Universidade Estadual de Maringá
Laboratório de estatística

Análise de cluster hierárquico aplicada à RMM

Wesley Oliveira Furriel RA:61493

Prof. Dr. Josmar Mazucheli

Maringá
2018

1 Introdução

O presente trabalho tem como objetivo verificar às semelhanças ou diferenças entre os municípios da região metropolitana de Maringá, a partir das áreas de ponderação definidas pelo IBGE. Para tal, foi utilizado a técnica multivariava de análise de cluster hierárquico, uma vez que, não havia um número pré definido de grupos a serem considerados. As variáveis utilizadas na investigação representam algumas características socioeconômicas da população residente na região e foram baseadas em um estudo, intitulado Tipologia Intraurbana, realizado pelo IBGE em 2017.

2 Metodologia

O objetivo da análise de *cluster* é identificar padrões de agrupamentos em observações multivariadas, de forma a encontrar grupos nos quais as informações dentro de cada cluster são similares, porém os clusters são distintos entre si. Para verificar as similaridades ou dissimilaridade entre os elementos de um conjunto de dados são utilizadas medidas de distância entre as observações, normalmente padronizadas de um conjunto de dados(RENCHER, 2003; JOHNSON, 2002). Considerando a abordagem de Clusters sob a ótica de *Machine Learning* as técnicas de agrupamento são enquadradadas no seguimento de aprendizado não supervisionado (unsupervised machine learning), que utiliza técnicas de agrupamento como K-means, Métodos Hierárquicos com um enfoque computacional.

É possível dividir a análise de cluster em dois tipos de métodos: hierárquicos e não hierárquicos. No caso do primeiro, não sabemos de previamente quantos clusters serão formados. A definição do número de agrupamentos será realizada pela análise gráfica do dendograma ou por métodos como elbow e silhouette (KASSAMBARA, 2017). Já os agrupamentos não hierarquicos são caracterizados pela necessidade de definir um número inicial de agrupamentos.

Os algoritmos utilizados para clusters hierárquicos geralmente são divididos em dois seguimentos, os aglomerativos e os divisivos. No caso dos aglomerativos as observações iniciam separadas e são agrupados em etapas, nas quais clusters similares vão sendo agrupados até a criação de um único cluster. Quanto aos divisivos as observações iniciam em um cluster são separadas até que cada elemento seja seu próprio cluster.

No caso deste trabalho foram utilizados métodos de cluster hierárquico, uma vez que o interesse foi o de verificar a existência de regiões com características distintas na Região Metropolitana de Maringá e o quão elas se distanciam de outros grupos formados,sem um número prévio de agrupamentos definido.

Tendo em vista que o interesse é agrupar itens de acordo com sua similaridade e separalos, segundo sua dissimilaridade é necessário o emprego de uma medida, para captar tais a proximidade entre as observações. Desse modo, no presente estudo foi utilizada a distância euclidiana, tendo em vista a natureza continua dos dados. Para definir a distância entre dois objetos no espaço considera-se os vetores p-dimensional $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$ e $\boldsymbol{y} = (y_1, y_2, ..., y_p)'$ decorrentes da mensuração de p variáveis contínuas com n observações. A distância Euclidiana é dada por:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{p} (x_i - y_i)} \tag{1}$$

No caso de variáveis contínuas outras distâncias como Minkowski e Manhattan (RENCHER, 2003; JOHNSON, 2002).

3 Métodos de Aglomeração

A maior parte dos algoritmos de agrupamento hierárquico são variações de métodos conhecidos como *linkage*, responsáveis por agrupamentos de informações similares criando grandes clusters e mensurando a distância entre estes (KASSAMBARA, 2017). Desse modo, apresentamos de forma sucinta os métodos empregados neste trabalho. Para tal, detona-se D_{ij} como a distância entre dois *clusters* C_i e C_j.

3.0.1 Single Linkage

O método Single Linkage mensura a distancia D_{ij} entre os *clusters* C_i e C_j, sua expressão é dada por

$$D_{ij} = \min_{x \in C_i, y \in C_j} (x, y)\,. \tag{2}$$

Uma desvantagem deste método é que os *clusters* podem ser forçados se agruparem devido a elementos próximos uns dos outros, mesmo muitos dos elementos em cada *cluster* estão consideravelmente distantes. Consequentemente, esse método fornece *clusters* irregulares e muito alongados.

3.0.2 Complete Linkage

Já o método Complete Linkage verifica a distância D_{ij} entre C_i e C_j é calculada por

$$D_{ij} = \max_{x \in C_i, y \in C_j} (x, y)\,. \tag{3}$$

Conforme discute Hastie, Tibshirani e Friedman (2009) o complete linkage representa o extremo oposto do single linkage, isto é, dois grupos são considerados próximas se todas as observações na sua união estão relativamente similares. Nesse sentido, as observações atribuídas a um *cluster* podem estar muito mais próximas a observações de outro *cluster* do que observações de seu próprio *cluster*.

3.0.3 Average Linkage

Neste método, a distancia D_{ij} entre dois *clusters* é definida por

$$D_{ij} = \frac{1}{n_i\, n_j} \sum_{x \in C_i, y \in C_j} (x, y) \tag{4}$$

em que n_i e n_j são o número de elementos nos *clusters* C_i e C_j, respectivamente.

Observa-se por (4) que a distância entre dois *clusters* é dada pela distancia média entre os pares de observações um em cada *cluster*. Este método tem tendência em juntar *clusters* com pequena variância e então produzir *clusters* com a mesma variância.

3.0.4 Método de Ward

Neste caso à distância entre dois *clusters* é definida por

$$D_{ij} = \frac{\|\overline{\mathbf{x}}_i - \overline{\mathbf{y}}_j\|^2}{\frac{1}{n_i} + \frac{1}{n_j}} \tag{5}$$

em que $\overline{\mathbf{x}}_i$ e $\overline{\mathbf{y}}_j$ são os vetores de médias dos *clusters* C_i e C_j, respectivamente e $\|\mathbf{x}\|$ denota a distancia Euclidiana do vetor $\mathbf{x}$ definida em (**??**).

O método de Ward consiste em calcular a soma de quadrados entre os dois *clusters* definida em (5). A cada estágio, os dois *clusters* que apresentarem o menor aumento na soma global de quadrados dentro dos *clusters* são agrupados (**?**). Em geral, este método produz *clusters* com aproximadamente o mesmo número de observações. Segundo (**?**) ele também é sensível a presença de outliers.

4 Número ótimo de *clusters*

Definir o número ótimo de clusters em métodos de agrupamento hierárquico é uma tarefa bastante complicada e sem uma resposta definida pela literatura, tal questão depende da medida de similaridade e do método de agrupamento empregados. Uma solução simples e amplamente utilizada é a análise do dendograma, entretanto, tal abordagem muitas vezes se dá de modo subjetivo, já que o dendograma pode ser separado de distintas formas. Tibshirani, Walther e Hastie(2001) introduzem a estatística de GAP, utilizada para estimar um número ótimo de clusters. Desse modo, considere um conjunto de observações $x_{ij}, i = 1, 2, ..., n \; e \; 1, 2, ..., p$, em que p representa o número de variáveis e n o número de observações. Considerando a distância Euclidiana quadrada, na qual $d_{ii'}$ denota a distância entre as observações i e i'. Suponha que k clusters foram formados $C_1, C_2, ..., C_k$, com C_r denotando o índice de observações no cluster r, e $n_r = |C_r|$, temos

$$D_r = \sum_{i,i' \in C_r} \tag{6}$$

sendo a soma dos pares das distâncias para todos os pontos no cluster r definida por

