologicas. El diagrama de caja, por ejemplo, muestra las variables para cada grupo de la partición de K conglomerados. La gráfica se obtienen a partir de los paquetes R, `ggplot2` (Wickham, 2016 [63]), `tidyverse` (Wickham et al., 2019 [64]) y `dplyr` (Wickham et al., 2020 [62]).

3.5.1. Índice de desigualdad colectiva

Se trata de un indicador estadístico regional, el índice de desigualdad colectiva es una medida que se puede descomponer y se define como el peso w atribuido al cálculo de la matriz de disimilitud D.

Para el cálculo del índice de desigualdad colectiva, se utilizará el PIB de las 23 microrregiones del Estado de Paraíba (H), que toma valores $h_1, h_2, \ldots, h_{23}$ con frecuencias absolutas $n_i, n_2, \ldots, n_{23}$ sobre una población finita de tamaño $N = 23$. Según la característica propuesta por Zaiger (1983) [46], una medida de desigual-

dad descompuesta viene dada por:

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

Siendo $f_i = n_i/N$ la frecuencia relativa y $\Gamma_\beta(h)$ una función definida, para el valor de $\beta < 0$, cuya función será: $\Gamma_\beta(h) = h^\beta - 1$. González y Céspedes (2004) [45] establece que el índice de desigualdad colectiva (IDC) es:

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

Con $\beta = -1$, pues se trata de una desigualdad colectiva.

3.5.2. Coeficiente de diversificación

El coeficiente de diversificación trata de medir el grado en que el valor de una notificación de tuberculosis en una microrregión proviene de una variedad más o menos acusada de diferentes variables (nuevos casos de tuberculosis o recaída, por ejemplo), o si, por el contrario, proviene de un número relativamente bajo de variables.

Si una microrregión tiene un alto coeficiente de especialización, su aparición está más influida por una variable específica, en cuyo caso, la diversificación es mínima. En cambio, si una microrregión se clasifica como diversificada, significa que su situación epidemiológica de Tuberculosis no depende mucho de ninguna variable específica: todas están igualmente influidas por el conjunto de variables, en cuyo caso la diversificación es máxima.

El coeficiente de diversificación de la microrregión i se define como sigue

(González y Céspedes, 2004 [45]):

$$CD_i = \frac{(\sum_{j=1}^{L} Y_{ij})^2}{L \sum_{j=1}^{L} Y_{ij}^2} \tag{3.8}$$

Donde CD es la magnitud de las variables socioepidemiológicas, cuyos datos están en forma de matriz, donde Y_{ii} es el valor que toma la variable socioepidemiológica $j(j = 1, \ldots, 4)$ en la microrregión $i(i = 1, ..., 23)$. La CD es una cantidad entre $1/L$ y 1, $\frac{1}{L} \leq CD_i \leq 1$, siendo $1/L$ cuando la diversificación es mínima y 1 cuando es máxima. Para normalizar este coeficiente entre cero y uno se utiliza la siguiente fórmula: $CD_i^{=} L/(L - 1(D_i - 1/L))$ o, de forma equivalente: $CD_i^{=} \frac{LCD_i - 1}{L-1}$.

3.5.3. Razón de incidencia estandarizada

Una medida simple del riesgo de enfermedad es la tasa de incidencia estandarizada (RIE). Para cada área $i, i = 1, \ldots, n = 223$, el RIE se define como la relación entre los conteos observados y los esperados (Moraga, 2019 [55]).

$$RIE_i = \frac{Y_i}{E_i} \tag{3.9}$$

Los conteos esperados E_i representan el número total de casos de tuberculosis que se esperaría si la población del municipio i se comportara como la población del Estado de Paraíba.

E_i se puede calcular utilizando la estandarización indirecta como $E_i = \sum_{j=1}^{m} r_j^{(s)} n_j^{(i)}$, donde $r_j^{(s)}$ es la tasa (número de casos dividido por la población) en el estrato j en la población estándar, y $n_j^{(i)}$ es la población en el estrato j del área i.

El RIE puede ser calculado usando las funciones `group_by()` y `summary()` del paquete `dplyr` (Wickham, et al. 2019 [54]).

Capítulo 4

Paquete `ClustGeo`

Un paquete es una colección de funciones, datos y código R que se almacenan en una carpeta conforme a una estructura bien definida, fácilmente accesible para R, siendo R un lenguaje de programación funcional con un enfoque en el análisis estadístico.

Desarollado por Chavent et al. 2017 [43] el paquete `Clustgeo` implementa un algoritmo de conglomerado jerárquico Ward-like que incluye restricciones espaciales/geográficas. El funcionamento del paquete si da en versiones de R (>= 3.0.0). Las principales funciones se desarrollan a continuación.

4.1. Elección empírica del parámetro de mezcla: `choicealpha`

Esta función calcula la proporción y proporción normalizada, respectivamente, de la inercia explicada de las particiones en K conglomerados obtenidas con el procedimiento `hclustgeo` tipo Ward para un rango de parámetros de mezcla alfa. Cuando la proporción (resp. proporción normalizada) de la inercia explicada basada en D_0 disminuye, la proporción (resp. proporción normalizada) de la inercia explicada basada en D_1 aumenta. El gráfico de estos criterios puede ayudar al usuario a elegir el parámetro de mezcla alfa.

`choicealpha(D0, D1, range.alpha, K, wt = NULL, scale = TRUE, graph = TRUE)`

Argumentos	D_0	un objeto de clase `dist` con las disimilitudes entre las n observaciones. La función `as.dist` puede utilizarse para transformar un objeto de la matriz de clases en un objeto de la clase `dist`.
	D_1	bjeto de clase `dist` con otras diferencias entre las mismas n observaciones.
	`range.alpha`	vector de los valores reales α_j (entre 0 y 1) considerados por el usuario en la cuadrícula G de tamaño J.
	K	número de conglomerados.
	wt	vector con los pesos de las observaciones. Por defecto, `wt = NULL` corresponde al caso en el que todas las observaciones están ponderadas por $1/n$.
	`scale`	si es Verdadero las dos matrices de disimilitud se escalan, es decir, se dividen por su máximo.
	`graph`	si es Verdadero los dos gráficos (proporción y proporción normalizada de inercia) se dibujan.

4.2. Hierarchical clustering with geographical contraints: `hclustgeo`

Esta función implementa el algoritmo de conglomerado jerárquico Ward-like que incluye restricciones de contigüidad suaves.

Este algoritmo toma como entrada dos matrices de disimilitud D_0 y D_1 y un parámetro de mezcla $\alpha \in [0, 1]$. Las disimilitudes pueden ser no euclidianas y los

pesos de las observaciones pueden no ser uniformes.

La matriz D_0 da las diferencias en el "espacio de características" (variables socioepidemiologicas, por ejemplo) y la matriz D_1 contiene las diferencias en el espacio de "restricción" (matriz de distancias geográficas o una matriz construida a partir de la matriz de contigüidad C, por ejemplo). Y por fin, el parámetro de mezcla α establece la importancia de la restricción en el procedimiento de agrupamiento.

```
hclustgeo(D0, D1 = NULL, alpha = 0, scale = TRUE, wt = NULL)
```

4.3. Pseudo-inercia de un conglomerado: `inertdiss`

La funcion `inert` calcula la inercia de un conglomerado, es decir, en un subconjunto de filas de una matriz de datos.

`inert(Z, indices = 1:nrow(Z), wt = rep(1/nrow(Z), nrow(Z)), M = rep(1, ncol(Z))),`

en que `Z` es la matriz de datos, `indices` es el vector que representa el subconjunto de filas, `wt` es el vector de peso y `M` es la matriz de distancia diagonal.

Aunque para calcular la pseudo-inercia, la función utilizada es:

`inertdiss(D, indices = NULL, wt = NULL)`, en que,

Argumentos	D	un objeto de classe `dist` con las disimilitudes entre las n observaciones. La función **as.dist** puede utilizarse para transformar un objeto de clase.
	indices	un vector con los índices del subconjunto de observaciones
	wt	vector con los pesos de las n observaciones

4.4. Gráfico del parámetro de mezcla: `plot.choicealpha`

Con la salida de la función `choicealpha` obtenemos el gráfico del criterio Q o $Qnorm$.

4.5. Medidas de agregación de Ward entre los singletons: `wardinit`

Esta función calcula las medidas de agregación de Ward entre pares de singleton, siendo la medida de agregación de Ward entre los singletons i y j ponderados por w_i y w_j es: $(w_i w_j)/({}_w i + w_j) d_{ij}^2$ donde d_{ij} es la diferencia entre i y j.
`wardinit(D, wt = NULL)`

4.6. Pseudo-inercia dentro del conglomerado basada en la disimilitud de una partición: `withindiss`

Esta función realiza la pseudo inercia dentro del conglomerado de una partición de una matriz de disimilitud.

`withindiss(D, part, wt = NULL)`, siendo `part` un vector con pertenencia a un conglomerado.

Capítulo 5

Detección de valores atípicos y distancias elegidas

Los datos espaciales consisten en valores de atributos empíricos (multivariados) asociados con coordenadas geográficas. Si bien los outliers globales son puntos de datos que se encuentran lejos de la mayor parte de los datos en el espacio, los outliers locales difieren en sus atributos no espaciales de las observaciones dentro de un vecindario restringido localmente (Filzmoser et al. 2014 [60]).

Por lo tanto, para introducir adecuadamente el concepto de un valor atípico espacial, se requiere una definición precisa de un vecindario espacial. Una opción común es definir la vecindad local, M_i, siendo $i = 1, \ldots, 223$ en el dominio espacial a través de una distancia máxima $d_{\text{máx}}$ alrededor de una observación M_i, para $i \in \{1, \ldots, 223\}$, que representa la i-ésima fila de la matriz de datos $M_{(n \times p)}$. Los puntos de datos $M_j \in N_j$ para $j \in (1, \ldots, i-1, i+1, \ldots, n)$ se consideran vecinos de M_i con $d_i, d_{\text{máx}}$, donde d_i, j denota la distancia entre los municipios M_i y M_j.

Sin embargo, siguiendo este enfoque, pueden surgir problemas en el área fronteriza, ya que están poco poblados por puntos de datos vecinos y el tamaño de la población del vecindario puede variar ampliamente. Alternativamente,

puede ser de interés considerar un número fijo k de vecinos que puede lograrse mediante el enfoque de k vecinos más cercanos (de *k-nearest neighbours*, kNN). Las distancias por pares entre M_i y todos los M_j restantes se ordenan como d_i, $(1) \leq d_i, (2) \leq \ldots \leq d_i, (n-1)$, y la vecindad local de M_i se define formalmente como $N_i = (M_j \in X : d_i, (j) \leq d_i, (k))$. Los puntos de datos asociados con las k distancias más pequeñas formarán entonces la vecindad restringida.

Un enfoque de Filzmoser et al. (2014) [60] considera el llamado grado de aislamiento de una observación M_i de una proporción preestablecida $(1-\beta)$ de su vecindad N_i.

$$\chi^2_{p;\alpha(i)} \left(MD^2(x_i) \right) \quad = \quad MD^2 \left(x_i, x_{([n(i)\beta])} \right) \tag{5.1}$$

donde la medida $\alpha(i)$ es indicativo de outliers local del município M_i. La fración $[n(i)\beta]$ expresa el número de puntos (municipios) similares dentro de $N_i, i = 1, \ldots, 223$.

Consideramos los métodos propuestos por Filzmoser et al. (2018) y su implementación en el paquete R `mvoutlier` para la detección de valores atípicos multivariados. Más específicamente, aplicamos la función de gráfico de cuantiles ajustada `aq.plot` a nuestros datos utilizando valores predeterminados, a saber, un cuantil de chisq de .975, un delta de .05 y un alfa de .05 (Filzmoser y Gschwandtner, 2018 [61]).

Capítulo 6

Distancia geografica entre los municípios de Paraíba

La matriz de disimilitud D_1 es calculada a partir de la matriz de distancia geográfica $D.geo$ en km entre los municipios del Estado de Paraíba.

La distancia geográfica entre los municipios es la distancia medida a lo largo de la superficie de la tierra. Fueran calculadas entre puntos que se definen por coordenadas geográficas en términos de latitud y longitud se basa en los centros de los municipios.

El cálculo de la distancia entre coordenadas geográficas se basa en cierto nivel de abstracción; no proporciona una distancia exacta, que es inalcanzable si se intenta dar cuenta de cada irregularidad en la superficie de la tierra (CFR, 2016 [56]). Las abstracciones comunes para la superficie entre dos puntos geográficos aquí consideradas son las superficies plana.

La matriz de distancia $D.geo$ se calcula entre los municípios, M_1 e M_2. Las coordenadas geográficas de los dos puntos, como pares (latitud, longitud), son $(Theta_1, \lambda_1)$ y (Θ_2, λ_2), respectivamente.

Los elementos de la matriz distancia tiene uma gran importancia en la análisis de conglomerado. Las largas distancias influyen en gran medida en la composición de los conglomerados. La distancia es bastante alteradas por outliers. Esos valores elevados destuercen a verdadera estructura y tornan los conglomerados derivados no representativos de la verdadera estructura de la población.

Por esa razón, un triaje preliminar en busca de outlierses es necesaria, pues en análisis de conglomerados las grandes distancias entre los municipios son sensibles y refliten en la composición de los dendrogramas. Los outliers pueden presentar (1) observaciones que pueden ser llamadas de verdaderas anomalías y que no son representativas de población general; (2) ítems de un determinado grupo, obtenidos de una mala muestra llevan a una mala representación de los grupos creados del análisis de conglomerados.

La distancia de Mahalanobis (DM) es una métrica de distancia efectiva que encuentra la distancia entre el punto y una distribución. Funciona con bastante eficacia en datos multivariados. La razón por la que DM es eficaz en datos multivariados es porque usa la covarianza entre variables para encontrar la distancia en Km entre los municipios. En otras palabras, Mahalanobis calcula la distancia entre el punto "M_1" y el punto "M_2"considerando la desviación estándar.

Para los cálculos en R de la matriz de distancia geográfica *D.geo* se utilizó de las funciones de paquetes `Imap` (Wallace, 2012 [57]), la función `ReplaceLowerOrUpperTriangle` del package `sgeostat2016` (Majure y Gebhardt, 2016 []) y `geosphere` (Hijmans, 2019 [59]).

Capítulo 7

Estudio de caso: conglomerado jerárquico Ward-like con restricciones espaciales en datos de tuberculosis

En el período 2001-2018 se reportaron 24.258 casos de tuberculosis en el Estado de Paraíba (Brasil), de los cuales 80 % fueron casos nuevos; el 65 % se curaron de la enfermedad; el 46,8 % tenían menos de diez años de escolaridad; el 63,2 % tenían entre 20 y 49 años y el 67 % eran hombres.

La meta establecida por la OMS era curar el 85 % de los nuevos casos de tuberculosis bacilífera para el 2020 (OMS, 2017), sin embargo, como se observa en los datos de 2018, Brasil no alcanza esta meta (71,4 %), y la situación es aún más crítica para el Estado de Paraíba (55,5 %) (Brasil, 2019 [49]). En un estudio realizado en 2016, se concluyó que en Brasil cuanto menor es el nivel de educación de los pacientes (menos de nueve años de educación formal), mayor es el número de nuevos casos de tuberculosis y mayores son las tasas de cura y abandono del tratamiento en todo el país (Camêlo et al., 2016 [50]).

Se analizan, mediante tres estudios, los casos notificados de tuberculosis en los

223 municipios del Estado de Paraíba en el período comprendido entre 2001 y 2018. Las variables son proporciones y se dividen en epidemiológicas y sociales.

Los datos se obtuvieron de una fuente secundaria, a través de la base de datos, registrada en el Sistema de Información de Enfermedades Notificables (SINAN, 2020 [52]) y disponible en la página web del Departamento de Informática del Sistema Único de Salud (DATASUS). Para el análisis de datos se utilizó el programa *R* versión 3.6.2 (R Core Team, 2019 [53]).

Como se trata de una encuesta de datos secundarios y no implica directamente a seres humanos, estos estudios no fueran sometidos a la evaluación del Comité de Ética en Investigación.

El principal diferencial entre los tres estudios son los pesos w de las observaciones que tomaron variables diferentes.

Los pesos $w = (w_1, \ldots, w_n)$ son un vector n-dimensional de los pesos de las observaciones como argumentos. Para cada uno de los estudios, utilizamos un w.

En el primero estudio, tenemos el vector w con pesos uniformes del Índice de Desigualdad Colectiva (IDC) para cada una de las 23 microrregiones del Capítulo 3.5.1. Las variables son proporciones y se dividen en epidemiológicas (casos nuevos y curados) y variables sociales como *los años de estudio* (menos de diez años de educación formal) y edad laboral (20-49).

También se calculó una matriz con las distancias geográficas entre los municipios y el peso w fue atribuido al cálculo de la matriz de disimilitud D del IDC y del Producto Interno Bruto (PIB) de los municipios. Este indicador forma parte del sistema de información que da cuenta de las unidades geopolíticas en los niveles local, regional y nacional, permitiendo examinar las condiciones regionales

y locales.

La recolección de datos se llevó a cabo durante febrero de 2020. Como unidades de análisis se utilizaron municipios y microrregiones.

El segundo y tercer estudio, Capítulo 3.5.2 y 3.5.3, el vector w asumen valores no uniformes, es decir, para cada municipio se calcula su valor de la respectiva variable. Es decir, el w del segundo estudio se tomó como la variable coeficiente de diversificación (CD) calculado para cada municipio. En el tercer estudio, el w fue la variable razón de incidencia estandarizada (RIE) de la tuberculosis, también calculada para los 223 municipios.

Cuando los pesos no son uniformes, el uso de la función `hclustgeo` del paquete Chavent, et al., 2017 [43] es claramente más conveniente que la función `hclust` del paquete `stats` (R Core Team (2020), [53]).

El segundo estudio utiliza variables epidemiológicas (casos nuevos y curados) y variables sociales como años de estudio (menos de diez años de educación formal) y edad laboral (20 -49). También se calculó una matriz con las distancias geográficas entre los municipios y el peso w atribuido al cálculo de la matriz de disimilitud D como el coeficiente de diversificación de la tuberculosis en el Estado de Paraíba.

La recolección de datos se llevó a cabo durante febrero de 2020. Municipios y microrregiones fueron usadas como unidades de análisis.

En el tercer estudio fue publicado en la revista *Mathematics* con factor de impacto JCR 1,747 (Camêlo-Aguiar, et al., 2020 [51]). Las variables consideradas fueron proporciones, divididas en epidemiológicas (casos nuevos, cura, muertes masculinas y femeninas) y variable social (pacientes con tuberculosis en edad activa (20-64)). También se calculó una matriz con las distancias geográficas entre los

municipios y el peso w no uniforme atribuido al cálculo de la matriz de disimilitud D, como la razón de incidencia estandarizada (SIR) de tuberculosis en el Estado de Paraíba. Los datos se recopilaron entre febrero y mayo de 2020.

En este capítulo ilustramos cómo la metodología de conglomerado jerárquico Ward-like presentada en el Capítulo 3 se puede aplicar en un contexto del mundo real.

7.1. Conglomerado jerárquico con restricciones espaciales

Sabemos que los determinantes socioeconómicos tienen un impacto sustancial en el control de las enfermedades infecciosas, por ello hemos incluido el índice de desigualdad colectiva (IDC), aunque ha influido en el aumento de la heterogeneidad entre los municipios debido a las desigualdades económicas entre ellos.

Los enfoques de agrupamiento son una herramienta útil para detectar patrones en conjuntos de datos y generar hipótesis sobre posibles relaciones. Por lo tanto, el papel del análisis de conglomerados es descubrir un cierto tipo de estructura natural en el conjunto de datos (Wierzchoń y Kłopotek, 2018 [48]).

La Figura 7.1 muestra el dendrograma de la matriz de disimilitudes D_0; es decir, las diferencias en el espacio característico de las variables socioepidemiológicas.

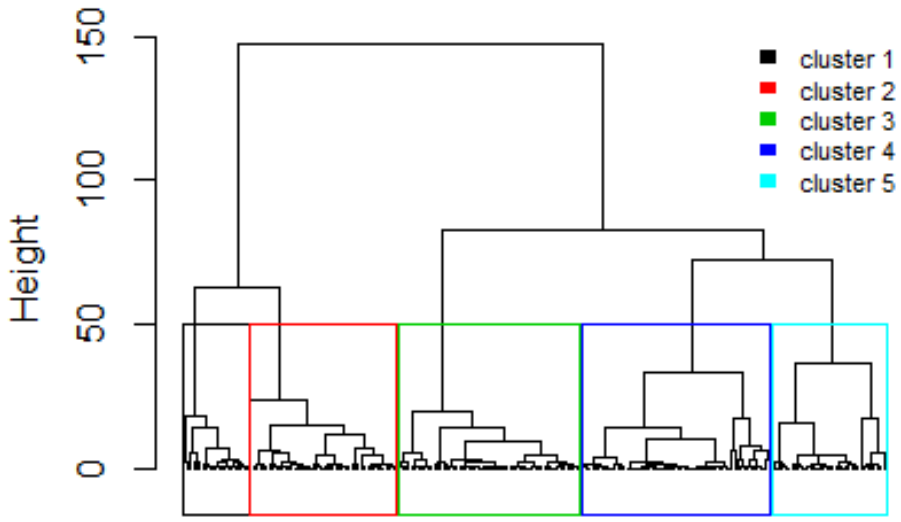

Figura 7.1: Dendrograma de los $n = 223$ municipios con base en las 4 variables socioepidemiológicas (es decir, usando solo D_0).

