



UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

CARLA CAROLINE CARVALHO SILVA

Análise de Agrupamentos aplicados à dados socioeconômicos de municípios paraibanos.

Campina Grande - PB

2017

CARLA CAROLINE CARVALHO SILVA

**Análise de Agrupamentos aplicados à dados
socioeconômicos de municípios paraibanos.**

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Ricardo Alves Olinda

Coorientador: Prof. Dr. João Gil de Luna

Campina Grande - PB

2017

É expressamente proibida a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano da dissertação.

S586a Silva, Carla Caroline Carvalho.
Análise de agrupamentos aplicados à dados socioeconômicos de municípios paraibanos [manuscrito] / Carla Caroline Carvalho Silva. - 2017.
35 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2017.
"Orientação: Prof. Dr. Ricardo Alves Olinda, Departamento de Estatística".
"Co-Orientação: Prof. Dr. João Gil de Luna, Departamento de Estatística".
1. Variáveis socioeconômicas. 2. Técnicas hierárquicas aglomerativas. 3. Distância de Mahalanobis. I. Título.
21. ed. CDD 519.5

CARLA CAROLINE CARVALHO SILVA

Análise de Agrupamentos aplicados à dados socioeconômicos de municípios paraibanos.

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

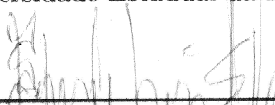
Área de Concentração:

Aprovado em: 09 de Agosto de 2017

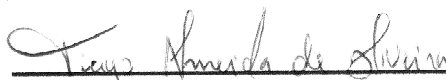
BANCA EXAMINADORA



Prof. Dr. Ricardo Alves
Olinda(Orientador)
Universidade Estadual da Paraíba



Prof. Dr. Edwirde Luiz Silva
Universidade Estadual da Paraíba



Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Resumo

O objetivo principal deste trabalho é fazer uma breve introdução da análise de agrupamento, baseada nas técnicas hierárquicas aglomerativas de ligação simples, completa, média e Ward, tendo por medida de proximidade a distância de Mahalanobis. Para auxiliar na decisão dos órgãos públicos, foram utilizadas variáveis socioeconômicas, com o intuito de verificar a semelhança dos municípios socioeconomicamente sem a restrição geográfica, como é o caso das mesorregiões. Foi considerado para o estudo apenas uma amostra de quarenta municípios, selecionados segundo a sua quantidade populacional, coletados por meio de banco de dados virtuais e manuseados através do software R(3.4.0). A partir dos métodos implementados no conjunto de dados observou-se um grau de similaridade considerável entre os municípios de João Pessoa e Campina Grande, mesmo pertencendo a mesorregiões diferentes.

Palavras-chave: Variáveis socioeconômicas; Técnicas hierárquicas aglomerativas; Distância de Mahalanobis.

Abstract

The main objective of this work is to make a brief introduction of the cluster analysis, based on the cluster analysis based on the agglomerative techniques of simple, complete, mean and Ward binding, taking as a measure of proximity to Mahalanobis distance. In order to aid in the decision of the public agencies, socioeconomic variables were used, in order to verify the similarity of the municipalities socioeconomically without the geographical restriction, as is the case of mesoregions. The study was considered a sample of forty municipalities, selected according to their population, collected through a virtual database and handled through R(3.4.0) software. From the methods implemented in the dataset, a considerable degree of similarity was observed between the cities of João Pessoa and Campina Grande, even though they belonged to different mesoregions.

Key-words: Socioeconomic variables; Agglomerative hierarchical techniques; Mahalanobis distance.

Lista de ilustrações

Figura 1 – Mortalidade infantil e Expectativa de vida do ano de 2010 no Brasil(adaptado do Google Imagens).	9
Figura 2 – Nomenclatura referente as Técnicas Hierárquicas Aglomerativas mais utilizadas.	19
Figura 3 – Gráfico de correlação referente as 8 variáveis estudadas(que são: V2=proporção de residencias com água encanada; V3=proporção de residencias que possuem coleta de lixo; V4=proporção de residencias que possuem rede de esgoto; V5= IDHM; V6=índice de Gini; V7=proporção de residencias com oito pessoas; V8=índice pluviométrico).	28
Figura 4 – Comportamento das variáveis em estudo, revelado por intermédio do Gráfico de perfil.	30
Figura 5 – Boxplot referentes as variáveis socioeconômicas.	31
Figura 6 – Dendrograma formado através da técnica hierárquica aglomerativa utilizando a ligação simples e a média como forma de agrupamento, e a distância de Mahalanobis como medida de proximidade.	32

Lista de tabelas

Tabela 1	– Estrutura referente aos valores observados, onde as linhas representam os elementos e as colunas as suas respectivas características (variáveis).	11
Tabela 2	– Tabela de contingência para os elemento f e t caracterizados pelas v variáveis binárias.	14
Tabela 3	– Coeficientes de similaridades entre dois elementos f e t , cujas realizações são respostas binárias em v variáveis, junto com o nome utilizado na literatura, explicação racional de seu significado e intervalo de variação correspondente.	15
Tabela 4	– Distâncias entre dois elementos f e t , levando em consideração apenas duas variáveis.	17
Tabela 5	– Resumo dos métodos hierárquicos aglomerativos, onde, N_f , N_t e N_w são os números de elementos nos grupos f , t e w .	20
Tabela 6	– Municípios mais populosos do estado da Paraíba, com suas representações numéricas e respectivas mesorregiões.	26
Tabela 7	– Resultado do teste de correlação de Spearman.	30
Tabela 8	– Comparação dos coeficientes de correlação cofenética provenientes dos agrupamentos formados com retirada de municípios que representam os valores discrepantes mais latentes.	32
Tabela 9	– Grupos formados por meio do Dendrograma gerado através da matriz de distâncias de Mahalanobis	33

Sumário

1	INTRODUÇÃO	8
2	REVISÃO DE LITERATURA	9
2.1	Seleção da variável estatística de agrupamento	10
2.2	Tratamento dos dados	11
2.2.1	Tratamento das variáveis (atributos)	12
2.2.2	Normalização (padronização ou estandardização)	12
2.3	Medida de Proximidade	13
2.3.1	Medidas de Similaridade	13
2.3.2	Medidas de Dissimilaridade	15
2.4	Técnicas de Agrupamento	18
2.4.1	Métodos Hierárquicos Aglomerativos	19
2.4.1.1	Ligação Simples	20
2.4.1.2	Ligação Completa	21
2.4.1.3	Ligação média	21
2.4.1.4	Ligação Ward	22
2.5	Métodos para encontrar o número g de clusters(ou número de corte)	22
2.6	Métodos de validação	23
2.6.1	Número de grupos	23
2.6.2	Técnica hierárquica	23
3	MATERIAL E MÉTODOS	25
3.1	Distância de Mahalanobis	26
3.2	Métodos hierárquicos aglomerativos	27
4	APLICAÇÃO	28
5	CONSIDERAÇÕES FINAIS	34
	REFERÊNCIAS	35

1 Introdução

A Análise multivariada é utilizada quando não existe independência entre as variáveis, o que é suposto na univariada, tendo portanto que analisar as suas correlações. Pode ser aplicada em várias áreas do conhecimento, inclusive na economia, saúde, sociologia, ou em todas elas ao mesmo tempo, utilizando de variáveis socioeconômicas. Os primeiros relatos a respeito desta análise foram obtidos através de Pearson(1901), Fisher(1928), Hotelling(1931), Wilks(1932) e Bartlett(1937), que viram a necessidade de utilizar múltiplas respostas na análise dos dados (SARTORIO, 2008, p.13). Várias foram as análises criadas com este intuito, entre elas estão: a análise de componentes principais, a análise fatorial, a análise discriminantes, a análise de variância multivariada (MANOVA), a análise de agrupamentos, entre outras.

A Análise de Agrupamentos (também chamada de análise de conglomerado, clusters ou de taxonomia, dependendo da linha de pesquisa) é uma técnica que auxilia na formação de grupos (de objetos, pessoas, árvores, animais e etc.) internamente homogêneos e externamente heterogêneos, por meio de um critério de aproximação relacionado a semelhança entre os objetos, através da comparação de seus elementos baseada na variável estatística de agrupamento. É considerada a única técnica multivariada que não utiliza da estimação da variável, mas deixa da forma como implementada pelo pesquisador (HAIR et al., 2009, p.430), podendo ser classificada como descritiva, ateorética e não inferencial, sendo mais utilizada como uma técnica exploratória, com formação de uma taxonomia (alocação dos elementos com base no conhecimento empírico), servindo para reduzir o número de elementos, como também para formular hipóteses (sobre a natureza dos dados) ou examinar às já estabelecidas. Entretanto, também é possível utiliza-la como para confirmar uma estrutura de dados já existente (RODRIGUES et al., 2009, p.327-329).

Na teoria a análise de clusters é dividida em algumas etapas: (1) Definir as variáveis; (2) Tratamento dos dados; (3) Escolha da medida de parença (ou proximidade); (4) Escolha da técnica de agrupamento (ou algoritmos); (5) Validar o agrupamento formado e interpretá-lo (REIS, 2001, p.290-291). Estas etapas não são independentes. Pois, pode ocorrer a necessidade de voltar às anteriores para corrigi-las, proporcionando um aprimoramento das posteriores (RODRIGUES et al., 2009, p.329).

O objetivo deste trabalho é aplicar as técnicas de agrupamento em variáveis socioeconômicas do estado da Paraíba; comparar diferentes métodos da análise hierárquica aglomerativa referente a variáveis socioeconômicas; aplicar os métodos que melhor retratam a realidade; e apresentar para os órgãos públicos, municípios, que apresentam similaridades socioeconômicas.

2 Revisão de Literatura

Conforme o Instituto Brasileiro de Geografia e Estatística (IBGE), o estado da Paraíba é dividido em quatro mesorregiões, que repartem os 223 municípios, levando em consideração as características econômicas, sociais e políticas, baseados nas dimensões referentes ao processo social determinante, o quadro natural como condicional e a rede de comunicação e lugares como elemento da articulação espacial, isto é, esta divisão teve por critério as características locais e a organização tanto socioeconômica quanto política.

“A socioeconomia é uma área de conhecimento que se propõe a estudar diferentes expressões, projetos e estratégias de convivência social, alicerçadas na expansão do plano democrático, em que os avanços econômicos estejam subordinados a benefícios sociais estendidos a toda a sociedade”. (SANTOS, 2014).

Quando trata-se de benefícios sociais para a população paraibana, é vital destacar as melhorias que vem ocorrendo em vários indicadores socioeconômicos, contudo, “o Estado ainda registra carências sociais relevantes, cuja solução se põe como desafio da mais alta prioridade nas ações de governo nos próximos anos”. Pois, mesmo com estas melhorias os dados mais recentes de alguns indicadores importantes, como, mortalidade infantil e expectativa de vida, ainda são considerados os piores do país (FIEP, 2010). Observe a Figura 1.

