Cluster hierárquico aplicado à RMM

Wesley Furriel

Universidade Estadual de Maringá, Departamento de Estatística, PR, Brasil

11 de Julho de 2018

Abordagem multivariada

- Uma área de estatística que analisa um conjunto de variáveis de forma simultânea;
- *Educação:* As escolas variam sua performance, e ao buscar razões para determinada variação pode ser útil agrupar as escolas de modo que seja possível entender o que as escolas tem em comum.
- *Violência:* Permite agrupar diferentes municípios conforme variáveis relacionadas a criminalidade(assalto, assassinato, estupro, entre outros).

Análise de Agrupamento (*Cluster*)

- O objetivo da análise de agrupamento é identificar padrões ou grupos similares em um conjunto de dados;
- A análise de agrupamentos estabelece medidas de similaridade para variáveis, a fim de apresentar os resultados de forma condensada facilitando o entendimento do fenômeno estudado;
- Nesse sentido, a análise de *cluster* é amplamente utilizada por possibilitar a interpretação individual de cada grupo e a relação que este grupo possui com os demais.

Cluster hierárquico

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

Objetivos

- Considerando a pesquisa “Tipologia intraurbana” realizada pelo IBGE em 2010 para a cidade de Rio de Janeiro.
- Os objetivos desta análise são:
  - Identificar os grupos de áreas de ponderações da região metropolitana de Maringá que são mais similares de acordo com variáveis que refletem a qualidade de vida da população;
  - Caracterizar a diversidade socioespacial intra RMM.

Pacotes utilizados e descrição do banco

- **Rstudio 1.1.447**
  - *dplyr, cluster, factoextra, ggcorrplot, ggplot, dendextend2, NbClust*.
- **Bancos de dados**
  - IBGE, Censo Demográfico
    - ⋆ Ano: 2010
    - ⋆ Número de observações: 52
    - ⋆ Número de variáveis: 11
    - ⋆ Informações agregadas por área de ponderação.
    - ⋆ Fonte: https://www.ibge.gov.br/

Padronização dos dados

A padronização pela Z

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j} \tag{1}$$

Outro método de padronização é dado pela fórmula de mínimos e máximos, em casos nos quais o máximo é o interesse temos

$$maxmin_{ij} = \frac{X_{ij} - max(X_j)}{min(X_j) - max(X_j)} \tag{2}$$

já quando o mínimo é o interesse

$$minmax_{ij} = \frac{X_{ij} - min(X_j)}{max(X_j) - min(X_j)} \tag{3}$$

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

Distância Euclidiana

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:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{p} (x_i - y_i)} \tag{4}$$

Métodos de Aglomeração(*Linkages*)

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.

Single Linkage

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

$$D_{ij} = \min_{x \in C_i, y \in C_j} (x, y) \,. \tag{5}$$

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.

Complete Linkage

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

$$D_{ij} = \max_{x \in C_i, y \in C_j} (x, y) \,. \tag{6}$$

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*.

Average Linkage

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

$$D_{ij} = \frac{1}{n_i\, n_j} \sum_{x \in C_i, y \in C_j} (x, y) \tag{7}$$

em que n_i e n_j são o número de elementos nos *clusters* C_i e C_j, respectivamente.
Observa-se por (7) 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.

Ward

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

$$D_{ij} = \frac{\overline{\mathbf{x}}_i - \overline{\mathbf{y}}_j{}^2}{\frac{1}{n_i} + \frac{1}{n_j}} \tag{8}$$

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 $\mathbf{x}$.

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 empregado. **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. Desse modo, alguns índices podem ser utilizados para obtenção do número ótimo de clusters

id	Índice	SAS	clValid	NbClust
1	CH (Calinski and Harabasz 1974)	x		x
2	CCC (Sarle 1983)	x		x
3	Pseudot2 (Duda and Hart 1973)	x		x
4	KL (Krzanowski and Lai 1988)			x
5	Gamma (Baker and Hubert 1975)			x
6	Gap (Tibshirani et al. 2001)			x
7	Silhouette (Rousseeuw 1987)		x	x
8	Hartigan (Hartigan 1975)			x
9	Cindex (Hubert and Levin 1976)			x
10	DB (Davies and Bouldin 1979)			x
11	Ratkowsky (Ratkowsky and Lance 1978)			x
12	Scott (Scott and Symons 1971)			x
13	Marriot (Marriot 1971)			x
14	Ball (Ball and Hall 1965)			x
15	Trcovw (Milligan and Cooper 1985)			x
16	Tracew (Milligan and Cooper 1985)			x
17	Friedman (Friedman and Rubin 1967)			x
18	Rubin (Friedman and Rubin 1967)			x
19	Dunn (Dunn 1974)		x	x
20	SDbw (Halkidi and Vazirgiannis 2001)			x
21	Hubert (Hubert and Arabie 1985)			x
22	Dindex (Lebart et al. 2000)			x
23	McClain (McClain and Rao 1975)			x
24	Tau (Rohlf 1974)			x
25	Frey(Frey and Van Groenewoud 1972)			x
26	Gplus (Rohlf 1974)			x
27	Ptbiserial (Milligan and Cooper 1985)			x
28	Beale (Beale 1969)			x
29	Duda(Duda and Hart 1973)			x
30	SDindex			x

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.

Table 1: Sample Size Statistics

Mean	698
Median	293
Std. Deviation	1697
Minimum	10
Maximum	20000

Table 2: Statistics on the Number of Variables

Mean	17
Median	15
Std. Deviation	11.48
Minimum	10
Maximum	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.

Variáveis

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

Dessa forma, para que $2^p \leq 52$, o número de variáveis deve ser $p = 5$, logo $n = 2^5 = 32$

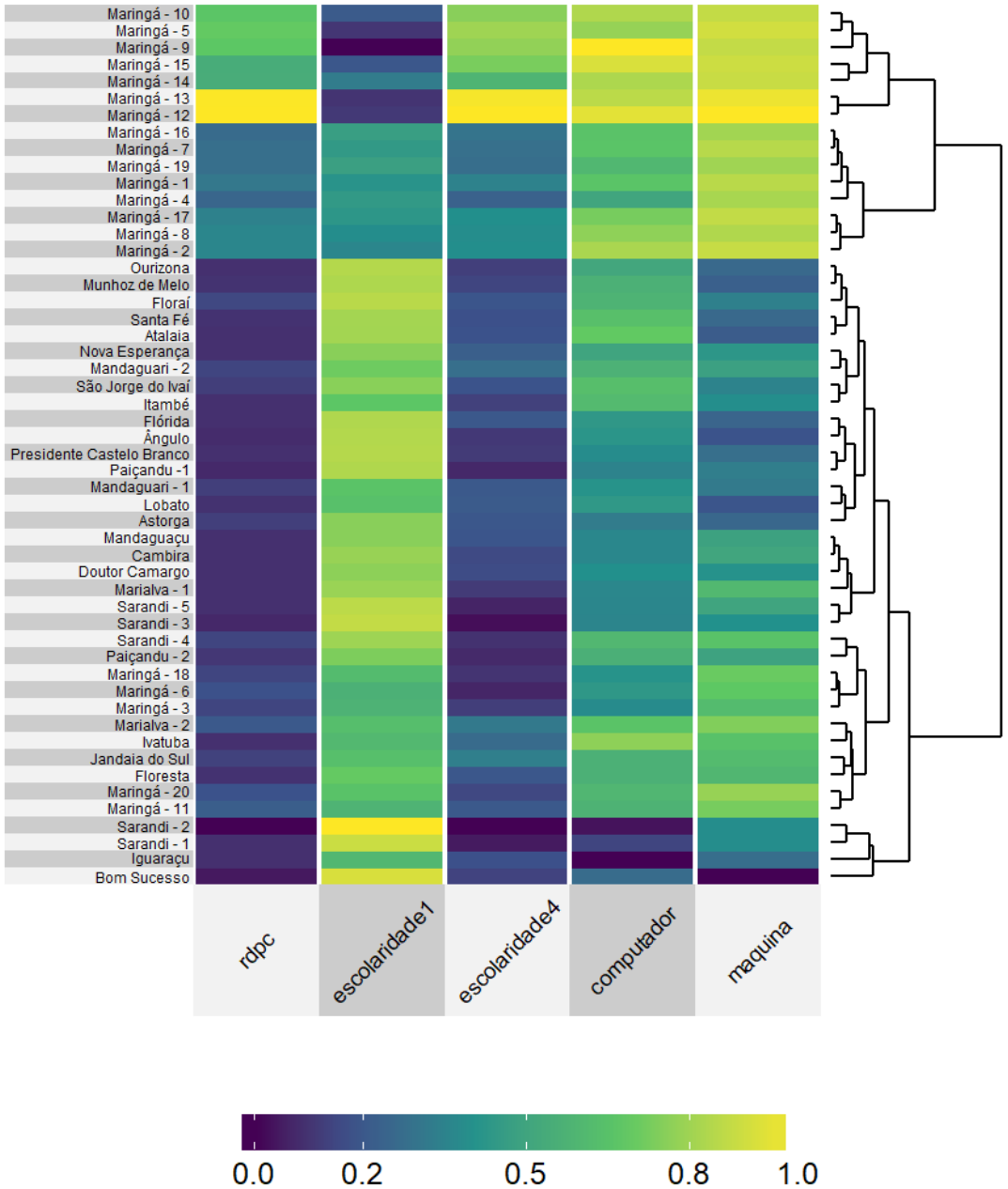
Correlação entre a distância Euclidiana e a Cofenética