La inspección visual del dendrograma en la Figura 7.1 sugiere retener $K = 5$ conglomerados. La partición correspondiente a los cinco clústeres se puede ver en el mapa de la Figura 7.2.

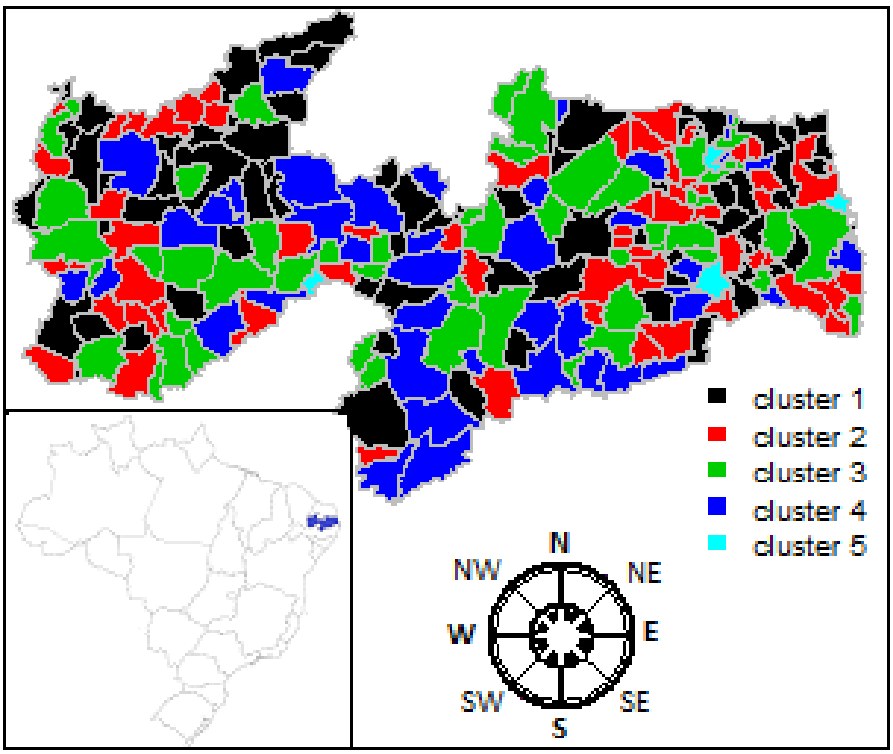

Figura 7.2: Mapa de la partición con $K = 5$ conglomerados solo basado en las variables socioepidemiológicas (es decir, usando solo D_0).

Los 223 municipios se agruparon en sus respectivos conglomerados según su similitud socioepidemiológica, a saber, conglomerado 1 (68 municipios), conglomerado 2 (58 municipios), conglomerado 3 (52 municipios), conglomerado 4 (41 municipios) y conglomerado 5 solo cuatro municipios.

Es interesante la interpretación de los conglomerados según las variables socioepidemiológicas iniciales. Geográficamente percibimos que los conglomerados están bastante dispersos según variables socioepidemiológicas; es decir, los conglomerados no son estrictamente contiguos.

El conglomerado 1 tiene la proporción más baja de años de estudio en pacientes

con tuberculosis en el área de estudio; por el contrario tiene mayor incidencia de nuevos casos.

Existe una alta proporción de casos nuevos y una baja proporción de pacientes con tuberculosis en edad laboral en el conglomerado 2.

El conglomerado 3 tiene alta tasa de casos nuevos, una baja tasa de escolaridad (por debajo del valor promedio del área de estudio) y la tasa más baja de pacientes con tuberculosis en edad activa en todos los conglomerados.

Tasas de cura más bajas y alta proporción de casos nuevos (aunque su proporción media es menor en comparación con otros grupos) se observa en el conglomerado 4.

El conglomerado 5 tiene alta tasa de personas en edad de trabajo, con baja escolaridad y la tasa media de curación es ligeramente superior a la de los casos nuevos.

Para obtener conglomerados geográficamente más compactos, introduciremos la matriz D_1 de distancias geográficas en `hclustgeo`. Para ello, es necesario que se seleccione un parámetro de mezcla α para mejorar la cohesión geográfica de los 5 conglomerados sin afectar negativamente a la cohesión socioepidemiológica.

En la Figura 7.3, tenemos el parámetro de mezcla $\alpha \in [0, 1]$ que define la importancia de D_0 y D_1 en el proceso de agrupamiento con cálculos separados para la homogeneidad socioeconómica y la cohesión geográfica de las particiones obtenidas para un rango de diferentes valores de α y los 5 grupos y la cohesión geográfica de las particiones obtenidas para un rango de diferentes valores de α y los 5 conglomerados.

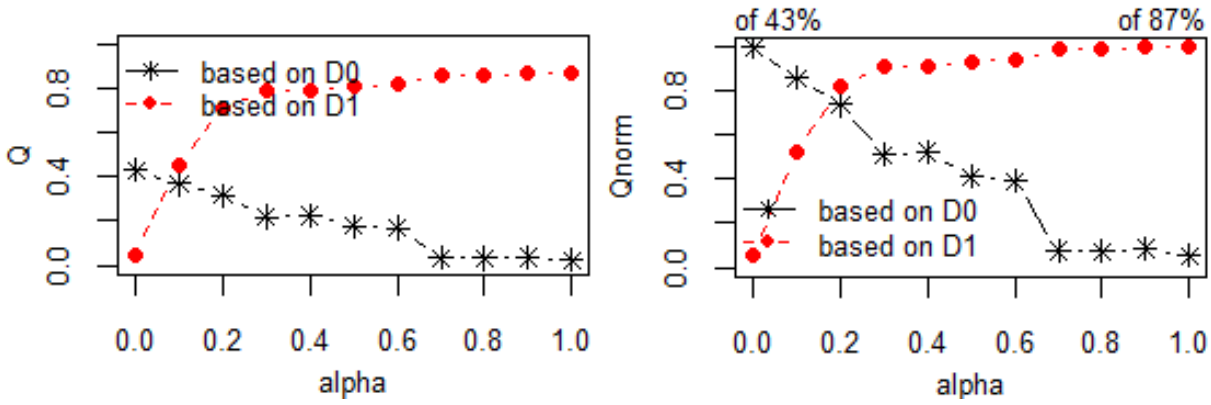

Figura 7.3: Elección de α para una partición en $K = 5$ conglomerados cuando D_1 son las distancias geográficas entre municipios. Izquierda: proporción de pseudo-inercias explicadas $Q_0(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1(P_K^\alpha)$ versus α (en línea discontinua). Derecha: proporción normalizada de pseudo-inercias explicadas $Q_0^*(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1^*(P_K^\alpha)$ versus α (en línea discontinua).

Al obtener la partición teniendo en cuenta las restricciones geográficas en la Figura 7.3, se muestra el valor α, que tiene como objetivo aumentar la contigüidad espacial. Cuando $\alpha = 0$ no se tienen en cuenta las disimilitudes geográficas y cuando $\alpha = 1$ son las distancias socioepidemiológicas las que no se tienen en cuenta, y los conglomerados se obtienen con las distancias geográficas únicamente.

La Figura 7.3 muestra la gráfica de la proporción de pseudo-inercia explicada calculada con D_0 (distancias socioepidemiológicas), que es igual a 0.43 cuando $\alpha = 0$ y disminuye cuando α aumenta (línea negra continua).

Por el contrario, la proporción de pseudo-inercia explicada calculada con D_1 (las distancias geográficas) es igual a 0,87 cuando $\alpha = 1$ y disminuye cuando α decrece (línea discontinua).

La gráfica de la proporción normalizada de inercias explicadas (Figura 7.3)

sugiere retener $\alpha = 0{,}1$ o 0.2. El valor $\alpha = 0,1$ favorece levemente la homogeneidad socioepidemiológica frente a la homogeneidad geográfica.

De acuerdo con la prioridad dada en esta solicitud a los aspectos socioepidemiológicos, la partición final obtenida con $\alpha = 0{,}1$, que corresponde a una pérdida de solo (1-0.76) 24 % de homogeneidad socioepidemiológica, y un aumento (1-0.36) 64 % en la homogeneidad geográfica.

La mayor cohesión geográfica de esta partición con D_0 y D_1 y $\alpha = 0{,}1$ se puede ver en la Figura 7.4.

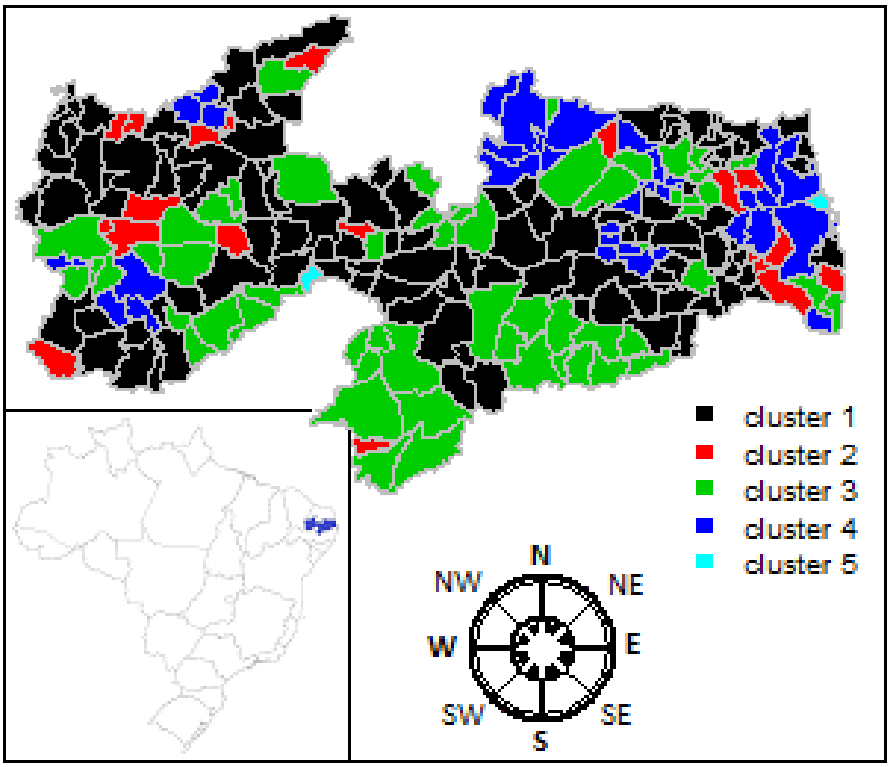

Figura 7.4: Mapa de la partición con $K = 5$ conglomerados basado en las distancias socioepidemiológicas D_0 y las distancias geográficas entre los municipios D_1 con $\alpha = 0,1$.

En la Figura 7.4 se percibe una ganancia en la homogeneidad espacial, principalmente en los conglomerados 1 y 3. Los conglomerados 2 y 4 se alteraron significativamente. La Figura 7 muestra los diagramas de caja de las variables para cada grupo de la partición (fila del medio).

El cambio en el conglomerado 4 (partición de la Figura 7.4) en relación con el conglomerado 4 (partición Figura 7.2) se debió principalmente a la variable de cura, con la tasa más baja en el área de estudio.

El conglomerado 2 (partición de la Figura 7.4) tiene proporción media más alta de pacientes con tuberculosis en edad de trabajo y tasa de escolarización más baja; lo contrario ocurre en el conglomerado 2 (partición de la Figura 7.2), una mayor tasa de escolaridad y la menor proporción media de edad laboral.

Los conglomerados 5 de la partición de la Figura 7.4 y el conglomerado 5 de la partición de la Figura 7.2 son semejantes.

La siguiente gráfica, Figura 7.5, muestra la elección de alfa para la partición.

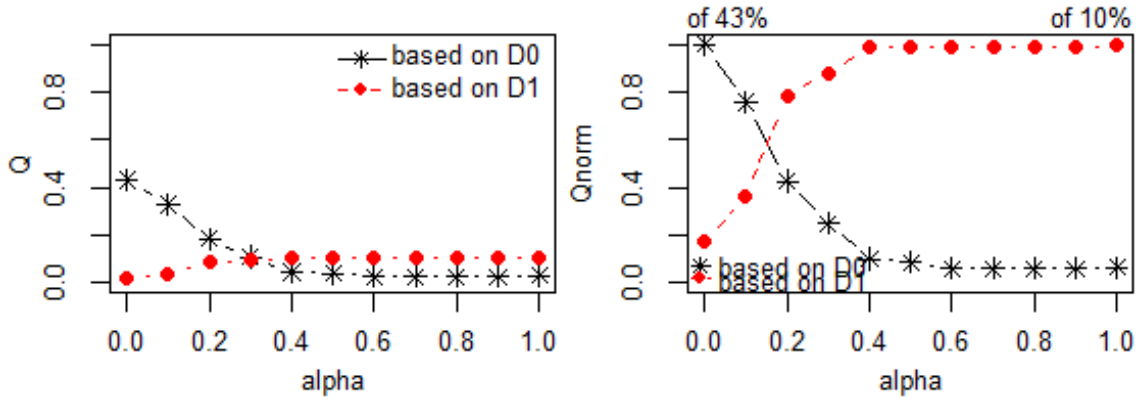

Figura 7.5: Elección de α para una partición en $K = 5$ conglomerados cuando D_1 es la matriz de disimilitud de vecindario entre municipios. Izquierda: proporción de pseudo-inercias explicadas $Q_0(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1(P_K^\alpha)$ versus α (en línea discontinua). Derecha: proporción normalizada de pseudo-inercias explicadas $Q_0^*(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1^*(P_K^\alpha)$ versus α (en línea discontinua).

A la derecha de la Figura 7.5, la gráfica de la proporción normalizada de inercias explicadas (es decir, $Q_0(P_K^\alpha)$ y $Q_1(P_K^\alpha)$) sugiere retener $\alpha = 0{,}2$ favoreciendo levemente la homogeneidad socioepidemiológica versus la homogeneidad geográfica.

Solo queda determinar esta partición final para $K = 5$ conglomerados y $\alpha = 0,2$. El mapa correspondiente se muestra en la Figura 7.6.

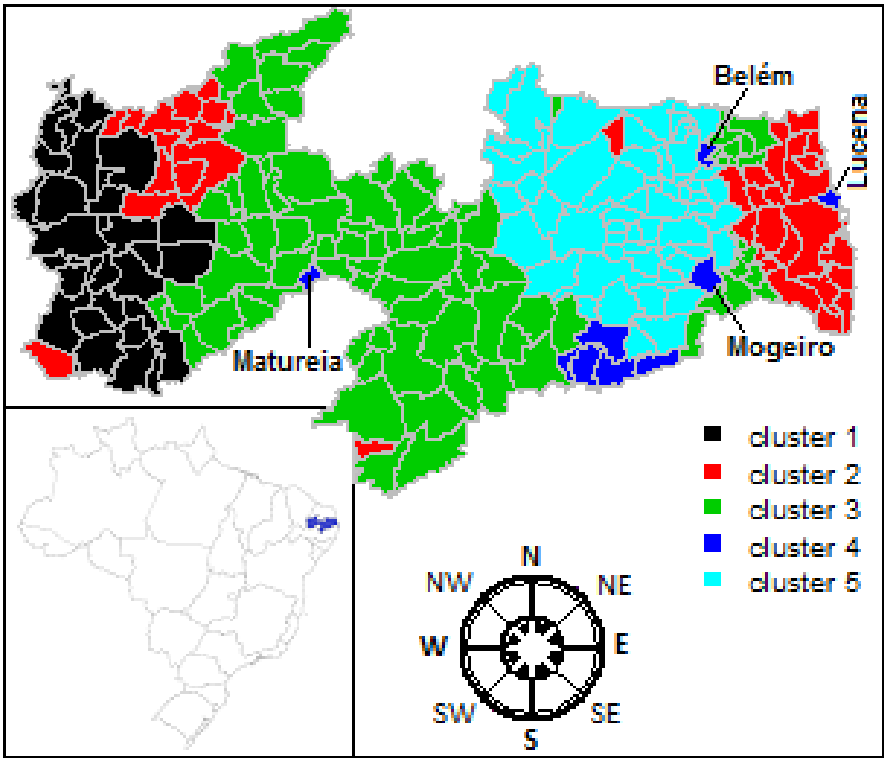

Figura 7.6: Mapa de la partición con $K = 5$ conglomerados basado en las distancias socioepidemiológicas D_0 y las distancias "vecinas" de los municipios D_1 con $\alpha = 0,2$.

La Figura 7.6 muestra que los conglomerados son espacialmente más compactos que los de la Figura 7.4.

Sin embargo, se sabe que este enfoque crea divergencias en la matriz de adyacencia, lo que le da más importancia a los barrios.

No obstante, como el enfoque se basa en restricciones suaves de contigüidad, los municipios que no son vecinos pueden estar en el mismo conglomerado según ocurre con los municipios de Lucena, Belém, Matureia y Mogeiro en el conglomerado 4. La calidad de la partición en la Figura 7.6 es ligeramente peor que el de la

partición de la Figura 7.4, según el criterio Q_0 (32,61 % frente a 36,98 %).

En el siguiente gráfico puede verse detalladamente la partición en términos de variables (socioepidemiológicas) de las Figuras 7.2, 7.4 y 7.6.

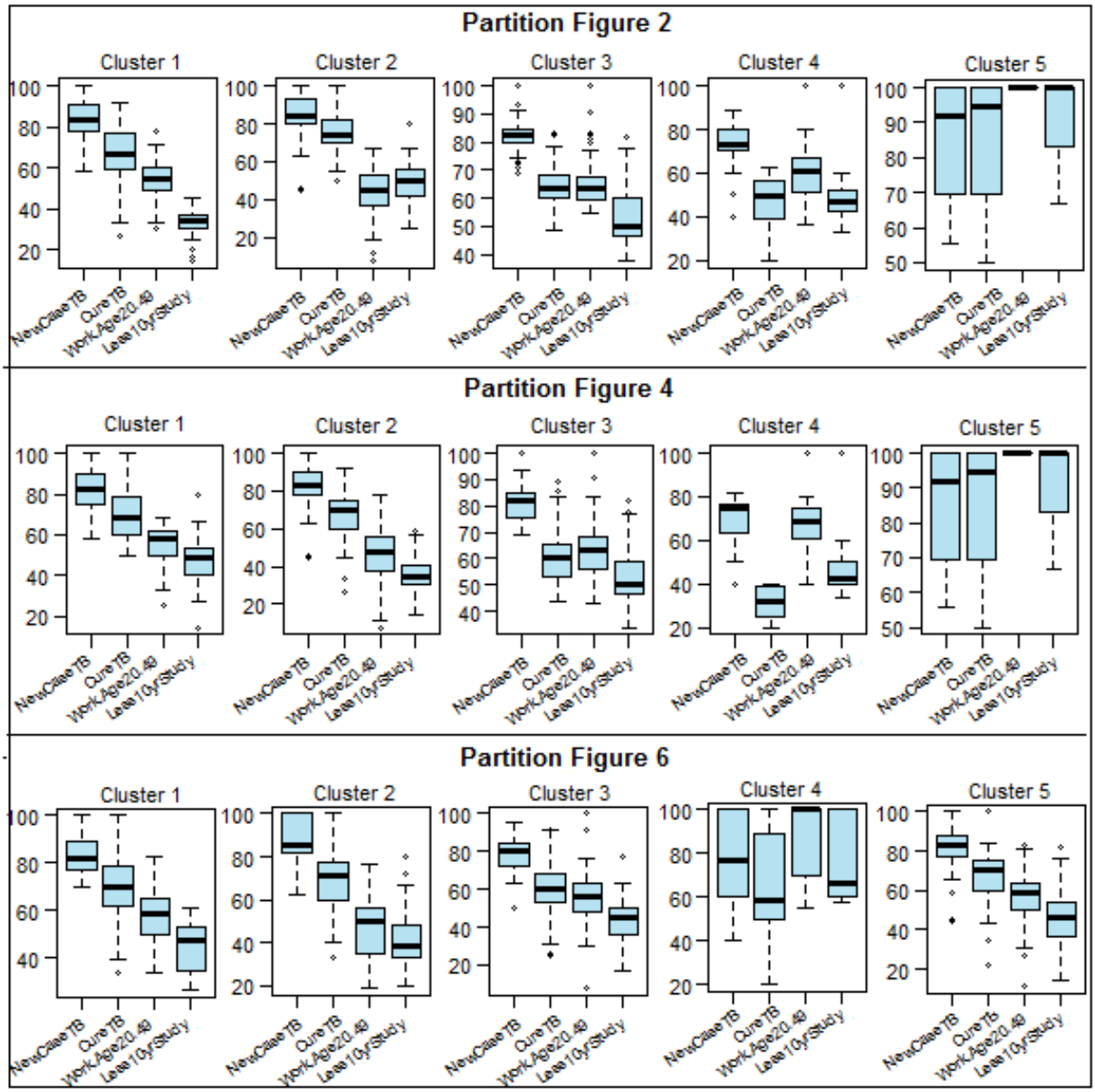

Figura 7.7: Comparación de las particiones finales Figura 7.2, Figura 7.4 y Figura 7.6 en términos de variables.

7.1.1. Conclusión

La aplicación del método de conglomerado jerárquico Ward-like resulta viable en estudios epidemiológicos, ya que permite considerar dos matrices simultáneamente, la primera con diferencias en el espacio de características (variables socioepidemiológicas) y la segunda con diferencias en el espacio de restricción (distancia geográficas) con un parámetro de mezcla alfa para mejorar la cohesión geográfica de los conglomerados sin afectar negativamente a la cohesión socioepidemiológica.

De este modo, al considerar las restricciones espaciales, el conglomerado jerárquico se vuelve aún más completo, una vez que detecta patrones en conjuntos de datos de diferentes dimensiones.

