Marlon Vinícius Alves de Araújo

Métodos de *Clustering* em Aprendizado de Máquinas Não Supervisionado

Niterói - RJ, Brasil

21 de setembro de 2021

Universidade Federal Fluminense

Marlon Vinícius Alves de Araújo

Métodos de *Clustering* em Aprendizado de Máquinas Não Supervisionado

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientadora: Profª. Draª. Karina Yuriko Yaginuma

Niterói - RJ, Brasil

21 de setembro de 2021

Universidade Federal Fluminense

Marlon Vinícius Alves de Araújo

Métodos de *Clustering* em Aprendizado de Máquinas Não Supervisionado

Monografia de Projeto Final de Graduação sob o título *"Métodos de Clustering em Aprendizado de Máquinas Não Supervisionado"*, defendida por Marlon Vinícius Alves de Araújo e aprovada em 21 de setembro de 2021, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profª. Draª. Karina Yuriko Yaginuma
Departamento de Estatística – UFF

Prof. Dr. Hugo Henrique Kegler dos Santos
Departamento de Estatística – UFF

Profª. Drª. Patrícia Lusié Velozo da Costa
Departamento de Estatística – UFF

Niterói, 21 de setembro de 2021

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

A658m Araújo, Marlon Vinícius Alves de
Métodos de clustering em aprendizado de máquinas não supervisionado / Marlon Vinícius Alves de Araújo ; Karina Yuriko Yaginuma, orientadora. Niterói, 2021.
89 f.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2021.

1. Clustering. 2. Cluster. 3. Método k-means. 4. Método complete linkage. 5. Produção intelectual. I. Yaginuma, Karina Yuriko, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

Atualmente, conforme a tecnologia avança, a quantidade de dados cresce exponencialmente, com milhões de terabytes de dados sendo gerados diariamente. Para obter informações a partir de um conjunto de dados, métodos de *machine learning*, ou aprendizado de máquinas, são utilizados para análises, previsões, resolução de problemas, de acordo com o que se busca extrair, automatizando o desenvolvimento de modelos analíticos. Porém, por mais que seja "fácil"o acesso há diversas bases de dados, em alguns casos, as bases não conterão todas as informações almejadas, como dados rotulados, ou categorizados. Isso acontece porque coletar dados anotados pode ser extremamente caro, custar muito tempo, e em certas situações, até mesmo impossível. Para lidar com essa ausência de informações desejadas, são utilizadas técnicas de aprendizado de máquinas não supervisionado, que auxiliam na detecção de padrões e percepções ocultas nos dados analisados. Entre diversos métodos, um dos mais importantes dentro de aprendizagem não supervisionada é o *clustering*, ou agrupamento, em que seus algoritmos processarão os dados, permitindo encontrar *clusters* (grupos) caso existam, de forma que os elementos dentro do mesmo *cluster* sejam o mais semelhante possível, e tenham menos ou nenhuma semelhança com os elementos de outro grupo. O objetivo deste trabalho é estudar e aplicar algoritmos de *clustering* em um conjunto de dados não rotulado, utilizando suas respectivas ferramentas na linguagem de programação R, verificando se os algoritmos são capazes de fornecer resultados eficientes e confiáveis.

Palavras-chave: *Clustering*. *Cluster*. Método *k-means*. Método *complete linkage*.

Dedicatória

Este trabalho é dedicado à todos que sempre estiveram ao meu lado nesta fase da minha vida, em especial à minha mãe e à minha companheira.

Agradecimentos

Primeiramente, agradeço por tudo à minha mãe Kátia Alves, pois só foi possível chegar até aqui graças a seu esforço, sua dedicação, em querer ver o seu filho formado. Agradeço à minha companheira Thawana Nogueira, por todo o amor, apoio, incentivo, paciência, nunca deixou de estar ao meu lado.

Agradeço à todos os meus amigos que fizeram parte da minha vida nesses últimos anos, em especial Carolina Lourenço e Gabriela de Barros, que sempre estiveram comigo desde o início da graduação. Agradeço aos meus amigos Ayron Borges, Simone Galdino, e Rayner Mello, pelos momentos de conversas e filas do bandejão. E aos meus parceiros Matheus Machado e Natan Vaz, que também passaram pelas mesmas dificuldades nos últimos períodos.

Agradeço à minha família, por sempre acreditar em mim. À minha tia Edna por sempre querer me ajudar no que fosse preciso. À minha prima-irmã Nathália por sempre me motivar, e à minha prima-irmã Giullia, que também sempre me motivou. E à minha tia Fernanda por sempre me incentivar à graduação.

Agradeço imensamente à professora Karina Yaginuma por ter aceitado ser minha orientadora, oferecendo toda atenção, suporte, motivação, para que fosse possível concluir este trabalho. Gostaria de agradecer ao professor Hugo por toda assistência prestada, e também à professora Patrícia por toda ajuda prestada.

Agradeço à todos que me apoiaram de alguma forma.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Neste Capítulo, inicialmente é apresentado a motivação para a utilização do aprendizado de máquinas não supervisionado. Na seção 1.2, são definidos os objetivos deste trabalho, e por último, a organização.

1.1 Motivação

Em determinadas situações, obter dados rotulados (dados com algum rótulo ou identificação especial) pode custar muito tempo e dinheiro, ou até ser impossível. Ou seja, nem sempre uma base de dados terá os dados de entrada com a saída correspondente (rótulo), fazendo com que o aprendizado de máquinas não supervisionado, que permite encontrar padrões nos dados, seja de grande importância para resolver tais casos, justamente por oferecer a possibilidade de obtenção de informações que são desejáveis para análises.

Portanto, como no dia a dia é muito mais fácil conseguir uma base de dados sem rótulos [4], os métodos não supervisionados auxiliam na descoberta de todos os tipos de padrões desconhecidos (caso existam), ajudam a encontrar recursos que podem ser úteis para categorização e podem dividir automaticamente o conjunto de dados em grupos com base em suas semelhanças, permitindo assim futuras análises de acordo com supostas necessidades.

Além disso, o aprendizado de máquinas não supervisionado pode funcionar como uma forma de pré-processamento, possibilitando o treinamento de modelos para um aprendizado de máquinas supervisionado, pois este pré-processamento geraria dados categorizados na base.

1.2 Objetivos

Os principais objetivos deste trabalho são:

- Entender os conceitos e métodos do aprendizado de máquinas não supervisionado com *clustering*;
- Compreender o desempenho e como a aplicação dos algoritmos de *clustering* agem sobre uma base de dados não rotulada;
- Entender e analisar os *clusters* gerados através dos métodos;
- Aprender como supostas análises poderiam ser feitas com os resultados adquiridos.

Com a utilização da linguagem de programação R [14][15], serão explorados e comparados algoritmos hierárquicos e não hierárquicos de *clustering*, e avaliar como os dados rotulados fornecem informações e futuras inferências.

1.3 Organização

O Capítulo 2 começa introduzindo os conceitos de aprendizados de máquinas, e a seguir contém os conceitos teóricos de *clustering*, dividido em 4 seções, abordando medidas de similaridade, os principais métodos de *clustering hierárquico*, e o método não hierárquico *k-means*. No Capítulo 3 são apresentados a aplicação dos algoritmos e seus resultados. Por fim, no Capítulo 4 são apresentadas as considerações finais.

2 Metodologia

Este Capítulo contém toda a parte teórica deste trabalho. A primeira seção apresenta uma breve descrição sobre aprendizado de máquinas. A seção 2.2 apresenta os conceitos de *clustering*. Na seção 2.3, são definidas as medidas de similaridade. Na seção 2.4, são apresentados os métodos de *clustering* hierárquicos. Por fim, a seção 2.5 define os métodos de *clustering* não hierárquicos, e apresenta o método *k-means*.

2.1 Aprendizado de Máquinas (*Machine Learning*)

O Aprendizado de Máquinas, também conhecido como aprendizado automático, é um método de análise de dados que utiliza algoritmos que o permitem extrair informações e identificar padrões nos dados, automatizando a construção de modelos analíticos, a fim de fazer previsões ou decisões.

O Aprendizado de Máquinas apresenta 3 tipos principais:

- **Supervisionado**: Quando a base de dados apresenta os dados de entrada e a saída correspondente (variável resposta). Aplica-se um particionamento na base, em conjuntos treino e teste, e um algoritmo de aprendizado supervisionado analisa o conjunto de dados treino, para produzir uma função inferida, sendo esta testada no conjunto de dados teste, verificando sua eficiência, podendo assim ser usada para gerar previsões razoáveis para a variável de interesse. Os algoritmos de aprendizado de máquinas supervisionado são aplicados de duas formas, divididos em problemas de regressão e classificação [8][13];

- **Não Supervisionado**: Quando a base de dados apresenta apenas os dados de entrada, sem nenhum rótulo (variável resposta). Nesta situação, o algoritmo não supervisionado gera uma função para descrever a estrutura oculta de dados não rotulados, encontrando padrões ou descobrindo grupos dos dados de entrada;

- **Por reforço**: Não há conjunto de treinamento, rotulados ou não. O modelo é treinado para tomar uma sequência de decisões com o objetivo de alcançar uma determinado meta, recebendo *feedbacks* (retornos) quanto a recompensas ou penalidades [9]. A meta seria maximizar a recompensa total.

Há outros tipos de aprendizado de máquinas como:

- **Semi-Supervisionado**: Fica entre o supervisionado e o não supervisionado, e inclui o problema de ambos. Caracteriza-se quando o conjunto de dados apresenta tanto dados rotulados quanto saídas ausentes, o que seria um conjunto de dados incompleto;

- **Transdução**: Também denominada como aprendizagem transdutiva, sendo o conjunto inteiro das instâncias do problema conhecido no momento do aprendizado, mas com parte dos objetivos ausente [16].

Em relação aos algoritmos de aprendizagem de máquinas, eles podem ser divididos em 3 categorias amplas, que representam justamente os 3 principais métodos de aprendizado: aprendizagem supervisionada, aprendizado sem supervisão e aprendizado de reforço. No entanto, como este trabalho é dedicado ao método de aprendizado de máquinas não supervisionado, sem estudar os conceitos dos demais métodos, podemos citar alguns dos principais algoritmos, sem citar os que pertencem ao aprendizado não supervisionado, pois veremos em 2.1.2. Alguns dos algoritmos são:

−Árvores de Decisão;	−Regressão Logística;
−*Suport Vectors Machine* (SVM);	−Perceptron;
−Redes Bayesianas;	−Classificador Naïve Bayes;
−Discriminação e Classificação;	−Reforço.

2.1.1 Aprendizado de Máquinas Não Supervisionado

O Aprendizado de Máquinas Não Supervisionado é uma técnica de aprendizado de máquinas em que os usuários não precisam supervisionar o modelo. Ao invés disso, permite que o modelo trabalhe por conta própria para descobrir padrões e informações que não foram detectados anteriormente, pois lida com dados não rotulados, como visto anteriormente. Os dados rotulados poderiam ser descritos como um conjunto de dados que possui uma identificação[5], uma "etiqueta"para as observações.

O aprendizado não supervisionado pode ser comparado ao aprendizado que ocorre no cérebro humano enquanto se aprende coisas novas, isso porque se assemelha muito a um ser humano que aprende a pensar por suas próprias experiências, o que o torna mais próximo da IA (Inteligência artificial) real [4]. Uma de suas utilidades é justamente encontrar *insights* [7], que seriam intuições ou percepções a partir dos dados.

Ao contrário do aprendizado supervisionado, o aprendizado não supervisionado não pode ser aplicado diretamente a um problema de regressão ou classificação [4], pois temos os dados de entrada, mas nenhum dado de saída correspondente, além de ser um método menos preciso e confiável, se comparado ao supervisionado.

2.1.2 Algoritmos de aprendizado não supervisionado

Os algoritmos de aprendizado não supervisionado permitem que os usuários realizem tarefas de processamento mais complexas em comparação ao aprendizado supervisionado, ou seja, sua complexidade computacional é maior. Entretanto, a aprendizagem não supervisionada pode ser menos precisa em comparação com outros métodos de aprendizagem, já que não sabem a saída exata com antecedência [4].

Alguns dos principais algoritmos de aprendizado não supervisionado são:

- *Clustering* Hierárquico:
  - *Single Linkage*;
  - *Complete Linkage*;
  - *Average Linkage*;
  - Ward;
- *Clustering* Não Hierárquico:
  - *K-means*;
- *k-Nearest Neighbors* KNN (k vizinhos mais próximos);
- Análise de Componentes Principais;
- Análise de Componentes Independentes;
- Redes Neurais;
- Detecção de anomalias;

- Algoritmo a priori.

O foco deste trabalho será em algoritmos de *clustering*, e seus conceitos serão explorados a partir da Seção 2.4. Mas podemos citar algumas aplicações destes demais algoritmos [7], como por exemplo:

- A detecção de anomalias pode descobrir pontos de dados incomuns em seu conjunto de dados. É útil para encontrar transações fraudulentas;
- Mineração de associação identifica conjuntos de itens que costumam ocorrer juntos em seu conjunto de dados;
- Modelos de variáveis latentes são amplamente usados para pré-processamento de dados. Como reduzir o número de recursos em um conjunto de dados ou decompor o conjunto de dados em vários componentes.

2.2 *Clustering*

A partir da seção 2.2 até o fim do Capítulo 2, o conteúdo deste trabalho é reproduzido baseando-se no livro *Applied multivariate statistical analysis*, dos autores Richard A. Johnson e Dean W. Wichern. Maiores informações podem ser encontradas em [10].

Em uma análise multivariada, compreender uma complexa natureza de relações é de extrema importância. Alguns procedimentos exploratórios possuem técnicas muito utilizadas para estudar os dados com o objetivo de encontrar uma estrutura de agrupamentos "naturais", podendo estes nos fornecer um meio informal para avaliar a dimensionalidade, identificar *outliers* e sugerir hipóteses interessantes sobre relacionamentos.

Clustering é uma técnica estatística usada para classificar elementos em grupos, de forma que elementos dentro de um mesmo *cluster* sejam muito parecidos, e os elementos em diferentes *clusters* sejam distintos entre si. O *clustering*, ou agrupamento, é uma técnica primitiva, pois nenhuma suposição é feita a respeito do número de grupos ou da estrutura dos grupos. O agrupamento é feito com base em semelhanças ou distâncias (dissimilaridades). As entradas necessárias são medidas de semelhança ou dados a partir dos quais as semelhanças podem ser calculadas.

A ideia central da análise de *cluster* é a possibilidade de efetuar a classificação dos objetos em grupos, pois seu objetivo básico é descobrir agrupamentos naturais das variáveis.

Por sua vez, devemos primeiro desenvolver uma escala quantitativa, com a função de medir a associação (similaridade) entre objetos.

Clustering é diferente dos métodos de classificação e análise fatorial. Enquanto a análise de *cluster* objetiva agregar objetos (e não variáveis), fazendo a agregação baseada na distância (proximidade), a classificação trabalha com um número conhecido de grupos, e o objetivo operacional é atribuir novas observações a um desses grupos. E já em análise fatorial, agrega-se variáveis e efetua os agrupamentos em base de padrões de variação (correlação) dos dados.

As técnicas de *clustering* essencialmente tentam formalizar o que os observadores humanos fazem tão bem em duas ou três dimensões, ou seja, mesmo sem o conhecimento preciso de um agrupamento natural, muitas vezes somos capazes de agrupar objetos em gráficos bidimensionais ou tridimensionais de forma visual.

2.3 Medidas de similaridade

Na maioria dos casos, para se produzir uma estrutura de grupo bem simples a partir de um conjunto de dados complexo, requer uma medida de "proximidade" ou "similaridade", e muitas vezes existe uma grande parcela de subjetividade envolvida na escolha de uma medida de similaridade. As considerações importantes incluem a natureza das variáveis (discretas, contínuas, binárias), escalas de medição (nominal, ordinal, intervalar, razão), e requer conhecimento do assunto.

Quando os itens, sejam unidades ou casos, são agrupados, a proximidade geralmente é indicada por algum tipo de distância. Entretanto, as variáveis são geralmente agrupadas com base em coeficientes de correlação ou como medidas de associação.

2.3.1 Coeficientes de distâncias e similaridade para pares de itens

Qualquer medida de distância $d(x, y)$ entre dois pontos x e y é válida desde que satisfaça as seguintes propriedades, onde z é qualquer outro ponto intermediário:

$$\begin{aligned} &d(x,y) = d(y,x) \\ &d(x,y) > 0, \text{ se } x \neq y \\ &d(x,y) = 0, \text{ se } x = y \\ &d(x,y) \leq d(x,z) + d(z,y) \quad \text{(desigualdade triangular).} \end{aligned} \tag{2.1}$$

Há diversas medidas de distância. A distância Euclidiana entre duas observações (itens) p-dimensionais, em que cada item pode ter p informações associadas a ele, sendo denotadas por $x' = [x_1, x_2..., x_p]$ e $y' = [y_1, y_2..., y_p]$, é

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} = \sqrt{(x - y)'(x - y)}. \quad (2.2)$$

Já a distância estatística entre as mesmas duas observações é da forma

$$d(x, y) = \sqrt{(x - y)'A(x - y)}. \quad (2.3)$$

Usualmente, $A = S^{-1}$, sendo S^{-1} é a matriz inversa de S, ou seja, quando multiplicamos as matrizes S e S^{-1}, temos como produto a matriz identidade I_n

$$S \times S^{-1} = I_n,$$

e S é a matriz de variâncias e covariâncias da amostra, dada como

$$\mathbf{S} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \dots & \sigma_n^2 \end{bmatrix},$$

sendo os elementos diagonais da matriz representando as variâncias das variáveis, e os elementos fora da diagonal contendo as covariâncias entre todos os possíveis pares de variáveis. A distância Euclidiana geralmente é a mais utilizada em *clustering*.

Existem outras medidas de distâncias utilizadas no problema de *clustering*. Uma dessas medidas de distância é a métrica de Minkowski

$$d(x, y) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{1/m}. \quad (2.4)$$

Quando $m = 1$, $d(x, y)$ mede a distância "quarteirão" entre dois pontos em p dimensões. Quando $m = 2$, essa métrica coincide com a distância Euclidiana. De forma geral, variar m muda o peso dado a diferenças maiores e menores.

Outras duas medidas populares adicionais de "distância" ou dissimilaridade são dadas pela métrica de Canberra e o coeficiente de Czekanowski. Ambas as medidas são definidas

apenas para variáveis não negativas.

$$\text{Métrica de Canberra} : d(x,y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}. \tag{2.5}$$

$$\text{Coeficiente de Czekanowski} : d(x,y) = 1 - \frac{2\sum_{i=1}^{p} min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}. \tag{2.6}$$

Sempre que possível, é aconselhável usar distâncias "verdadeiras", ou seja, qualquer medida de distância $d(x,y)$ que satisfaz as propriedades de distância de (2.1), para agrupar objetos. No entanto, a maioria dos algoritmos de *clustering* aceitará número de distância atribuídos subjetivamente que podem não satisfazer, por exemplo, a desigualdade triangular.

Quando os objetos não podem ser representados por medições p-dimensionais significativas, os pares de objetos são freqüentemente comparados com base na presença ou ausência de certas características. Objetos semelhantes têm mais características em comum do que objetos diferentes.

A presença ou ausência de uma característica pode ser descrita matematicamente pela introdução de uma *variável binária*, que assume o valor 1 se a característica estiver presente e o valor 0 se a característica estiver ausente. Para $p = 5$ variáveis binárias, por exemplo, as "pontuações" para dois objetos i e k podem ser organizadas da seguinte forma:

	Variável				
	1	2	3	4	5
Objeto i	1	0	0	1	1
Objeto k	1	1	0	1	0

Neste caso, existem duas correspondências 1-1, uma correspondência 0-0 e duas incompatibilidades.

Seja x_{ij} a pontuação (1 ou 0) da j-ésima variável binária no i-ésimo item e x_{kj} a pontuação (novamente, 1 ou 0) da j-ésima variável no k-ésimo objeto, onde $j = 1, 2, ..., p$. Consequentemente,

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{se } x_{ij} = x_{kj} = 1 \text{ ou } x_{ij} = x_{kj} = 0 \\ 1 & \text{se } x_{ij} \neq x_{kj} \end{cases}, \tag{2.7}$$

e a distância Euclidiana quadrada, $\sum_{i=1}^{p}(x_{ij} - x_{kj})^2$, fornecerá, nesse caso, uma contagem do número de incompatibilidades. Uma grande distância corresponderá a muitas incompatibilidades, ou seja, a presença de muitos objetos diferentes. Aplicando-se a distância Euclidiana no exemplo dado anteriormente, tem-se que a distância dos itens do exemplo possui o seguinte valor:

$$\begin{aligned}\sum_{i=1}^{5}(x_{ij} - x_{kj})^2 &= (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 \\ &= 2\end{aligned}$$

Embora uma distância baseada em (2.7) possa ser usada para medir a similaridade, é diferente de pesar as partidas 1-1 e 0-0 igualmente. Em alguns casos, uma correspondência 1-1 é uma indicação mais forte de similaridade do que uma correspondência 0-0.

Por exemplo, ao agrupar pessoas, quando duas pessoas sabem ler grego antigo tem-se uma evidência mais forte de similaridade do que a ausência dessa habilidade. Assim, pode ser razoável descontar as correspondências 0-0 ou mesmo desconsiderá-las completamente. Para permitir o tratamento diferencial das correspondências 1-1 e 0-0, são sugeridos diversos esquemas para definir coeficientes de similaridade.

Para introduzir esses esquemas, vamos organizar as frequências de correspondências e incompatibilidades para os objetos i e k na forma de uma tabela de contingência:

		Objeto k: 1	Objeto k: 0	Total
Objeto i	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$p = a+b+c+d$

Nesta tabela, a representa a frequência de correspondências 1-1 , b é a frequência de correspondências 1-0 e assim por diante. No exemplo citado anteriormente, tem-se que $a = 2$ e $b = c = d = 1$.

A Tabela 1 lista os coeficientes de similaridade comuns definidos em termos das frequências da tabela anterior, e um breve raciocínio segue cada definição.

Os três primeiros coeficientes na tabela são monotonicamente relacionados, ou seja, apresentam sempre o mesmo padrão. Em uma situação, suponhamos que o coeficiente 1 seja calculado para duas tabelas de contingência, Tabela I e Tabela II. Logo, se

Tabela 1: Coeficientes de semelhança para itens de *clustering* (p variáveis binárias)

Coeficiente	Raciocínio (justificativa)
1. $\frac{a+d}{p}$	Pesos iguais para correspondências 1-1 e 0-0
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Peso duplo para correspondências 1-1 e 0-0
3. $\frac{a+d}{a+d+2(b+c)}$	Peso duplo para pares incompatíveis
4. $\frac{a}{p}$	Nenhuma correspondência 0-0 no numerador
5. $\frac{a}{a+b+c}$	Nenhuma correspondência 0-0 no numerador ou denominador(as correspondências 0-0 são tratadas como irrelevantes)
6. $\frac{2a}{2a+b+c}$	Nenhuma correspondência 0-0 no numerador ou denominador. Peso duplo para correspondências 1-1
7. $\frac{a}{a+2(b+c)}$	Nenhuma correspondência 0-0 no numerador ou denominador. Peso duplo para pares incompatíveis
8. $\frac{a}{b+c}$	Proporção de correspondências para incompatibilidades com correspondências 0-0 excluídas

$(a_I + d_I)/p \geq (a_{II} + d_{II})/p$, também teremos $2(a_I + d_I)/[2(a_I + d_I) + b_I + c_I] \geq 2(a_{II} + d_{II})/[2(a_{II} + d_{II}) + b_{II} + c_{II})$, e o coeficiente 3 será pelo menos tão grande para a Tabela I quanto para a Tabela II. Os coeficientes 5, 6 e 7 também retêm suas ordens relativas.

A monotonicidade é importante, porque alguns procedimentos de *clustering* não são afetados caso a definição de similaridade seja alterada de alguma forma que deixe as ordenações relativas de similaridades inalteradas. Os procedimentos hierárquicos *single linkage* (ligação simples) e *complete linkage* (ligação completa), que são discutidos na Seção 2.4, não são afetados. Com esses métodos, qualquer escolha dos coeficientes 1, 2 e 3 na Tabela 1 produzirá os mesmos agrupamentos. E da mesma forma, qualquer escolha dos coeficientes 5, 6 e 7 resultará em agrupamentos idênticos.

Exemplo 2.1 ***Calcular os valores de um coeficiente de similaridade***

Suponha que há 5 indivíduos possuindo as seguites características:

	Altura (em polegadas)	Peso (em libras)	Cor dos olhos	Cor do cabelo	Lateralidade	Sexo
Indivíduo 1	68	140	verde	loiro	destro	Feminino
Indivíduo 2	73	185	castanho	castanho	destro	Masculino
Indivíduo 3	67	165	azul	loiro	destro	Masculino
Indivíduo 4	64	120	castanho	castanho	destro	Feminino
Indivíduo 5	76	210	castanho	castanho	canhoto	Masculino

As variáveis contínuas "Altura"e "Peso"são transformadas em variáveis binárias.

