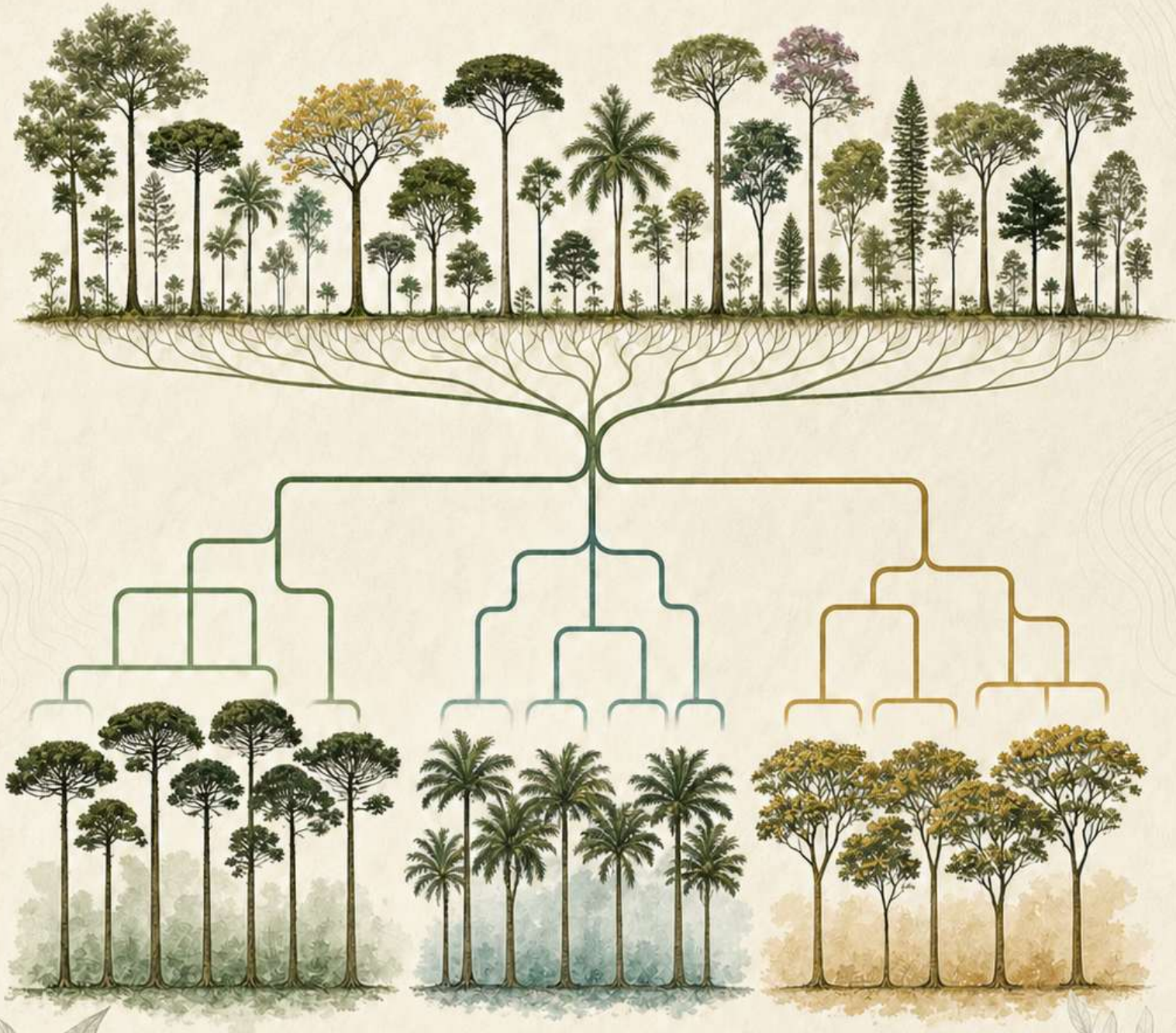
# ANÁLISE DE AGRUPAMENTO EM CIÊNCIA FLORESTAL COM APLICAÇÕES EM R



Rinaldo Luiz Caraciolo Ferreira
Agostinho Lopes de Souza
Rômulo Môra

# ANÁLISE DE AGRUPAMENTO EM CIÊNCIA FLORESTAL COM APLICAÇÕES EM R

UFRPE
UNIVERSIDADE FEDERAL RURAL
DE PERNAMBUCO - UFRPE

Editora Universitária da UFRPE

**Maria José de Sena**
Reitora
**Maria do Socorro de Lima Oliveira**
Vice-reitora
**Renata Valéria Regis de Sousa Gomes**
Pró-Reitora de Extensão, Cultura e Cidadania
**Rinaldo Aparecido Mota**
Pró-Reitor de Pós-Graduação
**Danielli Matias de Macedo Dantas**
Pró-Reitora de Ensino de Graduação
**Robson Wagner Albuquerque de Medeiros**
Pró-Reitor de Planejamento e Administração

**Tália de Azevedo Souto Santos**
Pró-Reitora de Gestão Estudantil e Inclusão

**Thieres George Freire da Silva**
Pró-Reitor de Pesquisa

**Renata Andrade de Lima e Souza**
Pró-Reitora de Gestão de Pessoas

**Elisabeth da Silva Araujo**
Diretora do Sistema de Bibliotecas da UFRPE

**Antão Marcelo Freitas Athayde Cavalcanti**
Diretor da Editora da UFRPE
**José Abmael de Araújo**
Coordenador Administrativo
**Marco Aurélio Cabral Pereira**
Editoração Eletrônica

**Autores**
Rinaldo Luiz Caraciolo Ferreira
Agostinho Lopes de Souza
Rômulo Môra

**Conselho Editorial**
Disson Soares dos Prazeres
Crislene Santos da Paixão
Lucas Rezende Valeriano
Renata de Farias Limeira Carvalho
Fábio Lima Santos

**Diagramação**
Rinaldo Luiz Caraciolo Ferreira

**Revisão Linguística**
Gaby Carvalho Alves

**Capa**
Rodrigo da Hora Caraciolo Ferreira

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Bibliotecária Suely Manzi – CRB/4 - 809

F383a Ferreira, Rinaldo Luiz Caraciolo
Análise de agrupamento em ciência florestal com aplicações em R / Rinaldo Luiz Caraciolo Ferreira, Agostinho Lopes de Souza, Rômulo Môra. – 1. ed. - Recife: EDUFRPE, 2026.
255 p.: il.
Inclui referências.

1. Análise multivariada 2. Análise de agrupamentos 3. Engenharia florestal 4. R (Linguagem de programação de computador) I. Souza, Agostinho Lopes de II. Môra, Rômulo III. Título

CDD 634.9

ISBN DIGITAL nº 978-65-02-25134-8

# ANÁLISE DE AGRUPAMENTO EM CIÊNCIA FLORESTAL COM APLICAÇÕES EM R

Rinaldo Luiz Caraciolo Ferreira
Agostinho Lopes de Souza
Rômulo Môra

Editora Universitária da UFRPE

1ª Edição

Recife, PE
2026

## Rinaldo Luiz Caraciolo Ferreira

Graduação em Engenharia Florestal pela Universidade Federal Rural de Pernambuco - UFRPE (1983), mestrado (1988) e doutorado (1997) em Ciências Florestais, área de Manejo Florestal, pela Universidade Federal de Viçosa - UFV e Pós-Doutorado na Universidad de Córdoba-Espanha - UCO (2009). Professor Titular da UFRPE na área de manejo florestal. Bolsista de Produtividade em Pesquisa A do CNPq.

## Agostinho Lopes de Souza

Graduação em Engenharia Florestal pela Universidade Federal de Viçosa - UFV (1976), mestrado em Ciências Florestais pela UFV (1981) e doutorado em Engenharia Florestal pela Universidade Federal do Paraná - UFPR (1989). Professor titular aposentado da UFV. Ex-Bolsista de Produtividade em Pesquisa A do CNPq

## Rômulo Môra

Graduação em Engenharia Florestal pela Universidade Federal do Espírito Santo - UFES (2008), graduação em Estatística pela Universidade Federal de Mato Grosso - UFMT (2023), mestrado em Ciências Florestais pela UFES (2011) e doutorado em Engenharia Florestal pela Universidade Federal do Paraná - UFPR (2015). Professor Associado II na área de manejo florestal da Universidade Federal Rural de Pernambuco - UFRPE.

À memória do Prof. Dr. Eufrázio de Souza Santos pela sua dedicação à Biometria e Estatística Aplicada, sendo a nossa inspiração para escrever este livro.

# PREFÁCIO

Cursos de análise estatística multivariada, em todos os ramos das ciências, são cada vez mais comuns em instituições de ensino e de pesquisa. O livro sobre Análise de Agrupamento em Ciência Florestal com Aplicações em R, escrito pelos renomados professores e pesquisadores Rinaldo Luiz Caraciolo Ferreira, Agostinho Lopes de Souza e Rômulo Môra, foi elaborado para ampliar as alternativas, resolver e preencher lacunas de estudantes, pesquisadores e profissionais da engenharia florestal, apresentando atualizações de conceitos clássicos, aplicações criativas e uma abordagem inovadora para lidar com a complexidade das florestas, por meio da aplicação do método estatístico multivariado de análise de agrupamento. Então, considerando a diversidade de aplicações dos conhecimentos advindos dessa técnica, esta obra apresenta-se como uma excelente contribuição ao tema, com a profundidade e o rigor técnico necessários.

Além do brilhante caráter prático aplicado à obra pelos autores, o livro apresenta uma abordagem teórica similar ao nível dos melhores compêndios existentes na área da estatística de análise multivariada. O livro busca reunir conhecimentos básicos e especializados da Análise Multivariada de Agrupamento aplicada em estudos de vários temas da ciência florestal, assim como detalhes essenciais sobre a utilização e interpretação dos resultados a partir do aplicativo "R".

Cada capítulo foi estruturado de modo a propiciar a aprendizagem gradual, com explicações passo a passo, ilustrações conceituais e aplicações discutidas e avaliadas. Além disso, são incentivadas práticas de análise com dados reais, promovendo o desenvolvimento do pensamento crítico e da autonomia do leitor na condução de estudos com análise estatística multivariada.

Um dos principais diferenciais do livro reside na integração entre teoria e prática. Foram incluídos estudos de caso, conjuntos de dados reais e exercícios que estimulam o desenvolvimento do raciocínio analítico e da interpretação crítica. Além disso, são discutidos aspectos fundamentais, como a preparação dos dados, tipos e escalas de variáveis, validação dos resultados e limitações dos métodos, elementos indispensáveis para a aplicação responsável das técnicas apresentadas.

É importante fazer um destaque especial de que, logo na introdução do livro, os autores apresentam de forma brilhante as principais definições, conceitos, princípios básicos e objetivos da análise de agrupamento, assim como discorrem sobre as etapas para estabelecer os critérios de classificação, avaliação e validação do processo de agrupamento, objetivando obter a

homogeneidade dentro dos grupos, inclusive destacando a definição apropriada de uma análise em um espaço p-dimensional.

Denota-se que o livro foi elaborado para servir como um guia prático e teórico para estudantes de graduação e pós-graduação, pesquisadores e profissionais das áreas de engenharia florestal, ecologia e ciências ambientais que necessitam dos conhecimentos e da teoria de estatística multivariada, em especial de análise de agrupamento. Sua abordagem apresenta uma forma didática, adequada aos conhecimentos básicos de estatística, podendo, dessa forma, ser utilizado como texto em diversos cursos de graduação e pós-graduação. De fato, o livro serve como instrumento para o aprimoramento das análises de diversas pesquisas florestais e, principalmente, em inventários florestais.

Esta obra reúne uma amplitude e uma riqueza de conhecimento bastante ampla, pois apresenta e descreve com detalhes oito importantes medidas de distâncias, oito coeficientes de similaridades para as variáveis qualitativas e quatro para variáveis quantitativas. Descreve nove métodos hierárquicos aglomerativos e o método divisivo de Cavalli-Sforza. Mostra três técnicas de partição ou otimização, assim como as técnicas de modelos de densidade de mistura fina. Além disso, discute a validação dos métodos de agrupamento, a determinação do número ideal de grupos e implementação no R.

A estatística é a matemática aplicada à análise de dados observados, o que demonstra a importância da matemática na formação do engenheiro florestal. A floresta não é apenas a soma de suas árvores, mas um sistema dinâmico onde a estrutura, a composição florística, as propriedades do solo e as variáveis climáticas operam em uma constante e intrincada rede de dependências. A floresta é um conjunto populacional vivo multidimensional, cujas variáveis e dependências formam a sua estrutura e estão sintetizadas em uma matriz denominada de covariâncias ou de correlações. Ao isolar variáveis, corre-se o risco de ignorar as correlações latentes e os padrões emergentes que definem a resiliência e a produtividade de um ecossistema florestal.

Cabe ao engenheiro florestal descobrir e diagnosticar, através da análise estatística multidimensional, as informações desejadas em seu trabalho de pesquisa. A análise estatística multivariada é uma lente fundamental para estudar as complexidades dos ecossistemas e suas interações. Variáveis ambientais, biológicas e socioeconômicas frequentemente estão inter-relacionadas, e sua análise conjunta possibilita uma visão mais abrangente e integrada da realidade. Diferentemente da análise univariada, a qual estuda variáveis como se os ecossistemas atuassem em espaços individuais independentes, a análise multivariada acolhe o desafio de estudar as interdependências entre as variáveis, haja vista que o ecossistema florestal é, por definição, um

sistema de alta dimensionalidade. Tentar compreender tal complexidade através de análises isoladas é como tentar apreciar uma sinfonia ouvindo apenas as notas de um instrumento.

A análise de agrupamento é, sem dúvida, uma técnica multivariada muito importante na ciência florestal, especialmente em inventários voltados ao manejo, pois permite a obtenção de uma pós-estratificação multidimensional quando o processo amostral é a amostragem sistemática. Por outro lado, a partir dos resultados da estratificação multidimensional, o pesquisador florestal tem a possibilidade de aprofundar os estudos por meio da aplicação de outros métodos multivariados, dentro de cada ecossistema, tais como análise de variância multivariada, análise de componentes principais, análise de fatores, análise discriminante e, principalmente, a análise de correlação canônica. A aplicação da análise de correlação canônica, dentro de cada subpopulação ou estrato, permitirá estudar as correlações ambientais dentro de cada bioma obtido pela análise de agrupamento. A versatilidade da análise de agrupamento é o que a torna fascinante. Ela encontra morada em vários ramos da biologia e na engenharia de recursos naturais.

Além disso, considerando a abrangência do livro, a obra servirá como suporte para análise dos dados do Inventário Florestal Nacional (IFN), que começou a ser executada, cujas informações não serão tratadas por meio de uma simples análise univariada, haja vista a grande diversidade florística e ambiental das florestas brasileiras. A aplicação da análise multivariada de agrupamento, quando usada a amostragem sistemática, como no caso do IFN, configura-se como um importante suporte técnico e científico para obter uma pós-estratificação multidimensional da floresta, a fim de separar e classificar tipologias florestais diversas.

É notória a complexidade das florestas tropicais, logo, o emprego da análise multivariada pode alavancar substancialmente o potencial de resposta dos inventários florestais, ampliando as possibilidades para um melhor uso e conservação das florestas. A partir da experiência demonstrada pelos autores, é possível observar, no texto, a boa qualidade dos exemplos e dos exercícios resolvidos. Além disso, o livro apresenta uma coleção de técnicas e metodologias estatísticas reunidas numa só obra. Dessa forma, professores, pesquisadores e usuários podem compreender e implementar facilmente cada uma das técnicas e ferramentas abordadas. Finalmente, esta obra pode ser destinada a todos aqueles que utilizam e estudam as técnicas multivariadas e seus benefícios nas diversas áreas de conhecimento.

Sinto-me honrado ao prefaciar esta obra e, destarte, gostaria de expressar a minha alegria e entusiasmo, pois tenho a plena convicção de que os numerosos e dedicados estudantes, pesquisadores e profissionais da Engenharia Florestal terão a maior facilidade, após consultar este livro, de obter resultados excepcionais nos seus trabalhos de pesquisa.

Parabenizo os autores pelas excelentes contribuições apresentadas, pois, ao concluir a leitura desta obra, confesso o meu entusiasmo pela amplitude dos conhecimentos adicionais apresentados e, indubitavelmente, este livro contribuirá para o avanço das análises estatísticas multivariadas na Engenharia Florestal brasileira, fortalecendo a tomada de decisão baseada em evidências reais e promovendo o uso sustentável dos recursos naturais.

Belém, 23 de julho de 2026

Professor Dr. Waldenei Travassos de Queiroz

# Sumário

# Introdução

A análise de agrupamento é, sem dúvida, uma das técnicas multivariadas mais utilizadas em Ciência Florestal, constatando-se sua aplicação em áreas como conservação da natureza, manejo florestal, silvicultura, técnicas e operações florestais e tecnologia e utilização de produtos florestais.

A análise tem por finalidade reunir, por algum critério de classificação, elementos ou unidades amostrais em grupos, de tal forma que existam homogeneidade dentro do grupo e heterogeneidade entre grupos (Johnson; Wichern, 2007; Mardia et al., 2023).

Na literatura, podem-se encontrar como sinônimos de análise de agrupamento, entre outros menos comuns, *cluster analysis* (termo em inglês), análise de *cluster* e análise de conglomerados (Queiroz, 2021). Vale ressaltar que dificuldades computacionais retardaram seu desenvolvimento até o final da década de 1950, quando a informatização resultou na proliferação de técnicas de agrupamento (Bailey, 1994).

Para a realização de uma análise de agrupamento, Malhotra (2019) sugere as etapas conforme a Figura 1.


**Figura 1.** Etapas para realização de uma análise de agrupamento. Adaptada de Malhotra (2019).

Na formulação do problema, deve-se decidir sobre a seleção dos elementos a serem agrupados, variáveis a serem utilizadas na análise e se elas devem ser padronizadas.

# 1. Conceito de Agrupamento

O termo agrupamento é o ato ou procedimento de reunir objetos ou elementos em grupos, isto é, o ato ou efeito de grupar. Assim, agrupamento é um conjunto de elementos contíguos de uma população estatística, como um grupo de indivíduos pertencentes a uma dada comunidade, o que dá uma ideia de distância física entre os elementos do conjunto, e não uma ideia da distância funcional entre eles (Everitt et al., 2011).

Logo, dado um conjunto **E** de **n** elementos, em que cada elemento é representado por um conjunto de medidas de **p** variáveis, deseja-se determinar os subconjuntos ou agrupamentos de elementos em que pode ser decomposto o conjunto **E**.

Assim, uma definição mais apropriada de agrupamento deve abordar a ideia de espaço **p**-dimensional e, portanto, deve englobar o conceito de partição do espaço **p**-dimensional em regiões adjacentes ou contíguas com considerável densidade de pontos (objetos ou indivíduos). Logo, agrupamento é um subconjunto de elementos de uma população, cuja imagem em um espaço euclidiano **p**-dimensional ($\mathbf{I_P}$) é formada por um conjunto de pontos de alta densidade, separado de outros conjuntos por regiões de baixa densidade.

# 2. Princípios Básicos

Considerando-se um conjunto **E** de **n** elementos pertencentes a populações distintas e que, em cada um dos elementos, são feitas **p** observações ou medidas referentes às **p** variáveis ou características, logo, os elementos de **E** poderão ser descritos por um vetor ponto de espaço **p**-dimensional, como:

$$X_{ij} = \left(X_{i1}, X_{i2}, \ldots, X_{ip}\right), \text{ para i = 1, 2, ..., n; j = 1, 2, ..., p.}$$

Em que: $\mathbf{X_{ij}}$ = **p** medida da **j**-ésima característica no **i**-ésimo elemento pertencente ao espaço **p**-dimensional.

Portanto, $X_{ij}$ forma um conjunto de **n** pontos no espaço euclidiano **p**-dimensional.

A questão primordial é obter grupos ou subconjuntos do conjunto **E**, de forma que cada elemento de um subconjunto pertença a um, e somente a um, subconjunto particular. Pela teoria dos conjuntos, pode-se dizer que a interseção entre dois subconjuntos tem de ser um conjunto vazio. Considerando-se um inteiro **g**, tal que $\mathbf{g \leq n}$, o problema reside em determinar **g** grupos, de forma que cada um de seus elementos pertença somente a um dado grupo ou subconjunto de **E**.

A obtenção dos **g** agrupamentos pode ser vista como a determinação de uma partição do espaço **p**-dimensional em regiões $\mathbf{R_k}$, em que: $\mathbf{R_i} \cap \mathbf{R_j} = 0$, para i = j, de modo que, se um determinado ponto $\boldsymbol{x_i} = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)$ pertence a uma região $\mathbf{R_k}$ (k = 1, 2, ..., **g**), não

pertencerá a nenhuma outra região. Dessa forma, a união (U) dos elementos pertencentes às **k** regiões é igual ao espaço **p**-dimensional ($\mathbf{I_P}$), ou $\bigcup_{k=1}^{g} \boldsymbol{R_k} = \boldsymbol{I_p}$. Assim, os elementos pertencentes ao subconjunto $\mathbf{R_k}$ constituem o que se conceitua como agrupamento.

## 3. Objetivos da Análise de Agrupamento

Como objetivos da utilização da análise de agrupamento, podem-se citar:

- Redução de grande massa de dados em grupos, para permitir sua análise;
- Análise exploratória de dados;
- Formulação de hipóteses sobre estrutura de dados; e
- Serve como técnica alternativa à análise estatística clássica.

Vale ressaltar que, com os avanços computacionais, é evidente que os objetivos e as finalidades da análise de agrupamento são constantemente ampliados.

O uso mais frequente da análise é no desenvolvimento de classificação. No entanto, a análise de agrupamento não é um procedimento padronizado e há armadilhas associadas ao seu uso inadequado, portanto, cuidados são necessários em sua aplicação (Gore Jr., 2000).

O processo de realização da técnica de agrupamento envolve basicamente três etapas: a primeira se relaciona com a estimativa de uma função de agrupamento entre as unidades amostrais; a segunda, com a adoção de um algoritmo para formação dos grupos; e, finalmente, a terceira, com a definição do número de grupos a serem formados.

# Capítulo 1

## Função de Agrupamento - Medida de Distância e Similaridade

A resposta à questão relativa à formação de grupos ou subconjuntos do conjunto **E** é obtida mediante a aplicação de um determinado critério de agrupamento. A maioria dos algoritmos emprega, para alcançar tal solução, medidas de similaridade ou de distância entre os elementos do conjunto **E** (Gama, 2022). Essas medidas, segundo Orlóci (1978) e Gama (2022), constituem o "*input*" para os vários algoritmos e definem uma função dos valores dos vetores representativos dos elementos de **E**, para os quais se procura determinar uma medida de similaridade ou de distância, sendo comumente denominadas, respectivamente, coeficiente de similaridade e medida de distância.

O termo semelhança, usado como sinônimo de similaridade, é uma propriedade mensurável de objetos, ou grupos de objetos, enquanto dissimilaridade é expressa como uma função das características que os objetos possuem, podendo os objetos representar espécies individuais, um povoamento ou alguma outra entidade (Orlóci, 1978).

Considere a matriz **X**, denominada de matriz de dados ou matriz primária:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{bmatrix}$$

Em que: **i** = 1, 2, ..., **p** espécies na amostra; e **j** = 1, 2, ..., **n** parcelas.

A maioria dos algoritmos utiliza, a partir do conjunto dos vetores, $x_i = [x_{i1}, x_{i2}, \cdots, x_{in}]$, representados pela matriz de dados $\mathbf{X}_{(\mathbf{p\ x\ n})}$, uma função denominada "função de agrupamento", cujos valores, calculados com base na matriz **X**, podem ser representados por uma matriz simétrica de distâncias (**D**) ou de similaridades (**S**). A matriz **D** ou a matriz **S** é de ordem **n** x **n** e tem todos os elementos da diagonal principal, respectivamente, 0 ou 1.

$$D = \begin{bmatrix} 0 & \cdots & D_{1n} \\ \vdots & \ddots & \vdots \\ D_{n1} & \cdots & 0 \end{bmatrix} \text{ ou } S = \begin{bmatrix} 1 & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{n1} & \cdots & 1 \end{bmatrix}$$

Em que: $\mathbf{D_{1n}}$ = distância entre o objeto 1 e o **n**-ésimo; $\mathbf{D_{n1}}$ = distância entre o **n**-ésimo e o objeto 1 = $\mathbf{D_{1n}}$; $\mathbf{D_{ii}}$ = 0 = distância entre o **i**-ésimo e ele mesmo; $\mathbf{S_{1n}}$ = similaridade entre o objeto 1 e o

**n**-ésimo; $\mathbf{S_{n1}}$ = similaridade entre o **n**-ésimo e o objeto 1 = $\mathbf{S_{1n}}$; **Sii** = 1 = similaridade entre o **i**-ésimo objeto e ele mesmo; **i** = 1, ...., n.

Para o melhor entendimento dos algoritmos de agrupamento, é necessário entender o que são o coeficiente de similaridade e a medida de distância entre dois elementos de **E**, isto é, entre dois pontos quaisquer do espaço euclidiano **p**-dimensional. Outro problema é aquele relativo à escolha das **p** variáveis que definem o vetor ponto de um espaço **p**-dimensional representativo de um elemento de **E**.

No caso específico de análise de vegetação, exemplo no presente capítulo, o problema primordial consiste na definição da matriz primária de dados. Pode-se conceber a matriz **X** de dados originais ou matriz primária como uma tabela de dupla entrada, em que as espécies são arranjadas em linhas e as parcelas, em colunas. Portanto, as colunas constituem as comunidades de plantas ou árvores que serão comparadas entre si, podendo cada uma delas provir de um conjunto **E** de parcelas ou de uma única parcela (Matteucci; Colma, 1982). Na interseção das linhas e colunas, situa-se o valor de abundância (densidade) ou de presença de cada atributo da comunidade. Se esse valor provém de um conjunto de unidades de amostra, então ele representa uma estimativa da média do atributo para a espécie considerada na dita comunidade.

Em se tratando de comparações numéricas de comunidades de plantas, é usual adotar técnicas estatísticas, das quais, partindo da matriz **X** primária (espécie/amostra), obtêm-se matrizes secundárias de similaridade ou de distância. Essas matrizes passam a constituir o "***input***" (entrada) de quase todos os sistemas numéricos e sistemas formais de classificação e ordenação de vegetação. Definidos os critérios de similaridade ou dissimilaridade entre parcelas de amostras, ou entre espécies na população amostral, classifica-se e ordena-se a vegetação.

Para melhor compreensão das funções de agrupamentos e das medidas de distância e de similaridade, é necessário entender o que é espaço amostral.

## 1.1. Espaço Amostral

Na Tabela 1, está representado um levantamento amostral da vegetação da Mata da Silvicultura, município de Viçosa-MG. Esses dados serão utilizados como exemplo ao longo deste livro para ilustrar a aplicação dos diferentes métodos, análises e procedimentos apresentados.

Neste exemplo, a amostra é constituída de 17 espécies (**p** = 17) e de 11 parcelas (**n** = 11). Logo, os dados da Tabela 1 podem ser representados pela seguinte matriz de dados X:

**Tabela 1**. Número de indivíduos de 17 espécies da Mata da Silvicultura, em parcelas de 20 x 50 m, município de Viçosa-MG

| **Espécies** | **Parcelas** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
| *Casearia decandra* Jacq. | 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 |
| *Anadenanthera peregrina* (L.) Speg. | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 |
| *Apuleia leiocarpa* (Vog.) J. F. Macbr. | 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 |
| *Mabea fistulifera* Mart. | 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 |
| *Anadenanthera macrocarpa* (Benth.) Brenan. | 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| *Platypodium elegans* Vog. | 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 |
| *Machaerium floridum* (Mart. ex Benth.) Ducke | 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 |
| *Copaifera langsdorffii* Desf. | 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 |
| *Ocotea pretiosa* (Nees) Mez. | 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 |
| *Cabralea cangerana* Saldanha | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 |
| *Piptadenia gonoacantha* (Mart.) J. F. Macbr. | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 |
| *Dalbergia nigra* (Vell.) Allemão ex Benth. | 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| *Luehea divaricata* Mart. | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 |
| *Cecropia hololeuca* Miq. | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| *Melanoxylum brauna* Schott. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| *Cedrela fissilis* Vell. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *Croton floribundus* Spreng. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

$$
X = \begin{vmatrix}
8 & 1 & 27 & 0 & 1 & 9 & 2 & 3 & 22 & 15 & 7 \\
0 & 0 & 0 & 0 & 0 & 0 & 12 & 1 & 17 & 1 & 9 \\
3 & 9 & 4 & 6 & 22 & 9 & 5 & 2 & 7 & 4 & 4 \\
6 & 3 & 3 & 4 & 29 & 12 & 0 & 4 & 4 & 4 & 4 \\
0 & 12 & 0 & 1 & 0 & 0 & 1 & 0 & 2 & 0 & 0 \\
0 & 0 & 1 & 1 & 9 & 1 & 0 & 0 & 5 & 11 & 1 \\
0 & 0 & 10 & 1 & 9 & 2 & 1 & 0 & 0 & 11 & 5 \\
1 & 1 & 0 & 2 & 1 & 13 & 0 & 0 & 0 & 3 & 1 \\
2 & 0 & 2 & 2 & 2 & 6 & 0 & 5 & 0 & 2 & 2 \\
1 & 0 & 0 & 2 & 0 & 0 & 1 & 6 & 2 & 3 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 1 & 0 & 5 \\
5 & 0 & 7 & 0 & 5 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 5 & 2 \\
7 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{vmatrix}
$$

Em que: cada linha corresponde a um vetor espécie ($\mathbf{X_p}$) e cada coluna, a um vetor parcela ($\mathbf{X_n}$).

Considerando as **n** (11) parcelas como um conjunto de pontos no espaço **p**-dimensional ($\mathbf{I_{17}}$), as posições relativas desses pontos estão na proporção das diferenças na composição de espécies nas parcelas (Orlóci, 1978). Os pontos e a função de semelhança $\boldsymbol{f}(\mathbf{j},\mathbf{k})$ que medem a posição espacial dos pontos constituem o **espaço amostral**. Essa função $\boldsymbol{f}(\mathbf{j},\mathbf{k})$ pode representar uma distância (dissimilaridade) ou pode ser uma medida de similaridade. Nesse caso, diz-se que $\boldsymbol{f}(\mathbf{j},\mathbf{k})$ é um parâmetro espacial.

## 1.1.1. Implementação de Dados em R para Análise de Agrupamento

Após a compreensão dos conceitos teóricos da análise de agrupamento, é crucial aplicar esses conhecimentos na prática, utilizando o *software* R. Reconhecido por sua robustez e flexibilidade, o R é uma ferramenta essencial para a implementação de técnicas de análise multivariada na comunidade científica.

Nesta seção, demonstraremos como manipular, armazenar e analisar dados de densidade de espécies florestais, utilizando a matriz de dados **X**, conforme apresentada na Tabela 1. O domínio dessas técnicas é fundamental para pesquisadores em estudos de vegetação e ecologia florestal que buscam aplicar métodos quantitativos.

## a) Instalação e Primeiros Passos no Ambiente R

Para começar a trabalhar com R, é fundamental ter o *software* instalado. O R e o RStudio (uma interface de desenvolvimento integrada - IDE) são ferramentas gratuitas e de código aberto, amplamente utilizadas.

## a.1) Instalação do R

Para realizar a instalação:

**a)** Acesse o site oficial do CRAN (The Comprehensive R Archive Network) para baixar o R: https://cran.r-project.org/.

**b)** Selecione o link correspondente ao seu sistema operacional (Windows, macOS ou Linux) e siga as instruções para baixar o instalador (Figura 2).