Por tanto, su aplicación se torna fundamental para un mejor conocimiento de la realidad socioepidemiológica y económica del municipio, ya que es una herramienta de análisis que permite tomar mejores decisiones en la elaboración de políticas públicas y acciones de salud más efectivas en la lucha contra la tuberculosis, ya que esta enfermedad está directamente relacionada con el gradiente socioeconómico, el nivel de pobreza y el contexto social.

Las dificultades del Estado de Paraíba y del propio Brasil con la tuberculosis, especialmente con la cura de nuevos casos bacilíferos, son preocupantes y el escenario podría ser aún peor, ya que la finaniación para la cura de la tuberculosis en Brasil ha venido disminuyendo significativamente desde 2018; en 2019, el presupuesto nacional de tuberculosis fue de solo 38 (millones de dólares), además de los cambios en la regulación de la inversión financiada con fondos federales en áreas estratégicas de salud y límites estrictos impuestos al crecimiento del gasto público hasta 2036.

7.2. Conglomerado jerárquico Ward-like con disimilitudes y pesos no uniformes

Es importante saber si la situación de la tuberculosis en el estado de Paraíba está diversificada o no. A partir de las variables socioepidemiológicas calcularemos el coeficiente de diversificación. Los valores del coeficiente de diversificación para 23 microrregiones se muestran en la Figura 7.8.

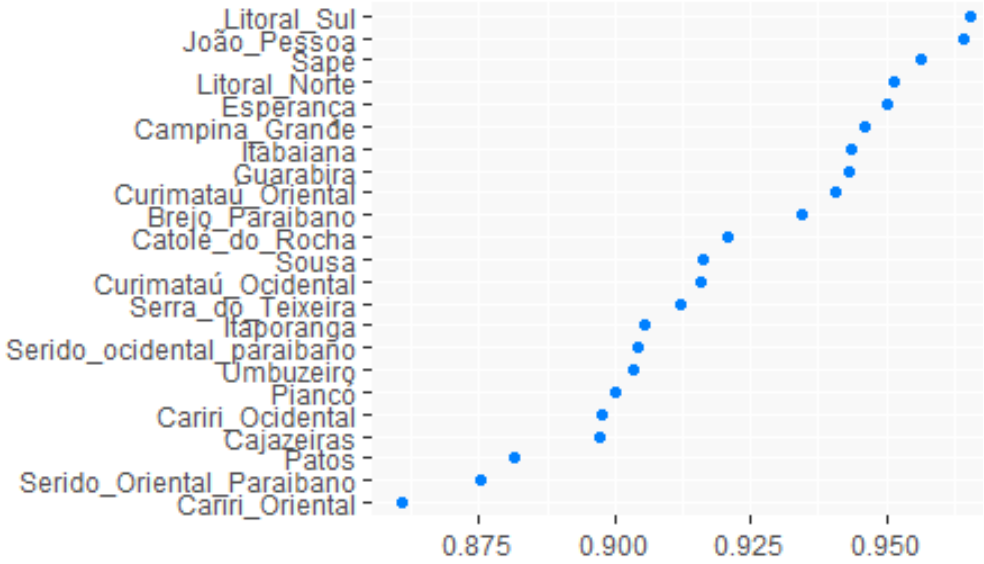

Figura 7.8: Valores del coeficiente de diversificación (CD) de variables socioepidemiológicas de las microrregiones, Paraíba, Brasil, 2001-2018.

Si una microrregión se clasifica como diversificada, su situación epidemiológica de tuberculosis no depende mucho de ninguna variable específica, es decir, todas están igualmente influenciadas por el conjunto de variables. Se puede observar en la Figura 7.8 que la mayoría de las microrregiones tienen una medida de diversificación cercana a 1, con valor mínimo de aproximadamente 0,68 y máximo de 0,967, Cariri Oriental y Litoral Sul, respectivamente. La diversificación en la microrregión Cariri Oriental se ve disminuida por la existencia de desigualdades

entre las variables epidemiológicas (nuevos casos y cura) y sociales (menos de diez años de educación formal y edad laboral (20-49)), centrándose más en una de ellas.

Este coeficiente de diversificación es el peso de la restricción sobre la calidad de las soluciones y está controlado por α, que define la importancia de la restricción en el procedimiento de agrupamiento.

La Figura 7.9 muestra el dendrograma de la matriz de disimilitud D_0: las diferencias en el espacio de características de las variables socioepidemiológicas y el mapa de partición correspondiente a los cinco conglomerados.

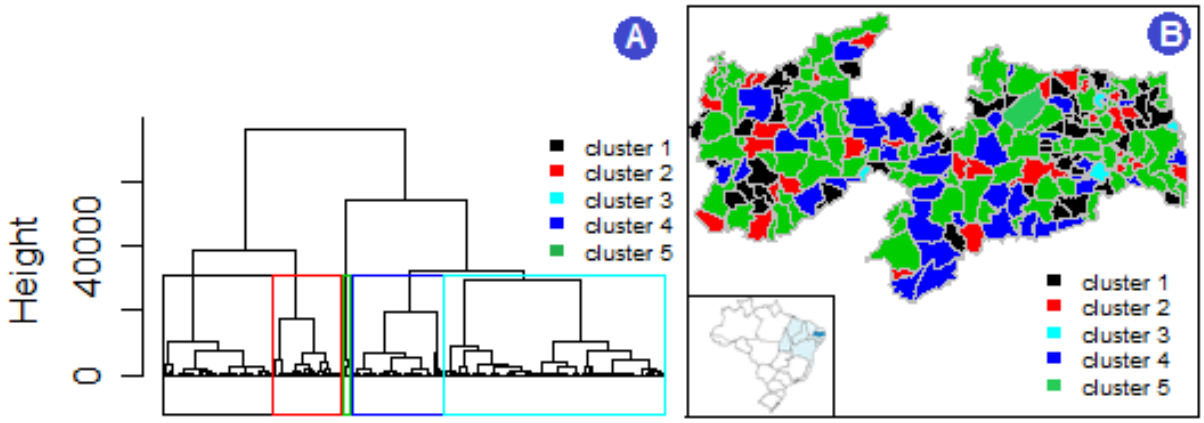

Figura 7.9: Valores del coeficiente de diversificación (CD) de variables socioepidemiológicas de las microrregiones, Paraíba, Brasil.

Según el criterio del método Ward-like, la Figura 7.9 (A) muestra el dendrograma de la matriz de distancias de los 223 municipios utilizando solo las cuatro variables socioepidemiológicas según las medidas de diversificación.

La inspección visual del dendrograma en la Figura 7.9 (A) sugiere retener $K = 5$ grupos. El mapa proporcionado presenta la partición correspondiente en cinco grupos Figura 7.9 (B).

Geográficamente percibimos conglomerados muy dispersos según las variables socioepidemiológicas; es decir, los conglomerados no son estrictamente contiguos. Se observa que los 5 conglomerados están bien distribuidos dentro del Estado de Paraíba.

La función `choicealpha` del paquete `ClustGeo` (Chavent et al, 2018 [44]) que encuentra un valor alfa para importancia relativa entre las matrices de disimilitud D_0 y D_1. Se consideró el valor $\alpha = 0{,}3$, ya que la partición tiene en cuenta las restricciones geográficas de la Figura 7.10.

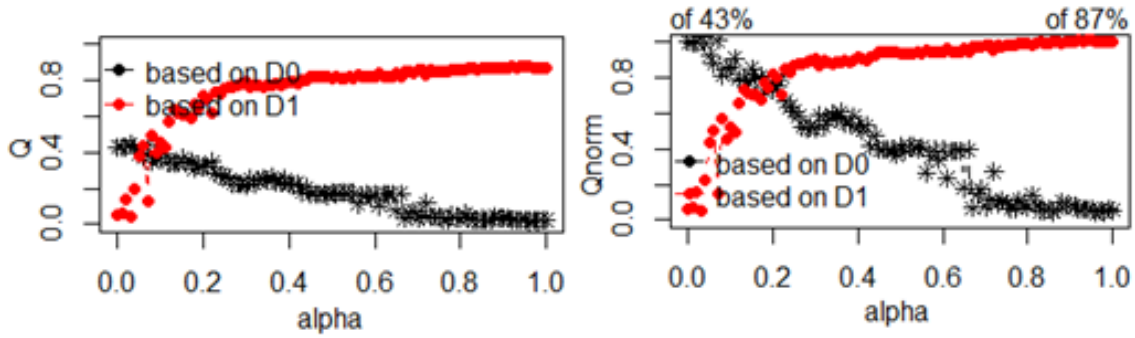

Figura 7.10: Elección de α para una partición en $K = 5$ conglomerados cuando D_1 son las distancias geográficas entre municipios. Izquierda: proporción de pseudo-inercias explicadas $Q_0(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1(P_K^\alpha)$ versus α (en línea discontinua). Derecha: proporción normalizada de pseudo-inercias explicadas $Q_0^*(P_K^\alpha)$ versus α (en línea continua negra) y $Q_1^*(P_K^\alpha)$ versus α (en línea discontinua).

Al obtener la partición teniendo en cuenta las restricciones geográficas en la Figura 7.10, se muestra el valor α que tiene como objetivo aumentar la contigüidad espacial, visto en detalle en la Tabla 7.1.

Tabla 7.1: Proporción normalizada de pseudo-inercias explicadas.

Valores de Alpha (α)	Q_0**norm**	Q_1**norm**
α=0,17	0,80773244	0,68104151
α=0,18	0,71786338	0,76987949
α=0,19	0,75936603	0,74331210
α=0,20	0,73858351	0,82422833
α=0,21	0,75132496	0,80288570

Cuando $\alpha = 0$ no se toman en cuenta las disimilitudes geográficas y cuando $\alpha = 1$ son las distancias socioepidemiológicas que no se toman en cuenta, luego, los conglomerados se obtienen solo con las distancias geográficas.

La gráfica de la Figura 7.10 (izquierda) parece sugerir la elección de $\alpha = 0,2$ que corresponde a una pérdida de solo (1-0,7385 = 26,11 %) de homogeneidad socioepidemiológica con el coeficiente de diversificación de cada municipio, y un aumento del 17,58 % en la homogeneidad geográfica.

La mayor cohesión geográfica de esta partición se puede ver en la Figura 7.11.

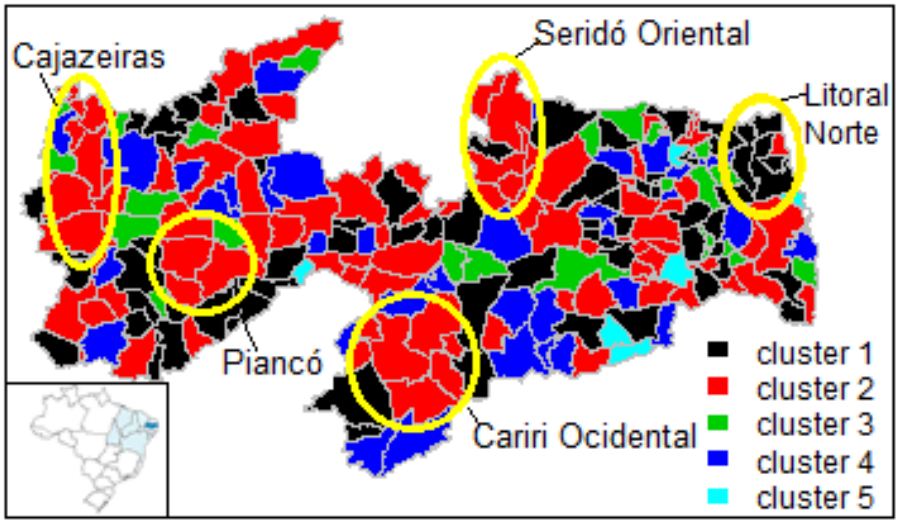

Figura 7.11: Mapa de la partición con $K = 5$ conglomerados basado en las distancias socioepidemiológicas D_0 y las distancias geográficas D_1 entre los municipios con $\alpha = 0,2$

En la Figura 7.11 se percibe nuevamente la homogeneidad espacial, principalmente en el conglomerado 2, a continuación aparece el conglomerado 1. Los municipios en círculos amarillos quedaron bien ubicados en las microrregiones de Cariri Ocidental, Piancó, Cajazeiras y Seridó Oriental para el conglomerado 2 y los municipios del Litoral Norte en el conglomerado 1.

Los cambios significativos ocurrieron principalmente en el conglomerado 3. La Figura 7.12 muestra los diagramas de caja de las variables para cada conglomerado de la partición de la Figura 7.11.

El conglomerado 1 presentó un comportamiento similar al conglomerado 2.

Parece que los conglomerados 3 y 5 se separaron en función de la proporción de pacientes con tuberculosis en edad laboral, porque los municipios del conglomerado 3 tienen menores proporciones de pacientes con tuberculosis en edad laboral y

con menos de diez años de estudio, por lo contrario, el conglomerado 5 presenta mayores proporciones de personas con menos de diez años de estudio en edad laboral.

El conglomerado 4 tiene la tasa de curación más baja de todos los grupos. Aunque tiene la proporción mediana más baja de casos nuevos, el conglomerado 5 tiene altas tasas de cura, mayor proporción de personas en edad de trabajo y con menos de diez años de escolaridad, en seis municipios, Maturéia, Gado Bravo, Mogeiro, Belém, Lucena y Umbuzeiro.

La partición en términos de variables (socioepidemiológicas) de la Figura 7.11 puede verse detalladamente en el siguiente gráfico:

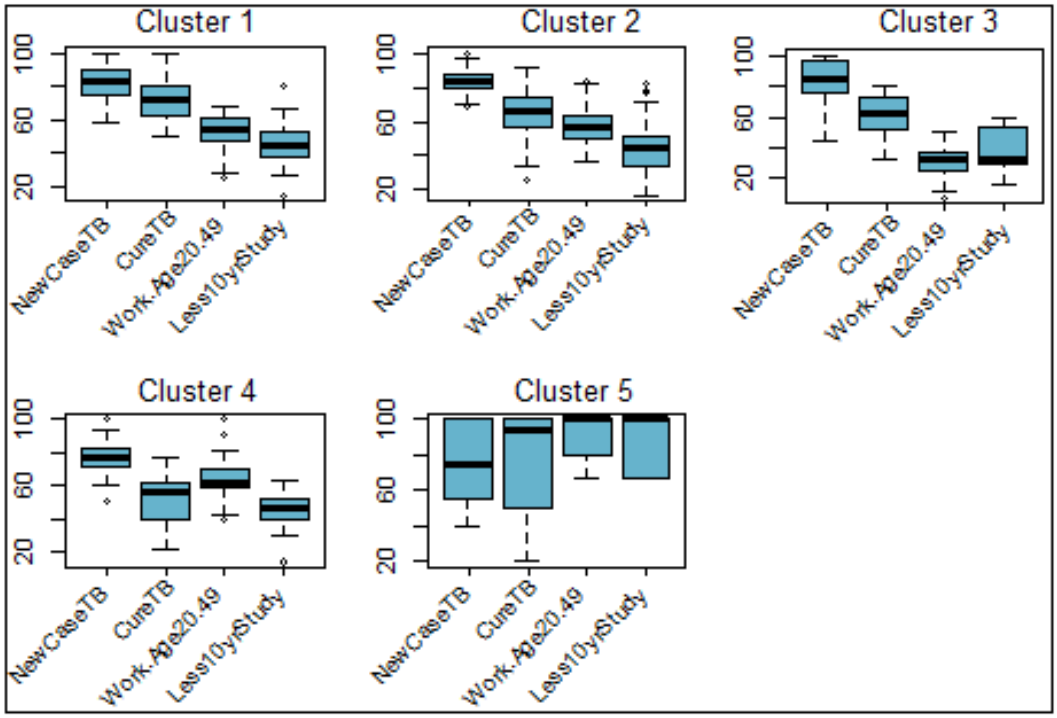

Figura 7.12: Comparación de conglomerados en la partición de la Figura 7.11 en términos de variables.

7.2.1. Conclusión

Al considerar las restricciones espaciales, el conglomerado jerárquico se vuelve aún más completo, ya que detecta patrones en conjuntos de datos de diferentes dimensiones. Por lo tanto, la aplicación del método Ward-Like se vuelve indispensable para una mejor comprensión de la realidad socioepidemiológica del Estado de Paraíba desde una perspectiva espacial.

7.3. Conglomerado jerárquico Ward-like con restricciones espaciales y tasa de incidencia estandarizada

Los análisis de normalidad permiten analizar cuánto difiere la distribución de los datos observados respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica.

En la Figura 7.13 los subgráficos (A, B) representan el diagrama de dispersión y la distribución de distancias de los puntos de datos compuestos, respectivamente. Los subgráficos (C, D) representan detecciones de valores atípicos utilizando diferentes cuantiles. Notamos la presencia de 5 valores atípicos, de un total de 223 municipios. Estos fueron los municipios Cacimba de Areia, Olivedos, Itaporanga, Santa Inês y Várzea.

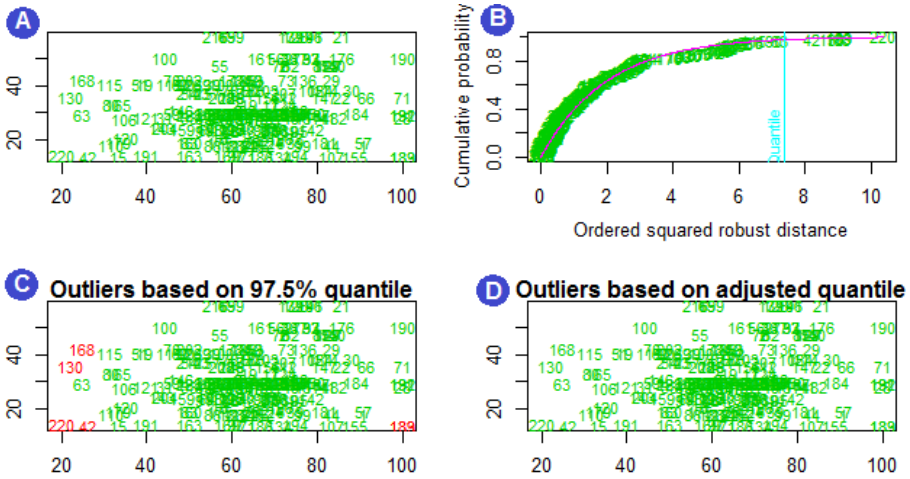

Figura 7.13: Dendrograma de los $n = 223$ municipios con base en las 5 variables socioepidemiológicas (es decir, que sólo utiliza D_0).

La Figura 7.14 muestra el dendrograma de la matriz de disimilitud D_0, es decir, las diferencias en el espacio de características de las variables socioepidemiológicas, que es la matriz de distancias de Manhattan entre los 223 municipios realizada $p = 5$ variables socioepidemiológicas.

Para elegir el número adecuado de K conglomerados, nos centramos en el dendrograma de Ward basado en las variables socioepidemiológicas $p = 5$, es decir, utilizando solo D_0.

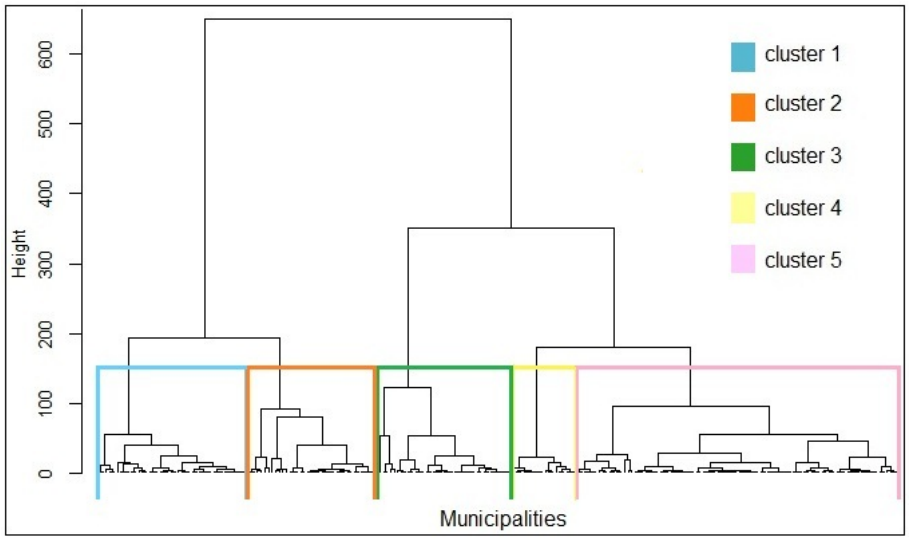

Figura 7.14: Dendrograma de los $n = 223$ municipios con base en las 5 variables socioepidemiológicas (es decir, que sólo utiliza D_0).

La inspección visual del dendrograma en la Figura 7.14 sugiere retener $K = 5$ conglomerados.

Los 223 municipios estan agrupados en sus respectivos conglomerados según similitud socioepidemiológica, a saber, conglomerado 1 con 42 municipios, el conglomerado 2 con 37 municipios; conglomerado 3 con 36 municipios; el conglomerado 4 con 90 municipios y el conglomerado 5 con solo 18 municipios. La partición correspondiente a los cinco grupos se muestra en el mapa presentado en la Figura 7.15.

Desde una perspectiva geográfica, percibimos conglomerados bien dispersos según variables las socioepidemiológicas; es decir, los conglomerados no son estrictamente contiguos. Es interesante la interpretación de los conglomerados según las variables socioepidemiológicas iniciales.

La Figura 7.20 muestra los diagramas de caja de las variables para cada conglo-

merado (fila superior).