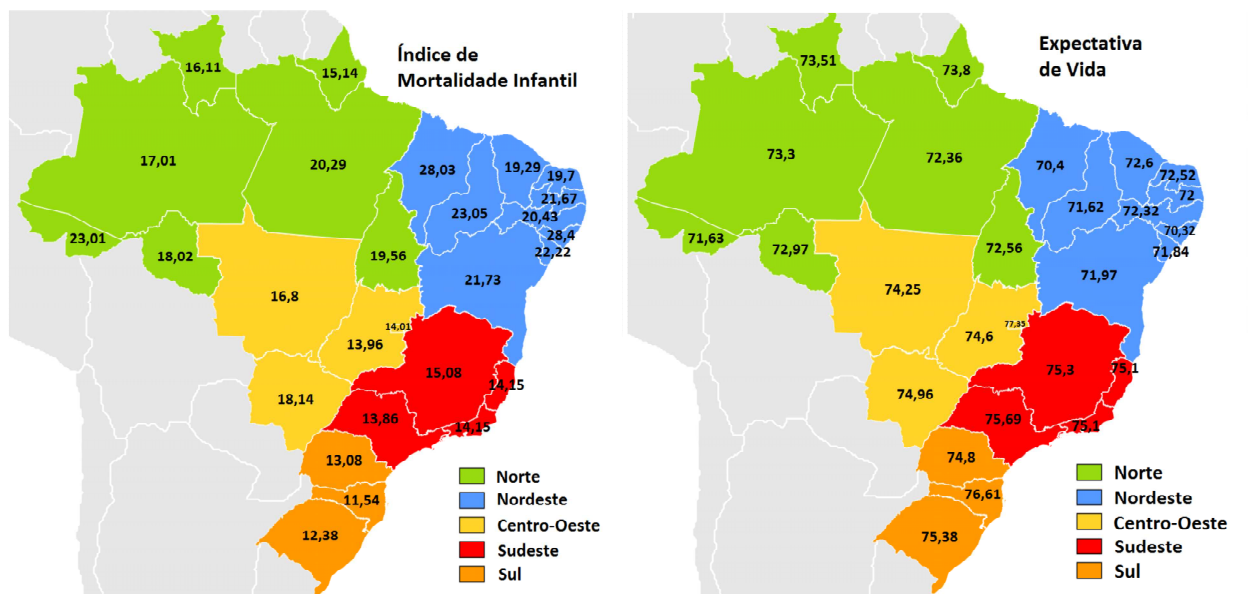


Figura 1 – Mortalidade infantil e Expectativa de vida do ano de 2010 no Brasil(adaptado do Google Imagens).

Logo, é interessante e importante estudar a semelhança entre os municípios de um estado, referente aos indicadores socioeconômicos, ainda em melhorias, sem limites geográficos, para facilitar no desdobramento das soluções que possivelmente serão apresentadas pelos políticos. Para este fim existe a análise de agrupamento.

Esta teve início na psicologia, a partir dos trabalhos de Zubin (1933) e Tryon (1939), na antropologia através do artigo de Driver e Kroeber (1932), na economia, Fisher (1969), na Geografia, Berry e Ray (1966) e na Ciência Política, Kaiser (1966). Em resumo, existiram diversos autores que foram pioneiros desta análise em suas áreas de conhecimento (BAILEY, 1975 apud FILHO; JUNIOR; ROCHA, 2012, p.111).

A Análise de Agrupamento está presente em muitos ramos da ciência. Na social, por exemplo, atua como instrumento de apoio na melhoria da qualidade de atendimento ao cliente de forma que os produtos sejam organizados da melhor forma possível, agrupando-os de acordo com a sua semelhança; na econômica, auxilia na tomada de decisão, por exemplo, descobrindo qual o estado que é mais vantajoso para a implantação de uma empresa, utilizando de variáveis como: renda, Índice de Desenvolvimento Humano Municipal (IDHM), taxa cobrada (Imposto), e etc.

As opiniões divergem a respeito da Análise de Agrupamento, quanto a ser um método classificatório ou não, pelo fato, dela não possuir o número de grupos pré-definido (como é o caso da análise discriminante), sendo intitulado desta forma como um aprendizado não-supervisionado, ou seja, não classificatório. Por exemplo, Mingotí (2013), Sartorio (2008), Vicini (2005) e Reis (2001) consideram AA classificatória, já Doni (2004) e Ferreira (2008) não. Entretanto, sabe-se que a AA possui os métodos não-hierárquicos, que necessitam saber previamente o número de grupos, isto é, antes da análise (FERREIRA, 2008, p.353,355).

“A análise de agrupamentos se assemelha à análise fatorial em seu objetivo de avaliar a estrutura” (HAIR et al., 2009, p.430). Entretanto, se diferenciam quando trata-se de limitações, pois a análise fatorial lida apenas com a formação de grupos de variáveis, enquanto que a análise de agrupamento pode lidar tanto com a formação de grupos de variáveis quanto de elementos (RODRIGUES et al., 2009, p.326). Além disso, a análise fatorial forma grupos baseando-se nos padrões de variação dos dados, isto é, na correlação, já a análise de agrupamento, pode tanto ser baseado na distância, como na correlação, dependendo do tipo de variável (HAIR et al., 2009, p.430).

2.1 Seleção da variável estatística de agrupamento

A escolha das variáveis deve ser feita baseada em aspectos teóricos, conceituais e práticos. Tendo em vista que estas devem tanto caracterizar os elementos como respeitar os objetivos da análise. Independentemente deste objetivo ser exploratório ou confirmatório os

grupos vão refletir a estrutura formada a partir das variáveis escolhidas. Portanto, a adição ou exclusão de variáveis relevantes podem ter um substancial efeito na solução resultante, como também a adição de variáveis irrelevantes, que podem provocar um aumento nas chances de criar *outliers* (RODRIGUES et al., 2009, p.328,330). Por isso, na seleção de variáveis o objetivo principal é escolher as que possuem maior relevância para o estudo, desprezando as demais (VALE, 2005, p.21).

Um procedimento (ou estratégia) que pode ser utilizado para este fim, é o de criar vários grupos utilizando a análise de clusters, onde os elementos em questão são as variáveis, que por sua vez, são colocadas e retiradas do estudo visando avaliar o efeito na solução final (KUBRUSLY, 2001, p.111-112). Escolhendo desta forma, as variáveis mais viáveis e conseqüentemente, as mais adequadas para o estudo. Após este procedimento ainda pode-se restringir a quantidade de variáveis, se necessário tendo por base a dificuldade, custo, precisão e importância prática (MINGOTI, 2013, p.190).

Se após o procedimento ainda exista um excesso de variáveis, inviabilizando a análise, pode-se excluir as que possuam alta ou baixa discriminação, isto é, as que possuam valores muito próximos em relação a todos os elementos ou muito distantes (ou ainda que são muito iguais ou muito diferentes). “Em último caso, podem-se ainda utilizar técnicas estatísticas para a redução da dimensionalidade da matriz de dados, tais como a Análise de Componentes Principais e a Análise Fatorial” (RODRIGUES et al., 2009, p.331).

2.2 Tratamento dos dados

Os dados coletados poderão ser representados em forma de tabela ou matriz ($e \times v$), em que as colunas(j) representem variáveis e as linhas(i) elementos, onde $j = 1, \dots, v$ e $i = 1, \dots, e$, da seguinte forma:

Tabela 1 – Estrutura referente aos valores observados, onde as linhas representam os elementos e as colunas as suas respectivas características (variáveis).

Elementos / Variáveis	X_1	X_2	\dots	X_v
E_1	x_{11}	x_{12}	\dots	x_{1v}
E_2	x_{21}	x_{22}	\dots	x_{2v}
\vdots	\vdots	\vdots	\ddots	\vdots
E_e	x_{e1}	x_{e2}	\dots	x_{ev}

Matriz representativa dos dados originais:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1v} \\ x_{21} & x_{22} & \cdots & x_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ x_{e1} & x_{e2} & \cdots & x_{ev} \end{bmatrix}$$

O tratamento dos dados tem por objetivo proporcionar a qualidade e eficiência no processo de agrupamento, tratando os *outliers* de forma que, os valores inválidos que se encontrem significativamente fora do esperado para uma variável sejam removidos; eliminando também os valores duplicados ou corrompidos, assim como, os valores faltantes ou inválidos (DONI, 2004, p.15). Entretanto, só faz sentido remover os valores faltantes se existirem muitos, contudo, existe outra forma para solucionar este problema, que é o de preencher estes valores com a média da respectiva variável (JESUS, 2015, p.16). Além de utilizar, se necessário, da transformação dos dados, que é subdividida em duas etapas: o tratamento de variáveis e a normalização (DONI, 2004, p.15).

2.2.1 Tratamento das variáveis (atributos)

As variáveis selecionadas podem tanto ser categóricas quanto quantitativas. As quantitativas assumem valores numéricos, podendo ser contínua (decimais) ou discreta (inteiros). As categóricas assumem valores finitos e podem ser binários, nominais ou ordinais, onde os elementos de suas variáveis precisam ser representados numericamente, para que os algoritmos possam ser utilizados (DONI, 2004, p.15). Sabendo que, mesmo com perda de informação, a variável quantitativa pode ser transformada em qualitativa, e a qualitativa em quantitativa (MINGOTI, 2013, p.160-161).

2.2.2 Normalização (padronização ou estandardização)

Antes do manuseio dos dados deve-se verificar se as variáveis possuem a mesma unidade de medida, caso contrário, será necessário padronizar, diminuindo assim, a variabilidade dos dados, de forma que estes possuam a mesma influência no processo. Em contrapartida, sem a sua utilização, à medida com maior escala torna-se dominante, influenciando na proximidade e conseqüentemente no agrupamento. Entretanto, existem casos, em que a diferença na influência das variáveis é significativa para o estudo, ou seja, podendo existir variáveis que possuam uma importância intrínseca superior, não sendo aconselhável a estandardização, pois este processo anula esta diferença, deixando todas as variáveis com mesmo peso (REIS, 2001, p.293).

Os processos de padronização mais utilizados são:

- i) z-score, calculado através da fórmula:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j},$$

onde, \bar{x}_j é a média e σ_j o desvio padrão das variáveis. Este método é de muita utilidade quando as médias ou variâncias dos dados em estudo, são diferentes (VALE, 2005, p.23). Também pode-se utilizar a divisão pelo desvio médio absoluto que é mais eficaz quando o conjunto de dados a ser analisado apresenta valores atípicos, neste caso, utiliza-se a equação do ponto acima, substituindo σ_j por s_j . Em que, o desvio médio absoluto para uma determinada variável V_j é calculado através da fórmula: $s_j = \sum_{i=1}^e |x_{ij} - \bar{x}_j|/e$, onde x_{ij} representa os valores referentes a cada elemento (E_i) em relação a variável em estudo e \bar{x}_j é a média associada a coluna j, isto é, a variável V_j (NUNES, 2006, p.30).

