ISSN 0103-7811
Novembro / 2020

DOCUMENTOS

133

Análise de *cluster* não supervisionado em R: agrupamento hierárquico

Embrapa

ISSN 0103-7811
Novembro/2020

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Territorial
Ministério da Agricultura, Pecuária e Abastecimento

DOCUMENTOS 133

Análise de *cluster* não supervisionado em R: agrupamento hierárquico

Rogério Resende Martins Ferreira
Fernando Antônio de Pádua Paim
Valéria Guimarães Silvestre Rodrigues
Gustavo Spadotti Amaral Castro

Embrapa Territorial
Campinas, SP
2020

Exemplares desta publicação podem ser adquiridos na:

Embrapa Territorial
Av. Soldado Passarinho, nº 303
Fazenda Chapadão
13070-115, Campinas, SP
Fone: (19) 3211.6200
www.embrapa.br/territorial
www.embrapa.br/fale-conosco/sac

Supervisão editorial
Suzi Carneiro e Bibiana Teixeira de Almeida

Revisão de texto
Bibiana Teixeira de Almeida

Normalização bibliográfica
Vera Viana dos Santos Brandão

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica e tratamento das ilustrações
Suzi Carneiro

Ilustração da capa
Designed by Creativeart / Freepik

1ª edição
1ª impressão (2020): versão on-line

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Territorial

Análise de *cluster* não supervisionado em R: agrupamento hierárquico /
Rogério Resende Martins Ferreira, Fernando Antônio de Pádua Paim, Valéria Guimarães Silvestre Rodrigues, Gustavo Spadotti Amaral Castro. - Campinas: Embrapa Territorial, 2020.
43 p.: il. ; (Documentos / Embrapa Territorial, ISSN 0103-7811; 133).

1. Análise de agrupamento. 2. Análise de dados. 3. Software R. I. Ferreira, Rogério Resende Martins. II. Paim, Fernando Antônio de Pádua. III. Rodrigues, Valéria Guimarães Silvestre. IV. Castro, Gustavo Spadotti Amaral. V. Título. VI. Série.

CDD 519.53

Vera Viana dos Santos Brandão (CRB 8/7283)

Autores

Rogério Resende Martins Ferreira
Agrônomo, Doutor em Agronomia, pesquisador da Embrapa Territorial, Campinas, SP

Fernando Antônio de Pádua Paim
Analista de Sistemas, Especialista em Análise de Sistemas, analista da Embrapa Territorial, Campinas, SP

Valéria Guimarães Silvestre Rodrigues
Geóloga, Doutora em Geociências, professora e pesquisadora na Universidade de São Paulo, São Carlos, SP

Gustavo Spadotti Amaral Castro
Engenheiro-agrônomo, Doutor em Agricultura, analista da Embrapa Territorial, Campinas, SP

Apresentação

Ao analisar uma base de dados, um dos principais desafios do analista é resumir a informação coletada. Em muitos casos, quando contamos com um grande número de observações, pode ser de interesse criar grupos. Dentro de cada grupo os elementos devem ser semelhantes entre si e diferentes dos elementos dentro dos outros grupos.

A análise de *clusters* é um procedimento da estatística multivariada que tenta agrupar um conjunto de dados em subgrupos homogêneos, chamados de agrupamentos. É uma técnica matemática que tem como finalidade revelar estruturas de classificação nos dados do mundo real. Os métodos hierárquicos da análise de *cluster* têm como principal característica um algoritmo capaz de fornecer mais de um tipo de partição dos dados. Ele gera vários agrupamentos possíveis, e um *cluster* pode ser mesclado a outro em determinado passo do algoritmo.

A maioria dos ambientes e softwares de análise estatística dispõem de opções para fazer análise de *cluster* e construção de dendrogramas. O software R conta com uma grande quantidade de funções e pacotes de trabalho para análise de agrupamento.

É nesse contexto que esta publicação se insere, ao descrever os principais conceitos para a aplicação de procedimentos estatísticos de análise não supervisionada que objetivam produzir agrupamentos hierárquicos com base na semelhança ou dissemelhança entre os objetos de estudo. O leitor perceberá a dificuldade associada ao processo e entenderá os usos principais da técnica.

Evaristo Eduardo de Miranda
Chefe-Geral da Embrapa Territorial

Sumário

Preceitos do aprendizado de máquina: modo e estratégia

A área de inteligência artificial nasceu em 1956, em uma conferência em Dartmouth College, EUA. Na proposta dessa conferência, submetida à fundação Rockfeller, constava a intenção dos autores de fazerem um estudo cujo tema era inteligência artificial. Essa área de pesquisa passou pela construção de sistemas de resolução de problemas gerais, de sistemas especialistas, projetados para simular o comportamento humano na resolução de um problema de domínio específico, e pela elaboração de metodologias para aquisição e análise de conhecimento (Ernst; Newell, 1969). Mais recentemente, o processo de mineração de dados vem sendo amplamente aplicado, com objetivo de descobrir conhecimento novo em bases de dados de diversos domínios e áreas de aplicação. Esse processo de exploração de dados e descoberta de conhecimento utiliza, entre outros, técnicas e conceitos de aprendizado de máquina.

A definição de aprendizado de máquina é quando um programa de computador aprende a partir de uma experiência E, relacionada a uma tarefa T e com uma métrica de performance P, e a sua performance em T, medida por P, melhora com a experiência E (Mitchell, 1997). Os objetivos do aprendizado de máquina são o desenvolvimento de técnicas computacionais que permitem simular o processo de aprendizado e a construção de sistemas capazes de adquirir conhecimento de maneira automática. Esses sistemas utilizam o domínio e os resultados de experiências anteriores para auxiliar o processo de decisão e melhorar o seu desempenho futuro.

Usualmente, em processos de aprendizado, o aprendiz utiliza o conhecimento que adquire para obter novos conhecimentos, que podem ser aprendidos de diversas maneiras. Cinco estratégias de aprendizado podem ser enumeradas segundo o grau de complexidade de inferência: hábito, instrução, dedução, analogia e indução. A primeira estratégia apresenta menor complexidade de inferência, ao passo que a estratégia indutiva exige maior esforço para o aprendizado (Monard; Baranauskas, 2003).

O aprendizado por indução é caracterizado pelo raciocínio que parte do específico para o geral. É um modo de inferência lógica que permite obter generalizações a partir de exemplos, para induzir um conceito que pode ou não preservar a verdade. Assim, mesmo que as premissas sejam verdadeiras, pode-se chegar a conclusões falsas. Justamente por ser uma estratégia de aprendizado complexa, uma vez que o aprendiz desempenha a maior parte do esforço para a aquisição do conhecimento, ela permite que conceitos muito mais amplos possam ser aprendidos e, portanto, constitui uma das estratégias de aprendizado de maior interesse para pesquisas relacionadas ao aprendizado de máquina (Kuhn; Johnson, 2016; Wickham; Grolemund, 2019).

Além da estratégia, na construção de um algoritmo de aprendizado, deve-se considerar o modo de aprendizado. Sob esse aspecto, o aprendizado indutivo pode ser dividido em supervisionado e não supervisionado. O que distingue esses dois modos de aprendizado é a presença ou a ausência do atributo classe que rotula os exemplos do conjunto de dados (Michalski; Stepp, 1983). No aprendizado supervisionado, esse rótulo é conhecido, ao passo que, no aprendizado não supervisionado, os exemplos não estão previamente rotulados. Estudos mais recentes têm apresentado um terceiro modo de aprendizado, denominado semissupervisionado, no qual poucos exemplos apresentam-se rotulados. Esse fato impossibilita o uso direto de algoritmos de aprendizado supervisionado, pois o método requer um número razoável de exemplos rotulados (Blum; Mitchell, 1998).

No aprendizado não supervisionado, também conhecido como aprendizado por observação e descoberta ou análise exploratória de dados, o conjunto de dados de entrada é composto por exemplos não rotulados (não há a informação sobre a classe associada a cada exemplo). Nesse

caso, são utilizados algoritmos para descobrir padrões nos dados a partir de alguma caracterização de regularidade, e esses padrões são denominados *clusters* (Zumel; Mount, 2020). A tarefa consiste em agrupar uma coleção de exemplos segundo alguma medida de similaridade, de modo que exemplos pertencentes ao mesmo *cluster* devem ser mais similares entre si e menos similares aos exemplos que constituem outros *clusters*.

Análise de *cluster* não supervisionada e agrupamento hierárquico

A modelagem supervisionada envolve a descoberta de padrões para prever o valor de uma variável-alvo especificada. A modelagem não supervisionada não se concentra na variável-alvo, e sim na regularidade do conjunto de dados para estimá-la (Provost; Fawcett, 2016).

Agrupamento ou *cluster* significa formar grupos de pontos considerando a semelhança entre eles. A ideia básica é identificar grupos cujos objetos internos sejam semelhantes, ficando os objetos não semelhantes em grupos distintos.

Os agrupamentos hierárquicos são formados por nós, que posteriormente são mesclados de forma repetitiva até serem fundidos em um único nó por meio de uma função matemática de similaridade ou distância escolhida, conhecida como função de ligação, como a distância euclidiana, por exemplo.

A análise de *clusters* inclui uma série de procedimentos estatísticos que podem ser usados para classificar objetos observando apenas as semelhanças ou dissemelhanças entre eles, sem definir previamente critérios de inclusão em qualquer agrupamento. Além da estruturação dos dados em grupos e da consequente redução da dimensão do espaço associado às novas variáveis, a comparação das propriedades de um objeto qualquer com as propriedades dos elementos dos subgrupos permite identificar o subgrupo onde incluí-lo, uma vez que elementos pertencentes ao mesmo subgrupo têm propriedades semelhantes.

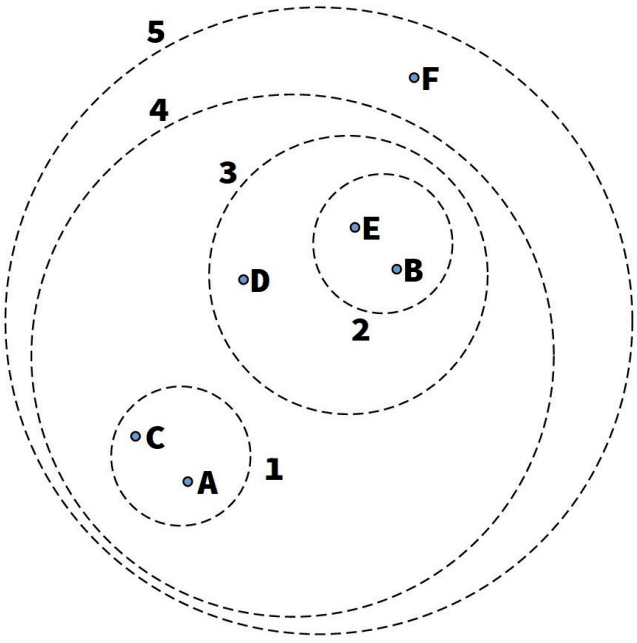

Figura 1. Seis pontos e seus possíveis agrupamentos formam uma hierarquia implícita (Provost; Fawcett, 2016).

Na Figura 1 são mostrados seis pontos, de A a F, dispostos em um plano (espaço bidimensional). Círculos rotulados de 1 a 5 são colocados sobre os pontos, para indicar os agrupamentos.