Seja X_{ij} a variável aleatória do indivíduo i de forma que:

$$\mathrm{X}_{i1} = \begin{cases} 1 & \text{altura} \geq 72 \text{ polegadas} \\ 0 & \text{altura} < 72 \text{ polegadas} \end{cases} \qquad \mathrm{X}_{i4} = \begin{cases} 1 & \text{é cabelo loiro} \\ 0 & \text{não é cabelo loiro} \end{cases}$$

$$\mathrm{X}_{i2} = \begin{cases} 1 & \text{peso} \geq 150 \text{ libras} \\ 0 & \text{peso} < 150 \text{ libras} \end{cases} \qquad \mathrm{X}_{i5} = \begin{cases} 1 & \text{mão direita (destro)} \\ 0 & \text{mão esquerda (canhoto)} \end{cases}$$

$$\mathrm{X}_{i3} = \begin{cases} 1 & \text{olhos castanhos} \\ 0 & \text{outros} \end{cases} \qquad \mathrm{X}_{i6} = \begin{cases} 1 & \text{feminino} \\ 0 & \text{masculino} \end{cases}$$

e então tem-se que essas variáveis assumirão os seguintes valores:

		X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}
Indivíduo	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

O número de correspondências e incompatibilidades são indicados na matriz bidimensional

		Indivíduo 2		
		1	0	Total
Indivíduo 1	1	1	2	3
	0	3	0	3
	Total	4	2	6

Empregando coeficiente de similaridade 1, que dá pesos iguais às correspondências, calculamos

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$$

Continuando com coeficiente de similaridade 1, vamos calcular o número de similaridades restantes para pares de indivíduos. Estes são exibidos na matriz simétrica 5X5

$$\begin{array}{cc} & \text{Indivíduo} \\ & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} \\ 1 \\ 2 \\ \text{Indivíduo} \quad 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 1 & & & & \\ 1/6 & 1 & 0 & & \\ 4/6 & 3/6 & 1 & & \\ 4/6 & 3/6 & 2/6 & 1 & \\ 0 & \textcircled{5/6} & 2/6 & 2/6 & 1 \end{bmatrix} \end{array}$$

Baseado nas magnitudes do coeficiente de similaridade, podemos concluir que os indivíduos 2 e 5 são os mais semelhantes e os indivíduos 1 e 5 são os menos semelhantes. Outros pares estão entre esses extremos. Dividi-se os indivíduos em dois subgrupos relativamente homogêneos com base nos números de similaridade, poderia formar os subgrupos (1 3 4) e (2 5).

Observe que $X_3 = 0$ *implica uma ausência de olhos castanhos, de modo que duas pessoas, uma com olhos azuis e outra com olhos verdes, resultarão em uma correspondência 0-0. Conseqüentemente, pode ser inapropriado usar o coeficiente de similaridade 1, 2 ou 3 porque esses coeficientes dão os mesmos pesos para correspondências 1-1 e 0-0.*

Descrevemos a construção de distâncias e semelhanças. Sempre é possível construir semelhanças a partir de distâncias. Por exemplo, podemos definir

$$\widetilde{s_{ik}} = \frac{1}{1 + d_{ik}}, \tag{2.8}$$

onde $0 < \widetilde{s_{ik}} \leq 1$ é a similaridade entre os objetos i e k, e d_{ik} é a distância correspondente.

No entanto, as distâncias que devem satisfazer as propriedades em (2.1) nem sempre podem ser construídas a partir de semelhanças, pois isso só pode ser feito se a matriz de semelhanças for definida não negativa, ou seja, a matriz de semelhanças apresenta todos os elementos maiores ou iguais a zero:

$$x_{ij} \geq 0 \qquad \forall i, j.$$

Com a condição definida não negativa, e com a similaridade máxima escalonada de modo que $s_{ii} = 1$,

$$d_{ik} = \sqrt{2(1 - \widetilde{s_{ik}})}, \tag{2.9}$$

tem as propriedades de uma distância.

2.3.2 Similaridades e medidas de associação para pares de variáveis

Em algumas aplicações, as medidas de similaridade são para as variáveis, e não para objetos, e essas variáveis devem ser agrupadas. Medidas de similaridade para variáveis geralmente assumem a forma de coeficientes de correlação de amostra. Além disso, em algumas aplicações de *clustering*, correlações negativas são substituídas por seus valores absolutos.

Quando as variáveis são binárias, os dados podem ser novamente organizados em

tabelas de contingência. Porém, agora são as variáveis que esboçam as categorias, ao invés dos objetos. Para cada par de variáveis, há n objetos categorizados na tabela. Com a codificação usual de 0 e 1, a tabela passa ser a seguinte:

		Variável k		
		1	0	Total
Variável i	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$n=a+b+c+d$

Por exemplo, a variável i é igual a 1 e a variável k é igual a 0 para b dos n objetos.

A fórmula usual de correlação entre 2 covariáveis utilizando o produto de funções dessa variável. aplicada às variáveis binárias na Tabela de contingência acima dá (ver em Exemplo 2.2):

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}. \tag{2.10}$$

Este número pode ser considerado como uma medida de similaridade entre as duas variáveis.

O coeficiente de correlação em (2.10) está relacionado com a estatística Qui-Quadrado $r^2 = \frac{\chi^2}{n}$ para testar a independência de duas variáveis categóricas. Para um n fixo, uma grande similaridade (ou correlação) é consistente com a presença de dependência.

Com a Tabela de contingência, podemos desenvolver medidas de associação ou similaridade exatamente análogas às listadas na Tabela 1. A única mudança necessária é a substituição de n (o número de objetos) por p (o número de variáveis).

Exemplo 2.2 ***Medindo as semelhanças de 11 idiomas***

Os significados das palavras mudam com o curso da história. No entanto, o significado dos números 1,2,3, ... representa uma exceção conspícua. Assim, uma primeira comparação de idiomas pode basear-se apenas nos numerais. Para a comparação, foram utilizados os 10 primeiros números em inglês, polonês, húngaro e em outras oito línguas europeias modernas (somente línguas que usam o alfabeto romano são consideradas, e os acentos, cedilhas, etc..., são omitidos).

As palavras para o número 1 em francês, espanhol e italiano começam com "u". Chamamos as palavras do mesmo número em duas línguas de concordantes se tiverem a mesma primeira letra e discordantes se não tiverem. Da Tabela 2, vemos as concordâncias

Tabela 2: Primeiras letras concordantes para números em 11 idiomas

	Ing	Nor	Din	Hol	Ale	Fra	Esp	Ita	Pol	Hun	Fin
Ing	10										
Nor	8	10									
Din	8	9	10								
Hol	3	5	4	10							
Ale	4	6	5	5	10						
Fra	4	4	4	1	3	10					
Esp	4	4	5	1	3	8	10				
Ita	4	4	5	1	3	9	9	10			
Pol	3	3	4	0	2	5	7	6	10		
Hun	1	2	2	2	1	0	0	0	0	10	
Fin	1	1	1	1	1	1	1	1	1	2	10

(frequências das primeiras iniciais correspondentes) para os números de 1 a 10. Vemos que o inglês e o norueguês têm a mesma primeira letra para 8 dos 10 pares de palavras. As demais frequências foram calculadas da mesma maneira. Ou seja, inglês, norueguês, dinamarquês, holandês e alemão parecem formar um grupo. Francês, espanhol, italiano e polonês podem ser agrupados, enquanto húngaro e finlandês parecem estar sozinhos.

O exame superficial da grafia dos numerais da Tabela 2 sugere que os primeiros cinco idiomas (inglês, norueguês, dinamarquês, holandês e alemão) são muito semelhantes. Francês, espanhol e italiano estão ainda mais de acordo. Húngaro e finlandês parecem se manter por si próprios, e polonês tem algumas das características das línguas em cada um dos subgrupos maiores.

Nos exemplos até agora, foram utilizadas impressões visuais de similaridade ou medidas de distância para formar grupos. Existem esquemas menos subjetivos para a criação de clusters.

2.4 Métodos de *clustering* hierárquico

Raramente há condições de examinar todas as possibilidades de agrupamento, mesmo com os melhores e mais rápidos computadores, pois isso demanda muito tempo. Por conta dessa situação, surgiu uma ampla variedade de algoritmos de *clustering* que encontram *clusters* "razoáveis" sem ter que avaliar todas as configurações.

As técnicas de *clustering* hierárquico prosseguem por uma série de fusões sucessivas ou uma série de divisões sucessivas. Os métodos hierárquicos aglomerativos começam com os objetos individuais. Portanto, existem inicialmente tantos *clusters* quanto objetos.

Os objetos mais semelhantes primeiramente são agrupados, e esses grupos iniciais são mesclados de acordo com suas semelhanças. Eventualmente, conforme a similaridade diminui, todos os subgrupos são fundidos em um único *cluster*.

Os métodos hierárquicos divisivos funcionam na direção oposta. Um único grupo inicial de objetos é dividido em dois subgrupos, de modo que os objetos de um subgrupo estão "longe" dos objetos do outro. Esses subgrupos são então divididos em subgrupos diferentes; o processo continua até que haja tantos subgrupos quanto objetos - isto é, até que cada objeto forme um grupo.

Os resultados dos métodos aglomerativos e divisivos podem ser exibidos na forma de um diagrama bidimensional, conhecido como dendrograma. O dendrograma ilustra as fusões ou divisões que foram feitas em níveis sucessivos.

Nesta seção, será dado preferência aos procedimentos hierárquicos aglomerativos e, em particular, aos métodos *linkage*.

Métodos *linkage* são adequados para itens de agrupamento, bem como variáveis. Isso não significa que será compatível para todos os procedimentos aglomerativos hierárquicos. É apresentado *single linkage* (distância mínima ou vizinho mais próximo), *complete linkage* (distância máxima ou vizinho mais distante) e *average linkage* (distância média). A fusão de *clusters* sob os três critérios *linkage* é ilustrada esquematicamente na Figura 1.

A partir da Figura 1, vemos que o *single linkage* resulta quando os grupos são fundidos de acordo com a distância entre seus membros mais próximos. O *complete linkage* ocorre quando os grupos são fundidos de acordo com a distância entre seus membros mais distantes. E no *average linkage*, os grupos são fundidos de acordo com a distância média entre os pares de membros nos respectivos conjuntos.

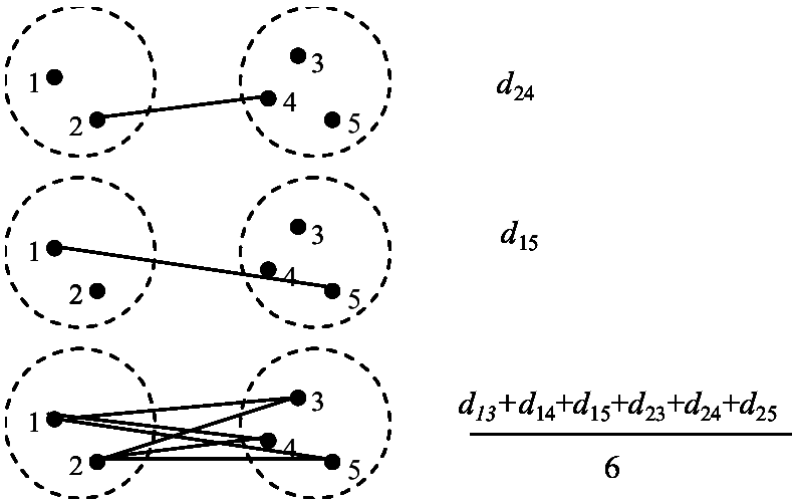


Figura 1: Distância *intercluster* (dissimilaridade) para *single linkage*, *complete linkage* e *average linkage*, respectivamente[6]

A seguir estão as etapas do algoritmo de *clustering* hierárquico aglomerativo para agrupar N objetos (itens ou variáveis):

1. Comece com N *clusters*, cada um contendo uma única entidade e uma matriz simétrica de distâncias (ou semelhanças) $N \times N$: $\mathbf{D} = d_{ik}$.

2. Procure a matriz de distância para o par de *clusters* mais próximo (mais semelhante), e essa distância entre os *clusters* "mais semelhantes" U e V será d_{uv}.

3. Mesclar os *clusters* U e V. Rotule o *cluster* recém-formado (UV), atualize as entradas na matriz de distância (a), excluindo as linhas e colunas correspondentes aos *clusters* U e V, e (b) adicionando uma linha e coluna dando as distâncias entre o *cluster* (UV) e os *clusters* restantes.

4. Repita as etapas 2 e 3 num total de $N - 1$ vezes (Todos os objetos estarão em um único *cluster* após o algoritmo terminar). Registre a identidade dos *clusters* que são mesclados e os níveis (distâncias ou semelhanças) nos quais as fusões ocorrem.

(2.11)

As ideias por trás de qualquer procedimento de *clustering* são provavelmente melhor transmitidas por meio de exemplos, que serão apresentados depois de breves discussões sobre a entrada e os componentes algorítmicos dos métodos *linkage*.

2.4.1 *Single Linkage*

As entradas para um algoritmo *single linkage* podem ser distâncias ou medidas de semelhanças apresentados na seção 2.3 entre pares de objetos. Os grupos são formados a partir de entidades individuais pela fusão de vizinhos mais próximos, onde o termo vizinho mais próximo conota a menor distância ou a maior semelhança.

Inicialmente, devemos encontrar a menor distância em $\mathbf{D} = \{d_{ik}\}$ e mesclar os objetos correspondentes, por exemplo, U e V, para obter o *cluster* (UV). Para a Etapa 3 do algoritmo geral de (2.11), as distâncias entre (UV) e qualquer outro cluster W são calculadas por

$$d_{(UV)W} = min\{d_{UW}, d_{VW}\}. \tag{2.12}$$

Aqui as quantidades d_{UW} e d_{VW} são as distâncias entre os vizinhos mais próximos de *clusters* U e W e *clusters* V e W, respectivamente.

Os resultados do *clustering single linkage* podem ser exibidos graficamente na forma de um dendrograma ou diagrama de árvore. Os ramos na árvore representam os *clusters*. Os ramos vêm juntos (fusão) em nós, cujas posições ao longo de um eixo de distância (ou semelhança) indicam o nível em que as fusões ocorrem.

Exemplo 2.3 ***Clustering usando single linkage***

Para ilustrar o algoritmo single linkage, nós consideramos as distâncias hipotéticas entre pares de 5 objetos do seguinte modo:

$$\boldsymbol{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

Tratando cada objeto como um cluster, começamos o agrupamento fundindo os dois itens mais próximos. Uma vez que

$$\min_{i \neq k}(\{d_{ik}\}) = d_{53} = 2$$

objetos 5 *e* 3 *são mesclados para formar o cluster* $(3, 5)$*. Para implementar o próximo nível de clustering, precisamos das distâncias entre o cluster* $(3, 5)$ *e os objetos restantes,*

1, 2 *e* 4. *As distâncias vizinhas mais próximas são*

$$d_{(3,5)1} = min\{d_{31}, d_{51}\} = min\{3, 11\} = 3$$
$$d_{(3,5)2} = min\{d_{32}, d_{52}\} = min\{7, 10\} = 7$$
$$d_{(3,5)4} = min\{d_{34}, d_{54}\} = min\{9, 8\} = 8$$

Excluindo as linhas e colunas de $\boldsymbol{D}$ *correspondentes aos objetos* 3 *e* 5, *e adicionando uma linha e coluna para o cluster* $(3, 5)$, *vamos obter uma nova matriz de distância*

$$\begin{array}{c} \\ (3,5) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (3,5) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ \textcircled{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

A menor distância entre pares de clusters é agora $d_{(3,5)1} = 3$, *e mesclamos o cluster* (1) *com o cluster* $(3, 5)$ *para obter o próximo cluster,* $(1, 3, 5)$. *Calculando*

$$d_{(1,3,5)2} = min\{d_{(3,5)2}, d_{12}\} = min\{7, 9\} = 7$$
$$d_{(1,3,5)4} = min\{d_{(3,5)4}, d_{14}\} = min\{8, 6\} = 6$$

descobrimos que a matriz de distância para o próximo nível de clustering é

$$\begin{array}{c} \\ (1,3,5) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (1,3,5) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

A distância mínima do vizinho mais próximo entre pares de clusters é $d_{42} = 5$, *e mesclamos os objetos* 4 *e* 2 *para obter o cluster* $(2, 4)$.

Neste ponto, temos dois clusters distintos, $(1, 3, 5)$ *e* $(2, 4)$. *A distância vizinha mais próxima é*

$$d_{(1,3,5)(2,4)} = min\{d_{(1,3,5)2}, d_{(1,3,5)4}\} = min\{7, 6\} = 6$$

A matriz de distância final torna-se

$$\begin{array}{c} \\ (1,3,5) \\ (2,4) \end{array} \begin{array}{c} \begin{array}{cc} (1,3,5) & (2,4) \end{array} \\ \begin{bmatrix} 0 & \\ \textcircled{6} & 0 \end{bmatrix} \end{array}$$

Consequentemente, os clusters $(1, 3, 5)$ *e* $(2, 4)$ *são mesclados para formar um único cluster*

de todos os cinco objetos, $(1, 2, 3, 4, 5)$*, quando a distância do vizinho mais próximo atinge 6.*

O dendrograma retratando o clustering hierárquico recém-concluído é mostrado na Figura 2. Os agrupamentos e os níveis de distância em que ocorrem são claramente ilustrados pelo dendrograma.

Em aplicações típicas de *clustering* hierárquico, os resultados intermediários, no qual os objetos são classificados em um número moderado de *clusters*, são os de interesse principal.

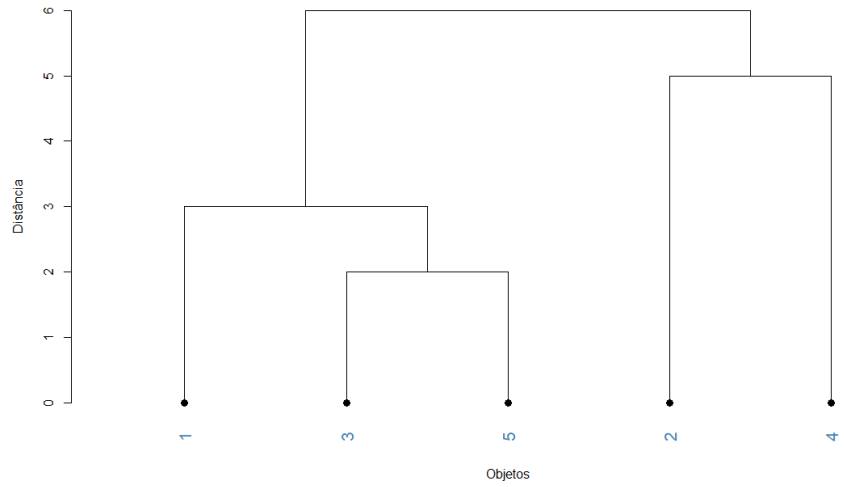


Figura 2: Dendrograma single linkage para distâncias entre 5 objetos

Exemplo 2.4 ***Clustering single linkage de 11 idiomas***

Considere a matriz de concordâncias na Tabela 2 representando a proximidade entre os números 1-10 em 11 idiomas. Para desenvolver uma matriz de distâncias, subtraímos as concordâncias da figura de concordância perfeita de 10 que cada língua tem consigo

mesma. As atribuições subsequentes de distâncias são

	Ing	*Nor*	*Din*	*Hol*	*Ale*	*Fra*	*Esp*	*Ita*	*Pol*	*Hun*	*Fin*
Ing	0										
Nor	2	0									
Din	2	(1)	0								
Hol	7	5	6	0							
Ale	6	4	5	5	0						
Fra	6	6	6	9	7	0					
Esp	6	6	5	9	7	2	0				
Ita	6	6	5	9	7	(1)	(1)	0			
Pol	7	7	6	10	8	5	3	4	0		
Hun	9	8	8	8	9	10	10	10	10	0	
Fin	9	9	9	9	9	9	9	9	9	9	0

Primeiro procuramos a distância mínima entre pares de idiomas (clusters). A distância mínima, 1, ocorre entre dinamarquês e norueguês, italiano e francês e italiano e espanhol. Numerando os idiomas na ordem em que aparecem no topo do matriz, temos

$$d_{32} = 1; \quad d_{86} = 1; \quad d_{87} = 1$$

Como $d_{76} = 2$*, podemos mesclar apenas os clusters* 8 *e* 6*, ou os clusters* 8 *e* 7*. Não podemos mesclar os clusters* 6*,* 7 *e* 8 *no nível 1. Escolhemos primeiro mesclar* 6 *e* 8 *e, em seguida, atualizar a matriz de distância e fundir* 2 *e* 3 *para obter os clusters* $(6, 8)$ *e* $(2, 3)$*. Cálculos computadorizados subsequentes produzem o dendrograma da Figura 3.*

A partir do dendrograma, vemos que o norueguês e o dinamarquês, e também o francês e o italiano, se agrupam no nível de distância mínima (semelhança máxima). Quando a distância permitida é aumentada, o inglês é adicionado ao grupo norueguês-dinamarquês, e espanhol se funde com o grupo francês-italiano. Observe que o húngaro e o finlandês são mais semelhantes entre si do que com os outros grupos de línguas. No entanto, esses dois clusters (idiomas) não se fundem até a distância entre os vizinhos mais próximos tenha aumentado substancialmente. Finalmente, todos os clusters de idiomas são fundidos em um único cluster na maior distância do vizinho mais próximo, 9*.*

Como o *single linkage* une os *clusters* pelo *link* mais curto entre eles, a técnica não pode discernir *clusters* mal separados, como podemos observar na Figura 4(a). Por outro lado, o *single linkage* é um dos poucos métodos de *clustering* que pode delinear *clusters*

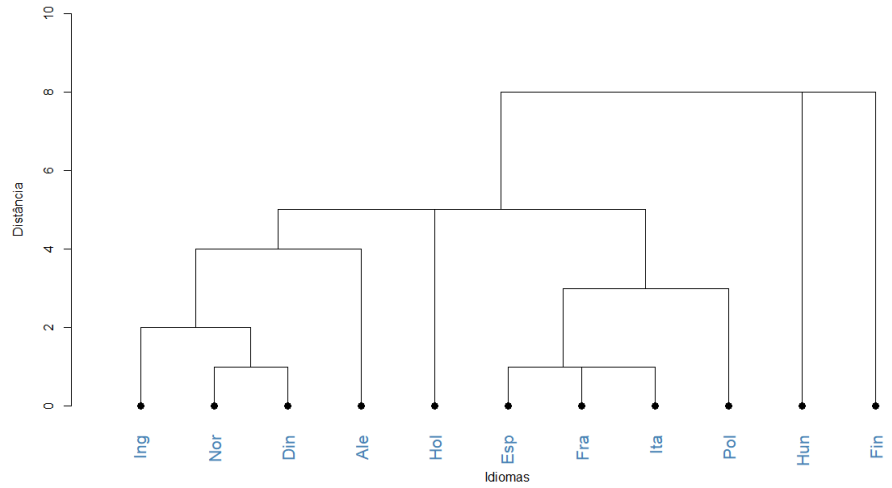
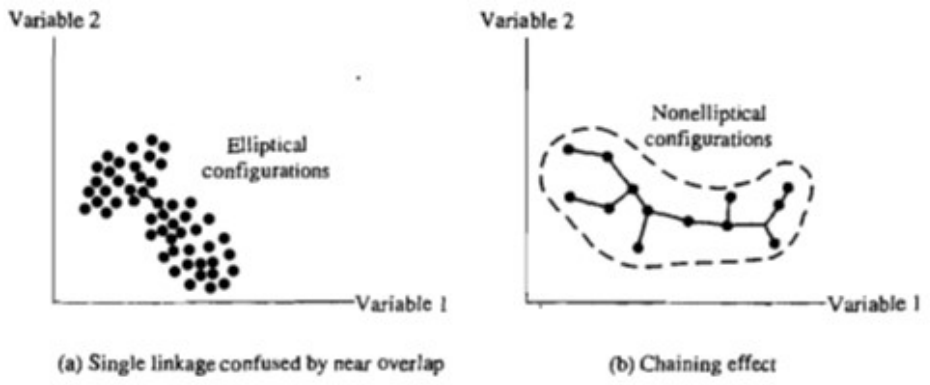


Figura 3: Dendrograma single linkage para distâncias entre números em 11 idiomas

não elipsoidais. A tendência do *single linkage* de separar longas "cordas" semelhantes a fios é conhecida como encadeamento, como na Figura 4(b). O encadeamento pode ser enganoso se os itens nas extremidades opostas da cadeia forem, de fato, muito diferentes.



Fonte: Johnson, Richard A. (2007)

Figura 4: *Clusters single linkage*

Os *clusters* formados pelo método *single linkage* não serão alterados pela sinalização de distância (similaridade) que fornece as mesmas ordenações relativas das distâncias iniciais (similaridades). Em particular, qualquer um de um conjunto de coeficientes de similaridade da Tabela 1 que são monotônicos entre si produzirá o mesmo *clustering*.

2.4.2 *Complete Linkage*

O *clustering complete linkage* procede da mesma forma que *clusters single linkage*, com uma importante exceção: em cada estágio, a distância (similaridade) entre os *clusters* é determinada pela distância (similaridade) entre os dois elementos, um de cada *cluster*, que estão "mais distantes". Assim, o *complete linkage* garante que todos os itens em um *cluster* estejam dentro de alguma similaridade de distância máxima entre si.

O algoritmo aglomerativo geral novamente começa encontrando a entrada mínima em $\mathbf{D} = \{d_{ik}\}$ e mesclando os objetos correspondentes, como U e V, para obter o *cluster* (UV). Para a Etapa 3 do algoritmo geral em (2.11), as distâncias entre (UV) e qualquer outro cluster W são calculadas por

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\}. \tag{2.13}$$

Aqui d_{UW} e d_{VW} são as distâncias entre o mais distantes membros de *clusters* U e W, e *clusters* V e W, respectivamente.