- ii) min-max cutoff, calculado através da fórmula:

$$\frac{x_{ij} - x_{min}}{x_{max} - x_{min}},$$

onde, x_{min} e x_{max} são o valor mínimo e máximo de cada variável, resultando em dados pertencentes ao intervalo [0,1] (VALE, 2005, p.23).

Outros processos são considerados igualmente validos, como, por exemplo, a padronização (utilizada no primeiro ponto) através da divisão pela média, pelo valor máximo ou mesmo pelo intervalo de variação (REIS, 2001, p.293).

2.3 Medida de Proximidade

A primeira etapa desta análise é a escolha de uma medida de parença para verificação de proximidade dos elementos. Esta é dividida em dois tipos, que são utilizados de acordo com a necessidade do pesquisador. A similaridade (quanto maior o valor, mais próximos estão os elementos) é utilizada para variáveis qualitativas e a dissimilaridade (quanto menor o seu valor, mais próximos estão os elementos) é utilizada para variáveis quantitativas (MINGOTI, 2013, p.157).

Nesta fase do procedimento, a matriz original ($e \times v$) é convertida em uma matriz quadrada ($e \times e$), através da medida de parença. Se as variáveis foram padronizadas, então, ao invés de utilizar a matriz original será utilizada a matriz padronizada, substituindo \mathbf{X} por \mathbf{Z} em todas as expressões matemáticas (VICINI, 2005, p.20).

2.3.1 Medidas de Similaridade

Os coeficientes de similaridade foram criados com o intuito de tratar das variáveis qualitativas, para verificar a proximidade de seus elementos, tendo em vista que, quanto

mais alto for o valor, mais próximo (ou mais similares) serão os elementos. Comparando-os de acordo com a presença ou ausência, que podem ser representados pela variável binária (0 e 1) (MINGOTI, 2013, p.160).

Tabela 2 – Tabela de contingência para os elemento f e t caracterizados pelas v variáveis binárias.

$E_t \backslash E_f$	1	0	Totais
1	a	b	a+b
0	c	d	c+d
Totais	a+c	b+d	v = a+b+c+d

Fonte: Retirada de (REIS, 2001, p.304).

Através da Tabela 2, pode-se observar a presença e ausência das características representadas pelas variáveis nos elementos f e t, onde o “a” representa o número de características que estão presentes nos dois elementos, isto é, o número de variáveis que possuem valor 1, o mesmo ocorre com o “d”, entretanto este representa o oposto, sendo o número de características que não estão presentes nos dois elementos, isto é, o número de variáveis que possuem o valor 0, já o “b” representa o número de características presentes no E_t e que não no E_f e o “c” representa, exatamente o oposto do “b”, o número de características ausentes no E_t e presentes no E_f . Na tabela 3 serão apresentados alguns coeficientes de similaridade, que utilizaram a teoria apresentada na tabela de contingência.

O coeficiente de similaridade produz uma matriz quadrada derivada dos elementos das variáveis estudadas. Transformando a matriz de dados originais \mathbf{X} na de similaridade \mathbf{C} .

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1v} \\ x_{21} & x_{22} & \cdots & x_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ x_{e1} & x_{e2} & \cdots & x_{ev} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} c_{(1,1)} & c_{(1,2)} & \cdots & c_{(1,e)} \\ c_{(2,1)} & c_{(2,2)} & \cdots & c_{(2,e)} \\ \vdots & \vdots & \ddots & \vdots \\ c_{(e,1)} & c_{(e,2)} & \cdots & c_{(e,e)} \end{bmatrix}$$

A matriz \mathbf{C} possui diagonal com valor 1, pois a proximidade de um elemento com ele mesmo é igual ao maior valor do intervalo de variação.

Tabela 3 – Coeficientes de similaridades entre dois elementos f e t, cujas realizações são respostas binárias em v variáveis, junto com o nome utilizado na literatura, explicação racional de seu significado e intervalo de variação correspondente.

Nome	C_{ft}	Explicação racional	Variação
Concordância Simples	$\frac{a+d}{v}$	Pesos iguais: 1-1 e 0-0	0-1
Sokal e Sneath	$\frac{2(a+d)}{2(a+d)+b+c}$	Pesos duplos: 1-1 e 0-0	0-1
Rogers e Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	Pesos duplos: 1-0 e 0-1	0-1
Russel e Rão	$\frac{a}{v}$	Ignora 0-0 no numerador	0-1
Jaccard	$\frac{a}{a+b+c}$	0-0 é irrelevante	0-1
Sorensen-Dice	$\frac{2a}{2a+b+c}$	0-0 é irrelevante e 1-1 tem duplo peso	0-1
Sem nome	$\frac{a}{a+2(b+c)}$	0-0 é irrelevante e duplo peso 1-0 e 0-1	0-1
Kulezynski	$\frac{a}{b+c}$	Razão entre concordâncias e discordâncias, excluindo 0-0	0-(v-1)
Sem nome	$\frac{a+d}{b+c}$	Razão entre concordâncias e discordâncias, incluindo 0-0	0-(v-1)
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Concordâncias positivas sobre adaptação da média geométrica de discordâncias	0-1
Baroni-Urbani-Buser	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	Concordâncias positivas e a média geométrica de concordâncias positivas e negativas	0-1
Ochiai II	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Razão entre produtos de concordâncias e média geométrica total modificada	0-1
Hamman	$\frac{(a+d)-(b+c)}{v}$	Diferença entre as proporções de concordâncias e discordâncias	-1-1
Yule	$\frac{ad-bc}{ad+bc}$	Diferença entre as proporções de ad e bc	-1-1
ϕ	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Produto de momento de correlação aplicado à variáveis binárias	-1-1

Fonte: Retirada de (FERREIRA, 2008, p.368)

2.3.2 Medidas de Dissimilaridade

“As medidas de distância são, na verdade, uma medida de dissimilaridade, com valores maiores denotando menor similaridade” (HAIR et al., 2009, p.442). Isto é, quanto

maior a distância menor será a proximidade entre os elementos.

Sejam três vetores, que possuem as características do elemento correspondente, \mathbf{f}, \mathbf{t} e $\mathbf{w} \in \mathfrak{R}^p$ e a matriz Ψ positiva definida (chamada de métrica), então a expressão geral para a distância quadrática entre os vetores \mathbf{f} e \mathbf{t} é dada por, $d^2(\mathbf{f}, \mathbf{t}) = \|\mathbf{f} - \mathbf{t}\|^2\Psi = (\mathbf{f} - \mathbf{t})^T\Psi(\mathbf{f} - \mathbf{t}) = \|r\|^2\Psi = r^T\Psi r$ (FERREIRA, 2008, p.92). Com isso, deve-se verificar se as seguintes propriedades são obedecidas (DONI, 2004, p.27):

$$\begin{cases} d(\mathbf{f}, \mathbf{t}) \geq 0, \text{ onde } d(\mathbf{f}, \mathbf{t}) = 0, \text{ se e somente se, } \mathbf{f} = \mathbf{t} \\ d(\mathbf{f}, \mathbf{t}) = d(\mathbf{t}, \mathbf{f}) \\ d(\mathbf{f}, \mathbf{t}) \leq d(\mathbf{f}, \mathbf{w}) + d(\mathbf{t}, \mathbf{w}). \end{cases}$$

A terceira propriedade mostrada através dos sistema acima, também chamada de desigualdade triangular é derivada da desigualdade de Cauchy-Schwarz, o que pode ser provado, mais abaixo (FERREIRA, 2008, p.92,93):

$$\begin{aligned} d^2(\mathbf{f}, \mathbf{t}) &= \|\mathbf{f} - \mathbf{t}\|^2\Psi \stackrel{(1)}{=} \|(\mathbf{f} - \mathbf{w} + \mathbf{w} - \mathbf{t})\|^2\Psi \stackrel{(2)}{=} (\mathbf{f} - \mathbf{w} + \mathbf{w} - \mathbf{t})^T\Psi(\mathbf{f} - \mathbf{w} + \mathbf{w} - \mathbf{t}) \\ &\stackrel{(3)}{=} [(\mathbf{f} - \mathbf{w}) + (\mathbf{w} - \mathbf{t})]^T\Psi[(\mathbf{f} - \mathbf{w}) + (\mathbf{w} - \mathbf{t})] \\ &\stackrel{(4)}{=} (\mathbf{f} - \mathbf{w})^T\Psi(\mathbf{f} - \mathbf{w}) + 2(\mathbf{f} - \mathbf{w})^T\Psi(\mathbf{w} - \mathbf{t}) + (\mathbf{w} - \mathbf{t})^T\Psi(\mathbf{w} - \mathbf{t}) \\ &\stackrel{(5)}{=} \|\mathbf{f} - \mathbf{w}\|^2\Psi + 2(\mathbf{f} - \mathbf{w})^T\Psi(\mathbf{w} - \mathbf{t}) + \|\mathbf{w} - \mathbf{t}\|^2\Psi \implies \\ d^2(\mathbf{f}, \mathbf{t}) &\stackrel{(6)}{\leq} \|\mathbf{f} - \mathbf{w}\|^2\Psi + 2\|(\mathbf{f} - \mathbf{w})\|\Psi\|(\mathbf{w} - \mathbf{t})\|\Psi + \|\mathbf{w} - \mathbf{t}\|^2\Psi = \\ &= (\|\mathbf{f} - \mathbf{w}\|^2\Psi + \|\mathbf{w} - \mathbf{t}\|^2\Psi)^2 \implies \\ d^2(\mathbf{f}, \mathbf{t}) &\stackrel{(7)}{\leq} (\|\mathbf{f} - \mathbf{w}\|\Psi + \|\mathbf{w} - \mathbf{t}\|\Psi)^2 \implies \\ d(\mathbf{f}, \mathbf{t}) &\leq \|\mathbf{f} - \mathbf{w}\|\Psi + \|\mathbf{w} - \mathbf{t}\|\Psi = d(\mathbf{f}, \mathbf{w}) + d(\mathbf{w}, \mathbf{t}) \end{aligned}$$

Na prova acima, pode-se observar alguns números entre parenteses que foram utilizados com o objetivo de explicar cada etapa, onde, (1) Acrescenta-se um \mathbf{w} e $-\mathbf{w}$, que se anulam (não possuindo influência sobre o resultado); (2) Considera-se $r = (\mathbf{f} - \mathbf{w} + \mathbf{w} - \mathbf{t})$; (3) Calcula-se a propriedade distributiva da multiplicação; (4) Substitui $(\mathbf{f} - \mathbf{w})^T\Psi(\mathbf{f} - \mathbf{w})$ por $\|\mathbf{f} - \mathbf{w}\|^2\Psi$ e $(\mathbf{w} - \mathbf{t})^T\Psi(\mathbf{w} - \mathbf{t})$ por $\|\mathbf{w} - \mathbf{t}\|^2\Psi$, de acordo com a formula: $d^2(\mathbf{f}, \mathbf{t}) = \|\mathbf{f} - \mathbf{t}\|^2\Psi = (\mathbf{f} - \mathbf{t})^T\Psi(\mathbf{f} - \mathbf{t})$; (5) Utiliza-se a formula da desigualdade de Cauchy-Schwarz, $\mathbf{h}^T\Psi\mathbf{l} = \|\mathbf{h}\|\Psi\|\mathbf{l}\|$; (6) Utiliza-se o produto notável do quadrado da soma; (7) Tirar a raiz quadrada de ambos os lados.