O diagrama mostra os aspectos essenciais do que é chamado de agrupamento hierárquico. Trata-se de um agrupamento porque há a formação de grupos de pontos por semelhança, e a única sobreposição entre os agrupamentos é quando um contém os demais. Devido à estrutura, os círculos representam uma hierarquia de agrupamentos, na qual o mais geral (nível mais alto – 5) é aquele que contém todos os

elementos, e o nível mais baixo corresponde aos próprios pontos. Remover os círculos em ordem decrescente produz uma coleção de diferentes agrupamentos, cada um com um número maior.

Usar a distância euclidiana (Figura 2) torna os pontos mais fortemente semelhantes, caso estejam mais próximos uns dos outros no plano. O gráfico em dendrograma mostra explicitamente a hierarquia dos agrupamentos. Ao longo do eixo x estão organizados (em nenhuma ordem particular, exceto para evitar cruzamento de linhas) os pontos de dados individuais. O eixo y representa a distância entre os agrupamentos. Na parte inferior (y = 0), cada ponto está em um agrupamento separado. Conforme y aumenta, diferentes agrupamentos caem dentro da restrição de distância: primeiro, A e C são agrupados, depois B e E são mesclados, em seguida, o agrupamento BE é mesclado com D, e assim por diante, até que todos os agrupamentos estejam mesclados no topo. Os números são junções dos dendrogramas e correspondem aos círculos numerados no diagrama superior.

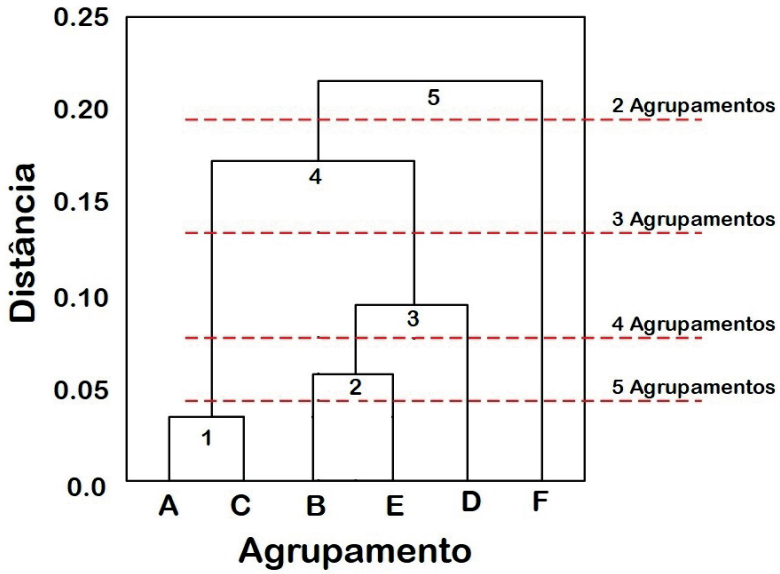

Figura 2. Dendrograma de seis pontos correspondendo aos agrupamentos com hierarquia explícita (Provost; Fawcett, 2016).

Cortar o dendrograma na linha identificada como “2 agrupamentos” resulta em dois grupos diferentes, neste caso o ponto F avulso e o grupo contendo todos os outros pontos (A, C, B, E, D, F). Esse corte no ponto “2 agrupamentos” corresponde à remoção do círculo 5. O corte na linha rotulada “3 agrupamentos” implica deixar três grupos abaixo da linha (AC, BED, F) no dendrograma e a remoção dos círculos 5 e 4. Os agrupamentos A e C formam um grupo fechado; os agrupamentos B, E e D formam outro agrupamento fechado e o agrupamento F continua destacando-se do resto do conjunto. Outro ponto a ser observado é que, uma vez que dois grupos são unidos em um determinado nível, eles permanecem assim nos níveis mais altos da hierarquia.

Como o eixo y representa a distância entre os agrupamentos, o dendrograma gera o subsídio para a linha de corte dos agrupamentos naturais. Observa-se no dendrograma da Figura 3 que há uma distância relativamente longa entre o agrupamento 3 (cerca de 0,10) e o agrupamento 4 (cerca de 0,17), sugerindo uma divisão aceitável. Além disso, nota-se que o ponto F se funde no ponto mais alto do dendrograma, o que pode indicar um “valor atípico” a ser investigado.

Ambos os exemplos mostram que o agrupamento hierárquico não apenas cria um “agrupamento” ou um único conjunto de grupos de objetos. Ele cria uma coleção de maneiras de agrupar os elementos. A quantidade de agrupamento pode ser analisada traçando-se uma linha horizontal através do dendrograma. Conforme a linha é movida para baixo, são obtidos diferentes agrupamentos com números crescentes de grupos, como mostra a Figura 3.

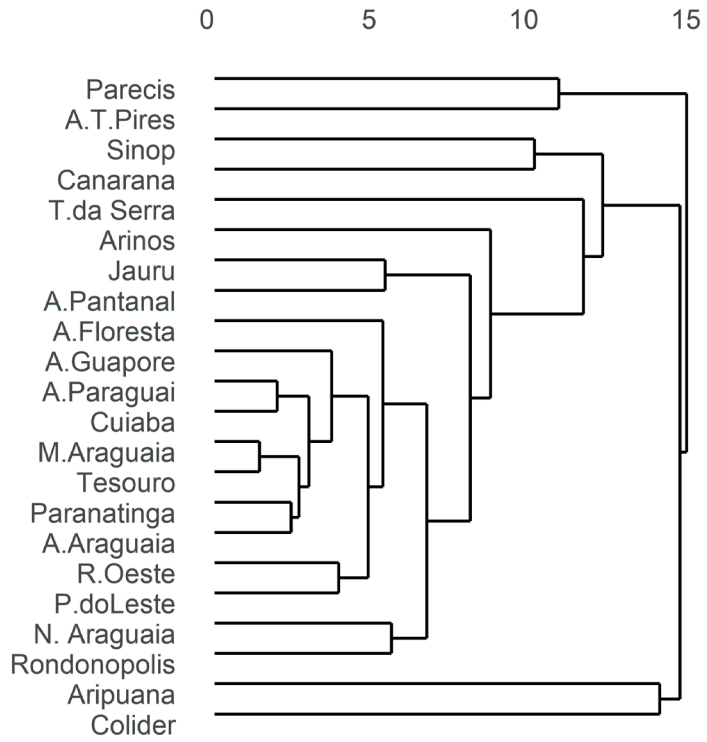

Figura 3. Exemplo de agrupamento hierárquico.

Seleção das Variáveis

A seleção das variáveis é um dos aspectos que influencia os resultados da análise de *cluster*. É necessária atenção no tipo de variáveis utilizadas e na escala.

Sobre o número de variáveis, existem opiniões divergentes: há relatos de que o aumento do número de variáveis resulta em melhor identificação dos *clusters* (Everitt, 1991) e, nas mesmas condições, relatos de fraca identificação dos *clusters* (Price, 1993).

Mas essas opiniões não são contrárias. As variáveis, nas caracterizações dos grupos, poderão ser satisfatórias para conjuntos específicos de dados, porém o uso de outros dados da mesma população e com as mesmas variáveis pode gerar um resultado da análise de *cluster* diferente. Por um lado, pode-se dizer que a identificação dos grupos é fraca. Por outro lado, com poucas variáveis é possível omitir informações importantes e obter poucos *clusters*, que serão interpolados de forma muito geral. Deve haver um número de variáveis que corresponde à situação mais “equilibrada” da análise de *clusters*, e que será robusta para outras bases de dados da mesma população.

A atribuição de peso às variáveis influencia a semelhança/dissemelhança entre os objetos, e consequentemente influencia a formação de *clusters*. Há quatro formas de atribuir pesos às variáveis

(Romesburg, 1984). Na primeira forma, o investigador pode deixar livremente algumas variáveis fora da matriz de dados original, isto é, atribuir-lhes peso igual a zero.

A segunda forma consiste em fazer uma análise de correlação que encontrará variáveis altamente correlacionadas. Se assim for, nas variáveis quantitativas mais correlacionadas, usa-se uma análise de componentes principais para obter um novo conjunto de variáveis não correlacionadas, que são as componentes principais (em menor número que as variáveis originais). A análise de componentes principais é uma técnica de estatística multivariada que pretende transformar um conjunto de variáveis relacionadas, X_1, X_2 ... X_p, em um outro que tenha um menor valor de variáveis. Essas novas variáveis, Y_1, Y_2 ... Y_p, não estão relacionadas e chamam-se componentes principais.

A terceira forma é escolher uma função de estandardização (transformação de variáveis). Com a estandardização, todas as variáveis têm o mesmo peso no que diz respeito às unidades de medida e variância. No entanto, existem situações nas quais podem existir variáveis com importância superior, a qual deve ser mantida.

Estandardizar é um meio de mudar os dados originais. Há outras duas formas de mudar os dados: transformá-los usando uma função de transformação, por exemplo $Z_{i,j}$ =log($X_{i,j}$) ou $Z_{i,j} = X_{i,j}^{1/2}$, identificando e depois removendo *outliers*. A função de estandardização usa parâmetros como a média da amostra e o desvio-padrão da amostra para a variável X_j, respectivamente.

Um problema que surge frequentemente na seleção de variáveis é a ausência de alguns valores nas variáveis selecionadas. Isso decorre do fato de a informação não ter sido conseguida quando a amostra foi recolhida ou de poder ter sido perdida. Na matriz de dados, pode ser colocado nos valores em falta um símbolo, por exemplo, NA (*not available*). Na matriz estandardizada, também é colocado NA nos valores que faltam.

A quarta forma consiste em atribuir pesos, para fazer com que as variáveis contribuam de uma forma que se baseia na semelhança entre objetos. O peso de uma variável pode ser aumentado por sua repetição na matriz. Outras medidas de variabilidade podem ser usadas para definir pesos, como o desvio-padrão e amplitude da amostra.

Medidas de distância

Dois objetos são próximos quando a sua dissemelhança ou distância é pequena ou a semelhança é grande. Há várias medidas de proximidade, algumas delas serão abordadas aqui.

Em um conjunto de dados com variáveis qualitativas, normalmente são usadas medidas de semelhança. Essas medidas geralmente têm valores pertencentes ao intervalo [0,1], embora por vezes sejam expressas em percentagem no intervalo [0,100].

Dois objetos *i* e *j* têm um coeficiente de semelhança igual a um, $S_{i,j}$=1, se têm valores idênticos para todas as variáveis. Dois objetos *i* e *j* têm um coeficiente de semelhança igual a zero, $S_{i,j}$=0, se diferem no máximo para todas as variáveis.

Vários coeficientes podem ser aplicados em um determinado estudo. A escolha depende do tipo de estudo, dos objetivos e da experiência e do bom senso do investigador. Diferentes coeficientes de semelhança podem originar resultados diferentes. Os coeficientes de semelhança mais usados

Tabela 1. Equações e intervalo de variação dos coeficientes de semelhança.

Coeficiente	Intervalo de variação
Jaccard: $C_{ij} = \frac{a}{a+b+c}$	[0,1]
Concordância simples: $C_{ij} = \frac{a+d}{a+b+c+d}$	[0,1]
Yule: $C_{ij} = \frac{ad-bc}{ad+bc}$	[-1,1]
Hamann: $C_{ij} = \frac{(a+d)-(b+c)}{(a+d)+(b+c)}$	[-1,1]
Sorenson: $C_{ij} = \frac{2a}{2a+b+c}$	[0,1]
Rogers e Tanimoto: $C_{ij} = \frac{a+d}{a+2(b+c)+d}$	[0,1]
Sokal e Sneath: $C_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$	[0,1]
Rogers e Tanimoto: $C_{ij} = \frac{a+d}{a+2(b+c)+d}$	[0,1]
Sokal e Sneath: $C_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$	[0,1]

para variações qualitativas são Jaccard, Concordância simples, Yule, Hamann, Sorenson, Rogers e Tanimoto, Sokal e Sneath (Tabela 1).