Exemplo 2.5 ***Clustering usando complete linkage***

Retornando à matriz de distância introduzida no Exemplo 2.3:

$$\boldsymbol{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

No primeiro palco, os objetos 3 *e* 5 *são mesclados, uma vez que são mais semelhantes. Isso dá o cluster* $(3, 5)$*. No estágio 2, calculamos*

$$\begin{aligned} d_{(3,5)1} &= max\{d_{31}, d_{51}\} = max\{3, 11\} = 11 \\ d_{(3,5)2} &= max\{d_{32}, d_{52}\} = max\{7, 10\} = 10 \\ d_{(3,5)4} &= max\{d_{34}, d_{54}\} = max\{9, 8\} = 9 \end{aligned}$$

e a matriz de distância modificada torna-se

$$\begin{array}{c} \\ (3,5) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (3,5) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

A próxima fusão ocorre entre os grupos mais semelhantes, 2 *e* 4*, para dar o cluster* $(2,4)$*. No estágio 3, temos*

$$d_{(2,4)(3,5)} = max\{d_{2(3,5)}, d_{4(3,5)}\} = max\{10, 9\} = 10$$
$$d_{(2,4)1} = max\{d_{21}, d_{41}\} = 9$$

e a matriz de distância

$$\begin{array}{c} \\ (3,5) \\ (2,4) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (3,5) & (2,4) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & \textcircled{9} & 0 \end{bmatrix} \end{array}$$

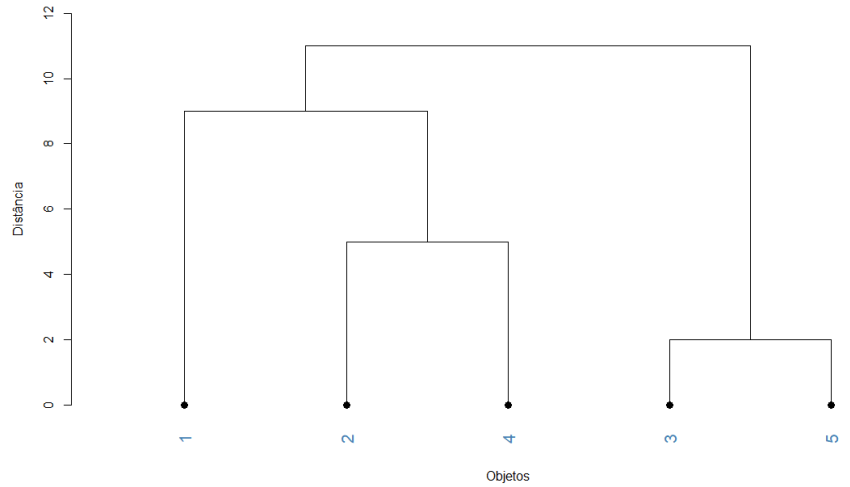


Figura 5: Dendrograma complete linkage para distâncias entre 5 objetos

A próxima fusão produz o cluster $(1,2,4)$*. No estágio final, os grupos* $(3,5)$ *e* $(1,2,4)$ *são mesclados como o único cluster* $(1,2,3,4,5)$ *no nível*

$$d_{(1,2,4)(3,5)} = max\{d_{1(3,5)}, d_{(2,4)(3,5)}\} = max\{11, 10\} = 11$$

O dendrograma é dado na Figura 5.

Comparando as Figuras 2 e 5, vemos que os dendrogramas para *single linkage* e *complete linkage* diferem na alocação do objeto 1 para grupos anteriores.

Exemplo 2.6 ***Clustering complete linkage de 11 idiomas***

No Exemplo 2.2, apresentamos uma matriz de distância para números em 11 idiomas. O algoritmo de Clustering complete linkage aplicado a esta matriz de distância produz o dendrograma mostrado na Figura 6.

Comparando as Figuras 6 e 3, vemos que ambos os métodos hierárquicos produzem os grupos de idiomas inglês-norueguês-dinamarquês e francês-italiano-espanhol. O polonês é mesclado com o francês-italiano-espanhol em um nível intermediário. Além disso, ambos os métodos mesclam o húngaro e o finlandês apenas no penúltimo estágio.

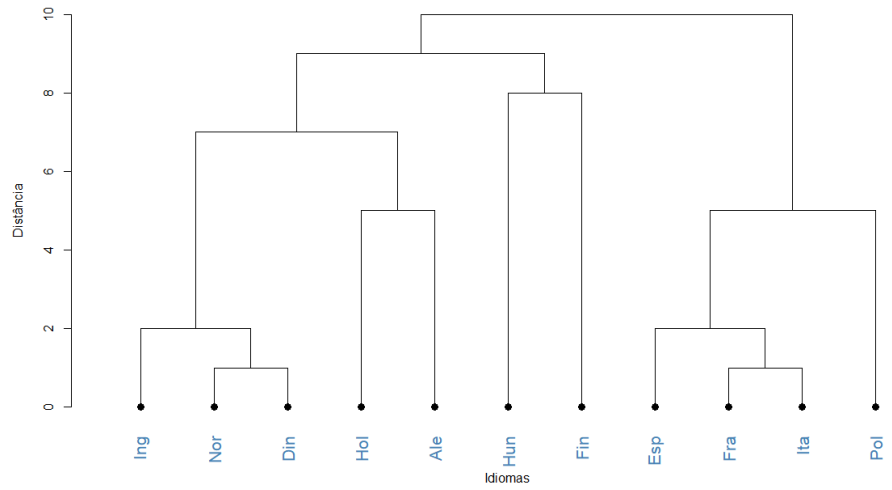


Figura 6: Dendrograma complete linkage para distâncias entre números em 11 idiomas

Entretando, os dois métodos lidam de forma diferente com o alemão e o holandês. Single linkage funde alemão e holandês a uma distância intermediária, e essas duas línguas permanecem um cluster até a fusão final. Já complete linkage mescla alemão com o grupo inglês-norueguês-dinamarquês num nível intermediário. Holandês permanece um cluster por si só até que seja mesclado com os grupos inglês-norueguês-dinamarquês-alemão e francês-italiano-espanhol-polonês em um nível de distância mais alto. A última mesclagem complete linkage envolve dois clusters. E no single linkage, a última mesclagem envolve três clusters.

2.4.3 *Average Linkage*

O average linkage trata a distância entre dois *clusters* como a distância média entre todos os pares de itens onde um membro de um par pertence a cada *cluster*.

Novamente, a entrada para o algoritmo de *average linkage* pode ser distâncias ou semelhanças, e o método pode ser usado para agrupar objetos ou variáveis. O algoritmo de *average linkage* procede da maneira do algoritmo geral de (2.11). Começamos pesquisando a matriz de distância $\mathbf{D} = \{d_{ik}\}$ para encontrar os objetos mais próximos (mais semelhantes), por exemplo, U e V. Esses objetos são mesclados para formar o *cluster* (UV). Para a Etapa 3 do algoritmo aglomerativo geral, as distâncias entre (UV) e o outro *cluster* W são determinadas por

$$d_{(UV)W} = \frac{\sum_{i \in I} \sum_{j \in J} d_{ik}}{N_{(UV)} N_W}, \tag{2.14}$$

onde d_{ik} é a distância entre o objeto i no *cluster* (UV) e o objeto k no *cluster* W, e $N_{(UV)}$ e N_W são o número de itens em *clusters* (UV) e W, respectivamente.

Exemplo 2.7 ***Clustering average linkage de 11 idiomas***

O algoritmo average linkage também foi aplicado para as "distâncias" entre 11 idiomas dado no Exemplo 2.4. O dendrograma resultante é exibido na Figura 7.

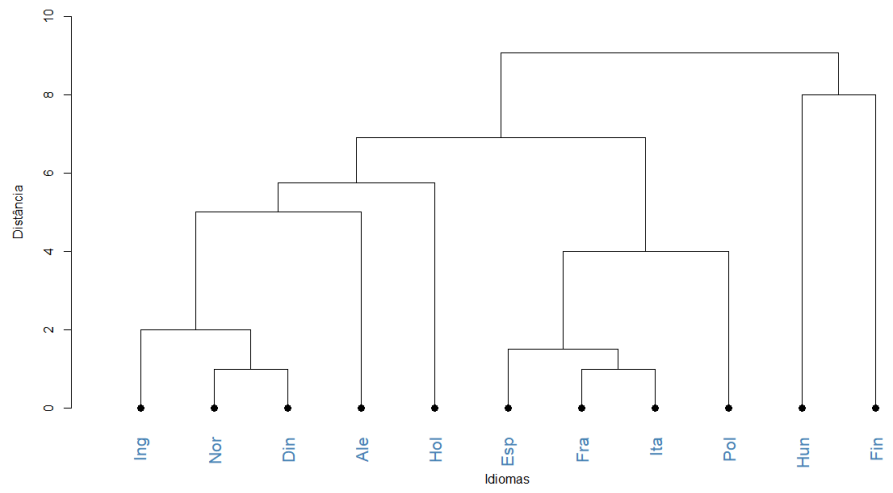


Figura 7: Dendrograma average linkage para distâncias entre números em 11 idiomas

Uma comparação do dendrograma na Figura 7 com o dendrograma de single linkage correspondente a Figura 3 e o dendrograma de complete linkage (Figura 6) indica que a

average linkage produz uma configuração muito parecida com a configuração do complete linkage. No entanto, como a distância é definida de forma diferente para cada caso, não é surpreendente que as fusões ocorram em níveis diferentes.

Exemplo 2.8 ***Clustering average linkage de utilidades públicas***

Um algoritmo average linkage aplicado às distâncias euclidianas entre 22 serviços públicos, onde pode ser visto na matriz de distâncias, produziu o dendrograma da Figura 8.

Matriz das distâncias entre 22 serviços públicos

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0																					
2	3.1	0																				
3	3.8	4.92	0																			
4	2.46	2.16	4.11	0																		
5	4.12	3.85	4.47	4.13	0																	
6	3.61	4.22	2.99	3.2	4.6	0																
7	3.9	3.45	4.22	3.97	4.6	3.35	0															
8	2.74	3.89	4.99	3.69	5.16	4.91	4.36	0														
9	3.25	3.96	2.75	3.75	4.49	3.73	2.8	3.59	0													
10	3.1	2.71	3.93	1.49	4.05	3.83	4.51	3.67	3.57	0												
11	3.49	4.79	5.9	4.86	6.46	6	6	3.46	5.18	5.08	0											
12	3.22	2.43	4.03	3.5	3.6	3.74	1.66	4.06	2.74	3.94	5.21	0										
13	3.96	3.43	4.39	2.58	4.76	4.55	5.01	4.14	3.66	1.41	5.31	4.5	0									
14	2.11	4.32	2.74	3.23	4.82	3.47	4.91	4.34	3.82	3.61	4.32	4.34	4.39	0								
15	2.59	2.5	5.16	3.19	4.26	4.07	2.93	3.85	4.11	4.26	4.74	2.33	5.1	4.24	0							
16	4.03	4.84	5.26	4.97	5.82	5.84	5.04	2.2	3.63	4.53	3.43	4.62	4.41	5.17	5.18	0						
17	4.4	3.62	6.36	4.89	5.63	6.1	4.58	5.43	4.9	5.48	4.75	3.5	5.61	5.56	3.4	5.56	0					
18	1.88	2.9	2.72	2.65	4.34	2.85	2.95	3.24	2.43	3.07	3.95	2.45	3.78	2.3	3	3.97	4.43	0				
19	2.41	4.63	3.18	3.46	5.13	2.58	4.52	4.11	4.11	4.13	4.52	4.41	5.01	1.88	4.03	5.23	6.09	2.47	0			
20	3.17	3	3.73	1.82	4.39	2.91	3.54	4.09	2.95	2.05	5.35	3.43	2.23	3.74	3.78	4.82	4.87	2.92	3.9	0		
21	3.45	2.32	5.09	3.88	3.64	4.63	2.68	3.98	3.74	4.36	4.88	1.38	4.94	4.93	2.1	4.57	3.1	3.19	4.97	4.15	0	
22	2.51	2.42	4.11	2.58	3.77	4.03	4	3.24	3.21	2.56	3.44	3	2.74	3.51	3.35	3.46	3.63	2.55	3.97	2.62	3.01	0

As variáveis utilizadas para calcular a matriz de distância estão na Tabela 3, onde descreve o nome de cada empresa de acordo com o seu código na matriz de distâncias:

Tabela 3: Empresas de serviços públicos

Código da empresa	**Nome da empresa de serviços públicos**
1	Arizona Public Service
2	Boston Edison Company
3	Central Louisiana Eletric Company
4	Commonwealth Edison Company
5	Consolidated Edison Company (N.Y.)
6	Florida Power & Light Company
7	Hawaiian Eletric Company
8	Idaho Power Company
9	Kentucky Utilities Company
10	Madison Gas & Eletric Company
11	Nevada Power Company
12	New England Eletric Company
13	Northern States Power Company
14	Oklahoma Gas & Eletric Company
15	Pacific Gas & Eletric Company
16	Puget Sound Power & Light Company
17	San Diego Gas & Eletric Company
18	The Southern Company
19	Texas Utilities Company
20	Wisconsin Eletric Power Company
21	United Illuminating Company
22	Virginia Eletric & Power Company

Concentrando-se nos clusters intermediários, vemos que as empresas de serviços tendem a se agrupar de acordo com a localização geográfica. Por exemplo, um cluster intermediário contém as empresas 1 *(Arizona Public Service),* 18 *(The Southern Company, principalmente Geórgia e Alabama),* 19 *(Texas Utilities Company) e* 14 *(Oklahoma Gas and Electric Company), todas elas localizadas em estados no sul dos Estados Unidos. Existem algumas exceções. O cluster* $(7, 12, 21, 15, 2)$ *contém empresas no litoral leste e no extremo oeste. Por outro lado, todas as empresas estão localizadas perto da costa. Note que a Consolidated Edison Company de Nova York e a San Diego Gas and Electric Company permanecem sozinhas até os estágios finais de fusão.*

Talvez não seja surpreendente que as empresas de serviços públicos com localizações (ou tipos de localizações) semelhantes se agrupem. Seria de se esperar que empresas regulamentadas na mesma área usassem, basicamente, o mesmo tipo de combustível(s) para usinas de energia e mercados comuns. Conseqüentemente, os tipos de geração, custos,

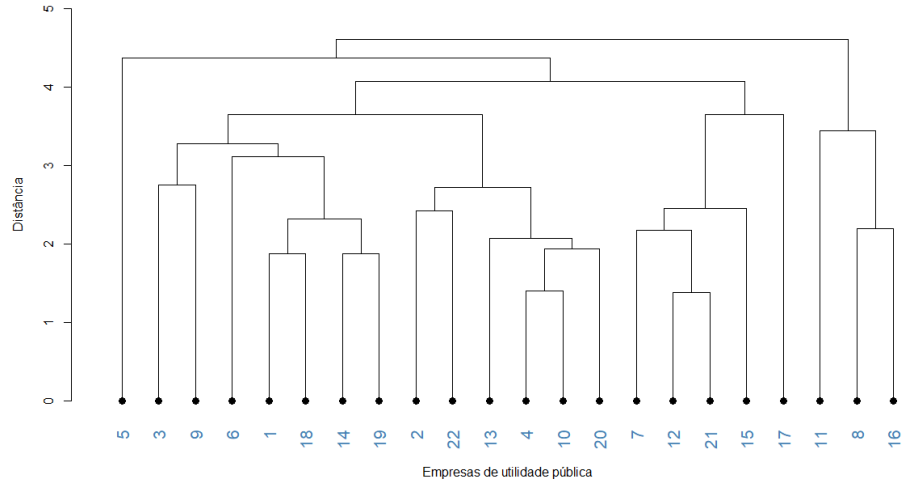


Figura 8: Dendrograma average linkage para distâncias entre 22 empresas de serviços públicos

taxas de crescimento e assim por diante devem ser relativamente homogêneos entre essas empresas. Isto aparentemente se reflete no clustering hierárquico.

Para *clustering average linkage*, mudanças na atribuição de distâncias (semelhanças) podem afetar o arranjo da configuração final dos *clusters*, mesmo que as mudanças preservem ordenações relativas.

2.4.4 Método de *clustering* hierárquico de Ward

Ward considerou procedimentos de *clustering* hierárquico com base na minimização da "perda de informação"ao juntar dois grupos. Este método é geralmente implementado com a perda de informação considerada um aumento no critério de soma de erros quadráticos, SEQ. Em primeiro lugar, para um determinado *cluster* k, deixe SEQ_k ser a soma dos desvios quadrados de cada item no *cluster* da média do *cluster* (centróide). Se houver atualmente K *clusters*, defina $SEQ = SEQ_1 + SEQ_2 + ... + SEQ_K$. Em cada etapa na análise, a união de todos os pares possíveis de *clusters* é considerada, e os dois *clusters* cuja combinação resulta no menor aumento no SEQ(perda mínima de informação) são unidos. Inicialmente, cada *cluster* consiste em um único item e, se houver N itens, $SEQ_k = 0, k = 1, 2.., N$, então $SEQ = 0$. No outro extremo, quando todos os *clusters* são combinados em um único grupo de N itens, o valor de SEQ é dado por

$$SEQ = \sum_{j=1}^{N} (x_j - \bar{x})'.(x_j - \bar{x}),$$

onde x_j é a medição multivariada associada ao j-ésimo item e $\bar{x}$ é a média de todos os itens.

Os resultados do método de Ward podem ser exibidos como um dendrograma. O eixo vertical fornece os valores de SEQ nos quais as fusões ocorrem.

O método de Ward é baseado na noção de que se espera que os grupos de observações multivariadas tenham uma forma aproximadamente elíptica. Isto é, um elemento é alocado aum determinado grupo, de forma que minimize a homogeneidade dentro dos grupos, ou seja, minimiza a soma dos quadrados dos erros dentro dos grupos. É um precursor hierárquico para métodos de *clustering* não hierárquicos que otimizam alguns critérios para dividir dados em um determinado número de grupos elípticos. Discuti-se procedimentos de *clustering* não hierárquicos na próxima seção.

2.4.5 Comentários finais - Procedimentos hierárquicos

Existem muitos procedimentos de *clustering* hierárquico aglomerativo além do *single linkage*, *complete linkage* e *average linkage*. Porém, todos os procedimentos aglomerativos seguem o algoritmo básico de (2.11).

Como acontece com a maioria dos métodos de *clustering*, as fontes de erro e variação não são formalmente consideradas nos procedimentos hierárquicos. Isso significa que um método de *clustering* será sensível a "outliers" ou a "pontos de ruído". No *clustering* hierárquico, não há provisão para uma realocação de objetos que podem ter sido agrupados "incorretamente"em um estágio inicial. Conseqüentemente, a configuração final de *clusters* deve sempre ser examinada cuidadosamente para ver se é sensato.

A estabilidade de uma solução hierárquica pode, às vezes, ser verificada aplicando o algoritmo de *clustering* antes e depois de pequenos erros (perturbações) terem sido adicionados às unidades de dados. Se os grupos forem muito distintos, os agrupamentos antes da perturbação e após a perturbação devem concordar.

Valores comuns (laços) na matriz de semelhança ou distância podem produzir várias soluções para um problema de *clustering* hierárquico. Ou seja, os dendrogramas correspondentes a diferentes tratamentos das semelhanças (distâncias) amarradas podem ser diferentes, particularmente nos níveis mais baixos. Este não é um problema inerente a qualquer método; ao invés disso, várias soluções ocorrem para certos tipos de dados. Soluções múltiplas não são necessariamente ruins, mas o usuário precisa saber de sua existência para que os agrupamentos (dendrogramas) possam ser interpretados

corretamente e diferentes agrupamentos (dendrogramas) comparados para avaliar sua sobreposição.

2.5 Métodos de *clustering* não hierárquicos

As técnicas de *clustering* não hierárquicos são projetadas para agrupar itens, ao invés de variáveis, para uma coleção de *clusters* K. O número de *clusters* K, pode ser especificado antecipadamente ou determinado como parte do procedimento de *clustering*. Como uma matriz de distâncias (semelhanças) não precisa ser determinada, e os dados básicos não precisam ser armazenados durante o período de comparação, os métodos não hierárquicos podem ser aplicados a conjuntos de dados muito maiores do que as técnicas hierárquicas.

Os métodos não hierárquicos podem começar de ambos (1) uma partição inicial de itens em grupos ou (2) um conjunto inicial de pontos de sementes, que formarão os núcleos dos *clusters*. Boas escolhas para configurações iniciais devem ser livres de vieses evidentes. Uma maneira de começar é selecionar aleatoriamente pontos de semente entre os itens ou particionar aleatoriamente os itens em grupos iniciais.

Nesta seção, discutiremos um dos procedimentos não hierárquicos mais populares, o método *K-means*.

2.5.1 Método *K-means*

O termo *K-means* é sugerido para descrever um algoritmo que atribui cada item ao *cluster* que tem o centróide mais próximo (média). Em sua versão mais simples, o processo é composto por estas três etapas:

1. Particionar os itens em K *clusters* iniciais.
2. Prossiga com a lista de itens, atribuindo um item ao *cluster* cujo centróide (média) é o mais próximo (a distância é geralmente calculada usando a distância Euclidiana com observações padronizadas ou não padronizadas). Recalcule o centróide para o *cluster* que recebe o novo item e para o *cluster* que perde o item.
3. Repita a Etapa 2 até que não ocorram mais reatribuições.

(2.15)

Ao invés de começar com uma partição de todos os itens em K grupos preliminares na Etapa 1, pode-se especificar K centróides iniciais (pontos de semente) e então prosseguir para a Etapa 2.

A atribuição final de itens aos *clusters* será, até certo ponto, dependente da partição inicial ou da seleção inicial de pontos de semente. A experiência sugere que a maioria das mudanças importantes na atribuição ocorre com a primeira etapa de realocação.

Exemplo 2.9 ***Clustering usando o método K-means***

Suponha que medimos duas variáveis X_1 e X_2 para cada um dos quatro itens A, B, C e D. Os dados são fornecidos na seguinte tabela:

Item	*Observações*	
	x_1	x_2
A	*5*	*3*
B	*-1*	*1*
C	*1*	*-2*
D	*-3*	*-2*

O objetivo é dividir esses itens em $K = 2$ clusters de forma que os itens dentro de um cluster fiquem mais próximos uns dos outros do que dos itens em diferentes clusters. Para implementar o método $K = 2$-means, particionamos arbitrariamente os itens em dois clusters, como (A, B) e (C, D), e calculamos as coordenadas $(\bar{x}_1, \bar{x}_2)$ do centróide do cluster (média). Assim, na Etapa 1, temos

Cluster	*Coordenadas do centróide*	
	$\bar{x}_1$	$\bar{x}_2$
(A,B)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(C,D)	$\frac{1+(-3)}{2} = -1$	$\frac{-2+(-2)}{2} = 2$

Na Etapa 2, calculamos a distância euclidiana de cada item dos centróides do grupo e reatribuímos cada item ao grupo mais próximo. Se um item for movido da configuração inicial, os centróides do cluster (médias) devem ser atualizados antes de continuar. A i-ésima coordenada, $i = 1, 2, ..., p$, do centróide é facilmente atualizada usando as fórmulas:

$$\bar{x}_{i,novo} = \frac{n\bar{x}_i + x_{ji}}{n+1} \qquad \textit{se o j-ésimo item é adicionado ao grupo}$$
$$\bar{x}_{i,novo} = \frac{n\bar{x}_i + x_{ji}}{n-1} \qquad \textit{se o j-ésimo item é removido do grupo}$$

Aqui n é o número de itens no grupo "antigo" com centróide $\bar{x}' = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)$.

Considere os clusters iniciais (A, B) *e* (C, D)*. As* ***coordenadas*** *dos centróides são* $(2, 2)$ *e* $(-1, -2)$ *respectivamente. Suponha que o item A, com as coordenadas* $(5, 3)$ *seja movido para o grupo* (C, D)*. Os novos grupos são* (B) *e* (A, C, D) *com centróides atualizados:*

Grupo (B) $\quad \bar{x}_{1,novo} = \frac{2(2)-5}{2-1} = -1 \quad \bar{x}_{2,novo} = \frac{2(2)-3}{2-1} = 1$, *as* ***coordenadas*** *de B*
Grupo (A, C, D) $\quad \bar{x}_{1,novo} = \frac{2(-1)+5}{2+1} = 1 \quad \bar{x}_{2,novo} = \frac{2(-2)+3}{2+1} = -0.33$

Retornando ao agrupamentos iniciais na Etapa 1, calculamos as distâncias quadradas

$d^2(A, (A, B)) = (5-2)^2 + (3-2)^2 = 10$
$d^2(A, (C, D)) = (5+1)^2 + (3+2)^2 = 61$ — *se A não é movido*

$d^2(A, (B)) = (5+1)^2 + (3-1)^2 = 40$
$d^2(A, (A, C, D)) = (5-1)^2 + (3+0.33)^2 = 27.09$ — *se A é movido para o grupo (C,D)*

Como A está mais próximo do centro de (A, B) *do que do centro de* (A, C, D)*, não é reatribuído.*

Continuando, consideramos reatribuir B. Obtemos

$d^2(B, (A, B)) = (-1-2)^2 + (1-2)^2 = 10$
$d^2(B, (C, D)) = (-1+1)^2 + (1+2)^2 = 9$ — *se B não é movido*

$d^2(B, (A)) = (-1-5)^2 + (1-3)^2 = 40$
$d^2(B, (B, C, D)) = (-1+1)^2 + (1+1)^2 = 4$ — *se B é movido para o grupo (C,D)*

Como B está mais próximo do centro de (B, C, D) *do que do centro de* (A, B)*, B é reatribuído ao grupo* (C, D)*. Agora temos os clusters* (A) *e* (B, C, D) *com coordenadas de centróide* $(5, 3)$ *e* $(-1, -1)$*, respectivamente.*

Verificamos C para reatribuição.