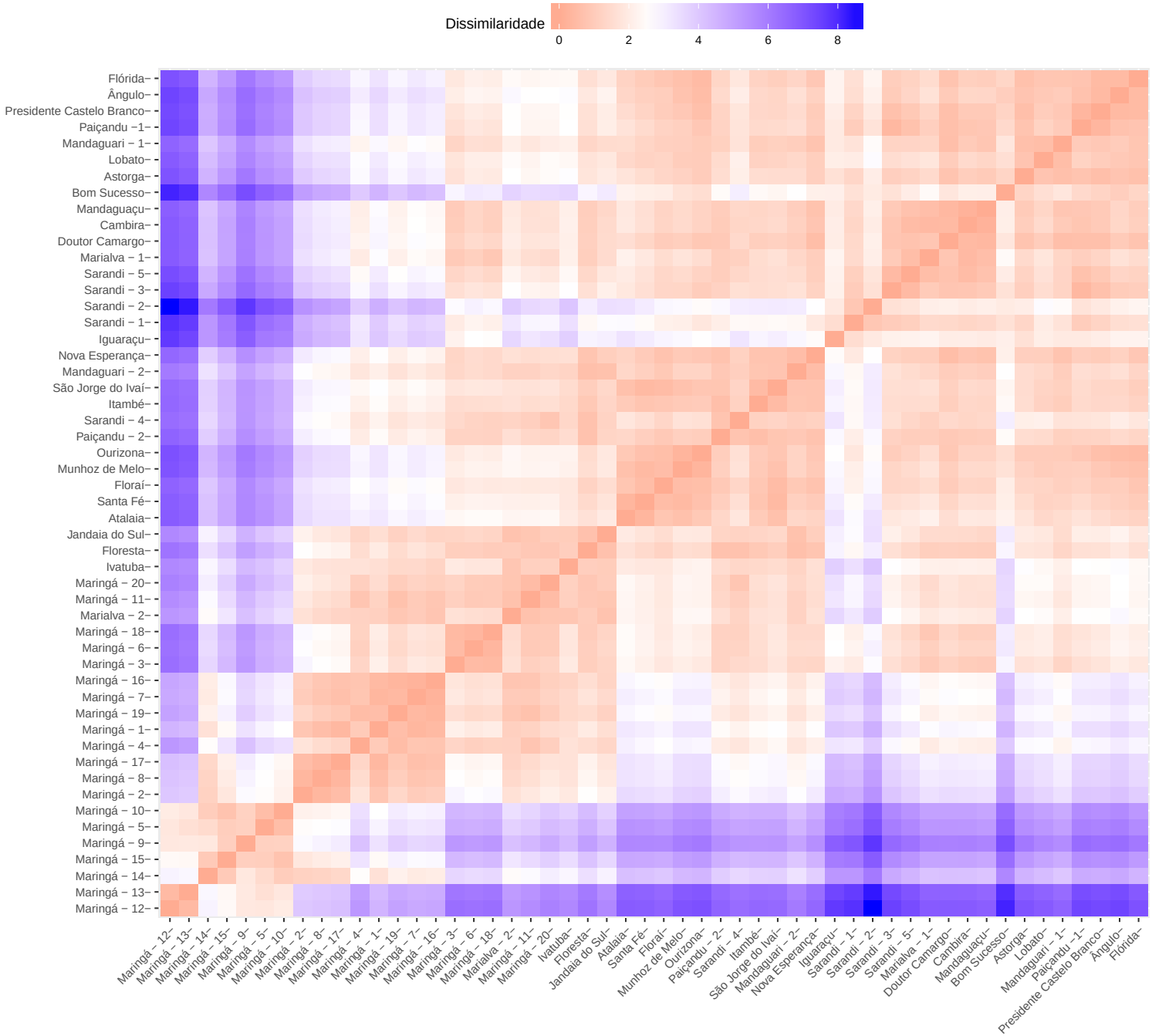
O coeficiente de correlação cofenético C é obtido calculando a correlação entre a matriz D e a matriz Z

$$C = Cor(D, Z) = \frac{\sum_{i<j}(D_{ij} - \bar{D})(Z_{ij} - \bar{Z})}{\sqrt{\sum_{i<j}(D_{ij} - \bar{D})^2 \sum_{i<j}(Z_{ij} - \bar{Z})^2}}$$

- D_{ij} é a distância Eucludiana entre i e j;
- Z_{ij} é a distância Cofenética entre i e j;
- $\bar{Z}$ e $\bar{D}$ são as respectivas médias.

Resultados


Maringá - 10
Maringá - 5
Maringá - 9
Maringá - 15
Maringá - 14
Maringá - 13
Maringá - 12
Maringá - 16
Maringá - 7
Maringá - 19
Maringá - 1
Maringá - 4
Maringá - 17
Maringá - 8
Maringá - 2
Ourizona
Munhoz de Melo
Floraí
Santa Fé
Atalaia
Nova Esperança
Mandaguari - 2
São Jorge do Ivaí
Itambé
Flórida
Ângulo
Presidente Castelo Branco
Paiçandu -1
Mandaguari - 1
Lobato
Astorga
Mandaguaçu
Cambira
Doutor Camargo
Marialva - 1
Sarandi - 5
Sarandi - 3
Sarandi - 4
Paiçandu - 2
Maringá - 18
Maringá - 6
Maringá - 3
Marialva - 2
Ivatuba
Jandaia do Sul
Floresta
Maringá - 20
Maringá - 11
Sarandi - 2
Sarandi - 1
Iguaraçu
Bom Sucesso
rdpc
escolaridade1
escolaridade4
computador
maquina
0.0
0.2
0.5
0.8
1.0


Dissimilaridade
0
2
4
6
8
Flórida
Ângulo
Presidente Castelo Branco
Paiçandu −1
Mandaguari − 1
Lobato
Astorga
Bom Sucesso
Mandaguaçu
Cambira
Doutor Camargo
Marialva − 1
Sarandi − 5
Sarandi − 3
Sarandi − 2
Sarandi − 1
Iguaraçu
Nova Esperança
Mandaguari − 2
São Jorge do Ivaí
Itambé
Sarandi − 4
Paiçandu − 2
Ourizona
Munhoz de Melo
Floraí
Santa Fé
Atalaia
Jandaia do Sul
Floresta
Ivatuba
Maringá − 20
Maringá − 11
Marialva − 2
Maringá − 18
Maringá − 6
Maringá − 3
Maringá − 16
Maringá − 7
Maringá − 19
Maringá − 1
Maringá − 4
Maringá − 17
Maringá − 8
Maringá − 2
Maringá − 10
Maringá − 5
Maringá − 9
Maringá − 15
Maringá − 14
Maringá − 13
Maringá − 12