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{7}$$

O procedimento para obtenção da estatística de CAP segue os seguintes passos

- Agrupar os dados observados, variando o número total de $k = 1, 2, ..., K$ e para cada k obter a medida exposta em (7);
- Em seguida, é necessário gerar B conjuntos de dados como referência, a partir de uma distribuição uniforme sobre o intervalo dos valores observados das variáveis. Para cada B conjuntos de dados são criados clusters e obtida a medida (7), assim $W*_{kb}$ em que $b = 1, 2, ..., B$. Dessa forma, à estatística de GAP estimada é obtida por

$$Gap(k) = (1/B) \sum_{b} log(W^*_{kb}) - log(W_k) \tag{8}$$

- Por fim, considerando $\bar{I} = (1/B)\sum_b log(W^*_{kb})$, o desvio padrão é determinado por

$$sd_k = \left[(1/B)\sum_b \left\{log(W^*_{kb}) - \bar{I}\right\}^2\right]^{1/2} \tag{9}$$

e defina $s_k = sd_k\sqrt{(1+1/B)}$. Partindo disso, o número de clusters é dado via

$$\hat{k} = min(k) \text{ de forma que } Gap(k) \geq Gap(k+1) - s_{k+1} \tag{10}$$

Além disso, vários métodos de seleção para um número ótimo de clusters foram propostos por Milligan e Cooper (1985) e implementados por Charrad et al. (2014) em uma biblioteca do*software R (R Core Team,117 2016).*

5 Comparação dos dendogramas

Para comparação de dendogramas é possível utilizar técnicas gráficas como expõe Kassambara (2017), para tal, dois dendogramas são colocados lado a lado, com os *labels* iguais conectados por linhas. A qualidade do alinhamento destes dendogramas pode ser medida usando a medida de *Entanglement*, em 1ue valores próximos a 1 indicam confusão total e 0 coesão total. Assim, um coeficiente baixo indica bom alinhamento. No que tange as medidas, pode-se destacar a correlação entre as distâncias cofenéticas e a medida de Barker. FOWLKES e MALLOW (1983) sugerem o índice B_k para verificar a similaridade entre dois *clusters* hierárquicos, com o mesmo número de variáveis. Considere que os dendrogramas A_1 e A_2 podem ser fracionados para produzirem $k = 2, \ldots, n-1$ *clusters*. De forma que, para cada valor de k a seguinte estrutura pode ser criada

$$M = [m_{ij}], \qquad i, j = 1, \ldots, k$$

em que a quantidade m_{ij} representa o número de observações comuns entre o i-ésimo cluster de A_1 e j-ésimo cluster de A_2. A medida de associação é definida então por

$$B_k = \frac{T_k}{\sqrt{P_k\, Q_k}} \tag{11}$$

de modo que

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n, \tag{12}$$

$$P_k = \sum_{i=1}^{k}\left(\sum_{j=1}^{k} m_{ij}\right)^2 - n, \tag{13}$$

$$Q_k = \sum_{j=1}^{k}\left(\sum_{i=1}^{k} m_{ij}\right)^2 - n. \tag{14}$$

Como B_k é calculado para vários valores de k e portanto, a similaridade entre os métodos pode ser exposta em um gráfico de B_k segundo k.

6 Número de variáveis e tamanho da amostra

Como expõe Dolnicar(2002) não há uma regra geral para definir o número máximo de variáveis passíveis análise, segundo determinado tamanho de amostra. Entretanto, é preciso considerar que análises com alto número de variáveis e poucas observações, podem gerar resultados pouco confiáveis. Segundo a autora, à maior parte dos trabalhos acadêmicos que utilizam técnicas de clusters, trabalham com amostras menores que 300, com tamanhos de amostras que variam entre 10 e 20.000, enquanto que o número de variáveis empregadas, oscilam entre 10 e 66. Desse modo, uma medida sugerida por Formann (1984) pode servir para avaliar o número mínimo de observações necessárias, para determinado número de variáveis, tal dimensionamento é obtido fazendo $n = 2^p$, em que p representa o número de variáveis estudadas.

7 Padronização dos dados

Ao analisar os dados deve-se observar se as variáveis foram medidas em unidades muito distintas entre si, pois grandes distinções entres as grandezas podem influenciar os resultados da análise. Desse modo, foram utilizados métodos de padronização dos dados. O primeiro deles, foi à padronização pela Z

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j} \tag{15}$$

Outro método de padronização é dado pela fórmula de mínimos e máximos, para tal é preciso ter mente se o valor de interesse é o mínimo ou máximo da variável e questão, em casos nos quais o máximo é o interesse temos

$$maxmin_{ij} = \frac{X_{ij} - max(X_j)}{min(X_j) - max(X_j)} \tag{16}$$

já quando o mínimo é o interesse

$$minmax_{ij} = \frac{X_{ij} - min(X_j)}{max(X_j) - min(X_j)} \tag{17}$$

para ambos os casos as variáveis serão padronizadas no intervalo entre 0 e 1.

8 Resultados e Discussões

A aplicação deste trabalho foi baseada na pesquisa Tipologia Intraurbana, realizada pelo IBGE em 2017. Dessa forma, às variáveis foram selecionado com base no estudo realizado, sendo oriundas do Censo Demográfico de 2010. O espaço geográfico considerado para a investigação foi à Região Metropolitana de Maringá, formada por 26 municípios, sendo a unidade amostral as 52 áreas de ponderação que compõe a região. As variáveis empregadas podem ser verificadas na Tabela 8

Código	**Descrição**
APOND	Área de ponderação
ESGOTO	% de pessoas cujo domicílio possui rede geral de esgoto ou pluvial, ou fossa séptica
AGUA	% de pessoas cujo domicílio possui água distribuída por rede geral de abastecimento
LIXO	% de pessoas cujo domicílio possui coleta de lixo diretamente por serviço de limpeza
DENSIDORM	% de pessoas em domicílios com densidade de até dois moradores por dormitório
RDPC	Mediana do rendimento domiciliar per capita
RDEP	Razão de dependência de menores de 15 anos (Pessoas de 0 a 14 anos / Pessoas de 15 a 64 anos)
ESCOLARIDADE1	% de pessoas sem instrução ou com fundamental incompleto e 18 anos ou mais de idade
ESCOLARIDADE4	% de pessoas com superior completo e 18 anos ou mais de idade
MAQUINA	% de pessoas em domicílios com existência de máquina de lavar
COMPUTADOR	% de pessoas em domicílios com existência de computador com acesso à Internet
ALVENARIA	% de pessoas em domicílios com alvenaria predominante nas paredes externas

Como é possível observar foram consideradas 11 variáveis para o estudo, todas agregadas por área de ponderação, isto é, os valores aqui considerados referem-se a proporção, mediana e razão das pessoas e domicílios segundo suas respectivas áreas de ponderação. Como se faz necessário considerar à discussão realizada por Dolnicar(2002) sobre o dimensionamento do conjunto de dados, duas análises foram realizadas, uma considerando todas as variáveis expostas e a segunda apenas com as variáveis que aparecem em negrito. Tendo em vista, à discussão acerca do tamanho mínimo de amostra segundo o número de variáveis para o emprego da análise de clusters hierárquicos. Foi realizado o cálculo sugerido por Formann(1985) de forma que $2^p \leq 52$, assim, foram selecionadas 5 variáveis das variáveis expostas em 8. Para a seleção, foram consideradas variáveis que expressem características individuais da população da apond, isto é, que não representem reflexos das políticas públicas e habitacionais da região, mas apenas fatores socioeconômicos de nível individual. Desse modo, $n = 2^p = 32$, como a amostra é de 52 aponds, as análises podem ser realizadas com folga.