Entretanto, a terceira condição é considerada suficiente e não, necessária, para a realização do agrupamento. Logo, pode-se sem grandes consequências formar agrupamentos com medidas que não formam uma métrica (FERREIRA, 2008, p.359).

A medida de dissimilaridade produz uma matriz quadrada das distâncias dos elementos das variáveis estudadas. Transformando a matriz de dados originais \mathbf{X} na matriz das distâncias \mathbf{D} (com diagonal possuindo valor 0).

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1v} \\ x_{21} & x_{22} & \cdots & x_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ x_{e1} & x_{e2} & \cdots & x_{ev} \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} d_{(1,1)} & d_{(1,2)} & \cdots & d_{(1,e)} \\ d_{(2,1)} & d_{(2,2)} & \cdots & d_{(2,e)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(e,1)} & d_{(e,2)} & \cdots & d_{(e,e)} \end{bmatrix}$$

As distancias mais utilizadas para variáveis quantitativas podem ser observadas através da Tabela 4 .

Tabela 4 – Distâncias entre dois elementos f e t, levando em consideração apenas duas variáveis.

Distâncias	função	Explicação
Euclidiana	$D(E_f, E_t) = \sqrt{(x_{1t} - x_{1f})^2 + (x_{2t} - x_{2f})^2}$, em que, $j=1,2; t, f \in i$ e $t > f$	É o comprimento da hipotenusa de um triângulo retângulo
Euclidiana quadrada (ou absoluta)	$D(E_f, E_t) = (x_{1t} - x_{1f})^2 + (x_{2t} - x_{2f})^2$, em que, $j=1,2; t, f \in i$ e $t > f$	É idêntica a euclidiana com excessão da raiz. Recomendada para os metodos centróide e Ward.
City-block (de Manhattan)	$D(E_f, E_t) = x_{1t} - x_{1f} + x_{2t} - x_{2f} $, em que, $j=1,2; t, f \in i$ e $t > f$	É a soma das diferenças absolutas das variáveis, isto é, a soma dos valores absolutos dos catetos de um triângulo retângulo.
Chebychev	$D(E_f, E_t) = \max(x_{1t} - x_{1f} , x_{2t} - x_{2f})$, em que, $j=1,2; t, f \in i$ e $t > f$	É a maior diferença entre dois elementos ao longo de todas as variáveis, em valor absoluto.
Mahalanobis (D^2)	$D(E_f, E_t) = \sqrt{(\vec{E}_t - \vec{E}_f)' \Sigma^{-1} (\vec{E}_t - \vec{E}_f)}$, em que, $t, f \in i$	É a distancia entre dois vetores de elementos (ou variáveis), através da raiz da forma quadratica $x'Ax$, onde, x é a sua diferença e A é a inversa da matriz de covariância residual.

Fonte: A segunda coluna foi retirada de (DONI, 2004, p.27-29) e (VALE, 2005, p.28), enquanto que, a terceira foi retirada de (HAIR et al., 2009, p.442).

De acordo com Rodrigues et al. (2009, p.336-337), a escolha da distância depende do tipo de escala da variável, para dados intervalares, a distância Euclidiana, Euclidiana Quadrada, Cosine, Correlação de Pearson, Chebyshev, Block, Minkowski e Customizada; para dados nominais, a Qui-quadrado e Phi-quadrado; e para dados binários a distância Euclidiana, Euclidiana Quadrada, Size Difference, Variance, Dispersion, Shape, Lambda, Jaccard, entre outros.

Para tentar escolher uma medida de proximidade adequada para os dados é necessário voltar a atenção para alguns pontos (HAIR et al., 2009, p.431):

- i) Diferentes escolhas a cerca da medida de distância, assim como da seleção das variáveis, e até mesmo da técnica utilizada, podem gerar diferentes soluções. Portanto, é aconselhável utilizar varias medidas e técnicas para comparar os seus respectivos resultados com padrões teóricos ou conhecidos.
- ii) Quando as variáveis são correlacionadas a medida mais adequada a ser utilizada é a de Mahalanobis, pois pondera as variáveis de forma igualitária, ajustando as correlações.

2.4 Técnicas de Agrupamento

As técnicas ou métodos de agrupamento são formas distintas de agrupar, utilizando a matriz de proximidade, formada a partir da medida de parença escolhida. Logo, diferentes agrupamentos podem ser formados. Essas formas distintas, se devem ao fato de que, existem diferentes meios de se definir proximidade entre indivíduos de um grupo, ou entre grupos de indivíduos. Para encontrar a técnica mais adequada para os dados ou problema em questão é necessário utilizar e comparar os resultados dos algoritmos entendendo as características individuais dos grupos, pois ainda não foi encontrado uma técnica que possa ser generalizada e aceita como a melhor (VICINI, 2005, p.24). Visto que, ainda não foi encontrado um algoritmo que apresente todas as características a seguir:

“(1)Ser capaz de lidar com dados com alta dimensionalidade; (2)Ser ‘escalável’ com o número de dimensões e com a quantidade de elementos a serem agrupados; (3)Habilidade para lidar com diferentes tipos de dados; (4) Capacidade de definir agrupamentos de diferentes tamanhos e formas; (5)Exigir o mínimo de conhecimento para determinação dos parâmetros de entrada; (6) Ser robusto à presença de ruído; (7)Apresentar resultado consistente independente da ordem em que os dados são apresentados” (DONI, 2004, p.27).

Existem dois tipos de técnicas de agrupamento, as hierárquicas e as particionais. As hierárquicas são divididas em aglomerativas e divisivas, da mesma forma, que as não-hierárquicas(ou particionais) são divididas em exclusivas e não-exclusivas. O método hierárquico aglomerativo inicia o seu procedimento com “e” grupos de apenas um elemento finalizando com um único grupo possuindo todos os elementos. Ao contrario do método hierárquico divisivo, que começa com todos os elementos em um único grupo e finaliza com “e” grupos de apenas um elemento (VALE, 2005, p.33,38). Entretanto, este trabalho será focado nos métodos hierárquicos aglomerativos.

2.4.1 Métodos Hierárquicos Aglomerativos

Os métodos aglomerativos são constituídos de algoritmos de agrupamento que são intitulados de métodos de encadeamento, métodos de erros de somas de quadrados (ou métodos de variância) e métodos centróides, segundo a sua finalidade (VICINI, 2005, p.26).

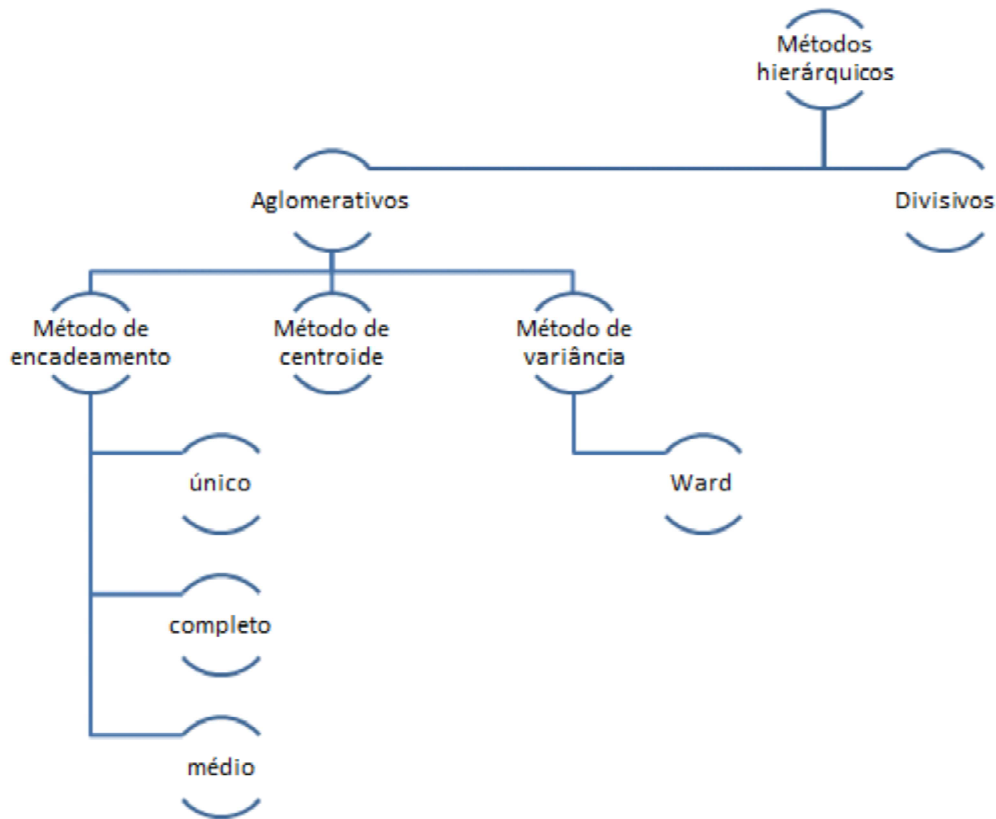


Figura 2 – Nomenclatura referente às Técnicas Hierárquicas Aglomerativas mais utilizadas.

Um algoritmo geral para este método seria (SARTORIO, 2008, p.55): (1) Inicia-se com o número de grupos ($g = e$) definido pela quantidade de elementos (ou variáveis) e com uma matriz de distância quadrada e simétrica (\mathbf{D}), derivada dos dados originais, de ordem e ; (2) Procura-se a menor distância na matriz \mathbf{D} , isto é, os grupos mais próximos (mais semelhantes), f e t , neste caso; (3) Define-se um novo grupo, através da união dos grupos mais parecidos, (ft) e calcula-se uma nova distância para esse grupo, utilizando tanto a distancia quanto o método escolhido, eliminando as linhas e colunas dos grupos f e t individuais e substituindo-as por uma linha e coluna com as distâncias entre o novo grupo e os demais. (4) Repetir esse processo até que todos os elementos façam parte de apenas um grupo, renomeando os grupos de acordo com as fusões e anotando as novas distâncias.