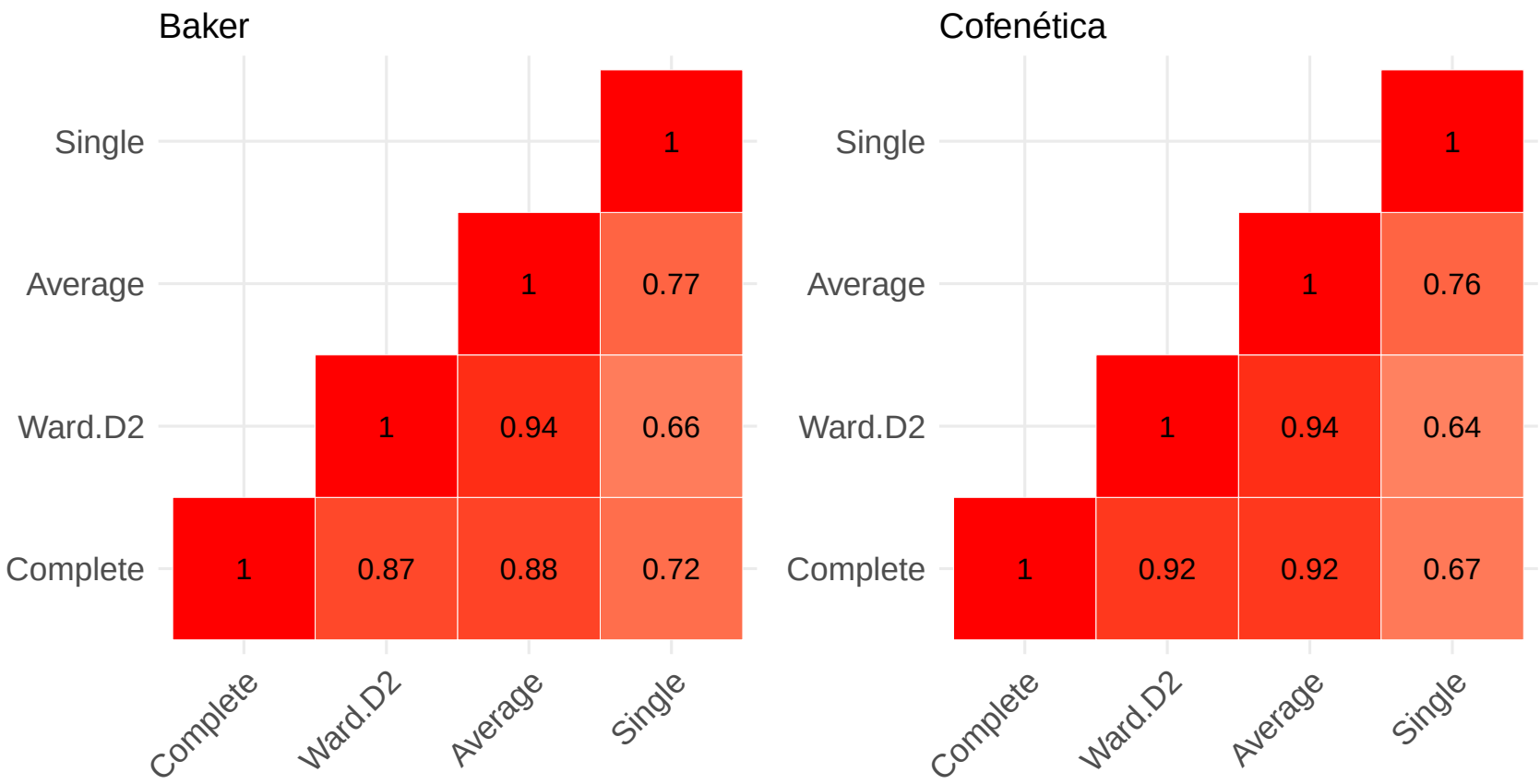
Formação dos clusters

Correlação entre as distâncias cofenéticas e euclidiana

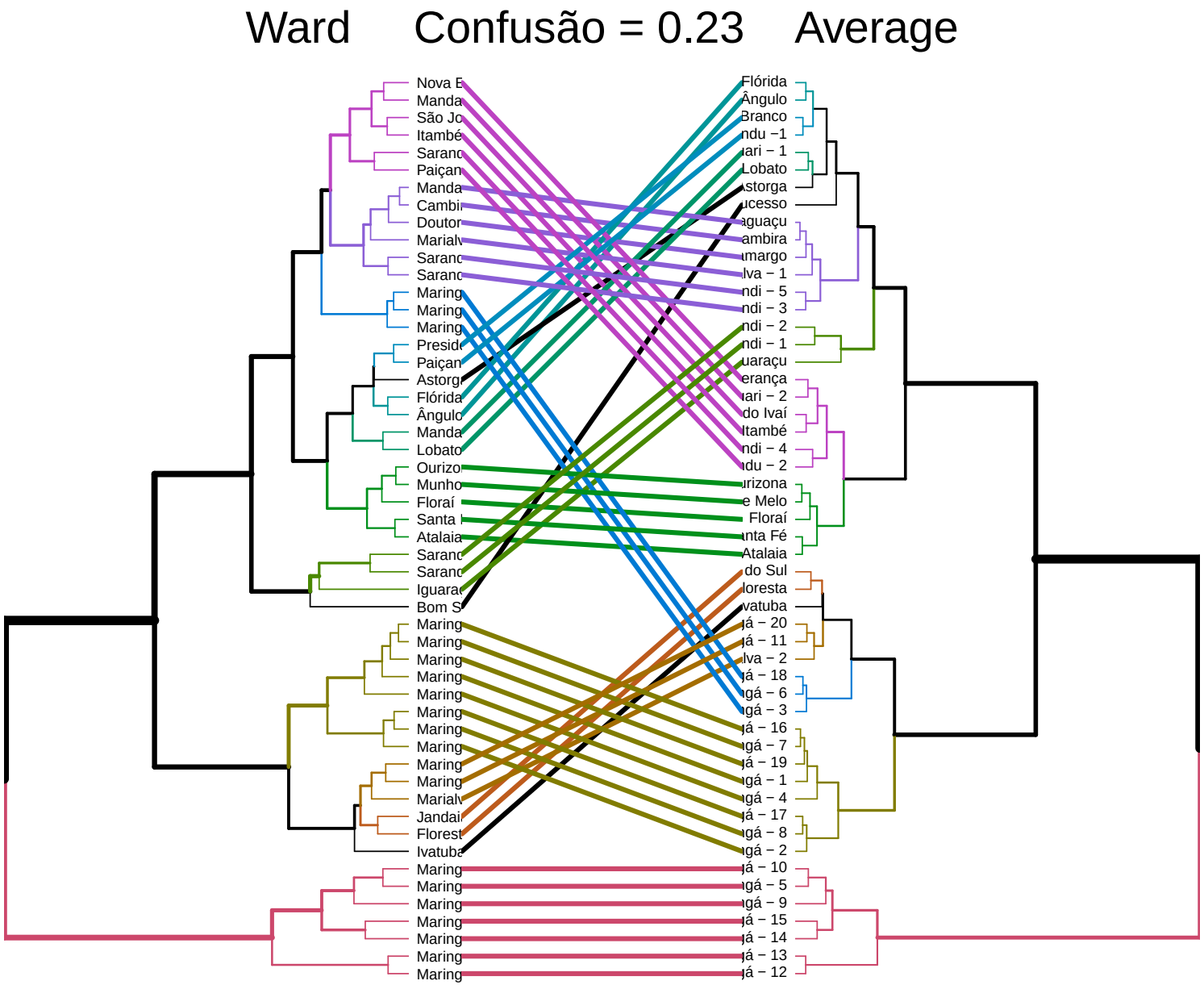
	5 variáveis		11 variáveis	
Linkages	$\hat{\rho}$	valor-p	$\hat{\rho}$	valor-p
Ward	0.8148	$<$0.001	0.6079	$<$0.001
Complete	0.8276	$<$0.001	0.7623	$<$0.001
Average	0.8490	$<$0.001	0.8247	$<$0.001
Single	0.7616	$<$0.001	0.5233	$<$0.001

Uma forma de verificar quão bem os clusters foram criados é pela correlação entre as distâncias Euclidiana e Cofenéticas. De modo que, valores superiores a **0.75** indicam boas formações(Kassambara, 2017).