**c)** Execute o arquivo baixado e siga as etapas do assistente de instalação. Na maioria dos casos, as opções padrão são adequadas.

**d)** Abrir o programa (Figura 3).


**Figura 2.** Sequência de etapas para *download* e instalação do R no sistema operacional Windows a partir do CRAN.

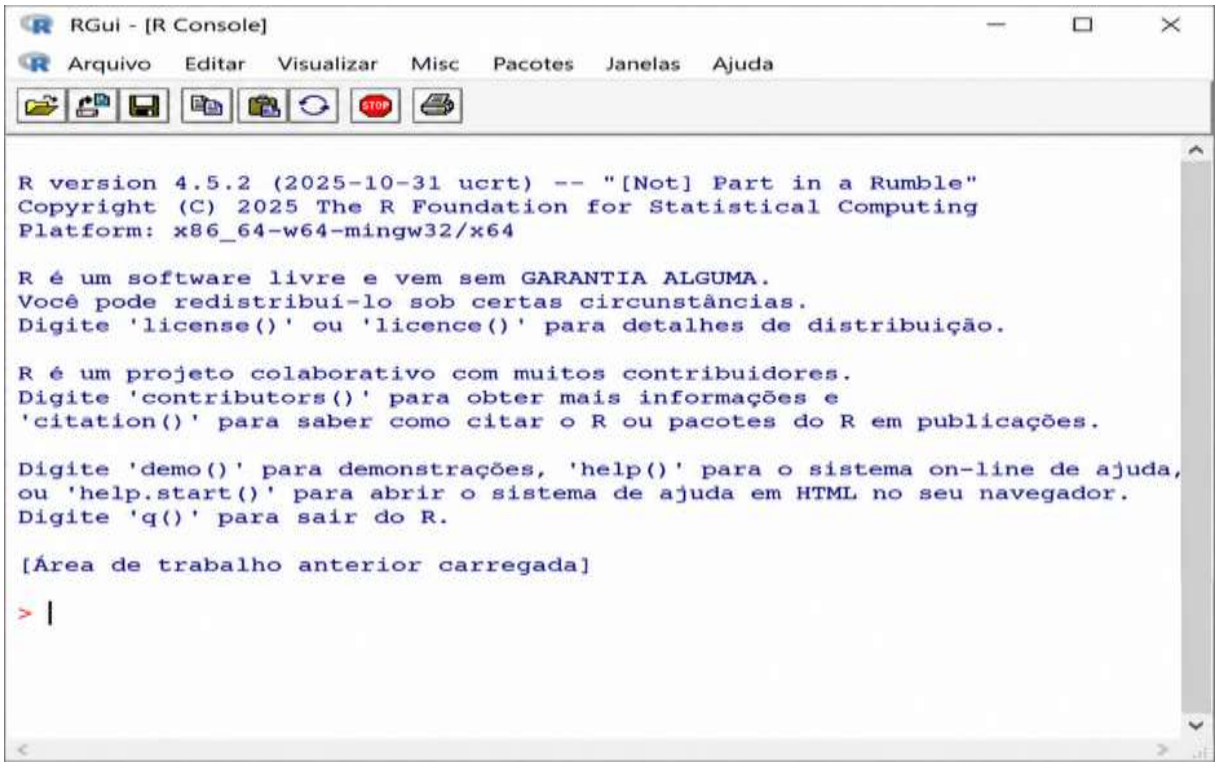



**Figura 3.** Tela inicial do ambiente R (RGui) após a abertura do *software* (Exemplo para Windows).

## a.2) Instalação do RStudio Desktop

**1)** Acesse o site oficial da Posit para baixar o RStudio Desktop: https://posit.co (Figura 4).

**2)** Execute o arquivo baixado e siga as etapas do assistente de instalação. As opções padrão geralmente são suficientes.

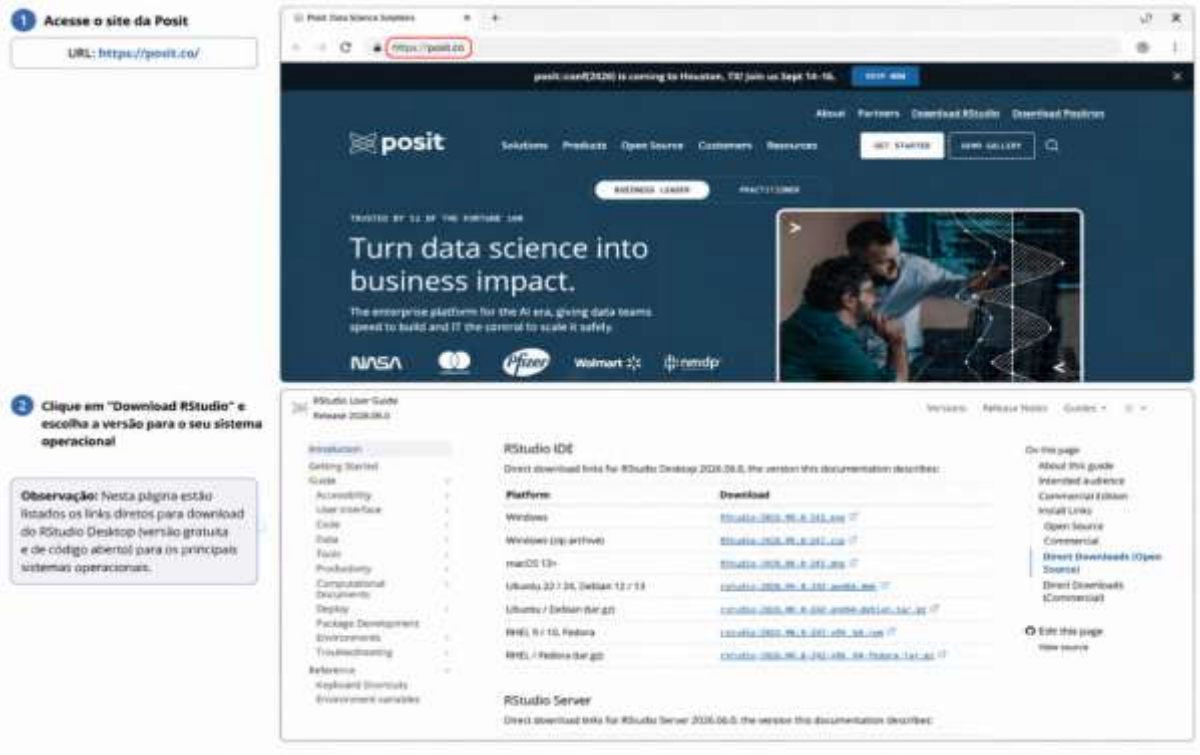



**Figura 4.** Sequência de etapas para *download* e instalação do RStudio no sistema operacional Windows a partir do CRAN.

## a.3) Primeiro Contato com o RStudio

Após a instalação de ambos os *softwares* (para uso do RStudio é obrigatória a instalação do R), abra o RStudio. Você verá uma interface dividida em painéis, conforme a Figura 5.

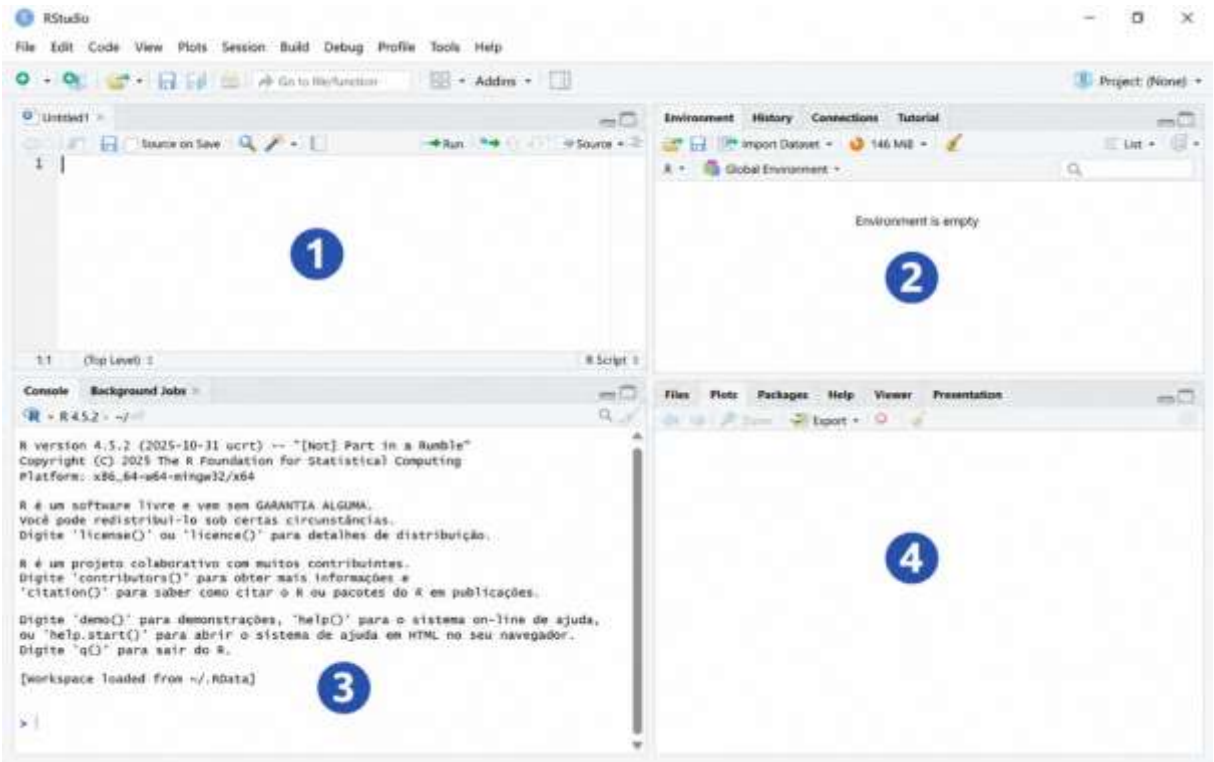



**Figura 5.** Principais componentes da interface do RStudio IDE: (1) Editor de scripts; (2) Ambiente e Histórico; (3) Console do R; e (4) Arquivos, Gráficos, Pacotes e Ajuda.

Para testar se a instalação foi bem-sucedida e se o RStudio está se comunicando corretamente com o R, digite uma operação simples no painel 'Console' e pressione *Enter*: (Figura 6).

```
2 + 2
```

O R retornará [1] 4, indicando o resultado da operação. Isso significa que o R está funcionando e pronto para ser utilizado.

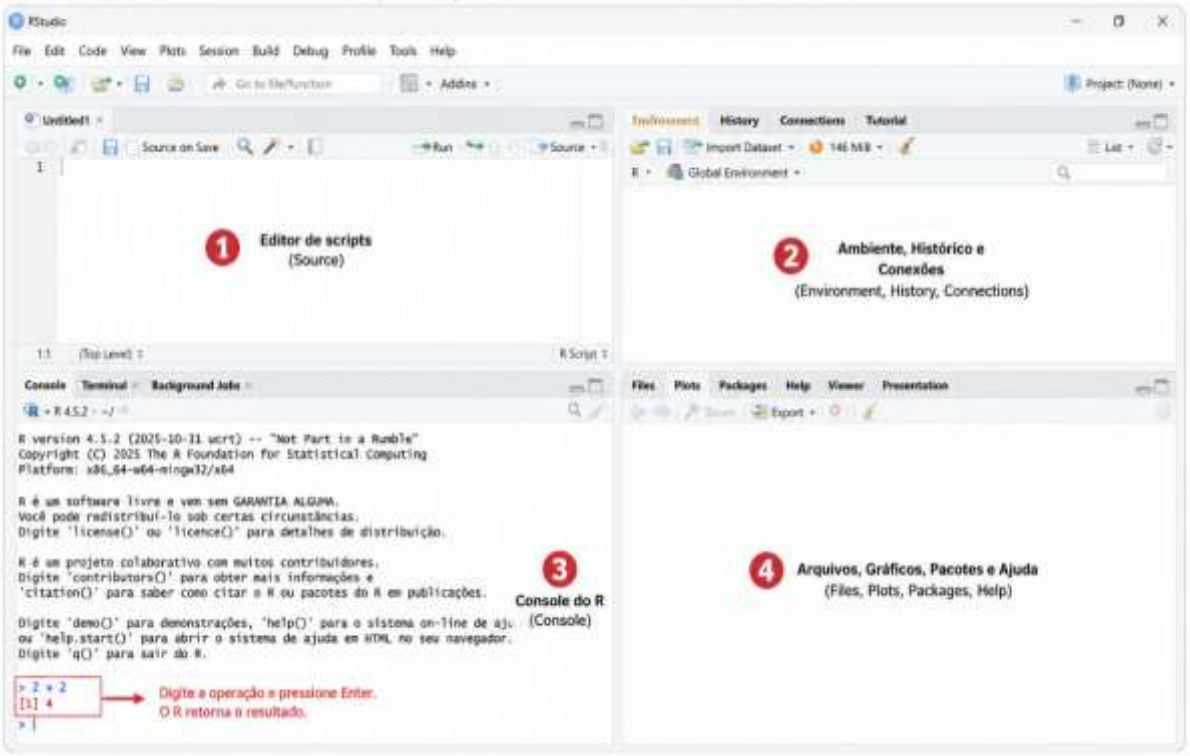



**Figura 6.** Verificação da instalação do RStudio utilizando uma operação simples no Console.

## b) Fundamentos Essenciais para Programação em R

Para iniciantes no ambiente R, é imprescindível compreender algumas regras básicas de sintaxe e operação. Essas diretrizes garantem a execução correta dos *scripts* e facilitam o desenvolvimento de análises:

**1) Diferenciação de Letras (*Case Sensitivity*):** O R distingue letras maiúsculas de minúsculas. Isso significa que Matriz e matriz são tratados como objetos distintos. Por exemplo:

```
Matriz <- 10  #Atribui valor 10 para o objeto Matriz
matriz <- 5 # Atribui o valor 5 para o objeto matriz
print(Matriz) # Retorna 10
print(matriz) # Retorna 5
```

**2) Estrutura de Dados:** Em qualquer conjunto de dados no R, as colunas representam as variáveis (ex: espécies, altura, diâmetro) e as linhas representam as unidades amostrais ou elementos (ex: indivíduos).

**3) Nomenclatura de Variáveis:** A escolha de nomes para as variáveis deve seguir algumas convenções para evitar erros e garantir a legibilidade:

- **Símbolos Matemáticos:** Evite usar símbolos de operação (+, -, *, /, ^) nos nomes das variáveis, pois o R interpretará como uma operação matemática.

- **Espaços:** Não utilize espaços. Prefira o *underline* (_) ou o ponto (.) para separar palavras (ex: densidade_especie ou densidade.especie).
- **Início do Nome:** Nomes de variáveis devem começar com letras. Evite iniciar com números ou símbolos especiais.

**4) Comentários:** Utilize o símbolo de cerquilha (#) para adicionar comentários ao seu código. Qualquer texto após este símbolo é ignorado pelo console, sendo útil para documentar e explicar a funcionalidade de cada comando. Exemplo:

```
# Este é um comentário explicando a linha de código abaixo
resultado <- 2 + 2 # Atribui a soma de 2 + 2 à variável resultado
```

**5) Uso de Pacotes:** Pacotes são coleções de funções e dados que estendem as capacidades do R. Seu uso envolve duas etapas principais:

- **Instalação:** A instalação de um pacote é um processo único por computador. No console do R, utilize *install.packages*("nome_do_pacote"). No RStudio, você pode usar o painel 'Packages' ou o mesmo comando no console.
- **Carregamento:** Após a instalação, o pacote deve ser carregado em cada nova sessão do R para que suas funções fiquem disponíveis. Utilize *library*(nome_do_pacote) para isso.

**6) Atribuição de Valores:** O operador preferencial para atribuir um valor ou resultado a uma variável no R é <- (ex: x <- 10). Embora o sinal de igual (=) possa funcionar em alguns contextos, o <- é o padrão da linguagem e é recomendado para consistência.

**7) Separador Decimal:** O R adota o padrão internacional, utilizando o ponto (.) como separador decimal. A vírgula (,) é utilizada para separar argumentos dentro de uma função.

**8) Diretório de Trabalho:** Antes de importar dados externos (como arquivos .csv ou .xlsx), é fundamental definir o diretório de trabalho, garantindo que o R esteja "olhando" para a pasta correta onde o arquivo está salvo. Isso pode ser feito de duas formas:

- No console do R: Utilize a função *setwd* ("caminho/para/sua/pasta") (substitua pelo caminho real da pasta).
- No RStudio: Navegue até *Session* > *Set Working Directory* > *Choose Directory* e selecione a pasta desejada.

**9) Fechamento de Parênteses e Aspas:** Toda estrutura é aberta e fechada com parênteses ( ). Em textos, aplicam-se as aspas (""). Se um parêntese ou aspa não for fechado, o R exibirá um sinal de + no console, indicando que o comando está incompleto.

**10) Organização de Arquivos:** Para evitar problemas com caminhos de arquivo e facilitar a reprodutibilidade, é altamente recomendável que todos os arquivos de dados, *scripts*

e resultados de uma análise estejam localizados na mesma pasta de trabalho. Isso simplifica a importação e exportação de dados no R.

## c) Obtendo Ajuda no R

O R possui um sistema de ajuda robusto. Se você precisar de informações sobre uma função específica, pode usar os seguintes comandos:

- *help*(nome_da_funcao): Abre a documentação completa da função.
- ?nome_da_funcao: Atalho para o comando *help*().

Por exemplo, para saber mais sobre a função *mean*() (média):

```
?mean
```

## d) Limpeza do Ambiente de Trabalho

É uma boa prática limpar o ambiente de trabalho regularmente para evitar conflitos entre objetos de análises anteriores. Você pode fazer isso de diversas formas:

**1) Limpar o console:**

- No RStudio: Clique no ícone da vassoura (*Clear Console*) no painel do Console, ou use o atalho Ctrl + L (Windows/Linux) ou Cmd + L (macOS).
- No R tradicional: Use o atalho Ctrl + L (Windows/Linux) ou Cmd + L (macOS).

**2) Remover todos os objetos do ambiente:**

Esse comando remove todas as variáveis e funções definidas na sessão atual do R, deixando o ambiente limpo para uma nova análise.

- No RStudio: Clique no ícone da vassoura (*Clear Objects*) no painel Environment, ou use o comando:

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

- No R tradicional: Use o comando:

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

## 1.1.2. Entrada de Dados no R para Análise de Agrupamento

Nesta seção abordamos como inserir e carregar dados no ambiente R, um passo fundamental para qualquer análise. Veremos desde a criação manual de uma matriz de dados

até a importação de arquivos de planilhas, garantindo que seus dados estejam prontos para processamento.

## a) Criação Manual de Dados no R

A primeira forma de trabalhar com dados no R é criá-los diretamente. Para nossa análise de agrupamento, usaremos a função *data.frame*() para construir uma estrutura de tabela, ideal para dados estatísticos. O exemplo a seguir mostra como criar a matriz de densidade, que chamaremos de X, diretamente no console do R.

### 1) Criando a Matriz Completa

O código abaixo cria a matriz de densidade, na qual as linhas representam as espécies e as colunas representam as parcelas (P1 a P11). Note que as espécies são inicialmente armazenadas em um vetor separado e, depois, incorporadas ao *data.frame*.

```
# Nomes das espécies
especies <- c("Casearia decandra Jacq.", "Anadenanthera peregrina (L.) Speg.",
              "Apuleia leiocarpa (Vog.) J. F. Macbr.", "Mabea fistulifera Mart.",
              "Anadenanthera macrocarpa (Benth.) Brenan.", "Platypodium elegans Vog.",
              "Machaerium floridum (Mart. ex Benth.) Ducke", "Copaifera langsdorffii Desf.",
              "Ocotea pretiosa (Nees) Mez.", "Cabralea cangerana Saldanha",
              "Piptadenia gonoacantha (Mart.) J. F. Macbr.", "Dalbergia nigra (Vell.) Allemão ex Benth.",
              "Luehea divaricata Mart.", "Cecropia hololeuca Miq.",
              "Melanoxylum brauna Schott.", "Cedrela fissilis Vell.",
              "Croton floribundus Spreng.")

# Criando a matriz de dados completa
matriz_densidade <- data.frame(Especies = especies,
                               P1 = c(8,0,3,6,0,0,0,1,2,1,0,5,0,7,0,0,0),
                               P2 = c(1,0,9,3,12,0,0,1,0,0,0,0,0,0,0,0,0),
                               P3 = c(27,0,4,3,0,1,10,0,2,0,0,7,1,0,0,0,1),
                               P4 = c(0,0,6,4,1,1,1,2,2,2,0,0,0,0,0,0,0),
                               P5 = c(1,0,22,29,0,9,9,1,2,0,0,5,0,0,0,0,0),
                               P6 = c(9,0,9,12,0,1,2,13,6,0,0,0,0,1,0,0,0),
                               P7 = c(2,12,5,0,1,0,1,0,0,1,6,0,2,0,0,1,0),
                               P8 = c(3,1,2,4,0,0,0,0,5,6,0,0,0,1,0,0,0),
                               P9 = c(22,17,7,4,2,5,0,0,0,2,1,0,0,0,0,0,0),
                               P10 = c(15,1,4,4,0,11,11,3,2,3,0,1,5,0,2,0,0),
                               P11 = c(7,9,4,4,0,1,5,1,2,1,5,0,2,0,1,0,0))

# Visualizando a estrutura dos dados
str(matriz_densidade)
```

```
'data.frame':	17 obs. of  12 variables:
 $ Especies: chr  "Casearia decandra Jacq." "Anadenanthera peregrina (L.) Speg." "Apuleia leiocarpa (Vog.) J. F. Macbr." "Mabea fistulifera Mart." "Anadenanthera macrocarpa (Benth.) Brenan." "Platypodium elegans Vog." "Machaerium floridum (Mart. ex Benth.) Ducke" "Copaifera langsdorffii Desf." "Ocotea pretiosa (Nees) Mez." "Cabralea cangerana Saldanha" "Piptadenia gonoacantha (Mart.) J. F. Macbr." "Dalbergia nigra (Vell.) Allemão ex Benth." "Luehea divaricata Mart." "Cecropia hololeuca Miq." "Melanoxylum brauna Schott." "Cedrela fissilis Vell." "Croton floribundus Spreng." “
```

```
 $ P1     : num  8 0 3 6 0 0 0 1 2 1 ...
 $ P2     : num  1 0 9 3 12 0 0 1 0 0 ...
 $ P3     : num  27 0 4 3 0 1 10 0 2 0 ...
 $ P4     : num  0 0 6 4 1 1 1 2 2 2 ...
 $ P5     : num  1 0 22 29 0 9 9 1 2 0 ...
 $ P6     : num  9 0 9 12 0 1 2 13 6 0 ...
 $ P7     : num  2 12 5 0 1 0 1 0 0 1 ...
 $ P8     : num  3 1 2 4 0 0 0 0 5 6 ...
 $ P9     : num  22 17 7 4 2 5 0 0 0 2 ...
 $ P10    : num  15 1 4 4 0 11 11 3 2 3 ...
 $ P11    : num  7 9 4 4 0 1 5 1 2 1 ...
```

```
# Visualizando as primeiras linhas da matriz
head(matriz_densidade)
```

| Espécies | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Casearia decandra* Jacq. | 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 |
| *Anadenanthera peregrina* (L.) Speg. | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 |
| *Apuleia leiocarpa* (Vog.) J. F. Macbr. | 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 |
| *Mabea fistulifera* Mart. | 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 |
| *Anadenanthera macrocarpa* (Benth.) Brenan | 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| *Platypodium elegans* Vog. | 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 |
| *Machaerium floridum* (Mart. ex Benth.) Ducke | 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 |
| *Copaifera langsdorffii* Desf. | 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 |
| *Ocotea pretiosa* (Nees) Mez. | 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 |
| *Cabralea cangerana* Saldanha | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 |
| *Piptadenia gonoacantha* (Mart.) J. F. Macbr. | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 |
| *Dalbergia nigra* (Vell.) Allemão ex Benth. | 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| *Luehea divaricata* Mart. | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 |
| *Cecropia hololeuca* Miq. | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| *Melanoxylum brauna* Schott. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| *Cedrela fissilis* Vell. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *Croton floribundus* Spreng. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

### 2) Criando a Matriz X (Numérica)

Para realizar a análise de agrupamento, é necessário utilizar uma matriz puramente numérica. O código abaixo demonstra como criar a matriz X, removendo a coluna de texto (espécies) e renomeando as linhas como E1, E2, ..., E17 para facilitar a identificação.

```
# Criando a matriz X apenas com dados numéricos
X <- data.frame(P1 = c(8, 0, 3, 6, 0, 0, 0, 1, 2, 1, 0, 5, 0, 7, 0, 0, 0),
                P2 = c(1, 0, 9, 3, 12, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0),
                P3 = c(27, 0, 4, 3, 0, 1, 10, 0, 2, 0, 0, 7, 1, 0, 0, 0, 1),
                P4 = c(0, 0, 6, 4, 1, 1, 1, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0),
                P5 = c(1, 0, 22, 29, 0, 9, 9, 1, 2, 0, 0, 5, 0, 0, 0, 0, 0),
                P6 = c(9, 0, 9, 12, 0, 1, 2, 13, 6, 0, 0, 0, 0, 1, 0, 0, 0),
                P7 = c(2, 12, 5, 0, 1, 0, 1, 0, 0, 1, 6, 0, 2, 0, 0, 1, 0),
                P8 = c(3, 1, 2, 4, 0, 0, 0, 0, 5, 6, 0, 0, 0, 1, 0, 0, 0),
                P9 = c(22, 17, 7, 4, 2, 5, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0),
                P10 = c(15, 1, 4, 4, 0, 11, 11, 3, 2, 3, 0, 1, 5, 0, 2, 0, 0),
                P11 = c(7, 9, 4, 4, 0, 1, 5, 1, 2, 1, 5, 0, 2, 0, 1, 0, 0))

# Nomeando as linhas como E1, E2, ..., E17
rownames(X) <- paste0("E", 1:17)

# Visualizando a matriz X
X
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| E1 | 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 |
| E2 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 |
| E3 | 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 |
| E4 | 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 |
| E5 | 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| E6 | 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 |
| E7 | 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 |
| E8 | 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 |
| E9 | 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 |
| E10 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 |
| E11 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 |
| E12 | 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| E13 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 |
| E14 | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| E15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| E16 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| E17 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## b) Importação de Dados de Planilhas para o R

É comum que os dados já estejam organizados em planilhas (como Excel, LibreOffice Calc ou Apple Numbers). Este guia prático demonstra como estruturar seus dados e importá-los de forma eficiente para o R.

### 1) Estrutura dos Dados na Planilha

Para um fluxo de trabalho organizado, é fundamental estruturar seus dados corretamente. Crie um arquivo de planilha (por exemplo, dados_densidade.xlsx ou dados_densidade.ods) com duas abas distintas. Nas Tabelas 2 e 3 ilustramos como estruturar seus dados nas duas planilhas.

**Planilha 1: Matriz_Densidade (Tabela 2)**

Esta aba deve conter a matriz de dados completa, incluindo uma coluna de identificação (como espécies) e as variáveis numéricas (P1, P2, ..., P11).

**Tabela 2.** Estruturação de dados na planilha 1(Matriz_Densidade – com coluna de espécies)

| Espécies | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Casearia decandra* Jacq. | 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 |
| *Anadenanthera peregrina* (L.) Speg. | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 |
| *Apuleia leiocarpa* (Vog.) J. F. Macbr. | 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 |
| *Mabea fistulifera* Mart. | 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 |
| *Anadenanthera macrocarpa* (Benth.) Brenan. | 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| *Platypodium elegans* Vog. | 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 |
| *Machaerium floridum* (Mart. ex Benth.) Ducke | 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 |
| *Copaifera langsdorffii* Desf. | 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 |
| *Ocotea pretiosa* (Nees) Mez. | 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 |
| *Cabralea cangerana* Saldanha | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 |
| *Piptadenia gonoacantha* (Mart.) J. F. Macbr. | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 |
| *Dalbergia nigra* (Vell.) Allemão ex Benth. | 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| *Luehea divaricata* Mart. | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 |
| *Cecropia hololeuca* Miq. | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| *Melanoxylum brauna* Schott. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| *Cedrela fissilis* Vell. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *Croton floribundus* Spreng. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## Planilha 2: Matriz_X (Numérica) (Tabela 3)

Esta aba deve conter apenas os dados numéricos, sem colunas de texto. Este formato é frequentemente necessário para análises estatísticas que exigem matrizes puramente numéricas, e suas linhas podem ser renomeadas posteriormente no R para E1, E2, ..., E17.

**Tabela 3**. Estruturação de dados na planilha 2 (Matrix X - apenas dados numéricos, sem coluna de espécies)

| P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 |
| 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 |
| 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 |
| 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 |
| 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 |
| 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 |
| 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 |
| 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 |
| 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 |
| 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 |
| 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

### 2) Fluxograma de Importação

Na Figura 7, é apresentado um fluxograma que resume os caminhos possíveis para importar dados para o R, dependendo do formato original do seu arquivo.

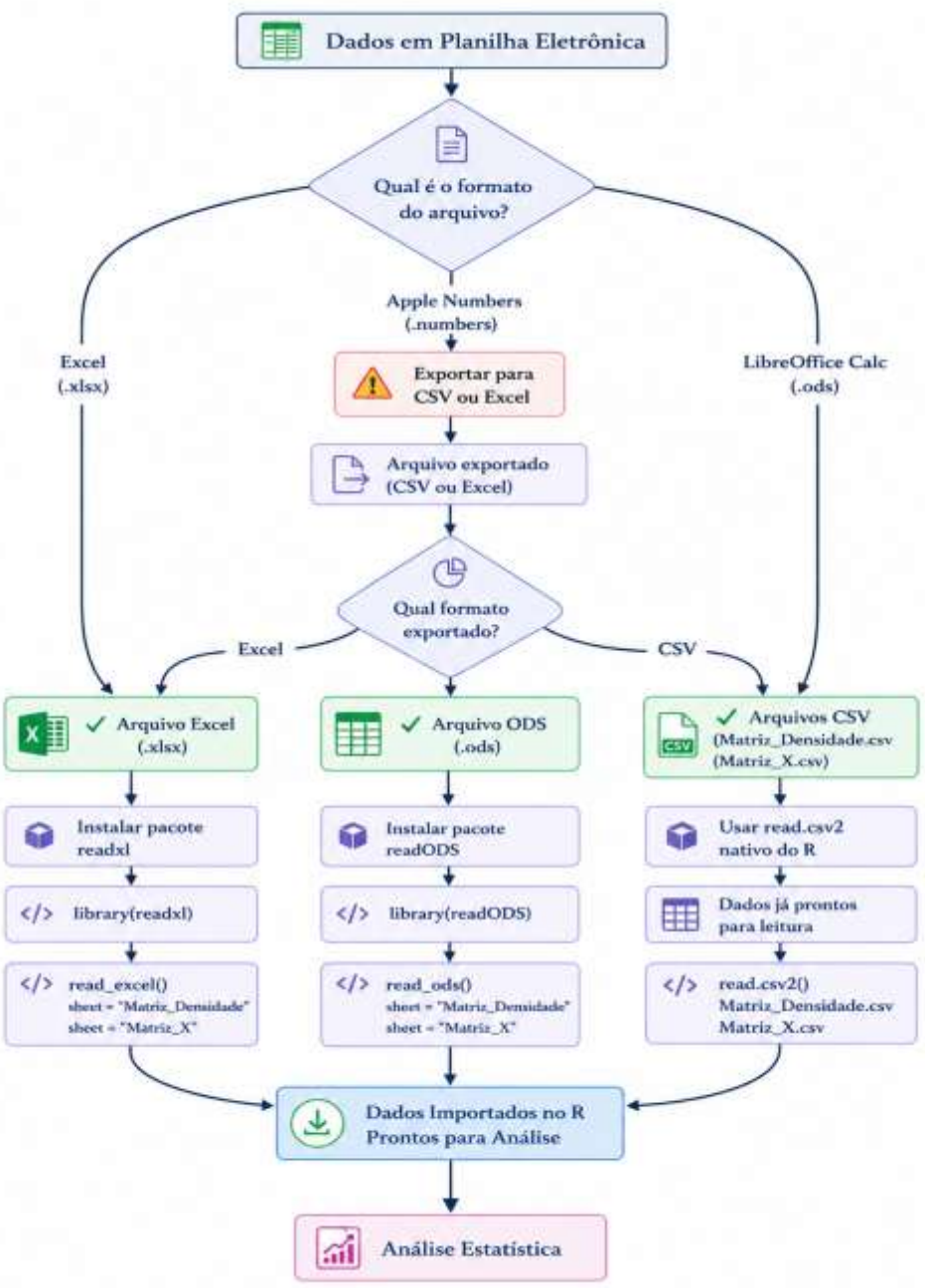



**Figura 7.** Fluxograma de importação de dados para o R, mostrando três caminhos possíveis dependendo do formato do arquivo.

### 3) Importação de Dados para o R

A importação de dados no R depende do formato do arquivo. O R pode ler arquivos .xlsx (Excel) e .ods (LibreOffice Calc) com o auxílio de pacotes específicos. O formato .numbers (Apple Numbers) não é lido diretamente, exigindo exportação prévia para um formato universal como CSV ou Excel.

Antes de importar, certifique-se de que o diretório de trabalho do R esteja configurado para a pasta onde seu arquivo de dados está localizado. Consulte a Seção 1.1.2 para revisar como definir o diretório de trabalho.

### 4) Lendo Arquivos do Excel (.xlsx):

Para ler arquivos do Excel, utilizamos o pacote *readxl*. Ele oferece uma forma simples e eficiente de importar dados, sem necessidade de conversão prévia.