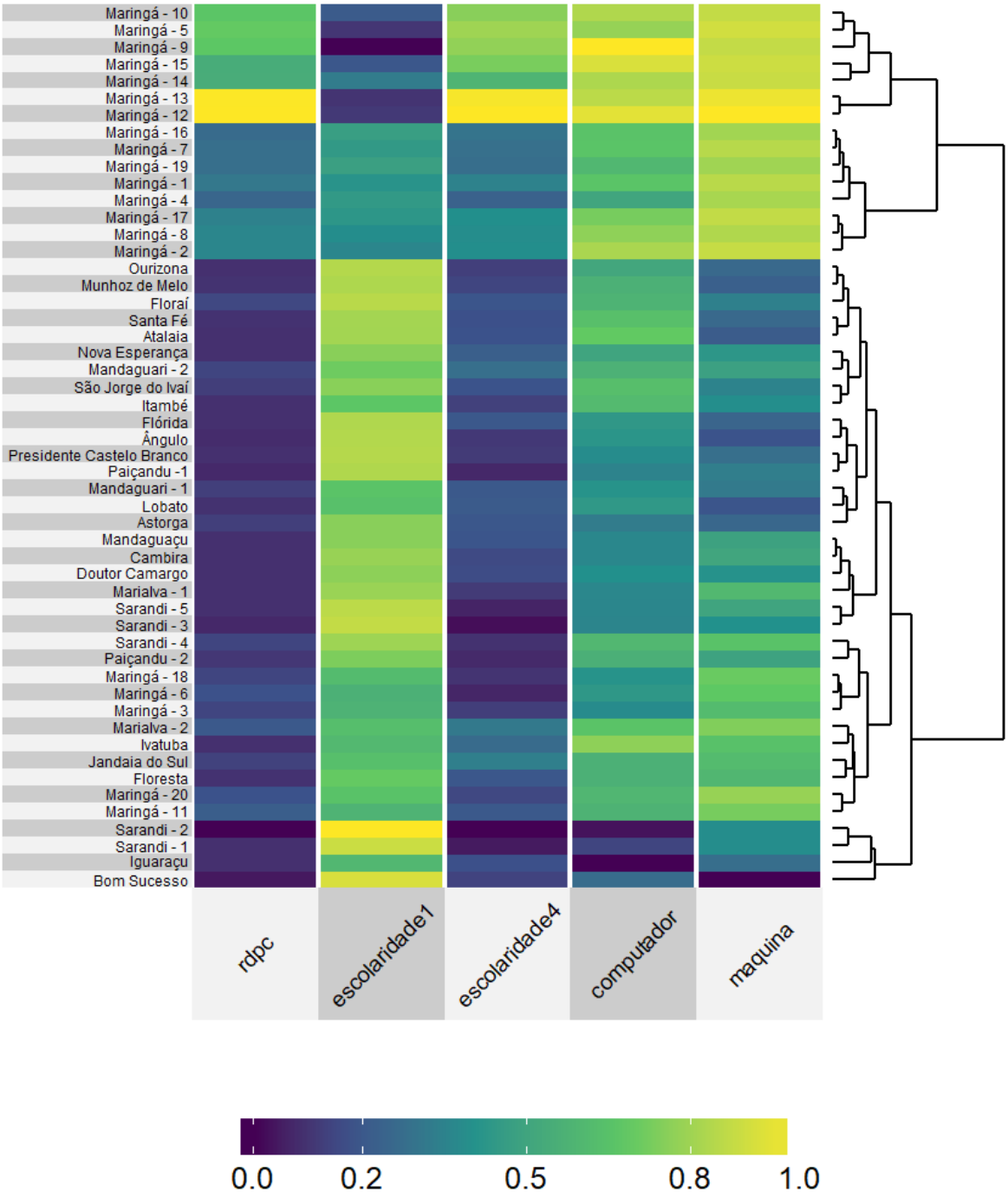


Figura 1: Valores padronizados das variáveis

Na figura 1 temos o heatmap das variáveis padronizadas pelo método de máximos e mínimos, segundo as aponds, de forma que valores próximos a 1 indicam boas condições socioeconômicas e próximos a 0, más.

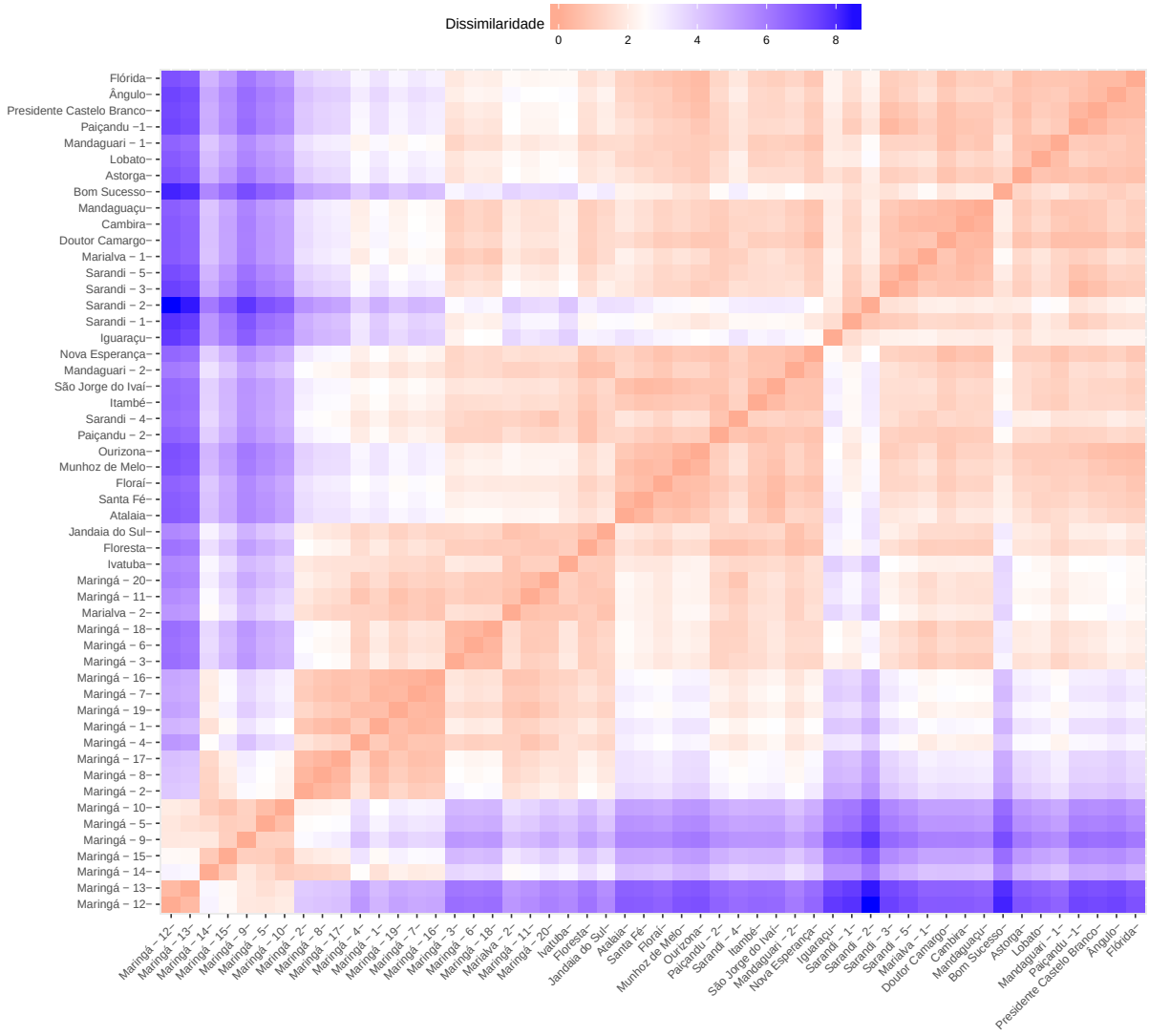


Figura 2: Matriz de distâncias

Pela Figura 2 é possível verificar à distância entre as áreas de ponderação da RMM, de modo que, valores vermelhos representam alta similaridade ($d(x_i, x_j) = 0$), enquanto que valores próximos ao azul alta dissimilaridade. Partindo disso, nota-se que às áreas de ponderação pertencentes a Maringá, se mostram bastante distintas das áreas dos demais municípios.