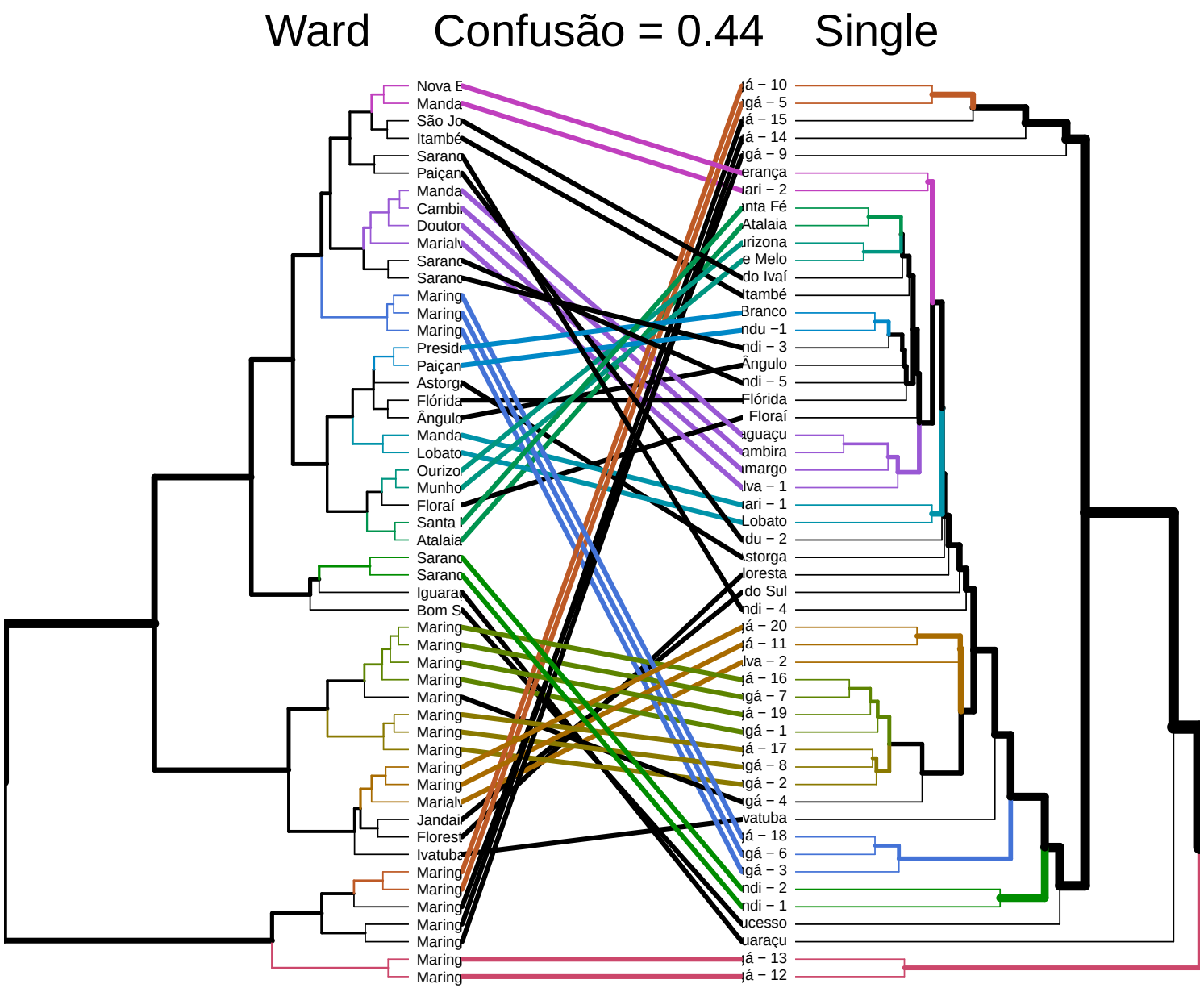
Comparação de clusters



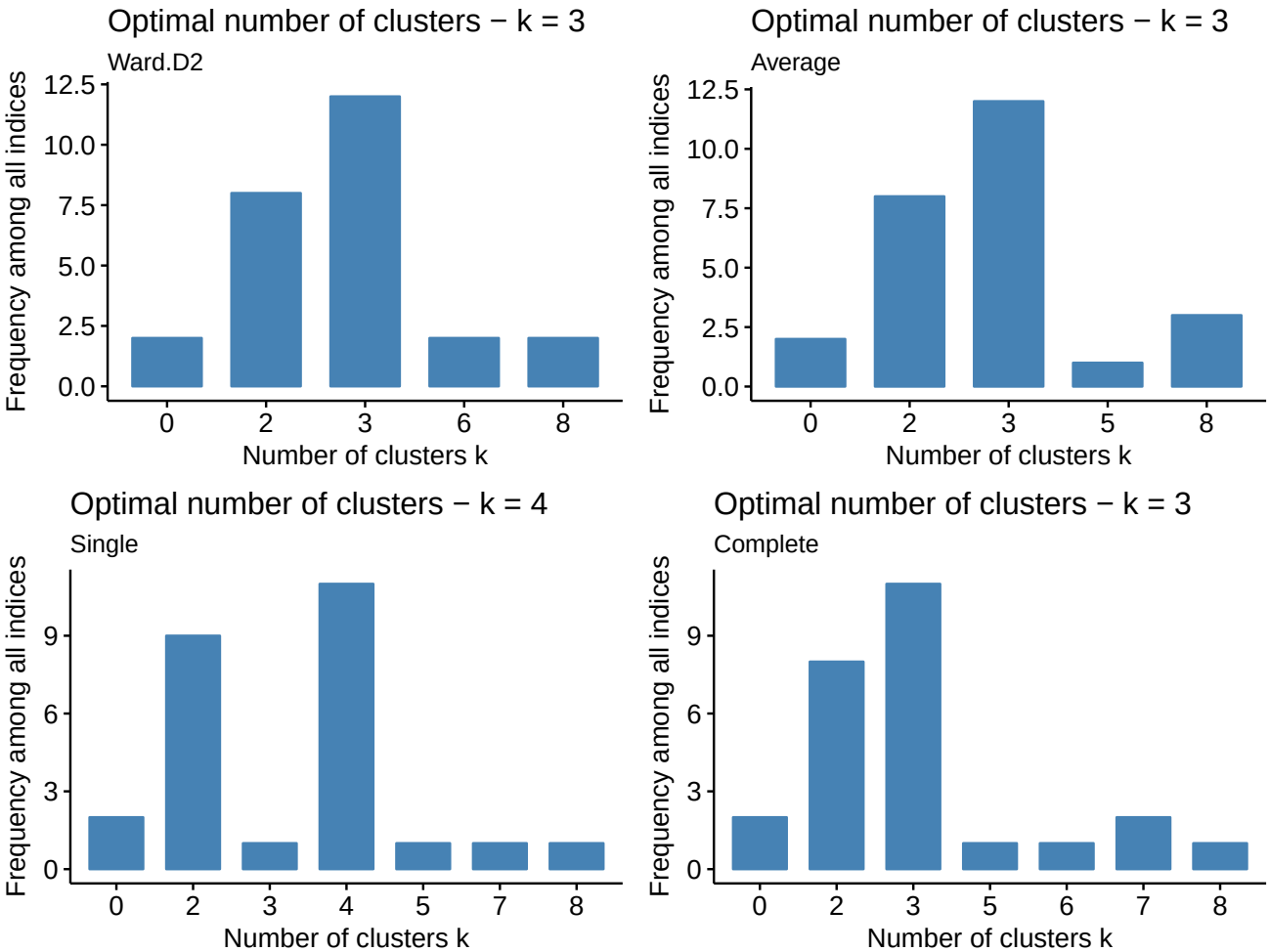
Comparação de clusters



Comparação de clusters



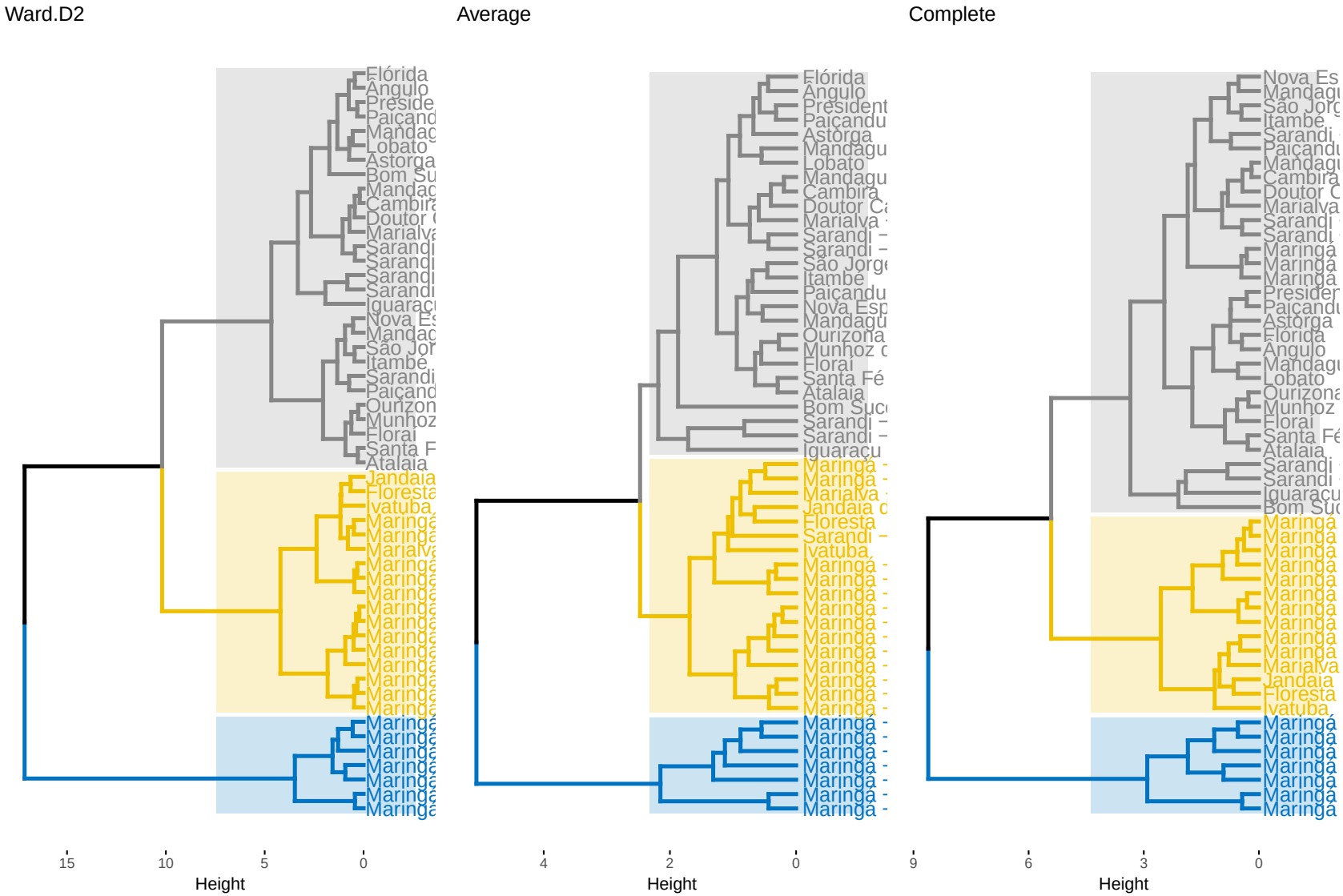