O que diferencia os métodos aglomerativos é o passo (3), onde redefine-se a matriz de proximidade a cada união de grupos (BARROSO; ARTES, 2003 apud SARTORIO, 2008, p.56). Que podem ser calculados através dos processos demonstrados na Tabela 5.

Tabela 5 – Resumo dos métodos hierárquicos aglomerativos, onde, N_f , N_t e N_w são os números de elementos nos grupos f , t e w .

Método	Distância	Características
Ligação simples	$d_{(ft)w} = \min(d_{fw}, d_{tw})$	Sensibilidade à ruídos. Encadeamento
Ligação completa	$d_{(ft)w} = \max(d_{fw}, d_{tw})$	Tendência a formar grupos compactos.
Ligação por média	$d_{(ft)w} = \frac{N_f d_{fw} + N_t d_{tw}}{N_f + N_t}$	Tendência a formar grupos com número de elementos similares.
Ligação por centróide	$d_{(ft)w} = \frac{N_f d_{fw} + N_t d_{tw}}{N_f + N_t} - \frac{N_f N_t d_{ft}}{(N_f + N_t)^2}$	Robustez à ruídos. Reversão.
Ligação por mediana	$d_{(ft)w} = \frac{d_{fw} + d_{tw}}{2} - \frac{d_{ft}}{4}$	Robustez à ruídos.
Ligação de ward	$d_{(ft)w} = \frac{(N_w + N_f)d_{fw} + (N_w + N_t)d_{tw} - N_w d_{ft}}{N_f + N_t + N_w}$	Sensibilidade à ruídos.

Fonte: Retirada de (DONI, 2004, p.46)

Os métodos de encadeamento podem ser utilizados tanto para variáveis quantitativas, como para qualitativas, entretanto, os métodos de centroide e de variância são apropriados apenas para variáveis quantitativas, pois são baseados na comparação de vetores de médias (BARROSO; ARTES, 2003 apud SARTORIO, 2008, p.57).

Por se tratar de um método hierárquico, os grupos formados através dos métodos hierárquicos aglomerativos podem ser representados graficamente por meio do dendrograma (ou diagrama de árvore) que é considerado uma síntese gráfica bi-dimensional, ocasionando em perda de informação, entretanto, é de grande valia na comparação e discussão dos agrupamentos (VICINI, 2005, p.14-15).

2.4.1.1 Ligação Simples

O método de ligação simples (ou de encadeamento único) baseia-se na distância mínima entre os grupos (VICINI, 2005, p.26), “ou seja, a distância entre os grupos é definida como sendo aquela entre os objetos mais parecidos entre esses grupos” (MEYER, 2002, p.16). Algumas características deste método são:

- i) Pode definir uma grande quantidade de padrões aglomerativos devido a sua versatilidade, possibilitando a ocorrência de problemas quando os grupos são mal delineados, podendo formar longas e sinuosas cadeias (HAIR et al., 2009, p.450);
- ii) Reuni objetos ao elemento mais próximo do grupo já formado, fazendo com que os objetos intermediários entre os grupos sejam facilmente aderidos a esses, ocorrendo um

encadeamento dos objetos, dificultando assim a separação dos grupos (VALENTIN, 1995, p.38).

- iii) “Os dendrogramas resultantes deste procedimento são geralmente pouco informativos, devido à informação dos indivíduos intermediários que não são evidentes” (MEYER, 2002, p.16).

2.4.1.2 Ligação Completa

A método ligação completa (ou de encadeamento completo) baseia-se na distância máxima entre os elementos (VICINI, 2005, p.26), ou seja, a distância entre os grupos é definida como sendo aquela entre os elementos mais diferentes entre esses grupos (MEYER, 2002, p.16-17). Logo, esta é o oposto da anterior. Algumas características deste método são:

- i) “Este método, geralmente, leva a grupos compactos e discretos, tendo os seus valores de similaridade relativamente pequenos” (MEYER, 2002, p.17);
- ii) Não possui problemas com encadeamento (HAIR et al., 2009, p.450), pois um elemento só se fundirá a um grupo se este for ligado a todos os elementos deste grupo, por conseguinte, a medida que os grupos crescem, torna-se cada vez mais difícil a aderência de elementos (VALENTIN, 1995, p.38).

Como os métodos de ligação simples e completa trabalham em direções opostas, pode-se verificar se o grupo é real, isto é, se ele é bem definido no espaço, comparando os seus resultados. Se foram semelhantes, é real, caso contrário, não (HOMESBURG, 1984 apud DONI, 2004, p.38).

2.4.1.3 Ligação média

O método de ligação média (ou de encadeamento médio) é calculado “pela distância média entre todos os pares de objetos (elementos) dos dois diferentes grupos” (FERREIRA, 2008, p.381). De acordo com Hair et al. (2009, p.451) este método possui as seguintes características:

- i) O que diferencia este método dos dois anteriores é que o seu algoritmo independe de valores externos, como pares mais próximos ou afastados (que é o caso do método de ligação simples e completa), sendo baseado nos elementos dos grupos (agregados) e não focando-se em um único par de membros externos, proporcionando a diminuição da influência das observações atípicas;
- ii) Tendem a produzir agregados (grupos) com pouca variação interna e com aproximadamente a mesma variância.

2.4.1.4 Ligação Ward

A ligação de Ward baseia-se na maximização da homogeneidade, ou melhor, diminuição da variância, através da soma dos quadrados da diagonal principal da matriz de covariância dentro de cada grupo (NUNES, 2006, p.34).

A agregação neste procedimento é baseada na minimização da soma interna de quadrados referente ao conjunto completo de agrupamentos, isto é, os grupos unidos(fundidos) minimizam o aumento da soma total de quadrados em todas as variáveis e agrupamentos (HAIR et al., 2009, p.452).

De acordo com Hair et al. (2009, p.452), este método possui as características a seguir:

- i) Não utiliza uma única medida de proximidade, mas a soma dos quadrados dentro dos grupos feita sobre todas as variáveis;
- ii) Por ser baseada na soma de quadrados, tende a combinar agrupamentos com pequenas quantidades de observações(elementos), além de ser facilmente distorcido por observações atípicas, isto é, valores discrepantes;
- iii) Tendem a juntar grupos com mesmo número de elementos.

2.5 Métodos para encontrar o número g de clusters(ou número de corte)

“A escolha do número final de grupos (g) em que o conjunto de dados deve ser repartido é subjetiva. Existem alguns métodos que podem ser utilizados para auxiliar na determinação de g ” (MINGOTI, 2013, p.179). Um deles é através da medida de heterogeneidade, que aumenta a cada combinação de grupos, podendo ser calculada através da média de todas as distâncias entre os elementos dentro dos agrupamentos. Quando esta torna-se elevada a medida em que unem-se dois agrupamentos, nota-se a possibilidade destes serem diferentes(distantes ou não similares), indicando, desta forma, o número g (HAIR et al., 2009, p.434-435).

Outra forma é utilizando do índice de Calinski-Hanabasz(CH) criado por Calinski e Harabasz em 1974, que compara a homogeneidade interna e o heterogeneidade externa dos grupos, através da formula:

$$CH(g) = \frac{e - g}{g - 1} \frac{SSB}{SSW},$$

onde, g é o número de grupos; e é o número de observações no conjunto de dados; SSB (Sum of Squares Between groups) é a soma dos quadrados entre grupos e SSW (Sum

of Squares Within) é soma dos quadrados dentro dos clusters. Ou seja, o SSW avalia a dispersão dentro dos clusters e o SSB avalia a dispersão entre grupos, comparando a soma da distância quadrática dentro dos grupos com a entre os grupos (JESUS, 2015, p.26). Entretanto, neste trabalho já se tem um número prévio de grupos ($g=4$).

2.6 Métodos de validação

2.6.1 Número de grupos

Mesmo já possuindo o número de grupos que foi previamente definido, este critério de escolha não é satisfatório por se tornar viesado. Entretanto, existe um método alternativo que compara, graficamente, o número de grupos já definidos com o coeficiente de fusão, que é o valor numérico (a distância ou semelhança) que unem os grupos, tendo como critério de escolha ótima, a não ocorrência de alterações significativas no coeficiente após a união de dois grupos. Entretanto, este também pode ser o seu ponto fraco, caso os pequenos saltos sejam frequentes na observação do gráfico, impossibilitando a visualização do melhor número de corte, isto é, inviabilizando a comparação do número de grupos já definidos com o número selecionado através do coeficiente de fusão (REIS, 2001, p.325).

Além destes, ainda existem vários outros índices que podem ser utilizados para validar e/ou encontrar o número de grupos, assim como os índices de: Davies Bouldin, C index, Dunn, Gamma, G plus, GDI, McClain Rao, PBM, Point Biserial, Ray Turi, SD e Xie Beni (JESUS, 2015, p.24-34).

2.6.2 Técnica hierárquica

Como não se tem um método fixado, tido como ótimo, ou até melhor do que os demais, faz-se comparações destes, para verificar qual se adéqua melhor aos dados, utilizando o dendrograma, que mostra graficamente a formação dos grupos, ocasionando em perda de informação. Para verificar, qual método obteve menor perda, isto é, qual dendrograma (ou grupos) mais se aproximam da realidade, utiliza-se do coeficiente de correlação cofenética, que avalia o grau de distorção provocado pela formação do dendrograma (VICINI, 2005, p.52).

A correlação cofenética é definida “como sendo a correlação entre as distâncias previstas e as efetivamente observadas. Quanto mais próxima de um, melhor será a qualidade do agrupamento” (SARTORIO, 2008, p.59), servindo para medir o grau de ajuste entre a matriz fenética e a cofenética. Esta pode ser calculada através da seguinte formula (VICINI, 2005, p.53,58):

$$r_{cof} = \frac{\sum_{f=1}^{v-1} \sum_{t=f+1}^v (d_{ft} - \bar{d})(k_{ft} - \bar{k})}{\sqrt{\sum_{f=1}^{v-1} \sum_{t=f+1}^v (d_{ft} - \bar{d})^2} \sqrt{\sum_{f=1}^{v-1} \sum_{t=f+1}^v (k_{ft} - \bar{k})^2}}, \text{ em que}$$

k_{ft} : valor da distância entre os elementos f e t, obtidos por meio da matriz \mathbf{K} , que é a matriz resultante após a utilização do método(matriz cofenética);

d_{ft} : valor da distância entre os elementos f e t, obtidos por meio da matriz \mathbf{D} , que é a matriz de dissimilaridade original(matriz fenética);

$$\bar{k} = \frac{2}{v(v-1)} \sum_{f=1}^{v-1} \sum_{t=f+1}^v k_{ft}; \quad \bar{d} = \frac{2}{v(v-1)} \sum_{f=1}^{v-1} \sum_{t=f+1}^v d_{ft}$$

Matricialmente seria,

$$r_{cof} = \frac{Cov(\mathbf{K}, \mathbf{D})}{\sqrt{\hat{Var}(\mathbf{K})\hat{Var}(\mathbf{D})}}, \text{ onde,}$$

$Cov(\mathbf{K}, \mathbf{D})$: covariância entre os elementos da matriz cofenética e fenética;

$\hat{Var}(\mathbf{K})$: variância dos elementos da matriz cofenética;

$\hat{Var}(\mathbf{D})$: variância dos elementos da matriz fenética.