$d^2(C, (A)) = (1-5)^2 + (-2-3)^2 = 41$
$d^2(C, (B, C, D)) = (1+1)^2 + (-2+1)^2 = 5$ — *se C não é movido*

$d^2(C, (A, C)) = (1-3)^2 + (-2-0.5)^2 = 10.25$
$d^2(C, (B, D)) = (1+2)^2 + (-2+0.5)^2 = 11.25$ — *se C é movido para o grupo (A)*

Como C está mais perto do centro do grupo (B, C, D) *do que do centro do grupo* (A, C)*, C não é movido. Continuando desta forma, descobrimos que não ocorrem mais reatribuições e os clusters* $K = 2$ *finais são* (A) *e* (B, C, D)*.*

Para os clusters finais, temos

	Distâncias quadradas para o grupo de centróides			
Cluster	*Item*			
	A	*B*	*C*	*D*
A	*0*	*40*	*41*	*89*
(B,C,D)	*52*	*4*	*5*	*5*

A soma dos quadrados dentro do cluster (soma das distâncias quadradas para o centróide) são

$$\textit{Cluster } A : 0$$
$$\textit{Cluster } (B, C, D) : 4 + 5 + 5 = 14$$

Equivalentemente, podemos determinar os clusters $K = 2$ *usando o critério*

$$min\ E = \sum d^2_{i,c(i)}$$

onde o mínimo é superior ao número de clusters $K = 2$ *e* $d^2_{i,c(i)}$ *é a distância quadrada do caso i do centróide (média) do cluster atribuído.*

Neste exemplo, existem sete possibilidades para clusters $K = 2$*:*

$$\begin{array}{ll} A, (B, C, D) & \\ B, (A, C, D) & \textit{(A,B),(C,D)} \\ C, (A, B, D) & \textit{(A,C),(B,D)} \\ D, (A, B, C) & \textit{(A,D),(B,C)} \end{array}$$

Para o par A, (B, C, D)*:*

$$\begin{array}{rl} A & d^2_{A,c(A)} = 0 \\ (B, C, D) & d^2_{B,c(B)} + d^2_{C,c(C)} + d^2_{D,c(D)} = 4 + 5 + 5 = 14 \end{array}$$

Consequentemente, $\sum d^2_{i,c(i)} = 0 + 14 = 14$

Para os pares restantes, você pode verificar que

$$\begin{aligned} B,(A,C,D) \qquad & \sum d^2_{i,c(i)} = 48.7 \\ C,(A,B,D) \qquad & \sum d^2_{i,c(i)} = 27.7 \\ D,(A,B,C) \qquad & \sum d^2_{i,c(i)} = 31.3 \\ (A,B),(C,D) \qquad & \sum d^2_{i,c(i)} = 28 \\ (A,C),(B,D) \qquad & \sum d^2_{i,c(i)} = 27 \\ (A,D),(B,C) \qquad & \sum d^2_{i,c(i)} = 51.3 \end{aligned}$$

Visto que o menor $\sum d^2_{i,c(i)}$ *ocorre para o par de clusters* (A) *e* (BCD)*, esta é a partição final .*

Para verificar a estabilidade do *clustering*, é desejável executar novamente o algoritmo com uma nova partição inicial. Uma vez que os *clusters* são determinados, as intuições relativas às suas interpretações são auxiliadas pelo rearranjo da lista de itens de forma que aqueles no primeiro *cluster* apareçam primeiro, aqueles no segundo *cluster* apareçam em seguida, e assim por diante. Uma tabela dos centróides do *cluster* (médias) e variâncias dentro do *cluster* também ajudam a delinear as diferenças do grupo.

Exemplo 2.10 ***Clustering K-means de serviços públicos***

Vamos retornar ao problema de clustering de serviços públicos usando os dados da Tabela 3. O algoritmo K-means para várias escolhas de K foi executado. Apresentamos um resumo dos resultados para $K = 4$ *e* $K = 5$*. Em geral, a escolha de um K em particular não é clara e depende do conhecimento do assunto, bem como de avaliações baseadas em dados. As avaliações baseadas em dados podem incluir a escolha de K de modo a maximizar a variabilidade relativa entre os clusters para a variabilidade dentro do cluster. O resumo é o seguinte:*

Distâncias entre centros de cluster

	1	2	3	4
1	0			
2	3.08	0		
3	3.29	3.56	0	
4	3.05	2.84	3.18	0

Distâncias entre centros de cluster

$K = 4$

Cluster	N° de empresas	Empresas
1	5	Idaho Power Co.(8), Nevada Power Co.(11), Puget Sound Power & Light Co.(16), Virginia Eletric Power Co.(22), Kentucky Utilities Co.(9)
2	6	Central Louisiana Eletric Co.(3), Oklahoma Gas & Eletric Co.(14), The Southern Co.(18), Texas Utilities Co.(19), Arizona Public Service(1), Florida Power & Light Co.(6)
3	5	New England Eletric Co.(12), Pacific Gas & Eletric Co.(15), San Diego Gas & Eletric Co.(17), United Illuminating Co.(21), Hawaiian Eletric Co.(7)
4	6	Consolidated Edison Co.(N.Y.)(5), Boston Edison Co. (2), Madison Gas & Eletric Co.(10), Northern States Power Co.(13), Wisconsin Eletric Power Co. (20), Commonwealth Edison Co.(4)

$K = 5$

Cluster	N° de firmas	Firmas
1	5	Nevada Power Co.(11), Puget Sound Power & Light Co.(16), Idaho Power Co.(8), Virginia Eletric Power Co. (22), Kentucky Utilities Co.(9)
2	6	Central Louisiana Eletric Co.(3), Texas Utilities Co.(19), Oklahoma Gas & Eletric Co.(14), The Southern Co. (18), Arizona Public Service(1), Florida Power & Light Co.(6)
3	5	New England Eletric Co.(12), Pacific Gas & Eletric Co.(15), San Diego Gas & Eletric Co.(17), United Illuminating Co.(21), Hawaiian Eletric Co.(7)
4	2	Consolidated Edison Co.(N.Y.)(5), Boston Edison Co.(2)
5	4	Commonwealth Edison Co.(4), Madison Gas & Eletric Co.(10) Northern States Power Co.(13), Wisconsin Eletric Power Co.(20)

$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{bmatrix}
0 & & & & \\
3.08 & 0 & & & \\
3.29 & 3.56 & 0 & & \\
3.63 & 3.46 & 2.63 & 0 & \\
3.18 & 2.99 & 3.81 & 2.89 & 0
\end{bmatrix}
\end{array}
$$

Analisando as empresas nos cinco clusters, é evidente que o método K-means fornece

resultados geralmente consistentes com o método hierárquico average linkage. Empresas com cluster de localizações geográficas comuns ou compatíveis. Custos de combustível e vendas anuais, por exemplo, podem ter alguma importância na distinção dos clusters.

Para verificar a estabilidade do agrupamento, é desejável executar novamente o algoritmo com uma nova partição inicial. Uma vez que os clusters são determinados, as intuições relativas às suas interpretações são auxiliadas reorganizando a lista de itens de forma que aqueles no primeiro agrupamento apareçam primeiro, os do segundo agrupamento apareçam em seguida, e assim por diante. Uma tabela dos centróides do cluster (médias) e variâncias dentro do cluster também ajuda a delinear as diferenças do grupo.

2.5.2 Comentários finais - procedimentos não hierárquicos

Existem fortes argumentos para não fixar o número de *clusters*, K, antecipadamente, incluindo o seguinte:

1. Se dois ou mais pontos de semente inadvertidamente estiverem dentro de um único *cluster*, os seus *clusters* resultantes serão mal diferenciados.

2. A existência de um outlier pode produzir pelo menos um grupo com itens muito dispersos.

3. Mesmo se a população for conhecida por consistir de K grupos, o método de amostragem deve ser tal que dados dos mais raros grupos não aparecem na amostra. Forçar os dados em K grupos levaria a *clusters* sem sentido.

Nos casos em que uma única execução do algoritmo exige que o usuário especifique K, é sempre uma boa ideia executar novamente o algoritmo para várias opções.

3 Análise dos Resultados

Este Capítulo apresenta os resultados das análises com a aplicação dos algoritmos de *clustering* em algumas bases de dados extraídas do Kaggle[1], utilizando a linguagem de programação R, que é uma linguagem de programação *open source* especificamente para computação estatística, e tanto os pacotes quanto as funções usadas para cada algoritmo são detalhadas no texto.

3.1 Base 1: *Humanitarian Aid to Underdeveloped Countries*

O primeiro banco de dados utilizado é "*Humanitarian Aid to Underdeveloped Countries*", que fala sobre a ONG HELP International, uma ONG humanitária internacional que é empenhada em combater a pobreza e fornecer às pessoas de países atrasados amenidades básicas e alívio durante desastres e calamidades naturais. Após recentes programas de financiamento, eles conseguiram arrecadar cerca de 10 milhões de dólares. Com esse dinheiro, o CEO (Chief Executive Officer) da ONG precisa decidir como usá-lo de forma estratégica e eficaz, e as questões significativas que surgem ao tomar essa decisão estão relacionadas principalmente à escolha dos países que mais precisam de ajuda[12].

A base contém 167 países, apresentando 9 variáveis socioeconômicas:

- child_mort: Morte de crianças menores de 5 anos de idade por 1000 nascidos vivos;
- exports: Exportações de bens e serviços per capita;
- health: Gasto total com saúde per capita;
- imports: Importação de bens e serviços per capita;
- Income: Renda líquida por pessoa;

[1] https://www.kaggle.com/

- Inflation: Aumento dos preços de bens e serviços. Quanto maior a taxa de inflação, menor é o poder de compra da moeda;
- life_expec: Número médio de anos que uma criança recém-nascida viveria se os padrões atuais de mortalidade permanecessem os mesmos;
- total_fer: Número de filhos que nasceriam de cada mulher se as atuais taxas de fertilidade por idade permanecessem as mesmas;
- gdpp: PIB per capita. Calculado como o PIB total dividido pela população total.

Começando com o método *k-means*, primeiramente aplica-se a função *fviz_nbclust*, do pacote *factoextra*, para encontrar o número de *clusters* ideais. A função possui três métodos diferentes: *Elbow* (método de "cotovelo"), *Silhouette* (método silhueta) e *Gap statistic* (método estatístico de lacunas). Para cada método, a função gera um gráfico com a informação do número ideal de *clusters*, indicado pela linha tracejada. E há um quarto gráfico produzido pela função *NbClust*, do pacote de mesmo nome *NbClust*, que mostra a quantidade de vezes que um número k de *clusters* é sugerido como o ideal.

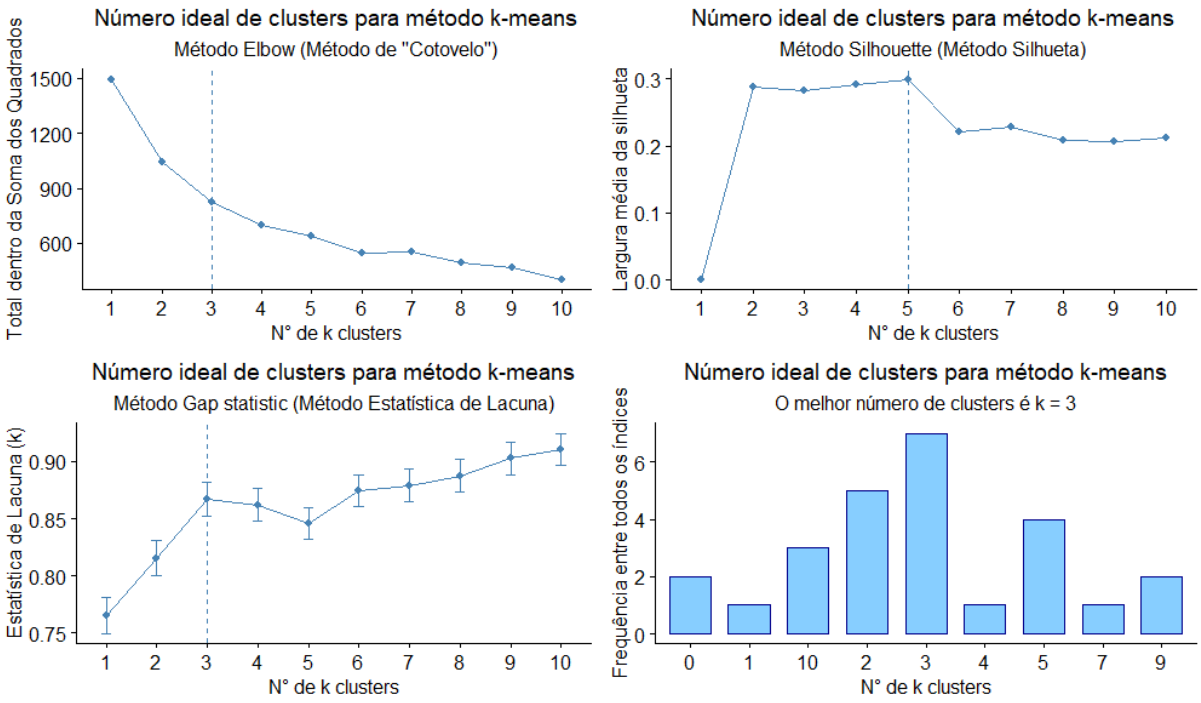


Figura 9: Verificando n° ideal de *clusters*

A Figura 9 mostra os 4 gráficos gerados a partir das funções. A ideia por trás do método *Elbow* é definir *clusters* de forma que a variação total dentro do *cluster* (ou soma quadrada total dentro do cluster) seja minimizada. Seu gráfico mostra a relação entre o

n° k de *clusters* e o total dentro da Soma dos Quadrados. Conforme o n° k de *clusters* aumenta, o total da soma tende a diminuir, e como esta soma mede a compactação do *cluster*, deseja-se que seja o menor possível[11]. Deve-se escolher um número de *clusters* que, ao adicionar outro *cluster*, não haja muita melhora no total dentro da Soma dos Quadrados. Pelo gráfico, a localização de uma dobra ("cotovelo"), indica o número ideal de *clusters*, neste caso $k = 3$.

O método de silhueta mede a qualidade de um agrupamento, determinando o quão bem cada objeto se encontra em seu *cluster*. Uma largura média alta da silhueta indica um bom agrupamento, e a localização do máximo é considerada como o número apropriado de *clusters*[11]. Neste caso, o número ideal encontrado foi 5.

Já a estatística de lacuna compara o total da variação dentro do *cluster* para diferentes valores de k com sua expectativa sob uma distribuição de referência nula apropriada[18]. A estimativa dos *clusters* ideais será o valor que maximiza a estatística de lacuna (ou seja, que produz a maior estatística de lacuna)[11]. O número k ideal encontrado foi 3, como no método de "cotovelo".

E a função *NbClust* fornece 30 índices para determinar o número de *clusters* e propõe ao usuário o melhor esquema de *clustering* a partir dos diferentes resultados obtidos pela variação de todas as combinações de número de *clusters*, medidas de distância e métodos de *clustering*[3]. Pelo gráfico, observa-se que realmente o número ideal de *clusters* é $k = 3$. No Apêndice 1, apresenta o cenário para o método k-means com 5 *clusters*.

Escolhido o número ideal de *clusters*, aplica-se a função *fviz_cluster*, gerando uma visualização dos *clusters* com seus respectivos países na Figura 10.

Para verificar o quão bem uma observação está agrupada, aplica-se a função *fviz_silhouette*, do pacote *factoextra*. A análise de silhueta mede e estima a distância média entre os *clusters*. O gráfico de silhueta exibe uma medida de quão próximo cada ponto em um *cluster* está de pontos nos *clusters* vizinhos[1].

Tabela 4: N° de elementos e silhueta média de cada *cluster*

cluster	N° de países	Largura média da silhueta
1	36	0,15
2	84	0,36
3	47	0,24

As observações com Si grande (próximo de 1) são muito bem agrupadas, enquanto

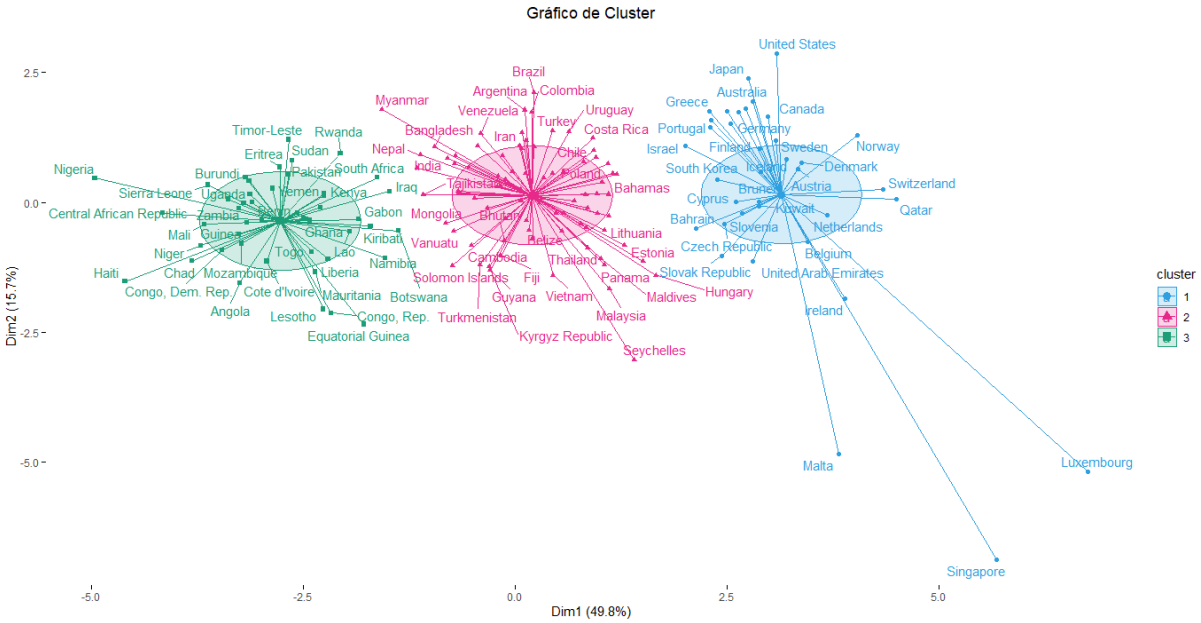


Figura 10: *Clustering k-means* aplicado nos 167 países

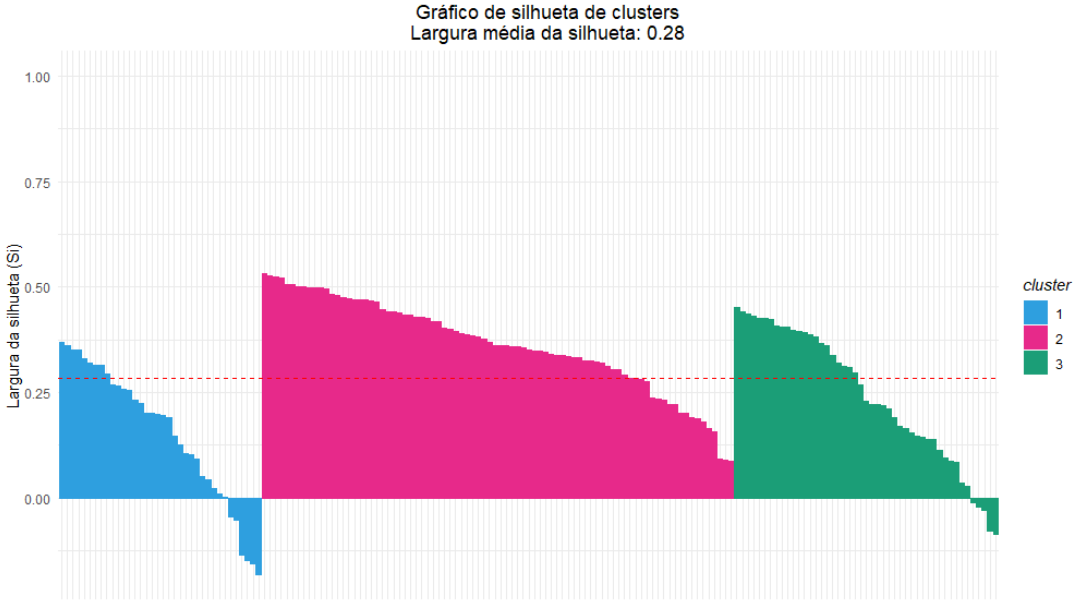


Figura 11: O coeficiente de silhueta de observações

um *Si* pequeno (em torno de 0) significa que a observação está entre dois *clusters*. E, quando as observações apresentam um *Si* negativo, provavelmente estas observações estão colocadas no *cluster* errado[2].

Observa-se na Figura 11 que os *clusters* 1 e 3 apresentam observações com *Si* negativo, indicando que estes países deveriam estar no *cluster* vizinho. A Tabela 5 mostra os países que podem estar agrupados de forma errada.

Tabela 5: Países agrupados no *cluster* errado

País	*cluster*	*cluster* vizinho	Largura da silhueta
Chipre	1	2	-0,044
Israel	1	2	-0,052
República Tcheca	1	2	-0,134
Coréia do Sul	1	2	-0,149
República Eslovaca	1	2	-0,155
Bahrein	1	2	-0,182
Laos	3	2	-0,012
África do Sul	3	2	-0,021
Namíbia	3	2	-0,029
Iraque	3	2	-0,078
Botswana	3	2	-0,086

Pela Tabela 5, conclui-se que todos os 11 países pertencentes aos *clusters* 1 e 3 com valores de Si negativos também podem pertencer ao *cluster* 2. Entretanto, estas observações serão mantidas em seus respectivos *clusters*, pois a análise de silhueta terá como objetivo apenas verificar a eficiência dos agrupamentos. Portanto, nenhuma observação será realocada em outro *cluster*.

Na Figura 10, o *cluster* 1 apresenta 3 países bem distantes de seu centróide: Malta, Luxemburgo e Singapura, indicando que estes países se diferem um pouco mais em relação aos demais países de seu grupo. O segundo *cluster* é o que mais apresenta elementos, com 84 países, conforme a Tabela 4. Observa-se os países de acordo com seu respectivo *cluster* na Figura 12, e os países na cor cinza não fazem parte da base de dados.

Percebe-se pelo mapa que o *cluster* 1 é composto por países desenvolvidos, como Estados Unidos, Canadá, países europeus, Japão, Coréia do Sul, Austrália e Nova Zelândia. O *cluster* 2 é composto por toda América do Sul (com exceção da Guiana Francesa), alguns países da América Central, grande parte do continente asiático, e a parte norte do continente africano, que seriam países do Grande Oriente Médio. Já o *cluster* 3 apresenta a maior parte do continente africano, e alguns países do Grande Oriente Médio na parte asiática.

Tendo conhecimento em relação aos países que compõem cada *cluster*, é preciso entender o comportamento destes *clusters*, sendo calculado a média de todas as variáveis dado pela Tabela 6 e a Tabela 7, para identificar as características de cada grupo, assim descrevendo as similaridades entre os países.

Analisando a Tabela 6 e a Tabela 7, observa-se a princípio que o *cluster* 3 é o que apre-

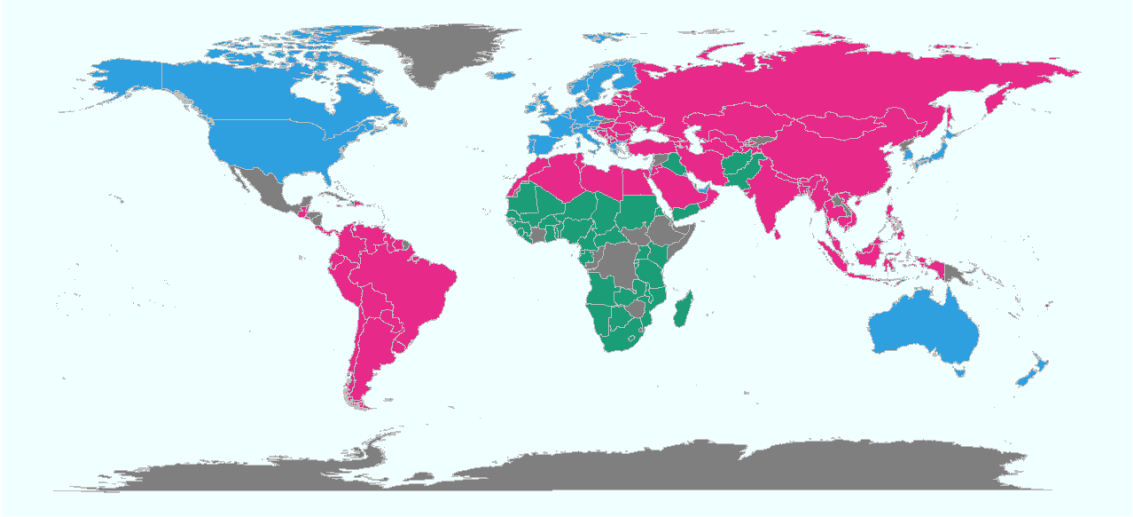


Figura 12: Representação dos países no mapa com relação ao *cluster*

Tabela 6: Média das variáveis de cada *cluster*

cluster	child_mort	exports	health	imports	income
1	5	58,739	8,808	51,492	45.672,220
2	21,927	40,244	6,201	47,473	12.305,590
3	92,962	29,151	6,389	42,323	3.942,404

Tabela 7: Média das variáveis de cada *cluster*

cluster	inflation	life_expec	total_fer	gdpp
1	2,671	80,128	1,753	42.494,440
2	7,601	72,814	2,308	6.486,452
3	12,020	59,187	5,008	1.922,383

senta a maior taxa média de mortalidade infantil, menos exportações de bens e serviços, menos arrecadam, possuí as maiores taxas de inflação, a menor expectativa de vida, a maior taxa de natalidade, e o menor PIB, indicando que os países pertecentes a este *cluster* são os mais pobres, que mais necessitam de ajuda. Tendo conhecimento dos países que compõem este *cluster*, visto na Figura 12, já era esperado que este grupo seria o que mais necessitaria de ajuda, pois sabe-se que o continente africano infelizmente apresenta uma extrema pobreza.