Número ótimo de clusters



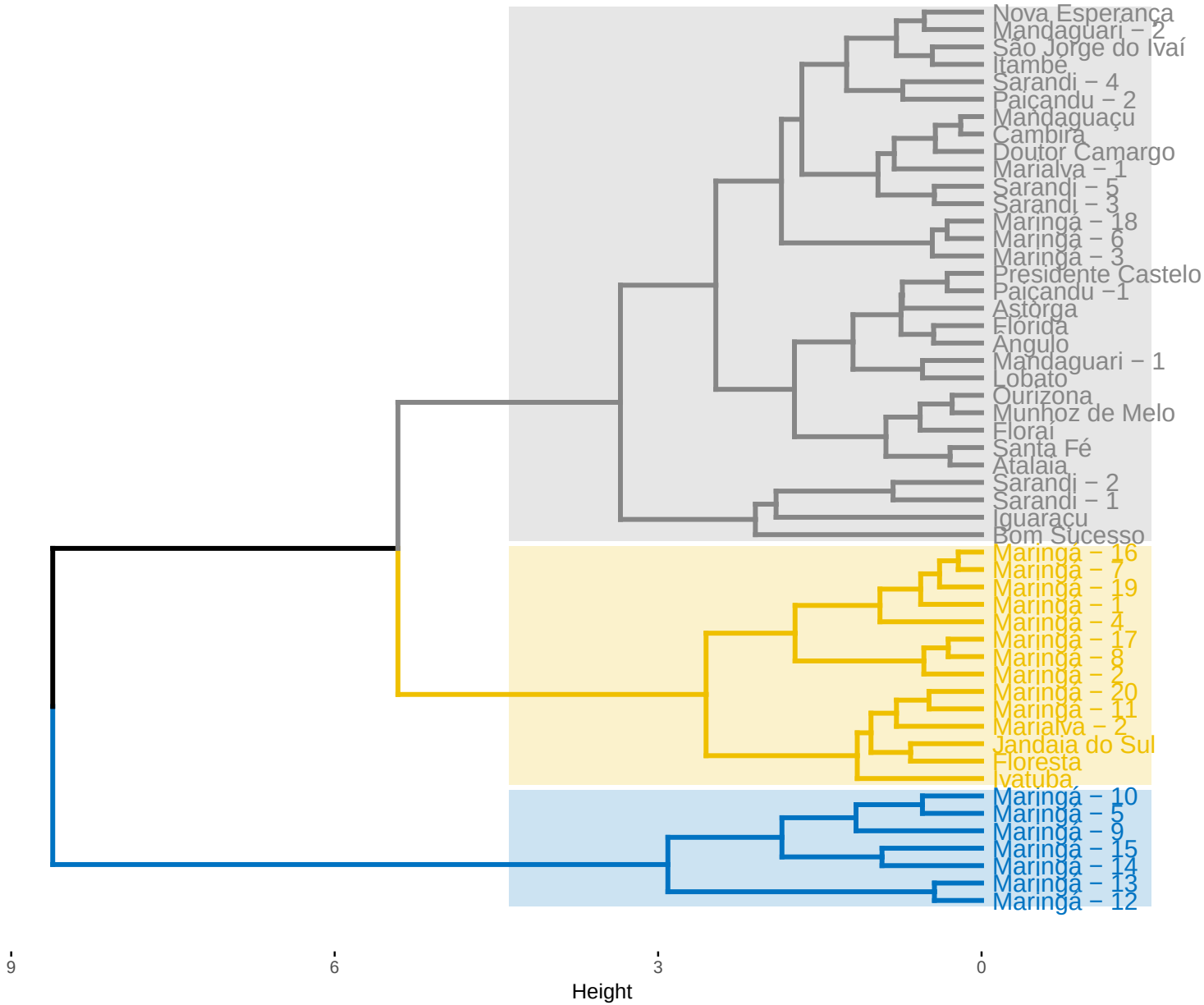
Medidas utilizadas pelo SAS

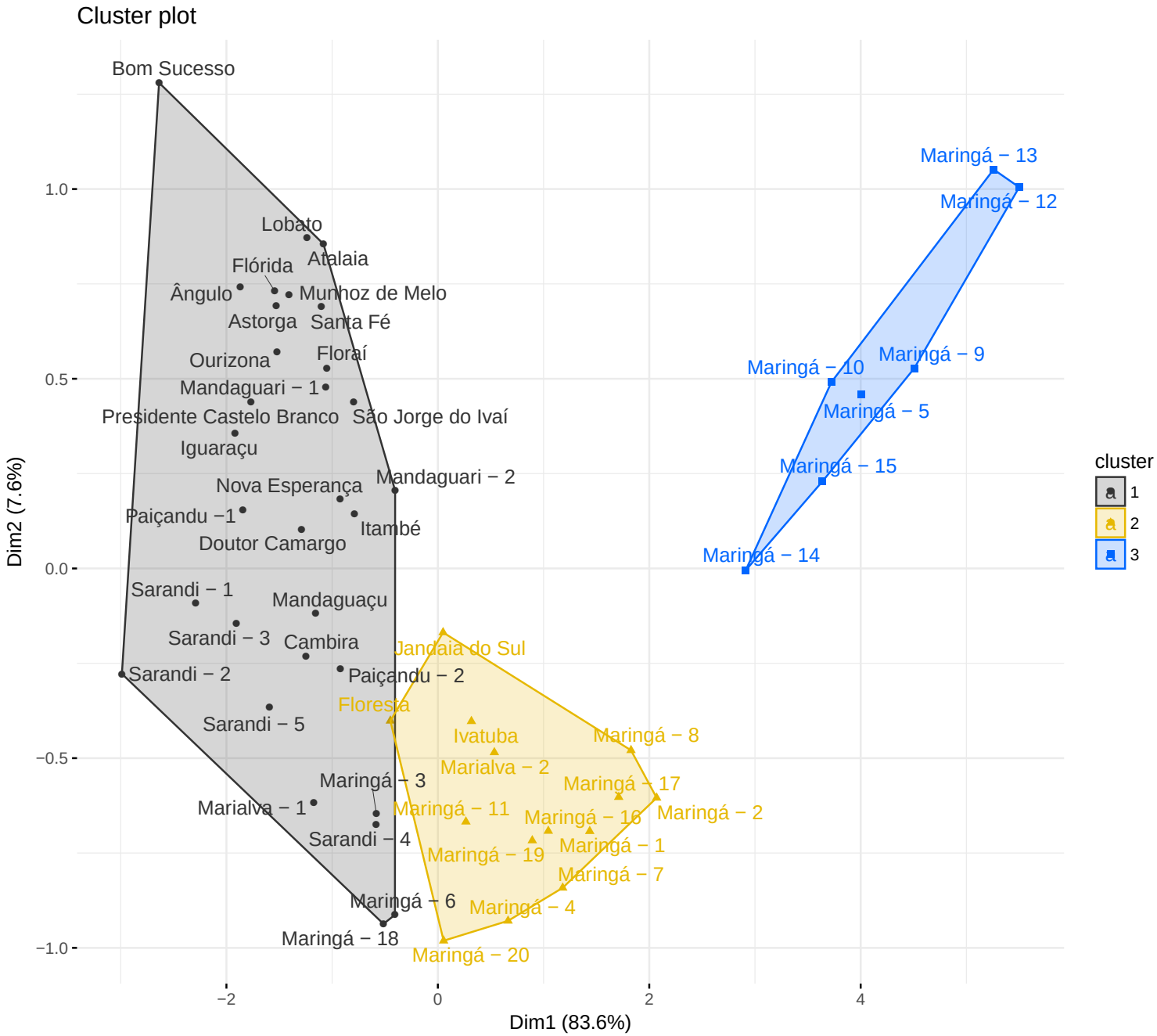
Índice	Valor	Número de clusters
PseudoT2	14.0489	3
CH	86.9732	3
CCC	34.5324	3