En el conglomerado 1, la tasa de mortalidad femenina es la más baja de todos los conglomerados, mientras que la mortalidad masculina tiene la media más alta. El conglomerado 2 tiene alta tasa de casos nuevos y cura, y tasa de mortalidad femenina más alta que en otros conglomerados. El conglomerado 3, más personas en edad laboral (20-64) y tienen la tasa superior al valor medio en el área de estudio, además de ser superior a otros conglomerados. De manera similar, el conglomerado 4 también tiene alta tasa de casos nuevos y edad promedio alta de pacientes con tuberculosis en edad laboral y también es mayor que el valor promedio del área de estudio. El conglomerado 5, presenta altas tasas de casos nuevos y personas en edad de trabajo, y la tasa de cura más baja de todos los demás conglomerados.

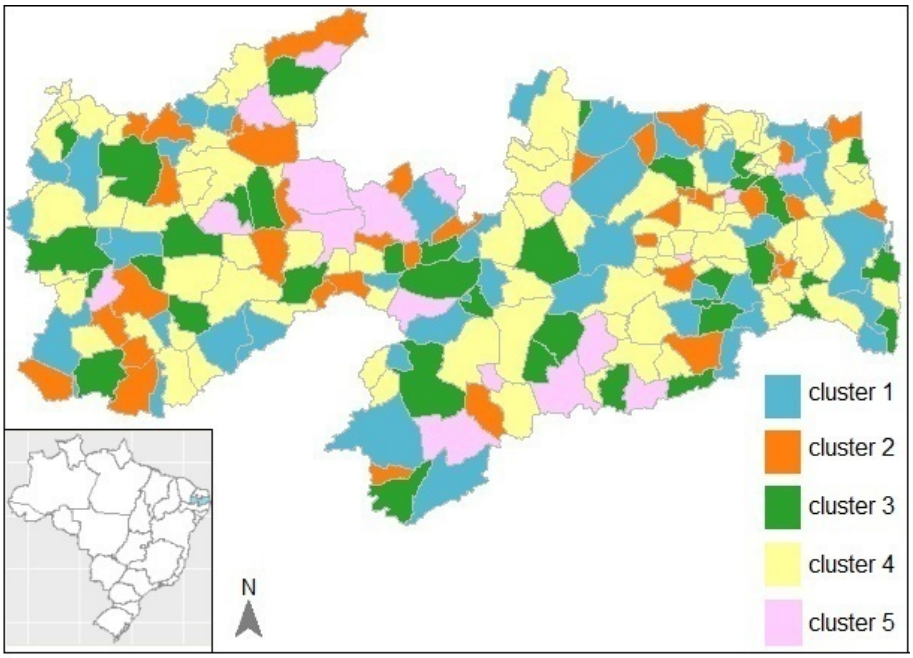

Figura 7.15: Mapa de la partición con $K = 5$ conglomerados solo basado en las variables socioepidemiológicas (es decir, usando solo D_0).

Introduciremos la matriz D_1 de distancias geográficas en `hclustgeo`, es decir,

una partición que tenga en cuenta las restricciones geográficas para obtener conglomerados geográficamente más compactos.

Para esto, es necesario que se seleccione un parámetro de mezcla α con el fin de mejorar la cohesión geográfica de los cinco conglomerados sin afectar adversamente la cohesión socioepidemiológica. En la Figura 7.16, tenemos el parámetro de mezcla $\alpha \in [0, 1]$ que define la importancia de D_0 y D_1 en el proceso de agrupamiento con cálculos separados para la homogeneidad socioepidemiológica y geográfica de las particiones obtenidas para un rango de valores diferentes de α y los cinco conglomerados.

La siguiente gráfica, Figura 7.16, muestra la elección de α para la partición, teniendo en cuenta las restricciones geográficas.

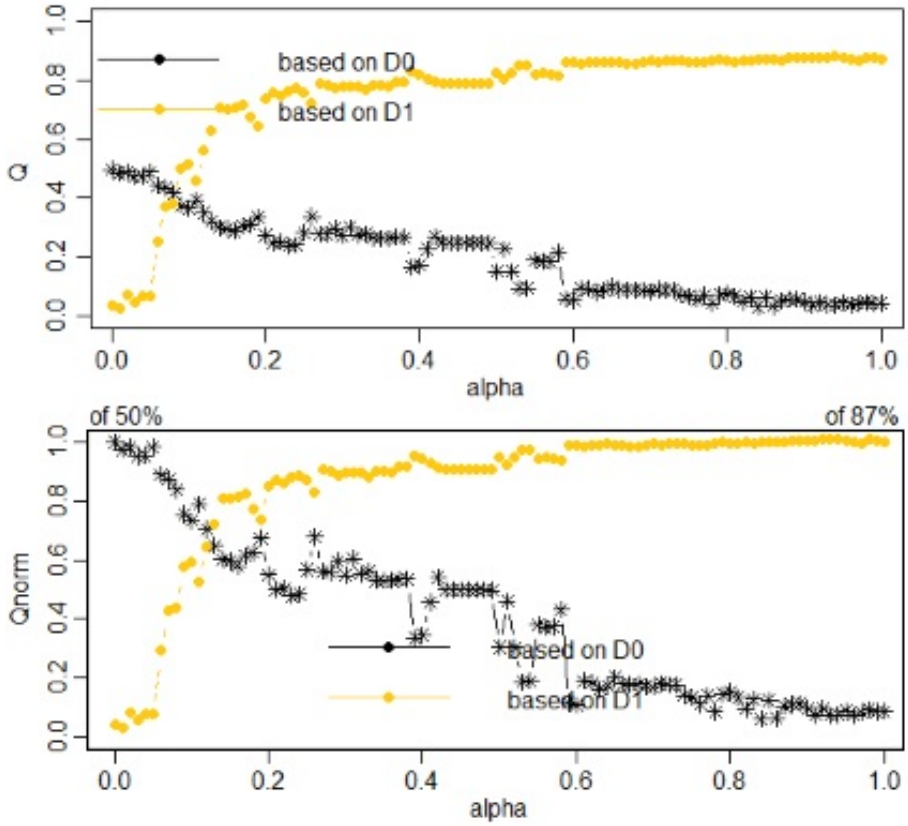

Figura 7.16: Elección de α para una partición en $K = 5$ conglomerados cuando D_1 son las distancias geográficas entre municipios. (**Arriba**) proporción de pseudo-inercias explicadas $Q_0(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1(P_K^\alpha)$ versus α (en línea discontinua dorada). (**Abajo**) proporción normalizada de pseudo-inercias explicadas $Q_0^*(P_K^\alpha)$ versus α (en línea negra continua) y $Q_1^*(P_K^\alpha)$ versus α (en línea discontinua dorada).

La Figura 7.16 muestra la gráfica de la proporción de pseudo-inercia explicada calculada con D_0 (las distancias socioepidemiológicas), que es igual a 0,50 cuando $\alpha = 0$ y disminuye cuando α aumenta (en línea negra continua). Por el contrario, la proporción de pseudo-inercia explicada calculada con D_1 (las distancias geográficas) es igual a 0,87 cuando $\alpha = 1$ y disminuye cuando α disminuye (línea discontinua).

La obtención de la partición teniendo en cuenta las restricciones geográficas con la proporción normalizada de inercias explicadas en la parte inferior de la Figura 7.16 (es decir, $Q_0^*(P_K^\alpha)$ y $Q_1^*(P_K^\alpha)$, muestra el valor α que tiene como objetivo aumentar la contiguidad espacial, como se ve en detalle en la Tabla 7.2.

Tabla 7.2: Proporción normalizada de pseudo-inercias explicadas.

Valores alfa	Q_0**norm**	Q_1**norm**
α=0,16	0,57808701	0,81167272
α=0,17	0,61407565	0,82247147
α=0,18	0,62478207	0,77433402
α=0,19	0,67296737	0,73850413
α=0,20	0,54877711	0,84948899

El valor de α es una compensación entre la pérdida de homogeneidad socioeconómica y la ganancia de cohesión geográfica. Cuando $\alpha = 0$, las diferencias geográficas no se tienen en cuenta, pero, cuando $\alpha = 1$, son las distancias socioepidemiológicas las que no se tienen en cuenta; los conglomerados se obtienen únicamente con las distancias geográficas.

La gráfica 7.16 (abajo) parece sugerir elegir $\alpha = 0,17$, que corresponde a una pérdida de solo (1-0,61407565 = 38,59 %) de homogeneidad socioepidemiológica con un SIR de cada municipio, y aumento del 17,75 % en la homogeneidad geográfica.

La mayor cohesión geográfica de esta partición se puede ver en la Figura 7.17.

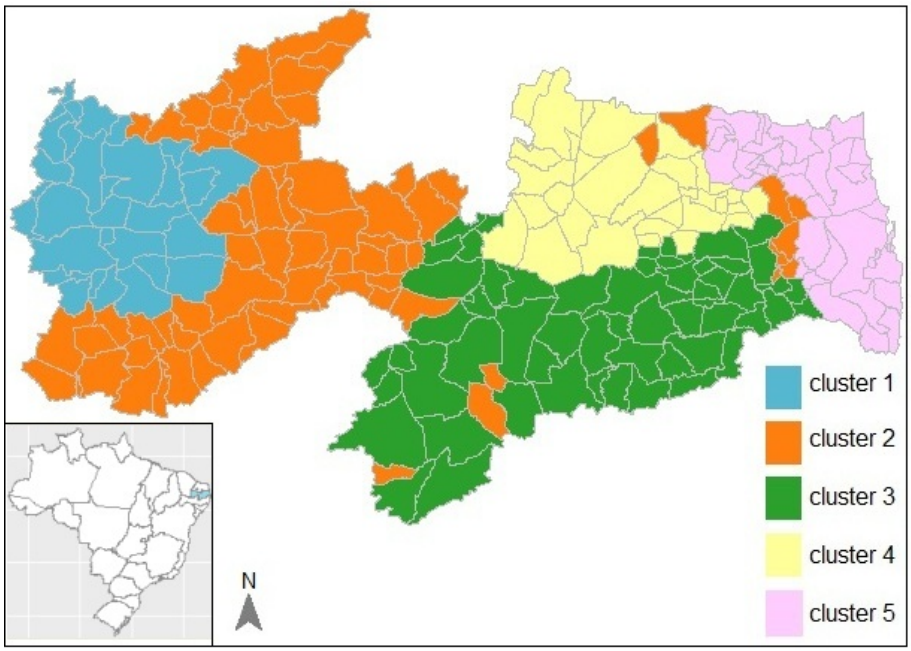

Figura 7.17: Mapa de la partición con $K = 5$ conglomerados basado en las distancias socioepidemiológicas D_0 y las distancias geográficas entre los municipios D_1 con $\alpha = 0,17$.

En la Figura 7.17, se percibe una ganancia significativa en la homogeneidad espacial. La Figura 7.20 muestra los diagramas de caja de las variables para cada conglomerado de la partición (fila central).

Los conglomerados 1, 3 y 4 parecen diferenciarse entre sí principalmente debido al ligero aumento en las muertes de pacientes masculinos en el conglomerado 4 y una mayor variación en la proporción de cura del conglomerado 3. El conglomerado 5 se diferencia del conglomerado 2 por el ligero aumento de las muertes, con mayores proporciones para los hombres, mientras que el conglomerado 2 tiene mayor número promedio de pacientes con tuberculosis en edad laboral y también mayor proporción de variación de cura.

La siguiente gráfica, Figura 7.18, muestra la elección de α para la partición,

teniendo en cuenta las restricciones de vecindad.

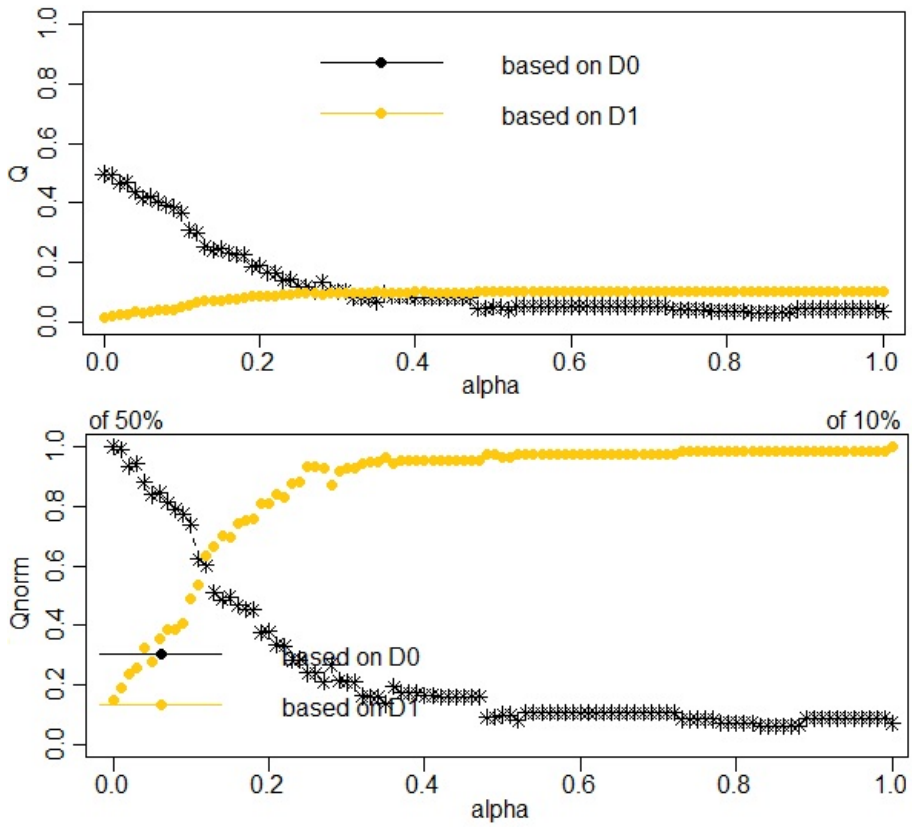

Figura 7.18: Elección de α para una partición en $K = 5$ conglomerados cuando D_1 es la matriz de disimilitud de vecindad entre municipios. (**Arriba**) proporción de pseudo-inercias explicadas $Q_0(P_K^\alpha)$ frente a α (en línea sólida negra) y $Q_1(P_K^\alpha)$ frente a α (en línea discontinua dorada). (**Abajo**) proporción normalizada de pseudoinercias explicadas $Q_0^*(P_K^\alpha)$ frente a α (en línea sólida negra) y $Q_1^*(P_K^\alpha)$ frente a α (en línea discontinua dorada).

En la parte inferior de la Figura 7.18, el gráfico de la proporción normalizada de inercias explicadas (es decir, $Q_0(P_K^\alpha)$ y $Q_1(P_K^\alpha)$) sugiere mantener $\alpha = 0,12$ favoreciendo ligeramente la homogeneidad socio-epidemiológica frente a la homogeneidad geográfica.

Sólo queda determinar esta partición final para $K = 5$ conglomerados y $\alpha = 0,12$. La Figura 7.19 proporciona el mapa correspondiente.

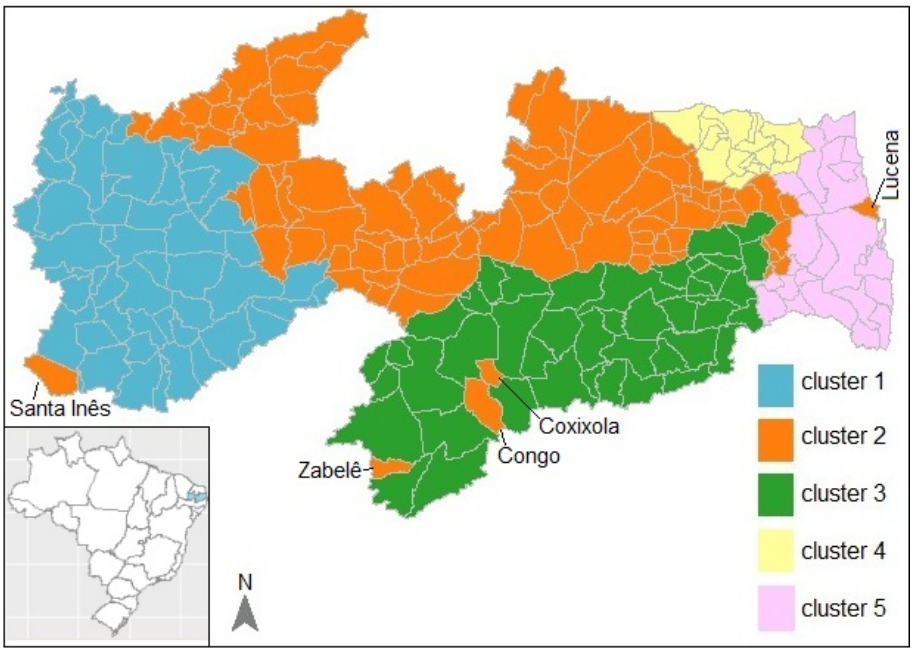

Figura 7.19: Mapa de la partición con $K = 5$ conglomerados basado en las distancias socio-epidemiológicas D_0 y las distancias de "vecindad" de los municipios D_1 con $\alpha = 0,12$.

La Figura 7.19 muestra que los conglomeradoss son espacialmente más compactos que los de la Figura 7.18. Sin embargo, se sabe que este enfoque crea divergencias en la matriz de adyacencia, lo que da más importancia a los vecinos.

Por lo tanto, como el enfoque se basa en restricciones de contigüidad suave, los municipios que no son vecinos pueden estar en la misma agrupación. Conforme ocurre con los municipios de Lucena, Coxixola, Congo, Zabelê y Santa Inés en el clúster 2. La calidad de la partición de la Figura 7.19 es ligeramente peor que la de la partición de la Figura 7.17, según el criterio Q_0 (61,41 % frente a 82,25 %).

La partición en términos de variables (socioepidemiológicas) de las Figuras 7.15, 7.17 y 7.19 puede verse detalladamente en el siguiente gráfico:

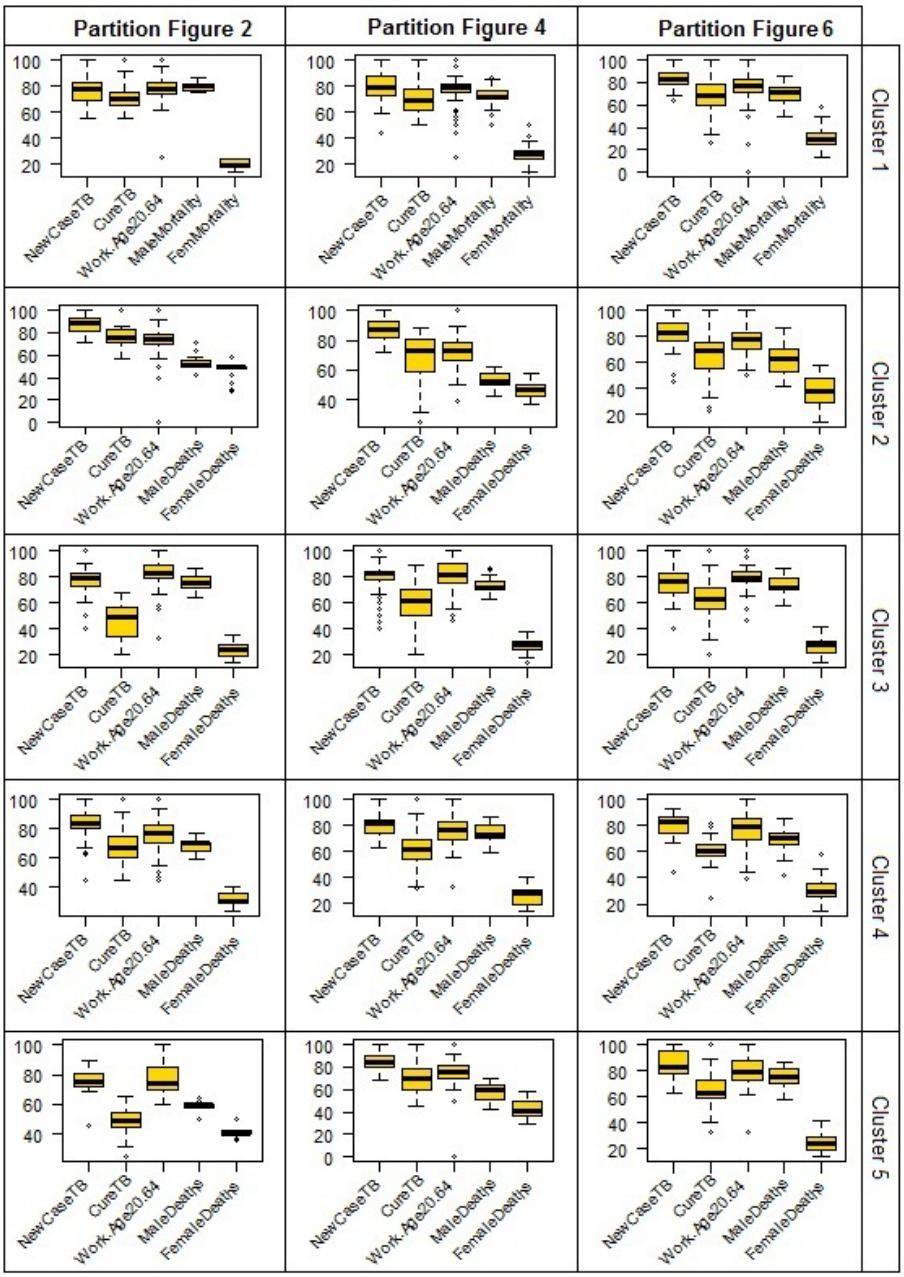

Figura 7.20: Comparación de conglomerados en la partición de las Figuras 7.15, 7.17 y 7.19 en términos de variables.