Por outro lado, o *clusters* 1 é o que apresenta as menores taxas de mortalidade infantil, o que mais exporta serviços e bens, as maiores taxas de gastos médios com saúde, o que

menos importa bens e serviços, o que mais arrecada, as menores taxas de inflação, a maior expectativa de vida, as menores taxas de natalidade, e o melhor PIB per capita, indicando que são os países mais ricos e menos necessitados.

Logo, pode-se afirmar que os países que mais precisam de ajuda da ONG são aqueles pertencentes ao *cluster* 3, seguidos pelos países que compõem o grupo 2, e por último, os que não passam por tantas dificuldades, os que pertencem ao *cluster* 1.

Feita as análises baseadas no método *k-means*, é aplicado o método hierárquico, e observar se os resultados são semelhantes ao método anterior. Aqui é apresentado e analisado o método *complete linkage*, os outros métodos podem ser vistos nos Apêndices 4 e 5.

Como feito no método *k-means*, são aplicadas as funções *fviz_nbclust* e *NbClust* para verificar o número ideal de *clusters*, dessa vez sendo voltado para o método hierárquico *complete linkage*.

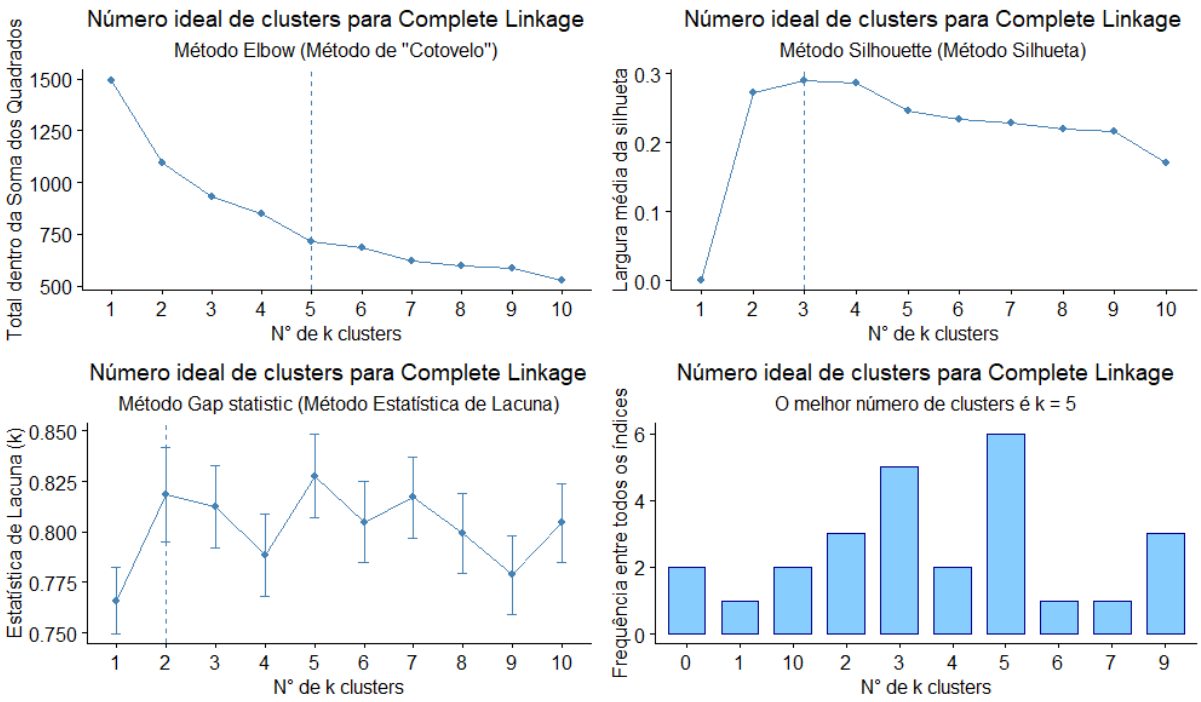


Figura 13: Verificando n° ideal de *clusters*

Na Figura 13, observa-se que o número ideal de grupos é 5, sendo $k = 3$ e $k = 2$ vindo em seguida. Os cenários para $k = 3$ e $k = 2$ estão nos Apêndices 2 e 3. Escolhido $k = 5$, aplica-se as funções *eclust* e *fviz_dend*, para calcular as distâncias entre os elementos, e criação do dendrograma.

Comparando com o número de *clusters* utilizado no método *k-means*, pode-se observar

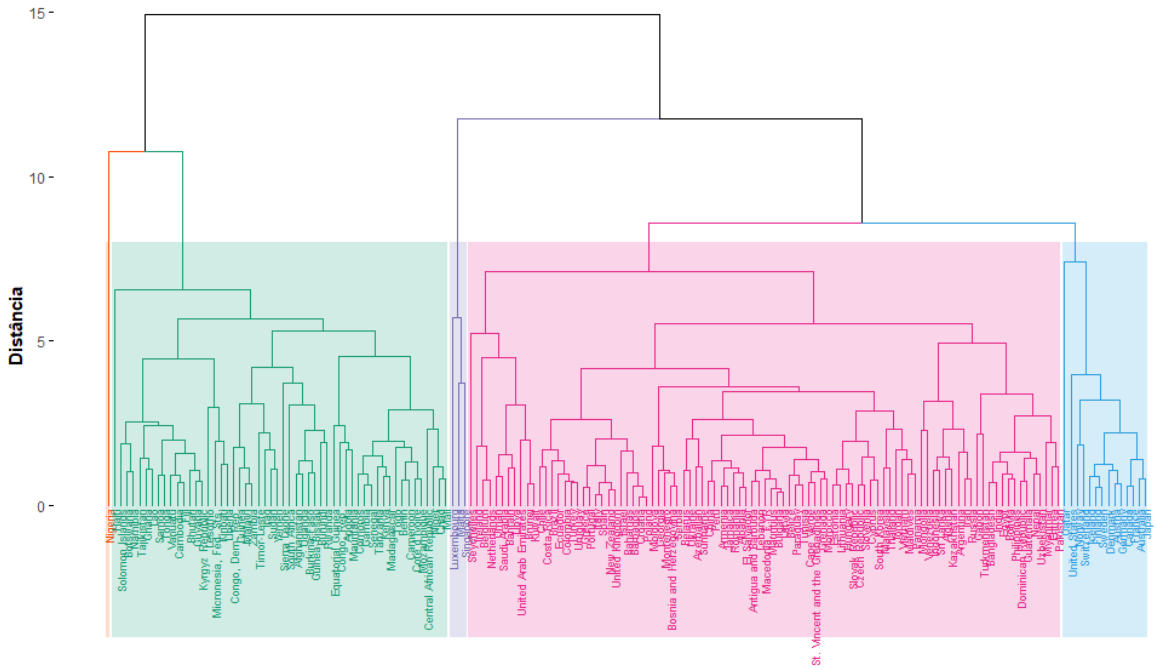


Figura 14: Dendrograma *Complete Linkage* dos 167 países

que este método acrescentou 2 grupos, contendo apenas 4 países (3 e 1, respectivamente), e o *cluster* 4 apresenta os 3 países que são justamente aqueles que pertenciam ao *cluster* 1 do método *k-means*, os países mais ricos, mas que eram os mais distantes do centróide: Luxemburgo, Malta e Singapura. Observa-se uma certa dificuldade de visualização, visto na Figura 14, por conta da quantidade de elementos da base, os 167 países, e vê-se um *cluster*, destacado na cor vermelha, formado apenas por Nigéria.

Como feito anteriormente no método *k-means*, é aplicado a função *fviz_silhouette* para visualizar o quão bem as observações estão agrupadas, e verificar quais países deveriam estar no *cluster* vizinho.

Dessa vez, há diversas observações do *cluster* 2 com a largura de silhueta negativa, visto pela Figura 15. O primeiro *cluster* também apresenta observações com Si negativo. Foram 22 países agrupados em *clusters* errados, como pode ser visto pela Tabela 8, o dobro de observações comparado ao método anterior.

Na Tabela 9, é apresentado a quantidade de países dentro de cada *cluster* e a largura média da silhueta indicando o quão bem cada *cluster* foi formado, e a Figura 16 mostra o mapa com esses novos grupos formado pelo método *complete linkage*.

Observa-se pelos mapas, Figuras 12 e 16, que houve algumas mudanças, pois, tirando os 4 países que foram divididos em 2 grupos, de 3 e 1 objetos respectivamente, percebe-

Tabela 8: Países agrupados no *cluster* errado

País	*cluster*	*cluster* vizinho	Largura da silhueta
Cambodja	1	2	-0,058
Tonga	1	2	-0,068
Samoa	1	2	-0,077
Guiana	1	2	-0,084
Quirguistão	1	2	-0,087
Fiji	1	2	-0,173
Butão	1	2	-0,226
Israel	2	3	-0,015
Brunei	2	3	-0,055
Kuwait	2	3	-0,065
Portugal	2	3	-0,098
Gabão	2	1	-0,115
Irlanda	2	3	-0,164
Grécia	2	3	-0,166
Paquistão	2	1	-0,207
Eritreia	2	1	-0,239
Espanha	2	3	-0,245
Bélgica	2	3	-0,257
Nova Zelândia	2	3	-0,294
Itália	2	3	-0,352
Reino Unido	2	3	-0,392
Holanda	2	3	-0,401

Tabela 9: N° de elementos e silhueta média de cada *cluster*

cluster	N° de países	Largura média da silhueta
1	54	0,25
2	95	0,21
3	14	0,42
4	3	0,40
5	1	0,00

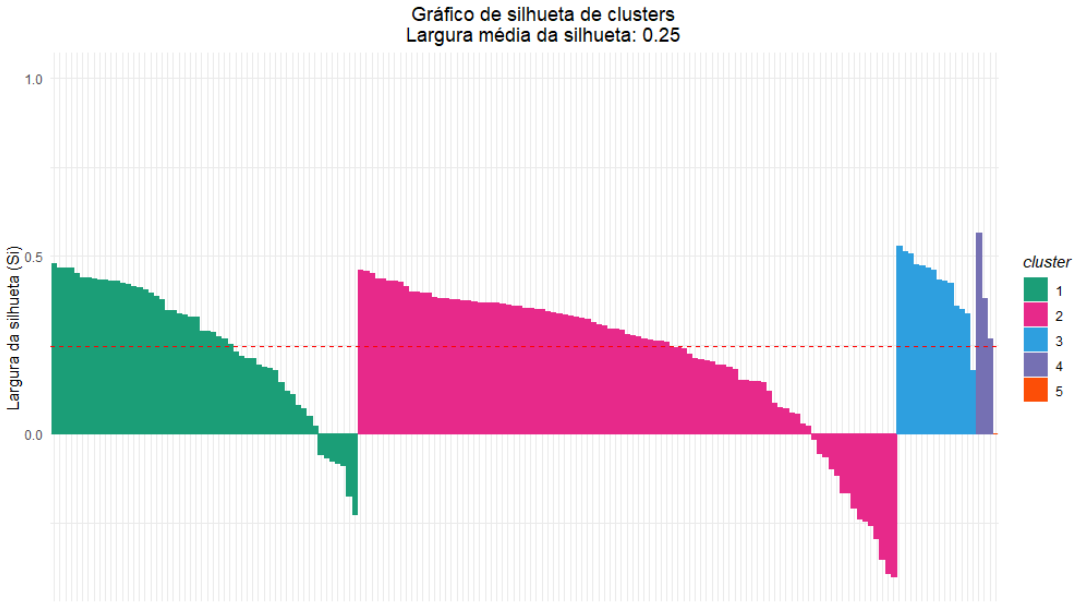


Figura 15: O coeficiente de silhueta de observações

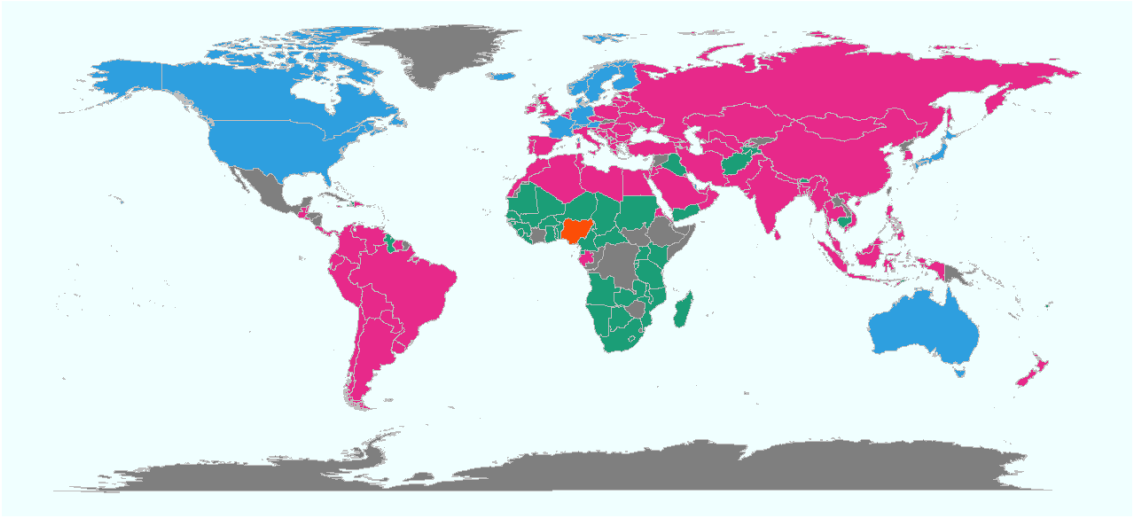


Figura 16: Representação dos países no mapa com relação ao *cluster*

se que há outros países que mudaram de *cluster*, os três maiores *clusters* do método hierárquico não são iguais aos *clusters* criados com o método *k-means*. E pela Tabela 8, observa-se que essa mudança no visual do mapa se deve ao fato de alguns países que estarem no *cluster* 2, mas deviam estar no *cluster* 3, como por exemplo, Portugal, Irlanda, Espanha, Bélgica, Nova Zelândia, Reino Unido e Holanda.

Para entender o comportamento desses novos grupos, cria-se novamente tabelas para compreender as características de cada *cluster*.

Tabela 10: Média das variáveis de cada *cluster*

cluster	child_mort	exports	health	imports	income
1	81,344	31,521	6,660	49,111	3.787,463
2	18,875	42,719	6,411	44,488	18.582
3	4,500	39,393	10,292	33,207	49.721,430
4	4,133	176	6,793	156,667	64.033,330
5	130	25,300	5,070	17,400	5.150

Tabela 11: Média das variáveis de cada *cluster*

cluster	inflation	life_expec	total_fer	gdpp
1	8,926	60,556	4,697	1.910,074
2	7,124	74,479	2,142	11.941,470
3	2,094	80,893	1,800	53.742,860
4	2,468	81,433	1,380	57.566,670
5	104	60,500	5,840	2.330

Analisando as Tabelas 10 e 11, chama a atenção os dados sobre a Nigéria, único país a compor o *cluster* 5, porque seus números são muito piores que a média dos países que compõem o primeiro *cluster*, pois se tratam dos países africanos que mais precisam de ajuda. A Nigéria possui a maior taxa de mortalidade infantil, menos exportam bens e serviços, menos investem em saúde, menos importam bens e serviços, a maior taxa de inflação e a maior taxa de natalidade, fazendo da Nigéria um dos países, ou se não o que mais, necessita de ajuda.

O *cluster* 4 chama a atenção pelo fato de possuir a maior taxa de exportações de bens e serviços, e a maior taxa de importações de bens e serviços, com médias muito maiores que todos os outros *clusters*, inclusive bem maiores que dos países que compõem o *cluster* 3, que baseado no mapa, são os países mais desenvolvidos do mundo. Além disso, estes três países que compõem o grupo 4, que não é possível de se ver no mapa da Figura 16, mas sabe-se que são Luxemburgo, Malta e Singapura, ainda são os que mais arrecadam, possuem a maior taxa de expectativa de vida, a menor taxa de natalidade, e o maior PIB. Portanto, são os três países que não necessitam de ajuda.

Em relação aos *clusters* 1, 2 e 3, nota-se que suas médias são parecidas, ou pelo menos se comportam de forma semelhante, aos *clusters* 3, 2 e 1, nesta ordem, criados do método *k-means*, fazendo com que exista indícios que os métodos conseguem trazer resultados bem semelhantes, pois como ocorrido no primeiro método, concluí-se que, por uma ordem de necessidade, deixando de lado a Nigéria, o *cluster* 1 do método hierárquico (semelhante

ao *cluster* 3 do *k-means*) é o que apresenta os países que mais precisam de apoio, seguido pelo *cluster* 2, e em seguida o *cluster* 3.

3.2 Base 2: *Simple Clustering Data ID Gender Income Spending*

O segundo banco de dados é o "*Simple Clustering Data ID Gender Income Spending*", uma base de dados fictícios bem simples[17] contendo informações sobre:

- ID: Identificação do cliente;
- Gender_Code: Gênero;
- Region: Região;
- Income: Renda (milhares/mês);
- Spending: Gastos (milhares/mês).

Esta base apresenta 1113 observações, porém, 23 observações são excluídas por apresentarem dados faltantes, totalizando 1090 observações estudadas. O conjunto de dados propicia a aplicação tanto do *clustering* com *k-means* quanto dos métodos hierárquicos, mas, como é um número grande de observações, a visualização dos dados pode ficar um pouco rasurada.

Como há poucas informações, pode-se entender um pouco o comportamento da base com algumas estatísticas descritivas, com o objetivo de descobrir alguns padrões, antes da aplicação dos métodos de *clustering*.

Na Figura 17, pode-se observar que o número de homens e mulheres são bem equilibrados, junto com o número de indivíduos que vivem em zonas urbanas ou rurais, não existe uma diferença significativa na quantidade de cada categoria.

Já em relação a renda e aos gastos dos indivíduos, percebe-se que quem mora no meio rural apresenta uma renda menor que as pessoas que residem no meio urbano, e chegando a ter gastos iguais ganhando menos, e também pode-se ver que não há diferença significativa nos ganhos e gastos com relação ao sexo do indivíduo.

Como feito na base anterior, é aplicado primeiramente o método *k-means*, e novamente são aplicadas as funções *fviz_nbclust* e *NbClust* para verificar o número ideal de *clusters*.

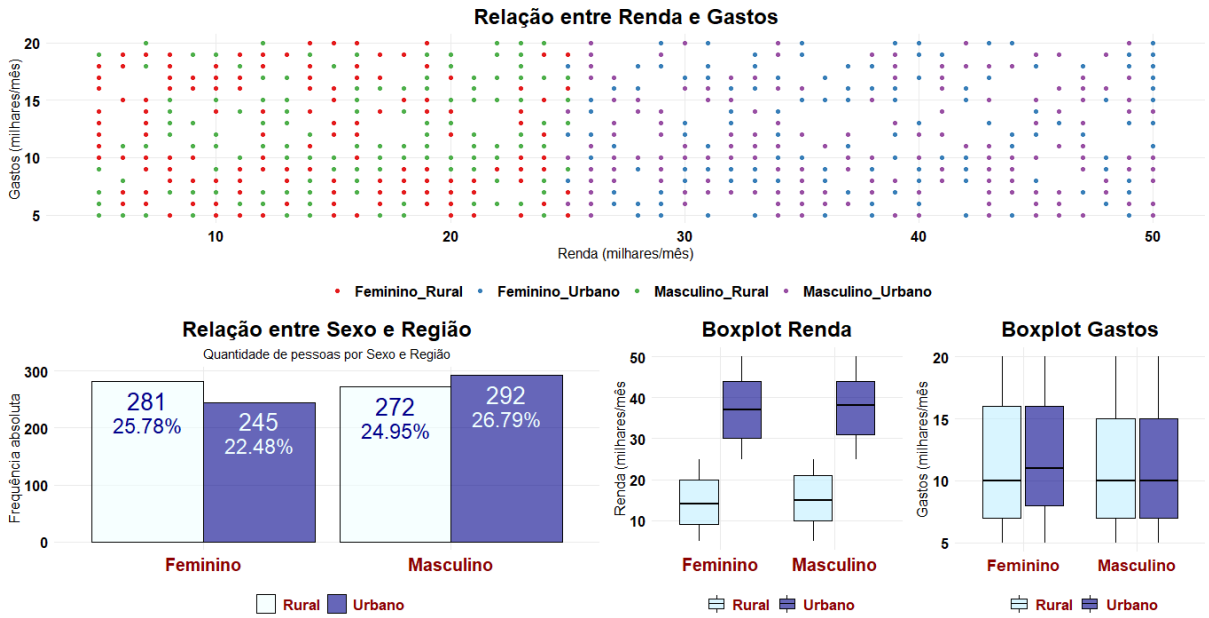


Figura 17: Estatísticas descritivas

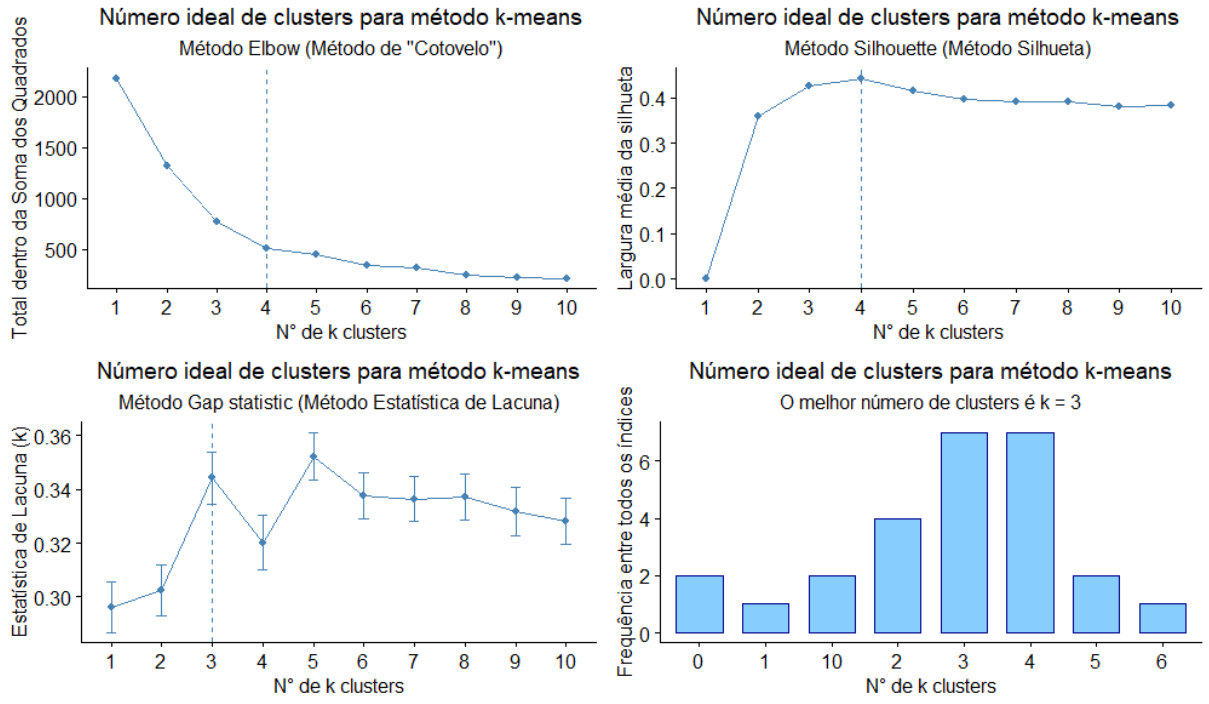


Figura 18: Verificando n° ideal de *clusters*

Observa-se pela Figura 18 que o número ideal de *clusters* é 4, podendo também ser $k = 3$, este aplicado no Apêndice 6. Aplicando a função *fviz_cluster*, tem-se a visualização dos *clusters*, com $k = 4$ Em seguida, a aplicação da análise de silhueta.

Diferente da base anterior, o gráfico de silhueta da Figura 20 mostra que o agrupamento desta base de dados é bem mais eficiente.