O coeficiente de Jaccard é muito usado na taxonomia numérica e na ecologia. O coeficiente de Sorenson é usado em botânica, em estudos de similaridade entre comunidades florestais. O coeficiente de Rogers e Tanimoto também é usado em botânica e na agropecuária. Por último, o coeficiente de Sokal e Sneath é usado em estudos sobre zoologia e genética.

Variáveis qualitativas e quantitativas

Para as variáveis qualitativas com dois ou mais níveis, a estratégia é decompor cada variável em variáveis binárias e construir um coeficiente de semelhança para as variações.

Variáveis ordinais são tratadas como variáveis normais, e cada uma deve ser decomposta em tantas variáveis binárias quantos forem os níveis que tiver. É construída uma tabela de duas entradas

para cada par de objetos, na qual são apresentados os números de níveis comuns a ambos os objetos, os números de níveis apresentados em somente um dos objetos, o número de níveis não apresentados em nenhum dos objetos. Em seguida, são calculados os coeficientes de semelhança para variáveis binárias.

Quando as variáveis quantitativas são contínuas, as medidas de proximidade entre objetos são normalmente medidas de dissemelhança ou medidas de distância (dy). É possível fazer a conversão de dados contínuos em dados categorizados e usar medidas para variáveis nominais. No entanto, cabe observar que há perda de informação na conversão.

Os coeficientes de dissemelhança para dados quantitativos mais utilizados são: distância euclidiana, distância Manhattan (ou *City Block*), distância Minkowski, coeficiente de correlação de Pearson e coeficiente de separação angular (cosseno) (Tabela 2).

Tabela 2. Lista de alguns coeficientes de dissemelhança para dados quantitativos.

Medida	Fórmulas
Distância euclidiana	$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
Distância Manhattan	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert ; \ \forall r \in R: r \geq 1$
Distância Minkowski	$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^r}$
Coeficiente de correlação de Pearson	$r_{if} = \frac{\sum_{k=1}^{p} (x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})}{\left[\sum_{k=1}^{p} (x_{ik} - \bar{x_i})^2 \sum_{k=1}^{p} (x_{jk} - \bar{x_j})^2\right]^{\frac{1}{2}}}$ $onde \ \overline{x_{s.}} = \frac{\sum_{k=1}^{p} x_{sk}}{p}$
Coeficiente de separação angular (cosseno)	$c_{ij} = \frac{\sum_{k=1}^{p} (x_{ik} x_{jk})}{\left[\sum_{k=1}^{p} x^2{}_{ik} \ \sum_{k=1}^{p} x^2{}_{jk}\right]^{\frac{1}{2}}}$

A distância euclidiana não deve ser usada quando as variáveis são medidas em unidades diferentes, quando são correlacionadas ou quando têm variâncias muito diferentes, porque nessas condições as variáveis intervêm com pesos muito diferentes na determinação dos valores de dissemelhança. Além disso, a dissemelhança é sensível à mudança de escala. Alterações na escala alteram os valores de distância, como também as ordens dos valores e, consequentemente, os resultados das análises de *clusters*.

A distância de Manhattan ou *City Block* mede a distância em uma configuração retilínea, uma configuração de cidade (quarteirões). Essa distância é menos afetada por *outliers* em relação à distância euclidiana e é de fácil interpretação.

A distância Minkowski é uma generalização das distâncias euclidiana e Manhattan.

O coeficiente de correlação de Pearson varia no intervalo [-1, 1], em que $r_{i,j}$ = 1 indica a semelhança máxima, mas não necessariamente a identidade entre as características dos objetos *i* e *j*, e $r_{i,j}$ = -1 indica o máximo de dissemelhança.

O coeficiente de separação angular (ou cosseno) define-se no intervalo [-1,1], em que: $C_{i,j}$= 1 indica que a semelhança é máxima (dissemelhança é mínima); $C_{i,j}$= -1 indica que a semelhança é mínima (dissemelhança é máxima); e $C_{i,j}$ é o cosseno do ângulo formado pelas duas semirretas que unem a origem aos respectivos objetos, representados com pontos no espaço.

Tanto o coeficiente de correlação de Pearson quanto o coeficiente de separação angular (ou cosseno) podem ser usados para quantificar semelhanças entre os objetos observados *i* e *j* no espaço de dimensão *p*.

Algoritmos de agrupamento

Os algoritmos de agrupamentos hierárquicos, conhecidos como *Sequential*, *Agglomerative, Hierarchic, Nonoverlapping Clustering Methods* (SAHN), são formados a partir de uma matriz de similaridade ou dissimilaridade, na qual é identificado o par de parcelas que mais se parecem. Nesse instante, o par é agrupado, formando uma parcela única. Esse processo requer uma nova matriz de similaridade ou dissimilaridade. Em seguida, identifica-se o par mais semelhante que formará o novo grupo, e assim sucessivamente, até que todas as parcelas fiquem reunidas em um grupo (Sneath; Sokal, 1973).

Os vários algoritmos de agrupamento diferem no modo como estimam distância entre grupo formado e outros grupos ou parcelas por agrupar. O processo de agrupamento de parcelas agrupadas depende da similaridade e dissimilaridade entre os grupos. Portanto, diferentes definições dessas distâncias podem resultar em diferentes soluções finais (Bussab et al., 1990).

A seguir, são apresentados diversos métodos de agrupamentos que fazem parte dos métodos SAHN. Vale salientar que não existe o que se possa chamar de melhor critério na análise de agrupamentos, embora alguns sejam mais indicados para determinadas situações do que outros (Kaufmann; Rosseeuw, 1990). É prática comum utilizar diversos critérios e comparar os resultados. Se os resultados forem semelhantes, é possível concluir que eles apresentam elevado grau de estabilidade, sendo, pois, confiáveis.

Os métodos mais comuns de agrupamento para determinar a distância entre agrupamentos são: ligação simples, ligação completa, centroides, mediana, médias das distâncias e método Ward (Anderberg, 1973).

Método da ligação simples (*single linkage*)

Também denominado "método do vizinho mais próximo" (*neighbourhoods*), foi proposto originalmente por Florek et al. (1951) e depois revisado por McQuitty (1960). É um dos algoritmos mais antigos e mais simples utilizados na literatura. Nele, as conexões entre parcelas e grupos, ou entre grupos, são feitas por ligações simples entre pares de parcelas.

O método da ligação simples, segundo Mardia et al. (1997), é uma técnica de hierarquização aglomerativa e tem, como uma de suas características, não exigir que o número de agrupamentos seja fixado a priori. Assim, seja $E = \{E_1, E_2, ... , E_p\}$ um conjunto de parcelas em que cada um é representado por um vetor X_i, para i = 1, 2, p pontos do espaço p-dimensional (lp). No caso de análise da vegetação, cada dimensão do espaço corresponde a uma espécie diferente. Desse modo, qualquer medida de distância estatística ou de similaridade pode ser empregada em tal algoritmo.

Supondo que tenham sido determinados todos os n(n - 1)/2, diferentes valores de $d_{i,j}$ ou $S_{i,j}$ (i = j = 1, 2... n) são representados na forma de uma matriz de distância (D_1) ou de similaridade (S_1). Este método produz grupos longos, comparados aos grupos formados por outros métodos de agrupamentos SAHN (Meyer et al., 2004).

Os dendrogramas resultantes desse procedimento são geralmente pouco informativos, devido à informação das parcelas intermediárias, que não são evidentes (Carlini-Garcia et al., 2001). Segundo Sneath e Sokal (1973), agrupamentos pelo método de ligação simples podem ser obtidos tanto pelo procedimento aglomerativo quanto pelo divisivo. Anderberg (1973) enumera as seguintes características do referido método:

- Em geral, grupos muito próximos podem não ser identificados;
- Permite detectar grupos de formas não elípticas;
- Apresenta pouca tolerância a *outliers*, por ter tendência a incorporá-los a um grupo existente;
- Apresenta bons resultados tanto para distância euclidiana quanto para outras distâncias;
- Tendência a formar longas cadeias (encadeamento).

Encadeamento é um termo que descreve a situação na qual um primeiro grupo, de uma ou mais parcelas, passa a incorporar, a cada interação, um grupo de apenas uma parcela. Assim, é formada uma longa cadeia, na qual se torna difícil definir um nível de corte para classificar as parcelas em grupos (Romesburg, 1984).

Método da ligação completa (*complete linkage*)

Este método foi introduzido em 1948, e é oposto ao método de ligação simples. É também denominado de método do elemento mais distante, e é uma das técnicas de hierarquização aglomerativa de maior aplicação na análise de agrupamento (Albuquerque et al., 2006). Como no método da ligação simples, aqui também não é exigida a fixação a priori do número de agrupamentos.

Segundo Bussab et al. (1990), no método da ligação completa, a dissimilaridade entre dois grupos é definida como aquela apresentada pelas parcelas de cada grupo que mais se parecem. Ou seja, formam-se todos os pares com um membro de cada grupo, e a similaridade ou dissimilaridade

entre os grupos é definida pelo par que mais se parece. Esse método geralmente produz grupos compactos e discretos, e tem valores de dissimilaridade relativamente grandes.

Kaufmann e Rosseeuw (1990) citam as seguintes características desse método:

- Apresenta bons resultados tanto para a distâncias euclidiana quanto para outras distâncias;
- Tendência a formar grupos compactos;
- Os *outliers* demoram a ser incorporados ao grupo.

O método descrito tem a desvantagem de poder produzir agrupamento diferente quando a dissimilaridade mínima ocorre para mais de um par de grupos e é necessário escolher um para ser unido.

Os métodos de ligação simples e ligação completa trabalham em direções opostas. Se apresentam resultados semelhantes, o grupo está bem definido no espaço, ou seja, o grupo é real. Todavia, se ocorre o contrário, é provável que os grupos não existam (Romesburg, 1984).

Método das médias das distâncias (*average linkage*)

Este método, também denominado de método das médias das ligações e método da média de grupo, foi proposto originalmente por Sokal e Michener (1958), e é uma ponderação entre os métodos de ligação simples e de ligação completa. Usa-se a similaridade ou dissimilaridade média das parcelas ou do grupo que se pretende unir a um grupo já existente. Há vários tipos de métodos, uma vez que há vários tipos de médias, quatro deles mais comuns, provenientes da combinação de dois critérios alternativos: agrupamento em função da média aritmética versus agrupamento com base no centroide, podendo ser ou não ponderados em ambos os casos.

Nos métodos de agrupamento com base na média aritmética, os coeficientes de similaridade ou dissimilaridade médios entre o indivíduo que se pretende agrupar e as parcelas do grupo já existente são calculados. O método do centroide busca o centroide das parcelas para construir grupos e medir a similaridade ou dissimilaridade relativa a esse ponto, entre qualquer parcela ou grupo candidato. Os métodos normalizados pretendem dar pesos iguais a todos os ramos do dendrograma, sendo que o número de parcelas que compõem cada ramo não é considerado (Bussab et al., 1990).