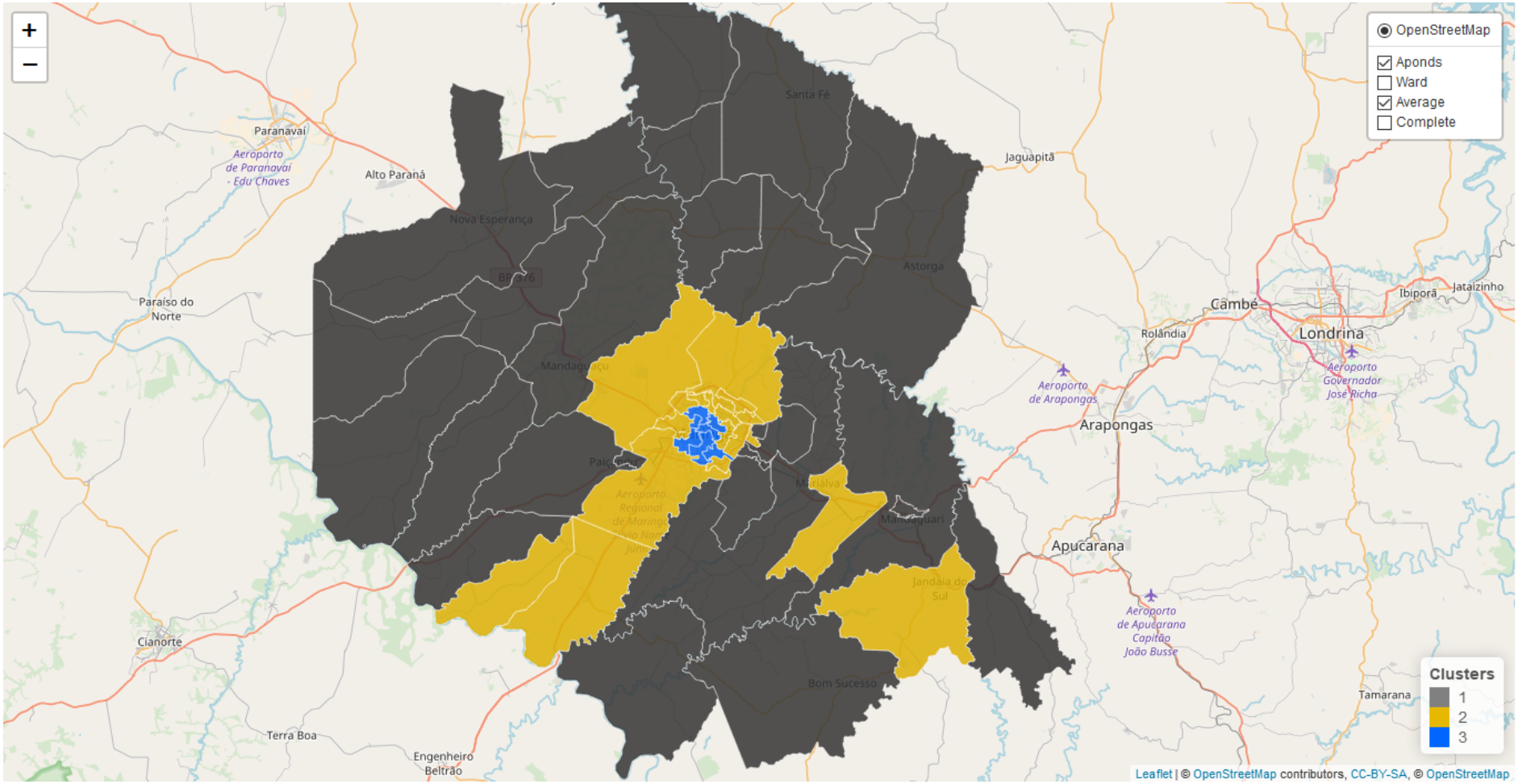
Dendogramas

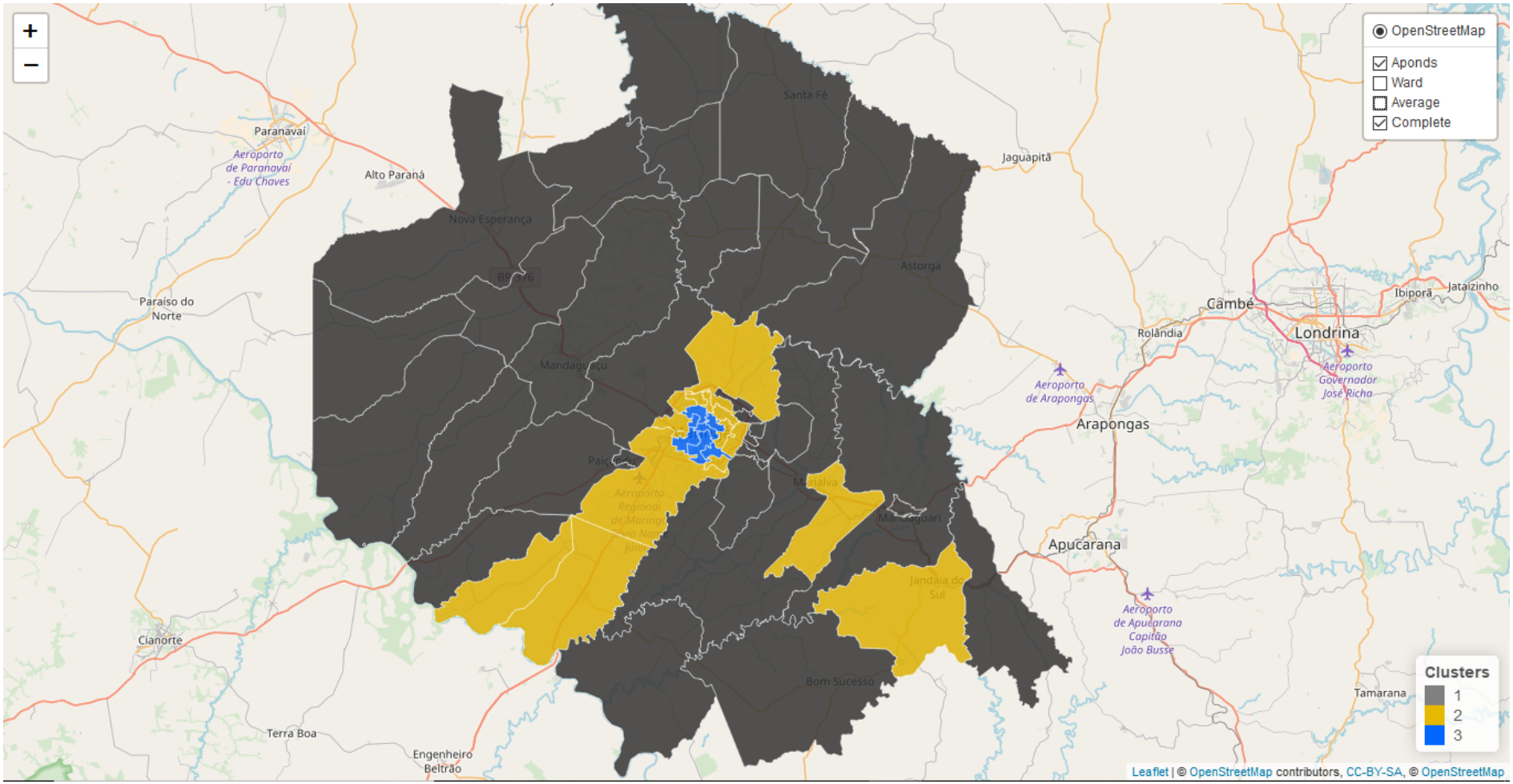


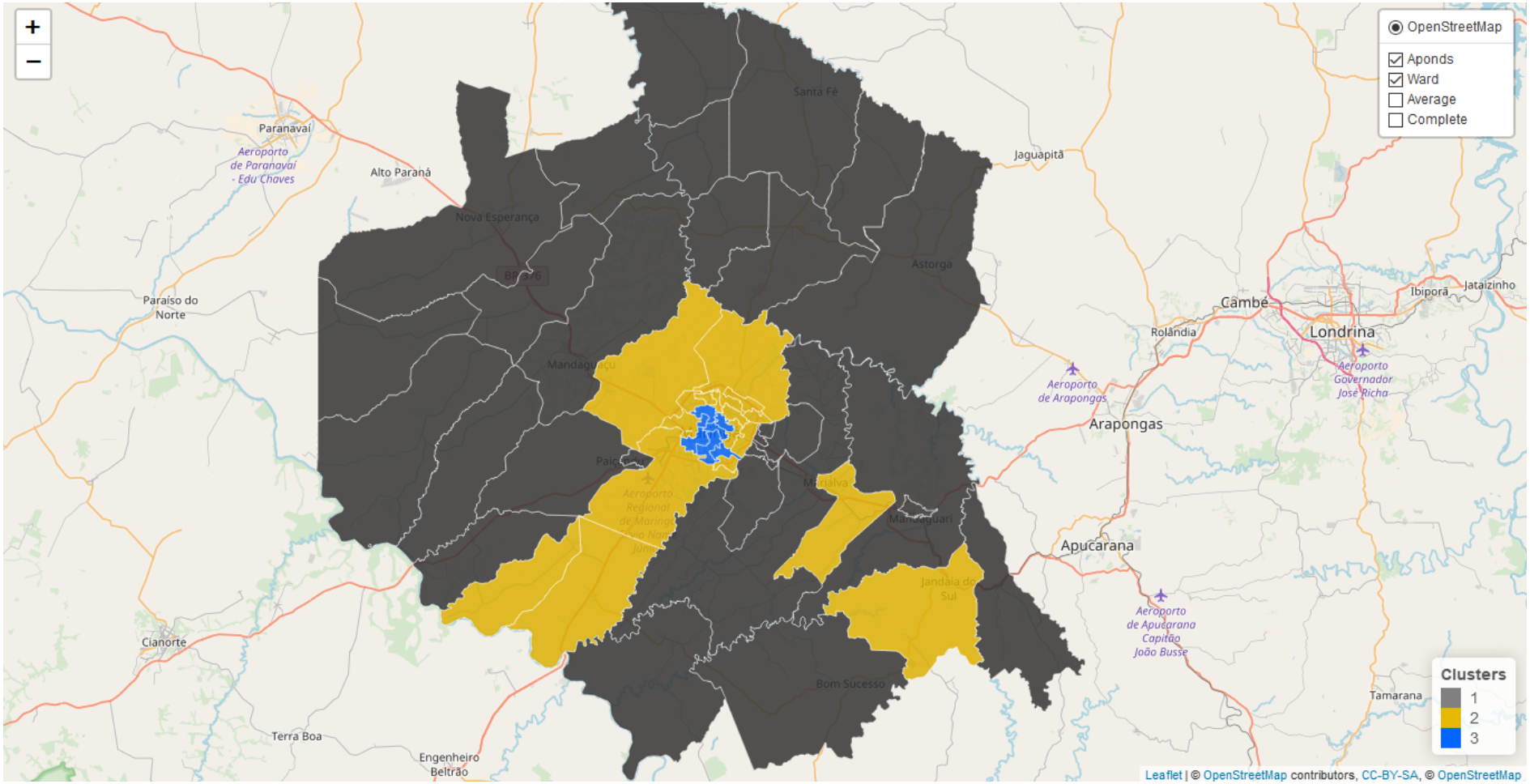
Complete

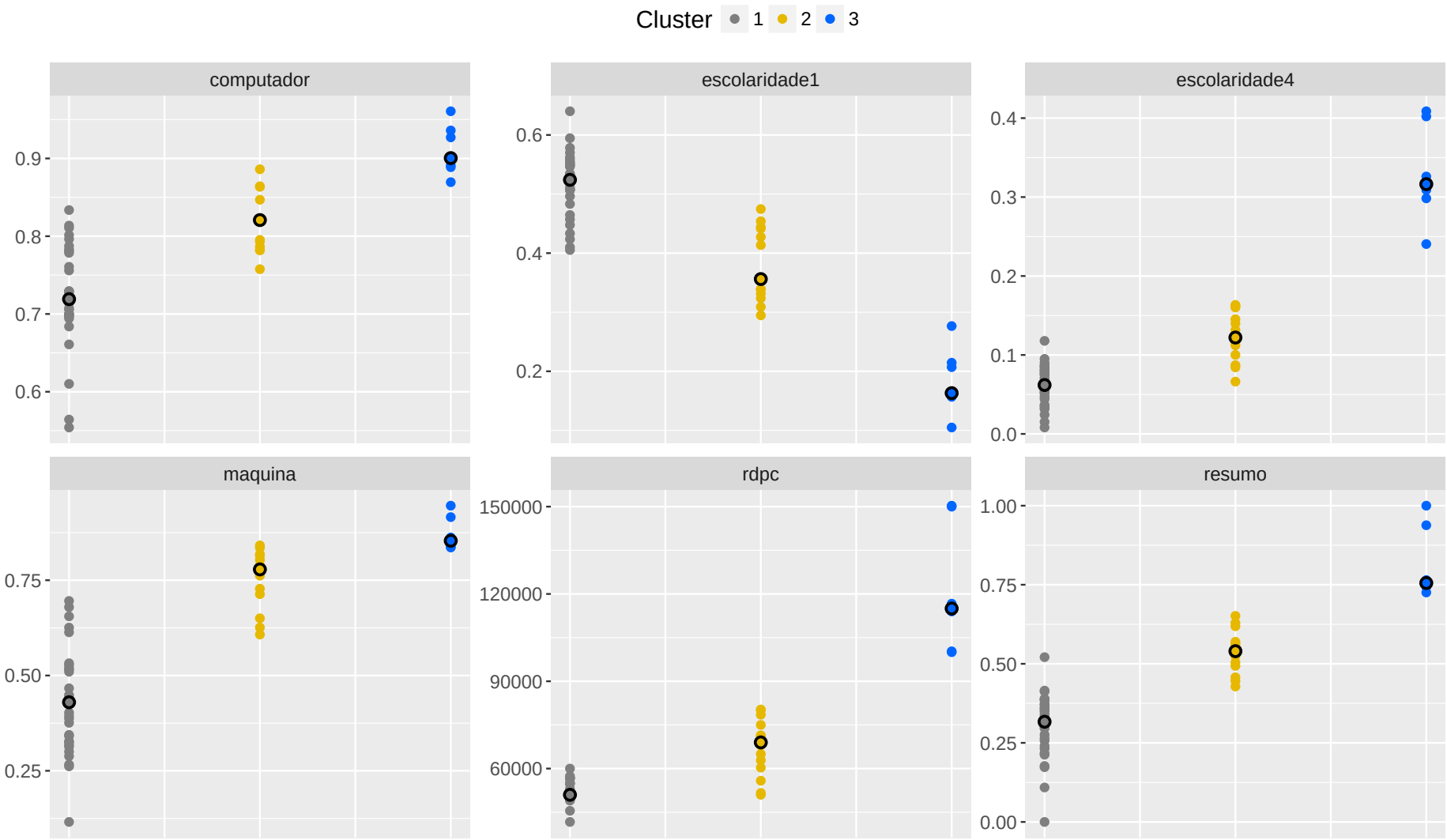

Nova Esperança
Mandaguari − 2
São Jorge do Ivaí
Itambé
Sarandi − 4
Paiçandu − 2
Mandaguaçu
Cambira
Doutor Camargo
Marialva − 1
Sarandi − 5
Sarandi − 3
Maringá − 18
Maringá − 6
Maringá − 3
Presidente Castelo
Paiçandu −1
Astorga
Flórida
Ângulo
Mandaguari − 1
Lobato
Ourizona
Munhoz de Melo
Floraí
Santa Fé
Atalaia
Sarandi − 2
Sarandi − 1
Iguaraçu
Bom Sucesso
Maringá − 16
Maringá − 7
Maringá − 19
Maringá − 1
Maringá − 4
Maringá − 17
Maringá − 8
Maringá − 2
Maringá − 20
Maringá − 11
Marialva − 2
Jandaia do Sul
Floresta
Ivatuba
Maringá − 10
Maringá − 5
Maringá − 9
Maringá − 15
Maringá − 14
Maringá − 13
Maringá − 12
9
6
3
0
Height


Cluster plot
Bom Sucesso
Lobato
Atalaia
Flórida
Ângulo
Munhoz de Melo
Astorga
Santa Fé
Ourizona
Floraí
Mandaguari − 1