Tabela 1: Correlação entre as distâncias cofenéticas e euclidiana

	5 variáveis		11 variáveis	
Linkages	$\hat{\rho}$	valor-p	$\hat{\rho}$	valor-p
Ward	0.8148	<0.001	0.6079	<0.001
Complete	0.8276	<0.001	0.7623	<0.001
Average	0.8490	<0.001	0.8247	<0.001
Single	0.7616	<0.001	0.5233	<0.001

Uma forma de verificar quão bem foi criando pode ser visto na Tabela 1, na qual verificamos à correlação de *pearson* da distância euclidiana e cofenética. Segundo, se o likage utilizado for válido a correlação entre estas duas distâncias deve ser forte, isto é,

acima de 0.75. É preciso ressaltar que o método Average, leva vantagem com este tipo de medida, dessa forma, as conclusões acerca de seu resultado devem ser interpretadas com parcimônia. Com isso, verificamos que quando utiliza-se 5 variáveis na análise todas as ligações mostram adequadas, enquanto que em uma análise com 11 variáveis apenas, Complete e Average apresentam altas correlações.

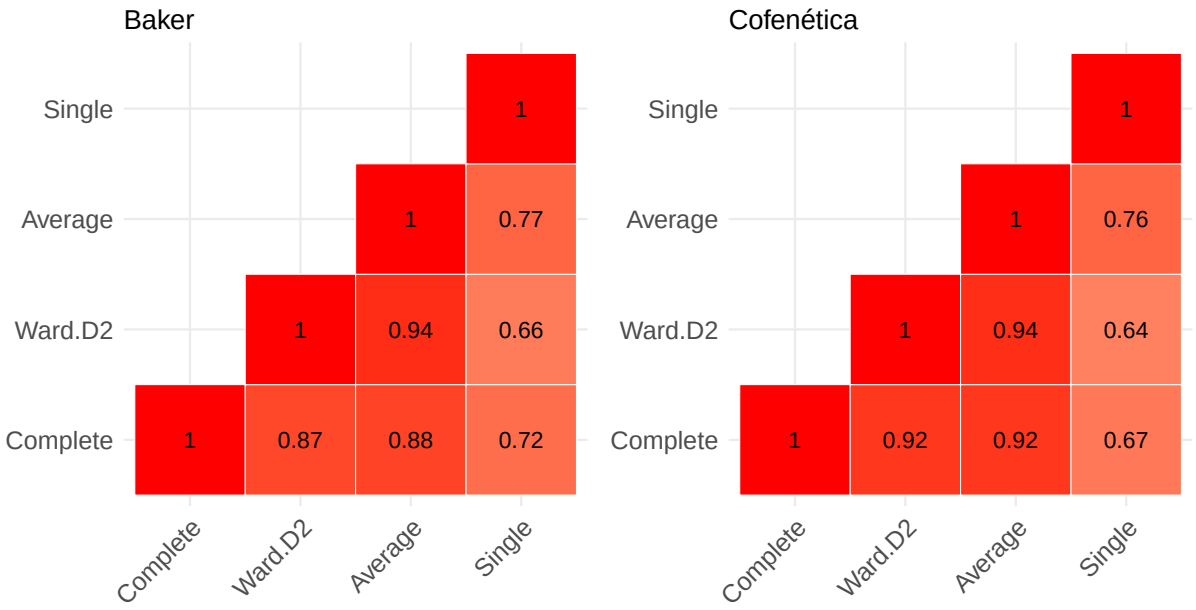


Figura 3: Medidas de Baker e Cofenética

Tendo em vista que todos os linkagens mostram-se adequados, realizou-se a correlação entre as medidas de Baker e as cofenéticas, segundo os *linkages* selecionados. Nota-se que, por ambos os testes, as medidas de Ward e Average, e também, Ward e Complete, apresentam uma estrutura bastante similar. Enquanto que Ward e Complete, quando comparadas a Single apresentam pouca similaridade.