Sneath e Sokal (1973) descrevem as quatro combinações possíveis para esses critérios descritos. Este método define a distância entre dois grupos como sendo a média das distâncias entre todos os pares de parcelas, sendo um em cada grupo. Este procedimento pode ser utilizado tanto para medidas de similaridade como de distância, contanto que o conceito de uma medida média seja aceitável. Os grupos são reunidos em um novo grupo quando a média das distâncias entre suas parcelas é mínima.

No método das médias das distâncias, define-se a distância entre dois grupos, i e j, como sendo a média das distâncias entre todos os pares de parcelas constituídos por parcelas dos dois grupos. A estratégia e o valor médio têm a vantagem de evitar valores extremos e de tomar em consideração toda a informação dos grupos.

Um grupo passa a ser definido como um conjunto de parcelas, em que cada um tem mais semelhanças, em média, com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo.

Kaufmann e Rosseeuw (1990) destacam as seguintes características desse método:

- Apresenta menor sensibilidade a *outliers*, comparado aos métodos de ligação simples e completa;
- Apresenta bons resultados tanto para a distância euclidiana quanto para outras distâncias;
- Revela tendência a formar grupos com número de parcelas similares.

Segundo Hartigan (1981) esse método também é inconsistente na detecção de grupos de "alta densidade". Todavia, Milligan e Cooper (1985), em um estudo comparativo envolvendo os métodos ligação simples, ligação completa e ligação média, classificaram o método da ligação média como o melhor, visto que esse último método tira proveito da homogeneidade do método de ligação completa e da estabilidade do método da ligação simples.

Método de Ward

Ward (1963) propôs um processo geral de classificação no qual n parcelas são progressivamente reunidas dentro de grupos, por meio da minimização de uma função objetiva para cada (n - 2) passo de fusão.

Inicialmente, para esse algoritmo, admitia-se que cada uma das parcelas constituía um único agrupamento. Considerando-se a primeira reunião de parcelas em um novo agrupamento, a soma dos desvios dos pontos representativos de suas parcelas em relação à média do agrupamento é calculada e dá uma indicação da homogeneidade do agrupamento formado. Esta medida fornece a "perda de informação" ocasionada quando as parcelas são reunidas em um agrupamento (Lattin et al., 2011).

A proposta de Borouche e Saporta (1982) demonstra quando as parcelas são pontos de um espaço euclidiano (Ip). A qualidade de uma partição é definida por sua inércia intragrupo ou por sua inércia intergrupo. Quando se parte de K+1 grupos para K grupos, ou seja, dois grupos são agrupados em um só, a inércia intergrupo só pode diminuir. A inércia intergrupo é a média da soma dos quadrados das distâncias entre os centros de gravidade de cada grupo e o centro de gravidade total.

O algoritmo de Ward baseia-se na perda de informação resultante do agrupamento das espécies e é medida por meio da soma dos quadrados dos desvios das parcelas individuais relativamente às médias dos grupos em que são classificadas. Cada grupo caracteriza-se por uma soma dos quadrados dos desvios de cada parcela do centroide do mesmo algoritmo (é uma soma dos numeradores dos estimadores das variâncias de cada variável dentro do grupo; é também a soma da distância euclidiana do quadrado de cada parcela do centroide). A distância entre dois grupos é definida como o aumento que se pronunciaria nessa soma de quadrados se ambos os grupos se agregassem para a formação de um único grupo. O algoritmo de Ward é atraente por se basear em uma medida com forte apelo estatístico e por gerar grupos que, assim como os do método do vizinho mais longe, apresentam alta homogeneidade interna (Barroso; Artes, 2003).

Romesburg (1984) cita as seguintes características do método ora descrito:

- Apresenta bons resultados tanto para distância euclidiana quanto para outras distâncias;
- Pode apresentar resultados insatisfatórios quando o número de parcelas em cada grupo é praticamente igual;

- Tem tendência a combinar grupos com poucas parcelas;
- É sensível à presença de *outliers*.

Os algoritmos de ligação simples, completa e média podem ser utilizados tanto para variáveis quantitativas quanto qualitativas, ao contrário dos métodos de centroide e de Ward, que são apropriados apenas para variáveis quantitativas, já que têm como base a comparação de vetores de médias (Barroso; Artes, 2003).

Números de *clusters*

A escolha do número de *clusters* pode ser aleatória, por meio de método hierárquico ou pela experiência e conhecimento do investigador.

Batista et al. (2011) afirmam que o número de grupos pode ser definido, a priori, por meio de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador ou por simplicidade ou, ainda, pode ser definido a posteriori, com base nos resultados da análise ou pela experiência do pesquisador.

De acordo com Lattin et al. (2011), para determinar o número apropriado de grupos, existem diversas abordagens possíveis: em primeiro lugar, o pesquisador pode especificar antecipadamente o número de agrupamentos. Talvez, por motivos teóricos e lógicos, esse número seja conhecido. O pesquisador pode, também, ter razões práticas para estabelecer o número de agrupamentos, com base no uso que pretende fazer dele. Em segundo lugar, o pesquisador pode especificar o nível de agrupamento de acordo com um critério. Se o critério de agrupamento for de fácil interpretação, tal como a média de similaridade ou dissimilaridade interna do agrupamento, é possível estabelecer certo nível que ditaria o número de agrupamentos. As distâncias entre os agrupamentos, em etapas sucessivas, podem servir de guia, e o pesquisador pode escolher interromper o processo, quando as distâncias excederem um valor estabelecido.

Uma terceira abordagem é representar, graficamente, a razão entre a variância total interna dos grupos e a variância entre os grupos em relação ao número de agrupamentos. O ponto em que surgir uma curva acentuada, um ponto de inflexão, poderá ser a indicação do número adequado de agrupamentos. Aumentar esse número para além desse ponto seria inútil, e diminuí-lo seria correr o risco de misturar parcelas diferentes.

Qualquer que seja a abordagem empregada, geralmente é aconselhável observar o padrão total de agrupamentos. Isso pode proporcionar uma medida da qualidade do processo de agrupamento e do número de agrupamentos que emergem nos diversos níveis do critério de agrupamento. De maneira geral, mais de um nível de agrupamento é relevante (Bertini et al., 2010).

O método hierárquico tem a vantagem de ser aplicado a grandes volumes de dados, uma vez que não é preciso calcular e armazenar uma nova matriz de dissemelhança em cada passo do algoritmo. Outra vantagem é sua capacidade de reagrupar os objetos em *clusters* diferentes daqueles que foram calculados inicialmente. A desvantagem é a definição do número de grupos.

Abordagem prática de análise exploratória de dados a partir do agrupamento hierárquico no software R

Considerando que já estão instalados o software R (http://cran-r-project.org/) e o software RStudio (http://www.rstudio.com/products/RStudio/), o próximo passo é instalar os pacotes por meio da função `install.packages( )`.

Pacotes:

- FactoMiner: o método proposto do pacote é a análise multivariada – análise de componentes principais, análise correspondente e agrupamento (*clustering*).
- ggplot2: criação de elegantes visualizações de dados usando a gramática de gráficos.
- `xlsx`: pacote contém funções para ler dados de arquivos do Excel.
- `factoextra`: função para extrair e visualizar os resultados de dados multivariados em diferentes pacotes do R.
- devtools: tem como objetivo facilitar o desenvolvimento de pacotes fornecendo funções R que simplificam as tarefas comuns.
- `dendextend`: função para estender objetos do dendrograma.

Os dados devem estar preparados na tabela, na qual as linhas devem representar as observações (indivíduos) e as colunas, as variáveis.

Neste trabalho são utilizados os dados da Produção Agrícola Municipal (PAM) do Instituto Brasileiro de Geografia e Estatística (IBGE), disponíveis na plataforma Sidra (https://sidra.ibge.gov.br/tabela/5457), mais especificamente a tabela 5457, com dados de área plantada ou destinada à colheita, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporárias e permanentes (IBGE, 2020).

As variáveis usadas são:

- área plantada ou destinada à colheita (hectares);
- produtos das lavouras temporárias e permanentes (71);
- ano de 2018;
- unidade territorial são as microrregiões do estado de Mato Grosso.

Algumas lavouras temporárias e permanentes não foram plantadas no ano de 2018 nas microrregiões, por isso foram excluídas (Tabela 3).

Tabela 3. Área plantada ou destinada à colheita nas microrregiões do estado de Mato Grosso no ano de 2018 (IBGE, 2020).

Microrregião MT	Feijão (em grão)	Girassol (em grão)	Goiaba	Guaraná (semente)	Laranja	Limão	Mamão	Mamona (baga)	Mandioca	Manga	Maracujá
Aripuanã	9.888	10.000		15	108	17	26		2.445		36
Alta Floresta	9			150	15	16	10		336		47
Colíder			36	7	64	40			1.760	60	78
Parecis	37.957	41.200	2		25	25		650	398	9	11
Arinos	13.430	1.250		13	19	2			380		46
Alto Teles Pires	99.050	2.800				29	10		845		20
Sinop	15.400		5	154	25	10	20		1.000		15
Paranatinga	11.000								370		9
Norte Araguaia	1.800								4.155		
Canarana	25.710				21		8	1214	1.228		
Médio Araguaia									170		
Alto Guaporé	1.740				25	10			595		
Tangará da Serra	6.678	2.000			14	17			1.530		5
Jauru	87				115	39	22		615		3
Alto Paraguai	150		2			2	2		400		3
Rosário Oeste	14		5			55	10		380	1	1
Cuiabá						14	6		1.640	3	8
Alto Pantanal	20		3		30	31	9		355		
Primavera do Leste	33.800					27	30		160		20
Tesouro	6.000								440		13
Rondonópolis	2.900					3	2		866		
Alto Araguaia	900								150		

Continua...

Tabela 3. Continuação.

Microrregião MT	Melancia	Melão	Milho (em grão)	Palmito	Pimenta-do-reino	Soja (em grão)	Sorgo (em grão)	Tangerina	Tomate	Urucum (semente)	Uva
Aripuanã	342	9	121.816	267	6	235.800			59	185	
Alta Floresta	84		16.890	181		31.525		10	4		
Colíder	110	12	65.500	43	1	159.476			7		4
Parecis	46		668.802	8		1.343.000	14.600		5		
Arinos	37	4	327.000	150	9	660.500	1.500	5	8		
Alto Teles Pires	311	1	1.378.930			2.264.440	4.600	5	9		35
Sinop	138	44	469.161		1	773.906	300	7	2		
Paranatinga	5		64.500			466.598					
Norte Araguaia			277.859			720.332	12.800				
Canarana	180	22	314.267	246		1.038.990	6.500		8		
Médio Araguaia			1.390			43.700	1.000				4
Alto Guaporé	52		26.255	25		59.686	7.070				
Tangará da Serra	95	4	52.301	16		109.137			40	15	
Jauru	245		5.109	38		8.664		8	16		
Alto Paraguai	17		23.900			63.500	700				
Rosário Oeste	33	52	7.168			27.040	170		6		
Cuiabá	62		32.350			48.438		2	3		
Alto Pantanal	86	15	2.734			16.024		20	17		
Primavera do Leste			183.000			490.000	6.000	5			4
Tesouro	7		98.975			243.602	3.100	1			4
Rondonópolis	30		194.600			461.764	7.120	9			2
Alto Araguaia			86.000			171.766	3.000				

Tratamento de dados

O primeiro erro que pode ser encontrado é a ausência de observações em algumas variáveis de uma determinada microrregião. Uma quantidade de dados menor do que é esperado pode prejudicar a confiabilidade dos resultados do estudo e, eventualmente, produzir resultados tendenciosos ou específicos para uma determinada microrregião. Observações ausentes podem decorrer de perda de informação, falta de resposta durante a coleta e também de valores discrepantes (*outliers*).