```
# Instale o pacote (apenas uma vez por computador):
install.packages("readxl")

# Carregue o pacote (toda vez que iniciar uma nova sessão do R):
library(readxl)

# Leia as planilhas do arquivo Excel:
matriz_densidade <- read_excel("dados_densidade.xlsx", sheet = "Matriz_Densidade")
X <- read_excel("dados_densidade.xlsx", sheet = "Matriz_X")

# Converta para data.frame comum
matriz_densidade <- as.data.frame(matriz_densidade)
X <- as.data.frame(X)

# Opcional: Renomear as linhas da matriz numérica
rownames(X) <- paste0("E", 1:nrow(X))

#Verifique os dados importados (exibindo as primeiras linhas):
print(head(matriz_densidade))
print(head(X))
```

### 5) Lendo Arquivos do LibreOffice Calc (.ods)

Para ler arquivos .ods, o pacote *readODS* é a ferramenta ideal, desenvolvido especificamente para esse formato.

```
# Instale o pacote (apenas uma vez):
install.packages("readODS")

# Carregue o pacote (toda vez que iniciar uma nova sessão do R):
library(readODS)

#Leia as planilhas do arquivo ODS:
matriz_densidade <- read_ods("dados_densidade.ods", sheet = "Matriz_Densidade")
X <- read_ods("dados_densidade.ods", sheet = "Matriz_X")

# Converta para data.frame comum
matriz_densidade <- as.data.frame(matriz_densidade)
X <- as.data.frame(X)

# Opcional: Renomear as linhas da matriz numérica
rownames(X) <- paste0("E", 1:nrow(X))

#Verifique os dados importados (exibindo as primeiras linhas):
print(head(matriz_densidade))
print(head(X))
```

**6) Lendo Arquivos do Numbers (via exportação)**

Como o R não lê arquivos .numbers nativamente, o primeiro passo é exportar seus dados do Numbers para o formato CSV ou Excel.

**Passo 1: Exportar do Numbers**

1. No Numbers, vá em: Arquivo → Exportar para → CSV... ou Excel...
2. Se escolher CSV, cada aba da planilha será salva como um arquivo .csv separado (ex: Matriz_Densidade.csv e Matriz_X.csv).
3. Se escolher Excel, um único arquivo .xlsx com ambas as abas das planilhas será criado.

**Passo 2: Importar para o R**

Após exportar, siga um dos métodos abaixo:

- **Opção A: Lendo os arquivos CSV**

A função `read.csv2()` é geralmente usada para arquivos CSV que utilizam ponto e vírgula (`;`) como separador de colunas, comum em sistemas configurados para o português.

```
# Leia os arquivos CSV
matriz_densidade <- read.csv2("Matriz_Densidade.csv")
X <- read.csv2("Matriz_X.csv")

# Opcional: Renomear as linhas da matriz numérica
rownames(X) <- paste0("E", 1:nrow(X))

# Verifique os dados importados
print(head(matriz_densidade))
print(head(X))
```

- **Opção B: Lendo o arquivo Excel (.xlsx)**

Se você exportou para o formato .xlsx, siga o procedimento, já descrito na Seção 1.1.2, subitem 4, para a leitura de arquivos do Excel.

## 1.2. Tipos e Escalas de Variáveis

Vale ressaltar que a escolha da medida de distância e/ou da similaridade depende do tipo e da escala da variável, podendo impor a necessidade de sua padronização, pois diferentes medidas de distância, tipo ou mudanças nas escalas das variáveis podem conduzir a diferentes soluções (Pohlmann, 2007).

### a) Tipos de variáveis

As características associadas aos elementos em um problema multivariado podem ser representadas por tipos diferentes de variáveis. Assim, a utilização de uma função de

agrupamento permite decidir se dois elementos, $\mathbf{E_i}$ **e** $\mathbf{E_j}$, do conjunto **E** pertencem ou não a um mesmo agrupamento (Gama, 2022).

As características associadas aos elementos $\mathbf{E_i}$ e $\mathbf{E_j}$ podem ser representadas por três tipos diferentes de variáveis:

a) Variável Dicotômica – é aquela que se apresenta de duas formas, como presença (+) e ausência (-). Exemplos: frequência de uma espécie/parcela e sanidade da árvore.

b) Variável Qualitativa – é aquela resultante de uma classificação por tipos ou atributos. Pode ser nominal ou ordinal.

A variável qualitativa nominal se caracteriza por dados que consistem apenas em nomes, rótulos ou categorias. Os dados não podem ser dispostos segundo um esquema ordenado. São utilizados símbolos, ou números, ou nomes para representar determinado tipo de dados, mostrando, assim, a qual grupo ou categoria eles pertencem. Exemplos: O conjunto de espécies: Cedro, Cassia e Ipê, quando existe só o nome, tais como: parcela **A**; espécie **i**.

A variável qualitativa ordinal ou por postos se caracteriza quando uma classificação for dividida em categorias ordenadas em graus convencionados, havendo uma relação entre as categorias do tipo “maior do que”, “menor do que”, “igual a”. Os dados por postos consistem em valores relativos atribuídos para denotar a ordem de primeiro, segundo, terceiro e, assim, sucessivamente. Por exemplo: tipo de iluminação da copa; qualidade de fuste; grau de infestação de cipó; e processo de regeneração natural (semente ou brotação). Pode ser, também, classificada como binária ou multicategórica. Os caracteres binários, por conveniência, são codificados em **0** e **1**, permitindo representar a ausência ou presença de certa característica. Os caracteres multicategóricos não podem ser ordenados, sendo codificados por diferentes símbolos ou letras, como **I**, **II**, **III**, ..., para categoria (Ex: cor ou sabor da casca em estudos de Dendrologia). Quando os caracteres multicategóricos podem ser ordenados, tem-se a variável ordinal (posição sociológica: estrato – inferior, médio e superior).

c) Variável Quantitativa – é aquela cujos valores são expressos em números. Pode ser discreta ou contínua. As variáveis discretas são obtidas mediante alguma forma de contagem (Ex: número de árvores $ha^{-1}$), ao passo que os valores das variáveis contínuas resultam, em geral, de uma medição, sendo, frequentemente, dados em alguma unidade de medida (Ex: volume [$m^3$], diâmetro [cm]). As variáveis quantitativas contínuas podem ser originadas ou não de um delineamento experimental.

## b) Tipos de Escalas

Com base nas relações entre atributos e sua representação simbólica na forma de uma combinação numérica, Stevens (1946) definiu quatros tipos de escala: nominal, ordinal, de intervalo e da razão. As duas primeiras são não métricas ou qualitativas, pois reconhecem em cada levantamento uma determinada qualidade ou propriedade, enquanto as duas últimas são escalas métricas ou quantitativas, capazes de refletir diferenças de grau ou quantidade.

A escala nominal permite identificar categorias mutuamente excludentes e exaustivas. As variáveis medidas nessa escala são denominadas de categóricas, as quais são classificadas entre duas ou mais categorias. Exemplos incluem grupos ecofisiológicos de espécies (pioneira, secundária inicial, secundária tardia, clímax), estratos (inferior, médio, superior). Os números que se associam a cada categoria não têm nenhum significado concreto nem relação de ordem alguma.

Na Ciência Florestal, outros exemplos de tipos de medida de escala nominal são as variáveis infestação de cipó, as quais poderiam receber **1** (presença) ou **2** (ausência), e a qualidade de uso, podendo receber **1** (serraria), **2** (lenha), **3** (carvão) ou **4** (não identificado). Por apresentar unicamente duas categorias, a variável infestação de cipó é dicotômica, enquanto a qualidade de uso é politômica.

A escala ordinal também permite representar categorias, porém, nesse caso, os números associados a cada uma obedecem a uma ordem. Um exemplo é a variável idade com respostas possíveis: **1** se é inferior a 5 anos; **2** se está entre 5 e 10; **3** se está entre 11 e 35; e **4** se é maior que 35.

A escala de intervalo ou intervalar apresenta as mesmas características da ordinal, porém a diferença entre dois níveis consecutivos é constante ao longo de toda escala. No entanto, essa escala carece de zero absoluto, o que impede afirmar que um valor determinado dela seja o dobro, o triplo ou, em geral, um múltiplo de outro valor da própria escala. Um exemplo é a escala de temperatura, cuja diferença entre a temperatura de 30ºC e 35ºC é a mesma quantidade de calor da diferença entre 40ºC e 45ºC.

Vale salientar que nem sempre é fácil a distinção entre uma escala ordinal e uma de intervalo.

A escala da razão supera o inconveniente mencionado para a de intervalo, pois apresenta um zero absoluto conhecido, o que possibilita realizar, com essas medidas, todas as operações matemáticas. A altura, o diâmetro ou volume são desse tipo, pois uma árvore com 5 m de altura é duas vezes menor do que uma com 10 m.

A identificação de que uma variável é métrica ou não métrica se reveste de especial importância no âmbito da análise multivariada, pois, dependendo do caráter da variável, define-se a técnica multivariada a ser aplicada. Assim, por exemplo, a análise fatorial requer variáveis métricas, enquanto a análise de correspondência necessita das não métricas.

No entanto, uma variável não métrica pode ser convertida em métrica por meio de variáveis fictícias binárias. Para isso, é necessário contar com um número de variáveis fictícias igual ao número de categorias da variável não métrica menos um. Por exemplo: suponhamos que se pretende transformar em métrica a variável volume da parte aérea da árvore, dividida em três categorias: **1** = fuste, **2** = galho e **3** = árvore (fuste + galho). A conversão poderia se efetuar por meio de duas variáveis fictícias, $\mathbf{F_1}$ e $\mathbf{F_2}$ (Tabela 4).

**Tabela 4**. Variáveis fictícias para transformar a variável volume da parte aérea de uma árvore em métrica

| Categoria | $F_1$ | $F_2$ |
|---|---|---|
| **Fuste** | 1 | 0 |
| **Galho** | 0 | 1 |
| **Árvore** | 0 | 0 |

Esse recurso das variáveis fictícias possibilita a aplicação com variáveis não métricas de técnicas que originalmente requerem o emprego de variáveis métricas, como, por exemplo, no caso de análise de regressão linear múltipla, na qual seria possível ajustar uma única equação capaz de estimar com precisão o volume do fuste, do galho e da árvore, com significado biológico, ou seja, a soma dos volumes estimados separadamente para o fuste e o galho seria igual ao volume da árvore. Um exemplo de uso está em Soares, Leite e Campos (2001), ao proporem um modelo de crescimento e produção (Equação 1), em que a variável fictícia **TX** e o diâmetro comercial assumem diferentes valores, o que permite, via modelo, proporcionar estimativas de volumes para diferentes condições de plantio (Tabela 5).

$$V = \beta_0 . I^{\beta_1} . B^{\beta_2} . S^{\beta_3} . \left(exp^{\beta_4 . TX/q}\right) . \left(1 - (d/q)^{1+\beta_5 . d}\right) + \varepsilon \qquad (1)$$

Em que: **V** = volume total ou comercial do fuste, em $m^3$; $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, ..., $\boldsymbol{\beta_5}$ = parâmetros do modelo; **I** = idade, em meses; **B** = área basal, em $m^2\ ha^{-1}$; **S** = índice de local, em metros; **TX** = variável binária, assumindo valor 0 e 1; **d** = diâmetro de uso da madeira, em cm; **q** = diâmetro do povoamento, em cm; **ε** = erro aleatório.

**Tabela 5**. Estimativas de volume para diferentes condições de plantio conforme a variável fictícia binária TX e o diâmetro comercial (d)

| **TX** | **d** (cm) | **Volume** ($m^3\ ha^{-1}$) |
|---|---|---|
| 0 | 0 | Volume total com casca |
| 0 | diferente de zero | Volume comercial com casca |
| 1 | 0 | Volume total sem casca |
| 1 | diferente de zero | Volume comercial sem casca |

Fonte: Soares, Leite e Campos (2001)

Finalmente, vale ressaltar que a classificação de Stevens (1946) é bastante difundida e utilizada em diversas áreas por sua simplicidade e por estar presente em *softwares* estatísticos. No entanto, há críticas na literatura, especialmente quanto ao fato de a escolha de determinado teste estatístico a ser aplicado para um vetor de dados não depender da escala de mensuração.

## 1.3. Medidas de Similaridade

Muitas medidas de similaridade ou semelhança têm sido propostas e utilizadas em análise multivariada (Matteuci; Colma, 1982; Johnson; Wichern, 2007), levando à pergunta: Qual medida escolher? Contudo, não é simples respondê-la. Por isso, na maioria das vezes, quando o tipo de dados é tal que algumas possíveis medidas de similaridade podem ser utilizadas, a escolha entre elas é quase sempre baseada na preferência e experiência do pesquisador (Sneath; Sokal, 1973).

Quando a magnitude da medida reflete a magnitude da similaridade, ela é denominada coeficiente ou índice de similaridade e, quando reflete a magnitude da dissimilaridade, é denominada medida de distância.

### 1.3.1. Coeficientes ou Índices de Similaridade

Qualquer uma das variáveis do tipo qualitativo, como as dicotômicas com várias alternativas, permite uma medida de similaridade entre dois elementos, $\mathbf{E_i}$ e $\mathbf{E_j}$, do conjunto **E** **p**-dimensional ($\mathbf{I_p}$). Os correspondentes vetores de medidas servem de base para o cálculo da medida de similaridade, desde que todos os seus componentes sejam variáveis qualitativas. Desse modo, pode-se definir como função de agrupamento uma função real não negativa, genericamente representada por: $S_{ij} = S(E_i, E_j)$, a qual, sendo $\mathbf{E_i}$ e $\mathbf{E_j}$ dois quaisquer elementos do conjunto **E**, define uma medida de similaridade se as seguintes propriedades forem satisfeitas:

1. $S(E_i,E_i) = 1$;
2. $S(E_i,E_j) = S(E_j,Ei)$; e
3. $0 \leq S\ (Ei,Ej) \leq 1$, para $i \neq j$.

Segundo Gama (2022), essa medida define um coeficiente de similaridade ou de associação entre os elementos $\mathbf{E_i}$ e $\mathbf{E_j}$ do conjunto **E**. Esse coeficiente mede a relação entre dois elementos de um conjunto **E**, com base nos dados dos valores observados das **p** variáveis do tipo qualitativo.

As funções de semelhança podem ser calculadas a partir de variáveis binárias ou qualitativas (presença/ausência) ou variáveis quantitativas (Orlóci, 1978; Matteucci; Colma,

1982). Quando se trata de comparações entre parcelas ou entre espécies, são calculadas por meio de Tabela de contingência 2 x 2 (Tabela 6).

**Tabela 6**. Frequência agrupada das espécies (parcelas) A (1) e B (2)

| **Espécie (Parcela)** | | **B (2)** | | **Total da Linha** |
|---|---|---|---|---|
| | | **Presença** | **Ausência** | |
| A (1) | **Presença** | a | b | a + b |
| | **Ausência** | c | d | c + d |
| **Total da Coluna** | | a + c | b + d | p = a + b + c + d |

Em que: **a** = número de parcelas em que as espécies A e B estão presentes ou número de espécies em que as parcelas 1 e 2 estão presentes; **b** = número de parcelas que contém só a espécie B ou número de espécies exclusivas da parcela 2; **c** = número de parcelas em que a espécie A aparece só ou o número de espécies exclusivas da parcela 1; **d** = número de parcelas em que as espécies A e B estão ausentes ou número de espécies ausentes de ambas as parcelas, simultaneamente; e **p** = número de amostras ou de espécies, respectivamente.

Para exemplificar a aplicação dos coeficientes de similaridade (Tabela 7), o 0 e 1 indicam, respectivamente, ausência e presença das espécies (1, 2, 3, ..., 17) nas parcelas (1, 2, 3, ..., 11). Para a obtenção das frequências agrupadas, são necessárias Tabelas de contingência 2 x 2. Assim, no exemplo prático, considerando que se quer agrupar parcelas (vetor coluna), seria necessária a confecção de 55 Tabelas (combinação de 11 espécies, duas a duas), semelhantes às Tabelas 8, 9 e 10.

**Tabela 7**. Presença (1) e ausência (0) de 17 espécies da mata da silvicultura, em parcelas de 20 x 50 m, município de Viçosa-MG

| Espécies | Parcelas | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| *Casearia decandra* Jacq. | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *Anadenanthera peregrina* Speg. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| *Apuleia leiocarpa* (Vog.) Macbr. | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *Mabea fistulifera* Mart. | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| *Anadenanthera macrocarpa* (Benth.) Brenan. | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| *Platypodium elegans* Vog. | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| *Machaerium floridum* (Benth.) Ducke | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| *Copaifera lansdorffii* Desf. | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| *Ocotea pretiosa* Mez. | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| *Cabralea cangerana* Saldanha | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| *Piptadenia gonoacantha* Macbr. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| *Dalbergia nigra* Allem. ex Benth. | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| *Luehea divaricata* Mart. | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| *Cecropia hololeuca* Miq. | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| *Melanoxylon brauna* Schott. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| *Cedrela fissilis* Vell. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *Croton floribundus* Spreng. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Os dados da Tabela 7 podem ser representados pela matriz X:

$$\mathbf{X} = \begin{vmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Em que: em cada linha e coluna da matriz **X**, descrevem-se, respectivamente, um vetor espécie ($\mathbf{X_p}$) e um vetor parcela ($\mathbf{X_n}$).

Como exemplos, são apresentadas as frequências agrupadas das parcelas 1 e 2 (Tabela 8), das parcelas 1 e 11 (Tabela 9) e das parcelas 10 e 11 (Tabela 10).

**Tabela 8**. Frequência agrupada das parcelas 1 e 2

| **Parcela** | | **2** | | **Total da Linha** |
|---|---|---|---|---|
| | | **Presença** | **Ausência** | |
| **1** | **Presença** | 4 (**a**) | 4 (**b**) | 8 (**a + b**) |
| | **Ausência** | 1 (**c**) | 8 (**d**) | 9 (**c + d**) |
| **Total da Coluna** | | 5 (**a + c**) | 12 (**b + d**) | 17 (**p = a + b + c + d**) |

**Tabela 9**. Frequência agrupada das parcelas 1 e 11

| **Parcela** | | **11** | | **Total da Linha** |
|---|---|---|---|---|
| | | **Presença** | **Ausência** | |
| **1** | **Presença** | 6 | 2 | 8 |
| | **Ausência** | 6 | 3 | 9 |
| **Total da Coluna** | | 12 | 5 | 17 |

**Tabela 10**. Frequência agrupada das parcelas 10 e 11

| **Parcela** | | **11** | | **Total da Linha** |
|---|---|---|---|---|
| | | **Presença** | **Ausência** | |
| **10** | **Presença** | 11 | 1 | 12 |
| | **Ausência** | 1 | 4 | 5 |
| **Total da Coluna** | | 12 | 5 | 17 |

## a) Seção Prática – Implementação da Matriz Binária

Após entender como o R funciona e como organizar os seus dados, o próximo passo é prepará-los para a análise. Nessa seção, veremos como lidar com dois tipos de dados: os que mostram quantidades (como a densidade de espécies) e os que indicam apenas presença ou ausência.

### 1) Preparação da matriz quantitativa (X)

Para a maioria das análises no R, é importante que as parcelas (os locais que você quer comparar) fiquem nas linhas e as espécies (as características que você está medindo) fiquem nas colunas. Lembre-se da matriz X que você criou anteriormente, com os números de densidade das espécies. Se a sua matriz X estiver com as espécies nas linhas e as parcelas nas colunas, você precisa "virá-la". No R, fazemos isso com a função *t*() (de transpor) e, depois, transformamos o resultado em uma matriz novamente, com *as.matrix*().

```
# Transpondo a matriz X para que as parcelas fiquem nas linhas
X_quantitativa <- as.matrix(t(X))

# Visualizando a matriz quantitativa transposta
print(X_quantitativa)
```

| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 | E17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 8 | 0 | 3 | 6 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 5 | 0 | 7 | 0 | 0 | 0 |
| P2 | 1 | 0 | 9 | 3 | 12 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P3 | 27 | 0 | 4 | 3 | 0 | 1 | 10 | 0 | 2 | 0 | 0 | 7 | 1 | 0 | 0 | 0 | 1 |
| P4 | 0 | 0 | 6 | 4 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P5 | 1 | 0 | 22 | 29 | 0 | 9 | 9 | 1 | 2 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| P6 | 9 | 0 | 9 | 12 | 0 | 1 | 2 | 13 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| P7 | 2 | 12 | 5 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 6 | 0 | 2 | 0 | 0 | 1 | 0 |
| P8 | 3 | 1 | 2 | 4 | 0 | 0 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| P9 | 22 | 17 | 7 | 4 | 2 | 5 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| P10 | 15 | 1 | 4 | 4 | 0 | 11 | 11 | 3 | 2 | 3 | 0 | 1 | 5 | 0 | 2 | 0 | 0 |
| P11 | 7 | 9 | 4 | 4 | 0 | 1 | 5 | 1 | 2 | 1 | 5 | 0 | 2 | 0 | 1 | 0 | 0 |

### 2) Criação e Ajuste da Matriz Binária

Além dos dados de densidade, podemos ter dados que mostram apenas se uma espécie está presente (1) ou ausente (0) em cada parcela. Isso é o que chamamos de dados binários. Para criar essa matriz no R, primeiro, definimos os dados e, depois, os ajustamos para a organização correta.

```
# Criando a tabela com os dados de presença (1) ou ausência (0)
# (Espécies nas linhas, Parcelas nas colunas)
dados_binarios <- data.frame(
  P1 = c(1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0),
  P2 = c(1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0),
  P3 = c(1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1),
  P4 = c(0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0),
  P5 = c(1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0),
  P6 = c(1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0),
  P7 = c(1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0),
  P8 = c(1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0),
  P9 = c(1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0),
  P10 = c(1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0),
  P11 = c(1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0)
)

# Visualizando a matriz binária bruta
print(dados_binarios)
```

| P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Para que a matriz binária (X_binaria) esteja na mesma organização que a X_quantitativa (parcelas nas linhas e espécies nas colunas), precisamos transpor dados_binarios e convertê-la para o formato de matriz. Além disso, renomearemos as colunas para **E1**, **E2**, ..., **E17**, para padronizar com a identificação das espécies.

```
# Transpondo a matriz binária para que as parcelas fiquem nas linhas
X_binaria <- as.matrix(t(dados_binarios))

# Renomeando as colunas da matriz binária para corresponder às espécies
colnames(X_binaria) <- paste0("E", 1:ncol(X_binaria))
```

```
# Visualizando a matriz binária ajustada
print(X_binaria)
```

| Parcela | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 | E17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| P2 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P3 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| P4 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P5 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| P6 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| P7 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| P8 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| P9 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| P10 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| P11 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |

## b) Coeficientes de similaridade para variáveis qualitativas

Entre os índices e coeficientes que utilizam dados qualitativos e que podem ser considerados como função de agrupamento, podem-se citar:

### 1) Quadrado médio de contingência (Índice de Orlóci)

O quadrado médio de contingência ($\mathbf{X^2_{hi}/n}$) de duas espécies, **h** e **i**, é definido, segundo Orlóci (1978), por meio da seguinte expressão:

$$X^2_{hi}/n = \frac{(ad - bc)^2}{R_1 R_2 C_1 C_2}$$

Em que: $\boldsymbol{R_1 = a + b}$; $\boldsymbol{R_2 = c + d}$; $\boldsymbol{C_1 = a + c}$; $\boldsymbol{C_2 = b + d}$; e $\boldsymbol{n = a + b + c + d}$

**Aplicação**

Com base na Tabela 8, temos **a** = 4; **b** = 4; **c** = 1; e **d** = 8, logo:

$\boldsymbol{R_1} = 4 + 4 = 8;\ \boldsymbol{R_2} = 1 + 8 = 9;\ \boldsymbol{C_1} = 4 + 1 = 5;\ \boldsymbol{C_2} = 4 + 8 = 12$ e $\boldsymbol{n} = 4 + 4 + 1 + 8 = 17.$

Assim, o quadrado médio das parcelas 1 e 2 será:

$$\frac{X^2_{1,2}}{n} = \frac{(4x8 - 4x1)^2}{8x9x5x12} = \frac{28^2}{4320} = \frac{784}{4320} = 0{,}18$$

Já na Tabela 10, temos **a** = 11; **b** = 1; **c** = 1; e **d** = 4, logo:

$\boldsymbol{R_1} = 11 + 1 = 12;\ \boldsymbol{R_2} = 1 + 4 = 5;\ \boldsymbol{C_1} = 11 + 1 = 12;\ \boldsymbol{C_2} = 1 + 4 = 5;$ e $\boldsymbol{n} = 11 + 1 + 1 + 4 = 17.$

Assim, o quadrado médio das parcelas 10 e 11 será:

$$\frac{X^2_{10,11}}{n} = \frac{(11x4 - 1x1)^2}{12x5x12x5} = \frac{43^2}{3600} = \frac{1849}{3600} = 0{,}51$$

De posse dos dados desse índice para todas as combinações de parcelas 2 a 2, monta-se uma matriz $X_{hi}^2/n$ de dimensões 11 x 11.

### 1.1. Passo a passo no R: Quadrado médio de contingência (Índice de Orlóci)

Com a matriz binária (X_binaria) pronta, agora podemos calcular o Índice de Orlóci, uma medida de similaridade entre as parcelas. Esse índice nos ajuda a entender o quão parecidas são as composições de espécies entre diferentes áreas. Para isso, vamos criar uma função no R que realiza o cálculo e, depois, aplicá-la a todos os pares de parcelas.

**a. Desenvolvendo a função do índice**

Para facilitar o cálculo, vamos criar uma "receita" no R, ou seja, uma função personalizada. Essa função receberá os dados de presença/ausência de duas parcelas e retornará o valor do Índice de Orlóci entre elas. A função *orloci* é definida como:

```
orloci <- function(x, y) {
  # Cria uma tabela de contingência entre as duas parcelas (x e y)
  tab <- table(x, y)
  # Calcula o número total de observações
  n <- sum(tab)
  # Calcula os valores esperados para a tabela de contingência
  esperado <- outer(rowSums(tab), colSums(tab)) / n
  # Calcula o Índice de Orlóci, tratando casos onde o esperado é zero
  sum(ifelse(esperado > 0, (tab - esperado)^2 / esperado, 0)) / n
}
```

**b. Calculando o índice**

Para analisar a similaridade entre todas as parcelas, precisamos calcular o Índice de Orlóci para cada par possível. O resultado será uma matriz de similaridade, em que cada valor indica o grau de semelhança entre duas parcelas. Utilizaremos a função *outer*() do R para aplicar nossa função *orloci* a todos os pares de linhas da X_binaria.

```
# Aplicando a função orloci para todos os pares de parcelas na X_binaria
matriz_orloci <- outer(
  1:nrow(X_binaria), # Itera sobre as linhas (parcelas) da X_binaria
  1:nrow(X_binaria), # Itera novamente sobre as linhas (parcelas) da X_binaria
  Vectorize(function(i, j) {
    # Para cada par (i, j), calcula o índice de Orlóci usando os dados das parcelas i e j
    orloci(X_binaria[i, ], X_binaria[j, ])
  })
)

# Nomeando as linhas e colunas da matriz de Orlóci com os nomes das parcelas
rownames(matriz_orloci) <- rownames(X_binaria)
colnames(matriz_orloci) <- rownames(X_binaria)
```

```
# Visualizando o resultado final da matriz de Orlóci, arredondado para duas casas decimais
print(round(matriz_orloci, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.18 | 0.03 | 0.09 | 0.28 | 0.28 | 0.09 | 0.42 | 0.00 | 0.12 | 0.01 |
| P2 | 0.18 | 1.00 | 0.01 | 0.18 | 0.18 | 0.18 | 0.01 | 0.06 | 0.18 | 0.02 | 0.02 |
| P3 | 0.03 | 0.01 | 1.00 | 0.03 | 0.43 | 0.17 | 0.03 | 0.00 | 0.00 | 0.18 | 0.03 |
| P4 | 0.09 | 0.18 | 0.03 | 1.00 | 0.28 | 0.28 | 0.00 | 0.03 | 0.09 | 0.12 | 0.12 |
| P5 | 0.28 | 0.18 | 0.43 | 0.28 | 1.00 | 0.58 | 0.09 | 0.03 | 0.00 | 0.37 | 0.12 |
| P6 | 0.28 | 0.18 | 0.17 | 0.28 | 0.58 | 1.00 | 0.09 | 0.17 | 0.00 | 0.12 | 0.12 |
| P7 | 0.09 | 0.01 | 0.03 | 0.00 | 0.09 | 0.09 | 1.00 | 0.00 | 0.17 | 0.01 | 0.03 |
| P8 | 0.42 | 0.06 | 0.00 | 0.03 | 0.03 | 0.17 | 0.00 | 1.00 | 0.17 | 0.08 | 0.08 |
| P9 | 0.00 | 0.18 | 0.00 | 0.09 | 0.00 | 0.00 | 0.17 | 0.17 | 1.00 | 0.01 | 0.12 |
| P10 | 0.12 | 0.02 | 0.18 | 0.12 | 0.37 | 0.12 | 0.01 | 0.08 | 0.01 | 1.00 | 0.51 |
| P11 | 0.01 | 0.02 | 0.03 | 0.12 | 0.12 | 0.12 | 0.03 | 0.08 | 0.12 | 0.51 | 1.00 |

## 2. Coeficiente de correlação puntual

O Coeficiente de correlação puntual (**CCP**) é calculado conforme a seguinte equação:

$$CCP_{A,B} = \frac{ad - bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$$

Esse índice leva em consideração as ausências conjuntas. Se **b** = 0 e **c** = 0, as espécies A e B estão associadas positivamente, e se **a** = 0 ou **d** = 0, as espécies A e B mostram associações negativas.