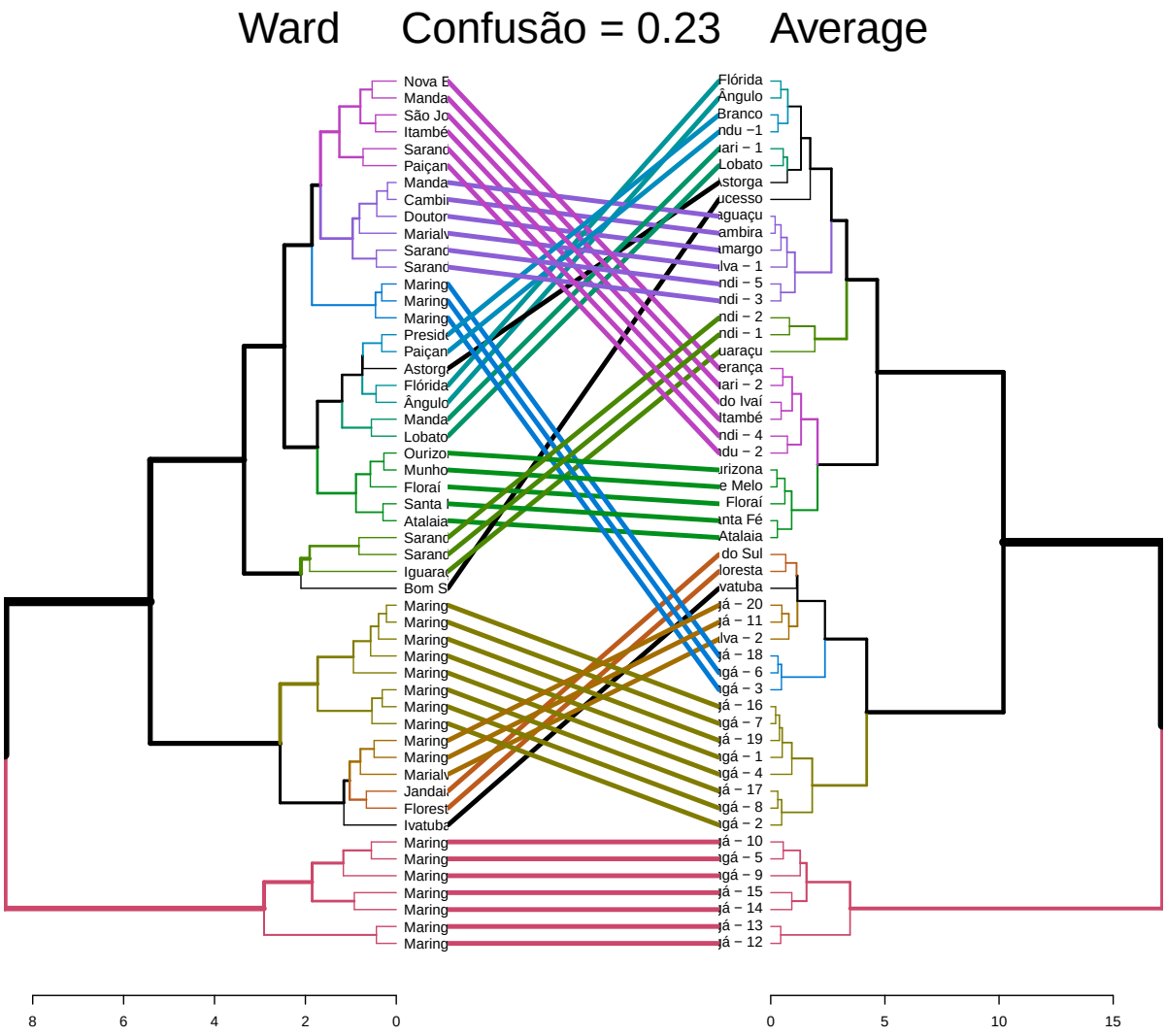


Figura 4: Comparação entre Ward e Average

Considerando as medidas apresentadas em 3 considerou-se os linkages que apresentaram maior e menor proximidade, para realizar uma investigação visual e calcular a medida de confusão(entanglement). esta medida varia no intervalo de 0 a 1, sendo qe valores próximos de 0 expressam pouca confusão, ou seja, similaridade entre os dendogramas e valores próximos de 1, indicam dendogramas distintos. Neste caso, verificamos que a medida de confusão retorna um valor bastante baixo 0.23, o que reforça a proximidade entre os linkages de Ward e Average.

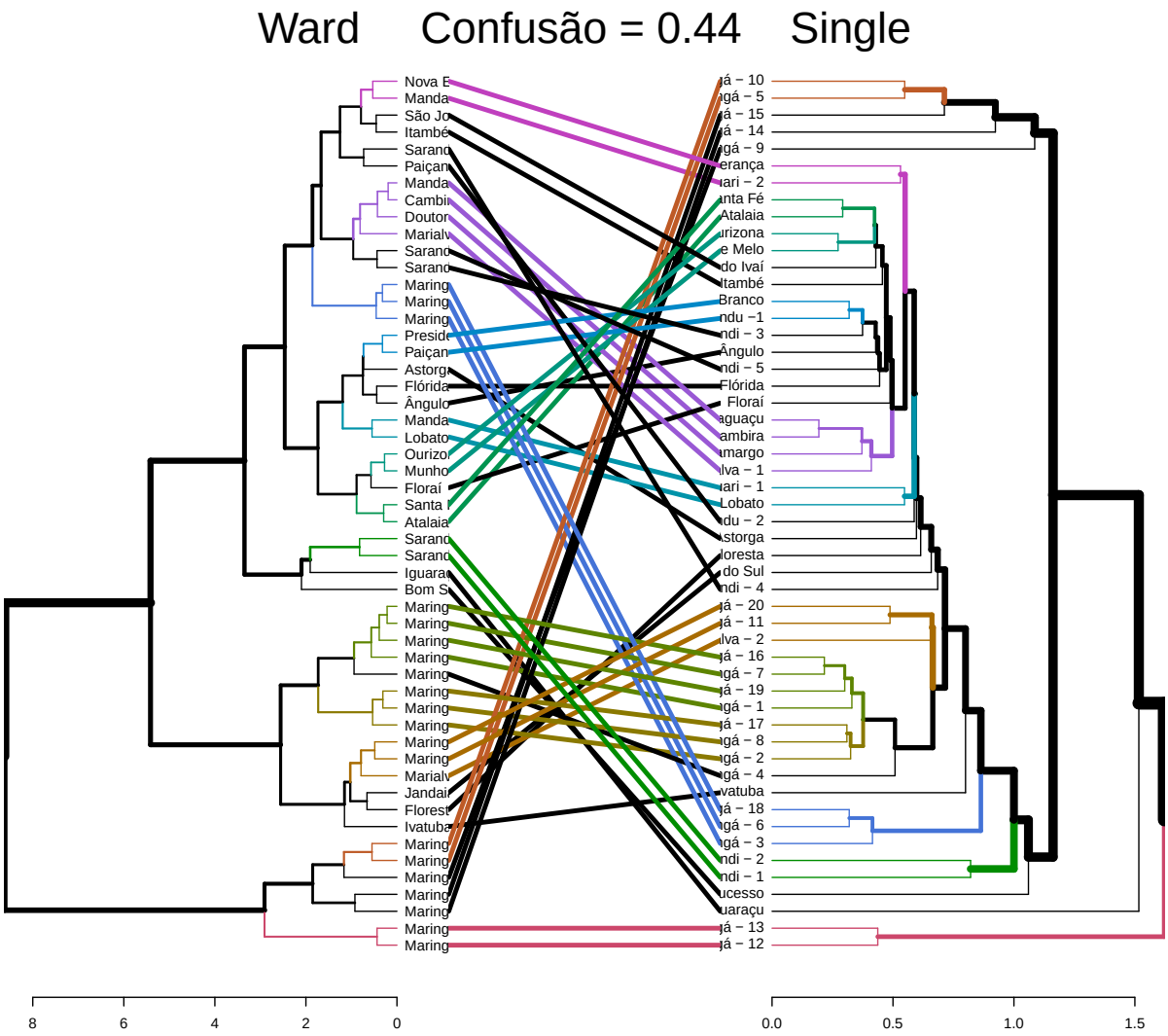


Figura 5: Comparação entre Ward e Single

Neste caso observamos a comparação gráficos dos dendogramas formados pelos linkage de Ward e Single. Nota-se que há uma mudança considerável entre os dois, além disso, a medida de confusão retornou valor de 0.44, fato que indica existência de consideráveis distinções entre ambos.