Figura 19: *Clustering k-means* aplicado nas 1090 observações

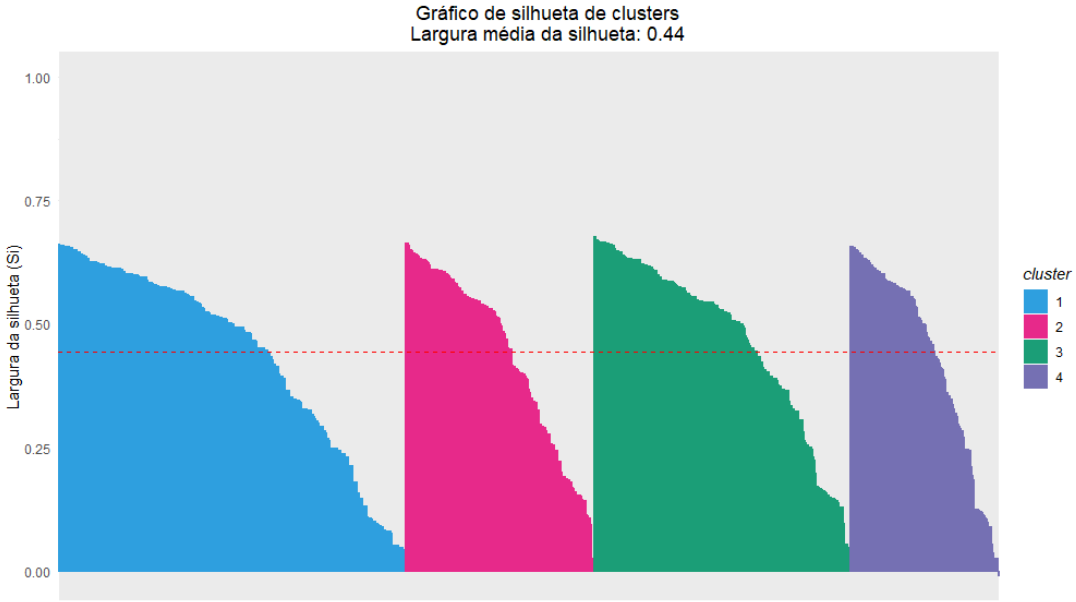


Figura 20: O coeficiente de silhueta de observações

Tabela 12: N° de elementos e silhueta média de cada *cluster*

cluster	N° de elementos	Largura média da silhueta
1	402	0,44
2	219	0,43
3	296	0,47
4	173	0,43

Tabela 13: Observações agrupadas no *cluster* errado

ID	*cluster*	*cluster* vizinho	Largura da silhueta
652	4	2	-0, 009

Na Figura 19, observa-se os 4 *clusters*, e com base na Tabela 12, o primeiro *cluster* é o que mais possuí elementos, com 402 observações, seguido do *cluster* 3, com 296, e o que apresenta menos elementos é o *cluster* 4. A Tabela 13 mostra que apenas uma observação foi agrupada de forma inapropriada.

Para entender o comportamento desses grupos, verifica-se as médias das variáveis dentro de cada *cluster* dado na Tabela 14:

Tabela 14: Média das variáveis de cada *cluster*

cluster	Income	Spending
1	15.905	8.231
2	16.187	16.489
3	39.125	8.128
4	39.549	16.890

O *cluster* 1 é caracterizado pelos indivíduos que apresentam ganhos de aproximadamente 16 mil ao mês, e gastos um pouco maiores que metade do que arrecadam. O segundo grupo apresenta renda média um pouco maior que 16 mil, mas gastam em média um pouco mais do que ganham. Já os *clusters* 3 e 4, são os que mais arrecadam, um pouco mais de 39 mil em média, mas no caso do terceiro grupo, são os que menos gastam, aproximadamente 1/5 do que arrecadam, e o grupo 4 gastam em média um pouco mais de 16 mil, semelhante aos gastos médios do *cluster* 2.

Como as características de cada *cluster* são conhecidas, verifica-se como as variáveis se comportam dentro de cada grupo criado:

Na Figura 21, conclui-se que as pessoas que menos arrecadam são majoritarimente do meio rural, como já havia sido notado nas estatísticas descritivas da Figura 17, antes da aplicação do *k-means*, e percebe-se um número maior de mulheres que apresentam gastos maiores do que sua renda.

Nos *clusters* 3 e 4, existem apenas indivíduos das zonas urbanas, justamente os que mais arrecadam, e nota-se que há uma quantidade maior de homens que gastam em média 1/5 de sua renda, em relação as mulheres. Observa-se também um equilíbrio no número

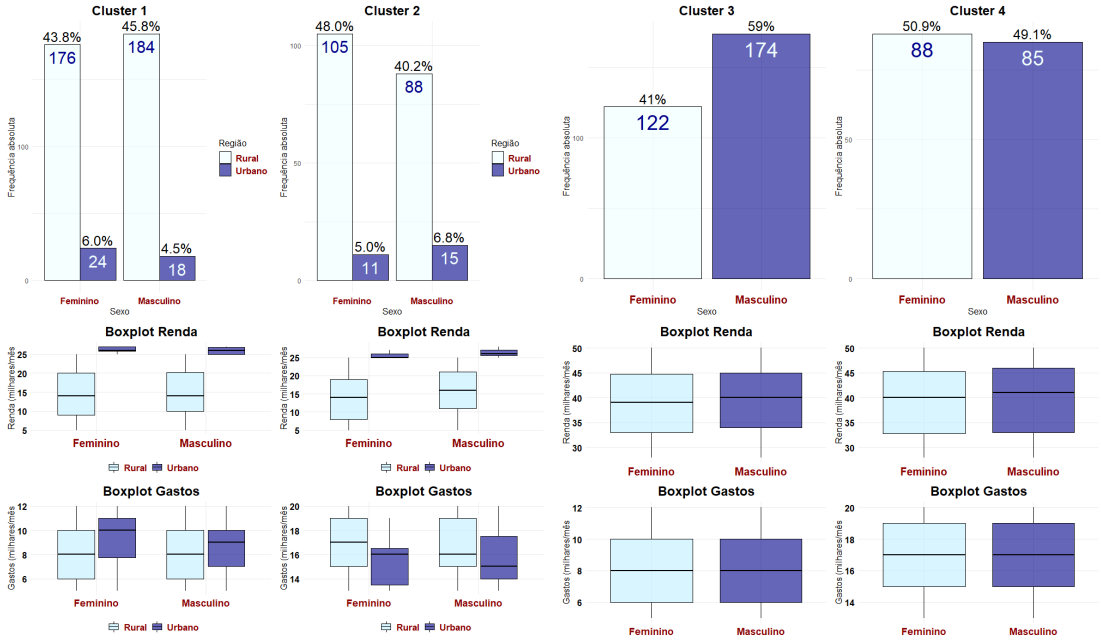


Figura 21: Estatísticas descritivas dos 4 clusters

de homens e mulheres no *cluster* 4.

Terminada as análises utilizando o método *k-means*, aplica-se o método hierárquico, e como feito na primeira base, o método hierárquico escolhido é o *complete linkage*.

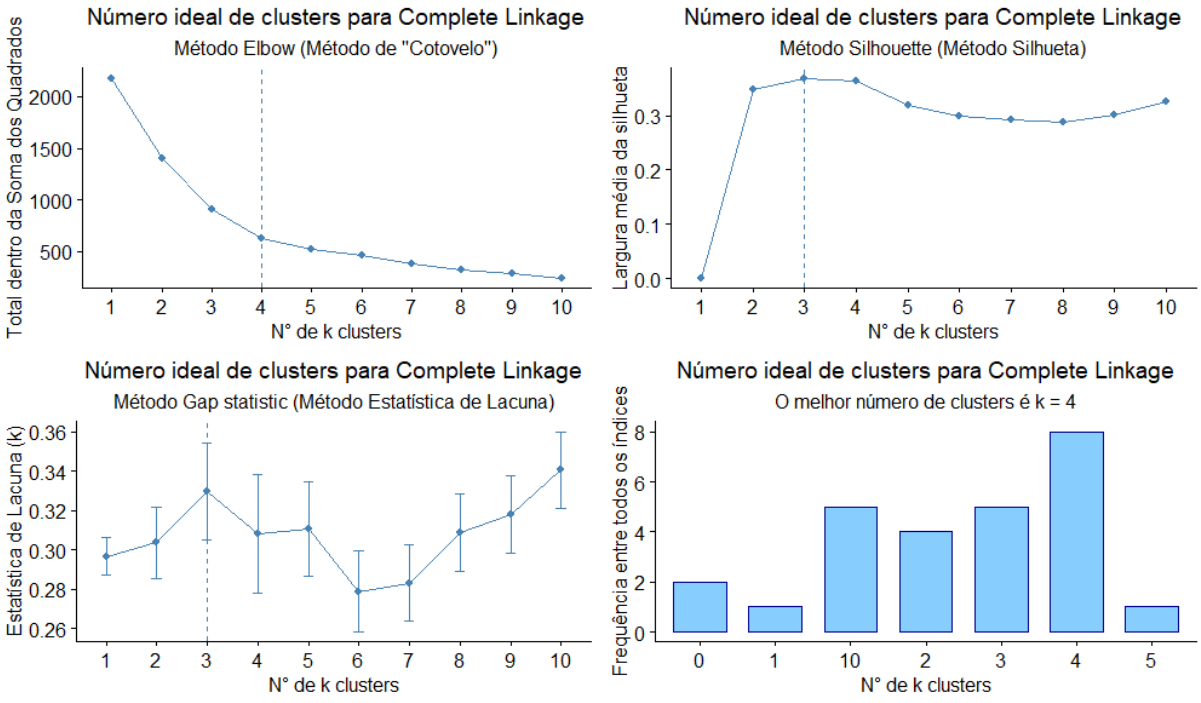


Figura 22: Verificando n° ideal de *clusters*

Com base na Figura 22, o número ideal de *clusters* a ser utilizado é 4, como no método

k-means. No Apêndice 7, encontra-se o cenário utilizando $k = 3$.

Novamente, aplica-se as funções *eclust*, *fviz_dend* e *fviz_silhouette*, para calcular as distâncias entre os elementos, a criação do dendrograma, e análise de silhueta.

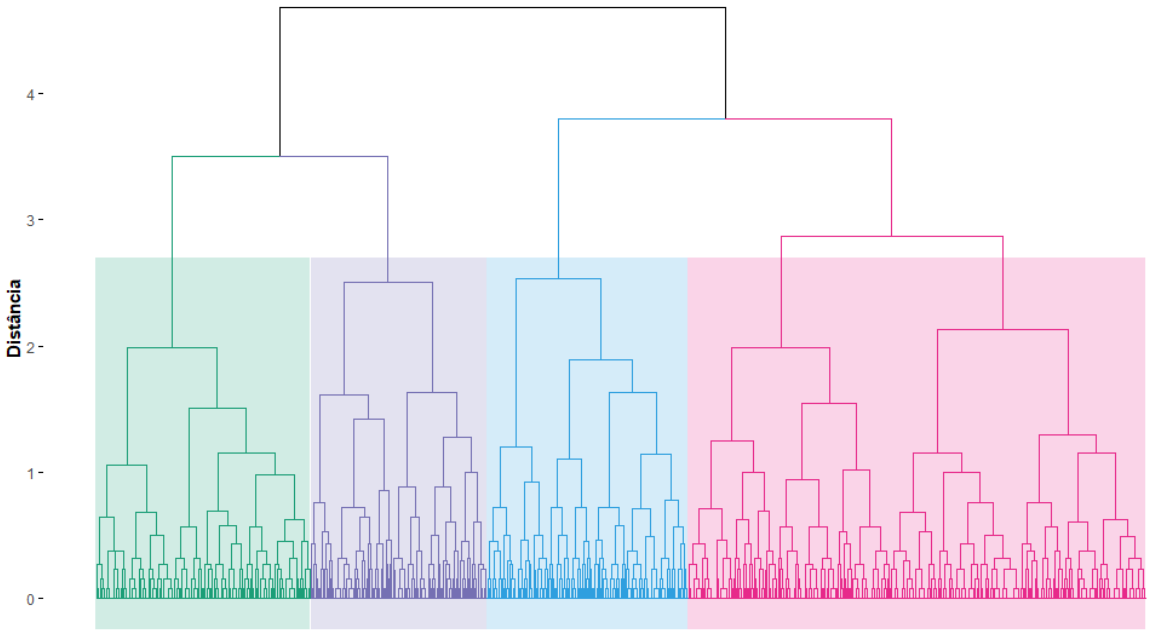


Figura 23: Dendrograma complete linkage

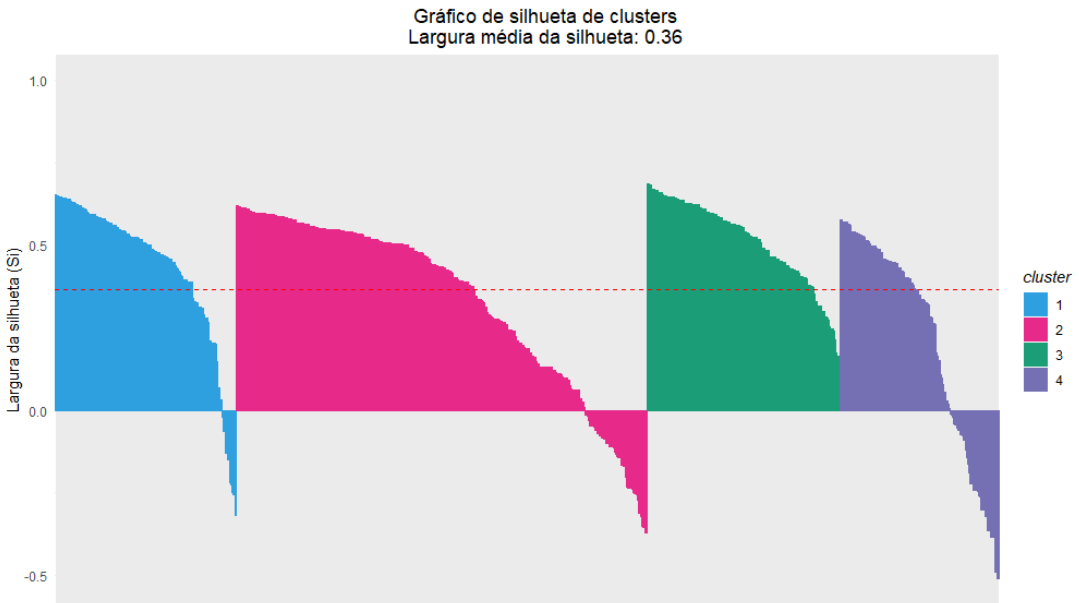


Figura 24: O coeficiente de silhueta de observações

Comparado ao o que aconteceu no método *k-means*, o gráfico de silhueta da Figura 24

mostra que o agrupamento feito pelo método *complete linkage* não foi tão eficiente, com diversas observações de Si negativo.

Tabela 15: N° de elementos e silhueta média de cada *cluster*

cluster	N° de elementos	Largura média da silhueta
1	209	0,43
2	475	0,33
3	223	0,51
4	183	0,21

Tabela 16: Número de observações agrupadas no *cluster* errado

cluster	*cluster* vizinho	Quantidade de observações
1	4	16
2	1	43
	3	29
4	1	10
	3	46

Pela Tabela 16, observa-se que 144 observações provavelmente estão no *cluster* errado, sendo o *cluster* 2 o que apresenta a maior quantidade de elementos que deviam ser de outro grupo, com 72 observações.

Pela Figura 23, pode-se ver que o *cluster* indicado pela cor rosa é o que mais possuí elementos, confirmado pela Tabela 15, com 475 elementos. Como visto na Figura 14, quanto mais observações uma base de dados possuir, maior será a dificuldade de compreensão do dendrograma. Neste caso em específico, a situação fica pior, pois esta base contém 1090 observações.

Para descobrir os padrões destes novos grupos, a Tabela 17 apresenta as médias das variáveis de cada *cluster*:

Tabela 17: Média das variáveis de cada *cluster*

cluster	Income	Spending
1	19.914	17.172
2	16.217	8.819
3	39.009	7.596
4	42.612	15.169

O *cluster* 1 é caracterizado pelos indivíduos que apresentam ganhos de aproximadamente 20 mil ao mês, e gastos um pouco maiores de 17 mil. O segundo grupo ganha em média 16 mil, e possuí gastos de quase 9 mil. Já os *clusters* 3 e 4, como no método *k-means*, são os que mais arrecadam. O grupo 3 arrecada um pouco mais de 39 mil em média, mas são os menos gastam com relação ao que arrecadam, gastando aproximadamente 1/5 do que arrecadam (quase 7.6 mil), e o grupo 4 gasta em média um pouco mais de 15 mil, mas é o grupo que mais arrecada, mais de 42 mil. Os *clusters* 3 e 4 do método *complete linkage* são bem similares aos encontrados pelo método *k-means*.

E para observar como se comporta as variáveis dentro de cada *cluster*, as frequências absoluta e relativa são expostas na Figura 25.

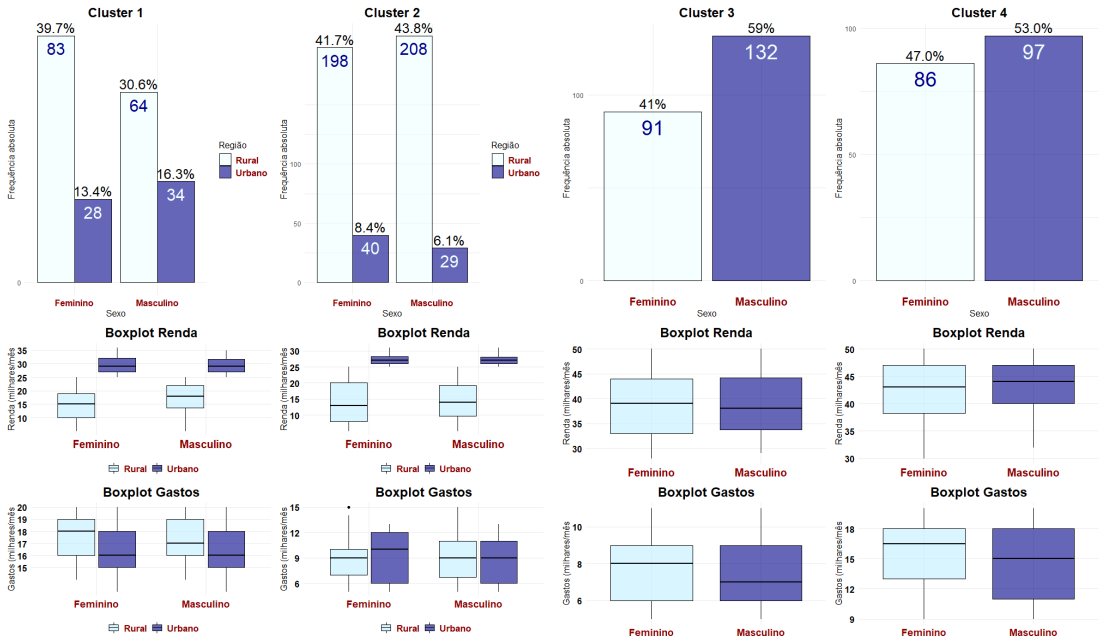


Figura 25: Estatísticas descritivas dos 4 clusters

O *cluster* 1 contém 209 pessoas: 111 mulheres (83 do meio rural e 28 da regiao urbana) e 98 homens (64 da zona rural e 34 da zona urbana), sendo caracterizados por arrecadar um pouco mais do que gastam, e sendo predominantes no meio rural. O *cluster* 2 contém 238 mulheres, sendo apenas 40 de zonas urbanas, e 237 homens, dos quais apenas 29 são das cidades. Este grupo é praticamente o mesmo número de mulheres e homens, e pouquíssimas pessoas de zonas urbanas, e são aqueles que gastam um pouco mais da metade do que arrecadam.

Já os *clusters* 3 e 4 só contém indivíduos das zonas urbanas, os que mais arrecadam, mais que o dobro dos dois grupos anteriores, e do *cluster* que menos gasta, o grupo 3, o número de homens é bem maior que o de mulheres, 132 e 91, respectivamente. E o

grupo 4, caracterizado por ser o que mais arrecada de todos os grupos, há também um predomínio maior de homens, 97 contra 86 das mulheres. Pode-se afirmar que os dados mostram uma triste realidade que ainda predomina na sociedade: majoritariamente, os homens ganham mais do que as mulheres, e por mais que estes dados não representem um cidadão comum, pois se trata de milhões de dólares de arrecadação, ainda assim mostra que os homens arrecadam mais que as mulheres.

De uma forma geral, os métodos *k-means* e *complete linkage* trouxeram resultados bem semelhantes para esta base de dados, mesmo o método hierárquico ter apresentado um número bem expressivo de observações que deveriam estar contidas no *cluster* vizinho. No entanto, principalmente pelo fato de terem tido o mesmo número k de partição inicial, os 4 grupos formados em cada método tiveram relações bem semelhantes se tratando de renda e gastos, e o número de componentes em cada *cluster*, de acordo com o padrão que cada grupo apresenta (ex: mais arrecadam, menos gastam), foram bem parecidos.

4 Conclusões

Um dos objetivos deste trabalho era de compreender o comportamento dos *clusters* gerados a partir dos métodos de *clustering*, e neste quesito, os dois métodos foram bem satisfatórios.

Em ambas bases de dados, tanto o método *k-means* como o método hierárquico *complete linkage*, conseguem fornecer as características de cada grupo, os elementos pertencentes a seu respectivo *cluster* (e os elementos que deveriam pertencer a outro *cluster*), gerando possibilidades de diversas análises estatísticas, sejam simples estatísticas descritivas como estudos mais aprofundados, como testes estatísticos ou modelos.

Com relação aos métodos hierárquicos, a impressão que se tem é de que ele seja muito mais apropriado a banco de dados pequenos, com no máximo 50 observações, pois a visualização de um dendrograma se torna muito confusa, muito rasurada. As bases utilizadas apresentam 167 e 1090 elementos, respectivamente, e mesmo 167 observações sendo uma quantidade razoável para uma base, sua compreensão visual se torna bem difícil. Uma opção, ou alternativa, para uma melhor visualização gráfica pode ser um sorteio de uma ou mais amostras de n elementos da base, para se obter um dendrograma de melhor compreensão, no entanto, muita informação pode acabar continuando oculta, ou se perder. A outra característica notável deste método é o fato de agrupar as observações com menos precisão, dada a diferença do número de elementos que foram agrupados no *cluster* errado pelo método hierárquico comparado aos números do método *k-means*.

Algo a ser destacado é o fato de que os métodos oferecem resultados bem semelhantes, mesmo quando cada método apresenta composições diferentes ou é aplicado com um número de *clusters* diferentes, como foi feito na primeira base de dados, onde o método *k-means* foi aplicado utilizando $k = 3$, e no método hierárquico foi utilizado $k = 5$. Pegando como exemplo a mesma base, em ambos os métodos de *clustering*, foi criado um grupo que continha os países mais desenvolvidos, e, mesmo os *clusters* apresentando comportamentos iguais, que no caso, seria de países que não precisariam de ajuda, os países que formavam o

cluster x do método *k-means* (representando os países ricos), não necessariamente seriam todos os mesmos países que formariam o *cluster y* do método hierárquico representando o grupo rico, no entanto, ainda assim apresentam caracteríticas muito semelhantes.

Portanto, concluí-se que ambos os métodos são bem satisfatórios para descobrir informações que sejam desconhecidas dentro de uma base de dados, ou até mesmo confirmar informações que não estariam inclusas na base, mas que de certa forma, elas já fossem esperadas de forma intuitiva. Pegando novamente como exemplo o caso dos países, em destaque os que mais precisam de ajuda, é intuitivo pensar que os países africanos são os que mais necessitam de ajuda, pois sabe-se de toda a pobreza e sofrimento que aqueles habitantes passam, e através das informações fornecidas pelos algoritmos, obtém-se a confirmação disto através de números, fazendo com que toda medida a ser tomada seja baseada em estatísticas, em fatos que comprovam as intuições pré-existentes.

Referências

[1] Clustering validation statistics: 4 vital things everyone should know - unsupervised machine learning. Disponível em: http://www.sthda.com/english/wiki/wiki.php?id_contents=7952.

[2] Hybrid hierarchical k-means clustering for optimizing clustering outputs - unsupervised machine learning. Disponível em: http://http://www.sthda.com/english/wiki/wiki.php?id_contents=8098#example-of-hierarchical-clustering.

[3] Nbclust: Nbclust package for determining the best number of clusters. Disponível em: https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust.

[4] Unsupervised machine learning. 2021. Disponível em: https://www.javatpoint.com/unsupervised-machine-learning.

[5] What is data labeling for machine learning? 2021. Disponível em: https://aws.amazon.com/pt/sagemaker/groundtruth/what-is-data-labeling/.

[6] Distance between two clusters for three hierarchical methods. 29 abr. 2021. Disponível em: https://www.researchgate.net/figure/Distance-between-two-clusters-for-three-hierarchical-methods-a-/single-linkage-b_fig1_228529077.

[7] *Entenda o aprendizado não supervisionado no Machine Learning*. ALIGER, 30 julho 2019. Disponível em: https://www.aliger.com.br/blog/machine-learning-entenda-o-que-e-aprendizado-nao-supervisionado/.

[8] Pedro Barros. *Aprendizagem de Maquina: Supervisionada ou Não Supervisionada?* Opensanca, 7 abril 2016. Disponível em: https://medium.com/opensanca/aprendizagem-de-maquina-supervisionada-ou-n%C3%A3o-supervisionada-7d01f78cd80a.

[9] Hugo Honda, Matheus Facure, and Peng Yaohao. *Os Três Tipos de Aprendizado de Máquina*. LAMFO, 27 julho 2017. Disponível em: https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/.