Na tabela das áreas plantadas (hectares) das lavouras temporárias e permanentes no ano de 2018 nas microrregiões do estado de Mato Grosso, foi acrescida uma constante 1 a todas as linhas nas quais faltavam coletas como tratamento de dados.

Os comandos e *scripts* são detalhados nos tópicos a seguir, e seus resultados são ilustrados nas Figuras 4 a 16.

Para fazer a análise de *cluster* das áreas plantadas em MT, o primeiro passo é salvar os arquivos ".R" no diretório de escolha para, em seguida, usar o comando "`setwd`", conforme os exemplos abaixo e seguindo o caminho desse seu diretório.

Em Linux e IOS:

```
>setwd("~/Documents/cluster/arquivo_08_2020")
```

```
Em Windows:
```

```
>setwd("C:\Documents\cluster\arquivo_08_2020\\")
```

```
rm(list=ls())
```

Para instalar pacotes:

```
install.packages("FactoMineR")
```

```
install.packages("ggplot2")
```

```
install.packages("ggdendro")
```

```
install.packages("xlsx")
```

```
install.packages("cluster")
```

```
install.packages("devtools")
```

```
install.packages("factoextra")
```

```
install.packages("stats")
```

```
install.packages("dendextend")
```

```
install.packages("igraph")
```

Outra forma de instalar os pacotes é:

```
install.packages (c("FactoMineR","ggplot2","ggdendro","xlsx","cluster","-
devtools", "factoextra","stats", "dendextend", "igraph"))
```

Caso os pacotes já estejam instalados, devem ser usados os comandos a seguir.

```
library(FactoMineR)
library(ggplot2)
library(ggdendro)
library(xlsx)
library(cluster)
library(devtools)
library(factoextra)
library(stats)
library(dendextend)
library(igraph)
options(prompt='R-> ')
options(digits=15)
options(scipen=15)
options(stringsAsFactors=FALSE)
```

Para ler os dados:

```
dados <<- read.table(
      file = 'area_plantada.csv',
      header = TRUE,
      sep = '\t',
      skip = 0,
      fill = TRUE,
      encoding = "UTF-8"
)
```

Para atribuir os nomes das microrregiões às linhas:

```
nomes_menores = c(
      "Aripuana","A.Floresta","Colider","Parecis",
      "Arinos","A.T.Pires","Sinop","Paranatinga",
```

```
      "N. Araguaia","Canarana","M.Araguaia","A.Guapore",
      "T.da Serra","Jauru","A.Paraguai","R.Oeste",
      "Cuiaba","A.Pantanal","P.doLeste","Tesouro",
      "Rondonopolis","A.Araguaia"
)

row.names(dados) <- nomes_menores
```

O procedimento `scale` procura os dados mais apropriados à aplicação de alguma técnica de análise, como métodos baseados em distância. A necessidade de usar scale pode ser consequência de vários fatores, como fazer com que cada atributo dos dados de entrada tenha o mesmo domínio.

Quanto às medidas de similaridade, há muitos métodos para calcular similaridade, tais como as distâncias euclidiana e Manhattan. No software R, é possível utilizar a função `dist( )` para computar a distância entre pares de objetos. Os resultados são chamados de distância ou matriz de dissimilaridade.

```
dados <- dados[,-1]
dados.scale <<- scale(dados)
dist.euclidean <<- dist(dados.scale,method='euclidean')
dist.minkowski <<- dist(dados.scale,method='minkowski')
dist.canberra <<- dist(dados.scale,method='canberra')
hc.ward <<- hclust(d=dist.euclidean,method='ward.D2')
hc.average <<- hclust(d=dist.euclidean,method='average')
hc.complete <<- hclust(d=dist.euclidean,method='complete')
dend.average <<- as.dendrogram(hc.average)
dend.ward <<- as.dendrogram(hc.ward)
```

Linkage

A função de ligação (*linkage*) obtém a informação da distância e retorna para a função `dist( )` grupos de pares de objetos em *clusters* (agrupamentos) baseados nas suas similaridades. Este processo é inteirado quando os dados originais são "ligados" em uma árvore hierárquica.

`hclust( )` pode ser usada como exemplo, a seguir:

```
res.hc <- hclust(d= res.dist, method = "ward.D2")
```

d: estrutura de dissimilaridade produzida pela função `dist( )`

O método de aglomeração (*linkage*) é usado para computar a distância entre agrupamentos (*clusters*). São usados `ward.D`, `Ward.D2`, `single`, `complete`, `average`, `mcquitty`, `median` ou `centroid`.

São vários métodos de agrupamentos (linkage methods). Os métodos mais comuns são descritos a seguir:

a) *Maximum* ou *complete linkage*: a distância entre dois *clusters* é definida como máximo valor da distância entre os elementos no *cluster* 1 e *cluster* 2. Isso tende a produzir agrupamentos (*clusters*) mais compactos.

b) *Minimum* ou *single linkage*: a distância entre dois agrupamentos (*clusters*) é definida como o mínimo valor entre os elementos no *cluster* 1 e *cluster* 2. Isso tende a produzir agrupamentos (*clusters*) mais longos (*loose clusters*).

c) *Mean* ou *average linkage*: a distância entre dois agrupamentos (*clusters*) é definida como a média da distância entre os elementos no *cluster* 1 e *cluster* 2.

d) *Centroid linkage*: a distância entre dois *clusters* é definida entre os centroides dos *cluster* 1 e *cluster* 2.

Os métodos mais utilizados, "*complete linkage*" e `Ward`"`s`", dependem do tipo de dados, do contexto do estudo, do conhecimento e da experiência do investigador.

Na prática, se os métodos abordados forem aplicados e representados em dendrogramas, para comparação dos resultados, isto é, para comparação do número de *clusters*, da sua composição e do nível de fusão dos *clusters*, é possível verificar se os dados apresentam ou não uma estrutura de grupos unânimes.

Para gerar as figuras:

```
graphp = data.frame(
      width = 2100,
      height = 2100,
      res = 600,
      pointsize = 12,
      quality = 100,
      bg = 'white'
)

ngroups = 9
#mycolors = terrain.colors(ngroups)
#mycolors = heat.colors(ngroups)
```

```
mycolors = rainbow(ngroups)
```

Para gerar o dendrograma ilustrado na Figura 3:

```
#fviz_dend(hc.ward,cex=0.6)
ggdendrogram(hc.ward,rotate=FALSE,size=2)
```

Dendrograma

Corresponde à representação gráfica de uma árvore hierarquizada utilizando a função `hclust( )`. O dendrograma pode ser produzido no R utilizando a função `plot(res.hc)`, na qual `res.hc` é a saída do `hclust( )`. A função `fviz_dend( )` no pacote factoextra R produz bons dendrogramas.

No dendrograma, a altura da fusão, provém do eixo vertical que indica a similaridade/distância entre dois objetos/*clusters*. No eixo vertical, os valores mais altos da fusão são os que menos têm similaridades entre os objetos. Neste exemplo, são cortados em nove grupos e diferenciados por cores (Figura 4).

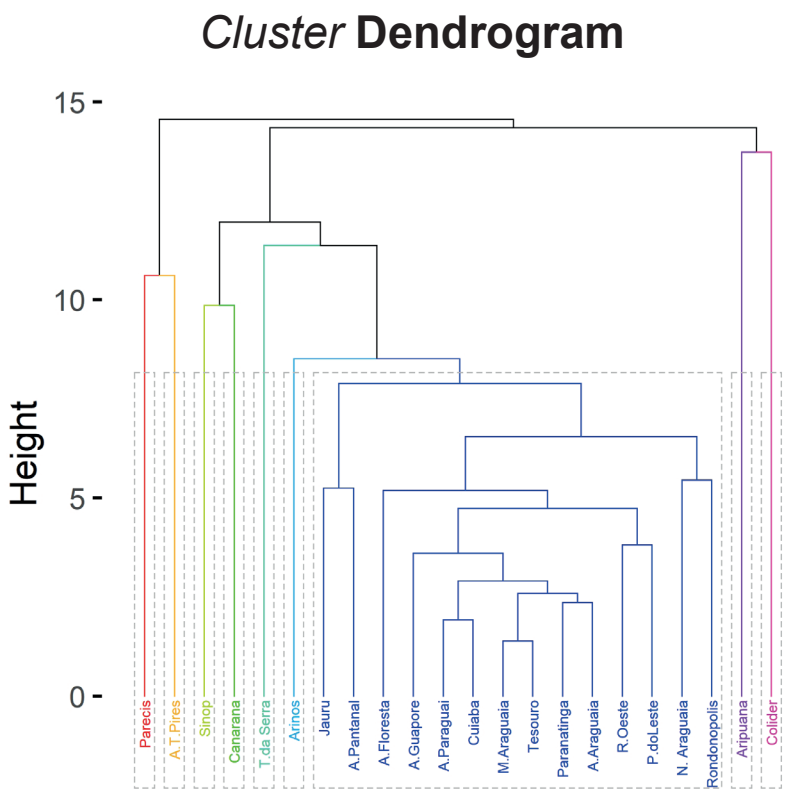

Figura 4. Dendrograma com grupos diferenciados por cores.

```
      fviz_dend(
            hc.ward,
            k = ngroups,     # cortar em k grupos
            k_colors = mycolors,
            color_labels_by_K = TRUE,        # Define cores no label dos
grupos
            cex = 0.3, # Define o tamanho do label
            lwd = 0.2,
            rect = TRUE           # Adiciona retângulos no entorno dos
grupos
      )
```

Agrupamentos hierárquicos divisivos (DIANA)

Os agrupamentos hierárquicos divisivos (DIANA) são o inverso dos agrupamentos por similaridade. Por exemplo, se G é um grupo já formado e o objetivo é dividi-lo em dois grupos de tal maneira que os dois grupos resultantes apresentem a maior dissimilaridade possível entre eles, inicialmente o algoritmo busca identificar um dado em G cuja dissimilaridade média com relação aos dados restantes seja máxima. O dado com dissimilaridade máxima é retirado de G e inserido em um novo grupo criado nesse momento, chamado tempG. Na sequência, para cada dado x ϵ G, é calculada a média dos valores de dissimilaridade de x, com todos os demais dados de G, média dos valores diss(x,y), y ϵ G. De maneira análoga, é calculada a média dos valores de dissimilaridade de x, com os dados pertencentes a tempG (Figura 5).

```
      dados.diana <- diana(dados,stand=TRUE)


      fviz_dend(
            dados.diana,
            cex = 0.3,
            k = ngroups,
            palette = mycolors
      )
```

Comparações de dendrogramas

Para comparar dois dendrogramas (Figura 6), é usada a função `dendextend` no software R.