7.3.1. Conclusión

Al considerar las restricciones espaciales/geográficas, el conglomerado jerárquico se vuelve aún más completo para detectar patrones en conjuntos de datos de diferentes dimensiones. De acuedo con los pesos que se den a las diferencias geográficas en esta combinación, la solución tendrá conglomerados más o menos contiguos espacialmente.

A través del resultado de este estudio, los pesos no uniformes w definidos por la tasa de incidencia estandarizada (SIR) de la tuberculosis contribuyeron a aumentar la claridad tanto desde el punto de vista espacial como socioepidemiológico.

Capítulo 8

Consideraciones finales y trabajo futuro

La aplicación del método Ward-like se vuelve indispensable para comprender la realidad socio-epidemiológica del Estado de Paraíba desde una perspectiva espacial, facilitando así las decisiones en el desarrollo de políticas públicas y acciones sanitarias más eficaces en la lucha contra la tuberculosis.

El trabajo futuro sería añadir otras variables socio-epidemiológicas y, en lugar de los municipios, a fin de utilizar las Regiones Sanitarias que son responsables de la organización, planificación y ejecución de las acciones y servicios de salud en el estado de Paraíba.

Capítulo 9

Considerações finais e trabalho futuro

La aplicación del método Ward-like se vuelve indispensable para comprender la realidad socio-epidemiológica del Estado de Paraíba desde una perspectiva espacial, facilitando así las decisiones en el desarrollo de políticas públicas y acciones sanitarias más eficaces en la lucha contra la tuberculosis.

El trabajo futuro sería añadir otras variables socio-epidemiológicas y, en lugar de los municipios, a fin de utilizar las Regiones Sanitarias que son responsables de la organización, planificación y ejecución de las acciones y servicios de salud en el estado de Paraíba.

Bibliografía

[1] Pinker, S. In *How the Mind Works*; The Penguin Press: London, 1997. Citado en la p. 163.

[2] Everitt, Brian; Hothorn, Torsten. *An Introduction to Applied Multivariate Analysis with R*; Springer New York Dordrecht Heidelberg: London, 2011.

[3] Speece, D. L.; McKinney, J. D.; Appelbaum, M. I. Classification of behaviour subtypes of learning-disable children *Journal of Education Psychology*, **1985**, *77*, 67-77. [Crossref]

[4] Punj, G.; Stewart, D. W. Cluster analysis in market research: review and suggestions for application. *Journal of Marketing Research*, **1983**, *20*, 134-148. [Crossref].

[5] Song, Q. X.; Merajver, S. D.; Li, J. Z. Cancer classification in the genomic era: five contemporary problems. *Human Genomics*, **2015**, *9 (27)*, 475-483. Doi: 10.1186/s40246-015-0049-8 [Crossref]

[6] Parra-Sánchez, José H.; Cardona-Rivas, Dora; Cerezo-Correa, María del Pilar. Análisis de conglomerados para el estudio de las desigualdades sociales por enfermedades cardiovasculares. *Rev. Salud Pública.*, **2017**, *19 (4)*, 475-483. [Crossref]

[7] Helmuth Spath. *Cluster analysis algorithms for data reduction and classification of objects*. John Wiley & sons, New York, 1980.

[8] Everitt, B.S. *Cluster Analysis*. Edward Arnold, London, 1992.

[9] Encinas, Luis Hernándes. In *Técnicas de taxonomía numérica*; La Muralla, S.A.: Madrid, 2001.

[10] Lattin, James; Carroll, J. D.; Green, P. E. In *Análise de dados multivariados*; Cengage Learning: São Paulo, 2011.

[11] Santos Neto. M.; Sousa, M. R.; da Silva, F. B. G.; Santos, F. S.; Ferreira, A. G. N.; Pascoal, L. M.; Costa, A. C. P. d.; Bezerra, J. M.; Serra, M. A. A. d. O.; Dias, I. C. C. M.; et al. Spatial distribution of tuberculosis cases in a priority Brazilian northeast municipality for control of the disease. *Int. J. Dev. Res.* **2017**, *7*, 10611. [Crossref]

[12] Gomathi, V. V.; S. Karthikeyan, 2014. Performance analysis of distance measures for computer tomography image segmentation. *Int. J. Comput. Technol. Applic.* **2014**, 5: 400-405. [Google Académico].

[13] Peña, D. *Analisis de datos multivariantes*. McGraw-Hill Interamericana de España/UNED: Madrid, 2002.

[14] Strauss, T.; von Maltitz, M. J. Generalising Ward's Method for Use with Manhattan Distances. *PLoS ONE* **2017**, *12*, e0168288. [Crossref].

[15] Peres, Sarajane Marques; Rocha, Thiago; Biscaro, Helton H.; Madeo, Renata Cristina B.; Boscarioli, Clodis. Tutorial sobre Fuzzy-c-Means e Fuzzy Learning Vector Quantization: Abordagens Híbridas para Tarefas de Agrupamento e Classificação. *Revista de Informática Teórica e Aplicada*. **2012**, *19(1)*, 120. [Crossref].

[16] López, César Pérez. *Métodos estadísticos avanzados con SPSS*. Thomson: Madrid, 2005; pp. 444.

[17] da Silva, Leandro Augusto; Peres, Sarajane Marques; Boscarioli, Clodis. In *Introdução à Mineração de Dados: com aplicações em R*. Elsevier: Rio de Janeiro, 2016.

[18] Hair, J. F.; Anderson, R. E.; Tatham, R. L.; Black, W. C. In *An Introduction to Applied Multivariate Analysis with R*; Ed. Prentice Hall, 1999.

[19] Manzano, Joaquin Aldas; Jimenez, Ezequiel Uriel. In *Analisis Multivariante Aplicado*; Thomson: Madrid, 2006.

[20] Sharma, Subhash. In *Applied Multivariate Methods Techniques*; John Wiley & Sons: New York, 1996.

[21] Johnson, Dallas E. In *Applied Multivariate Methods for Data Analysis*; Brooks Cole Publishing: New York, 1998.

[22] Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy: The principles and practice of numerical classification.*; Freeman W. H. and Co.: San Francisco, USA. 1973; 573p.

[23] King, B. Step-wise Clustering Procedures. *J. Am. Stat. Assoc.* **1967**, 69, pp. 86-101. [Google Académico].

[24] Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **1963**, *58(301)*, 236-244. [Crossref].

[25] Murtagh, F. Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly*, **1984**, *1*, 101–113. [Google Académico].

[26] Gower, J. C. *Measures of similarity, dissimilarity and distance*. Encyclopedia of Statistics. Vol 5. Johnson, NL, Kotz, S, Read CB (eds). Wiley: New York, 1985.

[27] Mucha, H. J. & Sofyan, H. *Cluster Analysis*. Sonderforschungsbereich 373, Discussion Paper 2000-49, Humboldt University at Berlin, 2000.

[28] Silva Camêlo, E. L.; Lisboa, P. J. G.; Sánchez, R. G.; Aguiar, D. C. *Principais técnicas de agrupamento: com aplicações em R*. Campina Grande: EDUEPB, 2020.

[29] Wolfson, M.; Zagros, M.; James, P. dentifying national types: a cluster analysis of politics, economics and conflict. *Journal of Peace Research*. **2004**, *41(5)*, pp. 607-623. [Google Académico].

[30] Gutiérrez, R.; González, A.; Torres, F.; Gallardo, J. A. *Técnicas de análisis de datos multivariables*.Tratamiento computacional. Servicio de Reprografía de la Facultad de Ciencias. Universidad de Granada: España, 1994.

[31] Gallardo, J. Métodos jerárquicos de análisis cluster *Curso de Diplomatura Estadística Teórico Práctico*; Universidad de Granada: Granada, **2011**. Disponible en: `https://www.ugr.es/~gallardo/pdf/cluster-3.pdf` [consulta: julio de 2020].

[32] Chavent, M.; Kuentz-Simonet, V.; Labenne, A.; Saracco, J. ClustGeo: An R package for hierarchical clustering with spatial constraints. *Comput. Stat*. **2018**, *33*, 1799–1822. [Crossref].

[33] Murtagh, F. In *Multidimensional clustering algorithms*; Compstat Lectures: Vienna, 1985a.

[34] Legendre, P. const.clust: Space-and Time-Constrained Clustering Package. R Package Version 1.2. 2011. Disponible en: `http://adn.biol.umontreal.ca/numericalecology/Rcode/` [consulta: mayo de 2020]. [Google Académico]

[35] Bécue-Bertaut, M.; Alvarez-Esteban, R.; Sánchez-Espigares, J.A. *Xplortext: Statistical Analysis of Textual Data R Package*. R Package Version 1.0. 2017. Version 0.9.9. Disponible en: `https://cran.r-project.org/package=Xplortext` [consulta: junio de 2020]. [Google Académico]

[36] Lance, G. N.; Williams, W. T. A General Theory of Classicatory Sorting Strategies. *The Computer Journal*.1967. Disponible en: `https://biocomparison.ucoz.ru/_ld/0/51_Lance_Willams_2.pdf` [consulta: junio de 2020]. []

[37] Ambroise, C.; Dang, M.; Govaert, G. Clustering of Spatial Data by the EM Algorithm. In *geoENV I-Geostatistics for Environmental Applicattions*; Soares, A.O., Gomez-Hernandez, J.J., Froidevaux, R., Eds.; Kluwer: Dordrecht, the Netherland, 1997; pp. 493–504. [Google Académico].

[38] Ambroise, C.; Govaert, G. Convergence of an EM-type algorithm for spatial clustering. *Pattern Recognit. Lett.* **1998**, *19*, 919–927. [Google Académico].

[39] Duque, J. C.; Dev, B.; Betancourt, A.; Franco, J. L. *ClusterPy: Library of Spatially Constrained Clustering Algorithms*. RiSE-Group (Research in Spatial Economics). 2011. Version 0.9.9. Disponible en: `http://www.rise-group.org/risem/clusterpy/` [consulta: mayo de 2020]. [Google Académico]

[40] Dehman, A.; Ambroise, C.; Neuvial, P. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinform.* **2015**, *16*, 148. [Google Académico].

[41] Aguiar, D. C.; Sánchez, R. G.; Silva Camêlo, E. L. Hierarchical clustering with spatial constraints in tuberculosis data. *IJDR* **2020**, *10*, 35374–35380. [Google Académico].

[42] Aguiar, D. C.; Sánchez, R. G.; Silva Camêlo, E. L. Ward-like hierarchical clustering with dissimilarities and non-uniform weights in cases of tuberculosis in Paraíba, Brazil. *IJDR*, **2020**, *10*, 35478–35483. https://www.journalijdr.com/ward-hierarchical-clustering-dissimilarities-and-non-uniform-weights-cases-tuberculosis-para %C3 %ADbaGoogle Académico].

[43] Chavent, M.; Kuentz-Simonet, V.; Labenne, A.; Saracco, J. *ClustGeo: Hierarchical Clustering with Spatial Constraints*. R Package Version 2.0. 2017. Disponible

en: https://CRAN.R-project.org/package=ClustGeo [consulta: mayo de 2020].

[44] Chavent, M.; Kuentz-Simonet, V.; Labenne, A.; Saracco, J. ClustGeo: An R package for hierarchical clustering with spatial constraints. *Comput. Stat*. **2018**, *33*, 1799–1822. [Crossref].

[45] González, F. P.; Céspedes, J. C. *Técnicas cuatitativas para el análisis regional*. Editorial Universidad de Granada: España, 2004.

[46] Zaiger, D. Inequalities for the Gini coefficient of composite populations. *Journal of Mathematical Economics*, **1983**, *12*. [Google Académico].

[47] Miele, V.; Picard, F.; Dray, S. Spatially constrained clustering of ecological networks. *Methods Ecol. Evol.* **2014**, *5*, 771–779. [Crossref].

[48] Wierzchoń, S.T.; Kłopotek, M.A. Cluster Analysis. *Modern Algorithms of Cluster Analysis*; Janusz, K., Ed.; Springer International Publishing AG: Cham, Switzerland, 2018; pp. 9–66.

[49] 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. https://portalarquivos2.saude.gov.br/images/pdf/2019/marco/22/2019-009.pdf [consulta: abril de 2020].

[50] Camêlo, Edwirde Luiz Silva; Aguiar, Dalila Camêlo; da Silva, Rosiane Davina; Figueiredo, Tânia Maria Ribeiro Monteiro de; Carmona, Andrés Gonzaléz; Sánchez, Ramón Gutiérrez. Tuberculosis in Brazil: New Cases, Healing and Abandonment in Relation to level of Education. *International Archives Of Medicine*, **2016**, *9(68)*, 1-9. [Crossref].

[51] Camêlo Aguiar, D.; Gutiérrez Sánchez, R.; Silva Camêlo, E. L. Hierarchical

Clustering with Spatial Constraints and Standardized Incidence Ratio in Tuberculosis Data. *Mathematics*, **2020**, *8*, 1478. [Crossref].

[52] SINAN. Sistema de Informação de Agravos de Notificação. In *Tuberculose–Casos Confirmados*; Ministério da Saúde, Brazil: Brasília, Barzil, 2020. Available online: http://www2.datasus.gov.br/ (accessed on 5 April 2020).

[53] R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Disponible: https://www.R-project.org/ [consulta: mayo de 2020].

[54] Wickham, Hadley; François, Romain; Henry, Lionel; Müller, Kirill. 2019. *Dplyr: A Grammar of Data Manipulation*. Disponible: https://CRAN.R-project.org/package=dplyr [consulta: mayo de 2020].

[55] Moraga, Paula. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC Biostatistics Series, 2019.

[56] CRF. *Reference points and distance computations*. Code of Federal Regulations (AnnualEdition). Title 47: Telecommunication. 73 (208), 2016.

[57] Wallace, John R. 2012. *Imap: Interactive Mapping*. R package version 1.32. Disponible: [consulta: mayo de 2020].

[58] Majure, J. J.; Gebhardt, A. 2016. *sgeostat: an object-oriented framework for geostatistical modeling in S+*. R package version 1.0-27. Disponible: [consulta: mayo de 2020].

[59] Hijmans, Robert J. 2019. *geosphere: Spherical Trigonometry*. R package version 1.5-10. Disponible: [consulta: mayo de 2020].

[60] Filzmoser, P.; Ruiz-Gazen, A.; Thomas-Agnan, C. Identification of local multivariate outliers. *Stat Pap*, **2014**, *55(1)*, 29-47. [Crossref].

[61] Filzmoser, P.; Gschwandtner, M. 2018. *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*. R package version 2.0.9. Disponible: [consulta: abril de 2020].

[62] Wickham, Hadley; François, Romain; Henry, Lionel; Müller, Kirill. 2020. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2. Disponible: [consulta: abril de 2020].

[63] Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. Disponible: [consulta: abril de 2020].

[64] Wickham, Hadley. Welcome to the tidyverse. *Journal of Open Source Software*, **2019**, *4(43)*, 1686. [Crossref].

9

Código fuente

```
# Artículo 3 - Conglomerado jerárquico Ward-like con restricciones
# espaciales y tasa de incidencia estandarizada

library(ClustGeo)
datPBm<-read.csv("TFD2020/datos.csv",dec=",",
     header = TRUE, sep=";", encoding="latin")
head(datPBm,4)
summary(datPBm)

# Variables socioepidemiologica
datPBmr<-datPBm[1:223, c(8,9,10,30,31), drop = FALSE]

## transformando en numérico
datPBmr[] <- lapply(datPBmr, function(x) as.numeric(as.character(x)))
str(datPBmr)
head(datPBmr,3)

boxplot(datPBmr)

## ------------------------------------------------
library(ClustGeo)
D0<- dist(datPBmr, method="manhattan")
```

```
D0 # distancia entre las variables socioepidemiologicas
n <- nrow(datPBmr); n

## ------------------------------------------------
library(rgdal)
D.mapPB<-readOGR(dsn="C:/ TFD2020/datos1",layer="PB_Mun97_region")
#  D.mapPB  = D.geo

## ------------------------------------------------
# wt recebe RIE - Razón de incidencia estandarizada
datPBr_SIR<-datPBm[1:223, 27 , drop = FALSE] # SIR 27
boxplot(datPBr_SIR, horizontal = TRUE)
head(datPBr_SIR,4)
wt1<- datPBr_SIR; wt1

## ------------------------------------------------
tree <- hclustgeo(D0, wt=wt1) # Recibiendo el RIE
plot(tree)
# Dividindo D^2/2n por el método de ward.D
###tree <- hclust(D^2/(2*n), method="ward.D")

## ------------------------------------------------
## si resulta que la suma de las alturas en el dendrograma
## es igual a la pseudo-inercia total del conjunto de datos:
inertdiss(D0, wt=wt1) # la pseudo-inercia de los datos
sum(tree$height) # Suma de la altura (dendrograma) de cada dendrograma

## ------------------------------------------------
# Delta reciebe la distancia de RIE
w<-wt1 # distancia de RIE
```

```
w==wt1

Delta <- D0
Delta
for (i in 1:(n-1)) {
  for (j in (i+1):n) {
    Delta[n*(i-1) - i*(i-1)/2 + j-i] <-
      Delta[n*(i-1) - i*(i-1)/2 + j-i]^2*w[i]*w[j]/(w[i]+w[j])
  }}

## ------------------------------------------------
## Geos - matriz D0 = socioepidemiologica
## Elección del número K de clústeres

tree <- hclustgeo(D0)

plot(tree,hang=-1,label=FALSE, xlab="",sub="",
main="Ward dendrogram with D0 only",cex.main=0.7,cex=0.7,
cex.axis=0.7,cex.lab=0.7)
rect.hclust(tree,k=5,border=c(4,5,3,2,1))
legend("topright", legend= paste("cluster",1:5), fill=1:5, cex=0.7,
bty="n",border="white")

## ------------------------------------------------------------------------
k<-5 # cut the dendrogram to get the partition in 5 conglomerados
P5 <- cutree(tree,k) # cut the dendrogram to get the partition in
# 5 conglomerados
sp::plot(D.mapPB, border="grey", col=P5) # plot an object of class sp
legend("bottomright", legend=paste("cluster", 1:k), fill=1:k, bty="n",
border="white", cex = 0.6)
```

```
table(P5)

## ------------------------------------------------------------------------
## Calculo de las distancias geograficas en km entre los municipios
# - matriz D1
M<-read.table("C:/ TFD2020/M.txt", header = TRUE, sep=";",
encoding="latin");M

library(geosphere)
library(Imap)

## Encontrar las distancias entre los municipios a partir de las
## coordenadas centrales
library(sgeostat)
ReplaceLowerOrUpperTriangle <- function(m, triangle.to.replace){
  if (nrow(m) != ncol(m)) stop("Supplied matrix must be square.")
  if      (tolower(triangle.to.replace) == "lower") tri <- lower.tri(m)
  else if (tolower(triangle.to.replace) == "upper") tri <- upper.tri(m)
  else stop("triangle.to.replace must be set to 'lower' or 'upper'.")
  m[tri] <- t(m)[tri]
  return(m)
}

GeoDistanceInMetresMatrix <- function(df.geopoints){
  # Returns a matrix (M) of distances between geographic points.
  # M[i,j] = M[j,i] = Distance between (df.geopoints$lat[i],
  df.geopoints$lon[i])
  # (df.geopoints$lat[j], df.geopoints$lon[j]).
  # The row and column names are given by df.geopoints$name.
```

```
  GeoDistanceInMetres <- function(g1, g2){
    # Returns a vector of distances. (But if g1$index > g2$index, returns
zero.)
    # The 1st value in the returned vector is the distance between g1[[1]]
    # and g2[[1]].
    # The 2nd value in the returned vector is the distance between g1[[2]]
    # and g2[[2]]. Etc.
    # Each g1[[x]] or g2[[x]] must be a list with named elements "index",
"lat"
    # and "lon".
    # E.g. g1 <- list(list("index"=1, "lat"=12.1, "lon"=10.1),
    list("index"=3,
    # "lat"=12.1, "lon"=13.2))
    DistM <- function(g1, g2){
      require("Imap")
      return(ifelse(g1$index > g2$index, 0, gdist(lat.1=g1$lat, lon.1=g1$lon,
      lon.2=g2$lon, units="m")))
    }
    return(mapply(DistM, g1, g2))
  }

  n.geopoints <- nrow(df.geopoints)
  # La columna de índice se utiliza para asegurarnos de que solo
  # hacemos cálculos para el triángulo superior de puntos.
  df.geopoints$index <- 1:n.geopoints

  # Create a list of lists
  list.geopoints <- by(df.geopoints[,c("index", "lat", "lon")],
  1:n.geopoints, function(x){return(list(x))})
```

```
  # Get a matrix of distances (in metres)
  mat.distances <- ReplaceLowerOrUpperTriangle(outer(list.geopoints,
      list.geopoints, GeoDistanceInMetres), "lower")

 # Set the row and column names
  rownames(mat.distances) <- df.geopoints$Locality
 colnames(mat.distances) <- df.geopoints$Locality
  return(mat.distances)
}

## ------------------------------------------------------------------------
## DM = la matriz D1
DM<-GeoDistanceInMetresMatrix(M) # Matriz distancia 223 x 223
DM<-DM/1000
DM<-round(DM,2); DM
dim(DM)  # 223  filas y 223 colunas
D1<-as.dist(DM)
k=5
# lista de los municipios del conglomerado 5
city_label <- as.vector(D.mapPB$"MUNIC_PIO")
city_label[which(P5 == 5)]  # Varia de 1 a 5 clustering
table(P5)

## ------------------------------------------------------------------------
cr <- choicealpha(D0, D1, range.alpha=seq(0, 1, 0.01), K=k, graph=TRUE)
cr
cr$Q # proporción de pseudo-inercia explicada
## alpha=0.19 0.33416683 0.64205719
## alpha=0.2  0.27249925 0.73854769 ## lo mejor
## alpha=0.21 0.24700761 0.75514687
```

```
## alpha=0.22 0.25018156 0.74815835
cr$Qnorm # proporción normalizeda de pseudo-inercia explicada
## alpha=0.16 0.57808701 0.81167272
## alpha=0.17 0.61407565 0.82247147 ## lo mejor
## alpha=0.18 0.62478207 0.77433402
## alpha=0.19 0.67296737 0.73850413
## alpha=0.2  0.54877711 0.84948899
## alpha=0.21 0.49744034 0.86858162
## alpha=0.22 0.50383225 0.86054332
## alpha=0.23 0.47787959 0.87914497

## ------------------------------------------------------------------------
treeD0 <- hclustgeo(D0, D1, alpha=0.27)

P5bis <- cutree(treeD0, k=k)
sp::plot(D.mapPB, border="grey", col=P5bis)
legend("bottomright", legend=paste("cluster", 1:k), fill=1:k, bty="n",
border="white", cex = 0.6)

city_label[which(P5bis == 5)]


## ------------------------------------------------------------------------
## 4.4 Obtención de una partición teniendo en cuenta las
## restricciones de vecindad

list.nb <- spdep::poly2nb(D.mapPB,
                row.names=rownames(datPBmr)) # lista de vecinos
list.nb
```

```
## lista los municipios en la posición de acuerdo con el shape en el
mapa
D.mapPB$MUNIC_PIO
## 116 - Campina grande
## 132  - Gado Bravo

city_label[list.nb[[1]]] # lista de los vecinos de Gado Bravo y
# Campina Grande
# La matriz de disimilitud D1 se construye a partir de la matriz de
# adyacencia
# matrix A con D1=1n/A

A <- spdep::nb2mat(list.nb, style="B"); A # construye la matriz de adyacencia
# con el nombre de los municipios
diag(A) <- 1
colnames(A) <- rownames(A) <- city_label
D1 <- 1-A
length(D1)
## D1[1:2, 1:5]
D1 <- as.dist(D1)

## ------------------------------------------------------------------------
## Elección del parámetro de mezcla alpha
cr <- choicealpha(D0, D1, range.alpha=seq(0, 1, 0.01), K=k, graph=TRUE)
cr$Q # proporción de pseudo-inercia explicada
cr$Qnorm # proporción normalizeda de pseudo-inercia explicada

## Partición final obtenida con alpha = 0.2 y 0.36
treeF <- hclustgeo(D0, D1, alpha=0.36)
P5ter <- cutree(treeF, k)
```

```
sp::plot(D.mapPB, border="grey", col=P5ter)
legend("bottomright", legend=paste("cluster", 1:5),
       fill=1:5, bty="n", border="white", cex = 0.6)

## ------------------------------------------------------------------------
## ------------------------------------------------------------------------
# Diagrama de cajas

library(ggplot2)
library(dplyr)
library(tidyverse)  # manipulacion de datos

datPBm<-read.csv("C:/ TFD2020/datos.csv", dec=",", header = TRUE,
sep=";", encoding="latin")
head(datPBm,4)

# Variables socioepidemiologicas
datPBmr<-datPBm[1:223, c(8,9,10, 30,31), drop = FALSE]
head(datPBmr,12)

## transformando en numérico
datPBmr[] <- lapply(datPBmr, function(x) as.numeric(as.character(x)))
str(datPBmr)
head(datPBmr,3)

### P5a - partición P5a p/ los 5 conglomerados
M<-datPBmr
M[,6]<-P5a
head(M)
attach(M)
```

```
par(mfrow=c(3,5))

A1<-M %>%
  filter(V6 == "1") %>%
  select(NewCaseTB, CureTB, Work.Age20.64,  MaleDeaths, FemaleDeaths)%>%
  ungroup()
boxplot(A1, horizontal = FALSE, las=2, col=(c("gold")) )

A2<-M %>%
  filter(V6 == "2") %>%
  select(NewCaseTB, CureTB, Work.Age20.64, MaleDeaths,FemaleDeaths)%>%
  ungroup()
boxplot(A2, horizontal = FALSE, las=2, col=(c("gold")) )

A3<-M %>%
  filter(V6 == "3") %>%
  select(NewCaseTB, CureTB, Work.Age20.64, MaleDeaths, FemaleDeaths)%>%
  ungroup()
boxplot(A3, horizontal = FALSE, las=2, col=(c("gold")) )

A4<-M %>%
  filter(V6 == "4") %>%
  select(NewCaseTB, CureTB, Work.Age20.64,  MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A4, horizontal = FALSE, las=2, col=(c("gold")) )

A5<-M %>%
  filter(V6 == "5") %>%
```

```
  select(NewCaseTB, CureTB, Work.Age20.64,  MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A5, horizontal = FALSE, las=2, col=(c("gold")) )

### P5bis - partición P5bis p/ los 5 conglomerados
M<-datPBmr
M[,6]<-P5bis
head(M)
attach(M)

par(mfrow=c(3,5))

library(ggplot2)
A1<-M %>%
  filter(V6 == "1") %>%
  select(NewCaseTB, CureTB, Work.Age20.64,  MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A1, horizontal = FALSE, las=2, col=(c("gold")) )

A2<-M %>%
  filter(V6 == "2") %>%
  select(NewCaseTB, CureTB, Work.Age20.64, MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A2, horizontal = FALSE, las=2, col=(c("gold")) )

A3<-M %>%
  filter(V6 == "3") %>%
```

```
  select(NewCaseTB, CureTB, Work.Age20.64, MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A3, horizontal = FALSE, las=2, col=(c("gold")) )

A4<-M %>%
  filter(V6 == "4") %>%
  select(NewCaseTB, CureTB, Work.Age20.64,  MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A4, horizontal = FALSE, las=2, col=(c("gold")) )

A5<-M %>%
  filter(V6 == "5") %>%
  select(NewCaseTB, CureTB, Work.Age20.64,  MaleDeaths,
  FemaleDeaths) %>%
  ungroup()
boxplot(A5, horizontal = FALSE, las=2, col=(c("gold")) )


## ------------------------------------------------------------------------
## ------------------------------------------------------------------------
# Valores atípicos
library(mvoutlier)
P1<-datPBm[,c(8,9,10,30,31)]
dim(P1)
# res <- aq.plot(datPBm[,c(8,9,10,30,31)], quan=1, alpha=0.01)
res <- aq.plot(datPBm[,c(8,9,10,30,31)],
               quan=1, alpha=0.05)
```