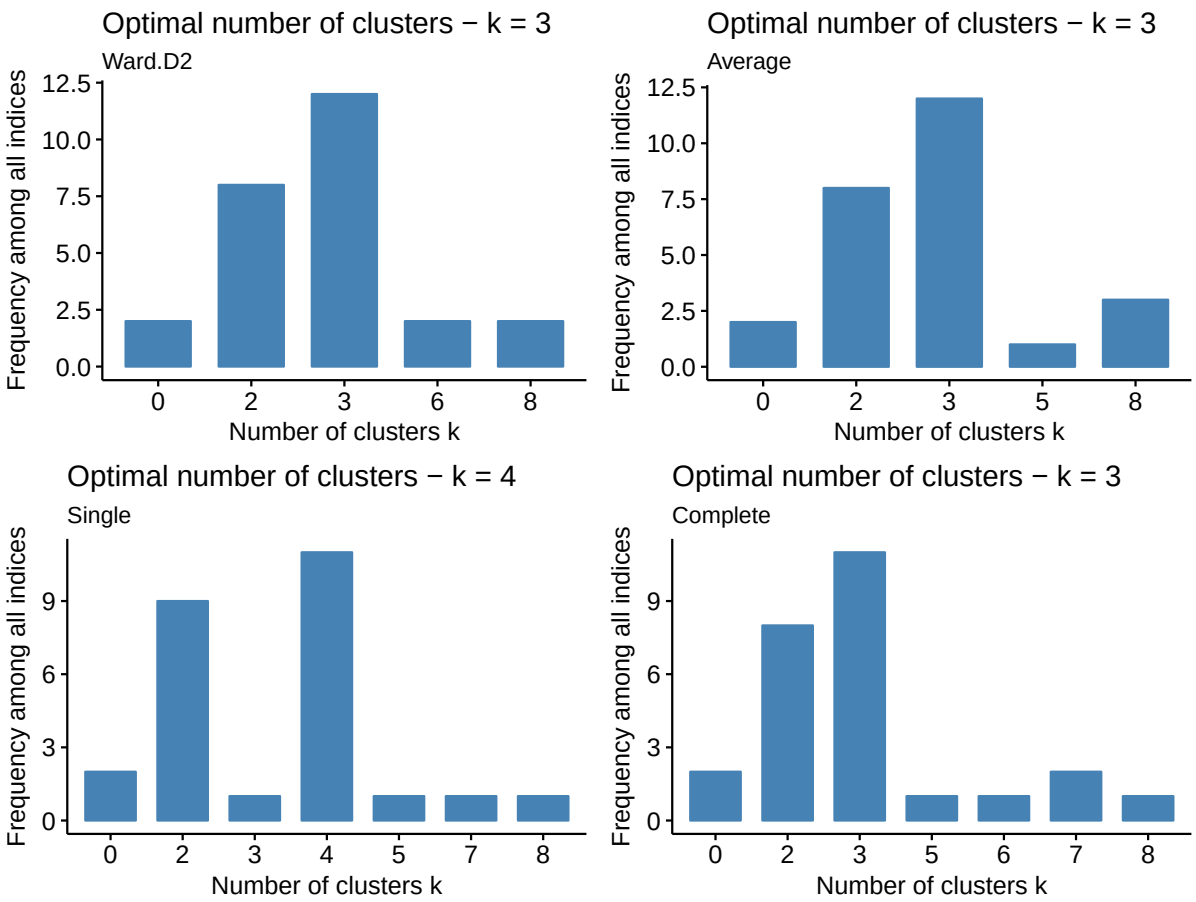


Figura 6: Número de clusters

Uma das fases mais importantes dos clusters hierárquicos se refere a definir o número de cluster, isto é, o ponto de corte do dendograma. Desse modo, foram utilizadas as trinta medidas implementadas no pacote NbClust e verificada à frequência do número de cluster sugeridas por elas, para cada linkage. Assim, com exceção do Single linkage, a maior frequência reside em 3 clusters.

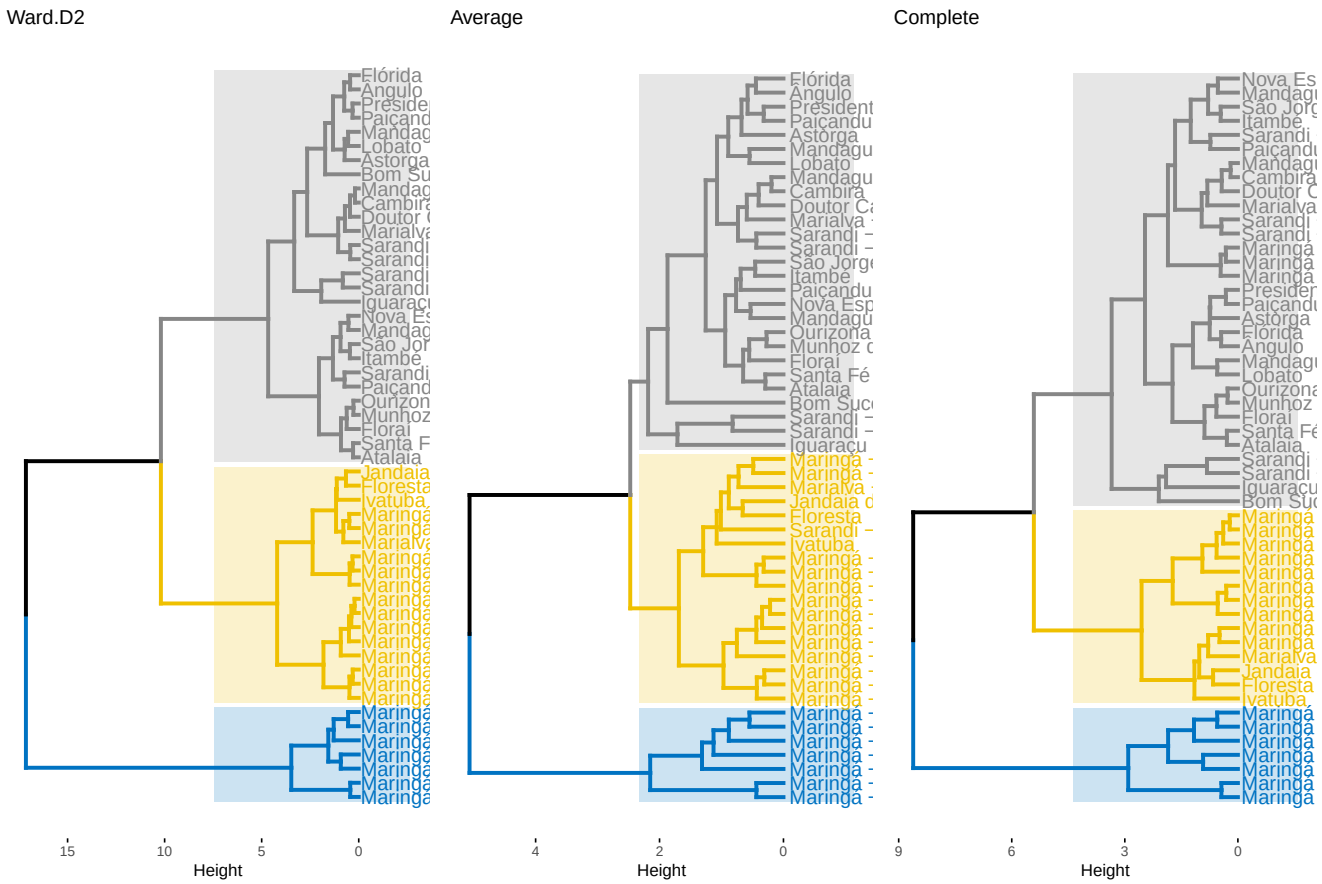


Figura 7: Dendogramas de com linkages Ward, Average e Complete

Quanto aos dendogramas, observa-se que considerado a criação de 3 clusters, à estrutura de organização das área de ponderação fica bastante próxima, entre eles. Principalmente considerando o cluster azul, nos quais as reigões de Maringá agrupadas foram sempre as mesmas.