3 Material e Métodos

As variáveis foram escolhidas segundo as causas que influenciam na diminuição (ou aumento) do índice de mortalidade infantil, que são: Saneamento básico (X_1 = água encanada; X_2 = coleta de lixo; X_3 = rede de esgoto), qualidade de vida (X_4 = IDHM; X_5 = índice de Gini; X_6 = proporção de residências com oito pessoas) e o índice pluviométrico (X_7), que foi acrescentada devido a seca. Também consideradas como variáveis socioeconômicas.

Alguns conceitos são necessários para compreensão no que se refere as variáveis que são:

1. IDHM (Índice de Desenvolvimento Humano Municipal): mede a qualidade de vida das pessoas por meio de três indicadores que são: longevidade, renda e educação, variando de 0 a 1.
2. Índice de Gini: mede o grau de concentração da renda, variando de 0 a 1, com 0 representando igualdade na renda dos pobres e dos ricos e 1 a total desigualdade entre as mesmas.
3. Índice Pluviométrico (medida em milímetros): é a soma de precipitação de água através da chuva em um dado local e período de tempo.

Os elementos (cidades) foram escolhidas por meio de um critério de seleção, que foi baseado nas cidades mais populosas, escolhendo dez de cada mesorregião (Meso1 = Sertão Paraibano, Meso2 = Borborema, Meso3 = Agreste Paraibano e Meso4 = Mata Paraibana) do estado da Paraíba, que podem ser visualizados na Tabela 6.

Os dados são referentes ao ano de 2010, isto é, o último ano do censo, estes foram coletados no banco de dados da Agência Executiva de Gestão das Águas do Estado da Paraíba (Aesa), do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Atlas do Desenvolvimento Humano no Brasil (Atlas Brasil) e manipulados por meio do software (R Core Team, 2017).

Tabela 6 – Municípios mais populosos do estado da Paraíba, com suas representações numéricas e respectivas mesorregiões.

Meso-região	Cidades	Meso-região	Cidades
Meso1	1.Patos	Meso3	21.Campina Grande
Meso1	2.Sousa	Meso3	22.Guarabira
Meso1	3.Cajazeiras	Meso3	23.Queimadas
Meso1	4.Pombal	Meso3	24.Esperança
Meso1	5.São Bento	Meso3	25.Alagoa Grande
Meso1	6.Catolé do Rocha	Meso3	26.Solânea
Meso1	7.Itaporanga	Meso3	27.Lagoa Seca
Meso1	8.Princesa Isabel	Meso3	28.Itabaiana
Meso1	9.São José de Piranhas	Meso3	29.Areia
Meso1	10.Conceição	Meso3	30.Bananeiras
Meso2	11.Monteiro	Meso4	31.João Pessoa
Meso2	12.Picuí	Meso4	32.Santa Rita
Meso2	13.Boqueirão	Meso4	33.Bayeux
Meso2	14.Juazeirinho	Meso4	34.Cabedelo
Meso2	15.Sumé	Meso4	35.Sapé
Meso2	16.Taperoá	Meso4	36.Mamanguape
Meso2	17.Santa Luzia	Meso4	37.Pedras de fogo
Meso2	18.Serra Branca	Meso4	38.Rio tinto
Meso2	19.Seridó	Meso4	39.Conde
Meso2	20.Barra de Santana	Meso4	40.Mari

3.1 Distância de Mahalanobis

“Uma medida da Distância Euclidiana usada que diretamente incorpora o procedimento de padronização é a Distância de Mahalanobis (\mathbf{D}^2)” (RODRIGUES et al., 2009, p.339).

A partir da distância generalizada ou ponderada, que é calculada através da formula $D(E_f, E_t) = \sqrt{(\vec{E}_t - \vec{E}_f)' \mathbf{B} (\vec{E}_t - \vec{E}_f)}$, pode-se obter informações sobre outras distâncias, pois se $\mathbf{B}=\mathbf{I}$ (Identidade), tem-se a distância euclidiana, assim como, se $\mathbf{B}=\mathbf{S}^{-1}$ (inversa da matriz de covariância amostral), tem-se a distância de Mahalanobis. Logo, nota-se que \mathbf{B} reflete o tipo de informação que deseja-se utilizar na ponderação das diferenças dos vetores E_f e E_t (MINGOTI, 2013, p.157-158).

Como neste caso, $\mathbf{B}=\mathbf{S}^{-1}$ leva-se em consideração através da ponderação, “as possíveis diferenças de variâncias e as relações lineares entre as variáveis, medidas em termos de covariância” (MINGOTI, 2013, p.158), isto é, além de desenvolver um processo de padronização, que pondera igualmente as variáveis, ainda ajusta as intercorrelações entre as mesmas, sendo portanto uma melhor opção quando as variáveis estão inter-relacionadas. Entretanto, se o objetivo do pesquisador for a ponderação desigual, a outros meios disponíveis (RODRIGUES et al., 2009, p.339).

Com isso, pode-se dizer que, a distância de Mahalanobis é calculada através da seguinte fórmula: $D(E_f, E_t) = \sqrt{(\vec{E}_t - \vec{E}_f)' \mathbf{S}^{-1} (\vec{E}_t - \vec{E}_f)}$. Sabendo que, se a matriz \mathbf{S}^{-1} for padronizada deixará de ser matriz de covariância amostral e passará a ser uma matriz de correlação amostral. Caso as correlações forem nulas, restará a matriz identidade, tornando a distância de Mahalanobis equivalente a euclidiana (VICINI, 2005, p.23), ocorrendo da mesma forma se a matriz \mathbf{S}^{-1} possuir covariâncias nulas e variâncias iguais para todas as variáveis (VALE, 2005, p.28).

“Admitindo-se distribuição multinormal p-dimensional, homogeneidade na matriz de variância-covariância nas unidades amostrais, pode-se chamar distância generalizada de Mahalanobis” (VICINI, 2005, p.23).

3.2 Métodos hierárquicos aglomerativos

Como forma de redefinir a matriz de proximidade foram utilizados os métodos de ligação simples, completa, média e Ward, comparando-se os resultados por meio do dendrograma. Por se tratar de uma sintaxe gráfica bi-dimensional o dendrograma possui perda de informações. Com o intuito de medir o grau da qualidade do agrupamento foi utilizado o Coeficiente de Correlação Cofenética.

4 Aplicação

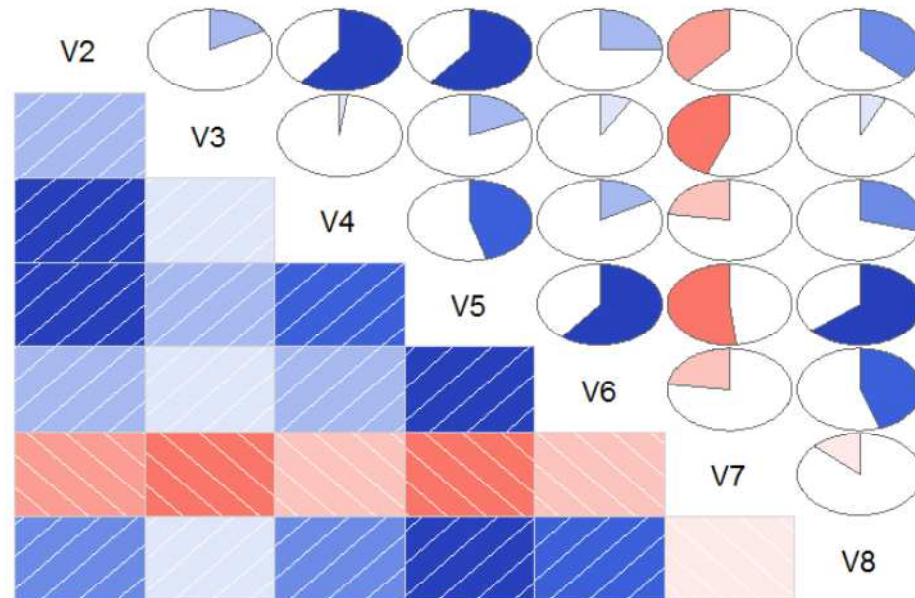


Figura 3 – Gráfico de correlação referente as 8 variáveis estudadas (que são: V2=proporção de residências com água encanada; V3=proporção de residências que possuem coleta de lixo; V4=proporção de residências que possuem rede de esgoto; V5=IDHM; V6=índice de Gini; V7=proporção de residências com oito pessoas; V8=índice pluviométrico).

O diagrama representado por meio da Figura 3, mostra que os vermelhos apresentam retas voltadas para baixo, enquanto que, os azuis voltados para cima, indicando que os vermelhos representam uma correlação linear negativa, e os azuis, em oposição, positivas. Também nota-se que quanto mais clara for a cor menor é a correlação entre as devidas variáveis, isto é, mais próxima de zero. Após estas considerações, observa-se que há uma quantidade maior de azuis que de vermelhos, expondo um número maior de correlações positivas. Porém, quando trata-se da quantidade de correlação existente entre as variáveis, pode-se dizer que há mais claros que escuros, isto é, ou a maioria das variáveis possuem baixa correlação entre si ou elas são independentes uma das outras.

Levando-se em consideração a quantidade, pode-se dizer que há mais azuis que vermelhos. Porém, quando olha-se a qualidade, isto é, a quantidade de correlação existente entre as variáveis pode-se dizer que há mais claros que escuros.