9

Artículos publicados en revistas científicas

A continuación, tenemos los tres artículos publicados en revistas científicas.

Artículo 3

Hierarchical Clustering with Spatial Constraints and Standardized Incidence Ratio in Tuberculosis Data

El tercer artículo *Hierarchical Clustering with Spatial Constraints and Standardized Incidence Ratio in Tuberculosis Data* fue publicado en la revista *Mathematics* 2020, 8, 1478, editada por el *Multidisciplinary Digital Publishing Institute* (MDPI) de acceso abierto, con indicador de calidad (cuartile Q1 y factor de impacto 1,747). doi.org/10.3390/math8091478

mathematics

MDPI

Article

Hierarchical Clustering with Spatial Constraints and Standardized Incidence Ratio in Tuberculosis Data

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

[1] Department of Statistics and Operational Research, Faculty of Science, University of Granada, Avda. Fuentenueva, S/N, 18071 Granada, Spain; ramongs@ugr.es

[2] Department of Statistics, State University of Paraíba, Rua Baraúnas, 351—Bairro Universitário, Campina Grande 58429-500, Brazil; edwirde@uepb.edu.br

* Correspondence: dalilacamelo@correo.ugr.es

Received: 26 June 2020; Accepted: 27 August 2020; Published: date

Abstract: In this paper, we propose presenting a solution based on socio-epidemiological variables of tuberculosis, considering a clustering with spatial/geographical constraints; and, determine a value of alpha that increases spatial contiguity without significantly deteriorating the quality of the solution based on the variables of interest, i.e. those of the feature space. 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 constraints space 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 standardized incidence ratio (SIR) of TB and that contributed significantly to the increase in clarity, both from a spatial and socio-epidemiological point of view. The method is shown to be feasible in epidemiological studies in the joint understanding of factors of different dimensions, aggregated from a spatial perspective. It is analysis tool that allows making a better understanding of the socio-epidemiological reality of the municipality.

Keywords: ward-like algorithm; spatial constraints; measure of risk; Tuberculosis; State of Paraíba, Brazil

1. Introduction

In exploratory data analysis, the statistician often uses clustering and visualization to improve his knowledge of the data [1]. In the viewing, he looks for some clusterings explaining some of the significant characteristics of the data.

The cluster analysis goal consists of distinguishing, in the data set to be analyzed, the groups, called clusters. In this paper, we study the hierarchical clustering (and not partitioning). Hierarchical cluster algorithm groups the data based on the distance between each one and looks for data within a cluster to be the most similar to each other. These groups are disjoint subsets of the data set. They have such a property that data that belong to different clusters differ among themselves much more than the data on the same cluster [2]. The difficulty of choosing the clustering method for grouping a set of n objects into k separate sets and the ideal number of clusters is well frequent among researchers. In Tuberculosis (TB) epidemiology, for instance, this challenge is excellent 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 [3]. Contiguity constraints (in space or time) are the most common; they occur when the objects in a cluster required not only to be similar to one other, but also to comprise a contiguous set of objects.

TB still poses a substantial global health threat, with some 10-million new cases per year. In Brazil, the estimate is that the incidence of TB is increasing after many years of decline due to the upward trend in the period of 2016–2018 [4]. TB incidence is disproportionately high among people in poverty [5]. The goal set by the World Health Organization (WHO) is to cure 85% of new bacilliferous TB cases by 2020 [6]; however, as observed in the 2018 data, Brazil (71.4%) falls short of reaching this goal [7]. In the State of Paraíba, the situation is even more critical [8] 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 Brazil's states with more than 4.018 million inhabitants according to 2019 estimates by the Brazilian Institute of Geography and Statistics [9].

The relationship between TB and social conditions demands an understanding of the dynamics of this aggravation and its occurrence in the territory [10]. This study aims to present a solution based on socio-epidemiological variables of TB with results being easily visualized on a map while using the ClustGeo package. This method uses Ward-like hierarchical clustering with non-Euclidean dissimilarities and non-uniform weights attributed to the standardized incidence ratio (SIR) of TB in the 223 municipalities of Paraíba and the importance of the constraint in the clustering procedure through the parameter α, responsible for controlling the weight of the constraint in the quality of the solution on the variables of interest.

Sometimes we wish to provide disease risk estimates in each of the areas that form partitions of the study region [11]. For instance, to identify changes in morbidity and/or mortality in time or to compare the incidence or prevalence. The standardized incidence ratio (SIR) is one simple measure of disease risk.

2. Material and Methods

2.1. 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 [12] and made available 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, divided into epidemiological (new cases, cure, male and female deaths) and social variable (active age (20–64) patients with TB). A matrix was also calculated with the geographic distances between the municipalities and the weight w non-uniform attributed to the calculation of the dissimilarity matrix D, as being the standardized incidence ratio (SIR) of TB in the State of Paraíba.

The data were collected between February–May 2020. Statistical analyses were undertaken in R version 3.6.2 [13]. This study was not submitted to the Research Ethics Committee's evaluation as it is a survey of secondary data and does not directly involve human beings.

2.2. Constrained hierarchical clustering

Usually, the researcher has difficulty of clustering a set of n objects into k disjoint clusters. Soon, many methods proposed finding 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.

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 ith and the jth objects are contiguous, and 0 if they are not [3]. An adjacency matrix used to find a connection between the borders of each municipality in the State of Paraíba. Accordingly, two clusters are regarded as contiguous if there are two

objects, one from each cluster, which is linked in the contiguity matrix. Several authors in different areas of knowledge have implemented of constrained clustering procedures [14–21]. For instance, Miele et al. [22] proposed a model-based spatially constrained method that embeds the geographical information within an EM regularization framework by adding some constraints to the maximum likelihood estimation of parameters. It is a partitioning method with neighbourhood constraints, while the Ward-like method [3] is a hierarchical clustering (and not partitioning) method, including spatial/geographical constraints (not necessarily neighbourhood constraints) [23].

2.3. Ward-Like hierarchical clustering

With algorithm similar to Ward, Ward-like is a constrained hierarchical clustering algorithm that optimizes the convex combination $D_\alpha = (1-\alpha)D_0 + \alpha D_1$ 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 $D_0 = [d_{0,ij}]$ is constructed from the Manhattan distance matrix between the 223 municipalities performed with the $p = 5$ variables socio-epidemiological, i.e., the matrix gives the differences in the feature space, and the dissimilarity matrix $D_1 = [d_{1,ij}]$ is constructed from the geographical distance between the 223 municipalities, i.e., the matrix D_1 gives 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 parameter α (the weight of this convex combination) gives the relative importance of D_0 as compared to D_1. This parameter controls the weight of the constraint on the quality of the solutions, i.e., for a given value of $\alpha[0;1]$, the mixing parameter α clearly controls the part of pseudo-inertia due to D_0 and D_1. The mixed pseudo inertia of the cluster C_k^α is defined as:

$$I_\alpha(C_k^\alpha) = (1-\alpha)\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{0,ij}^2 + {}^\alpha\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{1,ij}^2 \quad (1)$$

where $\mu_k^\alpha = \sum_{i\in C_k^\alpha} w_i$ is the weight of C_k^α; $d_{0;ij}$ and $d_{1;ij}$ are the normalized dissimilarity between observations i and j in D_0 and D_1, respectively. For the choice of α we will use two types of spatial constraints (geographical distances and neighborhood contiguity). For the last case, the dissimilarity matrix D_1 will be constructed from the corresponding adjacency matrix A, i.e., $D_1 = 1_n - A$ with $1_{n,ij} = 1\forall(i,j)$, a_{ij} equal to 1 if municipalities i and j are neighbourhood 0 otherwise, $a_{ii} = 1$ by convention. When α increases, the homogeneity that is calculated with D_0 decreases; conversely, the homogeneity calculated increases with D_1. Therefore, the idea is to determine a value of α, which increases the spatial the geographical homogeneity without deteriorating the quality of the solution on the variables of interest too much.

These homogeneities are measurable using the appropriate pseudo within-cluster inertias. To determine a suitable value for the mixing parameter α, let us assume that the dissimilarity matrix D_1 contains geographical distances between n municipalities, whereas the dissimilarity matrix D_0 contains distances that are based on a $n \times p_0$ data matrix X_0 of p_0 socio-epidemiologic variables measured on these n municipalities. Basically, the notion of proportion of the total mixed (pseudo) inertia explained by the partition P_K^α in K clusters is $Q_\beta(P_K^\alpha) = 1 - \frac{W_\beta(P_K^\alpha)}{W_\beta(P_1)} \in [0,1]$. When $\beta = 0$, the denominator $W_0(P_1)$ is the (pseudo) total inertia, and the numerator is the (pseudo) within-cluster inertia $W0(P_K^\alpha)$, both based on the D_0 dissimilarity matrix. Therefore, the higher the value of the $Q_0(P_K^\alpha)$ criterion, the more homogeneous is the P_K^α partition from the socio-epidemiological point of view; $\beta = 1$, the denominator $W_1(P_1)$ is the total inertia (pseudo) and the numerator is the (pseudo) inertia within the cluster $W_1(P_K^\alpha)$, both based on the D_1 dissimilarity matrix. Ergo, the higher the value of criterion $Q_1(P_K^\alpha)$, the more homogeneous is the partition P_K^α from the geographical point of view. When β assumes a value of $\beta \in]0,1[$, the denominator $W_\beta(P_1)$ is a

total mixed (pseudo) inertia, and it is not easy to interpret in practice and the numerator $W_\beta(P_K^\alpha)$ is the mixed (pseudo) inertia within the cluster.

With R package *ClustGeo* (version 2.0) developed by Chavent et al. [3], it is possible to implement this hierarchical clustering algorithm with geographical constraints and choose the mixing parameter α provided with two types of spatial constraints (geographical distances and neighbourhood contiguity). Let w_i be the weight of the ith observation for $i = 1, \ldots, n$. 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 is a wrapper of the usual function *hclust*. It performs the hierarchical clustering of *Ward.D*, using a dissimilarity matrix D (which is an object of the class *dist* , i.e., an object obtained with the *dist* function or a dissimilarity matrix transformed into an object of the class *dist* with the *as.dist* function) and the weights $w = (w_1, \ldots, w_n)$ of observations as arguments. Here, the standardized incidence ratio (SIR) Equation (4) of TB in the 223 municipalities of the State of Paraíba will be applied as non-uniform weights; ergo each municipality will have its weight. The sum of the heights in the dendrogram is equal to the total pseudo-inertia of the data set. The formula for pseudo-inertia of the Ward-like method is:

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

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 that belong to the cluster C_k. The function *hclustgeo* is a wrapper of the usual *hclust* function with the following arguments: (a) distance: D_0 (Manhattan distance). D_0 is the Manhattan distance matrix between the 223 municipalities performed with the $p = 5$ variables socio-epidemiological; (b) distance: D_1. The geographic distances between the municipalities; calculating a distance matrix for geographic points using R through packages: *sgeostat* (version 1.0–27) [24], *geosphere* (version 1.5–10) [25], and *Imap* (version 1.32) [26]. These functions calculate distance matrix for geographic for latitude and longitude points of the center of gravity of the municipalities; c) Members: $w = SIR_i$. The sum of the heights in the dendrogram is equal to the total pseudo-inertia of the data set Equation (2).

The spirit of the Ward-like hierarchical clustering is to aggregate the two clusters A and B from a given partition P_{K+1}^α in $K + 1$ clusters, to that the new partition has minimum mixed within-cluster inertia.

2.4. Manhattan Distance

We opted for the Manhattan distance, because the Ward method has already been generalized for use over non-Euclidean distances. According to Strauss and Maltitz [27], Ward's clustering algorithm can use it in conjunction with Manhattan distances.

$$d(i, j) = \sum_{(k=1)}^{n} |X_{ik} - X_{jk}| \tag{3}$$

where i and j are the municipalities with $k = 1, \ldots, n = 223$.

2.5. Standardized Incidence Ratio

One simple measure of disease risk is the standardized incidence ratio (SIR). For each area $i, i = 1, \ldots, n = 223$, the SIR is defined as the ratio of observed counts to the expected counts.

$$SIR_i = \frac{Y_i}{E_i} \tag{4}$$

The expected counts E_i represent the total number of TB cases that one would expect if the population of municipality i behaved the way the population of the State of Paraíba behaves. E_i can be calculated while using indirect standardization as $E_i = \sum_{j=1}^{m} r_j^{(s)} n_j^{(i)}$, where $r_j^{(s)}$ is the rate (number of cases divided by population) in stratum j in the standard population, and $n_j^{(i)}$ is the population in stratum j of area i.

3. Results and Discussion

Have been notified 24.258 TB cases in the State of Paraíba from 2001 to 2018. Of this total, 80% were new cases, 65% patients got cured, 46.8% had less than ten years of study, 81.3% were between working age (20–64), and 6,1% mortality, being men (4.2%) and women (1.9%). Clustering approaches are a useful tool to detect patterns in data sets and generate hypotheses regarding potential relationships. Therefore, the role of cluster analysis is to uncover a certain kind of natural structure in the data set [2].

Figure 1 shows the dendrogram of the dissimilarity matrix D_0, i.e., the differences in the feature space of socio-epidemiological variables, which is the Manhattan distance matrix between the 223 municipalities performed with $p = 5$ variables socio-epidemiological. To choose the suitable number K of clusters, we focus on the Ward dendrogram based on the $p = 5$ socio-epidemiological variables, that is using D_0 only.

Figure 1. Dendrogram of the $n = 223$ municipalities based on the 5 socio-epidemiologic variables (that is using D_0 only).

The visual inspection of the dendrogram in Figure 1 suggests retaining $K = 5$ clusters. The 223 municipalities grouped in their respective clusters according to socio-epidemiological similarity, namely, cluster 1 (42 municipalities), cluster 2 (37), cluster 3 (36), cluster 4 (90), and cluster 5, only 18 municipalities. The partition corresponding to the five clusters is shown on the map presented 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 A1 in Appendix A show the variable boxplots for each cluster (top row). Cluster 1, the female mortality rate is the lowest of all clusters, while male mortality has a higher median. Cluster 2 has a high rate of new cases and cure, and a higher female mortality rate than in other clusters. Cluster 3, people of working age (20–64) has a rate that is higher than the average value of the study area, as well as being higher than in other clusters. Similarly, cluster 4 also has a high rate of new cases and a high average age of TB patients of working age and is also greater than the average value of the study area. Cluster 5, high rates of new cases and people of working age, and the lowest cure rate in all other clusters.

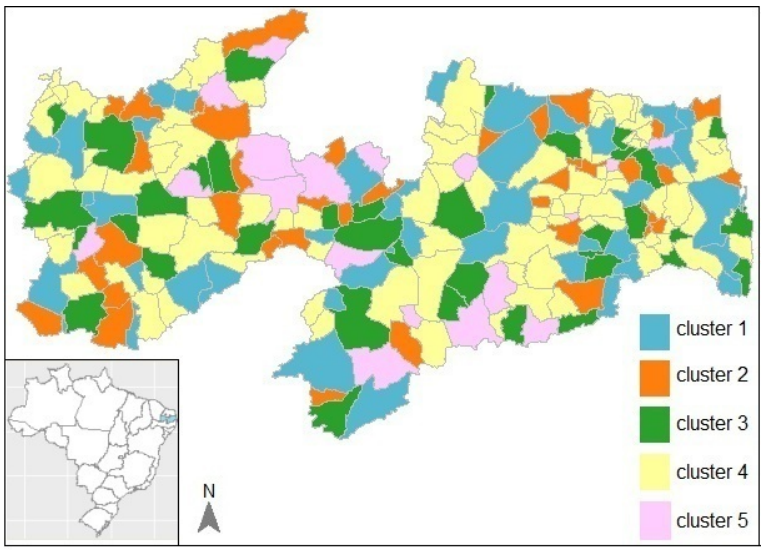

Figure 2. Map of the partition with 5 clusters only based on the socio-epidemiological variables (that is using D_0 only).