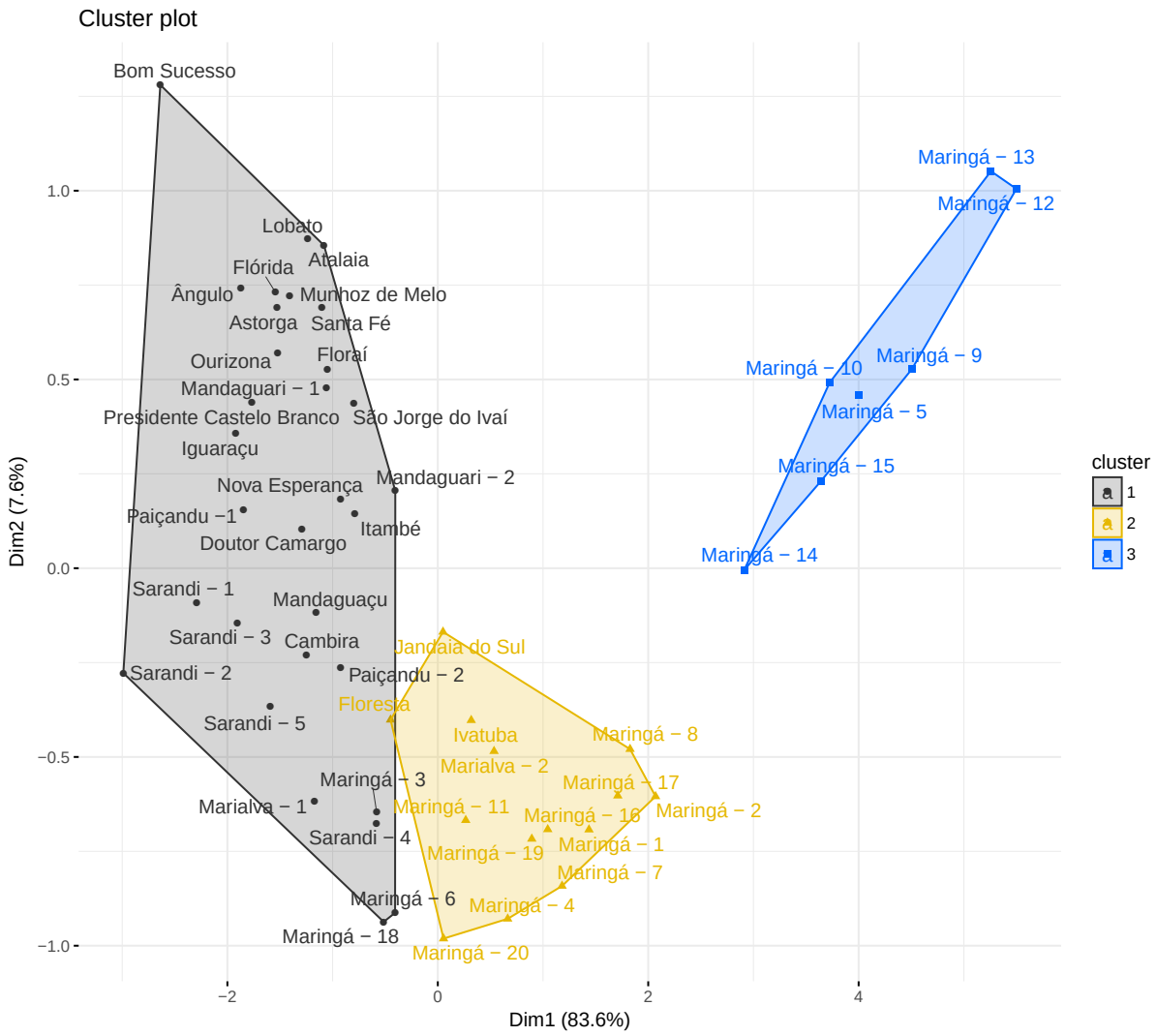
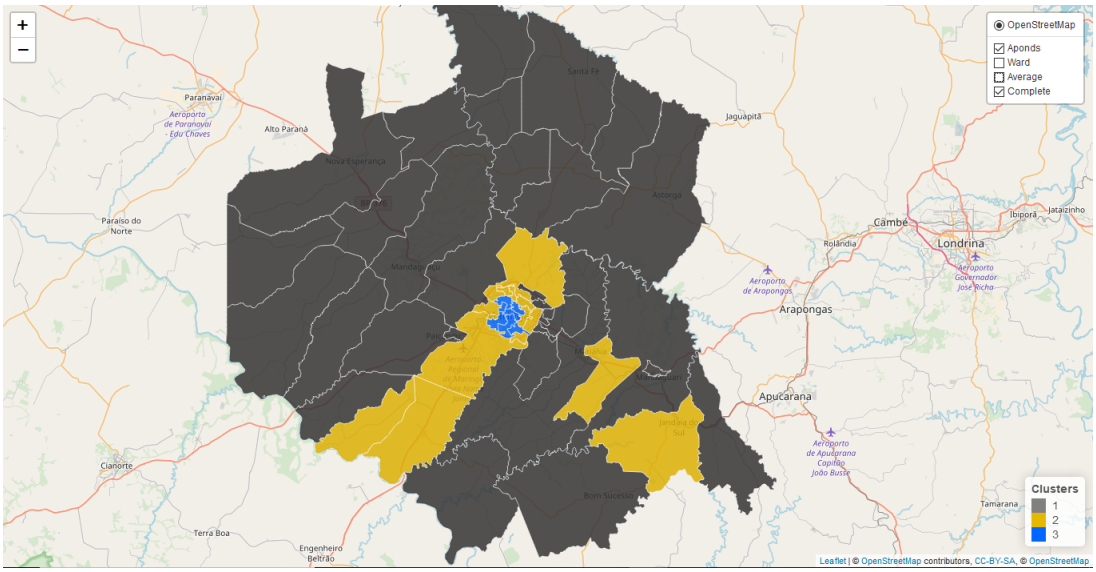
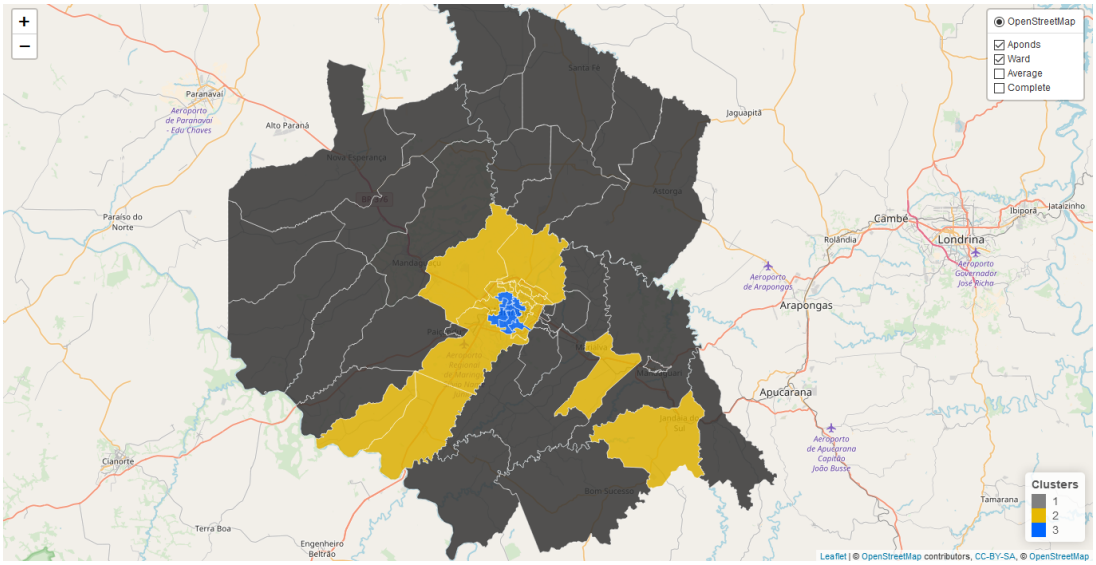
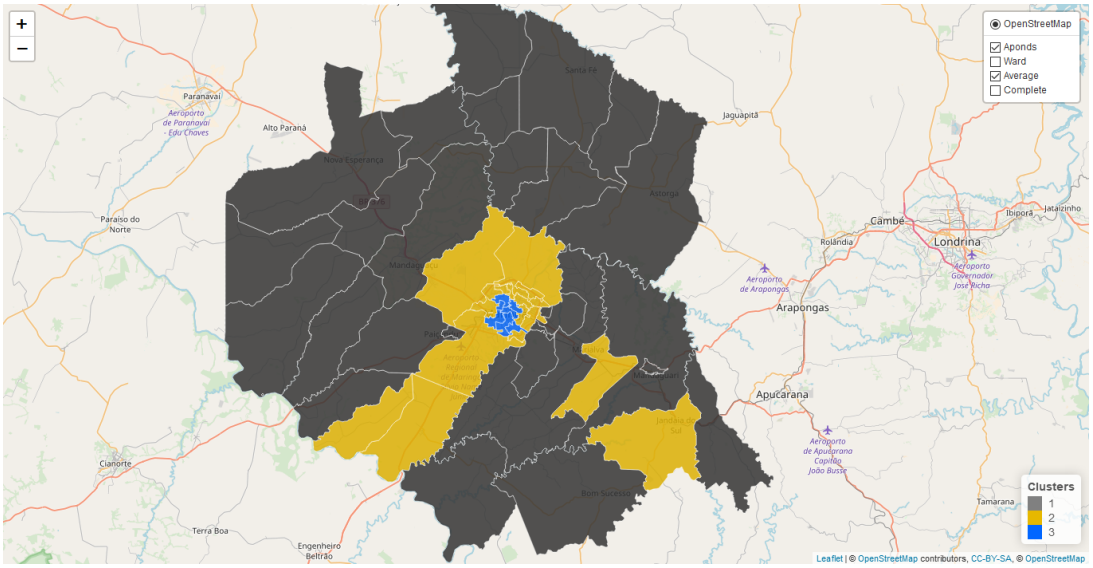