O pacote `dendextend` contém várias funções para comparar dendrogramas. Aqui são enfocadas duas funções:

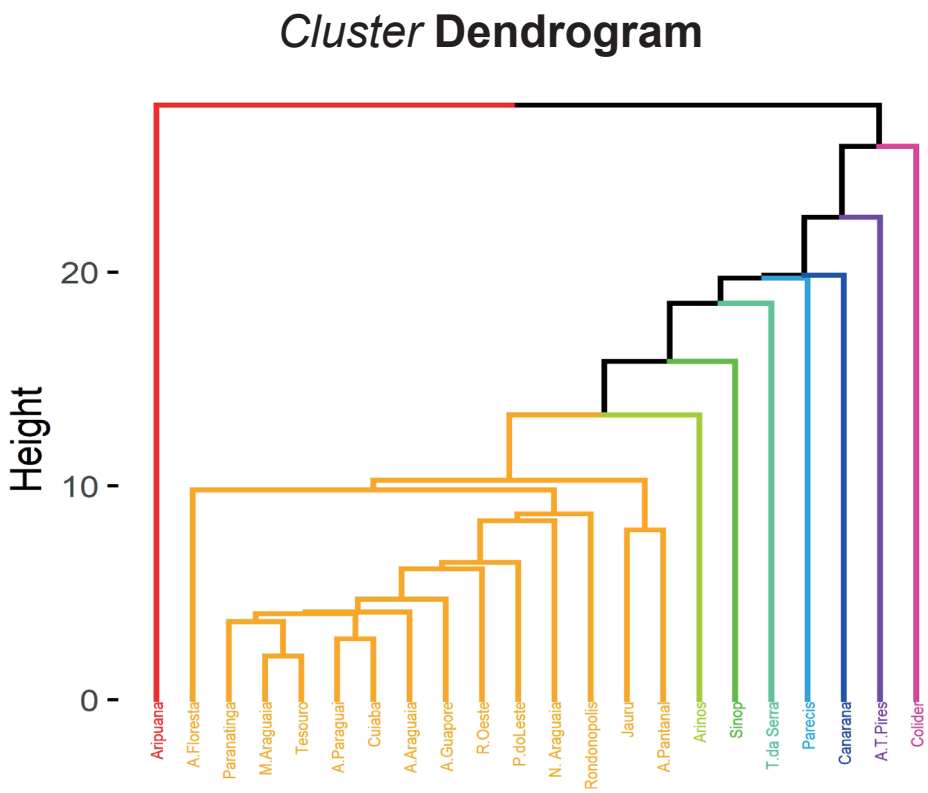

Figura 5. Agrupamentos hierárquicos divisivos (DIANA) por dissimilaridade.

`tanglegram( )`, para comparação visual de dois dendrogramas.

`cor.dendlist( )`, para computar a correlação entre dois dendrogramas.

```
dl <- dendlist(
  dend.ward %>%
    set("labels_col", value=mycolors,k=ngroups) %>%
    set("branches_lty", 1) %>%
    set("branches_k_color", value=mycolors,k=ngroups),
  dend.average %>%
    set("labels_col", value=mycolors,k=ngroups) %>%
    set("branches_lty",1) %>%
    set("branches_k_color", value=mycolors,k=ngroups)
)

tanglegram(
     dl,
     common_subtrees_color_lines = FALSE,
```

```
        highlight_distinct_edges = TRUE,
        highlight_branches_lwd = FALSE,
        margin_inner = 7,
        lwd = 2
)
```

Visualizar a comparação de dois dendrogramas

Para visualizar a comparação de dois dendrogramas (Figura 7), é utilizado o pacote *dendextend*:

`untangle( )`: procura o melhor leiaute para alinhar a lista no dendrograma utilizando o método heurístico.

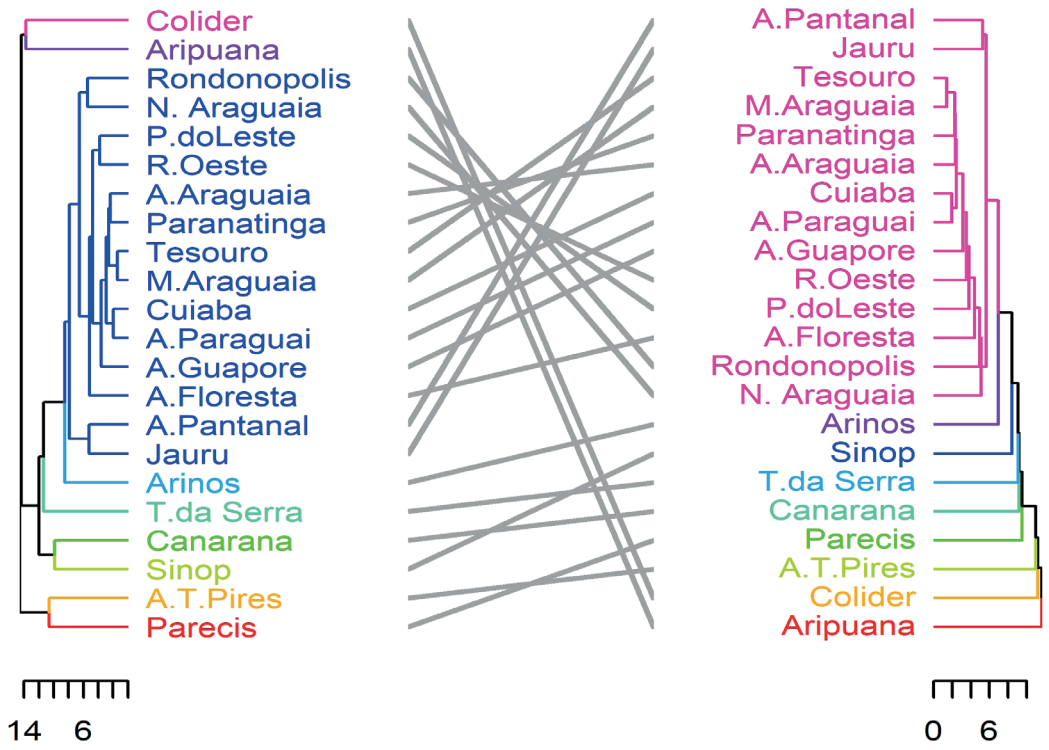

Figura 6. Comparações de dendrogramas.

`tanglegram( )`: plota dois dendrogramas, um ao lado do outro, e os *labels* (rótulos) são conectados por linhas.

`entanglement( )`: computa a qualidade do alinhamento em duas árvores. `Entanglement` é uma medida entre 1 (`entanglement` completo) e 0 (sem `entanglement`). Um baixo coeficiente de `entanglement` corresponde um bom alinhamento.

```
dl <- dendlist(
  dend.ward %>%
    set("labels_col", value=mycolors,k=ngroups) %>%
    set("branches_lty", 1) %>%
    set("branches_k_color", value=mycolors,k=ngroups),
  dend.average %>%
    set("labels_col", value=mycolors,k=ngroups) %>%
    set("branches_lty",1) %>%
    set("branches_k_color", value=mycolors,k=ngroups)

) %>%
untangle(method = "step1side")
```

É possível customizar o `tanglegram` usando outras opções:

```
tanglegram(
     dl,
     highlight_distinct_edges = FALSE,
     highlight_branches_lwd = FALSE,
     common_subtrees_color_lines = FALSE,
     common_subtrees_color_branches = TRUE,
     margin_inner = 7,
     lwd = 2
)
```

Para produzir uma matriz de correlação entre os dendrogramas:

```
cor.dendlist(dl, method = "cophenetic")
                    [,1]                [,2]
[1,] 1.000000000000000 0.979063043171907
[2,] 0.979063043171907 1.000000000000000
```

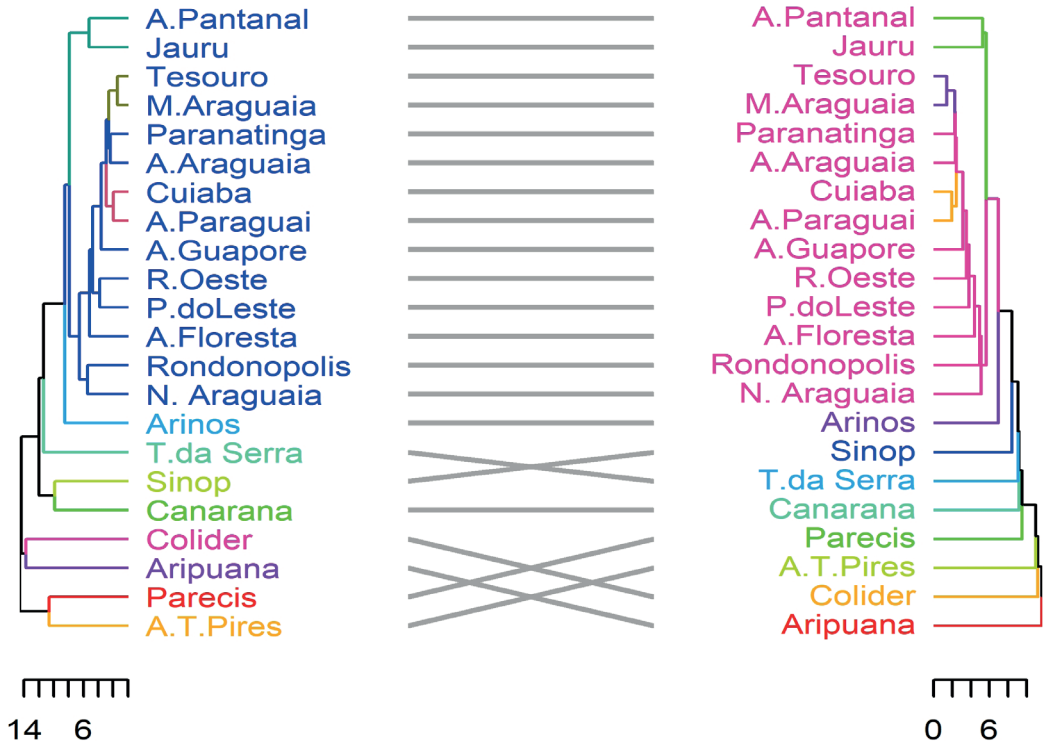

Figura 7. Customização da comparação de dois dendrogramas.

A função "`cor.dendlist( )`" é usada na correlação. Os valores variam de -1 a 1. A correlação positiva perfeita é igual 1 e a correlação negativa perfeita é igual a -1. O zero indica ausência de correlação.

Os métodos hierárquicos da análise de *cluster* impõem uma estrutura hierárquica aos dados. É necessário verificar o tipo de estrutura, se é aceitável ou introduz uma distorção inaceitável das relações nas proximidades originais entre os objetos.

Uma forma de validar o agrupamento consiste em comparar a matriz de proximidade original com a matriz de proximidade derivada. O coeficiente de correlação assim definido recebe o nome de correlação cofenético, e a matriz obtida, matriz cofenética.

Visualização de dendrogramas

Para visualizar o dendrograma, são utilizadas as seguintes funções e pacotes do R:

- “`fviz_dend( )`”, no pacote `factoextra`, para criar facilmente no `ggplot2`;
- “*`dendextend`*”, pacote para manipular os dendrogramas.