We will introduce the matrix D_1 of geographical distances into *hclustgeo*, i.e., a partition taking into account the geographical constraints in order to obtain geographically more compact clusters. For this, it is necessary that a mixing parameter is selected α to improve the geographical cohesion of the five 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-epidemiologic homogeneity and the geographic cohesion of the partitions obtained for a range of different values of α and the five clusters.

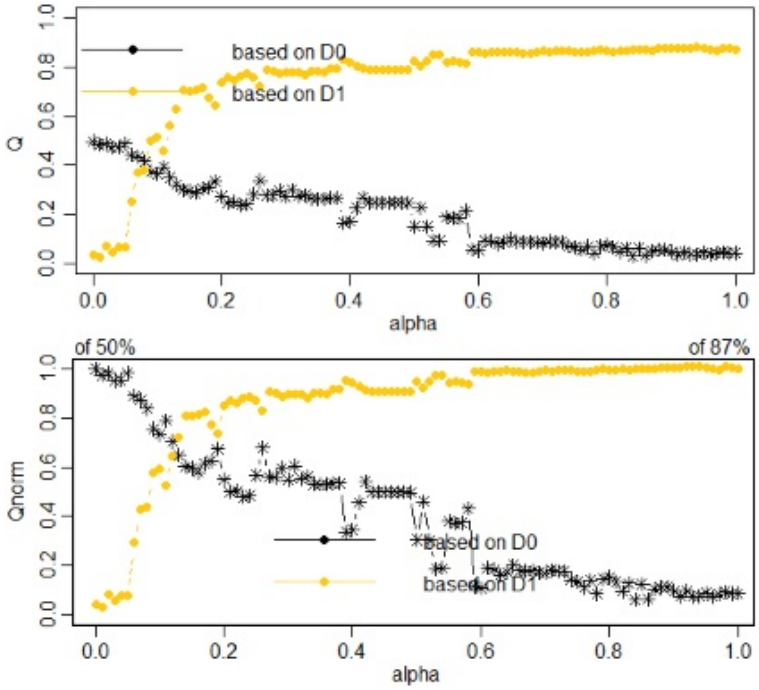

Figure 3. Choice of α for a partition in $K = 5$ clusters when D_1 is the geographical distances between municipalities. (**Top**) proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in gold dashed line). (**Bottom**) normalized proportion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

The next plot, Figure 3, shows the choice of α for partition, taking into account the geographical constraints. Figure 3 gives the plot of the proportion of explained pseudo-inertia calculated with D_0 (the socio-epidemiological distances), which is equal to 0.50 when $\alpha = 0$ and decreases when α increases (in solid black line). On the contrary, the proportion of explained pseudo-inertia calculated with D_1 (the geographical distances) is equal to 0.87 when $\alpha = 1$ and it decreases when α decreases (dashed line).

The obtaining of the partition taking into account the geographic constraints with the normalized proportion of explained inertias at the bottom of Figure 3 (i.e., $Q_0^*(P_K^\alpha)$ and $Q_1^*(P_K^\alpha)$, shows the value α that aims to increase the spatial contiguity, as seen in detail in Table 1.

Table 1. Normalized proportion of explained pseudo-inertias.

Alpha Values	Q0norm	Q1norm
α=0.16	0.57808701	0.81167272
α=0.17	0.61407565	0.82247147
α=0.18	0.62478207	0.77433402
α=0.19	0.67296737	0.73850413
α=0.20	0.54877711	0.84948899

The value of α is a trade-off between the loss of socio-economic homogeneity and the gain of geographic cohesion. When $\alpha = 0$, the geographical dissimilarities are not taken into account. When $\alpha = 1$, it is the socio-epidemiologic distances that are not taken into account; the clusters are obtained with the geographical distances only. The plot presented in Figure 3 (bottom) would appear to suggest choosing $\alpha = 0.17$, which corresponds to a loss of only (1 − 0.61407565 = 38.59%) of socio-epidemiologic with a SIR of each municipality, and 17.75% increase in geographical homogeneity.

208 The increased geographical cohesion of this partition can be seen in Figure 4.

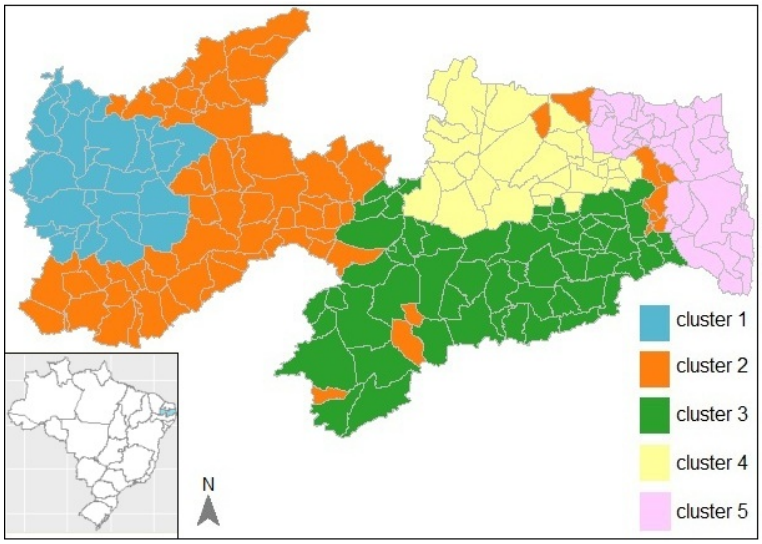

Figure 4. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the geographical distances between the municipalities D_1 with $\alpha = 0.17$.

209 In Figure 4, a gain significative in spatial homogeneity is perceived. Figure A1 presented in Appendix
210 A shows the boxplots of the variables for each cluster of the partition (middle row). Clusters 1, 3, and 4
211 seem to differentiate among themselves mainly due to the slight increase in deaths of male patients in
212 cluster 4 and greater variation in the cure proportion of cluster 3. Cluster 5 differed from cluster 2 by
213 the slight increase in deaths, with greater proportions for males, whereas cluster 2 has a higher average
214 number of working-age TB patients and greater variation proportion of cure.
215 The next plot, Figure 5, shows the choice of α for partition, taking into account the neighborhood constraints.

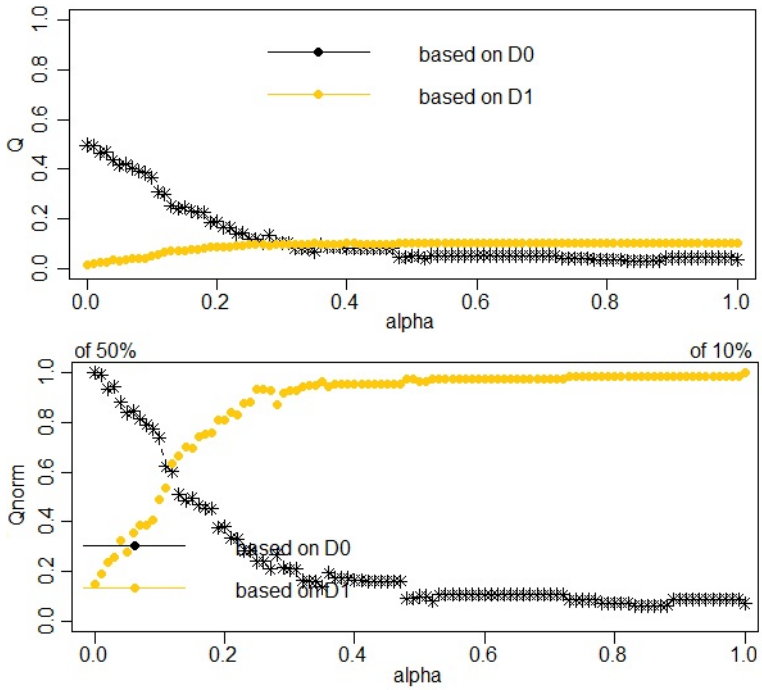

Figure 5. Choice of α for a partition in $K = 5$ clusters when D_1 is the neighborhood dissimilarity matrix between municipalities. (**Top**) proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in gold dashed line). (**Bottom**) normalized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

At the bottom of Figure 5, the plot of the normalized proportion of explained inertias (i.e., $Q_0(P_K^\alpha)$ and $Q_1(P_K^\alpha)$) suggests retaining $\alpha = 0.12$ slightly favoring socio-epidemiological homogeneity versus geographical homogeneity.

It remains only to determine this final partition for $K = 5$ clusters and $\alpha = 0.12$. Figure 6 provides the corresponding map.

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. Thereupon, as the approach is based on soft contiguity restrictions, municipalities that are not neighbours may be in the same clustering, according occurs with the municipalities of Lucena, Coxixola, Congo, Zabelê and Santa Inês in cluster 2. 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 (61.41% versus 82.25%).

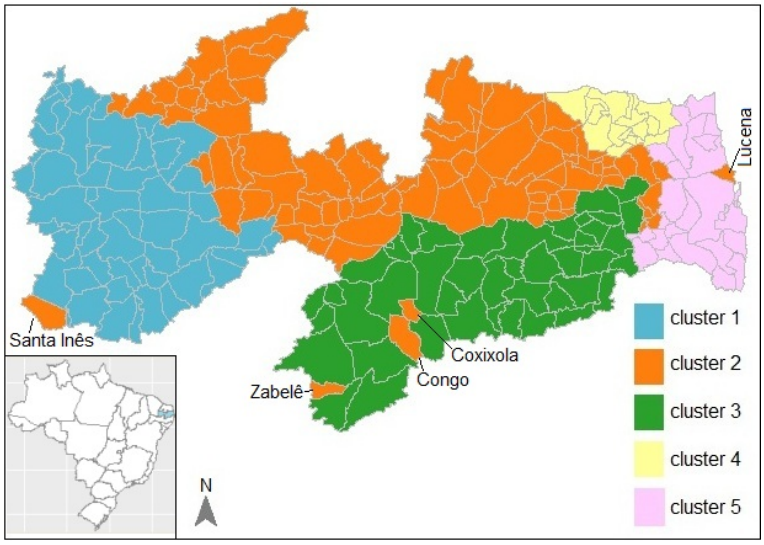

Figure 6. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the "neighborhood" distances of the municipalities D_1 with $\alpha = 0.12$.

4. Conclusions

When considering spatial/geographical constraints, the hierarchical grouping becomes even more complete in detecting patterns in data sets of different dimensions. According to the weights that are given to the geographical differences in this combination, the solution will have more or less spatially contiguous clusters. Through our results, the non-uniform weights w defined by the standardized incidence ratio (SIR) of TB contributed to the increase in clarity both from a spatial and socio-epidemiological point of view.

Therefore, the application of the Ward–Like method becomes indispensable in understanding the socio-epidemiological reality of the State of Paraíba from a spatial perspective, thus facilitating decisions in the development of public policies and more effective health actions in the fight against tuberculosis.

Future work would be to add new socio-epidemiological variables and, instead of the municipalities, use the Health Regions that are responsible for the organization, planning, and execution of health actions and services in the state of Paraíba.

Author Contributions: Resources, D.C.A., R.G.S. and E.L.S.C.; Supervision, R.G.S. and E.L.S.C.; Writing—review & editing, D.C.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DATASUS	Department of the Unified Health System
IBGE	Brazilian Institute of Geography and Statistics
SIR	standardized incidence ratio
TB	tuberculosis
WHO	World Health Organization

Appendix

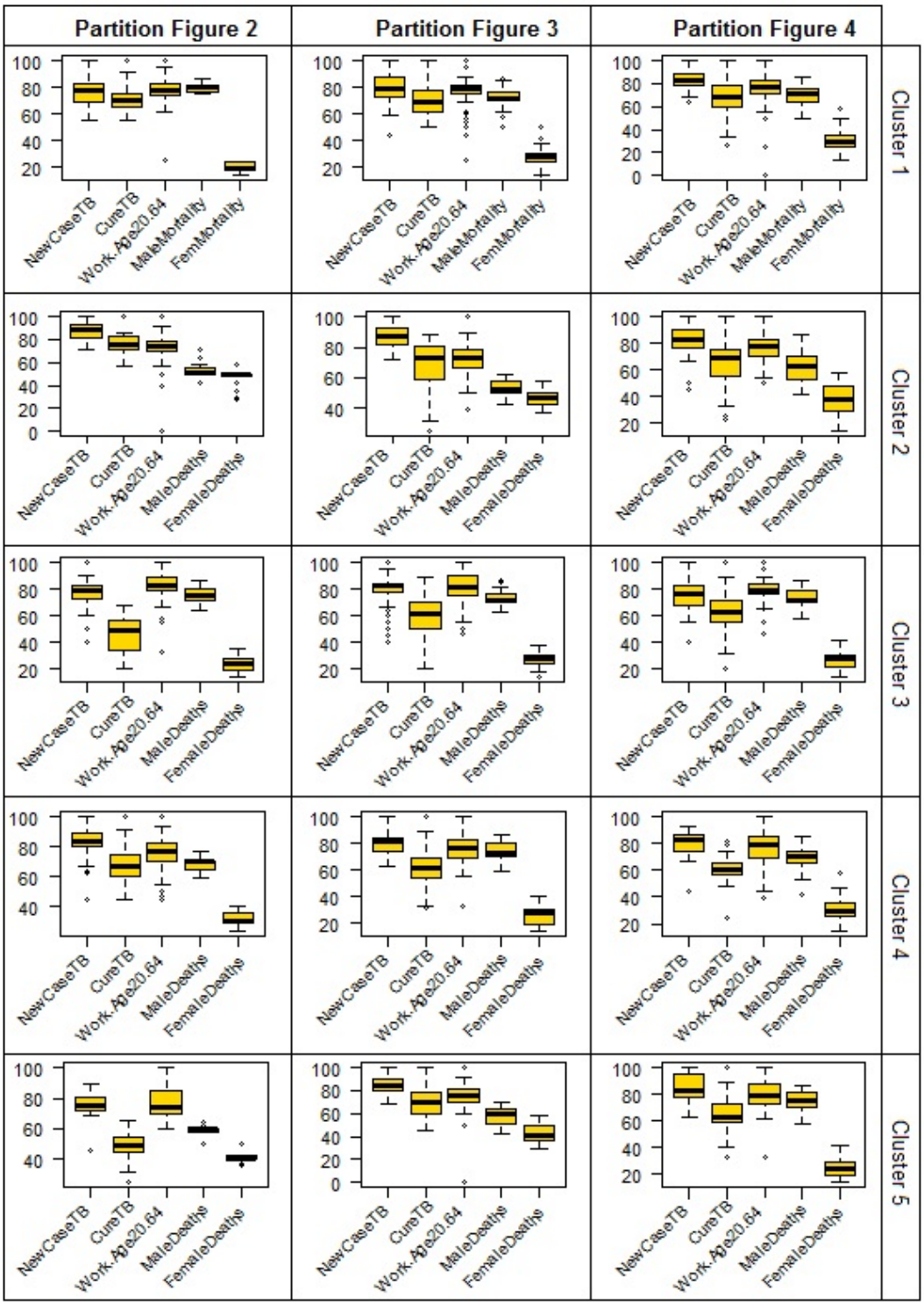

Figure A1. Comparison of clusters in the partition of Figures 2–4 in terms of variables.

References

1. Vandewalle, V. Multi-Partitions Subspace Clustering. *Mathematics* **2020**, *8*, 597–615. [Google Scholar] [Crossref]
2. Wierzchoń, S.T.; Kłopotek, M.A. Cluster Analysis. In *Modern Algorithms of Cluster Analysis*; Janusz, K., Ed.; Springer International Publishing AG: Cham, Switzerland, 2018; pp. 9–66.
3. Chavent, M.; Kuentz-Simonet, V.; Labenne, A.; Saracco, J. *ClustGeo: Hierarchical Clustering with Spatial Constraints*. R Package Version 2.0. 2017. Available online: https://CRAN.R-project.org/package=ClustGeo (accessed on 17 May 2020).
4. WHO—World Health Organization. *Global Tuberculosis Report 2019*; World Health Organization: Geneva, Switzerland, 2019. [Crossref]
5. Reis-Santos, B.; Shete, P.; Bertolde, A.; Sales, C.M.; Sanchez, M.N.; Arakaki-Sanchez, D.; Andrade, K.B.; Gomes, M.G.M.; Boccia, D.; Lienhardt, C.; et al. Tuberculosis in Brazil and Cash Transfer Programs: A Longitudinal Database Study of the Effect of Cash Transfer on Cure Rates. *PLoS ONE* **2019**, *14*, e0212617. [Crossref]
6. World Health Organization. *Global Tuberculosis Report 2017*; World Health Organization: Geneva, Switzerland, 2017. [Google Scholar]
7. Ministério da Saúde. *Brasil Livre da Tuberculose: Evolução dos Cenários Epidemiológicos e Operacionais da Doença*; Boletim Epidemiológico. Secretaria de Vigilância em Saúde: Brasília-Brasil, March 2019; Volume 50. [Google Scholar]
8. Aguiar, D.C.; Silva, Camelo, E.L.; Carneiro, R.O. Análise estatística de indicadores da tuberculose no Estado da Paraíba. *Rev. Aten. Saúde.* **2019**, *17*, 5–12. [Crossref]
9. IBGE. Instituto brasileiro de geografia e Estatística. *Paraíba–Panorama*, 2019. Available online: https://cidades.ibge.gov.br (accessed on 8 May 2020).
10. Santos Neto. M.; Sousa, M.R.; da Silva, F.B.G.; Santos, F.S.; Ferreira, A.G.N.; Pascoal, L.M.; Costa, A.C.P.d.; Bezerra, J.M.; Serra, M.A.A.d.O.; Dias, I.C.C.M.; et al. Spatial distribution of tuberculosis cases in a priority Brazilian northeast municipality for control of the disease. *Int. J. Dev. Res.* **2017**, *7*, 10611. [Crossref]
11. Moraga, P. Small Area Disease Risk Estimation and Visualization Using R. *R J* **2018**, *10*, 495–506. [Crossref]
12. SINAN. Sistema de Informação de Agravos de Notificação. In *Tuberculose–Casos Confirmados*; Ministério da Saúde, Brazil: Brasília, Barzil, 2020. Available online: http://www2.datasus.gov.br/ (accessed on 5 April 2020).
13. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019. Available online: https://www.R-project.org/ (accessed on 15 May 2020).
14. Duque, J.C.; Dev, B.; Betancourt, A.; Franco, J.L. *ClusterPy: Library of Spatially Constrained Clustering Algorithms*. RiSE-Group (Research in Spatial Economics). 2011. Version 0.9.9. Available online: http://www.rise-group.org/risem/clusterpy/ (accessed on 15 May 2020). [Google Scholar]
15. Bécue-Bertaut, M.; Alvarez-Esteban, R.; Sánchez-Espigares, J.A. *Xplortext: Statistical Analysis of Textual Data R Package*. R Package Version 1.0. 2017. Version 0.9.9. Available online: https://cran.r-project.org/package=Xplortext (accessed on 7 June 2020). [Google Scholar]
16. Dehman, A.; Ambroise, C.; Neuvial, P. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinform.* **2015**, *16*, 148. [Google Scholar].
17. Legendre, P. const.clust: Space-and Time-Constrained Clustering Package. R Package Version 1.2. 2011. Available online: http://adn.biol.umontreal.ca/numericalecology/Rcode/ (acessed on 20 May 2020).
18. Ambroise, C.; Govaert, G. Convergence of an EM-type algorithm for spatial clustering. *Pattern Recognit. Lett.* **1998**, *19*, 919–927. [Google Scholar].
19. Ambroise, C.; Dang, M.; Govaert, G. Clustering of Spatial Data by the EM Algorithm. In *geoENV I-Geostatistics for Environmental Applicattions*; Soares, A.O., Gomez-Hernandez, J.J., Froidevaux, R., Eds.; Kluwer: Dordrecht, the Netherland, 1997; pp. 493–504.
20. Aguiar, D.C.; Sánchez, R.G.; Silva Camêlo, E.L. Hierarchical clustering with spatial constraints in tuberculosis data. *IJDR* **2020**, *10*, 35374–35380. [Google Scholar].
21. Aguiar, D.C.; Sánchez, R.G.; Silva Camêlo, E.L. Ward-like hierarchical clustering with dissimilarities and non-uniform weights in cases of tuberculosis in Paraíba, Brazil. *IJDR* **2020**, *10*, 35478–35483. [Google Scholar].