Figura 8:

Uma forma alternativa de visualização dos clusters, quando temos mais de duas variáveis é por meio das dimensões dos componentes principais, técnica utilizada para reduzir à dimensionalidade dos dados. Primeiramente observa-se que os componentes 1 e 2 explicam 91.2% da variabilidade dos dados.


+
−
● OpenStreetMap
☑ Aponds
☐ Ward
☐ Average
☑ Complete
Paranavaí
Aeroporto de Paranavaí - Edu Chaves
Alto Paraná
Nova Esperança
Santa Fé
Jaguapitã
Astorga
Paraíso do Norte
Mandaguaçu
Cambé
Ibiporã
Jataizinho
Rolândia
Londrina
Aeroporto Governador José Richa
Aeroporto de Arapongas
Arapongas
Apucarana
Aeroporto de Apucarana Capitão João Busse
Cianorte
Bom Sucesso
Tamarana
Terra Boa
Engenheiro Beltrão
Clusters
1
2
3
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap


+
−
● OpenStreetMap
☑ Aponds
☑ Ward
☐ Average
☐ Complete
Paranavaí
Aeroporto de Paranavaí - Edu Chaves
Alto Paraná
Nova Esperança
Santa Fé
Jaguapitã
Astorga
Paraíso do Norte
Mandaguaçu
Cambé
Ibiporã
Jataizinho
Rolândia
Londrina
Aeroporto Governador José Richa
Aeroporto de Arapongas
Arapongas
Apucarana
Aeroporto de Apucarana Capitão João Busse
Cianorte
Bom Sucesso
Tamarana
Terra Boa
Engenheiro Beltrão
Clusters
1
2
3
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap


+
−
● OpenStreetMap
☑ Aponds
☐ Ward
☑ Average
☐ Complete
Paranavaí
Aeroporto de Paranavaí - Edu Chaves
Alto Paraná
Nova Esperança
Santa Fé
Jaguapitã
Astorga
Paraíso do Norte
Mandaguaçu
Cambé
Ibiporã
Jataizinho
Rolândia
Londrina
Aeroporto Governador José Richa
Aeroporto de Arapongas
Arapongas
Apucarana
Aeroporto de Apucarana Capitão João Busse
Cianorte
Bom Sucesso
Tamarana
Terra Boa
Engenheiro Beltrão
Clusters
1
2
3
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap

Por fim, é possível verificar a formação dos clusters no território da RMM, nota-se que a região central de Maringá forma o mesmo cluster para todos os likages utilizados. Além disso, os métodos de Average e Ward, retornam formações idênticas, enquanto que o Complete linkage desagrega três aponds de Maringá do cluster 2 para o 1.

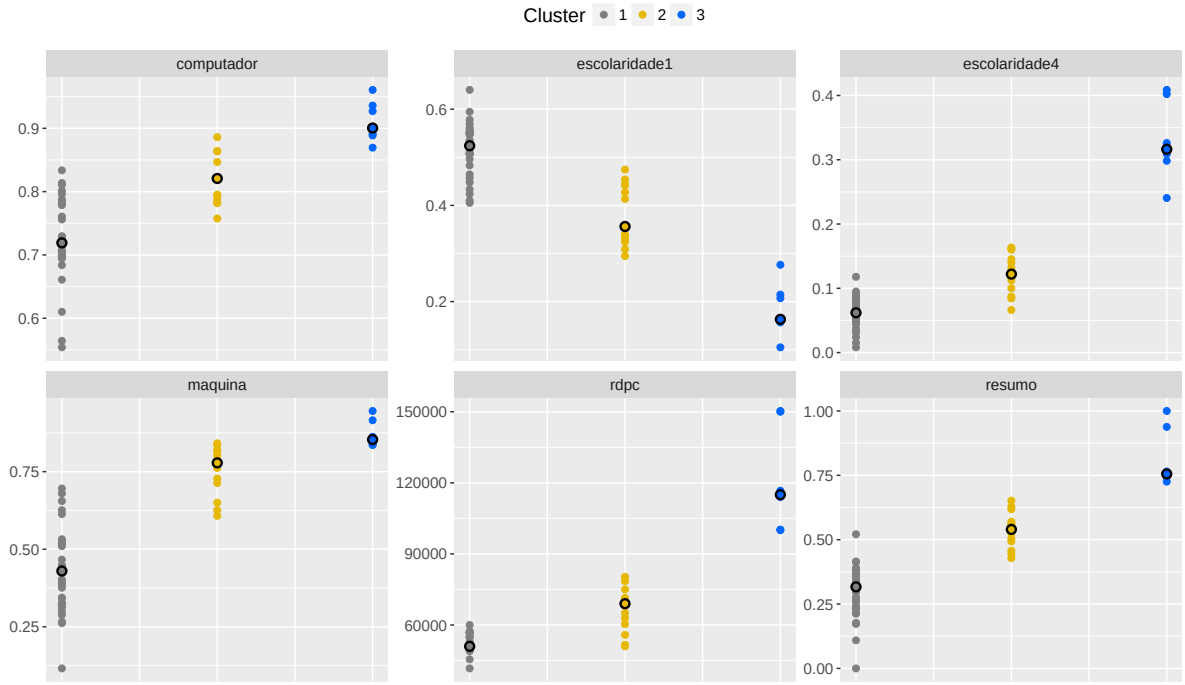


Figura 9:

Por fim verificamos as medidas descritivas das variáveis utilizadas. Nota-se que o cluster 3 apresenta as maiores médias, enquanto que o cluster 1 as menores. Além disso, pela medida resumo criada a partir da soma dos valores padronizados das variáveis, verificamos que no geral a região com as melhores condições socioeconômicas é a central de Maringá, como esperado. Para atingir os objetivos desejados foram utilizados os softwares Rstudio 1.1.447, com os pacotes *NbClust, factoextra, dendextend, clValid, ggplot2.*

Referências

[1] Rencher, Alvin C. Methods of multivariate analysis. Vol. 492. John Wiley & Sons, 2003..

[2] J ohnson, Richard A., and Dean Wichern. Multivariate analysis. John Wiley & Sons, Ltd, 2002.

[3] Kassambara, Alboukadel. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. Vol. 1. STHDA, 2017.

[4] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. The elements of statistical learning. Springer, New York, NY, 2009. 485-585.

[5] MILLIGAN, Glenn W.; COOPER, Martha C. An examination of procedures for determining the number of clusters in a data set. Psychometrika, v. 50, n. 2, p. 159-179, 1985.

[6] CHARRAD, Malika et al. Package NbClust. Journal of Statistical Software, v. 61, p. 1-36, 2014.

[7] Dolnicar, S. (2002). A review of unquestioned standards in using cluster analysis for data-driven market segmentation.