Presidente Castelo Branco
São Jorge do Ivaí
Iguaraçu
Nova Esperança
Mandaguari − 2
Paiçandu −1
Itambé
Doutor Camargo
Sarandi − 1
Mandaguaçu
Sarandi − 3
Cambira
Jandaia do Sul
Sarandi − 2
Paiçandu − 2
Floresta
Sarandi − 5
Ivatuba
Maringá − 8
Marialva − 2
Maringá − 17
Maringá − 3
Maringá − 11
Maringá − 16
Maringá − 2
Marialva − 1
Sarandi − 4
Maringá − 19
Maringá − 1
Maringá − 7
Maringá − 6
Maringá − 4
Maringá − 18
Maringá − 20
Maringá − 13
Maringá − 12
Maringá − 9
Maringá − 10
Maringá − 5
Maringá − 15
Maringá − 14
cluster
1
2
3
Dim2 (7.6%)
Dim1 (83.6%)
1.0
0.5
0.0
−0.5
−1.0
−2
0
2
4

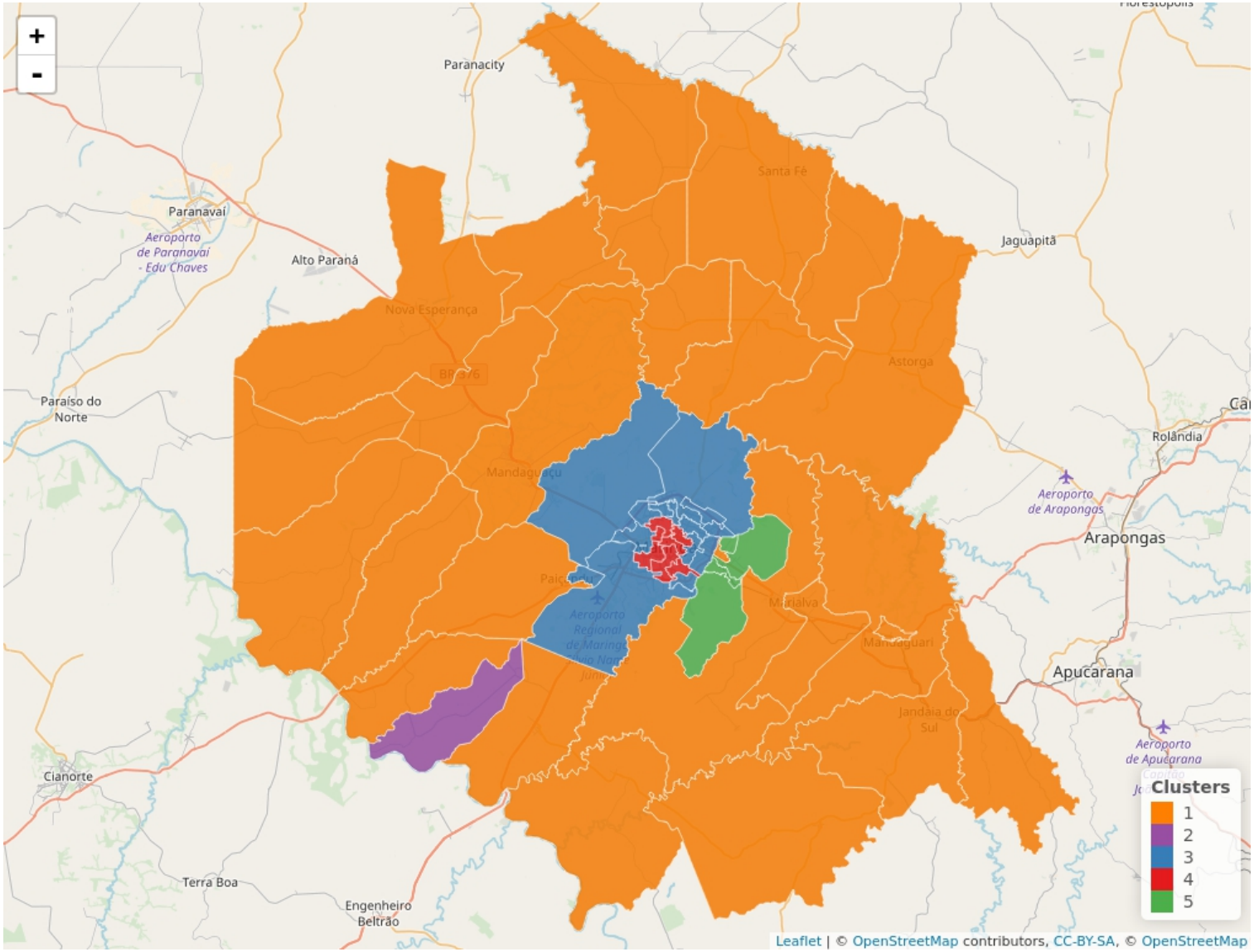

OpenStreetMap
Aponds
Ward
Average
Complete
Clusters
1
2
3
Paranavaí
Aeroporto de Paranavaí - Edu Chaves
Alto Paraná
Santa Fé
Jaguapitã
Nova Esperança
Astorga
Paraíso do Norte
Cambé
Ibiporã
Jataizinho
Rolândia
Londrina
Aeroporto Governador José Richa
Aeroporto de Arapongas
Arapongas
Apucarana
Aeroporto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


OpenStreetMap
Aponds
Ward
Average
Complete
Clusters
1
2
3
Paranavaí
Aeroporto de Paranavaí - Edu Chaves
Alto Paraná
Santa Fé
Jaguapitã
Nova Esperança
Astorga
Paraíso do Norte
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
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap


OpenStreetMap
Aponds
Ward
Average
Complete
Clusters
1
2
3
Paranavaí
Aeroporto de Paranavaí - Edu Chaves
Alto Paraná
Santa Fé
Jaguapitã
Nova Esperança
Astorga
Paraíso do Norte
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
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap


Cluster 1 2 3
computador
escolaridade1
escolaridade4
maquina
rdpc
resumo
0.9
0.8
0.7
0.6
0.6
0.4
0.2
0.4
0.3
0.2
0.1
0.0
0.75
0.50
0.25
150000
120000
90000
60000
1.00
0.75
0.50
0.25
0.00

Considerando as 11 variáveis e 5 clusters



Referências

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

Muito obrigado!