**Aplicação**

Para as parcelas 1 e 2, temos **a** = 4; **b** = 4; **c** = 1; e **d** = 8 (Tabela 8, página 26), logo:

$$\mathrm{CCP}_{1,2} = \frac{4x8-4x1}{[(4+4)(4+1)(4+8)(1+8)]^{1/2}} = \frac{28}{(4320)^{1/2}} = 0{,}43$$

Já para as parcelas 10 e 11 (Tabela 10, página 26), temos **a** = 11; **b** = 1; **c** = 1; e **d** = 4, assim:

$$\mathrm{CCP}_{10,11} = \frac{11x4-1x1}{[(11+1)(11+1)(1+4)(1+4)]^{1/2}} = \frac{43}{(3600)^{1/2}} = 0{,}72$$

### 2.1. Passo a passo no R: Coeficiente de correlação puntual

#### a) Calculando o índice

Para calcular a correlação entre todas as parcelas, utilizaremos a função *cor*() do R. Como nossa matriz X_binaria tem as parcelas nas linhas e as espécies nas colunas, precisamos transpor essa matriz (*t*(X_binaria)) antes de aplicar a função *cor*(). Isso fará com que as parcelas se tornem as colunas, permitindo que *cor*() calcule a associação entre elas.

```
# Calculando a matriz de correlação entre as parcelas
# Transpomos X_binaria para que as parcelas (linhas) se tornem colunas para a função cor()
matriz_corr_puntual <- cor(t(X_binaria))
```

```
# Tratamento de casos sem variação: Substitui valores 'NA' (Not Available) por 0.
# Isso ocorre quando uma parcela não tem variação (ex: todas as espécies ausentes ou presentes).
matriz_corr_puntual[is.na(matriz_corr_puntual)] <- 0

# Visualizando o resultado final da matriz de correlação, arredondado para duas casas decimais
print(round(matriz_corr_puntual, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.43 | 0.18 | 0.29 | 0.53 | 0.53 | -0.29 | 0.65 | 0.06 | 0.35 | 0.09 |
| P2 | 0.43 | 1.00 | 0.09 | 0.43 | 0.43 | 0.43 | 0.09 | 0.25 | 0.43 | 0.13 | 0.13 |
| P3 | 0.18 | 0.09 | 1.00 | 0.18 | 0.65 | 0.42 | -0.18 | 0.07 | -0.06 | 0.43 | 0.17 |
| P4 | 0.29 | 0.43 | 0.18 | 1.00 | 0.53 | 0.53 | -0.06 | 0.17 | 0.29 | 0.35 | 0.35 |
| P5 | 0.53 | 0.43 | 0.65 | 0.53 | 1.00 | 0.76 | -0.29 | 0.17 | 0.06 | 0.61 | 0.35 |
| P6 | 0.53 | 0.43 | 0.42 | 0.53 | 0.76 | 1.00 | -0.29 | 0.41 | 0.06 | 0.35 | 0.35 |
| P7 | -0.29 | 0.09 | -0.18 | -0.06 | -0.29 | -0.29 | 1.00 | 0.07 | 0.42 | -0.09 | 0.17 |
| P8 | 0.65 | 0.25 | 0.07 | 0.17 | 0.17 | 0.41 | 0.07 | 1.00 | 0.41 | 0.28 | 0.28 |
| P9 | 0.06 | 0.43 | -0.06 | 0.29 | 0.06 | 0.06 | 0.42 | 0.41 | 1.00 | 0.09 | 0.35 |
| P10 | 0.35 | 0.13 | 0.43 | 0.35 | 0.61 | 0.35 | -0.09 | 0.28 | 0.09 | 1.00 | 0.72 |
| P11 | 0.09 | 0.13 | 0.17 | 0.35 | 0.35 | 0.35 | 0.17 | 0.28 | 0.35 | 0.72 | 1.00 |

## 3. Coeficiente de comunidade de Jaccard

O coeficiente de comunidade de Jaccard (**CCJ**) relaciona o número de espécies comuns e o número de espécies encontradas nas duas amostras que se comparam, cuja fórmula é:

$$CCJ_{A,B} = \frac{a}{a+b+c}$$

Assim, o CCJ para as parcelas 1 e 2, **a** = 4, **b** = 4 e **c** =1 (Tabela 8, página 26), e 10 e 11, **a** = 11, **b** = 1 e **c** =1 (Tabela 10, página 26), respectivamente, seria:

$CCJ_{1,2} = \frac{4}{4+4+1} = 0{,}44$ e $CCJ_{10,11} = \frac{11}{11+1+1} = 0{,}85$

Pela análise desse coeficiente, pode-se verificar que, se a associação é total, $CCJ_{A,B} = 1{,}0$, e, se não, $CCJ_{A,B} = 0$.

Esse índice também é conhecido como coeficiente de similaridade de Agrell e Iverson.

### 3.1 Passo a passo no R: Coeficiente de Jaccard

#### a. Calculando a matriz de dissimilaridade

Para calcular a dissimilaridade (o oposto da similaridade) de Jaccard, utilizaremos a função *vegdist*() do pacote *vegan*. Por padrão, quando pedimos o método "*jaccard*", ele nos dá a dissimilaridade. Isso significa que, quanto maior o valor, menos parecidas são as parcelas. Usaremos a matriz X_binaria (aquela com 1s e 0s de presença/ausência) para esse cálculo.

```
# Carregamos o pacote 'vegan' para poder usar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(vegan)

# Calculamos a matriz de "diferença" (dissimilaridade) de Jaccard
```

```
# 'binary = TRUE' indica que estamos trabalhando com dados binários
dist_jaccard <- vegdist(X_binaria, method = "jaccard", binary = TRUE)

# Visualizamos a matriz de dissimilaridade, arredondada para duas casas decimais
print(round(as.matrix(dist_jaccard), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.56 | 0.58 | 0.55 | 0.40 | 0.40 | 0.79 | 0.33 | 0.67 | 0.46 | 0.57 |
| P2 | 0.56 | 0.00 | 0.73 | 0.56 | 0.56 | 0.56 | 0.73 | 0.67 | 0.56 | 0.69 | 1.00 |
| P3 | 0.58 | 0.73 | 0.00 | 0.58 | 0.30 | 0.45 | 0.71 | 0.67 | 0.69 | 0.38 | 0.50 |
| P4 | 0.55 | 0.56 | 0.58 | 0.00 | 0.40 | 0.40 | 0.69 | 0.64 | 0.55 | 0.46 | 0.46 |
| P5 | 0.40 | 0.56 | 0.30 | 0.40 | 0.00 | 0.22 | 0.79 | 0.64 | 0.67 | 0.33 | 0.46 |
| P6 | 0.40 | 0.56 | 0.45 | 0.40 | 0.22 | 0.00 | 0.79 | 0.50 | 0.67 | 0.46 | 0.46 |
| P7 | 0.79 | 0.73 | 0.71 | 0.69 | 0.79 | 0.79 | 0.00 | 0.67 | 0.45 | 0.60 | 0.50 |
| P8 | 0.33 | 0.67 | 0.67 | 0.64 | 0.64 | 0.50 | 0.67 | 0.00 | 0.50 | 0.54 | 0.54 |
| P9 | 0.67 | 0.56 | 0.69 | 0.55 | 0.67 | 0.67 | 0.45 | 0.50 | 0.00 | 0.57 | 0.46 |
| P10 | 0.46 | 0.69 | 0.38 | 0.46 | 0.33 | 0.46 | 0.60 | 0.54 | 0.57 | 0.00 | 0.15 |
| P11 | 0.57 | 0.69 | 0.50 | 0.46 | 0.46 | 0.46 | 0.50 | 0.54 | 0.46 | 0.15 | 0.00 |

### b. Calculando a matriz de similaridade

Como a função *vegdist*() nos fornece a dissimilaridade, para obter a similaridade (um número de 0 a 1 que mostra o quão parecidas são as parcelas), basta fazer uma conta simples: **Similaridade = 1 - Dissimilaridade**. Isso transformará os valores de diferença em valores de semelhança.

```
# Transformamos a "diferença" (dissimilaridade) em "similaridade"
matriz_similaridade_jaccard <- 1 - dist_jaccard

# Visualizamos a matriz de similaridade de Jaccard, arredondada para duas casas decimais
print(round(as.matrix(matriz_similaridade_jaccard), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.44 | 0.42 | 0.45 | 0.60 | 0.60 | 0.21 | 0.67 | 0.33 | 0.54 | 0.43 |
| P2 | 0.44 | 0.00 | 0.27 | 0.44 | 0.44 | 0.44 | 0.27 | 0.33 | 0.44 | 0.31 | 0.00 |
| P3 | 0.42 | 0.27 | 0.00 | 0.42 | 0.70 | 0.55 | 0.29 | 0.33 | 0.31 | 0.62 | 0.50 |
| P4 | 0.45 | 0.44 | 0.42 | 0.00 | 0.60 | 0.60 | 0.31 | 0.36 | 0.45 | 0.54 | 0.54 |
| P5 | 0.60 | 0.44 | 0.70 | 0.60 | 0.00 | 0.78 | 0.21 | 0.36 | 0.33 | 0.67 | 0.54 |
| P6 | 0.60 | 0.44 | 0.55 | 0.60 | 0.78 | 0.00 | 0.21 | 0.50 | 0.33 | 0.54 | 0.54 |
| P7 | 0.21 | 0.27 | 0.29 | 0.31 | 0.21 | 0.21 | 0.00 | 0.33 | 0.55 | 0.40 | 0.50 |
| P8 | 0.67 | 0.33 | 0.33 | 0.36 | 0.36 | 0.50 | 0.33 | 0.00 | 0.50 | 0.46 | 0.46 |
| P9 | 0.33 | 0.44 | 0.31 | 0.45 | 0.33 | 0.33 | 0.55 | 0.50 | 0.00 | 0.43 | 0.54 |
| P10 | 0.54 | 0.31 | 0.62 | 0.54 | 0.67 | 0.54 | 0.40 | 0.46 | 0.43 | 0.00 | 0.85 |
| P11 | 0.43 | 0.31 | 0.50 | 0.54 | 0.54 | 0.54 | 0.50 | 0.46 | 0.54 | 0.85 | 0.00 |

## 4. Coeficiente de comunidade de Sorensen

O coeficiente de comunidade de Sorensen (**CCS**) tem a mesma interpretação que o Jaccard, e nenhum deles leva em conta as ausências conjuntas que podem ser indicativas de associação. O CCS pode ser obtido conforme a seguinte expressão:

$$CCS_{A,B} = \frac{2a}{2a + b + c}$$

Esse coeficiente também é conhecido como índice de similaridade de Dice ou Dice-Sorensen.

Os coeficientes de Jaccard e Sorensen apresentam valor "1" se todas as espécies são comuns ou as amostras são idênticas, e valor "0" se não há espécies comuns ou as amostras são completamente distintas. Ambos omitem as concordâncias negativas.

**Aplicação**

Na Tabela 8 (página 26), temos **a** = 4, **b** = 4 e **c** =1 para as parcelas 1 e 2, e, na Tabela 10 (página 26), **a** = 11, **b** = 1 e **c** =1 para as parcelas 10 e 11, assim:

$CCS_{1,2} = \frac{2x4}{2x4+4+1} = 0{,}62$ e $CCS_{10,11} = \frac{2x11}{2x11+1+1} = 0{,}92$

## 4.1 Passo a passo no R: Coeficiente de Sorensen

### a) Calculando a matriz de similaridade

Para obter a similaridade de Sorensen no R, utilizamos a função *vegdist*() do pacote *vegan* com o método de Bray-Curtis (*method = "bray"*) aplicado a dados binários (*binary = TRUE*). Como essa função retorna à dissimilaridade, subtraímos o resultado de 1 para obter a matriz de similaridade.

```
# Carregamos o pacote 'vegan'
library(vegan)

# Calculamos a similaridade de Sorensen (1 - Dissimilaridade)
# O método "bray" com 'binary = TRUE' equivale ao índice de Sorensen
matriz_similaridade_sorensen <- 1 - vegdist(X_binaria, method = "bray", binary = TRUE)

# Visualizamos a matriz de similaridade, arredondada para duas casas decimais
print(round(as.matrix(matriz_similaridade_sorensen), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.62 | 0.59 | 0.63 | 0.75 | 0.75 | 0.35 | 0.80 | 0.50 | 0.70 | 0.60 |
| P2 | 0.62 | 0.00 | 0.43 | 0.62 | 0.62 | 0.62 | 0.43 | 0.50 | 0.62 | 0.47 | 0.47 |
| P3 | 0.59 | 0.43 | 0.00 | 0.59 | 0.82 | 0.71 | 0.44 | 0.50 | 0.47 | 0.76 | 0.67 |
| P4 | 0.63 | 0.62 | 0.59 | 0.00 | 0.75 | 0.75 | 0.47 | 0.53 | 0.63 | 0.70 | 0.70 |
| P5 | 0.75 | 0.62 | 0.82 | 0.75 | 0.00 | 0.88 | 0.35 | 0.53 | 0.50 | 0.80 | 0.70 |
| P6 | 0.75 | 0.62 | 0.71 | 0.75 | 0.88 | 0.00 | 0.35 | 0.67 | 0.50 | 0.70 | 0.70 |
| P7 | 0.35 | 0.43 | 0.44 | 0.47 | 0.35 | 0.35 | 0.00 | 0.50 | 0.71 | 0.57 | 0.67 |
| P8 | 0.80 | 0.50 | 0.50 | 0.53 | 0.53 | 0.67 | 0.50 | 0.00 | 0.67 | 0.63 | 0.63 |
| P9 | 0.50 | 0.62 | 0.47 | 0.63 | 0.50 | 0.50 | 0.71 | 0.67 | 0.00 | 0.60 | 0.70 |
| P10 | 0.70 | 0.47 | 0.76 | 0.70 | 0.80 | 0.70 | 0.57 | 0.63 | 0.60 | 0.00 | 0.92 |
| P11 | 0.60 | 0.47 | 0.67 | 0.70 | 0.70 | 0.70 | 0.67 | 0.63 | 0.70 | 0.92 | 0.00 |

## 5. Índice de informação

O índice de informação ($\mathbf{I_{(j+k)}}$) é empregado em alguns sistemas de classificação. Ele mede o ganho de informação obtido quando se unem duas amostras, e se baseia na teoria da

informação. Por definição, a informação contida em uma única parcela é zero. Quando se unem duas parcelas idênticas, o ganho de informação é zero. Caso contrário, o subconjunto formado contém informação dada pelas espécies não comuns. Quando se unem dois subconjuntos de amostras (**j** e **k**), a informação contida no subconjunto (**l** = **j** + **k**) é maior que a simples soma de informação contida em cada um dos subconjuntos, isto é, $\mathbf{I_l = I_j + I_k + I}$.

Para se proceder à fusão de duas amostras ou parcelas, a fórmula é:

$$I_{(1+2)} = 2(b + c).ln2 = I$$

Para a fusão de dois subconjuntos **j** e **k**, tem-se:

$$I_{(j+k)} = MNln(M) - \sum_{i=1}^{N} a_i \ln(a_i) + (M - a_i)x\ln(M - a_i)$$

Em que: $\boldsymbol{a_i}$= número de parcelas em que aparece o atributo i; **N** = número total de atributos (espécies); e **M** = número total de amostras do subconjunto (**j** + **k**).

A segunda equação se reduz à primeira quando se agrupam duas amostras.

Esse índice provém do conceito de entropia e pode ser considerado como uma medida de desordem ou de heterogeneidade dentro do conjunto de amostras.

**Aplicação**

Temos que **b** = 4 e **c** = 1 (Tabela 8 página 26) para as parcelas 1 e 2, e **b** = 1 e **c** = 1 (Tabela 10, página 26) para as parcelas 10 e 11, dessa forma:

$I_{(1+2)} = 2(4 + 1)xln2 = 6{,}93$ e $I_{(10+11)} = 2(1 + 1)xln2 = 2{,}77$

Logo, para o Índice de Informação, teremos uma matriz 11 x 11, conforme a saída do R abaixo.

## 5.1 Passo a passo no R: Índice de Informação

### a) Calculando a Matriz do Índice de Informação

Para calcular o Índice de Informação entre as parcelas, utilizaremos uma fórmula específica, que mede a diferença entre elas. O resultado será uma matriz em que cada valor representa o grau de dissimilaridade entre um par de parcelas. Usaremos a matriz X_binaria, na qual as parcelas estão nas linhas e as espécies nas colunas.

```
# Calculando a matriz do Índice de Informação entre as parcelas
# A função outer() aplica a fórmula para cada par de linhas (parcelas) da X_binaria
matriz_indice_informacao <- outer(
  1:nrow(X_binaria), # Itera sobre as linhas (parcelas) da X_binaria
  1:nrow(X_binaria), # Itera novamente sobre as linhas (parcelas) da X_binaria
  Vectorize(function(i, j) {
    # Fórmula do Índice de Informação para cada par de parcelas
```

```
    sum(abs(X_binaria[i, ] - X_binaria[j, ])) * 2 * log(2)
  })
)

# Nomeando as linhas e colunas da matriz do Índice de Informação com os nomes das parcelas
rownames(matriz_indice_informacao) <- rownames(X_binaria)
colnames(matriz_indice_informacao) <- rownames(X_binaria)

# Visualizando o resultado final da matriz do Índice de Informação, arredondado para duas casas decimais
print(round(matriz_indice_informacao, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 6.93 | 9.70 | 8.32 | 5.55 | 5.55 | 15.25 | 4.16 | 11.09 | 8.32 | 11.09 |
| P2 | 6.93 | 0.00 | 11.09 | 6.93 | 6.93 | 6.93 | 11.09 | 8.32 | 6.93 | 12.48 | 12.48 |
| P3 | 9.70 | 11.09 | 0.00 | 9.70 | 4.16 | 6.93 | 13.86 | 11.09 | 12.48 | 6.93 | 9.70 |
| P4 | 8.32 | 6.93 | 9.70 | 0.00 | 5.55 | 5.55 | 12.48 | 9.70 | 8.32 | 8.32 | 8.32 |
| P5 | 5.55 | 6.93 | 4.16 | 5.55 | 0.00 | 2.77 | 15.25 | 9.70 | 11.09 | 5.55 | 8.32 |
| P6 | 5.55 | 6.93 | 6.93 | 5.55 | 2.77 | 0.00 | 15.25 | 6.93 | 11.09 | 8.32 | 8.32 |
| P7 | 15.25 | 11.09 | 13.86 | 12.48 | 15.25 | 15.25 | 0.00 | 11.09 | 6.93 | 12.48 | 9.70 |
| P8 | 4.16 | 8.32 | 11.09 | 9.70 | 9.70 | 6.93 | 11.09 | 0.00 | 6.93 | 9.70 | 9.70 |
| P9 | 11.09 | 6.93 | 12.48 | 8.32 | 11.09 | 11.09 | 6.93 | 6.93 | 0.00 | 11.09 | 8.32 |
| P10 | 8.32 | 12.48 | 6.93 | 8.32 | 5.55 | 8.32 | 12.48 | 9.70 | 11.09 | 0.00 | 2.77 |
| P11 | 11.09 | 12.48 | 9.70 | 8.32 | 8.32 | 8.32 | 9.70 | 9.70 | 8.32 | 2.77 | 0.00 |

## 6. Distância binária de Sokal

A distância binária de Sokal **($dBS_{A,B}$)** indica a proporção de atributos não coincidentes nos dois indivíduos. Quanto maior esse número, menos semelhantes serão os indivíduos, sendo, portanto, uma medida de dissimilaridade. Sua amplitude varia entre **0** e **1**. Uma série de outros coeficientes derivam desse conceito.

Essa distância é obtida por meio da seguinte equação:

$$dBS_{A,B} = \left[\frac{(b+c)}{(a+b+c+d)}\right]^{1/2}$$

**Aplicação**

Para as Parcelas 1 e 2 (Tabela 8, página 26), temos **a** = 4, **b** = 4, **c** = 1 e **d** = 8, logo:

$$dBS_{1,2} = \left[\frac{(4+1)}{(4+4+1+8)}\right]^{1/2} = 0{,}54$$

Já para as Parcelas 10 e 11 (Tabela 10, página 26), temos **a** = 11, **b** = 1, **c** = 1 e **d** = 4, assim:

$$dBS_{10,11} = \left[\frac{(1+1)}{(11+1+1+4)}\right]^{1/2} = 0{,}34$$

## 6.1 Passo a passo no R: Distância binária de Sokal

### a) Desenvolvendo a função do índice

A Distância de Sokal é uma medida de dissimilaridade que nos mostra a proporção de espécies que não aparecem juntas em duas parcelas. Quanto maior o valor, mais diferentes são as parcelas. Para calculá-la, podemos criar uma função personalizada no R que aplica a fórmula da distância de Sokal para cada par de parcelas. Usaremos a matriz X_binaria, na qual as parcelas estão nas linhas e as espécies, nas colunas.

```
# Definindo o número de espécies (colunas da X_binaria)
n_especies <- ncol(X_binaria)

# Criando a função para calcular a Distância de Sokal entre duas parcelas
sokal_dist <- function(parcela1, parcela2) {
  # Calcula a diferença absoluta entre as presenças/ausências das espécies
  diferencas <- abs(parcela1 - parcela2)
  # Aplica a fórmula da Distância de Sokal
  sqrt(sum(diferencas) / n_especies)
}
```

### b) Calculando a Matriz de Distância de Sokal

Para obter a Distância de Sokal para todas as combinações de parcelas, aplicaremos a função *sokal_dist* a cada par de linhas da nossa matriz X_binaria. O resultado será uma matriz de dissimilaridade, na qual cada valor indica o grau de diferença entre as parcelas.

```
# Calculando a matriz de Distância de Sokal entre as parcelas
# A função outer() aplica a função sokal_dist para cada par de linhas (parcelas) da X_binaria
matriz_dist_sokal <- outer(
  1:nrow(X_binaria), # Itera sobre as linhas (parcelas) da X_binaria
  1:nrow(X_binaria), # Itera novamente sobre as linhas (parcelas) da X_binaria
  Vectorize(function(i, j) {
    # Aplica a função sokal_dist para os dados das parcelas i e j
    sokal_dist(X_binaria[i, ], X_binaria[j, ])
  })
)

# Nomeando as linhas e colunas da matriz de Distância de Sokal com os nomes das parcelas
rownames(matriz_dist_sokal) <- rownames(X_binaria)
colnames(matriz_dist_sokal) <- rownames(X_binaria)

# Visualizando o resultado final da matriz de Distância de Sokal, arredondado para duas casas decimais
print(round(matriz_dist_sokal, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.54 | 0.64 | 0.59 | 0.49 | 0.49 | 0.80 | 0.42 | 0.69 | 0.59 | 0.69 |
| P2 | 0.54 | 0.00 | 0.69 | 0.54 | 0.54 | 0.54 | 0.69 | 0.59 | 0.54 | 0.73 | 0.73 |
| P3 | 0.64 | 0.69 | 0.00 | 0.64 | 0.42 | 0.54 | 0.77 | 0.69 | 0.73 | 0.54 | 0.64 |
| P4 | 0.59 | 0.54 | 0.64 | 0.00 | 0.49 | 0.49 | 0.73 | 0.64 | 0.59 | 0.59 | 0.59 |
| P5 | 0.49 | 0.54 | 0.42 | 0.49 | 0.00 | 0.34 | 0.80 | 0.64 | 0.69 | 0.49 | 0.59 |
| P6 | 0.49 | 0.54 | 0.54 | 0.49 | 0.34 | 0.00 | 0.80 | 0.54 | 0.69 | 0.59 | 0.59 |
| P7 | 0.80 | 0.69 | 0.77 | 0.73 | 0.80 | 0.80 | 0.00 | 0.69 | 0.54 | 0.73 | 0.64 |
| P8 | 0.42 | 0.59 | 0.69 | 0.64 | 0.64 | 0.54 | 0.69 | 0.00 | 0.54 | 0.64 | 0.64 |
| P9 | 0.69 | 0.54 | 0.73 | 0.59 | 0.69 | 0.69 | 0.54 | 0.54 | 0.00 | 0.69 | 0.59 |
| P10 | 0.59 | 0.73 | 0.54 | 0.59 | 0.49 | 0.59 | 0.73 | 0.64 | 0.69 | 0.00 | 0.34 |
| P11 | 0.69 | 0.73 | 0.64 | 0.59 | 0.59 | 0.59 | 0.64 | 0.64 | 0.59 | 0.34 | 0.00 |

## 7. Coeficiente de concordância simples

O Coeficiente de concordância simples ($\mathbf{S_{A,B}}$) é um dos coeficientes que derivam da distância binária de Sokal, ou seja, uma primeira transformação é consiste em construir um coeficiente de similaridade, sendo a proposta mais usada a proporção de coincidências, dada por:

$$S_{A,B} = \frac{a+d}{a+b+c+d}$$

Assim, os valores maiores significam maior similaridade entre os indivíduos, variando entre 0 e 1.

**Aplicação**

Para as Parcelas 1 e 2 (Tabela 8, página 26), temos **a** = 4, **b** = 4, **c** = 1 e **d** = 8, e, para as Parcelas 10 e 11 (Tabela 10, página 26), temos **a** = 11, **b** = 1, **c** = 1 e **d** = 4, então:

$S_{1,2} = \frac{4+8}{4+4+1+8} = 0{,}71$ e $S_{10,11} = \frac{11+4}{11+1+1+4} = 0{,}88$

### 7.1 Passo a passo no R: Coeficiente de concordância simples

### a) Calculando a Matriz de Similaridade

O Coeficiente de Concordância Simples ($\mathbf{S_{A,B}}$) mede a proporção de concordâncias entre duas parcelas, levando em conta tanto as espécies presentes em ambas quanto as ausentes em ambas. Para calcular a matriz de similaridade, utilizaremos o pacote *proxy*, que possui a medida "*simple matching*" implementada.

```
# Carregamos o pacote para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(proxy)

# Calculamos a matriz de similaridade pelo método da Concordância Simples
# Como nossa matriz X_binaria já está com as parcelas nas linhas,
# a função simil() calculará a similaridade entre elas diretamente.
matriz_sim_concordancia <- simil(
```

```
  X_binaria,
  method = "simple matching"
)

# Exibimos a matriz de similaridade, arredondada para duas casas decimais
print(round(as.matrix(matriz_sim_concordancia), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00* | 0.71 | 0.59 | 0.65 | 0.76 | 0.76 | 0.35 | 0.82 | 0.53 | 0.65 | 0.53 |
| P2 | 0.71 | 1.00 | 0.53 | 0.71 | 0.71 | 0.71 | 0.53 | 0.65 | 0.71 | 0.47 | 0.47 |
| P3 | 0.59 | 0.53 | 1.00 | 0.59 | 0.82 | 0.71 | 0.41 | 0.53 | 0.47 | 0.71 | 0.59 |
| P4 | 0.65 | 0.71 | 0.59 | 1.00 | 0.76 | 0.76 | 0.47 | 0.59 | 0.65 | 0.65 | 0.65 |
| P5 | 0.76 | 0.71 | 0.82 | 0.76 | 1.00 | 0.88 | 0.35 | 0.59 | 0.53 | 0.76 | 0.65 |
| P6 | 0.76 | 0.71 | 0.71 | 0.76 | 0.88 | 1.00 | 0.35 | 0.71 | 0.53 | 0.65 | 0.65 |
| P7 | 0.35 | 0.53 | 0.41 | 0.47 | 0.35 | 0.35 | 1.00 | 0.53 | 0.71 | 0.47 | 0.59 |
| P8 | 0.82 | 0.50 | 0.50 | 0.53 | 0.53 | 0.67 | 0.50 | 1.00 | 0.67 | 0.63 | 0.63 |
| P9 | 0.53 | 0.71 | 0.47 | 0.65 | 0.53 | 0.53 | 0.71 | 0.71 | 1.00 | 0.53 | 0.65 |
| P10 | 0.65 | 0.47 | 0.71 | 0.65 | 0.76 | 0.65 | 0.47 | 0.59 | 0.53 | 1.00 | 0.88 |
| P11 | 0.53 | 0.47 | 0.59 | 0.65 | 0.65 | 0.65 | 0.59 | 0.59 | 0.65 | 0.88 | 1.00 |

* Na saída do R, o coeficiente relativo à mesma parcela sai *NA*, o qual foi modificado para 1.00, pois corresponde à máxima similaridade.

## 8. Coeficiente de concordâncias positivas (Coeficiente de Russel e Rao)

O coeficiente de concordâncias positivas também é conhecido como Coeficiente de Russel e Rao (Russel; Rao, 1940). Ele varia de 0 a 1, em que 0 indica que não há concordância positiva entre os objetos e 1 indica que eles são idênticos.

Nos casos em que se deseja medir a similaridade, baseando-se apenas na presença da característica, e não na ausência, tem-se o seguinte coeficiente:

$$CP_{A,B} = \frac{a}{a+b+c+d}$$

**Aplicação**

Para as Parcelas 1 e 2, temos **a** = 4, **b** = 4, **c** = 1 e **d** = 8 (Tabela 8, página 26), e, para as Parcelas 10 e 11, temos **a** = 11, **b** = 1, **c** = 1 e **d** = 4 (Tabela 10, página 26), então:

$CP_{1,2} = \frac{4}{4+4+1+8} = 0{,}24$ e $CP_{10,11} = \frac{11}{11+1+1+4} = 0{,}65$

### 8.1 Passo a passo no R: Coeficiente de concordâncias positivas (Coeficiente de Russel e Rao)

#### a) Calculando a Matriz de Similaridade

O Coeficiente de Russel e Rao mede a proporção de espécies que estão presentes simultaneamente em duas parcelas em relação ao número total de espécies avaliadas. Quanto maior o número de espécies compartilhadas, maior a similaridade. Para calcular a matriz de

similaridade entre as parcelas, utilizaremos o pacote *proxy*, que possui a medida "Russel" implementada.

```
# Carregamos o pacote para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(proxy)

# Calculamos a matriz de similaridade pelo coeficiente de Russel e Rao
# Como nossa matriz X_binaria já está com as parcelas nas linhas,
# a função simil() calculará a similaridade entre elas diretamente.
matriz_sim_russel_rao <- simil(
  X_binaria,
  method = "Russel"
)

# Exibimos a matriz de similaridade, arredondada para duas casas decimais
print(round(as.matrix(matriz_sim_russel_rao), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00* | 0.24 | 0.29 | 0.29 | 0.35 | 0.35 | 0.18 | 0.35 | 0.24 | 0.41 | 0.35 |
| P2 | 0.24 | 1.00 | 0.18 | 0.24 | 0.24 | 0.24 | 0.18 | 0.18 | 0.24 | 0.24 | 0.24 |
| P3 | 0.29 | 0.18 | 1.00 | 0.29 | 0.41 | 0.35 | 0.24 | 0.24 | 0.24 | 0.47 | 0.41 |
| P4 | 0.29 | 0.24 | 0.29 | 1.00 | 0.35 | 0.35 | 0.24 | 0.24 | 0.29 | 0.41 | 0.41 |
| P5 | 0.35 | 0.24 | 0.41 | 0.35 | 1.00 | 0.41 | 0.18 | 0.24 | 0.24 | 0.47 | 0.41 |
| P6 | 0.35 | 0.24 | 0.35 | 0.35 | 0.41 | 1.00 | 0.18 | 0.29 | 0.24 | 0.41 | 0.41 |
| P7 | 0.18 | 0.18 | 0.24 | 0.24 | 0.18 | 0.18 | 1.00 | 0.24 | 0.35 | 0.35 | 0.41 |
| P8 | 0.35 | 0.18 | 0.24 | 0.24 | 0.24 | 0.29 | 0.24 | 1.00 | 0.29 | 0.35 | 0.35 |
| P9 | 0.24 | 0.24 | 0.24 | 0.29 | 0.24 | 0.24 | 0.35 | 0.29 | 1.00 | 0.35 | 0.41 |
| P10 | 0.41 | 0.24 | 0.47 | 0.41 | 0.47 | 0.41 | 0.35 | 0.35 | 0.35 | 1.00 | 0.65 |
| P11 | 0.35 | 0.24 | 0.41 | 0.41 | 0.41 | 0.41 | 0.41 | 0.35 | 0.41 | 0.65 | 1.00 |

* Na saída do R, o coeficiente relativo à mesma parcela sai *NA*, o qual foi modificado para 1.00, pois corresponde à máxima similaridade.