O dendrograma customizado na posição horizontal é mostrado na Figura 8.

```
fviz_dend(
      hc.ward,
      cex = 0.3,
      main = 'Dendrogram - ward.D2',
      xlab = 'Objects',
      ylab = 'Distance',
      sub = ''
)
```

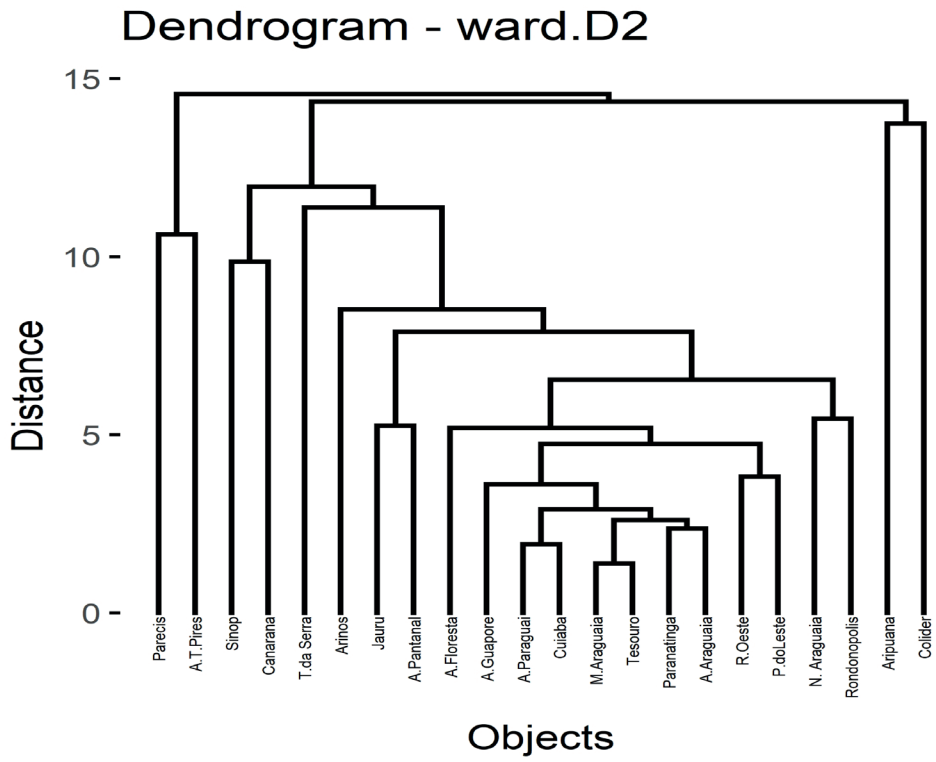

Figura 8. Customização do dendrograma na posição horizontal.

O dendrograma customizado na posição vertical é mostrado na Figura 9.

```
fviz_dend(
      hc.ward,
      cex = 0.3,
      horiz = TRUE
)
```

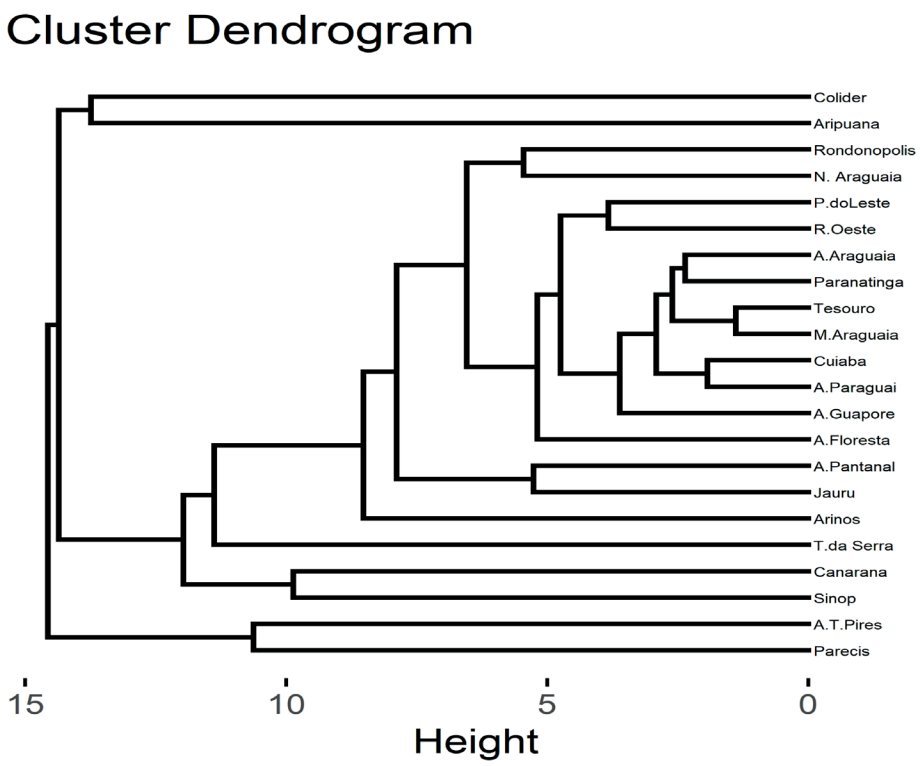

Figura 9. Customização do dendrograma na posição vertical.

Outros tipos de customização possíveis são mostrados nas Figuras 10 a 16.

```
fviz_dend(
      hc.ward,
      k = ngroups,
      k_colors = mycolors,
      color_labels_by_k = TRUE,
      cex = 0.5,
      rect = TRUE,
```

```
        rect_border = mycolors,
        rect_fill = TRUE
)
```

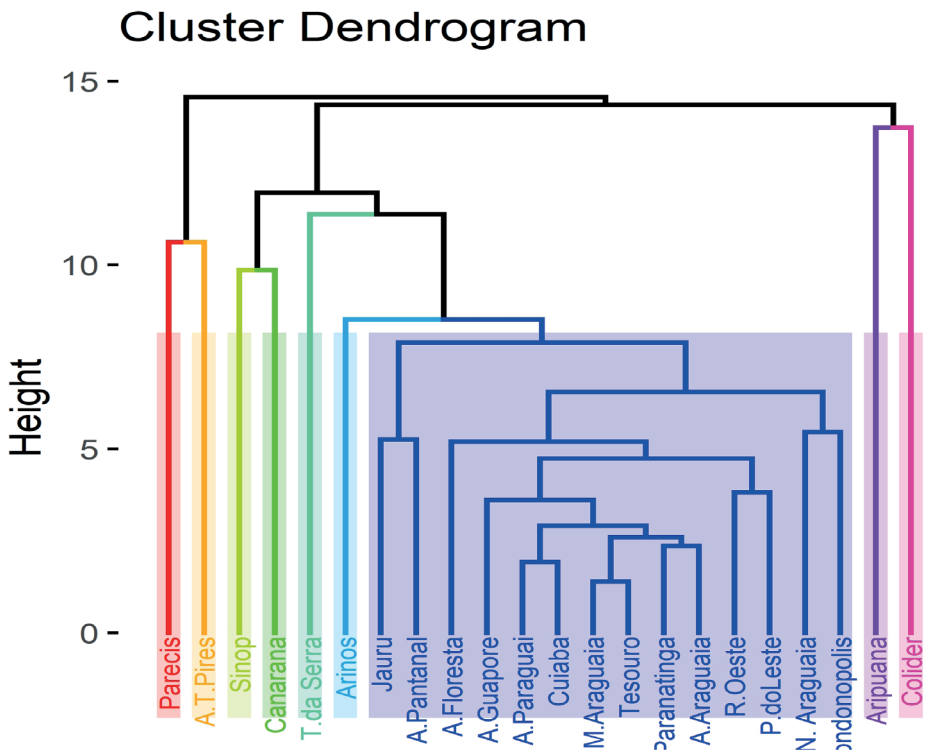

Figura 10. Customização do dendrograma separando nove grupos por cores e acrescentando retângulos ao redor de cada grupo.

```
fviz_dend(
        hc.ward,
        k = ngroups,
        k_colors = mycolors,
        color_labels_by_k = TRUE,
        cex = 0.3,
        ggtheme = theme_gray()
)
```

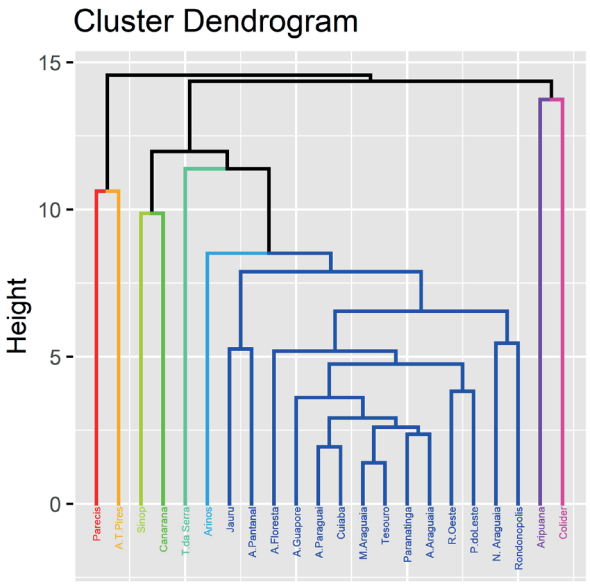

Figura 11. Customização do dendrograma separando nove grupos por cores e com fundo retangular cinza.

```
fviz_dend(
      hc.ward,
      cex = 0.3,
      k = ngroups,
      k_colors = mycolors
)
fviz_dend(
      hc.ward,
      k = ngroups,
      k_colors = mycolors,
      cex = 0.3,
      horiz = TRUE,
      rect = TRUE,
      rect_border = mycolors,
      rect_fill = TRUE
)
```

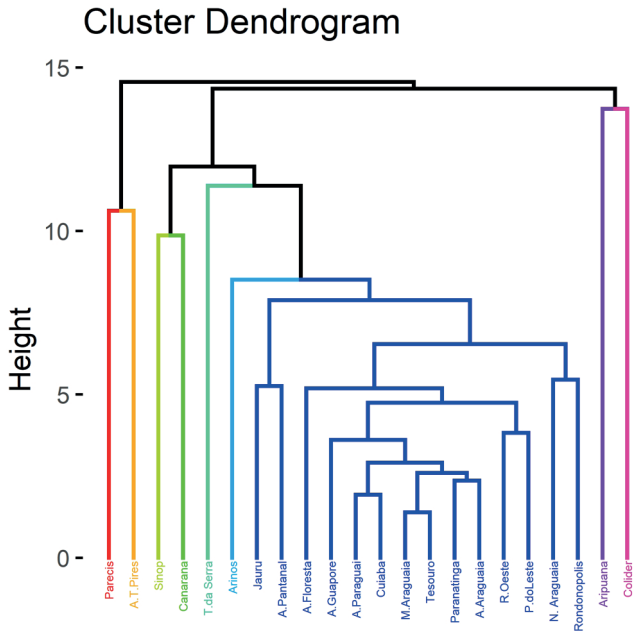

Figura 12. Customização do dendrograma separando nove grupos por cores e sem fundo cinza.