22. Miele, V.; Picard, F.; Dray, S. Spatially constrained clustering of ecological networks. *Methods Ecol. Evol.* **2014**, *5*, 771–779. [Crossref].
23. Chavent, M.; Kuentz-Simonet, V.; Labenne, A.; Saracco, J. ClustGeo: An R package for hierarchical clustering with spatial constraints. *Comput. Stat*. **2018**, *33*, 1799–1822. [Crossref].
24. Majure, J.J.; Gebhardt, A. *sgeostat: An Object-Oriented Framework for Geostatistical Modeling in S+*; R Package Version 1.0-27. 2016. Available online: https://CRAN.R-project.org/package=sgeostat (accessed on 7 June 2020). [Google Scholar].
25. Hijmans, R.J. *Geosphere: Spherical Trigonometry*; R Package Version 1.5-10. 2019. Available online: https://CRAN.R-project.org/package=geosphere (accessed on 7 June 2020). [Google Scholar].
26. Wallace, J.R. *Imap: Interactive Mapping*; R Package Version 1.32. 2012. Available onlinen: https://CRAN.R-project.org/package=Imap (accessed on 7 June 2020). [Google Scholar].
27. Strauss, T.; von Maltitz, M.J. Generalising Ward's Method for Use with Manhattan Distances. *PLoS ONE* **2017**, *12*, e0168288. [Crossref].

Artículo 2

Ward-like hierarchical clustering with dissimilarities and non-uniform weights in cases of tuberculosis in Paraíba, Brazil

El segundo artículo *Ward-like hierarchical clustering with dissimilarities and non-uniform weights in cases of tuberculosis in Paraíba, Brazil* fue publicado en la revista *International Journal of Development Research*, Vol. 10, Issue, 04, pp. 35478-35483, April, 2020 de acceso abierto con factor de impacto SJIF 2019: 7.012. doi.org/10.37118/ijdr.18753.04.2020

ISSN: 2230-9926

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

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):

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

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

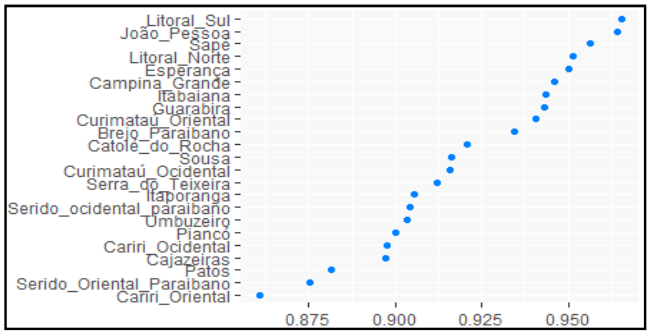

Figure 1. Values of the diversification coefficient (DC) of socio-epidemiological variables of microregions, Paraíba, Brazil

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
$\alpha = 0.17$	0.80773244	0.68104151
$\alpha = 0.18$	0.71786338	0.76987949
$\alpha = 0.19$	0.75936603	0.74331210
$\alpha = 0.20$	0.73858351	0.82422833
$\alpha = 0.21$	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,

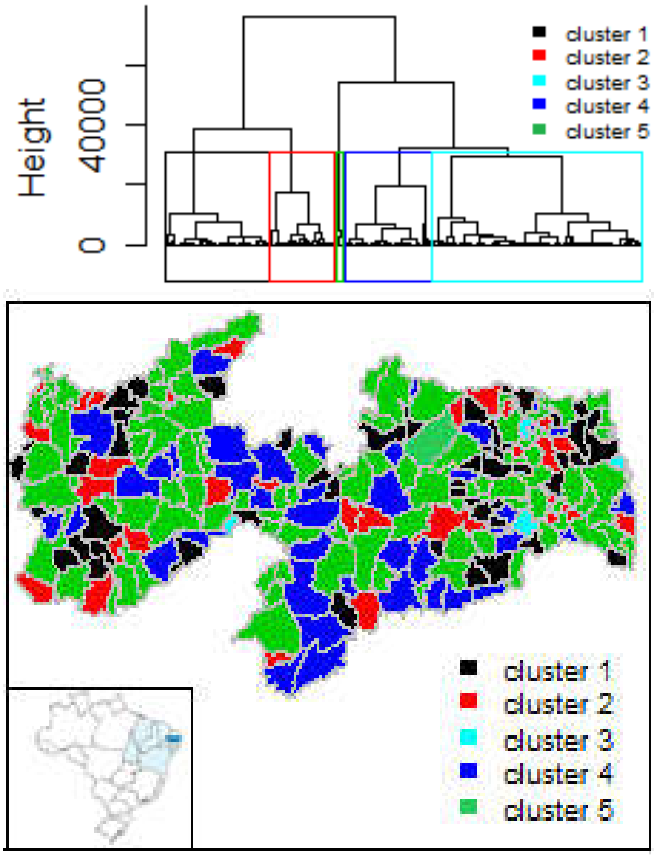

Figura 2. a) Dendrogram of the $n = 223$ municipalities based on the *4* socio-epidemiologic variables (that is using D_0 only). b) Map of the partition with 5 clusters only based on the socio-epidemiological variables for diversification coefficient

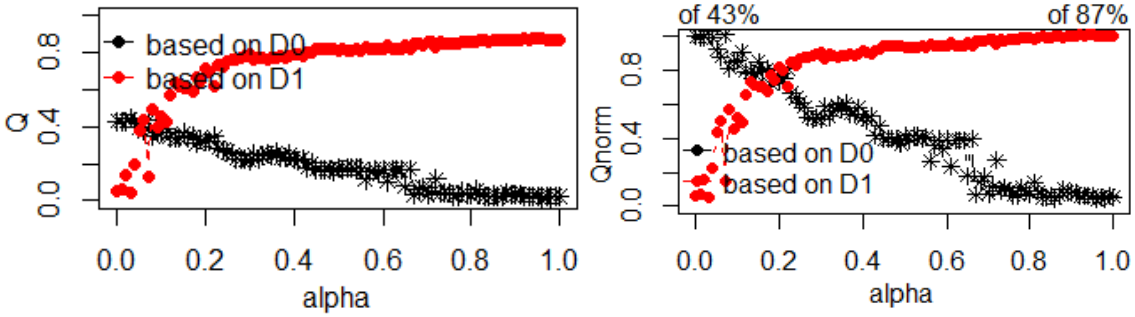

Figure 3. Choice of α for a partition in $K = 5$ clusters when D_1 is the geographical distances between municipalities. Left: proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in dashed line). Right: normalized proportion of explained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

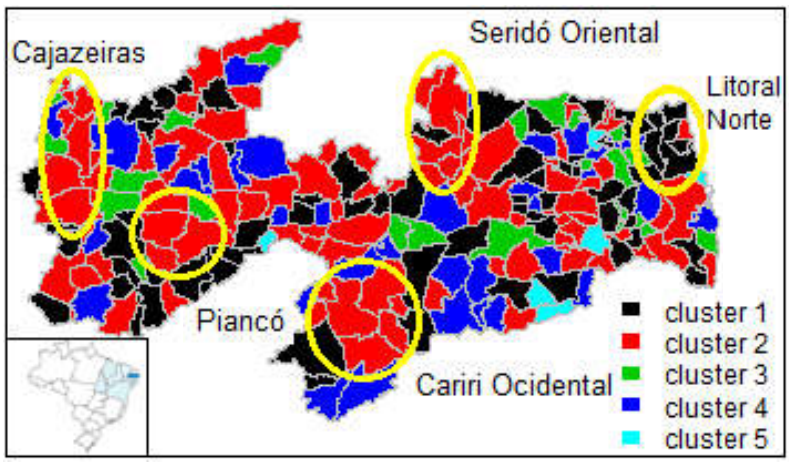

Figure 4. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the geographical distances between the municipalities D_1 with $\alpha = 0.2$.

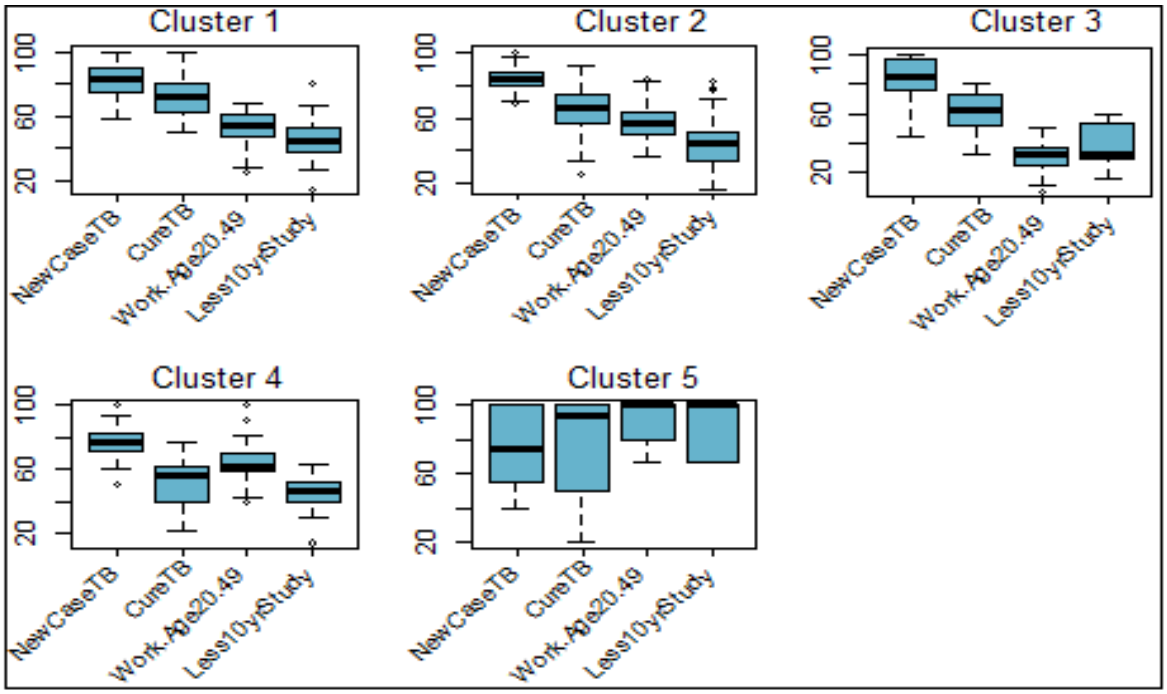

Figure 5. Comparison of clusters in the partition of Figure 5 in terms of variables

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.

****

Artículo 1

Hierarchical clustering with spatial constraints in tuberculosis data

El primer artículo *Hierarchical clustering with spatial constraints in tuberculosis data* fue publicado en la revista *International Journal of Development Research*, Vol. 10, Issue, 04, pp. 35374-35380, April, 2020 con factor de impacto SJIF 2019: 7.012. doi.org/10.37118/ijdr.18706.04.2020

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:

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

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 $h_1, h_2, \dots, h_{23}$ with absolute frequencies $n_i, n_2, \dots, n_{23}$ 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.

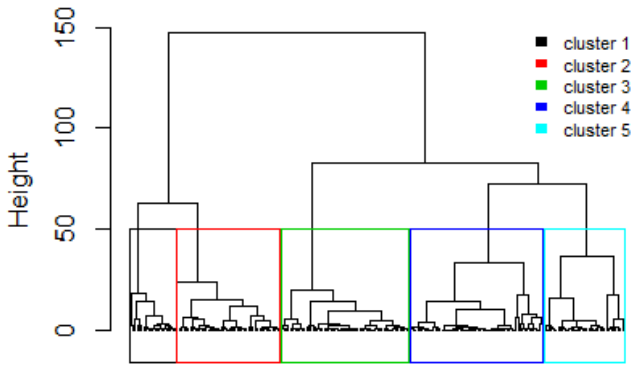

Figure 1. Dendrogram of the $n = 223$ municipalities based on the *4* socio-epidemiologic variables (that is using D_0 only).

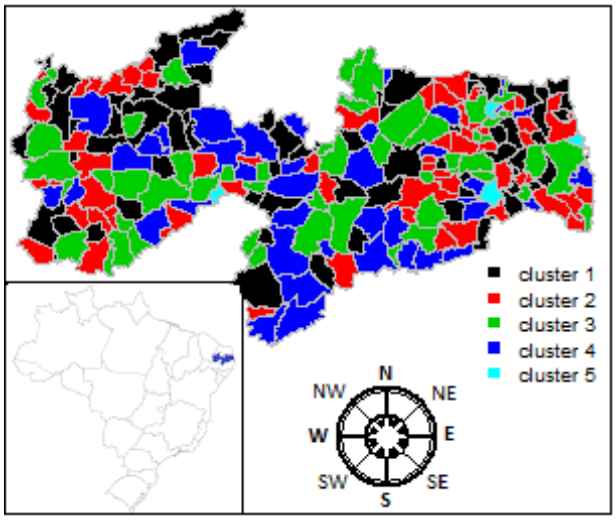

Figure 2. Map of the partition with K= 5 clusters only based on the socio-epidemiological variables (that is using D_0 only)

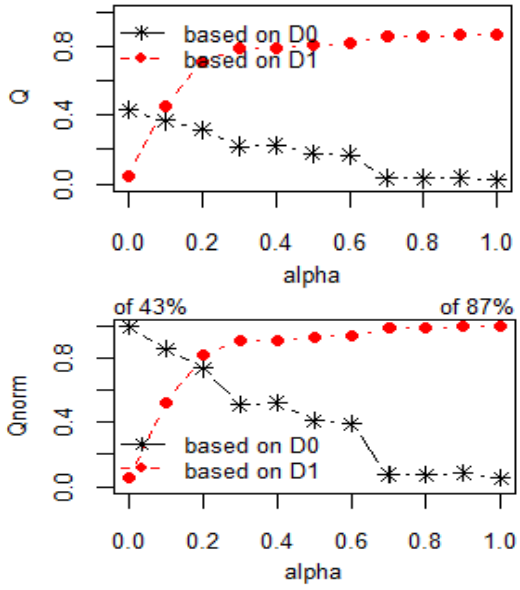

Figure 3. Choice of α for a partition in $K = 5$ clusters when D_1 is the geographical distances between municipalities. Left: proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in dashed line). Right: normalized proportion of explained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

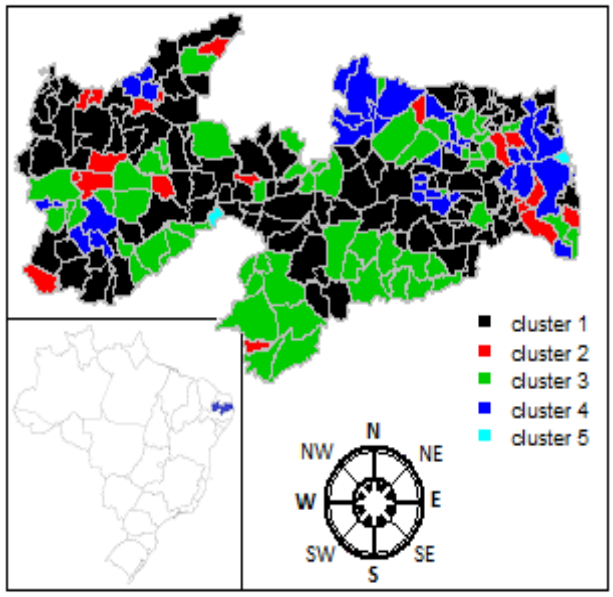

Figure 4. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the geographical distances between the municipalities D_1 with $\alpha = 0.1$.

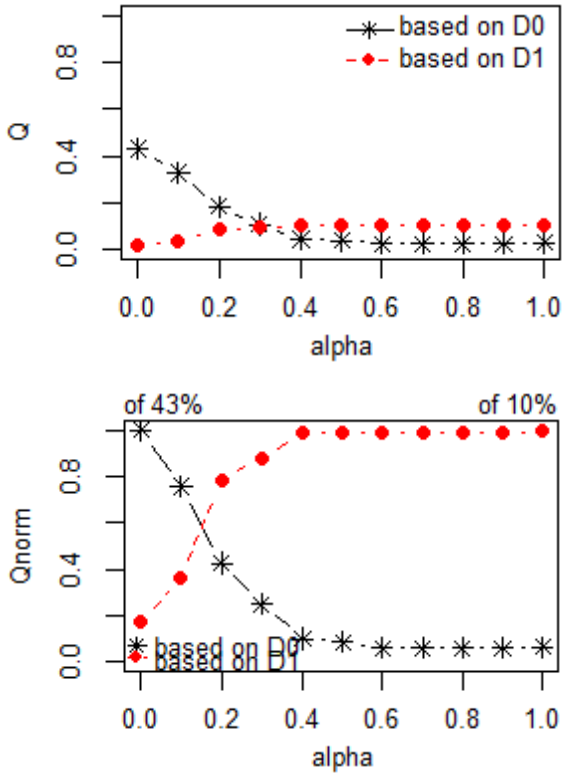

Figure 5. Choice of α for a partition in $K = 5$ clusters when D_1 is the neighborhood dissimilarity matrix between municipalities. Left: proportion of explained pseudo-inertias $Q_0(P_K^\alpha)$ versus α (in black solid line) and $Q_1(P_K^\alpha)$ versus α (in dashed line). Right: normalized proportion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

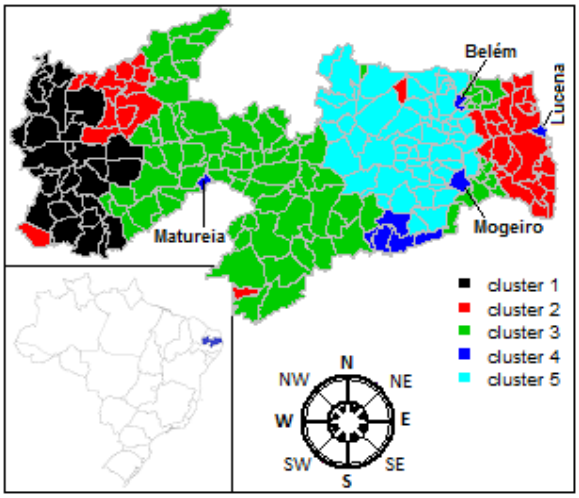

Figure 6. Map of the partition with $K = 5$ clusters based on the socio-epidemiological distances D_0 and the "neighborhood" distances of the municipalities D_1 with $\alpha = 0.2$.

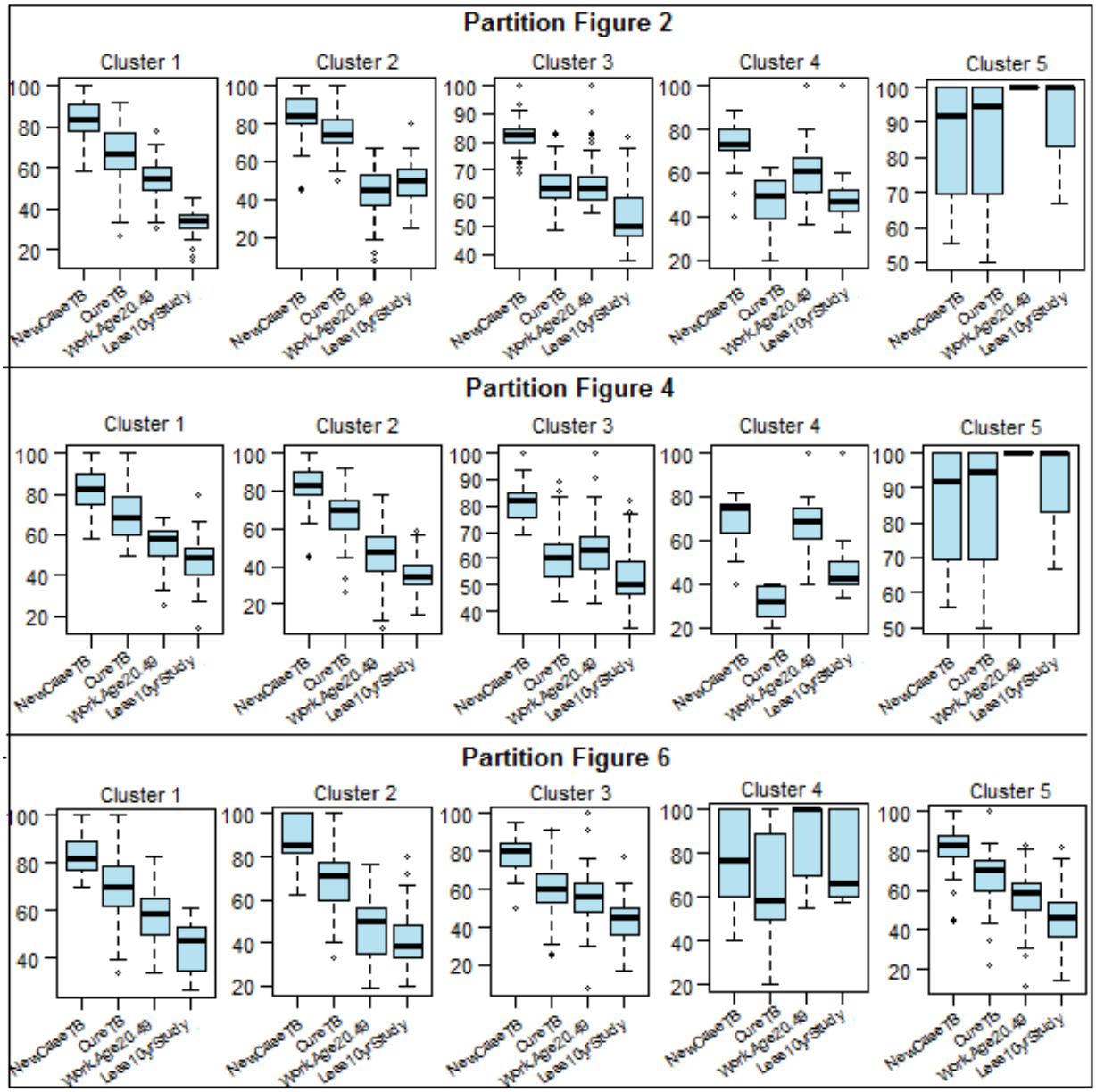

Figure 7. Comparison of the final partitions Figure 2, Figure 4 and Figure 5 in terms of variables.

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.

****