## 9. Outros coeficientes ou índices de similaridade

A partir dos coeficientes ou índices de similaridade descritos dos itens **1** a **8**, vários outros foram formulados, tentando ressaltar propriedades específicas (Tabela 11). Vale ressaltar que a maioria dos coeficientes sofre influência de codificação.

Para escolha de um coeficiente de similaridade, Romesburg (2004) levanta os seguintes pontos a serem considerados:

1. Os coeficientes de Jaccard, concordância simples e Sorenson são muito utilizados. Talvez isso decorra de suas bases conceituais claras: todas são proporções significativas.

2. Como esperado, o compartilhamento de variáveis comuns (**a**, **b**, **c**, **d**) tende a correlacionar os coeficientes.

3. Devido às suas propriedades matemáticas, em alguns casos, os coeficientes estão perfeitamente correlacionados, independentemente dos dados utilizados com eles.

**Tabela 11.** Alguns coeficientes de similaridade para variáveis binárias

| Coeficientes | Expressão | Intervalo de Variação | Exemplo* |
|---|---|---|---|
| Rogers e Tanimoto | $\frac{a+d}{a+2(b+c)+d}$ | [0;1] | 0,55 |
| Sokal e Sneath | $\frac{2(a+d)}{2(a+d)+b+c}$ | [0;1] | 0,83 |
| Ochiai | $\frac{a}{[(a+b)x(a+c)]^{1/2}}$ | [0;1] | 0,63 |
| Baroni-Urbani e Buser | $\frac{a+(ad)^{1/2}}{a+b+c+(ad)^{1/2}}$ | [0;1] | 0,66 |
| Hamman | $\frac{(a+d)-(b+c)}{a+b+c+d}$ | [-1;1] | 0,41 |
| Yule | $\frac{ad-bc}{ad+bc}$ | [-1;1] | 0,78 |
| Ochiai II | $\frac{ad}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$ | [0;1] | 0,49 |
| Indice II | $\frac{1}{2}x\left[\left(\frac{a}{a+b}\right)+\left(\frac{a}{a+c}\right)\right]$ | [0;1] | 0,65 |
| Indice III | $\frac{1}{4}x\left[\left(\frac{a}{a+b}\right)+\left(\frac{a}{a+c}\right)+\left(\frac{d}{d+b}\right)+\left(\frac{d}{d+c}\right)\right]$ | [0;1] | 0,71 |
| Sokal e Michener | $\frac{(a+b)}{a+b+c+d}$ | [0;1] | 0,47 |
| Phi (j) | $\frac{ad-bc}{[(a+b)(a+c)(b+d)+(c+d)]^{1/2}}$ | [-1;1] | 0,43 |
| Kulczynski I | $\frac{a}{b+c}$ | [0;∞] | 0,80 |
| Kulczynski II | $\frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$ | [0;1] | 0,65 |
| Anderberg | $\frac{a}{a+2(b+c)}$ | [0;1] | 0,29 |
| Pearson | $\frac{(ad-bc)}{(a+b)(c+d)(a+c)(b+d)}$ | [-1;1] | 0,006 |
| McConnaughy | $\frac{a^2-bc}{(a+b)(a+c)}$ | [-1;1] | 0,30 |

Em que: (1) **a** =1:1; **b** =1:0; **c** = 0:1; e **d** = 0:0. * Cálculos realizados com base nos dados da Tabela 8 (página 26).
Fonte: Adaptada a partir de Bussab et al. (1990), Romesburg (2004) e Cruz e Carneiro (2013).

4. Os coeficientes de similaridade são mais comumente expressos como funções das variáveis **a**, **b**, **c**, **d**. Há, no entanto, uma segunda maneira de expressá-los que usa símbolos diferentes para alcançar os mesmos resultados.

5. Há muitos coeficientes de similaridade para atributos qualitativos, cada um com suas propriedades matemáticas, as quais devem ser consideradas pelo pesquisador.

6. Embora a escolha de um coeficiente de similaridade deva ser guiada, na medida do possível, pela lógica, há uma tendência entre os pesquisadores, dentro do seu campo, de fazer uso habitual de certos coeficientes que outros, de campo distinto, estão utilizando, mesmo quando faltam razões lógicas.

Vale ressaltar, ainda, que as variáveis binárias (**a**, **b**, **c** e **d**) podem ser simétricas e não simétricas. As simétricas não possuem preferência na codificação; o resultado não sofre alterações quando os códigos são alterados, sendo assim, **a** e **d** têm a mesma função. Já nas

assimétricas, a modificação altera os resultados, ou seja, **a** (1:1) indica similaridade e **d** (0:0) não a indica, necessariamente (Frei, 2006).

Na Tabela 11, estão ilustrados coeficientes quanto à influência da codificação, além deles, pode-se afirmar que os de Jaccard e Sorensen são sensíveis. Além disso, desses coeficientes, sabe-se que alguns fazem com que o **d** (0:0) contribua para a semelhança entre objetos, enquanto outros não, logo, a contribuição de **d** depende do objetivo da pesquisa e do contexto do problema estudado (Romesburg, 2004).

## c) Coeficientes de similaridade para variáveis quantitativas

Entre os índices e coeficientes que utilizam dados quantitativos e que podem ser considerados como função de agrupamento, podem-se citar:

### 1. Coeficiente de Ellenberg

O coeficiente de similaridade de Ellenberg (**CE**), apresentado por Matteucci e Colma (1982), é obtido por meio da expressão:

$$CE_{AB} = \frac{\sum_t (X_{Aj} + X_{Bj})}{\sum_t (X_{Aj} + X_{Bj}) + 2(\sum_u X_{Aj} + \sum_v X_{Bj})}$$

Em que: **t** = subconjunto de amostras em que as espécies **A** e **B** aparecem juntas; **u** = subconjunto de amostras em que a espécie A aparece só; e **v** = subconjunto de amostras em que a espécie B aparece só; $\mathbf{x_{Aj}}$ e $\mathbf{x_{Bj}}$ = quantidades das espécies **A** e **B** na **j**-ésima parcela, respectivamente. Se as espécies **A** e **B** aparecem sempre juntas, então $\mathbf{CE_{(A,B)}}$ = 1, caso contrário, $\mathbf{CE_{(A,B)}}$ = 0.

**Aplicação**

Em relação às parcelas 1 e 2 (Tabela 3; página 17), os vetores coluna transpostos (parcelas) $\boldsymbol{X'_{1j}}$ e $\boldsymbol{X'_{2j}}$ fornecem os seguintes subconjuntos:

$$X'_{1j} = [8\ 0\ 3\ 6\ 00\ 0\ 0\ 1\ 2\ 1\ 0\ 5\ 0\ 7\ 0\ 0\ 0]$$

$$X'_{2j} = [1\ 0\ 9\ 3\ 12\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

$$t' = \begin{bmatrix} 8 & 3 & 6 & 1 \\ 1 & 9 & 3 & 1 \end{bmatrix}$$

$$u' = [2\ 1\ 5\ 7]$$

$$v = [12]$$

$$\sum_t (X_{1j} + X_{2j}) = (8+1) + (3+9) + (6+3) + (1+1) = 32$$

$$\sum_u X_{1j} = 2 + 1 + 5 + 7 = 15$$

$$\sum_u X_{2j} = 12 = 12$$

$$CE_{12} = \frac{32}{32 + 2(15 + 12)} = 0{,}37$$

Já em relação às parcelas 10 e 11 (Tabela 3), teremos:

$$X'_{10j} = [15\ \ 1\ \ 4\ \ 4\ \ 0\ 11\ 11\ \ 3\ \ 2\ \ 3\ \ 0\ \ 1\ \ 5\ \ 0\ \ 2\ \ 0\ \ 0]$$

$$X'_{11j} = [\ 7\ \ 9\ \ 4\ \ 4\ \ 0\ \ 1\ \ 5\ \ 1\ \ 2\ \ 1\ \ 5\ \ 0\ \ 2\ \ 0\ \ 1\ \ 0\ \ 0]$$

$$t' = \begin{bmatrix} 15\ 01\ 04\ \ 04\ 11\ 11\ 03\ 02\ 03\ 05\ 02 \\ 07\ 09\ 04\ \ 04\ 01\ 05\ 01\ 02\ 01\ 02\ 01 \end{bmatrix}$$

$$u' = [1]$$

$$v = [5]$$

$\sum_t (X_{10j} + X_{11j}) = (15 + 7) + (1 + 9) + (4 + 4) + +(4 + 4) + (11 + 1) + (11 + 5) + (3 + 1) + (2 + 2) + (3 + 1) + (5 + 2) + (2 + 1)$

$$\sum_t (X_{10j} + X_{11j}) = 98$$

$$\sum_u X_{1j} = 1$$

$$\sum_u X_{2j} = 5$$

$$CE_{1011} = \frac{98}{98 + 2(1 + 5)} = 0{,}89$$

## 1.1 Passo a passo no R: Coeficiente de Ellenberg

### a) Desenvolvendo a função do coeficiente

O Coeficiente de Ellenberg mede a similaridade entre duas parcelas, considerando a abundância das espécies. Ele valoriza as espécies compartilhadas e leva em conta as espécies exclusivas de cada parcela. Para calculá-lo, criaremos uma função personalizada no R.

```
# Criando a função para calcular o Coeficiente de Ellenberg entre duas parcelas
ellenberg_sim <- function(parcela1, parcela2) {
  # Identifica espécies presentes em ambas as parcelas
  comum <- (parcela1 > 0) & (parcela2 > 0)
  # Identifica espécies presentes apenas na parcela1
  so_a  <- (parcela1 > 0) & (parcela2 == 0)
  # Identifica espécies presentes apenas na parcela2
  so_b  <- (parcela1 == 0) & (parcela2 > 0)

  # Soma das abundâncias das espécies comuns nas duas parcelas
  sum_t <- sum(parcela1[comum] + parcela2[comum])
```

```
  # Soma das abundâncias das espécies exclusivas da parcela1
  sum_u <- sum(parcela1[so_a])
  # Soma das abundâncias das espécies exclusivas da parcela2
  sum_v <- sum(parcela2[so_b])

  # Denominador da fórmula de Ellenberg
  den <- sum_t + 2 * (sum_u + sum_v)

  # Retorna o coeficiente (evita divisão por zero)
  if (den == 0) 0 else sum_t / den
}
```

**b) Calculando a Matriz de Similaridade de Ellenberg**

Para obter o Coeficiente de Ellenberg para todas as combinações de parcelas, aplicaremos a função *ellenberg_sim* a cada par de colunas da nossa matriz X (matriz quantitativa), pois as parcelas estão organizadas nas colunas.

```
# Calculando a matriz de similaridade de Ellenberg entre as parcelas
# A função outer() aplica a função ellenberg_sim para cada par de colunas (parcelas) da X
matriz_sim_ellenberg <- outer(
  1:ncol(X), # Itera sobre as colunas (parcelas) da matriz X
  1:ncol(X), # Itera novamente sobre as colunas (parcelas) da matriz X
  Vectorize(function(i, j) {
    # Aplica a função ellenberg_sim para os dados das colunas (parcelas) i e j
    ellenberg_sim(X[, i], X[, j])
  })
)

# Nomeando as linhas e colunas da matriz de similaridade de Ellenberg com os nomes das parcelas
rownames(matriz_sim_ellenberg) <- colnames(X)
colnames(matriz_sim_ellenberg) <- colnames(X)

# Visualizando o resultado da matriz de similaridade de Ellenberg, arredondado para duas casas decimais
print(round(matriz_sim_ellenberg, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.37 | 0.61 | 0.38 | 0.62 | 0.81 | 0.19 | 0.77 | 0.40 | 0.44 | 0.36 |
| P2 | 0.37 | 1.00 | 0.41 | 0.70 | 0.48 | 0.56 | 0.36 | 0.30 | 0.54 | 0.29 | 0.28 |
| P3 | 0.61 | 0.41 | 1.00 | 0.29 | 0.97 | 0.66 | 0.43 | 0.48 | 0.46 | 0.86 | 0.60 |
| P4 | 0.38 | 0.70 | 0.29 | 1.00 | 0.81 | 0.68 | 0.21 | 0.58 | 0.27 | 0.52 | 0.41 |
| P5 | 0.62 | 0.48 | 0.97 | 0.81 | 1.00 | 0.91 | 0.22 | 0.52 | 0.56 | 0.85 | 0.68 |
| P6 | 0.81 | 0.56 | 0.66 | 0.68 | 0.91 | 1.00 | 0.20 | 0.53 | 0.44 | 0.80 | 0.67 |
| P7 | 0.19 | 0.36 | 0.43 | 0.21 | 0.22 | 0.20 | 1.00 | 0.43 | 0.75 | 0.50 | 0.74 |
| P8 | 0.77 | 0.30 | 0.48 | 0.58 | 0.52 | 0.53 | 0.43 | 1.00 | 0.71 | 0.42 | 0.60 |
| P9 | 0.40 | 0.54 | 0.46 | 0.27 | 0.56 | 0.44 | 0.75 | 0.71 | 1.00 | 0.64 | 0.77 |
| P10 | 0.44 | 0.29 | 0.86 | 0.52 | 0.85 | 0.80 | 0.50 | 0.42 | 0.64 | 1.00 | 0.89 |
| P11 | 0.36 | 0.28 | 0.60 | 0.41 | 0.68 | 0.67 | 0.74 | 0.60 | 0.77 | 0.89 | 1.00 |

## 2. Coeficiente de correlação de Pearson

O Coeficiente de correlação de Pearson (**r**) é obtido por meio da seguinte equação:

$$r = \frac{\sum_{j=1}^{p}\left(X_{Aj} - \bar{X}_{Aj}\right)\left(X_{Bj} - \bar{X}_{Bj}\right)}{\left[\sum_{j=1}^{p}\left(X_{Aj} - \bar{X}_{Aj}\right)^2 \sum_{j=1}^{p}\left(X_{Bj} - \bar{X}_{Bj}\right)^2\right]^{1/2}}$$

Em que: $\bar{\boldsymbol{X}}_{\boldsymbol{A}}$e $\bar{\boldsymbol{X}}_{\boldsymbol{B}}$ são, respectivamente, os valores médios das espécies **A** e **B** no conjunto de **p** parcelas amostrais. É importante ressaltar que, quando se empregam valores **0** e **1**, isto é, ausência (-) e presença (+), o coeficiente obtido é o coeficiente de correlação puntual, apresentado anteriormente. O seu emprego como medida de similaridade é restrito aos casos em que as variáveis são do tipo quantitativas ou não codificadas. Além do mais, não atende à terceira propriedade que define uma medida de similaridade ($0 \leq \mathbf{S\,(E_i,E_j)} \leq 1$, para $\mathbf{i \neq j}$.

### Aplicação

Em relação às parcelas 1 e 2 (Tabela 3, página 17), temos os vetores coluna transpostos (parcelas) $X'_{1j}$ e $X'_{2j}$:

$$X'_{1j} = [8\ 0\ 3\ 6\ 00\ 0\ 0\ 1\ 2\ 1\ 0\ 5\ 0\ 7\ 0\ 0\ 0]$$
$$X'_{2j} = [1\ 0\ 9\ 3\ 12\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

Assim:

$$\sum_{j=1}^{p}\left(X_{1j} - \bar{X}_{1j}\right)\left(X_{2j} - \bar{X}_{2j}\right) = \sum_{j=1}^{p} X_{1j}X_{2j} - \frac{\sum_{j=1}^{p} X_{1j} \sum_{j=1}^{p} X_{2j}}{p} = \left(54 - \frac{33x26}{17}\right) = 3{,}53$$

$\sum_{j=1}^{p} X_{1j}X_{2j} = (8x1) + (0x0) + (3x9) + (6x3) + (0x12) + (0x0) + (0x0) + (1x1) + (2x0) + (1x0) + (0x0) + (5x0) + (0x0) + (7x0) + (0x0) + (0x0) + (0X0)$

$$\sum_{j=1}^{p} X_{1j}X_{2j} = 54$$

$$\sum_{j=1}^{p} X_{1j} = 8 + 0 + 3 + 6 + 0 + 0 + 0 + 1 + 2 + 1 + 0 + 5 + 0 + 7 + 0 + 0 + 0 = 33$$

$$\sum_{j=1}^{p} X_{2j} = 1 + 0 + 9 + 3 + 12 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 26$$

$$\sum_{j=1}^{p}\left(X_{1j} - \bar{X}_{1j}\right)\left(X_{2j} - \bar{X}_{2j}\right) = \left(54 - \frac{33x26}{17}\right) = 3{,}53$$

$$\sum_{j=1}^{p}\left(X_{1j} - \bar{X}_{1j}\right)^2 = \sum_{j=1}^{p} X_{1j}^2 - \frac{\left(\sum_{j=1}^{p} X_{1j}\right)^2}{p}$$

$$\sum_{j=1}^{p} X_{1j}^2 = 8^2 + 0^2 + 3^2 + 6^2 + 0^2 + 0^2 + 0^2 + 1^2 + 2^2 + 1^2 + 0^2 + 5^2 + 0^2 + 7^2 + 0^2 + 0^2 + 0^2$$

$$\sum_{j=1}^{p} X_{1j}^2 = 189$$

$$\sum_{j=1}^{p} \left(X_{1j} - \bar{X}_{1j}\right)^2 = \left(189 - \frac{(33)^2}{17}\right) = 124{,}94$$

$$\sum_{j=1}^{p} \left(X_{2j} - \bar{X}_{2j}\right)^2 = \sum_{j=1}^{p} X_{2j}^2 - \frac{\left(\sum_{j=1}^{p} X_{2j}\right)^2}{p}$$

$$\sum_{j=1}^{p} X_{2j}^2 = 1^2 + 0^2 + 9^2 + 3^2 + 12^2 + 0^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2$$

$$\sum_{j=1}^{p} X_{2j}^2 = 236$$

$$\sum_{j=1}^{p} \left(X_{2j} - \bar{X}_{2j}\right)^2 = \left(236 - \frac{(26)^2}{17}\right) = 196{,}24$$

$$r_{1,2} = \frac{\sum_{j=1}^{p} \left(X_{1j} - \bar{X}_{1j}\right)\left(X_{2j} - \bar{X}_{2j}\right)}{\left[\sum_{j=1}^{p} \left(X_{1j} - \bar{X}_{1j}\right)^2 \sum_{j=1}^{p} \left(X_{2j} - \bar{X}_{2j}\right)^2\right]^{1/2}}$$

$$r_{12} = \frac{3{,}53}{[124{,}94x196{,}24]^{1/2}} = 0{,}02$$

Já em relação às parcelas 10 e 11 (Tabela 3, página 17), teremos:

$$X'_{10j} = [15\ \ 1\ \ 4\ \ 4\ \ 0\ 11\ 11\ \ 3\ \ 2\ \ 3\ \ 0\ \ 1\ \ 5\ \ 0\ \ 2\ \ 0\ \ 0]$$

$$X'_{11j} = [\ 7\ \ 9\ \ 4\ \ 4\ \ 0\ \ 1\ \ 5\ \ 1\ \ 2\ \ 1\ \ 5\ \ 0\ \ 2\ \ 0\ \ 1\ \ 0\ \ 0]$$

Assim:

$$\sum_{j=1}^{p} \left(X_{10j} - \bar{X}_{10j}\right)\left(X_{11j} - \bar{X}_{11j}\right) = \sum_{j=1}^{p} X_{10j}X_{11j} - \frac{\sum_{j=1}^{p} X_{10j} \sum_{j=1}^{p} X_{11j}}{p} = \left(234 - \frac{62x42}{17}\right) = 80{,}82$$

$$\sum_{j=1}^{p} X_{10j}X_{11j} = (15x7) + (1x9) + (4x4) + (4x4) + (0x0) + (11x1) + (11x5) + (3x1) + (2x2) + (3x1) + (0x5) + (1x0) + (5x2) + (0x0) + (2x1) + (0x0) + (0x0)$$

$$\sum_{j=1}^{p} X_{10j}X_{11j} = 234$$

$$\sum_{j=1}^{p} \left(X_{10j} - \bar{X}_{10j}\right)\left(X_{11j} - \bar{X}_{11j}\right) = 80{,}82$$

$$\sum_{j=1}^{p} X_{10j} = 15 + 1 + 4 + 4 + 0 + 11 + 11 + 3 + 2 + 3 + 0 + 1 + 5 + 0 + 2 + 0 + 0 = 62$$

$$\sum_{j=1}^{p} X_{11j} = 7 + 9 + 4 + 4 + 0 + 1 + 5 + 1 + 2 + 1 + 5 + 2 + 0 + 1 + 0 + 0 + 0 = 42$$

$$\sum_{j=1}^{p}\left(X_{10j} - \bar{X}_{10j}\right)\left(X_{11j} - \bar{X}_{11j}\right) = \left(234 - \frac{62x42}{17}\right) = 80{,}82$$

$$\sum_{j=1}^{p}\left(X_{10j} - \bar{X}_{10j}\right)^2 = \sum_{j=1}^{p} X_{10j}^2 - \frac{\left(\sum_{j=1}^{p} X_{10j}\right)^2}{p}$$

$\sum_{j=1}^{p} X_{10j}^2 = 15^2 + 1^2 + 4^2 + 4^2 + 0^2 + 11^2 + 11^2 + 3^2 + 2^2 + 3 + 0^2 + 1^2 + 5^2 + 0^2 + 2^2 + 0^2 + 0^2 = 546$

$$\sum_{j=1}^{p}\left(X_{10j} - \bar{X}_{10j}\right)^2 = \left(546 - \frac{(62)^2}{17}\right) = 319{,}88$$

$$\sum_{j=1}^{p}\left(X_{11j} - \bar{X}_{11j}\right)^2 = \sum_{j=1}^{p} X_{11j}^2 - \frac{\left(\sum_{j=1}^{p} X_{11j}\right)^2}{p}$$

$\sum_{j=1}^{p} X_{11j}^2 = 7^2 + 9^2 + 4^2 + 4^2 + 0^2 + 1^2 + 5^2 + 1^2 + 2^2 + 1^2 + 5^2 + 2^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2 = 224$

$$\sum_{j=1}^{p}\left(X_{11j} - \bar{X}_{11j}\right)^2 = \left(224 - \frac{(42)^2}{17}\right) = 120{,}24$$

$$r_{1,2} = \frac{\sum_{j=1}^{p}\left(X_{10j} - \bar{X}_{10j}\right)\left(X_{11j} - \bar{X}_{11j}\right)}{\left[\sum_{j=1}^{p}\left(X_{10j} - \bar{X}_{10j}\right)^2\left(X_{11j} - \bar{X}_{11j}\right)^2\right]^{1/2}}$$

$$r_{12} = \frac{80{,}82}{[319{,}88x120{,}24]^{1/2}} = 0{,}41$$

### 2.1. Passo a passo no R: Coeficiente de correlação de Pearson

#### a) Calculando o coeficiente

O Coeficiente de Correlação de Pearson mede a intensidade e a direção da relação linear entre duas parcelas com base na abundância das espécies. Seus valores variam de -1 a +1: valores próximos de +1 indicam forte associação positiva; valores próximos de -1 indicam forte associação negativa; e valores próximos de 0 indicam ausência de relação linear. Para calcular a matriz de correlação, utilizaremos a função *cor()* do R, especificando o método “Pearson”.

Como nossa matriz X (quantitativa) já está organizada com as parcelas nas colunas, a função *cor*() calculará a correlação entre as parcelas diretamente, sem a necessidade de transposição.

```
# Calculamos a matriz de correlação de Pearson entre as parcelas
# A função cor() compara as colunas da matriz por padrão
matriz_corr_pearson <- cor(
  X, # Usamos a matriz X diretamente, pois as parcelas estão nas colunas
  method = "pearson"
)

# Exibimos a matriz de correlação, arredondada para duas casas decimais
print(round(matriz_corr_pearson, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.02 | 0.59 | 0.15 | 0.33 | 0.48 | -0.20 | 0.38 | 0.37 | 0.25 | 0.13 |
| P2 | 0.02 | 1.00 | -0.03 | 0.53 | 0.35 | 0.22 | 0.07 | -0.01 | 0.08 | -0.11 | -0.05 |
| P3 | 0.59 | -0.03 | 1.00 | -0.08 | 0.10 | 0.35 | -0.04 | 0.18 | 0.64 | 0.76 | 0.47 |
| P4 | 0.15 | 0.53 | -0.08 | 1.00 | 0.79 | 0.64 | -0.03 | 0.45 | -0.01 | 0.04 | 0.06 |
| P5 | 0.33 | 0.35 | 0.10 | 0.79 | 1.00 | 0.57 | -0.05 | 0.27 | 0.07 | 0.24 | 0.21 |
| P6 | 0.48 | 0.22 | 0.35 | 0.64 | 0.57 | 1.00 | -0.13 | 0.37 | 0.26 | 0.32 | 0.25 |
| P7 | -0.20 | 0.07 | -0.04 | -0.03 | -0.05 | -0.13 | 1.00 | -0.06 | 0.56 | -0.12 | 0.76 |
| P8 | 0.38 | -0.01 | 0.18 | 0.45 | 0.27 | 0.37 | -0.06 | 1.00 | 0.23 | 0.12 | 0.18 |
| P9 | 0.37 | 0.08 | 0.64 | -0.01 | 0.07 | 0.26 | 0.56 | 0.23 | 1.00 | 0.50 | 0.76 |
| P10 | 0.25 | -0.11 | 0.76 | 0.04 | 0.24 | 0.32 | -0.12 | 0.12 | 0.50 | 1.00 | 0.41 |
| P11 | 0.13 | -0.05 | 0.47 | 0.06 | 0.21 | 0.25 | 0.76 | 0.18 | 0.76 | 0.41 | 1.00 |

### 3. Percentagem de similaridade de Czekanowiski

A percentagem de similaridade de Czekanowiski (**PS**) é calculada conforme a expressão a seguir:

$$PS_{(1,2)} = \frac{2[\sum_{i=1}^{M} min(X_{i1}, X_{i2})]}{\sum_{i=1}^{N}(X_{i1} + X_{i2})}$$

Em que: $\mathbf{X_{i1}}$ e $\mathbf{X_{i2}}$ são, respectivamente, as quantidades de cada espécie nas amostras 1 e 2; min ($\mathbf{X_{i1}}$, $\mathbf{X_{i2}}$) corresponde ao valor mínimo das quantidades de cada espécie que é comum a ambas as parcelas; e as somas dos quocientes compreendem todas as **i** de **1** a **N**. Se $\mathbf{x_{i1}}$ e $\mathbf{x_{i2}}$ se expressam como proporções do total, isto é, são padronizados pelo total das parcelas, a percentagem de similaridade simplifica-se a:

$$PS_{(1,2)} = min \sum_{i=1}^{N} (Y_{i1}, Y_{i2})$$

Em que: $Y_{i1} = \frac{X_{i1}}{\sum_{i=1}^{N} X_{i1}}$ e $Y_{i2} = \frac{X_{i2}}{\sum_{i=1}^{N} X_{i2}}$

**Aplicação**

Considerando as Parcelas 1 e 2 (Tabela 3, página 17), temos:

$$X'_{1j} = [8\ 0\ 3\ 6\ 00\ 0\ 0\ 1\ 2\ 1\ 0\ 5\ 0\ 7\ 0\ 0\ 0]$$

$$X'_{2j} = [1\ 0\ 9\ 3\ 12\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

Caso 1:

$$PS_{(1,2)} = \frac{2[\sum_{i=1}^{M} min(X_{i1}, X_{i2})]}{\sum_{i=1}^{N}(X_{i1} + X_{i2})}$$

$$PS_{(1,2)} = \frac{2[1+0+3+3+0+0+0+1+0+0+0+0+0+0+0+0+0]}{(8+1)+(0+0)+\cdots+(0+7)+(0+0)+(0+0)+(0+0)} = 0{,}27$$

Caso 2:

$$PS_{(1,2)} = min \sum_{i=1}^{N} (Y_{i1}, Y_{i2})$$

Em que:

$Y_{11} = \frac{X_{11}}{\sum_{i=1}^{N} X_{11}} = \frac{8}{33} = 0{,}24 \ldots y_{171} = \frac{X_{171}}{\sum_{i=1}^{N} X_{171}} = \frac{0}{33} = 0$

$Y_{12} = \frac{X_{12}}{\sum_{i=1}^{N} X_{12}} = \frac{1}{26} = 0{,}04 \ldots y_{172} = \frac{X_{172}}{\sum_{i=1}^{N} X_{172}} = \frac{0}{26} = 0$

Assim:

$Y'_{i2} = [0{,}04\ 0{,}00\ 0{,}35\ 0{,}11\ 0{,}46\ 0{,}00\ 0{,}00\ 0{,}04\ 0{,}00\ 0{,}00\ 0{,}00\ 0{,}00\ 0{,}00\ 0{,}00\ 0{,}00\ 0{,}00\ 0{,}00]$

$PS_{(1,2)} = (0{,}04 + 0 + 0{,}09 + 0{,}11 + 0 + 0 + 0{,}03 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) = 0{,}27$

### 3.1. Passo a passo no R: similaridade de Czekanowski

#### a) Calculando a Matriz de Dissimilaridade de Bray-Curtis

A Dissimilaridade de Bray-Curtis é uma medida muito usada para comparar parcelas com base na abundância das espécies. Ela nos mostra o quão diferentes duas áreas são, considerando a quantidade de cada espécie. Os valores variam de 0 (parcelas idênticas) a 1 (parcelas completamente diferentes).

Para calcular essa medida, utilizaremos a função *vegdist()* do pacote *vegan* com o argumento *method = "bray"*. Como a nossa matriz X (quantitativa) tem as parcelas nas colunas, precisamos "virar" a matriz (t(X)) para que a função *vegdist()* compare as parcelas corretamente.

```
# Carregamos o pacote para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(vegan)

# Calculamos a matriz de dissimilaridade de Bray-Curtis
# Transpomos a matriz X (t(X)) para que as parcelas fiquem nas linhas e sejam comparadas
matriz_dist_bray <- vegdist(
  t(X),
  method = "bray"
)

# Exibimos a matriz de dissimilaridade, arredondada para duas casas decimais
print(round(as.matrix(matriz_dist_bray), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.73 | 0.52 | 0.58 | 0.68 | 0.51 | 0.81 | 0.53 | 0.66 | 0.58 | 0.52 |
| P2 | 0.73 | 0.00 | 0.80 | 0.52 | 0.73 | 0.65 | 0.75 | 0.75 | 0.70 | 0.80 | 0.74 |
| P3 | 0.52 | 0.80 | 0.00 | 0.71 | 0.62 | 0.61 | 0.81 | 0.74 | 0.48 | 0.37 | 0.53 |
| P4 | 0.58 | 0.52 | 0.71 | 0.00 | 0.69 | 0.56 | 0.69 | 0.52 | 0.65 | 0.61 | 0.55 |
| P5 | 0.68 | 0.73 | 0.62 | 0.69 | 0.00 | 0.57 | 0.87 | 0.82 | 0.75 | 0.56 | 0.70 |
| P6 | 0.51 | 0.65 | 0.61 | 0.56 | 0.57 | 0.00 | 0.81 | 0.60 | 0.63 | 0.57 | 0.56 |
| P7 | 0.81 | 0.75 | 0.81 | 0.69 | 0.87 | 0.81 | 0.00 | 0.77 | 0.52 | 0.76 | 0.34 |
| P8 | 0.53 | 0.75 | 0.74 | 0.52 | 0.82 | 0.60 | 0.77 | 0.00 | 0.71 | 0.64 | 0.59 |
| P9 | 0.66 | 0.70 | 0.48 | 0.65 | 0.75 | 0.63 | 0.52 | 0.71 | 0.00 | 0.49 | 0.47 |
| P10 | 0.58 | 0.80 | 0.37 | 0.61 | 0.56 | 0.57 | 0.76 | 0.64 | 0.49 | 0.00 | 0.44 |
| P11 | 0.52 | 0.74 | 0.53 | 0.55 | 0.70 | 0.56 | 0.34 | 0.59 | 0.47 | 0.44 | 0.00 |

**b) Calculando a Matriz de Similaridade de Czekanowski**

A Similaridade de Czekanowski é o complemento da Dissimilaridade de Bray-Curtis. Isso significa que, para obter a similaridade (o quão parecidas as parcelas são), basta subtrair os valores da dissimilaridade de 1. Os valores variam de 0 (parcelas completamente diferentes) a 1 (parcelas idênticas).