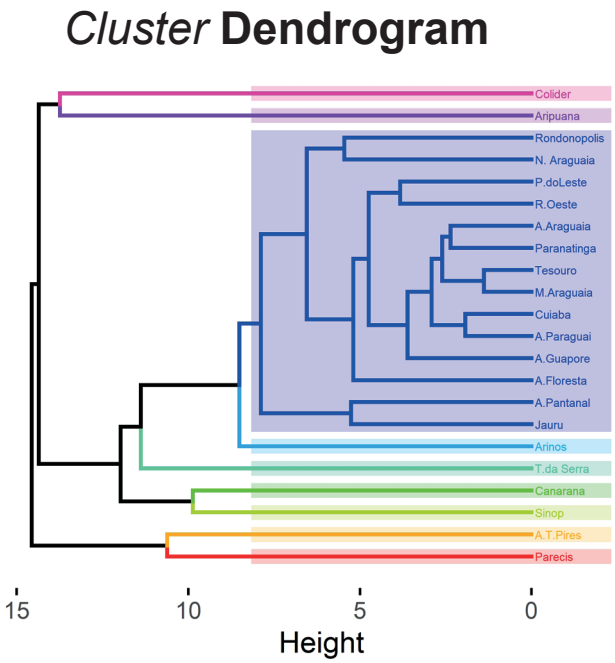

Figura 13. Customização do dendrograma separando nove grupos na vertical por cores e acrescentando a forma retangular ao redor de cada grupo.

```
fviz_dend(
      hc.ward,
      cex = 0.4,
      k = ngroups,
      k_colors = mycolors,
      type = 'circular'
)
```

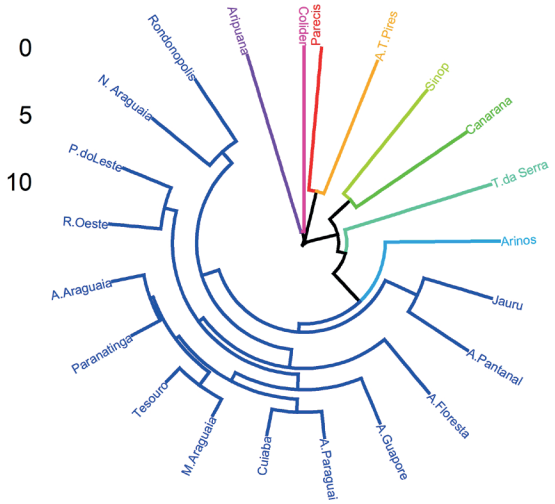

Figura 14. Customização do dendrograma na forma circular separando nove grupos por cores.

```
fviz_dend(
      hc.ward,
      k = ngroups,
      k_colors = mycolors,
      cex = 0.3,
      type = 'phylogenic',
      repel = TRUE
)
```

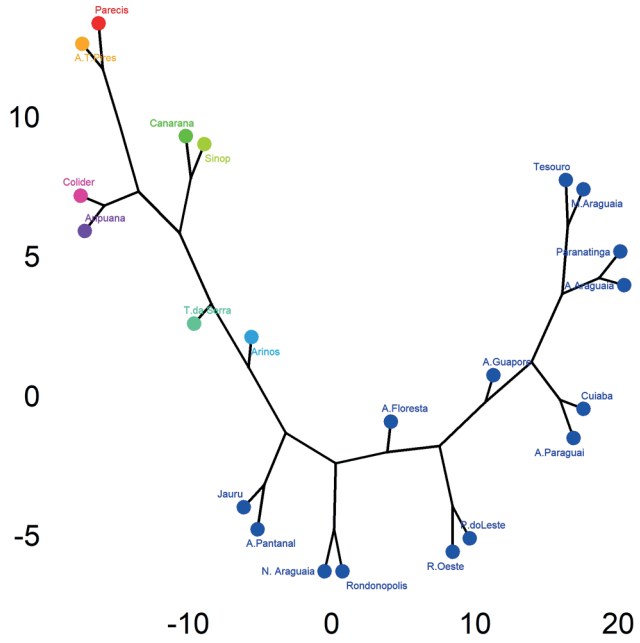

Figura 15. Customização do dendrograma na forma filogenética tipo árvore separando nove grupos por cores.

```
fviz_dend(
      hc.ward,
      k = ngroups,
      k_colors = mycolors,
      cex = 0.3,
      type = 'phylogenic',
      repel = TRUE,
      phylo_layout = 'layout.gem'
)
```

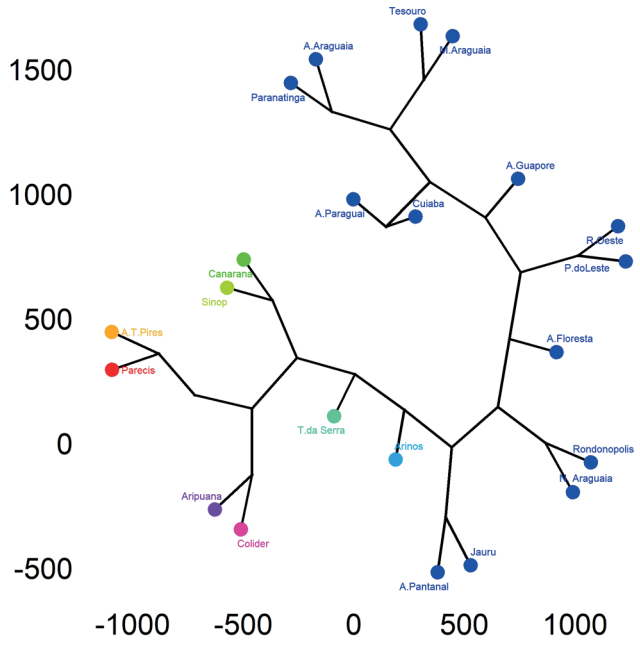

Figura 16. Customização do dendrograma na forma filogenética separando nove grupos por cores com leiaute genético.

Referências

ALBUQUERQUE, M. A.; FERREIRA, R. L. C.; SILVA, J. A. A.; SANTOS, E. S.; STOSIC, B.; SOUZA, A. L. Estabilidade em análise de agrupamento: estudo de caso em ciência florestal. **Revista Árvore**, Viçosa-MG, v. 30, n. 2, p. 257-265, 2006.

ANDERBERG, M. R. Hierarchical clustering methods. In: ANDERBERG, M. R. (Ed.) **Cluster analysis for applications**. London: Academic Press, 1973. p. 132-156.

BARROSO, L. P.; ARTES, R. Análise multivariada. In: REUNIÃO ANUAL DA RBES E SEAGRO, 48., 100., Lavras. **Curso**. Lavras: Departamento de Ciências Exatas, 2003. 155 p.

BATISTA, F. J.; JARDIM, M. A. G.; MEDEIROS, T. D. S.; MAGALHÃES, I. L. Comparação florística estrutural de duas florestas de várzea no estuário amazônico, Pará, Brasil. **Revista Árvore**, Viçosa-MG, v. 35, n. 2, p.289-298, 2011.

BERTINI, C. H. M.; ALMEIDA, W. S. de; SILVA, P. M. da; SILVA, J. W. L.; TEÓFILO, E. M. Análise multivariada e índice de seleção não identificada de genótipos superiores de feijão-caupi. **Acta Scientiarum Agronomy**, v. 32, n. 4, p. 613-619, 2010.

BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with cotraining. In: ANNUAL CONFERENCE ON COMPUTATIONAL LEARNING THEORY, 11., New York. **Proceedings of the Eleventh**. New York: ACM, 1998. p. 92-100.

BOROUCHE, J. M.; SAPORTA. G. **Análise de dados**. Zahar Editores. Rio de Janeiro, 1982. 116 p.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. Introdução à análise de agrupamentos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA (SINAPE), 9., 1990, **Anais**... São Paulo: Associação Brasileira de Estatística, 1990. p.72-75.

CARLINI-GARCIA, L. A.; VENCOVSKY, R.; COELHO, A. S. G. Método bootstrap aplicado em níveis de reamostragem na estimação de parâmetros genéticos populacionais. **Scientia Agricola**, v. 58, n. 4, p. 785-793, out./dez., 2001.

ERNST, G. W.; NEWELL, A. The problem solving structure of GPS. In: ERNST, G. W.; NEWELL, A. (Ed.). **GPS**: a case study in generality and problem solving. New York, NY: Front Cover, Academic Press; Business & Economics, 1969. p. 35-44.

EVERITT, B.; DUNN, G. *Cluster* analysis. In: EVERITT, B.; DUNN, G. (Ed.) **Applied Multivariate Data Analysis**. London, UK. Arnold, 1991. p.125-158.

FLOREK, K.; LUCASZEWICZ, J.; PERKAL, J.; STEINHAUS, H.; ZUBRZYCKI, S. Sur la Liaison et la Division des Points d'un Ensemble Fini. **Colloquium Mathematicae**, v. 2, p. 282 -285, 1951.

HARTIGAN, J. A. Consistency of Single Linkage for High-Density Clusters. J**ournal of the American Statistical** Association, v. 76, p. 388 - 394, 1981.

IBGE. **PAM - Produção Agrícola Municipal, 2018**. Disponível em: https://sidra.ibge.gov.br/tabela/5457. Acesso em: 02 ago. 2020.

KAUFMANN, L.; ROUSSEEUW, P. J. Clustering large applications (Program Clara). In: KAUFMANN, L.; ROUSSEEUW, P. J. (Ed.) **Finding groups in data**: an introduction to cluster analysis. New York: John Wiley, 1990. p.126-163.

KUHN, M.; JOHNSON, K. Classification Trees and Rule-Based Models. In: KUHN, M.; JOHNSON, K. (Ed.) **Applied Predictive Modeling**. New York, NY: Springer, 2016. p. 329- 366.

LATTIN, J. M.; DOUGLAS, C.; PAUL, E. G. Análise de componentes principais. In: LATTIN, J. M.; DOUGLAS, C.; PAUL, E. G. (Ed.) **Análise dos dados multivariados**. São Paulo: Cengage Learning, 2011. p. 67-101.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. Cluster analysis. In: MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. (Ed.) **Multivariate analysis**. London: Academic Press, 1997. p.360–393.

MCQUITTY, L. L. Hierarchical syndrome analysis for the isolation of types. **Educational and Psychological Measurement**, v. 20, p. 55 – 67, 1960.

MEYER, A. S.; GARCIA, A. A. F.; SOUZA, A. P.; SOUZA JUNIOR, C. L. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L.). **Genetics and Molecular Biology**, v. 27, n. 1, p. 83 - 91, 2004.

MICHALSKI, R. S.; STEPP, R. Automated construction of classifications: conceptual clustering versus numerical taxonomy. **IEE Transactions on Pattern Analysis and Machine Intelligence**, v. 5, p. 219-243, 1983.

MILLIGAN, G. N.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, New York, v. 50, n. 2, p. 159-179, 1985.

MITCHELL, T. Concept learning and the general to specific ordering. In: MITCHELL, T. (Ed.) **Machine Learning**. New York, NY, USA, McGraw-Hill, 1997. p. 22-45.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos de aprendizado de máquina. In: REZENDE, S.O. **Sistemas inteligentes**: fundamentos e aplicações. Barueri, SP: Manole, 2003. p. 89-94.

PRICE, L. J. Identifying cluster overlap with NORMIX population memberships probabilities. **Multivariate Behavorial Research**, v. 28, p. 235-262, 1993.

PROVOST, F.; FAWCETT, T. Agrupamento hierárquico. In: PROVOST, F.; FAWCETT, T. (Ed.) **Data Science para Negócios**. Rio de Janeiro, RJ: Alta Books, 2016. p.165-181.

ROMESBURG, C. H. How to make classifications. In: ROMESBURG, C. H. (Ed.). **Cluster analysis for researchers**. Belmont: Lifetime Learning Publications, 1984. p. 203-216.

SNEATH, P. H. A.; SOKAL, R. R. **Numeric taxonomy**: the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973. p. 15-20.

SOKAL, R. R., MICHENER, C. D. A statistical method for evaluating systematic relationships. **University of Kansas science bulletin**, n. 38, p. 109 - 143, 1958.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236-244, 1963.

WICKHAM, H.; GROLEMUND, G. Construção de modelos. In: WICKHAM, H., GROLEMUND, G. (Ed.) **R para Data Science, importe, arrume, transforme, visualize e modele dados**. Rio de Janeiro: Alta Books, 2019. p. 375-376.

ZUMEL, N.; MOUNT, J. Unsupervised methods. In: ZUMEL, N.; MOUNT, J. (Ed.) **Practical Data Science with R**. New York, NY: Ed. Manning, 2020. p. 311-339.

Embrapa
Territorial

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

CGPE XXXXX