[10] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

[11] kassambara. Determining the optimal number of clusters: 3 must know methods. 2017. Disponível em: http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determiningthe-optimal-number-of-clusters-3-must-know-methods/.

[12] Manish Kumar. Humanitarian aid to underdeveloped countries. Disponível em: `https://www.kaggle.com/hellbuoy/pca-kmeans-hierarchical-clustering`, 06 novembro 2019.

[13] Wlademir Ribeiro Prates. *Aprendizado de máquina: supervisionado e não supervisionado*. CIÊNCIA E NEGÓCIOS, 15 set 2018.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

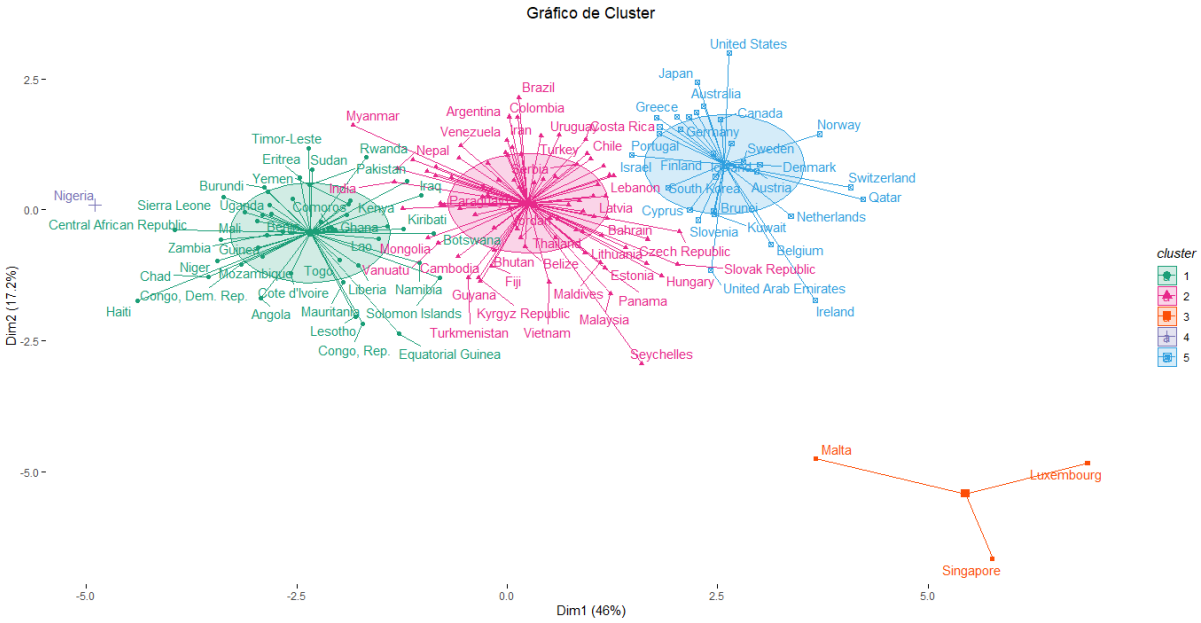
[15] Christian P. Robert and George Casella. *Introducing Monte Carlo Methods with R (Use R)*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2009.

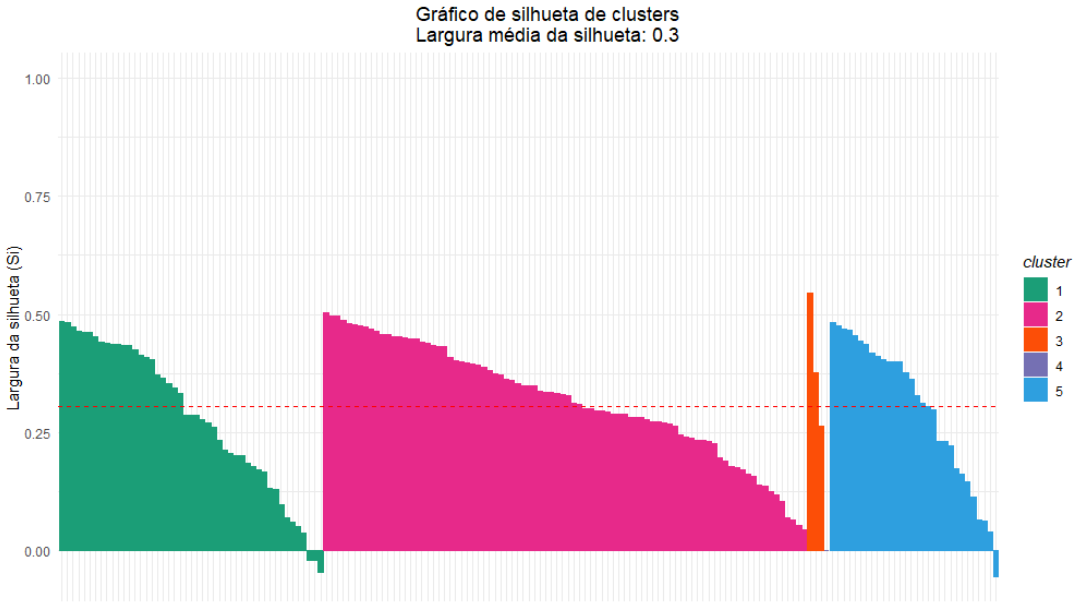
[16] Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach. 2 ed, 2003.

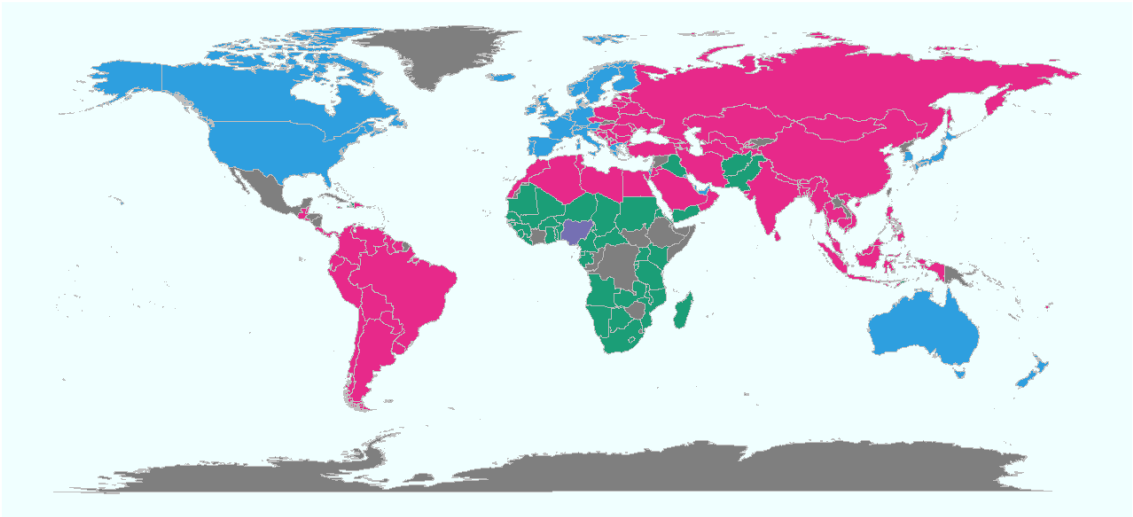
[17] Harriman Samuel Saragih. Simple clustering data id gender income spending. Disponível em: `https://www.kaggle.com/harrimansaragih/clustering-data-id-gender-income-spending`, 25 março 2021.

[18] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society Series B*, 63 (2):411–423, 2001.

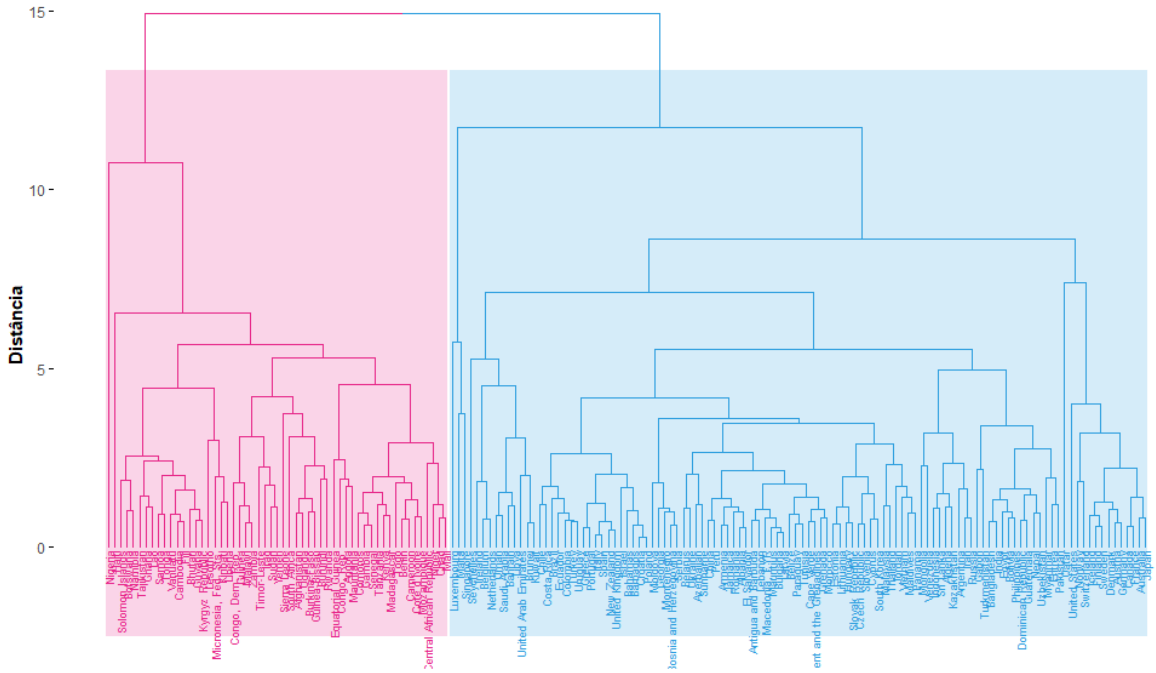
APÊNDICE 1 – Base 1: Método *k-means* com $k = 5$






cluster
1
2
3
4
5

APÊNDICE 2 – Base 1: Método *complete linkage* com $k = 2$



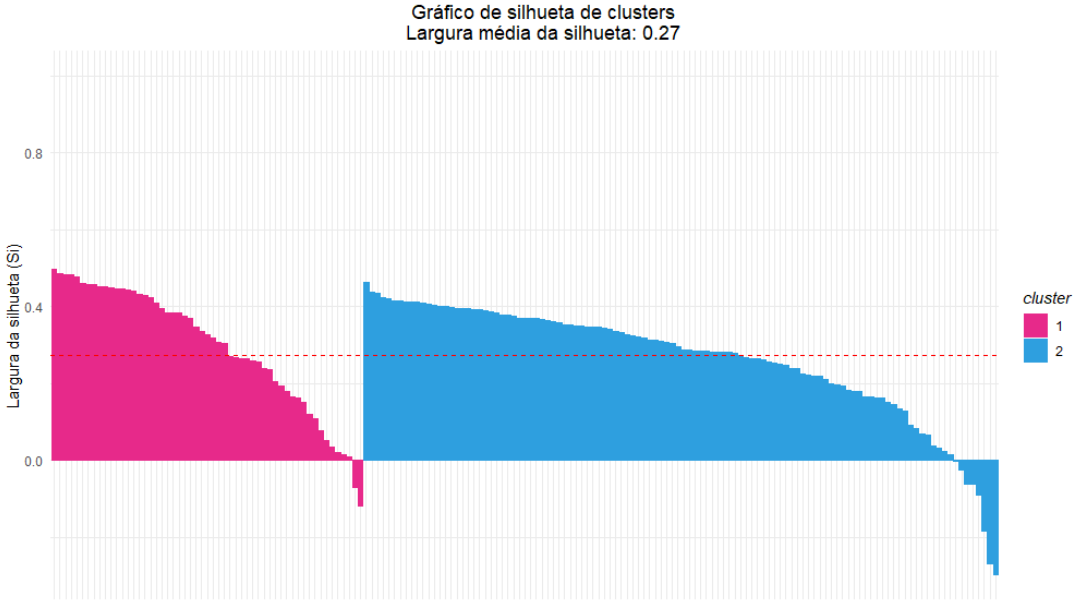
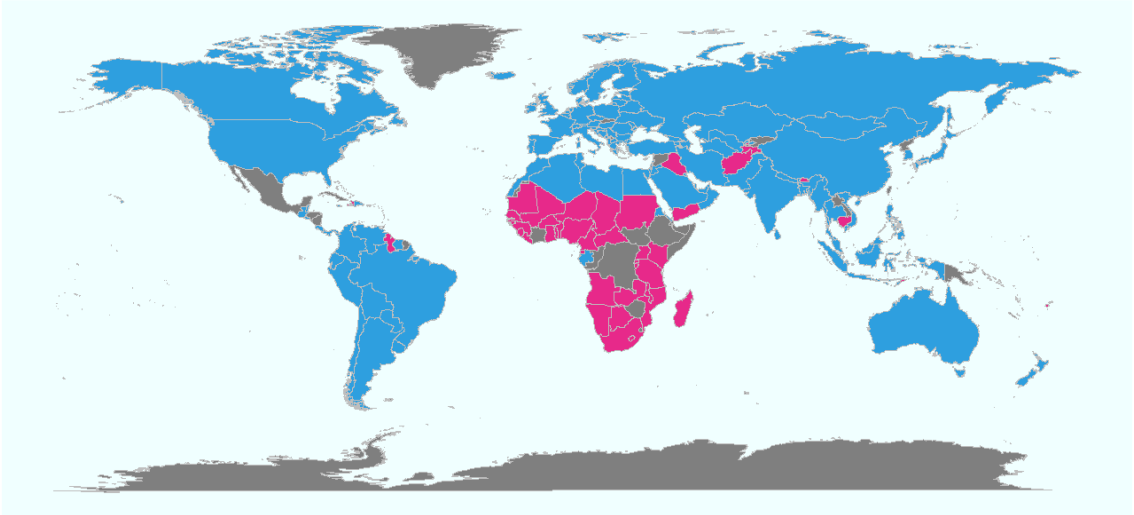
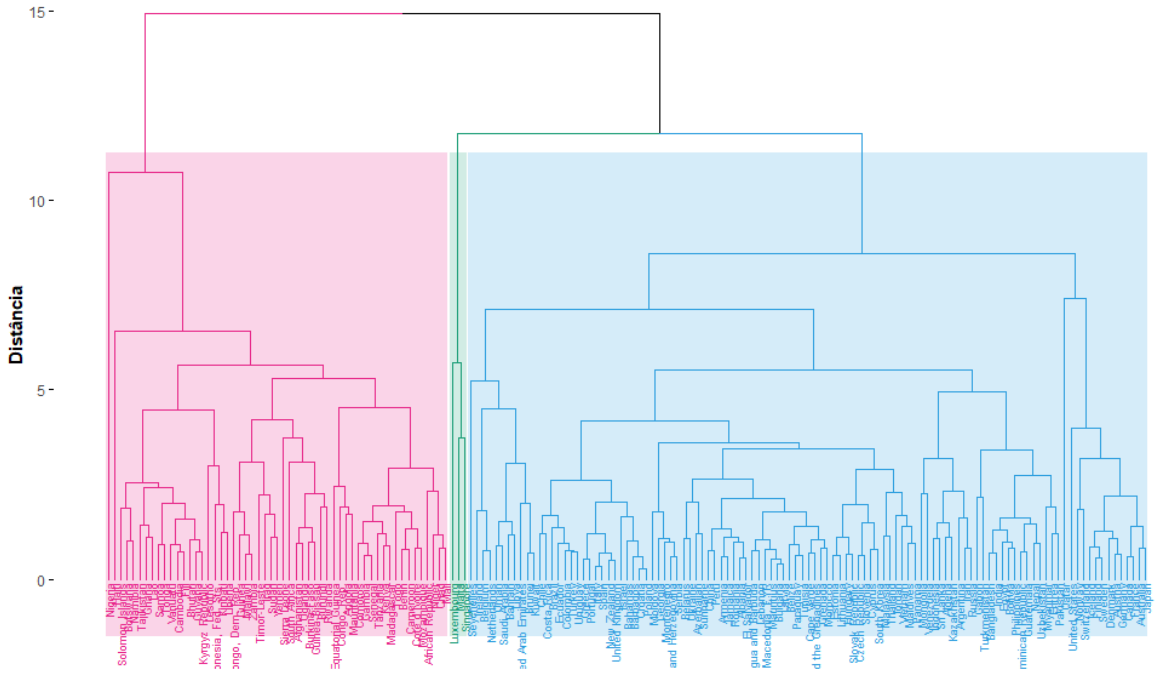


Gráfico de silhueta de clusters
Largura média da silhueta: 0.27
Largura da silhueta (Si)
0.8
0.4
0.0
cluster
1
2


cluster
1
2

APÊNDICE 3 – Base 1: Método *complete linkage* com $k = 3$



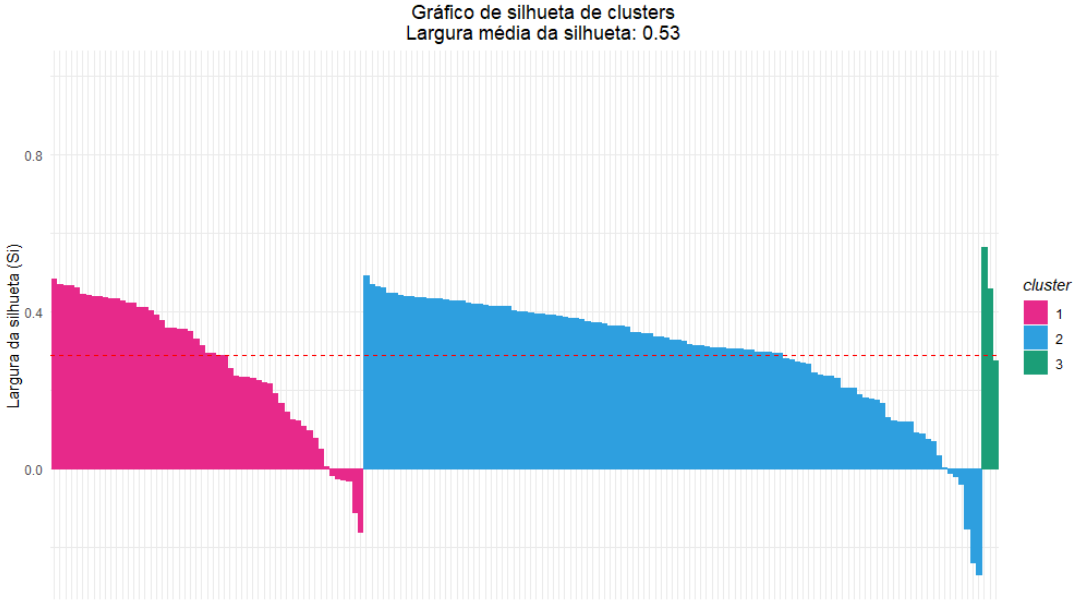
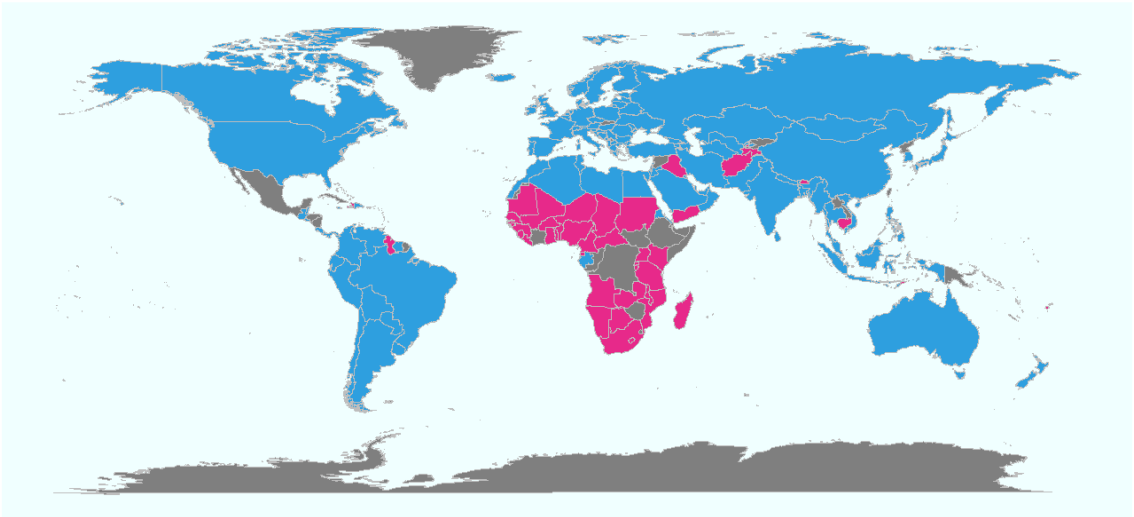
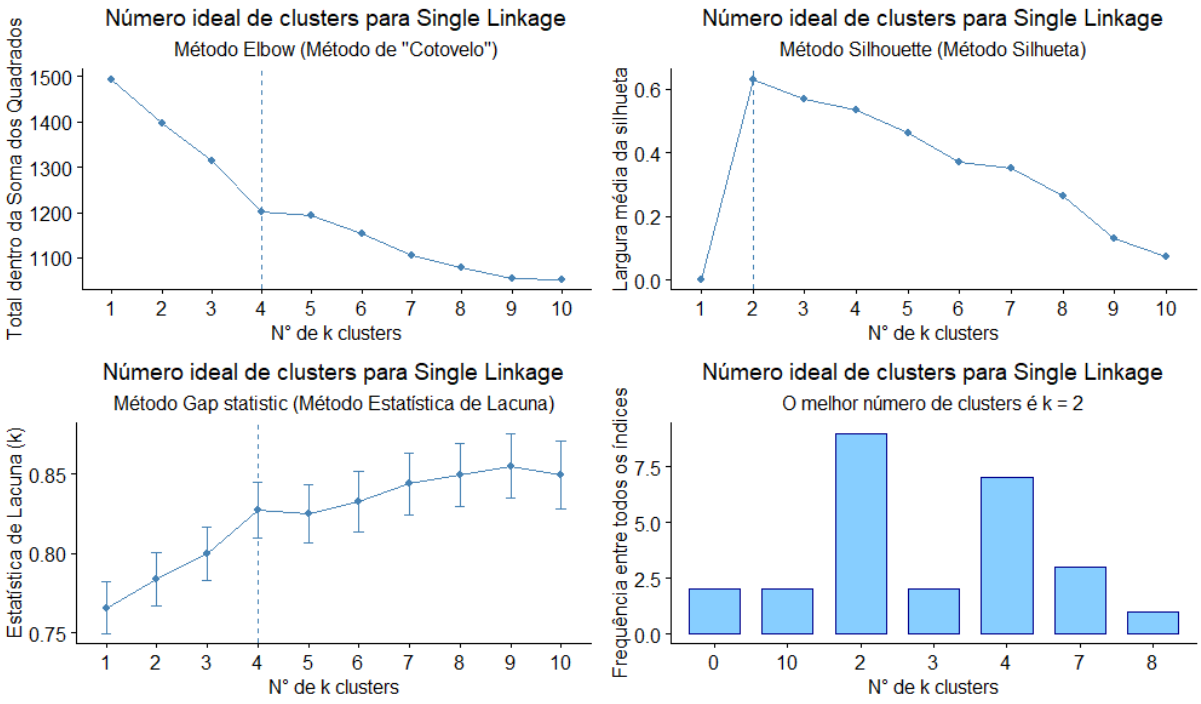