```
# Transformamos a dissimilaridade de Bray-Curtis em similaridade de Czekanowski
matriz_sim_czekanowski <- 1 - as.matrix(matriz_dist_bray)

# Exibimos a matriz de similaridade, arredondada para duas casas decimais
print(round(matriz_sim_czekanowski, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.27 | 0.48 | 0.42 | 0.32 | 0.49 | 0.19 | 0.47 | 0.34 | 0.42 | 0.48 |
| P2 | 0.27 | 1.00 | 0.20 | 0.48 | 0.27 | 0.35 | 0.25 | 0.25 | 0.30 | 0.20 | 0.26 |
| P3 | 0.48 | 0.20 | 1.00 | 0.29 | 0.38 | 0.39 | 0.19 | 0.26 | 0.52 | 0.63 | 0.47 |
| P4 | 0.42 | 0.48 | 0.29 | 1.00 | 0.31 | 0.44 | 0.31 | 0.48 | 0.35 | 0.39 | 0.45 |
| P5 | 0.32 | 0.27 | 0.38 | 0.31 | 1.00 | 0.43 | 0.13 | 0.18 | 0.25 | 0.44 | 0.30 |
| P6 | 0.49 | 0.35 | 0.39 | 0.44 | 0.43 | 1.00 | 0.19 | 0.40 | 0.37 | 0.43 | 0.44 |
| P7 | 0.19 | 0.25 | 0.19 | 0.31 | 0.13 | 0.19 | 1.00 | 0.23 | 0.48 | 0.24 | 0.66 |
| P8 | 0.47 | 0.25 | 0.26 | 0.48 | 0.18 | 0.40 | 0.23 | 1.00 | 0.29 | 0.36 | 0.41 |
| P9 | 0.34 | 0.30 | 0.52 | 0.35 | 0.25 | 0.37 | 0.48 | 0.29 | 1.00 | 0.51 | 0.53 |
| P10 | 0.42 | 0.20 | 0.63 | 0.39 | 0.44 | 0.43 | 0.24 | 0.36 | 0.51 | 1.00 | 0.56 |
| P11 | 0.48 | 0.26 | 0.47 | 0.45 | 0.30 | 0.44 | 0.66 | 0.41 | 0.53 | 0.56 | 1.00 |

## 4. Coeficiente de similaridade de Gower

O coeficiente de similaridade de Gower (**CG**) se baseia na proporção da variação em relação à maior discrepância possível, dada conforme a seguir:

$$CG_{(A,B)} = -log_{10}\left[1 - \frac{1}{p}\sum_{i=1}^{p}\left(\frac{|X_i(A) - X_i(B)|}{X_{imax} - X_{imin}}\right)\right]$$

Esse é um coeficiente de fácil utilização para dados quantitativos ou mesmo qualitativos (Gama, 2022). Mais detalhes podem ser encontrados em Gower (1971).

### Aplicação

Consideremos os vetores coluna (parcelas 1 e 2) transpostos $X'_{1j}$ e $X'_{2j}$da matriz X:

$$X'_{1j} = [8\ 0\ 3\ 6\ 00\ 0\ 0\ 1\ 2\ 1\ 0\ 5\ 0\ 7\ 0\ 0\ 0]$$
$$X'_{2j} = [1\ 0\ 9\ 3\ 12\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

Logo:

$$\left|X'_{1j} - X'_{2j}\right| = [7\ 0\ 6\ 3\ 12\ 0\ 0\ 0\ 2\ 1\ 0\ 5\ 0\ 7\ 0\ 0\ 0]$$

A máxima discrepância é determinada em relação aos dados referentes aos vetores linha (espécie), assim, para Espécie 1, teremos:

$$X_{i1} = [8\ 1\ 27\ 0\ 1\ 9\ 2\ 3\ 22\ 15\ 7]$$

Assim:

$$X_{1max} - X_{1min} = 27 - 0 = 27$$

Seguindo o mesmo raciocínio, para as demais espécies, teremos:

$X_{2max} - X_{2min} = 17 - 0 = 17$; $X_{3max} - X_{3min} = 22 - 2 = 20$; $X_{4max} - X_{4min} = 29 - 0 = 29$; $X_{5max} - X_{5min} = 12 - 0 = 12$; $X_{6max} - X_{6min} = 11 - 0 = 11$; $X_{7max} - X_{7min} = 11 - 0 = 11$; $X_{8max} - X_{8min} = 13 - 0 = 13$; $X_{9max} - X_{9min} = 6 - 0 = 6$; $X_{10max} - X_{10min} = 6 - 0 = 6$; $X_{11max} - X_{11min} = 6 - 0 = 6$; $X_{12max} - X_{12min} = 7 - 0 = 7$; $X_{13max} - X_{13min} = 5 - 0 = 0$; $X_{14max} - X_{14min} = 7 - 0 = 7$; $X_{15max} - X_{15min} = 2 - 0 = 2$; $X_{16max} - X_{16min} = 1 - 0 = 1$; $X_{17max} - X_{17min} = 1 - 0 = 1$

Dessa forma:

$$CG_{(1,2)} = -log_{10}\left[1 - \frac{1}{17}\left(\frac{7}{27} + \frac{0}{17} + \frac{6}{20} + \frac{3}{29} + \frac{12}{12} + \frac{0}{11} + \frac{0}{11} + \frac{0}{13} + \frac{2}{6} + \frac{1}{6} + \frac{0}{6} + \frac{5}{7} + \frac{0}{5} + \frac{7}{7} + \frac{0}{2} + \frac{0}{1} + \frac{0}{1}\right)\right]$$

$$CG_{(1,2)} = 0{,}11$$

## 4.1. Passo a passo no R: Coeficiente de similaridade de Gower

### a) Calculando a Matriz de Dissimilaridade de Gower

A Distância de Gower é uma medida de dissimilaridade utilizada para comparar parcelas com base na abundância das espécies. Ela quantifica o grau de diferença entre cada par de parcelas. Para realizar esse cálculo, utilizaremos a função *daisy*() do pacote *cluster*.

Como nossa matriz X (quantitativa) tem as parcelas nas colunas e a função *daisy*() espera que as observações (parcelas) estejam nas linhas, precisamos transpor a matriz X (*t*(X)) antes de aplicar a função.

```
# Carregamos o pacote para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(cluster)

# Calculamos a matriz de dissimilaridade de Gower
# Transpomos a matriz X (t(X)) para que as parcelas fiquem nas linhas e sejam comparadas
dist_gower <- daisy(
  t(X),
  metric = "gower"
)

# Exibimos a matriz de dissimilaridade de Gower, arredondada para duas casas decimais
print(round(as.matrix(dist_gower), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.23 | 0.27 | 0.16 | 0.28 | 0.24 | 0.35 | 0.20 | 0.29 | 0.38 | 0.28 |
| P2 | 0.23 | 0.00 | 0.34 | 0.12 | 0.31 | 0.23 | 0.28 | 0.21 | 0.22 | 0.41 | 0.28 |
| P3 | 0.27 | 0.34 | 0.00 | 0.28 | 0.30 | 0.35 | 0.44 | 0.35 | 0.34 | 0.35 | 0.33 |
| P4 | 0.16 | 0.12 | 0.28 | 0.00 | 0.26 | 0.17 | 0.24 | 0.12 | 0.18 | 0.29 | 0.19 |
| P5 | 0.28 | 0.31 | 0.30 | 0.26 | 0.00 | 0.31 | 0.47 | 0.36 | 0.37 | 0.35 | 0.37 |
| P6 | 0.24 | 0.23 | 0.35 | 0.17 | 0.31 | 0.00 | 0.39 | 0.20 | 0.31 | 0.40 | 0.30 |
| P7 | 0.35 | 0.28 | 0.44 | 0.24 | 0.47 | 0.39 | 0.00 | 0.32 | 0.25 | 0.47 | 0.19 |
| P8 | 0.20 | 0.21 | 0.35 | 0.12 | 0.36 | 0.20 | 0.32 | 0.00 | 0.25 | 0.36 | 0.27 |
| P9 | 0.29 | 0.22 | 0.34 | 0.18 | 0.37 | 0.31 | 0.25 | 0.25 | 0.00 | 0.36 | 0.25 |
| P10 | 0.38 | 0.41 | 0.35 | 0.29 | 0.35 | 0.40 | 0.47 | 0.36 | 0.36 | 0.00 | 0.28 |
| P11 | 0.28 | 0.28 | 0.33 | 0.19 | 0.37 | 0.30 | 0.19 | 0.27 | 0.25 | 0.28 | 0.00 |

### b) Aplicando a transformação logarítmica

Após o cálculo da dissimilaridade de Gower, aplicaremos uma transformação logarítmica para aumentar a sensibilidade a diferenças elevadas. Essa transformação é dada pela fórmula: $D_{ij} = -log10(1 - d_{ij})$, onde $d_{ij}$ é a dissimilaridade de Gower entre as parcelas i e j.

Para evitar problemas com log10(0) caso 1 - $d_{ij}$ seja zero, usamos *pmax*(1 - *as.matrix(dist_gower)*, 1e-10), que garante que o valor mínimo dentro do logaritmo seja um número muito pequeno (1e-10), e não zero.

```
# Aplicamos a transformação logarítmica na matriz de dissimilaridade de Gower
# pmax(..., 1e-10) garante que não teremos logaritmo de zero
matriz_dist_gower_log <- -log10(
  pmax(1 - as.matrix(dist_gower), 1e-10)
)

# Exibimos a matriz de distâncias transformadas, arredondada para duas casas decimais
print(round(matriz_dist_gower_log, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.11 | 0.14 | 0.08 | 0.14 | 0.12 | 0.19 | 0.10 | 0.15 | 0.21 | 0.14 |
| P2 | 0.11 | 0.00 | 0.18 | 0.06 | 0.16 | 0.11 | 0.14 | 0.10 | 0.11 | 0.23 | 0.14 |
| P3 | 0.14 | 0.18 | 0.00 | 0.14 | 0.16 | 0.19 | 0.25 | 0.19 | 0.18 | 0.19 | 0.17 |
| P4 | 0.08 | 0.06 | 0.14 | 0.00 | 0.13 | 0.08 | 0.12 | 0.06 | 0.09 | 0.15 | 0.09 |
| P5 | 0.14 | 0.16 | 0.16 | 0.13 | 0.00 | 0.16 | 0.27 | 0.19 | 0.20 | 0.19 | 0.20 |
| P6 | 0.12 | 0.11 | 0.19 | 0.08 | 0.16 | 0.00 | 0.21 | 0.10 | 0.16 | 0.22 | 0.15 |
| P7 | 0.19 | 0.14 | 0.25 | 0.12 | 0.27 | 0.21 | 0.00 | 0.16 | 0.13 | 0.27 | 0.09 |
| P8 | 0.10 | 0.10 | 0.19 | 0.06 | 0.19 | 0.10 | 0.16 | 0.00 | 0.13 | 0.19 | 0.14 |
| P9 | 0.15 | 0.11 | 0.18 | 0.09 | 0.20 | 0.16 | 0.13 | 0.13 | 0.00 | 0.19 | 0.13 |
| P10 | 0.21 | 0.23 | 0.19 | 0.15 | 0.19 | 0.22 | 0.27 | 0.19 | 0.19 | 0.00 | 0.14 |
| P11 | 0.14 | 0.14 | 0.17 | 0.09 | 0.20 | 0.15 | 0.09 | 0.14 | 0.13 | 0.14 | 0.00 |

## 1.3.2. Medidas de Distância

O tipo de função de agrupamento, ou de semelhança, de emprego mais corrente na análise de agrupamento é aquele representado por medidas de distância. Essa função permite verificar se dois elementos, $\mathbf{E_i}$ e $\mathbf{E_j}$, podem ser alocados a um mesmo grupo, de acordo com um critério especificado (Gama, 2022).

A alocação é obtida mediante o cálculo de uma medida de distância entre os pontos do espaço **p**-dimensional. O termo espaço métrico é aplicado a um conjunto de pontos, para os quais se determinam distâncias mútuas, que atendem a certas regras.

Qualquer medida de distância, $\mathbf{d_{(P,Q)}}$, entre dois pontos **P** e Q é válida se os seguintes axiomas são satisfeitos, sendo **R** um ponto intermediário (Johnson; Wichern, 2007; Mardia et al., 2023):

1. $\mathbf{d_{(P,Q)}} = \mathbf{d_{(Q,P)}}$ (simetria);
2. $\mathbf{d_{(P,Q)}} > 0$, se $\mathbf{P} \neq \mathbf{Q}$ (positividade);
3. $\mathbf{d_{(P,Q)}} = 0$ se e somente se $\mathbf{P} = \mathbf{Q}$; e
4. $\mathbf{d_{(P,Q)}} \leq \mathbf{d_{(P,R)}} + \mathbf{d_{(Q,R)}}$ (desigualdade triangular).

Além disso, é esperado que $\mathbf{d_{(P,Q)}}$ aumente quando a dissimilaridade entre **P** e **Q** aumentar.

Conforme o axioma da desigualdade triangular, a distância entre duas parcelas ou unidades de amostras não pode ser maior que a soma de suas distâncias individuais a uma terceira parcela.

Vários pesquisadores têm usado funções de distâncias que não se enquadram na categoria de métrica e verificaram que, se **j** e **k** são definidas como duas unidades amostrais, $f(\mathbf{j}, \mathbf{k})$ é uma medida abstrata que depende da composição de espécies nas parcelas **j** e **k**, mas não usa uma distância no terreno (Orlóci, 1978).

Em termos de análise de vegetação, os axiomas acima são definidos do seguinte modo:

1. Se $\mathbf{A} = \mathbf{B}$, então $\mathbf{d_{(A,B)}} = 0$;
2. Se $\mathbf{A} \neq \mathbf{B}$, então $\mathbf{d_{(A,B)}} > 0$;
3. $\mathbf{d_{(A,B)}} = \mathbf{d_{(B,A)}}$; e
4. $\mathbf{d_{(A,B)}} \leq \mathbf{d_{(A,C)}} + \mathbf{d_{(B,C)}}$ (desigualdade triangular).

Para quaisquer **A**, **B** e **C** no espaço multidimensional ($\mathbf{I_p}$).

O axioma 1 indica que parcelas com idêntica composição florística são indistinguíveis; o axioma 2 indica que esta distância é positiva quando $\mathbf{A} \neq \mathbf{B}$; o axioma 3 indica que a ordem na qual as parcelas são comparadas não tem nenhuma importância; e o axioma 4 é a desigualdade triangular (Orlóci, 1978).

A função de distância que satisfaz a todos os axiomas do espaço métrico é dita ser uma métrica. No entanto, a função que satisfaz somente os três primeiros axiomas é dita ser uma semimétrica.

## 1.3.2.1. Algumas Medidas de Distâncias

### a. Distância euclidiana

A expressão mais familiar de distância euclidiana ($\mathbf{d_{j,k}}$) é:

$$d_{j,k} = \sqrt{\sum_{h=1}^{n} \left(X_{hj} - X_{hk}\right)^2}$$

Em que: **p** é o número de parcelas na amostra; e $\mathbf{X_{hj}}$ e $\mathbf{X_{hk}}$ são, respectivamente, as quantidades da **h**-ésima espécie (**h** = 1, 2, .., n) nas parcelas **j** e **k**. Vale ressaltar que a distância semelhante pode ser formulada para espécies.

### Aplicação

Com base na Tabela 3 (página 17), exemplificamos os cálculos da distância euclidiana, considerando as Parcelas 1 e 2, 1 e 3 e 10 e 11, como segue:

$d_{1,2} = [(8-1)^2 + (0-0)^2 + (3-9)^2 + (6-3)^2 + (0-12)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (2-0)^2 + (1-0)^2 + (0-0)^2 + (5-0)^2 + (0-0)^2 + (7-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2]^{1/2} = 17{,}8$

$d_{1,3} = [(8-27)^2 + (0-0)^2 + (3-4)^2 + (6-3)^2 + (0-0)^2 + (0-1)^2 + (0-10)^2 + (1-0)^2 + (2-2)^2 + (1-0)^2 + (0-0)^2 + (5-7)^2 + (0-1)^2 + (7-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2]^{1/2} = 23{,}0$

**...**

$d_{10,11} = [(15-7)^2 + (1-9)^2 + (4-4)^2 + (4-4)^2 + (0-0)^2 + (11-1)^2 + (11-5)^2 + (3-1)^2 + (2-2)^2 + (3-1)^2 + (0-5)^2 + (1-0)^2 + (5-2)^2 + (0-0)^2 + (2-1)^2 + (0-0)^2 + (0-0)^2]^{1/2} = 17{,}5$

Assim, obtém-se uma matriz de distância euclidiana **D**:

$$\mathbf{D} = \begin{bmatrix} 0{,}00 & 17{,}8 & 23{,}0 & 12{,}5 & 33{,}9 & 17{,}3 & 18{,}5 & 11{,}3 & 24{,}8 & 20{,}0 & 14{,}7 \\ & 0{,}00 & 31{,}7 & 12{,}0 & 34{,}4 & 21{,}8 & 18{,}3 & 16{,}2 & 29{,}4 & 25{,}6 & 18{,}7 \\ & & 0{,}00 & 29{,}6 & 41{,}8 & 27{,}0 & 30{,}9 & 27{,}9 & 22{,}4 & 17{,}9 & 24{,}2 \\ & & & 0{,}00 & 32{,}3 & 17{,}3 & 14{,}7 & 7{,}7 & 28{,}3 & 21{,}5 & 13{,}5 \\ & & & & 0{,}00 & 28{,}7 & 38{,}6 & 35{,}6 & 41{,}4 & 34{,}8 & 34{,}7 \\ & & & & & 0{,}00 & 24{,}6 & 19{,}0 & 27{,}6 & 21{,}5 & 19{,}4 \\ & & & & & & 0{,}00 & 15{,}5 & 22{,}4 & 24{,}4 & 8{,}8 \\ & & & & & & & 0{,}00 & 26{,}7 & 21{,}2 & 13{,}3 \\ & & & & & & & & 0{,}00 & 22{,}8 & 19{,}2 \\ & & & & & & & & & 0{,}00 & 17{,}5 \\ & & & & & & & & & & 0{,}00 \end{bmatrix}$$

A distância euclidiana é a métrica de maior utilização como função (medida) de agrupamento e apresenta grande facilidade de cálculo. Ela define a distância entre duas parcelas como uma simples soma de **p** diferenças ou desvios ao quadrado. Geralmente, essa métrica é utilizada para variáveis padronizadas.

Pode-se observar que a fórmula de distância euclidiana não tem um limite superior fixo e, então, pode produzir distâncias incomparáveis. Por isso, é normalmente preferível trabalhar com unidades relativas para as quais se pode decidir se um dado valor de distância é maior ou menor que outro. O tamanho de **d(j,k)** depende da magnitude das diferenças reais das quantidades de espécies entre as parcelas. Portanto, em determinadas situações, duas parcelas podem parecer mais semelhantes que duas outras que possuem espécies em comum, como:

$$X'_{8i} = [03\ \ 1\ \ 2\ \ 4\ \ 0\ \ 00\ \ 00\ \ 0\ \ 5\ \ 6\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0]$$

$$X'_{10i} = [15\ \ 1\ \ 4\ \ 4\ \ 0\ \ 11\ \ 11\ \ 3\ \ 2\ \ 3\ \ 0\ \ 1\ \ 5\ \ 0\ \ 2\ \ 0\ \ 0]$$

$$X'_{11i} = [07\ \ 9\ \ 4\ \ 4\ \ 0\ \ 01\ \ 05\ \ 1\ \ 2\ \ 1\ \ 5\ \ 0\ \ 2\ \ 0\ \ 1\ \ 0\ \ 0]$$

Em termos de distância, a parcela 8 é mais semelhante à parcela 11, posto que $d_{(8,\,11)}$ = 13,27 e $d_{(10,11)}$ = 17,55.

Esse problema é contornado calculando-se as distâncias relativas do seguinte modo:

1. Normalizar o vetor parcela. Tomando a parcela 8, como exemplo, tem-se:

$$\|X'_{8i}\| = \sqrt{3^2 + 1^2 + 2^2 + 4^2 + 0^2 + 0^2 + 0^2 + 0^2 + 5^2 + 6^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2}$$

$\|X'_{8i}\| = 2\sqrt{23}$, que representa a norma do vetor.

Assim, vetor normalizado será $X'_{8i*} = \frac{X'_{8i}}{2\sqrt{23}}$, ou seja:

$X'_{8i*} = [3/2\sqrt{23}\ \ 1/2\sqrt{23}\ \ 1/\sqrt{23}\ \ 2/\sqrt{23}\ \ 0\ \ 0\ \ 0\ \ 0\ \ 5/2\sqrt{23}\ \ 3/\sqrt{23}\ \ 0\ \ 0\ \ 0\ \ 1/2\sqrt{23}\ \ 0\ \ 0\ \ 0]$

2. Calcular a distância euclidiana com base nos vetores normalizados. Segundo Orlóci (1978), a distância assim calculada é denominada de **distância corda** e pode ser diretamente obtida pela seguinte fórmula:

$$d_{j,k} = \sqrt{2\left(1 - \frac{q_{jk}}{\sqrt{q_{jj}q_{kk}}}\right)}$$

Em que: $\boldsymbol{q_{jk} = \sum_{h=1}^{p} X_{hj}X_{hk}}$; $\boldsymbol{q_{jj} = \sum_{h=1}^{p} X_{hj}^2}$; e $\boldsymbol{q_{kk} = \sum_{h=1}^{p} X_{hk}^2}$; **h** = 1, 2, ... , **p** espécies na amostra.

A distância corda é caracterizada pelas seguintes propriedades:

a. $0 \leq \mathbf{d_{j,k}} \leq \sqrt{2}$;

b. $\mathbf{d_{j,k}} = 0$, se $\mathbf{X_{hj}/X_{ij}} = \mathbf{X_{hk}/X_{ik}}$; e

c. $\mathbf{d_{j,k}} = \sqrt{2}$, se $X_{hj} = 0$, isso implica que, se $\mathbf{X_{hk}} = 0$, logo $\mathbf{X_{hj}} > 0$. Tudo isso indica que, para a distância ser zero, as duas parcelas não necessitam ter as mesmas espécies em iguais quantidades. É suficiente que as quantidades estejam nas mesmas proporções.

**Aplicação**

Considerando as Parcelas 1 e 2 (Tabela 3, página 17), teremos:

$q_{12} = \sum_{h=1}^{17} X_{h1}X_{h2} = 8x1 + 0x0 + 3x9 + 6x3 + 0x12 + 0x0 + 0x0 + 1x1 + 2x0 + 1x0 + 0x0 + 7x0 + 0x0 + 0x0 + 0x0 = 54$

$q_{11} = \sum_{h=1}^{17} X_{h1}^2 = 8^2 + 0^2 + 3^2 + 6^2 + 0^2 + 0^2 + 0^2 + 1^2 + 2^2 + 1^2 + 0^2 + 5^2 + 0^2 + 7^2 + 0^2 + 0^2 + 0^2 = 189$

$q_{22} = \sum_{h=1}^{17} X_{h2}^2 = 1^2 + 0^2 + 9^2 + 3^2 + 12^2 + 0^2 + 0^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 = 236$

$$d_{j,k} = \sqrt{2\left(1 - \frac{q_{jk}}{\sqrt{q_{jj}q_{kk}}}\right)} \quad \Rightarrow \quad d_{1,2} = \sqrt{2\left(1 - \frac{54}{\sqrt{189x236}}\right)} = 1{,}22$$

Foi verificado anteriormente que, em termos de distância euclidiana, a parcela 8 é mais semelhante à parcela 11, posto que $d_{(8,\ 11)} = 13{,}3$ e $d_{(10,11)} = 17{,}5$. Comparando esse resultado com a distância corda, observa-se que a parcela 10 é mais semelhante à 11, pois $d_{(8,\ 11)} = 1{,}02$ e $d_{(10,11)} = 0{,}70$, o que confirma que a distância euclidiana é alterada com a mudança da escala de medições.

Outra forma de padronização de variáveis é, por exemplo, fazendo $\mathbf{W_h = 1/S_h}$, em que $\mathbf{S_h}$ é o desvio-padrão da **h**-ésima variável. Então, a distância euclidiana padronizada será:

$$d_{j,k} = \sqrt{\sum_{h=1}^{p} W_h \left(X_{hj} - X_{hk}\right)^2}$$

Há outras formas de padronização de menor utilização, como a **padronização por amplitude:**

$$W_h = \left[\frac{1}{Max_{jk}\left(X_{hj} - X_{hk}\right)^2}\right] \qquad \text{e } W_h = \left[\frac{1}{Max_{jk}(X_{hj}) - Min_{jk}(X_{hk})}\right]$$

Em resumo, os inconvenientes apresentados pela distância euclidiana são o fato de ela ser alterada com a mudança da escala de medições e com o número de caracteres estudados e não levar em conta o grau de correlação entre eles. Assim, para contornar o problema de escala, tem sido recomendável a padronização dos dados, e, para contornar a influência do número de caracteres, utiliza-se a **distância euclidiana média padronizada (**Cruz et al., 2012), descrita a seguir:

Seja: $\boldsymbol{x_{ij}} = \frac{\boldsymbol{X_{ij}}}{S(X_j)}$, em que: $\mathbf{S(X_j)}$ é o desvio-padrão dos dados da **j**-ésima variável, então:

$$d_{j,k} = \sqrt{\frac{1}{p}\sum_{h=1}^{p} \left(x_{hj} - x_{hk}\right)^2}$$

Assim, $\mathbf{d_{jk}}$ é a distância euclidiana média com base em dados padronizados.

## Aplicação

Para calcular a distância euclidiana média padronizada, foram obtidos os resultados da Tabela 12.

**Tabela 12**. Médias e desvios padrão do número de indivíduos de 17 espécies da Mata da Silvicultura, em parcelas de 20 x 50 m, município de Viçosa-MG

| Parcela | Média | Desvio Padrão |
|---|---|---|
| P1 | 1,94 | 2,79 |
| P2 | 1,53 | 3,50 |
| P3 | 3,29 | 6,73 |
| P4 | 1,12 | 1,69 |
| P5 | 4,59 | 8,52 |
| P6 | 3,12 | 4,69 |
| P7 | 1,82 | 3,17 |
| P8 | 1,29 | 1,99 |
| P9 | 3,53 | 6,42 |
| P10 | 3,65 | 4,51 |
| P11 | 2,47 | 2,74 |

Assim, a padronização de $\mathbf{X_{i1}}$ e $\mathbf{X_{i2}}$ será:

Em relação às parcelas 1 e 2 (Tabela 3, página 17), temos os vetores coluna transpostos (parcelas):

$$X'_{1i} = [8\ 0\ 3\ 6\ 0\ 0\ 0\ 1\ 2\ 1\ 0\ 5\ 0\ 7\ 0\ 0\ 0]$$

$$X'_{2i} = [1\ 0\ 9\ 3\ 12\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

Dessa forma:

$$X'_{1i*} = \left[\frac{8}{2,79}\ \frac{0}{2,79}\ \frac{3}{2,79}\ \frac{6}{2,79}\ \frac{0}{2,79}\ \frac{0}{2,79}\ \frac{0}{2,79}\ \frac{1}{2,79}\ \frac{2}{2,79}\ \frac{1}{2,79}\ \frac{0}{2,79}\ \frac{5}{2,79}\ \frac{0}{2,79}\ \frac{7}{2,79}\ \frac{0}{2,79}\ \frac{0}{2,79}\ \frac{0}{2,79}\right]$$

$$X'_{2i*} = \left[\frac{1}{3,50}\ \frac{0}{3,50}\ \frac{9}{3,50}\ \frac{3}{3,50}\ \frac{12}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{1}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\right]$$

$$X'_{1i*} = [2,87\ 0,00\ 1,08\ 2,15\ 0,00\ 0,00\ 0,00\ 0,36\ 0,72\ 0,36\ 0,00\ 1,79\ 0,00\ 2,51\ 0,00\ 0,00\ 0,00]$$

E:

$$X'_{2i*} = [0,29\ 0,00\ 2,57\ 0,86\ 3,43\ 0,00\ 0,00\ 0,29\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00]$$

Logo:

$$d_{1,2} = \left(\frac{1}{17}[(2,87-0,29)^2 + (0-0)^2 + (1,08-2,57)^2 + (2,15-0,86)^2 + (0-3,43)^2 + (0-0)^2 + (0-0)^2 + (0,36-0,29)^2 + (0,72-0)^2 + (0,36-0)^2 + (0-0)^2 + (1,79-0)^2 + (0-0)^2 + (2,51-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2]\right)^{1/2} = 1,38$$

Em relação às parcelas 2 e 3 (Tabela 3, página 17), temos os vetores coluna transpostos (parcelas):

$$X'_{2i} = [1\ 0\ 9\ 3\ 12\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

$$X'_{3i} = [27\ 0\ 4\ 3\ 0\ 1\ 10\ 0\ 2\ 0\ 0\ 7\ 1\ 0\ 0\ 0\ 1]$$

Dessa forma:

$$X'_{2i*} = \left[\frac{1}{3,50}\ \frac{0}{3,50}\ \frac{9}{3,50}\ \frac{3}{3,50}\ \frac{12}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{1}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\ \frac{0}{3,50}\right]$$

$$X'_{3i*} = \left[\frac{27}{6,73}\ \frac{0}{6,73}\ \frac{4}{6,73}\ \frac{3}{6,73}\ \frac{0}{6,73}\ \frac{1}{6,73}\ \frac{10}{6,73}\ \frac{0}{6,73}\ \frac{2}{6,73}\ \frac{0}{6,73}\ \frac{0}{6,73}\ \frac{7}{6,73}\ \frac{1}{6,73}\ \frac{0}{6,73}\ \frac{0}{6,73}\ \frac{0}{6,73}\ \frac{1}{6,73}\right]$$

$$X'_{2i*} = [0,29\ 0,00\ 2,57\ 0,86\ 3,43\ 0,00\ 0,00\ 0,29\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00\ 0,00]$$

E:

$$X'_{3i*} = [4,01\ 0,00\ 0,59\ 0,45\ 0,00\ 0,15\ 1,49\ 0,00\ 0,30\ 0,00\ 0,00\ 1,04\ 0,15\ 0,00\ 0,00\ 0,00\ 0,15]$$

Logo:

$$d_{2,3} = \left(\frac{1}{17}[(0,29-4,01)^2 + (0-0)^2 + (2,57-0,59)^2 + (0,86-0,45)^2 + (3,43-0)^2 + (0-0,15)^2 + (0-1,49)^2 + (0,29-0)^2 + (0-0,30)^2 + (0-0)^2 + (0-0)^2 + (0-1,04)^2 + (0-0,15)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0,15)^2]\right)^{1/2} = 1,39$$

Finalmente, em relação às parcelas 10 e 11 (Tabela 3, página 17), temos os vetores coluna transpostos (parcelas):

$$X'_{10i} = [15\ 1\ 4\ 4\ 0\ 11\ 11\ 3\ 2\ 3\ 0\ 1\ 5\ 0\ 2\ 0\ 0]$$

$$X'_{11i} = [7\ 9\ 4\ 4\ 0\ 1\ 5\ 1\ 2\ 1\ 5\ 0\ 2\ 0\ 1\ 0\ 0]$$

Dessa forma:

$$X'_{10i*} = \left[\frac{15}{4,51}\ \frac{1}{4,51}\ \frac{4}{4,51}\ \frac{4}{4,51}\ \frac{0}{4,51}\ \frac{11}{4,51}\ \frac{11}{4,51}\ \frac{3}{4,51}\ \frac{2}{4,51}\ \frac{3}{4,51}\ \frac{0}{4,51}\ \frac{1}{4,51}\ \frac{5}{4,51}\ \frac{0}{4,51}\ \frac{2}{4,51}\ \frac{0}{4,51}\ \frac{0}{4,51}\right]$$

$$X'_{11i*} = \left[\frac{7}{2,74}\ \frac{9}{2,74}\ \frac{4}{2,74}\ \frac{4}{2,74}\ \frac{0}{2,74}\ \frac{1}{2,74}\ \frac{5}{2,74}\ \frac{1}{2,74}\ \frac{2}{2,74}\ \frac{1}{2,74}\ \frac{5}{2,74}\ \frac{0}{2,74}\ \frac{2}{2,74}\ \frac{0}{2,74}\ \frac{1}{2,74}\ \frac{0}{2,74}\ \frac{0}{2,74}\right]$$

$$X'_{10i*} = [3,32\ 0,22\ 0,89\ 0,89\ 0,00\ 2,44\ 2,44\ 0,66\ 0,44\ 0,66\ 0,44\ 0,00\ 0,22\ 1,11\ 0,44\ 0,00\ 0,00]$$

E:

$$X'_{11i*} = [2,55\ 3,28\ 1,46\ 1,46\ 0,00\ 0,36\ 1,82\ 0,36\ 0,73\ 0,36\ 1,82\ 0,00\ 0,73\ 0,00\ 0,36\ 0,00\ 0,00]$$

Logo:

$$d_{10,11} = \left(\frac{1}{17}[(3,32-2,55)^2 + (0,22-3,28)^2 + (0,89-1,46)^2 + (0,89-1,46)^2 + (0-0)^2 + (2,44-0,36)^2 + (2,44-1,82)^2 + (0,67-0,36)^2 + (0.44-0,73)^2 + (0,36-0,36)^2 + (0,44-1,82)^2 + (0-0)^2 + (0,22-0,73)^2 + (1,11-0)^2 + (0,44-0,36)^2 + (0-0)^2 + (0-0)^2]\right)^{1/2} = 1,06$$

A forma mais comum de apresentação da distância euclidiana é a matricial (Gama, 2022). Para isso, consideram-se os vetores $\mathbf{X_h}$ e $X_j$, representativos dos elementos $\mathbf{E_h}$ e $\mathbf{E_j}$, pontos do espaço **p**-dimensional. Então:

$\mathbf{d}_{\mathbf{j,k}} = (X_h - X_j).(X_h - X_j)'$.