Por conseguinte, pode-se concluir que a correlação entre X_2 e X_3 é quase nula, mais especificamente 0.02, que pode ser considerada uma independência entre estas variáveis,

levando a crer que, não existe correlação entre a quantidade de residências com coleta de lixo e a quantidade de residências com rede de esgoto. Seguida da correlação entre X_2 e X_5 e X_2 e X_7 , que possuem os menores valores, mostrando que X_2 só possui uma correlação moderada, e negativa, com X_6 , mais especificamente, que há uma relação entre a quantidade de residências com coleta de lixo e a número de residências com oito pessoas, ou seja, quanto menor for a quantidade de residências com coleta de lixo, maior será a proporção de residências com oito pessoas. Em contrapartida, as variáveis mais correlacionadas do estudo, porém, moderadamente, são X_1 com X_3 e X_4 , revelando que as casas que possuem água encanada geralmente dispõem de uma melhor qualidade de vida, assim como, de uma rede de esgoto; e V5(IDHM) com X_1 , X_3 , X_5 , X_6 e X_7 , expondo, uma relação entre o IDHM e as outras variáveis(o que pode ser explicado, por se tratar de uma variável que mede a qualidade de vida da população, utilizando de indicadores como renda, longevidade e educação), exceto X_2 . Os valores numéricos destas correlações estão descritos na Tabela 7. Entretanto, sabe-se que estas correlações necessitam ter um peso estatístico, isto é, precisa-se verificar a significância dessas correlações.

Para saber qual teste de correlação utilizar, procedeu-se à comparação de dois testes para verificar a normalidade, o teste de Shapiro-Wilk e Anderson Darling, o que resultou na concordância, mostrando que apenas a variável X_6 (Proporção de residências com 8 pessoas) possui distribuição normal, excluindo a possibilidade do teste de correlação de Pearson. Como a maioria das variáveis é não paramétrica, resta a possibilidade de dois testes, o teste de correlação de Kendall, que utiliza os dados coletados ao longo do tempo e o teste de correlação Spearman, que utiliza os escores. Por este motivo foi utilizado o teste de Spearman, já que os dados coletados são restritos ao ano de 2010.

Não foi utilizado o teste de normalidade multivariado, devido ao fato dos testes de correlação utilizados serem univariados, além, deste não ser o foco da análise, já que foi utilizado apenas como critério de seleção do teste de correlação mais adequado.

De acordo com a Tabela 7, boa parte das hipótese de nulidade das correlações entre as variáveis foi rejeitada, ao nível de 5% de significância, validando assim, a utilização da análise multivariada, pois esta inclui a correlação existente entre as variáveis em sua análise, sendo, portanto, mais adequada.

A Figura 4a refere-se ao perfil de todas as variáveis. Neste, nota-se que os valores de X_7 superam o das outras variáveis, supondo assim que sua média será alta e provavelmente o seu desvio padrão também. Pois, os seus valores estão entre 458,2 à 4743,4, além de possuir dois pontos discrepantes, referentes aos municípios de João Pessoa(31) e Campina Grande(21), com valores 4743,4 e 3186,3, os outros elementos possuem valores abaixo de 2000, dando a entender que nesses municípios o índice pluviométrico é maior, isto é, há uma ocorrência maior de chuvas. Na Figura 4b, quando este é retirado, percebe-se a diminuição dos valores da escala, de mil foi para 0,2, este gráfico é dividido nos dois

Tabela 7 – Resultado do teste de correlação de Spearman.

Variáveis		Estatística de teste	ρ	valor p
X_1	X_2	9284,9	0,12899	0,4276
X_1	X_3	4344	0,59249	0,0001
X_1	X_4	4392,6	0,58793	0,0001
X_1	X_5	8871,1	0,16781	0,3006
X_1	X_6	14625	-0,37196	0,0181
X_1	X_7	7040	0,33958	0,0326
X_2	X_3	9927	0,06876	0,6733
X_2	X_4	8097,9	0,24034	0,1352
X_2	X_5	9352,8	0,12262	0,4509
X_2	X_6	14324	-0,34373	0,0299
X_2	X_7	11631	-0,09109	0,5762
X_3	X_4	5963,2	0,44059	0,0044
X_3	X_5	9601	0,09934	0,5419
X_3	X_6	13090	-0,22792	0,1572
X_3	X_7	9136	0,14296	0,3775
X_4	X_5	6300,4	0,40896	0,0088
X_4	X_6	17343	-0,62695	0,00002
X_4	X_7	6502,4	0,39001	0,0128
X_5	X_6	13165	-0,23497	0,1444
X_5	X_7	7934,9	0,25563	0,1114
X_6	X_7	11639	-0,09186	0,5729

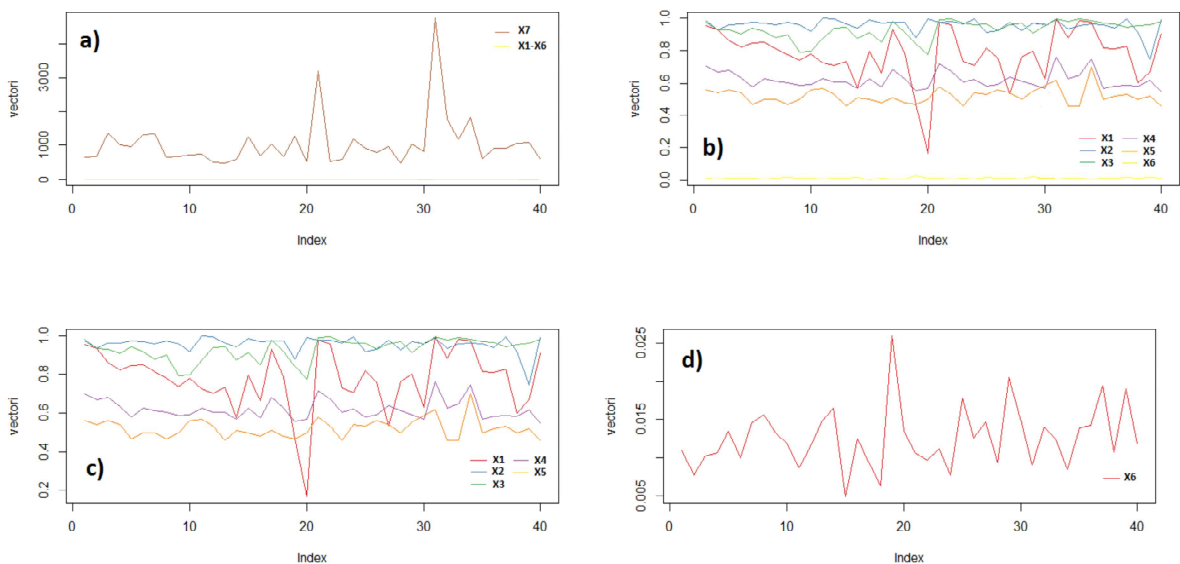


Figura 4 – Comportamento das variáveis em estudo, revelado por intermédio do Gráfico de perfil.

próximos, para uma melhor visualização, podendo assim observar, na Figura 4c, as variáveis

de X_1 a X_5 , onde a X_1 mostra-se mais variante, expondo um ponto discrepante, indicando um município, que neste caso é Barra de Santana(20), que possui menor acesso a água encanada, mais especificamente uma proporção de 83,27%, abastecendo por meio de água encanada apenas 16,73% das residencias. O último gráfico da Figura 3d apresenta valores bem menores para X_6 , mudando assim a escala do gráfico de 0,2 foi para 0,01, apresentando também um ponto discrepante referente ao município de Seridó(19), com valor 0,0259, isto é, 2,59% das residencias neste local possuem 8 pessoas. Logo, pode-se dizer que a variável X_6 possui os menores valores, enquanto que, a X_7 os maiores valores. O que pode ser confirmado através da Figura 5,

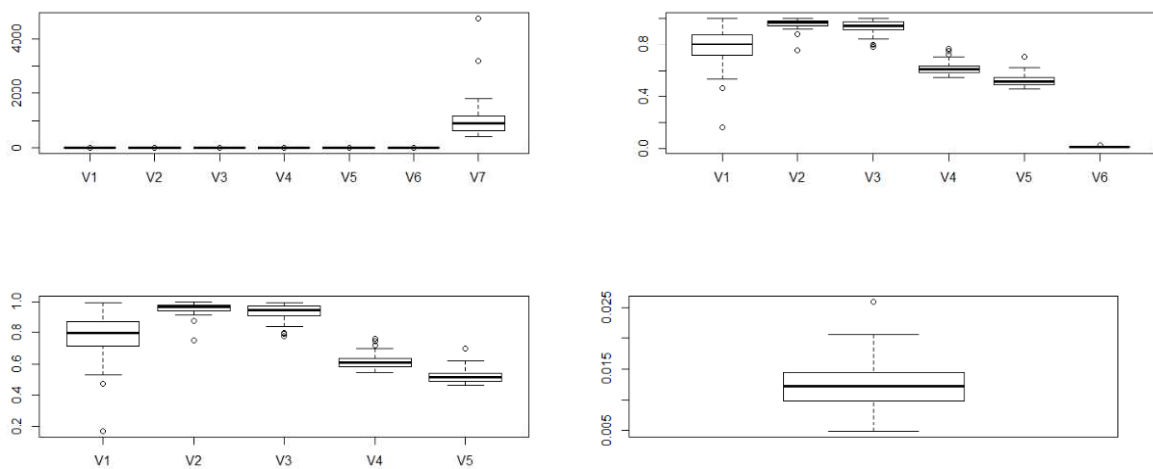


Figura 5 – Boxplot referentes as variáveis socioeconômicas.

O primeiro gráfico mostra o boxplot de todas as variáveis estudadas, entretanto, só pode ser visualizado o da variável X_7 (índice pluviométrico) devido aos seus altos valores, por este motivo o seu boxplot foi retirado no segundo, para facilitar a visualização dos outros, que por sua vez, foi dividido em dois gráficos, pois o X_6 possui valores muito pequenos. Quando compara-se a Figura 4 com a 5 nota-se uma semelhança na quantidade de pontos discrepantes mais latentes.

Tanto a mudança de escala presente na Figura 4, quanto a diferença visivelmente significativa da média e da variância, tornam previsível a padronização, que em geral, é considerada quando as variáveis possuem unidades diferentes, com o intuito de minimizar os efeitos produzidos pelas diferenças de escalas. Entretanto, como será utilizado a distância de Mahalanobis como meio para descobrir a proximidade dos elementos, não se vê a necessidade de padronizar as variáveis separadamente, já que esta distância possui a padronização em sua composição, padronizando os dados em sua própria execução .

Para verificar se algum dos valores discrepantes mais latentes disponíveis por meio da Figura 4 influenciam negativamente no ajuste, isto é, na qualidade do agrupamento, foi

utilizado da comparação de vários coeficientes de correlação cofenética de agrupamentos, através da retirada individual dos municípios representados por esses valores, podendo ser melhor visualizados através do Tabela 8.

Tabela 8 – Comparação dos coeficientes de correlação cofenética provenientes dos agrupamentos formados com retirada de municípios que representam os valores discrepantes mais latentes.