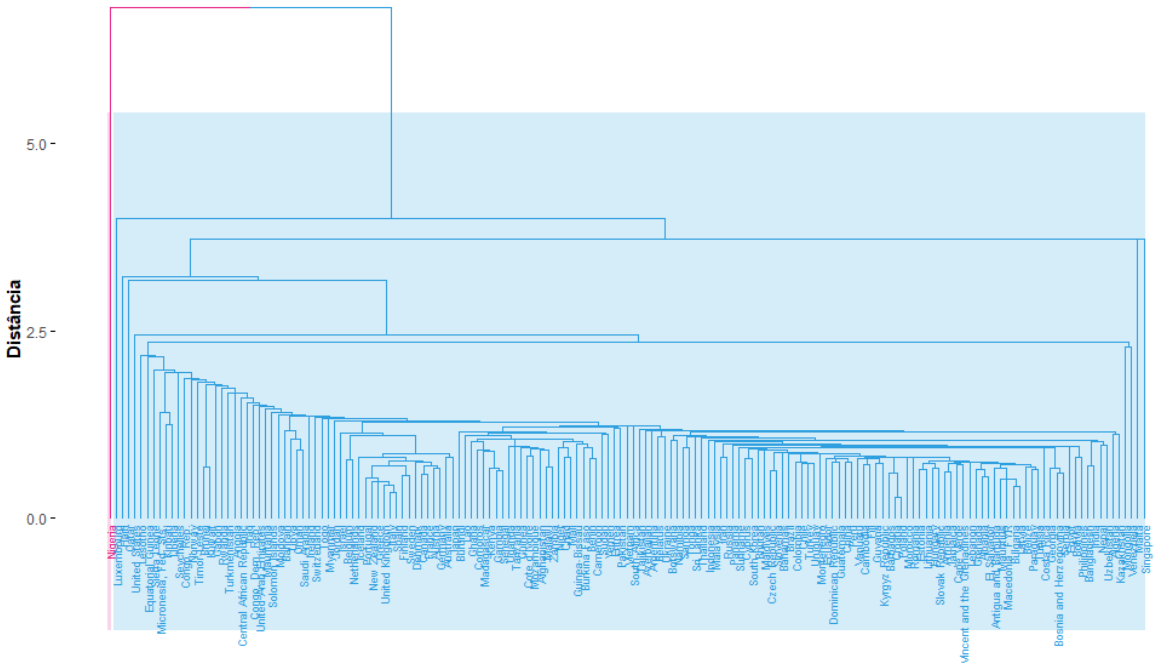
Gráfico de silhueta de clusters
Largura média da silhueta: 0.53
Largura da silhueta (Si)
0.8
0.4
0.0
cluster
1
2
3


cluster
1
2
3

APÊNDICE 4 – Base 1: Método *single linkage*



4.1 $k = 2$


Distância
5.0
2.5
0.0

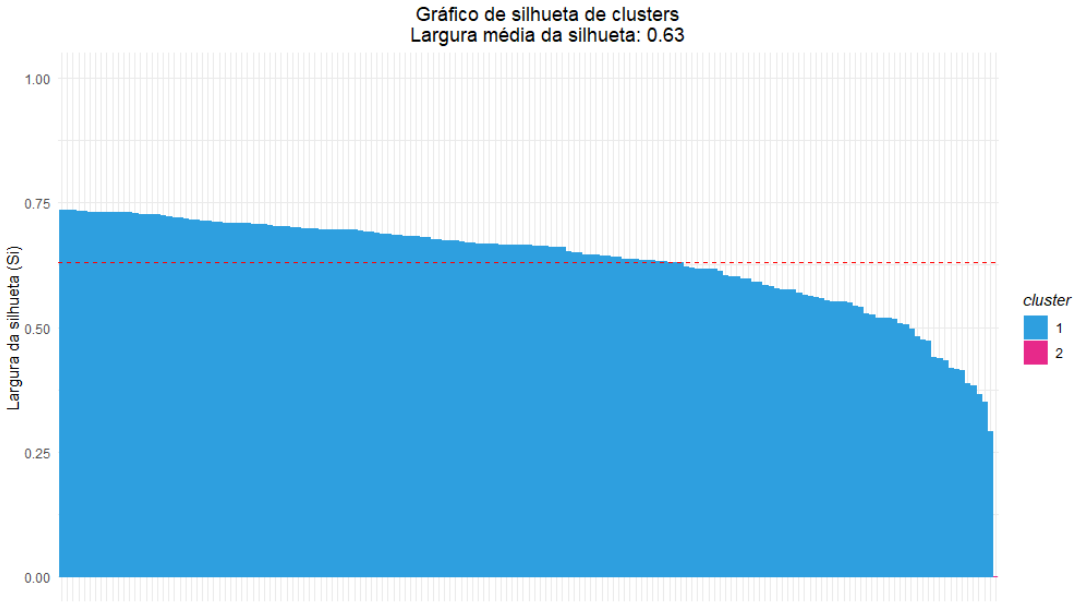
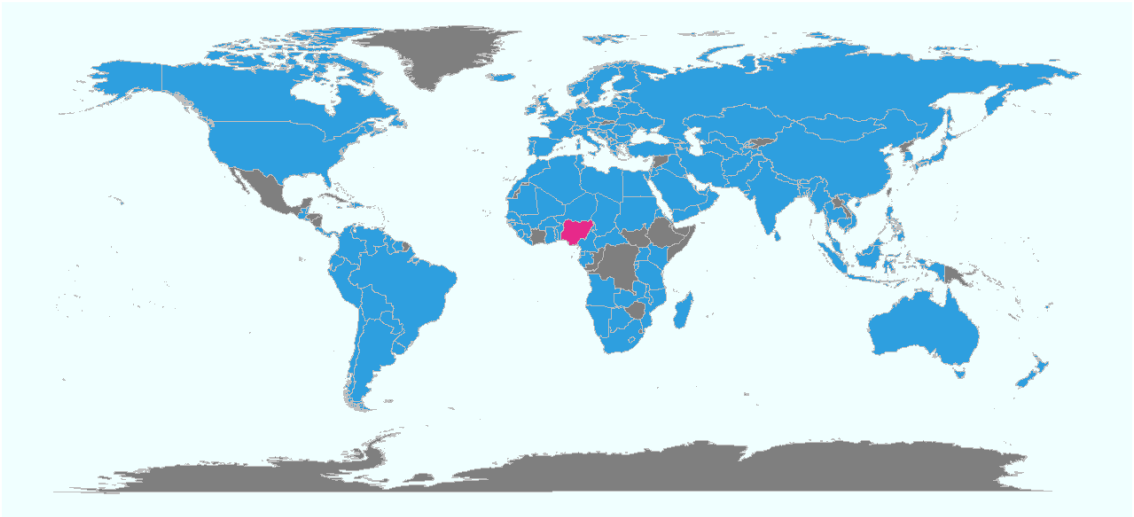
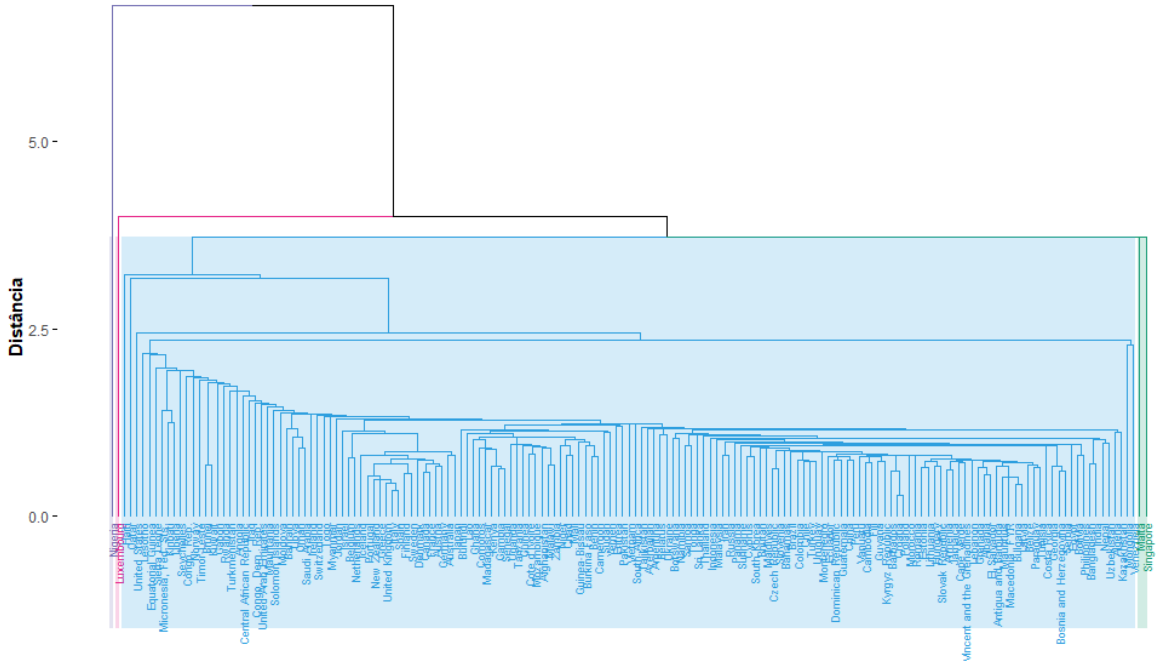


Gráfico de silhueta de clusters
Largura média da silhueta: 0.63
Largura da silhueta (Si)
1.00
0.75
0.50
0.25
0.00
cluster
1
2


cluster
1
2

4.2 $k = 4$


Distância
5.0
2.5
0.0

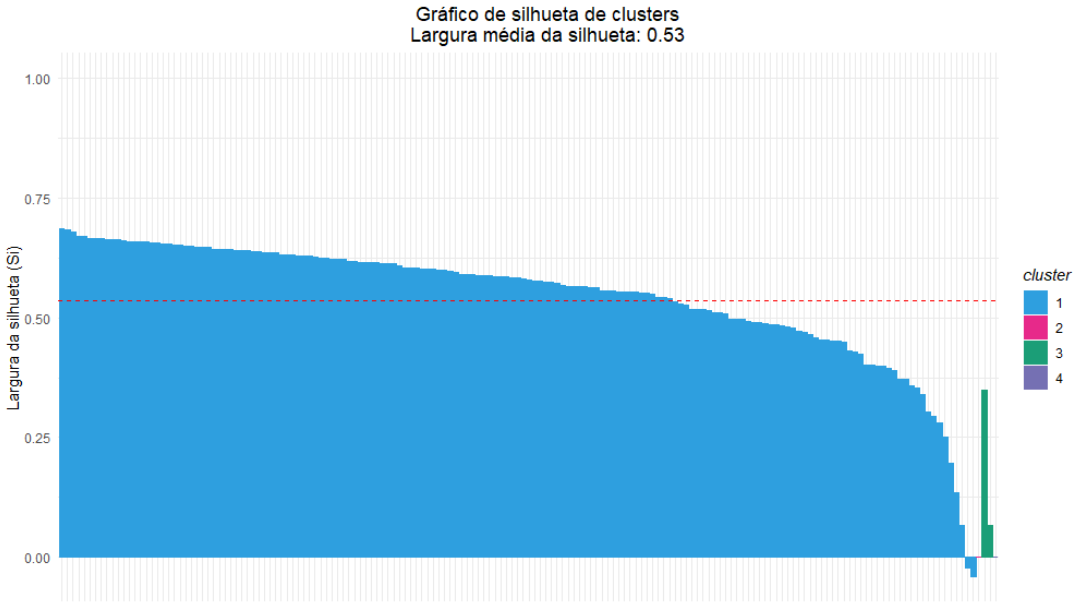
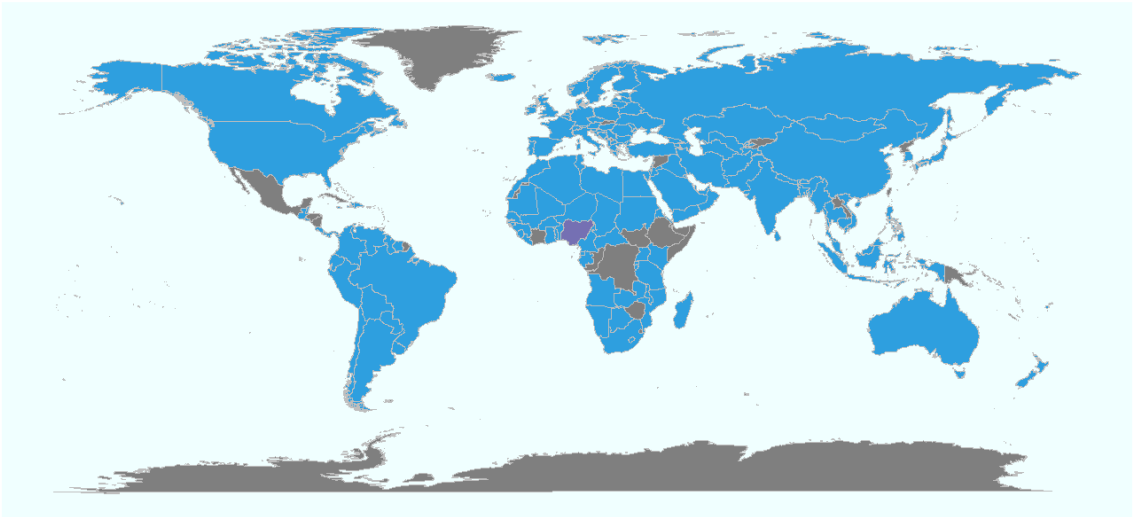
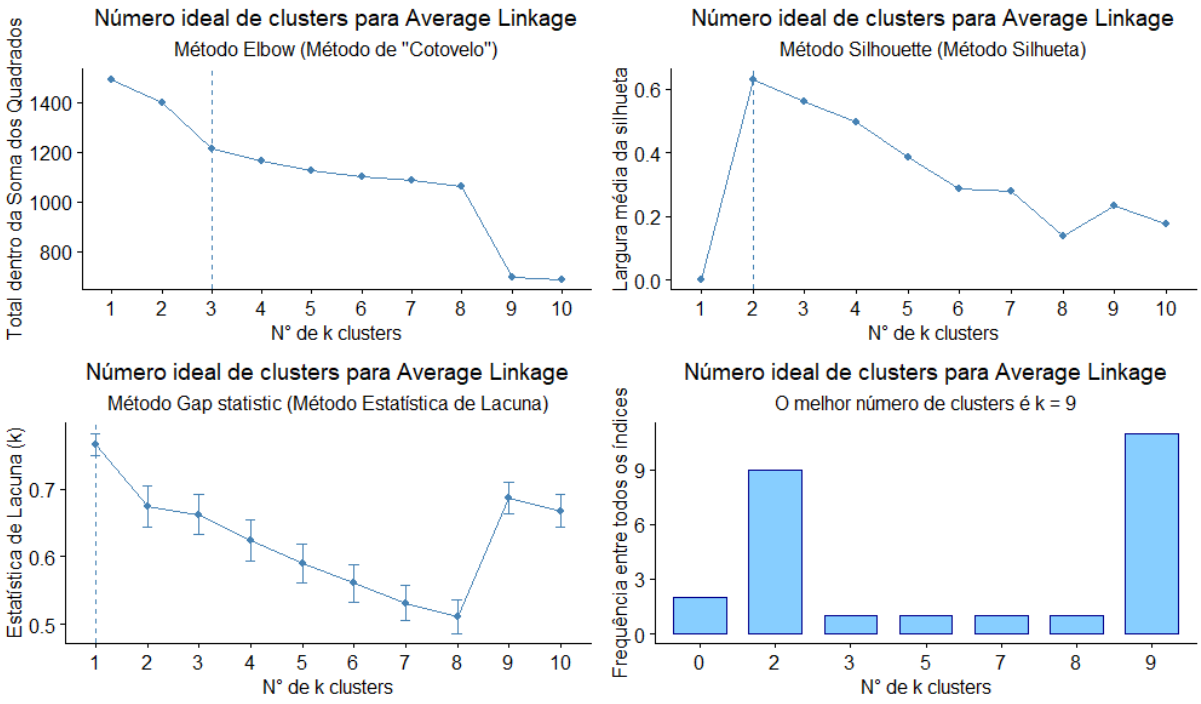


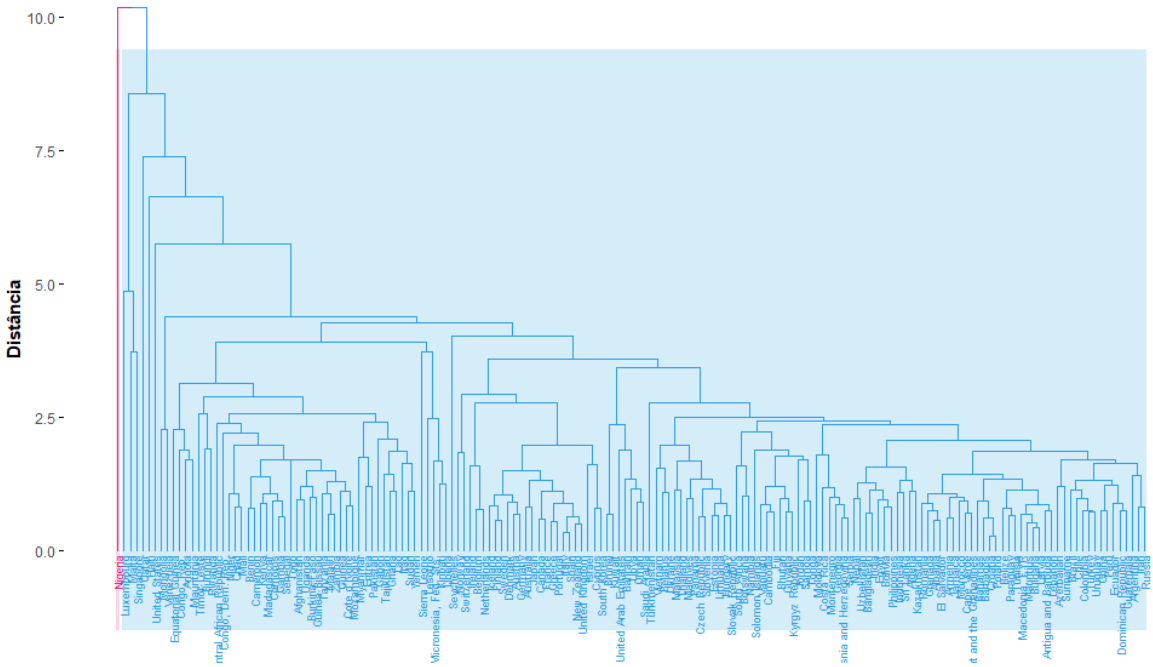
Gráfico de silhueta de clusters
Largura média da silhueta: 0.53
Largura da silhueta (Si)
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
4


cluster
1
2
3
4

APÊNDICE 5 – Base 1: Método *average linkage*



5.1 $k = 2$


Distância
10.0
7.5
5.0
2.5
0.0

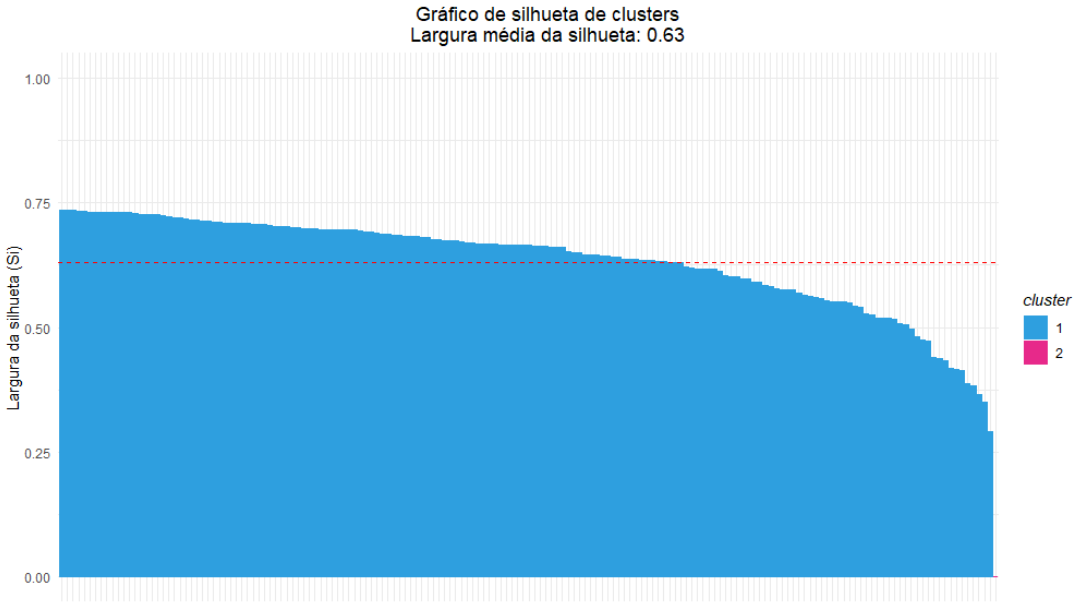
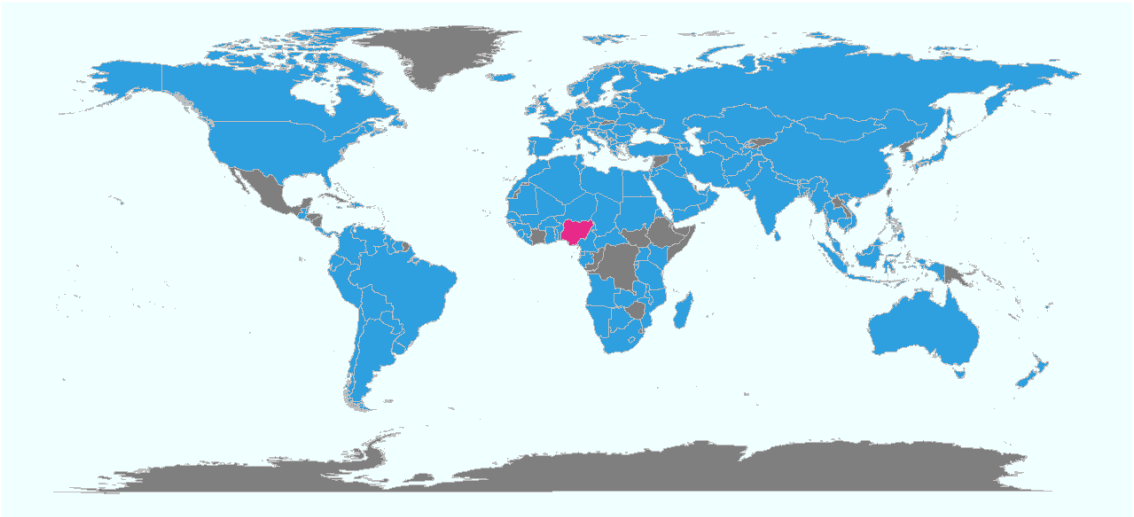
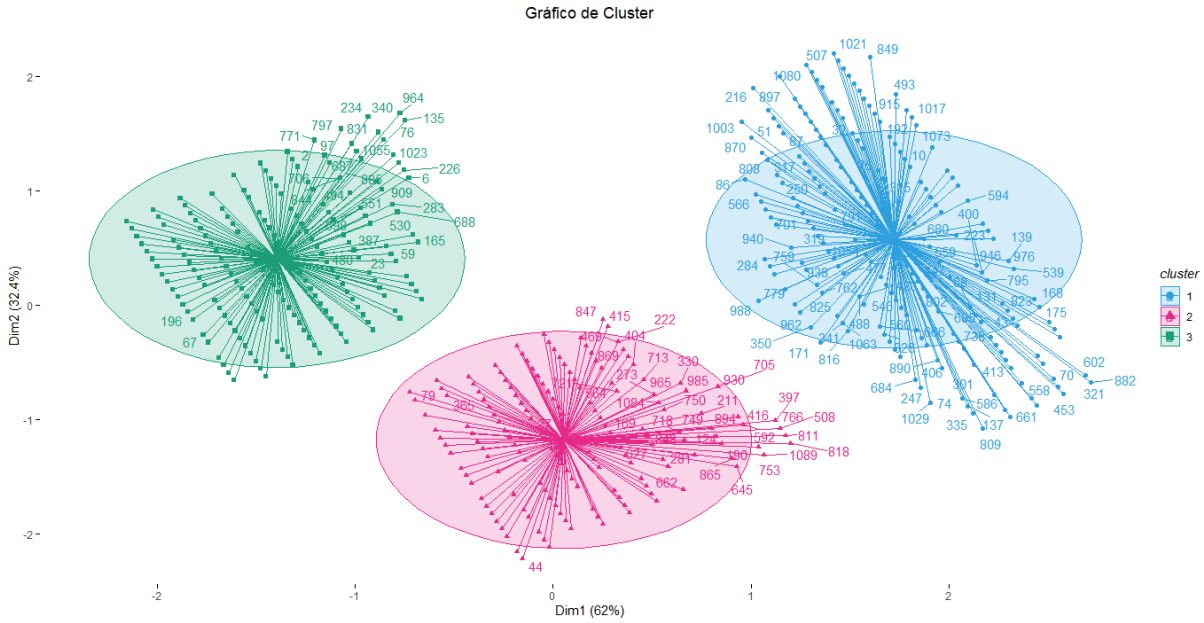
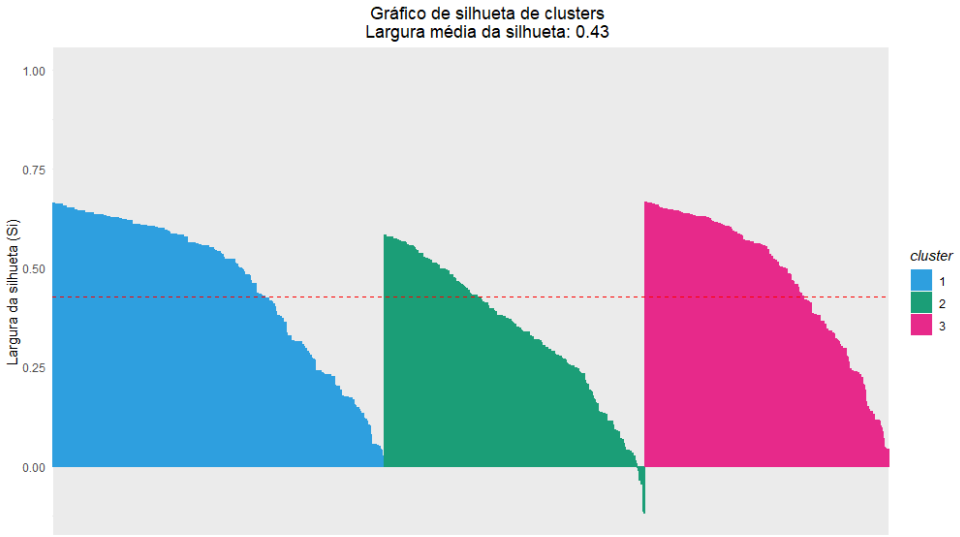


Gráfico de silhueta de clusters
Largura média da silhueta: 0.63
Largura da silhueta (Si)
1.00
0.75
0.50
0.25
0.00
cluster
1
2


cluster
1
2

APÊNDICE 6 – Base 2: Método *k-means* com $k = 3$





APÊNDICE 7 – Base 2: Método *complete linkage* com $k = 3$