Usando o desvio padrão das respectivas variáveis como fator de ponderação, a expressão padronizada é:

$\mathbf{d}_{\mathbf{(j,k)}} = (X_{h*} - X_{j*}).(X_{h*} - X_{j*})'$.

Para muitas variáveis aleatórias independentes, o quadrado da distância euclidiana entre dois elementos de **E** é a soma de variáveis independentes, podendo-se aplicar o teorema do limite central (Hartigan, 1975). O conjunto de distâncias obtidas é assintoticamente normal multivariada, portanto, as variâncias e as covariâncias podem ser calculadas com base na matriz de dados básicos.

## Passo a passo no R: Distância Euclidiana e suas variações

### 1. Distância Euclidiana

A Distância Euclidiana é a forma mais simples de medir a diferença entre duas parcelas, considerando a abundância de cada espécie. Ela calcula a "linha reta" entre os pontos que representam as parcelas em um espaço multidimensional. No R, a função *dist*() realiza esse cálculo.

Como nossa matriz X (quantitativa) tem as parcelas nas colunas e a função *dist*() espera que os itens a serem comparados (nossas parcelas) estejam nas linhas, precisamos transpor a matriz X (*t*(X)) antes de aplicar a função.

```
# Calculamos a Distância Euclidiana entre as parcelas
# Transpomos a matriz X (t(X)) para que as parcelas fiquem nas linhas e sejam comparadas
matriz_euclidiana <- as.matrix(dist(t(X)))

# Exibimos a matriz de distâncias, arredondada para uma casa decimal
print(round(matriz_euclidiana, 1))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.0 | 17.8 | 23.0 | 12.5 | 33.9 | 17.3 | 18.5 | 11.3 | 24.8 | 20.0 | 14.7 |
| P2 | 17.8 | 0.0 | 31.7 | 12.0 | 34.4 | 21.8 | 18.3 | 16.2 | 29.4 | 25.6 | 18.7 |
| P3 | 23.0 | 31.7 | 0.0 | 29.6 | 41.8 | 27.0 | 30.9 | 27.9 | 22.4 | 17.9 | 24.2 |
| P4 | 12.5 | 12.0 | 29.6 | 0.0 | 32.3 | 17.3 | 14.7 | 7.7 | 28.3 | 21.5 | 13.5 |
| P5 | 33.9 | 34.4 | 41.8 | 32.3 | 0.0 | 28.7 | 38.6 | 35.6 | 41.4 | 34.8 | 34.7 |
| P6 | 17.3 | 21.8 | 27.0 | 17.3 | 28.7 | 0.0 | 24.6 | 19.0 | 27.6 | 21.5 | 19.4 |
| P7 | 18.5 | 18.3 | 30.9 | 14.7 | 38.6 | 24.6 | 0.0 | 15.5 | 22.4 | 24.4 | 8.8 |
| P8 | 11.3 | 16.2 | 27.9 | 7.7 | 35.6 | 19.0 | 15.5 | 0.0 | 26.7 | 21.2 | 13.3 |
| P9 | 24.8 | 29.4 | 22.4 | 28.3 | 41.4 | 27.6 | 22.4 | 26.7 | 0.0 | 22.8 | 19.2 |
| P10 | 20.0 | 25.6 | 17.9 | 21.5 | 34.8 | 21.5 | 24.4 | 21.2 | 22.8 | 0.0 | 17.5 |
| P11 | 14.7 | 18.7 | 24.2 | 13.5 | 34.7 | 19.4 | 8.8 | 13.3 | 19.2 | 17.5 | 0.0 |

## 2. Distância Corda

A Distância Corda é uma variação da Euclidiana que "normaliza" os dados de abundância de cada parcela antes de calcular a distância. Isso significa que ela ajusta os valores para que a abundância total de espécies em cada parcela seja igual a 1. Essa normalização ajuda a comparar parcelas que podem ter diferentes tamanhos ou abundâncias totais, focando mais na composição relativa das espécies. O valor máximo que a Distância Corda pode ter é *sqrt*(2) (aproximadamente 1,4142).

Utilizaremos a função *decostand()* do pacote *vegan* para a normalização e, em seguida, a função *dist()* para calcular a distância Euclidiana.

```
# Carregamos o pacote 'vegan' para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(vegan)

# Normalizamos a matriz X. Transpomos X (t(X)) para que as parcelas fiquem nas linhas
# e a normalização seja aplicada por parcela.
X_norm <- decostand(t(X), method = "normalize")

# Calculamos a Distância Euclidiana entre as parcelas normalizadas
dist_corda <- dist(X_norm)

# Exibimos a matriz de distâncias, arredondada para duas casas decimais
print(round(as.matrix(dist_corda), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 1.22 | 0.79 | 1.05 | 0.98 | 0.84 | 1.30 | 0.91 | 0.95 | 0.97 | 1.02 |
| P2 | 1.22 | 0.00 | 1.30 | 0.85 | 1.02 | 1.10 | 1.22 | 1.25 | 1.21 | 1.28 | 1.23 |
| P3 | 0.79 | 1.30 | 0.00 | 1.26 | 1.19 | 0.99 | 1.27 | 1.11 | 0.75 | 0.63 | 0.89 |
| P4 | 1.05 | 0.85 | 1.26 | 0.00 | 0.54 | 0.68 | 1.20 | 0.85 | 1.20 | 1.09 | 1.06 |
| P5 | 0.98 | 1.02 | 1.19 | 0.54 | 0.00 | 0.79 | 1.26 | 1.04 | 1.19 | 1.03 | 1.04 |
| P6 | 0.84 | 1.10 | 0.99 | 0.68 | 0.79 | 0.00 | 1.27 | 0.93 | 1.03 | 0.93 | 0.96 |
| P7 | 1.30 | 1.22 | 1.27 | 1.20 | 1.26 | 1.27 | 0.00 | 1.23 | 0.81 | 1.23 | 0.59 |
| P8 | 0.91 | 1.25 | 1.11 | 0.85 | 1.04 | 0.93 | 1.23 | 0.00 | 1.06 | 1.06 | 1.01 |
| P9 | 0.95 | 1.21 | 0.75 | 1.20 | 1.19 | 1.03 | 0.81 | 1.06 | 0.00 | 0.83 | 0.60 |
| P10 | 0.97 | 1.28 | 0.63 | 1.09 | 1.03 | 0.93 | 1.23 | 1.06 | 0.83 | 0.00 | 0.82 |
| P11 | 1.02 | 1.23 | 0.89 | 1.06 | 1.04 | 0.96 | 0.59 | 1.01 | 0.60 | 0.82 | 0.00 |

### 3. Distância Euclidiana Média Padronizada

Essa é outra variação da Distância Euclidiana que padroniza os dados de forma diferente: cada valor de abundância é dividido pelo desvio padrão da própria parcela. Isso é útil para comparar parcelas que podem ter diferentes níveis de variação interna nas abundâncias das espécies, garantindo que espécies com maior variabilidade não dominem a medida de distância.

```
# Calculamos o desvio padrão para cada parcela (coluna da matriz X)
desvios <- apply(X, 2, sd)

# Evitamos divisão por zero, substituindo desvios padrão iguais a zero por 1
# (para parcelas sem variação, a padronização não terá efeito)
desvios[desvios == 0] <- 1

# Padronizamos a matriz X: cada abundância é dividida pelo desvio padrão da sua parcela
# A função sweep() aplica a divisão por coluna (MARGIN = 2)
X_pad <- sweep(X, 2, desvios, "/")

# Calculamos a Distância Euclidiana entre as parcelas padronizadas
# Transpomos X_pad (t(X_pad)) para que as parcelas fiquem nas linhas
# Dividimos por sqrt(nrow(X)) para obter a média padronizada
matriz_euclidiana_media <- as.matrix(dist(t(X_pad))) / sqrt(nrow(X))

# Exibimos a matriz de distâncias, arredondada para duas casas decimais
print(round(matriz_euclidiana_media, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 1.38 | 0.91 | 1.25 | 1.13 | 0.99 | 1.51 | 1.08 | 1.10 | 1.19 | 1.29 |
| P2 | 1.38 | 0.00 | 1.40 | 0.96 | 1.11 | 1.23 | 1.33 | 1.39 | 1.32 | 1.49 | 1.48 |
| P3 | 0.91 | 1.40 | 0.00 | 1.43 | 1.31 | 1.13 | 1.41 | 1.25 | 0.83 | 0.75 | 1.09 |
| P4 | 1.25 | 0.96 | 1.43 | 0.00 | 0.62 | 0.80 | 1.38 | 1.00 | 1.37 | 1.33 | 1.33 |
| P5 | 1.13 | 1.11 | 1.31 | 0.62 | 0.00 | 0.91 | 1.41 | 1.18 | 1.32 | 1.23 | 1.27 |
| P6 | 0.99 | 1.23 | 1.13 | 0.80 | 0.91 | 0.00 | 1.46 | 1.09 | 1.18 | 1.14 | 1.21 |
| P7 | 1.51 | 1.33 | 1.41 | 1.38 | 1.41 | 1.46 | 0.00 | 1.41 | 0.91 | 1.47 | 0.75 |
| P8 | 1.08 | 1.39 | 1.25 | 1.00 | 1.18 | 1.09 | 1.41 | 0.00 | 1.21 | 1.29 | 1.27 |
| P9 | 1.10 | 1.32 | 0.83 | 1.37 | 1.32 | 1.18 | 0.91 | 1.21 | 0.00 | 1.00 | 0.76 |
| P10 | 1.19 | 1.49 | 0.75 | 1.33 | 1.23 | 1.14 | 1.47 | 1.29 | 1.00 | 0.00 | 1.06 |
| P11 | 1.29 | 1.48 | 1.09 | 1.33 | 1.27 | 1.21 | 0.75 | 1.27 | 0.76 | 1.06 | 0.00 |

**b. Distância absoluta (Manhattan)**

A distância absoluta (**$d_{ij}$**) é outra métrica que utiliza a diferença entre os valores de uma variável, definida por:

$$d_{ij} = W_h \sum_{h=1}^{p} |X_{hi} - X_{hj}|$$

Em que: **$W_h$'s** representam as ponderações para as variáveis. Os valores mais usados são os da equiponderação **$W_h$ = 1** ou da média **$W_h$ = 1/p**. Essa medida é conhecida como a "métrica do quarteirão" (métrica *"city-block"*).

**Aplicação**

Utilizando $W_h = 1$, teremos:

$d_{12} = \sum_{h=1}^{p} |X_{h1} - X_{h2}| = [|8-1| + |0-0| + |3-9| + |6-3| + |0-12| + |0-0| + |0-0| + |1-1| + |2-0| + |1-0| + |0-0| + |5-0| + |0-0| + |7-0| + |0-0| + |0-0| + |0-0|] = 43$

Para $W_h = 1/p$, será:

$d_{12} = \frac{1}{17}\sum_{h=1}^{p} |X_{h1} - X_{h2}| = \frac{1}{17}[|8-1| + |0-0| + |3-9| + |6-3| + |0-12| + |0-0| + |0-0| + |1-1| + |2-0| + |1-0| + |0-0| + |5-0| + |0-0| + |7-0| + |0-0| + |0-0| + |0-0|] = 2{,}53$

E:

$d_{1011} = \sum_{h=1}^{p} |X_{hi10} - X_{h11}| = [|15-7| + |1-9| + |4-4| + |4-4| + |0-0| + |11-1| + |11-5| + |3-1| + |2-2| + |3-1| + |0-5| + |1-0| + |5-2| + |0-0| + |2-1| + |0-0| + |0-0|] = 46$

Para $W_h = 1/p$, será:

$d_{12} = \frac{1}{17}\sum_{h=1}^{p} |X_{h10} - X_{h11}| = \frac{1}{17}[|15-7| + |1-9| + |4-4| + |4-4| + |0-0| + |11-1| + |11-5| + |3-1| + |2-2| + |3-1| + |0-5| + |1-0| + |5-2| + |0-0| + |2-1| + |0-0| + |0-0|] = 2{,}71$

**Passo a passo no R: Distância Absoluta (Manhattan) e suas variações**

**1. Distância absoluta**

A Distância Absoluta, também conhecida como Distância de Manhattan, mede a diferença total de abundância entre duas parcelas. Ela soma todas as diferenças (sempre

positivas) de abundância para cada espécie. Quanto maior o valor, mais diferentes as parcelas são.

Para calcular essa medida, utilizaremos a função *vegdist()* do pacote *vegan* com o argumento *method = "manhattan"*. Como nossa matriz X (quantitativa) tem as parcelas nas colunas e a função *vegdist()* espera que as observações (parcelas) estejam nas linhas, precisamos transpor a matriz X (t(X)) antes de aplicar a função.

```
# Carregamos o pacote 'vegan' para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(vegan)

# Calculamos a Distância Absoluta (Manhattan) entre as parcelas
# Transpomos a matriz X (t(X)) para que as parcelas fiquem nas linhas e sejam comparadas
matriz_dist_absoluta <- vegdist(
  t(X),
  method = "manhattan"
)

# Exibimos a matriz de distâncias, arredondada para uma casa decimal
print(round(as.matrix(matriz_dist_absoluta), 1))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0 | 43 | 47 | 30 | 75 | 44 | 52 | 29 | 61 | 55 | 39 |
| P2 | 43 | 0 | 66 | 23 | 76 | 51 | 43 | 36 | 60 | 70 | 50 |
| P3 | 47 | 66 | 0 | 53 | 84 | 67 | 71 | 58 | 56 | 44 | 52 |
| P4 | 30 | 23 | 53 | 0 | 67 | 40 | 34 | 21 | 51 | 49 | 33 |
| P5 | 75 | 76 | 84 | 67 | 0 | 75 | 95 | 82 | 104 | 78 | 84 |
| P6 | 44 | 51 | 67 | 40 | 75 | 0 | 68 | 45 | 71 | 65 | 53 |
| P7 | 52 | 43 | 71 | 34 | 95 | 68 | 0 | 41 | 47 | 71 | 25 |
| P8 | 29 | 36 | 58 | 21 | 82 | 45 | 41 | 0 | 58 | 54 | 38 |
| P9 | 61 | 60 | 56 | 51 | 104 | 71 | 47 | 58 | 0 | 60 | 48 |
| P10 | 55 | 70 | 44 | 49 | 78 | 65 | 71 | 54 | 60 | 0 | 46 |
| P11 | 39 | 50 | 52 | 33 | 84 | 53 | 25 | 38 | 48 | 46 | 0 |

## 2. Distância absoluta média

A Distância Absoluta Média nos dá a diferença média por espécie entre as parcelas. Para obtê-la, simplesmente dividimos a Distância Absoluta total pelo número total de espécies (n_especies). Isso ajuda a comparar parcelas independentemente do número de espécies que elas possuem.

```
# 'n_especies' é o número de espécies (linhas na nossa matriz X original)
n_especies <- nrow(X)

# Calculamos a Distância Absoluta Média dividindo a matriz total pelo número de espécies
matriz_dist_absoluta_media <- matriz_dist_absoluta / n_especies
```

```
# Exibimos a matriz de distâncias médias, arredondada para duas casas decimais
print(round(as.matrix(matriz_dist_absoluta_media), 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 2.53 | 2.76 | 1.76 | 4.41 | 2.59 | 3.06 | 1.71 | 3.59 | 3.24 | 2.29 |
| P2 | 2.53 | 0.00 | 3.88 | 1.35 | 4.47 | 3.00 | 2.53 | 2.12 | 3.53 | 4.12 | 2.94 |
| P3 | 2.76 | 3.88 | 0.00 | 3.12 | 4.94 | 3.94 | 4.18 | 3.41 | 3.29 | 2.59 | 3.06 |
| P4 | 1.76 | 1.35 | 3.12 | 0.00 | 3.94 | 2.35 | 2.00 | 1.24 | 3.00 | 2.88 | 1.94 |
| P5 | 4.41 | 4.47 | 4.94 | 3.94 | 0.00 | 4.41 | 5.59 | 4.82 | 6.12 | 4.59 | 4.94 |
| P6 | 2.59 | 3.00 | 3.94 | 2.35 | 4.41 | 0.00 | 4.00 | 2.65 | 4.18 | 3.82 | 3.12 |
| P7 | 3.06 | 2.53 | 4.18 | 2.00 | 5.59 | 4.00 | 0.00 | 2.41 | 2.76 | 4.18 | 1.47 |
| P8 | 1.71 | 2.12 | 3.41 | 1.24 | 4.82 | 2.65 | 2.41 | 0.00 | 3.41 | 3.18 | 2.24 |
| P9 | 3.59 | 3.53 | 3.29 | 3.00 | 6.12 | 4.18 | 2.76 | 3.41 | 0.00 | 3.53 | 2.82 |
| P10 | 3.24 | 4.12 | 2.59 | 2.88 | 4.59 | 3.82 | 4.18 | 3.18 | 3.53 | 0.00 | 2.71 |
| P11 | 2.29 | 2.94 | 3.06 | 1.94 | 4.94 | 3.12 | 1.47 | 2.24 | 2.82 | 2.71 | 0.00 |

### c. Distância de Minkowsky

A distância de Minkowsky ($\mathbf{M_{ij}}$) é a generalização da distância absoluta. É obtida pela seguinte expressão:

$$M_{ij} = \sqrt[k]{\left[W \sum_{h=1}^{p} \left|X_{hi} - X_{hj}\right|^{k}\right]}$$

Para **p** = 17 e **k** = 1; e **p** = 17 e **k** = 2, essa distância passa a ser a distância absoluta ($\mathbf{W_h}$ = **1/p**) e a distância euclidiana média, respectivamente.

### Passo a passo no R: Distância de Minkowski

A Distância de Minkowski é uma medida de dissimilaridade versátil que pode se adaptar a diferentes tipos de distâncias, dependendo do valor de um parâmetro k. Ela é como uma "distância coringa", pois, com **k** = 1, torna-se a Distância Absoluta Média (Manhattan) e, com **k** = 2, torna-se a Distância Euclidiana Média.

Para calcular essas distâncias, utilizaremos a função *dist()* do R. Como nossa matriz X (quantitativa) tem as parcelas nas colunas e a função *dist()* espera que as observações (parcelas) estejam nas linhas, precisaremos transpor a matriz X (t(X)) antes de aplicar a função.

### 1. Distância de Minkowski com k = 1 (Distância Absoluta Média)

Quando definimos **k** = 1 (ou **p** = 1 no argumento da função *dist())*, a Distância de Minkowski se comporta como a Distância Absoluta Média (Manhattan). Para obter a versão "média", dividimos o resultado pelo número de espécies (n_especies).

```
# Definimos o número de espécies (linhas da nossa matriz X)
n_especies <- nrow(X)

# Calculamos a Distância de Minkowski com k=1 (Absoluta Média)
# Transpomos a matriz X (t(X)) para que as parcelas fiquem nas linhas
matriz_minkowski_k1 <- as.matrix(
  dist(
    t(X),
    method = "minkowski",
    p = 1
  )
) / n_especies

# Exibimos a matriz de distâncias, arredondada para duas casas decimais
print(round(matriz_minkowski_k1, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 2.53 | 2.76 | 1.76 | 4.41 | 2.59 | 3.06 | 1.71 | 3.59 | 3.24 | 2.29 |
| P2 | 2.53 | 0.00 | 3.88 | 1.35 | 4.47 | 3.00 | 2.53 | 2.12 | 3.53 | 4.12 | 2.94 |
| P3 | 2.76 | 3.88 | 0.00 | 3.12 | 4.94 | 3.94 | 4.18 | 3.41 | 3.29 | 2.59 | 3.06 |
| P4 | 1.76 | 1.35 | 3.12 | 0.00 | 3.94 | 2.35 | 2.00 | 1.24 | 3.00 | 2.88 | 1.94 |
| P5 | 4.41 | 4.47 | 4.94 | 3.94 | 0.00 | 4.41 | 5.59 | 4.82 | 6.12 | 4.59 | 4.94 |
| P6 | 2.59 | 3.00 | 3.94 | 2.35 | 4.41 | 0.00 | 4.00 | 2.65 | 4.18 | 3.82 | 3.12 |
| P7 | 3.06 | 2.53 | 4.18 | 2.00 | 5.59 | 4.00 | 0.00 | 2.41 | 2.76 | 4.18 | 1.47 |
| P8 | 1.71 | 2.12 | 3.41 | 1.24 | 4.82 | 2.65 | 2.41 | 0.00 | 3.41 | 3.18 | 2.24 |
| P9 | 3.59 | 3.53 | 3.29 | 3.00 | 6.12 | 4.18 | 2.76 | 3.41 | 0.00 | 3.53 | 2.82 |
| P10 | 3.24 | 4.12 | 2.59 | 2.88 | 4.59 | 3.82 | 4.18 | 3.18 | 3.53 | 0.00 | 2.71 |
| P11 | 2.29 | 2.94 | 3.06 | 1.94 | 4.94 | 3.12 | 1.47 | 2.24 | 2.82 | 2.71 | 0.00 |

**2. Distância de Minkowski com k = 2 (Distância Euclidiana Média Padronizada)**

Quando definimos **k** = 2 (ou **p** = 2 no argumento da função *dist())*, a Distância de Minkowski se torna a Distância Euclidiana Média Padronizada. Para obter a versão "média", dividimos o resultado pela raiz quadrada do número de espécies (sqrt(n_especies)).

```
# Calculamos a Distância de Minkowski com k=2 (Euclidiana Média Padronizada)
# Usamos a matriz X_pad que é a matriz de dados padronizada
# Transpomos a matriz X_pad (t(X_pad_)) para que as parcelas fiquem nas linhas
matriz_minkowski_k2 <- as.matrix(
  dist(
    t(X_pad),
    method = "minkowski",
    p = 2
  )
) / sqrt(n_especies)

# Exibimos a matriz de distâncias, arredondada para duas casas decimais
print(round(matriz_minkowski_k2, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 1.38 | 0.91 | 1.25 | 1.13 | 0.99 | 1.51 | 1.08 | 1.10 | 1.19 | 1.29 |
| P2 | 1.38 | 0.00 | 1.40 | 0.96 | 1.11 | 1.23 | 1.33 | 1.39 | 1.32 | 1.49 | 1.48 |
| P3 | 0.91 | 1.40 | 0.00 | 1.43 | 1.31 | 1.13 | 1.41 | 1.25 | 0.83 | 0.75 | 1.09 |
| P4 | 1.25 | 0.96 | 1.43 | 0.00 | 0.62 | 0.80 | 1.38 | 1.00 | 1.37 | 1.33 | 1.33 |
| P5 | 1.13 | 1.11 | 1.31 | 0.62 | 0.00 | 0.91 | 1.41 | 1.18 | 1.32 | 1.23 | 1.27 |
| P6 | 0.99 | 1.23 | 1.13 | 0.80 | 0.91 | 0.00 | 1.46 | 1.09 | 1.18 | 1.14 | 1.21 |
| P7 | 1.51 | 1.33 | 1.41 | 1.38 | 1.41 | 1.46 | 0.00 | 1.41 | 0.91 | 1.47 | 0.75 |
| P8 | 1.08 | 1.39 | 1.25 | 1.00 | 1.18 | 1.09 | 1.41 | 0.00 | 1.21 | 1.29 | 1.27 |
| P9 | 1.10 | 1.32 | 0.83 | 1.37 | 1.32 | 1.18 | 0.91 | 1.21 | 0.00 | 1.00 | 0.76 |
| P10 | 1.19 | 1.49 | 0.75 | 1.33 | 1.23 | 1.14 | 1.47 | 1.29 | 1.00 | 0.00 | 1.06 |
| P11 | 1.29 | 1.48 | 1.09 | 1.33 | 1.27 | 1.21 | 0.75 | 1.27 | 0.76 | 1.06 | 0.00 |

### d. Coeficiente de similaridade de Cattell

O coeficiente de similaridade de Cattell (Cattell, 1944) é dado por:

$$\mathbf{C_{ij}} = \left[\frac{2(p - 2/3) - d_{ij}^2}{2(p - 2/3) + d_{ij}^2}\right]$$

Em que: **d** é a distância euclidiana média com as variáveis padronizadas para terem variância igual a 1.

### Aplicação

Utilizando a distância euclidiana média padronizada para parcelas 1 e 2, teremos:

$$C_{12} = \left[\frac{2(17 - 2/3) - 1{,}38^2}{2(17 - 2/3) + 1{,}38^2}\right] = 0{,}89$$

Já para as parcelas 10 e 11, teremos:

$$C_{1011} = \left[\frac{2(17 - 2/3) - 1{,}06^2}{2(17 - 2/3) + 1{,}06^2}\right] = 0{,}93$$

Uma medida derivada desse é o coeficiente de similaridade de Cattell e Coulter (Cattell; Coulter, 1966):

$$C_{ij} = \left[\frac{2\sqrt{p} - d_{ij}^2}{2\sqrt{p} + d_{ij}^2}\right]$$

Assim, para as parcelas 1 e 2, teremos:

$$C_{12} = \left[\frac{2\sqrt{17} - 1{,}38^2}{2\sqrt{17} + 1{,}38^2}\right] = 0{,}62$$

E, para parcelas 10 e 11:

$$C_{1011} = \left[\frac{2\sqrt{17} - 1{,}06^2}{2\sqrt{17} + 1{,}06^2}\right] = 0{,}76$$

Mais detalhes sobre estes coeficientes podem ser encontrados em Comarck (1971).

**Passo a passo no R: Coeficiente de similaridade de Cattell e de Cattell e Coulter**

**1. Coeficiente de Similaridade de Cattell**

O Coeficiente de Similaridade de Cattell utiliza a Distância Euclidiana Média Padronizada (calculada na seção anterior) e o número de espécies para determinar a similaridade entre as parcelas. Ele é calculado a partir de uma fórmula que transforma a distância em similaridade.