Metodos/ Municípios	Sem JP(31)	Sem BS(20)	Sem Se- ridó(19)	Sem CG(21)	Todos
Simple	0,7107	0,7248	0,7551	0,8432	0,7496
Completa	0,3461	0,6066	0,6454	0,6759	0,6022
Média	0,7254	0,7922	0,7977	0,8599	0,7904
Ward	0,4324	0,3575	0,3863	0,4119	0,4036

A partir da Tabela 8, que mostra a retirada dos municípios individualmente e consequentemente a sua influência na qualidade do ajuste por serem pontos discrepantes, foi observado que após a retirada do município de Campina Grande, houve um aumento significativo no coeficiente de correlação cofenética(CCC) referente a todos os métodos utilizados, mostrando uma forte influência negativa na qualidade do ajuste, podendo interferir na qualidade e/ou validade dos agrupamentos. Se este fosse um estudo apenas teórico sem intenção de aplicabilidade, seria decidido pela sua retirada, entretanto, como não é o caso, optou-se por sua permanência, devido a sua grande importância no estado. Por este motivo será utilizado todos os elementos, voltando-se para última coluna, que por sinal, mostra que o método que mais se ajusta aos dados, é o da média, pois este possui o maior valor de CCC, dando a entender, neste caso, que os grupos formados através deste método se aproximam mais da realidade que o dos outros métodos. Observe na Figura 6, o dendrograma dos dois métodos que obtiveram o melhor ajuste.

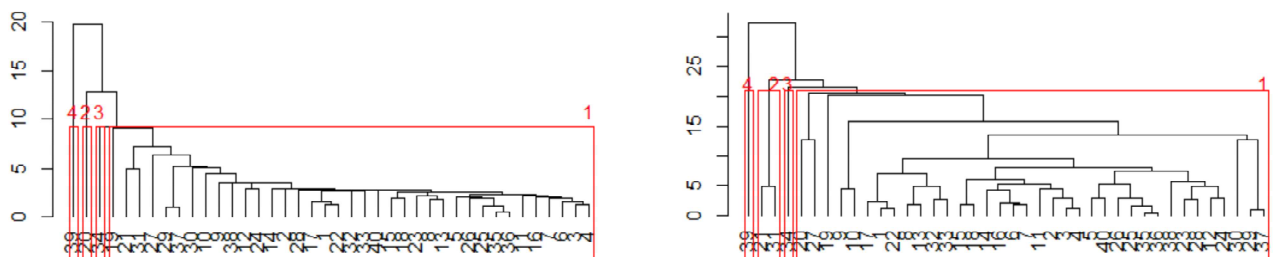


Figura 6 – Dendrograma formado através da técnica hierárquica aglomerativa utilizando a ligação simples e a média como forma de agrupamento, e a distância de Mahalanobis como medida de proximidade.

Ao comparar os dois dendrogramas presentes na Figura 6, observa-se que dois grupos são iguais, mais especificamente, o 3 e 4, que representam os municípios de Conde(39) e Cabedelo(34), que pode ser explicado devido a proximidade no valor do CCC, entretanto, como o método da média possui o maior valor deste coeficiente, então, a análise será restrita a ele.

Quanto mais perto de zero, mais homogêneos e próximos estarão os grupos. Sabendo deste fato, vê-se no segundo dendrograma apresentado na Figura 6, que Sapé(35) e Mamanguape(36), são os elementos mais próximos, e por isso, os primeiros a formarem grupo, seguidos por Areia(29) e Pedras de Fogo(37), em contrapartida, Conde(39) e Cabedelo(34) não formaram grupos até o corte, tornando-se os municípios mais diferentes e distantes, assim como o grupo que possui os elementos Campina Grande(21) e João Pessoa(31). Através da Tabela 9, pode-se observar melhor os quatro grupos formados, que são os apresentados dentro da borda vermelha do dendrograma.

Tabela 9 – Grupos formados por meio do Dendrograma gerado através da matriz de distâncias de Mahalanobis

Método/ Grupos	G1	G2	G3	G4
Ligação média	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,18,19,20,22,23,24,25,26,27,28, 29,30,32,33,35,36,37,38,40	21,31	34	39

Mostra-se que os grupos formados são diferentes das quatro mesorregiões, já que estes não são restritos pelos territórios, entretanto, quando fala-se de cidades metropolitanas, pode-se dizer que duas delas fizeram parte de um único grupo, Campina Grande(21) e João Pessoa(31), contudo, é interessante observar que as outras cidades metropolitanas não se expuseram, mas ficaram no maior grupo, dando a entender que socioeconomicamente são similares com os elementos deste.

Através dos elementos dos grupos, podemos supor alguns motivos que levariam a similaridade socioeconômica dos municípios, como por exemplo, para o grupo dois, poderia ser levado em consideração a sua grande proximidade, por serem os dois primeiros em quase tudo, isto é, além deles serem considerados os maiores centros econômicos da Paraíba e possuírem um aeroporto em sua redondeza, dispõem dos maiores números de habitantes do estado, e maiores valores do PIB, entre outras coisas; já o terceiro grupo, possui apenas um elemento, que pode ser diferente dos outros, devido a sua importação e exportação marítima, através do seu porto, que por sinal é considerado o mais oriental do país, sendo beneficiado pelo maior PIB per capita do estado.

5 Considerações Finais

Após a aplicação das técnicas de agrupamento em variáveis socioeconômicas de uma amostra de quarenta municípios do estado da Paraíba, foi comparada a perda de informação resultante da aplicação dos métodos de ligação simples, completa, média e ward, através do coeficiente de correlação cofenética, mostrando que os métodos simples e da média, obtiveram melhor resultado, propondo desta forma, que estes retratam melhor a realidade. Entretanto, os resultados sugerem que o método mais adequado é o da média, já que este, possui o maior valor, resultando na formação de grupos, em que, seus elementos(municípios) são similares socioeconomicamente, onde o primeiro possui a maioria dos municípios, o segundo é composto por João Pessoa e Campina Grande, o terceiro, Cabedelo e o quarto, Conde. Mostrando que o município Conde, é o mais distante, ou seja, o mais diferente.

Referências

- BAILEY, K. D. Cluster analysis. *Sociological Methodology*, v. 6, p. 59–128, 1975. Citado na página 10.
- BARROSO, L. P.; ARTES, R. Análise multivariada. *Seagro e Rbras*, 2003. Citado 2 vezes nas páginas 19 e 20.
- DONI, M. V. *Análise de Cluster: Métodos Hierárquicos e de Particionamento*. Dissertação (Mestrado) — Pontifícia Universidade Católica do rio de Janeiro, 2004. Citado 6 vezes nas páginas 12, 16, 17, 18, 20 e 21.
- FERREIRA, D. F. *Estatística Multivariada*. Lavras - Minas Gerais: Universidade Federal de Lavras, 2008. Citado 4 vezes nas páginas 10, 15, 16 e 21.
- FIEP. Perfil socioeconômico da Paraíba 2010. *FIEP*, 2010. Citado na página 9.
- FILHO, D. B. F.; JUNIOR, J. A. da S.; ROCHA, E. C. da. Classificando regimes políticos utilizando análise de conglomerados. *Opinião pública*, v. 18, p. 109–128, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-62762012000100006>. Citado na página 10.
- HAIR, J. F.; BLACK, W. C.; J. BABIN, B.; ANDERSON, R. E.; TATHAM, R. L. *Análise Multivariada de dados*. Belo Horizonte- Minas Gerais: Universidade de Minas Gerais, 2009. Citado 8 vezes nas páginas 8, 10, 15, 17, 18, 20, 21 e 22.
- HOMESBURG, C. H. Cluster analysis. *Belmont: Lifetime Learning*, 1984. Citado na página 21.
- JESUS, C. S. S. *Clustering aplicado à Bolsa de Valores de Lisboa*. Dissertação (Mestrado) — Isep, 2015. Citado 2 vezes nas páginas 12 e 23.
- KUBRUSLY, L. S. Um procedimento para calcular índices a partir de uma base de dados multivariada. *Pesquisa Operacional*, v. 21, p. 107–117, 2001. ISSN 1678-5142. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-74382001000100007>. Citado na página 11.
- MEYER, A. da S. *Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes*. Dissertação (Mestrado) — Escola Superior de Agricultura Luiz de Queiroz, São Paulo, 2002. Citado 2 vezes nas páginas 20 e 21.
- MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada uma abordagem aplicada*. Minas Gerais: UFMG, 2013. Citado 6 vezes nas páginas 11, 12, 13, 14, 22 e 26.
- NUNES, S. G. calves. *Contribuição da análise de clusters para a identificação de diferentes fenótipos na retinopatia diabética*. Dissertação (Mestrado) — Universidade de Coimbra, Coimbra, 2006. Disponível em: <https://www.uc.pt/en/fmuc/ibili/copy_of_Staff/SN>. Citado 2 vezes nas páginas 13 e 22.

- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <<https://www.R-project.org/>>. Citado na página 25.
- REIS, E. *Estatística Multivariada Aplicada*. Lisboa: Silabo, 2001. Citado 5 vezes nas páginas 8, 12, 13, 14 e 23.
- RODRIGUES, A.; COELHO, A. C.; PAULO, E.; ALMEIDA, F. C. de; BEZERRA, F. A.; CUNHA, J. V. A. da; ANTUNES, J.; FILHO, J. M. D.; DINIZ, J. A.; SANTOS, J. dos; CORRAR, L. J.; POHLMANN, M. C.; MARIO, P. do C.; HERDEIRO, R. F. C.; NAKAO, S. H. *Análise multivariada para os cursos de administração, Ciências Contábeis e Economia*. São Paulo: FINECAFI - Fundação Instituto de Pesquisas Contábeis, Atuariais e Financeiras, 2009. Citado 5 vezes nas páginas 8, 10, 11, 17 e 26.
- SANTOS, L. M. L. dos. *Socioeconomia*. Atlas, 2014. ISBN 9788522487868. Disponível em: <<https://www.amazon.com/Socioeconomia-Solidariedade-Economia-Organiza%C3%A7%C3%B5es-Portuguese/dp/8522487863?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=8522487863>>. Citado na página 9.
- SARTORIO, S. D. *Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R*. Dissertação (Mestrado) — Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2008. Citado 4 vezes nas páginas 8, 19, 20 e 23.
- VALE, M. N. do. *Agrupamento de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos*. Dissertação (Mestrado) — Pontifícia Universidade Católica, 2005. Citado 5 vezes nas páginas 11, 13, 17, 18 e 27.
- VALENTIN, J. L. Agrupamento e ordenação. *Oecologia brasilienses*, v. 2, p. 27–55, 1995. Disponível em: <<https://revistas.ufrj.br/index.php/oa/article/view/5553/>>. Citado na página 21.
- VICINI, L. *Análise Multivariada da Teoria à Prática*. Rio Grande do Sul: Universidade Federal de Santa Maria, 2005. Citado 7 vezes nas páginas 13, 18, 19, 20, 21, 23 e 27.