```
# Definimos o número de espécies (linhas da nossa matriz X)
n_especies <- nrow(X)

# Calculamos a constante 'k' para o Coeficiente de Cattell
k_cattell <- 2 * (n_especies - 2/3)

# Calculamos a matriz de similaridade de Cattell
# Usamos a matriz_euclidiana_media calculada anteriormente
matriz_sim_cattell <- (k_cattell - matriz_euclidiana_media^2) /
  (k_cattell + matriz_euclidiana_media^2)

# Exibimos a matriz de similaridade, arredondada para duas casas decimais
print(round(matriz_sim_cattell, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.89 | 0.95 | 0.91 | 0.92 | 0.94 | 0.87 | 0.93 | 0.93 | 0.92 | 0.90 |
| P2 | 0.89 | 1.00 | 0.89 | 0.94 | 0.93 | 0.91 | 0.90 | 0.89 | 0.90 | 0.87 | 0.87 |
| P3 | 0.95 | 0.89 | 1.00 | 0.88 | 0.90 | 0.93 | 0.89 | 0.91 | 0.96 | 0.97 | 0.93 |
| P4 | 0.91 | 0.94 | 0.88 | 1.00 | 0.97 | 0.96 | 0.89 | 0.94 | 0.89 | 0.89 | 0.90 |
| P5 | 0.92 | 0.93 | 0.90 | 0.97 | 1.00 | 0.95 | 0.89 | 0.92 | 0.90 | 0.91 | 0.91 |
| P6 | 0.94 | 0.91 | 0.93 | 0.96 | 0.95 | 1.00 | 0.88 | 0.93 | 0.92 | 0.92 | 0.91 |
| P7 | 0.87 | 0.90 | 0.89 | 0.89 | 0.89 | 0.88 | 1.00 | 0.88 | 0.95 | 0.88 | 0.97 |
| P8 | 0.93 | 0.89 | 0.91 | 0.94 | 0.92 | 0.93 | 0.88 | 1.00 | 0.91 | 0.90 | 0.91 |
| P9 | 0.93 | 0.90 | 0.96 | 0.89 | 0.90 | 0.92 | 0.95 | 0.91 | 1.00 | 0.94 | 0.97 |
| P10 | 0.92 | 0.87 | 0.97 | 0.89 | 0.91 | 0.92 | 0.88 | 0.90 | 0.94 | 1.00 | 0.93 |
| P11 | 0.90 | 0.87 | 0.93 | 0.90 | 0.91 | 0.91 | 0.97 | 0.91 | 0.97 | 0.93 | 1.00 |

**2. Coeficiente de Similaridade de Cattell e Coulter**

Esse é uma variação do coeficiente de Cattell, que também se baseia na Distância Euclidiana Média Padronizada, mas com uma pequena diferença na constante **k** da fórmula. Essa alteração ajusta a sensibilidade do coeficiente a diferentes cenários de dados.

```
# Definimos o número de espécies (linhas da nossa matriz X)
n_especies <- nrow(X)

# Calculamos a constante 'k' para o Coeficiente de Cattell e Coulter
k_cattell_coulter <- 2 * sqrt(n_especies)

# Calculamos a matriz de similaridade de Cattell e Coulter
```

```
# Usamos a matriz_euclidiana_media calculada anteriormente
matriz_sim_cattell_coulter <- (k_cattell_coulter - matriz_euclidiana_media^2) /
  (k_cattell_coulter + matriz_euclidiana_media^2)

# Exibimos a matriz de similaridade, arredondada para duas casas decimais
print(round(matriz_sim_cattell_coulter, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 1.00 | 0.62 | 0.82 | 0.67 | 0.73 | 0.79 | 0.57 | 0.75 | 0.74 | 0.71 | 0.66 |
| P2 | 0.62 | 1.00 | 0.62 | 0.79 | 0.74 | 0.69 | 0.65 | 0.62 | 0.65 | 0.57 | 0.58 |
| P3 | 0.82 | 0.62 | 1.00 | 0.60 | 0.66 | 0.73 | 0.61 | 0.68 | 0.85 | 0.87 | 0.75 |
| P4 | 0.67 | 0.79 | 0.60 | 1.00 | 0.90 | 0.85 | 0.62 | 0.78 | 0.62 | 0.64 | 0.64 |
| P5 | 0.73 | 0.74 | 0.66 | 0.90 | 1.00 | 0.82 | 0.61 | 0.71 | 0.65 | 0.69 | 0.67 |
| P6 | 0.79 | 0.69 | 0.73 | 0.85 | 0.82 | 1.00 | 0.59 | 0.75 | 0.71 | 0.73 | 0.70 |
| P7 | 0.57 | 0.65 | 0.61 | 0.62 | 0.61 | 0.59 | 1.00 | 0.61 | 0.82 | 0.59 | 0.87 |
| P8 | 0.75 | 0.62 | 0.68 | 0.78 | 0.71 | 0.75 | 0.61 | 1.00 | 0.70 | 0.66 | 0.67 |
| P9 | 0.74 | 0.65 | 0.85 | 0.62 | 0.65 | 0.71 | 0.82 | 0.70 | 1.00 | 0.78 | 0.87 |
| P10 | 0.71 | 0.57 | 0.87 | 0.64 | 0.69 | 0.73 | 0.59 | 0.66 | 0.78 | 1.00 | 0.76 |
| P11 | 0.66 | 0.58 | 0.75 | 0.64 | 0.67 | 0.70 | 0.87 | 0.67 | 0.87 | 0.76 | 1.00 |

**e. Coeficiente de Bray-Curtis**

O coeficiente de Bray-Curtis (**$BC_{ij}$**) é obtido pela seguinte expressão:

$$BC_{ij} = \left[\frac{1}{p}\left(\frac{\sum_h^p |X_{hi} - X_{hj}|}{\sum_h^p (X_{hi} + X_{hj})}\right)\right]$$

**Aplicação**

$$BC_{12} = \left[\frac{1}{p}\left(\frac{\sum_h^p |X_{h1} - X_{h2}|}{\sum_h^p (X_{h1} + X_{h2})}\right)\right] = \left[\frac{1}{17}\left(\frac{43}{59}\right)\right] = 0{,}04$$

E

$$BC_{1011} = \left[\frac{1}{p}\left(\frac{\sum_h^p |X_{h10} - X_{h11}|}{\sum_h^p (X_{h10} + X_{h11})}\right)\right] = \left[\frac{1}{17}\left(\frac{46}{104}\right)\right] = 0{,}03$$

**Passo a passo no R: Dissimilaridade de Bray-Curtis**

**1. Calculando a Dissimilaridade de Bray-Curtis**

A Dissimilaridade de Bray-Curtis é uma medida comum em Ecologia para comparar parcelas com base na abundância das espécies. A versão apresentada aqui é padronizada, dividindo a dissimilaridade obtida pelo número total de espécies. Isso permite que o valor final seja expresso em relação ao conjunto total de espécies analisadas.

Para calcular essa medida, utilizaremos a função *vegdist()* do pacote *vegan* com o argumento *method = "bray"*. Como nossa matriz X (quantitativa) tem as parcelas nas colunas e a função *vegdist()* espera que as observações (parcelas) estejam nas linhas, precisamos

transpor a matriz X (*t(X)*) antes de aplicar a função. O resultado será, então, dividido pelo número de espécies (*nrow(X)*).

```
# Carregamos o pacote para utilizar suas funções
# (Execute este comando apenas uma vez por sessão do R)
library(vegan)

# Definimos o número de espécies (linhas da nossa matriz X)
n_especies <- nrow(X)

# Calculamos a Dissimilaridade de Bray-Curtis e a dividimos pelo número de espécies
# Transpomos a matriz X (t(X)) para que as parcelas fiquem nas linhas e sejam comparadas
matriz_dist_bray_padronizada <- as.matrix(
  vegdist(
    t(X),
    method = "bray"
  )
) / n_especies

# Exibimos a matriz de dissimilaridade padronizada, arredondada para duas casas decimais
print(round(Matriz_dist_bray_padronizada, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.04 | 0.03 | 0.03 | 0.04 | 0.03 | 0.05 | 0.03 | 0.04 | 0.03 | 0.03 |
| P2 | 0.04 | 0.00 | 0.05 | 0.03 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 | 0.04 |
| P3 | 0.03 | 0.05 | 0.00 | 0.04 | 0.04 | 0.04 | 0.05 | 0.04 | 0.03 | 0.02 | 0.03 |
| P4 | 0.03 | 0.03 | 0.04 | 0.00 | 0.04 | 0.03 | 0.04 | 0.03 | 0.04 | 0.04 | 0.03 |
| P5 | 0.04 | 0.04 | 0.04 | 0.04 | 0.00 | 0.03 | 0.05 | 0.05 | 0.04 | 0.03 | 0.04 |
| P6 | 0.03 | 0.04 | 0.04 | 0.03 | 0.03 | 0.00 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 |
| P7 | 0.05 | 0.04 | 0.05 | 0.04 | 0.05 | 0.05 | 0.00 | 0.05 | 0.03 | 0.04 | 0.02 |
| P8 | 0.03 | 0.04 | 0.04 | 0.03 | 0.05 | 0.04 | 0.05 | 0.00 | 0.04 | 0.04 | 0.03 |
| P9 | 0.04 | 0.04 | 0.03 | 0.04 | 0.04 | 0.04 | 0.03 | 0.04 | 0.00 | 0.03 | 0.03 |
| P10 | 0.03 | 0.05 | 0.02 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 | 0.03 | 0.00 | 0.03 |
| P11 | 0.03 | 0.04 | 0.03 | 0.03 | 0.04 | 0.03 | 0.02 | 0.03 | 0.03 | 0.03 | 0.00 |

### f. Distância de Canberra

A distância de Canberra ($\mathbf{DC_{ij}}$) foi proposta por Lance e Williams (1967). A utilização dessa medida em análise de agrupamento tem o objetivo de avaliar a resistência de cultivares, segundo as suas curvas de progresso de doença, em complemento aos procedimentos convencionais (Thompson; Reis, 1979).

Essa medida é expressa por:

$$DC_{ij} = \left[\sum_{h=1}^{p}\left(\frac{\left|X_{hi} - X_{hj}\right|}{\left|X_{hi}\right| + \left|X_{hj}\right|}\right)\right]$$

**Aplicação**

No cálculo da Distância de Canberra, não se inclui as espécies ausentes em ambas as parcelas. No caso das parcelas 1 e 2, consideramos apenas as Espécies 1, 3, 4, 5, 8, 9, 10, 12 e 14, logo:

$$DC_{12} = \frac{|X_{11}-X_{12}|}{|X_{11}|+|X_{12}|} + \frac{|X_{31}-X_{32}|}{|X_{31}|+|X_{32}|} + \frac{|X_{41}-X_{42}|}{|X_{41}|+|X_{42}|} + \frac{|X_{51}-X_{52}|}{|X_{51}|+|X_{52}|} + \frac{|X_{81}-X_{82}|}{|X_{81}|+|X_{82}|} + \frac{|X_{91}-X_{92}|}{|X_{91}|+|X_{92}|} + \frac{|X_{101}-X_{102}|}{|X_{101}|+|X_{102}|} + \frac{|X_{121}-X_{122}|}{|X_{121}|+|X_{122}|} + \frac{|X_{141}-X_{142}|}{|X_{141}|+|X_{142}|}$$

$$DC_{12} = \frac{|8-1|}{|1|+|8|} + \frac{|3-9|}{|3|+|9|} + \frac{|6-3|}{|6|+|3|} + \frac{|0-12|}{|0|+|12|} + \frac{|1-1|}{|1|+|1|} + \frac{|2-0|}{|2|+|0|} + \frac{|1-0|}{|1|+|0|} + \frac{|5-0|}{|5|+|0|} + \frac{|7-0|}{|7|+|0|}$$

$$DC_{12} = 0{,}78 + 0{,}50 + 0{,}33 + 1{,}00 + 0{,}00 + 1{,}00 + 1{,}00 + 1{,}00 + 1{,}00 = 6{,}61$$

Já para as parcelas 10 e 11, excluiremos dos cálculos as Espécies 5, 16 e 17:

$$DC_{1011} = \frac{|15-7|}{|15|+|7|} + \frac{|1-9|}{|1|+|9|} + \frac{|4-4|}{|4|+|4|} + \frac{|4-4|}{|4|+|4|} + \frac{|11-1|}{|11|+|1|} + \frac{|11-5|}{|11|+|5|} + \frac{|3-1|}{|3|+|1|} + \frac{|2-2|}{|2|+|2|} + \frac{|3-1|}{|3|+|1|} + \frac{|0-5|}{|0|+|5|} + \frac{|1-0|}{|1|+|0|} + \frac{|5-2|}{|5|+|2|} + \frac{|2-1|}{|2|+|1|}$$

$$DC_{1011} = 0{,}36 + 0{,}80 + 0{,}00 + 0{,}00 + 0{,}83 + 0{,}38 + 0{,}50 + 0{,}00 + 0{,}50 + 1{,}00 + 1{,}00 + 0{,}43 + 0{,}33 = 6{,}13$$

**Passo a passo no R: Distância de Canberra**

**1. Desenvolvendo a Função do Coeficiente**

A Distância de Canberra é uma medida de dissimilaridade baseada nas diferenças proporcionais entre as abundâncias das espécies. Para cada espécie, calcula-se a razão entre a diferença absoluta de abundância e a soma das abundâncias observadas nas duas parcelas. Essas razões são, então, somadas para compor a distância final.

```
# Criando a função para calcular a Distância de Canberra entre duas parcelas
canberra_dist_livro <- function(parcela1, parcela2) {
  # Garante que os vetores são numéricos
  xi <- as.numeric(parcela1)
  xj <- as.numeric(parcela2)

  # Calcula a soma das razões (diferença absoluta / soma das abundâncias)
  distancia_final <- sum(
    ifelse(
      (abs(xi) + abs(xj)) > 0, # Evita divisão por zero
      abs(xi - xj) / (abs(xi) + abs(xj)),
      0
    )
  )

  return(distancia_final)
}
```

## 2. Calculando a Matriz de Distância de Canberra

Para obter a Distância de Canberra para todas as combinações de parcelas, aplicaremos a função *canberra_dist_livro* a cada par de colunas da nossa matriz X (matriz quantitativa), pois as parcelas estão organizadas nas colunas.

```
# Calculando a matriz de Distância de Canberra entre as parcelas
# A função outer() aplica a função canberra_dist_livro para cada par de colunas (parcelas) da X
Matriz_dist_canberra_livro <- outer(
  1:ncol(X), # Itera sobre as colunas (parcelas) da matriz X
  1:ncol(X), # Itera novamente sobre as colunas (parcelas) da matriz X
  Vectorize(function(i, j) {
    # Aplica a função canberra_dist_livro para os dados das colunas (parcelas) i e j
    canberra_dist_livro(X[, i], X[, j])
  })
)

# Nomeando as linhas e colunas da matriz de distância de Canberra com os nomes das parcelas
dimnames(Matriz_dist_canberra_livro) <- list(
  colnames(X),
  colnames(X)
)

# Visualizando o resultado final da matriz de distância de Canberra, arredondado para duas casas decimais
print(round(matriz_dist_canberra_livro, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 6.61 | 8.19 | 7.20 | 6.19 | 7.00 | 11.85 | 5.75 | 9.40 | 8.31 | 8.41 |
| P2 | 6.61 | 0.00 | 9.31 | 6.52 | 6.23 | 7.26 | 9.47 | 7.28 | 6.90 | 10.90 | 10.28 |
| P3 | 8.19 | 9.31 | 0.00 | 8.16 | 6.45 | 7.65 | 12.12 | 9.70 | 10.18 | 7.73 | 8.40 |
| P4 | 7.20 | 6.52 | 8.16 | 0.00 | 7.26 | 6.27 | 9.42 | 8.43 | 7.08 | 8.27 | 7.53 |
| P5 | 6.19 | 6.23 | 6.45 | 7.26 | 0.00 | 6.43 | 12.76 | 9.52 | 10.47 | 7.69 | 9.29 |
| P6 | 7.00 | 7.26 | 7.65 | 6.27 | 6.43 | 0.00 | 12.26 | 6.73 | 9.71 | 9.79 | 8.80 |
| P7 | 11.85 | 9.47 | 12.12 | 9.42 | 12.76 | 12.26 | 0.00 | 10.19 | 7.55 | 12.48 | 8.57 |
| P8 | 5.75 | 7.28 | 9.70 | 8.43 | 9.52 | 6.73 | 10.19 | 0.00 | 7.70 | 8.76 | 9.68 |
| P9 | 9.40 | 6.90 | 10.18 | 7.08 | 10.47 | 9.71 | 7.55 | 7.70 | 0.00 | 9.93 | 8.76 |
| P10 | 8.31 | 10.90 | 7.73 | 8.27 | 7.69 | 9.79 | 12.48 | 8.76 | 9.93 | 0.00 | 6.13 |
| P11 | 8.41 | 10.28 | 8.40 | 7.53 | 9.29 | 8.80 | 8.57 | 9.68 | 8.76 | 6.13 | 0.00 |

### g. Coeficiente de Sokal e Sneath

O coeficiente de Sokal e Sneath ($SS_{ij}$) apresenta variações (Sokal e Sneath, 1963), dentre elas, temos a dada pela seguinte expressão:

$$SS_{ij} = \left[\frac{1}{p}\sum_{h=1}^{p}\left(\frac{X_{hi} - X_{hj}}{X_{hi} + X_{hj}}\right)^2\right]^{1/2}$$

Outra variação do coeficiente de Sokal e Sneath pode ser observada na Tabela 11 (página 41).

**Aplicação**

No cálculo do Coeficiente de Sokal e Sneath, não se incluem as espécies ausentes em ambas as parcelas. Assim, para seu cálculo, considerando as parcelas 1 e 2, as Espécies 2, 6, 7, 11, 13, 15, 16, e 17 não foram consideradas (Tabela 3, página 17).

$$SS_{12} = \left[\frac{1}{17}\left\{\left(\frac{X_{11}-X_{12}}{X_{11}+X_{12}}\right)^2 + \left(\frac{X_{31}-X_{32}}{X_{31}+X_{32}}\right)^2 + \left(\frac{X_{41}-X_{42}}{X_{41}+X_{42}}\right)^2 + \left(\frac{X_{51}-X_{52}}{X_{51}+X_{52}}\right)^2 + \left(\frac{X_{81}-X_{82}}{X_{81}+X_{82}}\right)^2 + \left(\frac{X_{91}-X_{92}}{X_{91}+X_{92}}\right)^2 + \left(\frac{X_{101}-X_{102}}{X_{101}+X_{102}}\right)^2 + \left(\frac{X_{121}-X_{122}}{X_{121}+X_{122}}\right)^2 + \left(\frac{X_{141}-X_{142}}{X_{141}+X_{142}}\right)^2\right\}\right]^{1/2}$$

$$SS_{12} = \left[\frac{1}{17}\left\{\left(\frac{8-1}{8+1}\right)^2 + \left(\frac{3-9}{3+9}\right)^2 + \left(\frac{6-3}{6+3}\right)^2 + \left(\frac{0-12}{0+12}\right)^2 + \left(\frac{1-1}{1+1}\right)^2 + \left(\frac{2-0}{2+0}\right)^2 + \left(\frac{1-0}{1+0}\right)^2 + \left(\frac{5-0}{5+0}\right)^2 + \left(\frac{7-0}{7+0}\right)^2\right\}\right]^{1/2}$$

$$SS_{12} = \left[\frac{1}{17}\{0{,}60 + 0{,}25 + 0{,}11 + 1{,}00 + 0{,}00 + 1{,}00 + 1{,}00 + 1{,}00 + 1{,}00\}\right]^{1/2}$$

$$SS_{12} = \left[\frac{1}{17}.5{,}97\right]^{1/2} = 0{,}59$$

**Passo a passo no R: Coeficiente de Sokal e Sneath**

**1. Desenvolvendo a Função do Coeficiente**

O Coeficiente de Sokal e Sneath é uma medida de dissimilaridade que utiliza a raiz quadrada da média dos quadrados das diferenças proporcionais entre as abundâncias das espécies. Ele dá um peso maior para as diferenças maiores, resultando em um valor que indica o quão diferentes as parcelas são.

```
# Criando a função para calcular o Coeficiente de Sokal e Sneath entre duas parcelas
sokal_sneath_dist_livro <- function(parcela1, parcela2) {
  # Garante que os vetores são numéricos
  xi <- as.numeric(parcela1)
  xj <- as.numeric(parcela2)

  # Calcula as razões das diferenças proporcionais
  razoes <- ifelse(
    (xi + xj) > 0, # Evita divisão por zero
    (xi - xj) / (xi + xj),
    0
  )

  # Retorna a raiz quadrada da média dos quadrados das razões
  return(sqrt(mean(razoes^2)))
}
```

## 2. Calculando a Matriz de Dissimilaridade de Sokal e Sneath

Para obter o Coeficiente de Sokal e Sneath para todas as combinações de parcelas, aplicaremos a função sokal_sneath_dist_livro a cada par de colunas da nossa matriz X (matriz quantitativa), pois as parcelas estão organizadas nas colunas.

```
# Calculando a matriz de Dissimilaridade de Sokal e Sneath entre as parcelas
# A função outer() aplica a função sokal_sneath_dist_livro para cada par de colunas (parcelas) da X
matriz_dist_sokal_sneath_livro <- outer(
  1:ncol(X), # Itera sobre as colunas (parcelas) da matriz X
  1:ncol(X), # Itera novamente sobre as colunas (parcelas) da matriz X
  Vectorize(function(i, j) {
    # Aplica a função sokal_sneath_dist_livro para os dados das colunas (parcelas) i e j
    sokal_sneath_dist_livro(X[, i], X[, j])
  })
)

# Nomeando as linhas e colunas da matriz de dissimilaridade com os nomes das parcelas
dimnames(matriz_dist_sokal_sneath_livro) <- list(
  colnames(X),
  colnames(X)
)

# Visualizando o resultado final da matriz de dissimilaridade de Sokal e Sneath, arredondado para duas casas decimais
print(round(matriz_dist_sokal_sneath_livro, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.59 | 0.66 | 0.61 | 0.57 | 0.59 | 0.82 | 0.52 | 0.71 | 0.65 | 0.69 |
| P2 | 0.59 | 0.00 | 0.73 | 0.59 | 0.59 | 0.63 | 0.72 | 0.63 | 0.61 | 0.77 | 0.76 |
| P3 | 0.66 | 0.73 | 0.00 | 0.67 | 0.58 | 0.62 | 0.82 | 0.73 | 0.75 | 0.63 | 0.67 |
| P4 | 0.61 | 0.59 | 0.67 | 0.00 | 0.61 | 0.55 | 0.73 | 0.67 | 0.62 | 0.66 | 0.63 |
| P5 | 0.57 | 0.59 | 0.58 | 0.61 | 0.00 | 0.54 | 0.85 | 0.72 | 0.76 | 0.62 | 0.70 |
| P6 | 0.59 | 0.63 | 0.62 | 0.55 | 0.54 | 0.00 | 0.83 | 0.59 | 0.72 | 0.70 | 0.62 |
| P7 | 0.82 | 0.72 | 0.82 | 0.73 | 0.85 | 0.83 | 0.00 | 0.75 | 0.62 | 0.82 | 0.68 |
| P8 | 0.52 | 0.63 | 0.73 | 0.67 | 0.72 | 0.59 | 0.75 | 0.00 | 0.64 | 0.68 | 0.71 |
| P9 | 0.71 | 0.61 | 0.75 | 0.62 | 0.76 | 0.72 | 0.62 | 0.64 | 0.00 | 0.73 | 0.66 |
| P10 | 0.65 | 0.77 | 0.63 | 0.66 | 0.62 | 0.70 | 0.82 | 0.68 | 0.73 | 0.00 | 0.51 |
| P11 | 0.69 | 0.76 | 0.67 | 0.63 | 0.70 | 0.62 | 0.68 | 0.71 | 0.66 | 0.51 | 0.00 |

### h. Distância euclidiana generalizada ou $D^2$ de Mahalanobis

A análise multivariada desempenha um papel crucial em estudos ecológicos, permitindo a compreensão de padrões complexos de similaridade e dissimilaridade entre unidades amostrais ao considerar simultaneamente múltiplas variáveis, como a abundância de diferentes espécies (Legendre & Legendre, 2012). No entanto, dados ecológicos de comunidades frequentemente exibem características que desafiam as abordagens estatísticas tradicionais, tais como alta dimensionalidade, presença de muitas espécies raras, abundância de zeros,

correlações intrínsecas entre espécies, número limitado de unidades amostrais (**n** < **p**) e heterogeneidade espacial (Borcard, Gillet & Legendre, 2018).

Essas particularidades tornam inadequado o uso de medidas de distância euclidianas simples, como a Euclidiana, a de Manhattan, a de Canberra ou mesmo a de Bray-Curtis em certos contextos inferenciais. Tais medidas não conseguem capturar a complexa estrutura de covariância presente nos dados ecológicos, levando a interpretações potencialmente errôneas da similaridade ou dissimilaridade (McCune & Grace, 2002).

A Distância de Mahalanobis (**D²**) surge como uma alternativa robusta e poderosa, pois incorpora explicitamente a variância de cada variável, a covariância entre as variáveis e a estrutura geométrica multivariada do conjunto de dados (Mahalanobis, 1936). Desenvolvida por Prasanta Chandra Mahalanobis em 1936, essa medida de distância é particularmente útil em contextos em que as variáveis estão correlacionadas e possuem diferentes escalas de medida, pois padroniza os dados pela matriz de covariância (De Maesschalck, Jouan-Rimbaud & Massart, 2000).

A distância de Mahalanobis (**$D^2$**), ou distância euclidiana generalizada, é dada pela seguinte expressão:

$$D_{ij}^2 = \left(X_i - X_j\right)' \Sigma^{-1} \left(X_i - X_j\right)$$

Em que: $\sum^{-1}$ é a inversa da matriz de covariância residual de **X**; e **$D^2$** tem a característica de ser invariante para qualquer transformação linear não singular. Ao contrário da distância euclidiana, **$D^2$** pode ser utilizada quando existe correlação entre as variáveis. Caso os coeficientes de correlação sejam nulos, o valor de **$D^2$** equivale à distância euclidiana para variáveis padronizadas.

Para o cálculo de **$D^2$**, supõe-se a existência de distribuição multinormal **p**-dimensional e a homogeneidade da matriz de covariância residual das unidades amostrais, restringindo-se, portanto, o seu uso. Entretanto, considerável robustez à violação dessas hipóteses já foi demonstrada, o que faz da distância de Mahalanobis uma opção de grande utilidade, principalmente pelo fato de **$D^2$** ter analogia com outras técnicas multivariadas (Cruz et al., 2012). Além disso, a distância de Mahalanobis considera a variabilidade dentro de cada unidade amostral, e não somente a medida de tendência central, sendo, portanto, uma medida mais aceitável quando as unidades amostrais constituem um conjunto de indivíduos e, principalmente, quando as variáveis são correlacionadas.

A distância generalizada das parcelas **j** e **k** é definida por (Orlóci, 1978):

$$D_{jk}^2 = \left(X_j - X_k\right)' S^{-1} \left(X_j - X_k\right)$$

Em que: $\mathbf{X_j}$ e $\mathbf{X_k}$ representam, respectivamente, os vetores parcelas **j** e **k**, e $\mathbf{S^{-1}}$ simboliza a matriz inversa de variância-covariância residual das espécies.

No entanto, a aplicação da distância de Mahalanobis em dados ecológicos de alta dimensionalidade (nos quais o número de variáveis, **p**, é maior que o número de observações, **n**) enfrenta desafios significativos. Nesses cenários, a matriz de covariância amostral torna-se mal condicionada ou até singular, impedindo sua inversão confiável e, consequentemente, o cálculo da distância de Mahalanobis (Gotelli & Ellison, 2004). Para contornar esse problema, técnicas de regularização da matriz de covariância são empregadas. Entre elas, o método *Shrinkage*, notadamente o estimador de Ledoit-Wolf, oferece uma solução eficaz ao combinar a matriz de covariância amostral com uma matriz-alvo mais estável, garantindo a invertibilidade e a estabilidade numérica (Ledoit & Wolf, 2004).

Neste item será apresentado, de forma exaustiva, o cálculo completo da distância de Mahalanobis regularizada por *Shrinkage*, utilizando uma matriz ecológica de abundância de espécies como exemplo. Todo o procedimento será demonstrado matematicamente, incluindo a transformação dos dados, a construção e a centralização da matriz de covariância, o cálculo de variâncias e covariâncias, a regularização por *shrinkage* (com foco no método de Ledoit-Wolf), a inversão da matriz de covariância e, finalmente, o cálculo e a interpretação ecológica da distância $D^2$ de Mahalanobis.

Para exemplificar o cálculo da distância de Mahalanobis com *shrinkage*, foi utilizada uma matriz ecológica de abundância com 17 espécies e 11 parcelas. A matriz original, em que as linhas representavam espécies e as colunas parcelas, foi transposta no *software* R para seguir a convenção de linhas como observações e colunas como variáveis. Assim, a matriz final **X** ficou com dimensão 11×17, correspondendo a 11 parcelas e 17 espécies.

Na Tabela 13, apresenta-se novamente a matriz de abundância de espécies, em que cada elemento $X_{ij}$ corresponde à abundância da espécie j na parcela **i**. Essa matriz constitui a base para as transformações e cálculos realizados posteriormente.

Dados ecológicos de abundância frequentemente apresentam características que violam pressupostos de métodos estatísticos multivariados, como normalidade e homogeneidade de variâncias. Esses dados geralmente possuem distribuições assimétricas, variâncias heterogêneas, predominância de espécies muito abundantes e grande quantidade de zeros, o que pode comprometer as análises e distorcer as relações entre espécies e parcelas.

**Tabela 13.** Matriz de abundância de espécies na Mata da Silvicultura, Viçosa-MG

| Parcela | Espécie | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **E1** | **E2** | **E3** | **E4** | **E5** | **E6** | **E7** | **E8** | **E9** | **E10** | **E11** | **E12** | **E13** | **E14** | **E15** | **E16** | **E17** |
| **P1** | 8 | 0 | 3 | 6 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 5 | 0 | 7 | 0 | 0 | 0 |
| **P2** | 1 | 0 | 9 | 3 | 12 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P3** | 27 | 0 | 4 | 3 | 0 | 1 | 10 | 0 | 2 | 0 | 0 | 7 | 1 | 0 | 0 | 0 | 1 |
| **P4** | 0 | 0 | 6 | 4 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P5** | 1 | 0 | 22 | 29 | 0 | 9 | 9 | 1 | 2 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| **P6** | 9 | 0 | 9 | 12 | 0 | 1 | 2 | 13 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **P7** | 2 | 12 | 5 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 6 | 0 | 2 | 0 | 0 | 1 | 0 |
| **P8** | 3 | 1 | 2 | 4 | 0 | 0 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **P9** | 22 | 17 | 7 | 4 | 2 | 5 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P10** | 15 | 1 | 4 | 4 | 0 | 11 | 11 | 3 | 2 | 3 | 0 | 1 | 5 | 0 | 2 | 0 | 0 |
| **P11** | 7 | 9 | 4 | 4 | 0 | 1 | 5 | 1 | 2 | 1 | 5 | 0 | 2 | 0 | 1 | 0 | 0 |

Em que: **E1** = *Casearia decandra* Jacq.; **E2** = *Anadenanthera peregrina* (L.) Speg.; **E3** = *Apuleia leiocarpa* (Vog.) J. F. Macbr.; **E4** = *Mabea fistulifera* Mart.; **E5** = *Anadenanthera macrocarpa* (Benth.) Brenan.; **E6** = *Platypodium elegans* Vog.; **E7** = *Machaerium floridum* (Mart. ex Benth.) Ducke; **E8** = *Copaifera langsdorffii* Desf.; **E9** = *Ocotea pretiosa* (Nees) Mez.; **E10** = *Cabralea cangerana* Saldanha; **E11** = *Piptadenia gonoacantha* (Mart.) J. F. Macbr.; **E12** = *Dalbergia nigra* (Vell.) Allemão ex Benth.; **E13** = *Luehea divaricata* Mart.; **E14** = *Cecropia hololeuca* Miq.; **E15** = *Melanoxylum brauna* Schott.; **E16** = *Cedrela fissilis* Vell.; e **E17** = *Croton floribundus* Spreng.

Para reduzir esses problemas, aplicou-se a transformação pela raiz quadrada, dada por:

$$X'_{ij} = \sqrt{X_{ij}}$$

Essa transformação diminui a influência excessiva de espécies muito abundantes, contribui para a estabilização das variâncias e reduz a assimetria dos dados. Embora não garanta normalidade, ela torna a distribuição mais equilibrada e adequada para análises multivariadas baseadas em distâncias e covariâncias.

Para ilustrar a aplicação da transformação pela raiz quadrada, consideram-se alguns valores da matriz original:

• Exemplo 1: abundância igual a 8 para a espécie **E1** na parcela **P1**. Após a transformação: $\sqrt{8} \approx 2{,}83$;

• Exemplo 2: abundância igual a 27 para a espécie **E1** na parcela **P3**. Após a transformação: $\sqrt{27} \approx 5{,}20$; e

• Exemplo 3: abundância igual a 22 para a espécie **E1** na parcela **P9**. Após a transformação: $\sqrt{22} \approx 4{,}69$.

A matriz de abundância após a aplicação da transformação pela raiz quadrada é apresentada na Tabela 14. Observa-se que os valores originais foram substituídos por suas respectivas raízes quadradas, enquanto os valores iguais a zero permaneceram inalterados. Essa transformação reduz a influência de espécies muito abundantes e contribui para a estabilização da variância dos dados, tornando-os mais adequados para análises multivariadas.

**Tabela 14.** Matriz de abundância de espécies após a aplicação da transformação pela raiz quadrada, Mata da Silvicultura, Viçosa-MG

| Parcela | Espécie | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **E1** | **E2** | **E3** | **E4** | **E5** | **E6** | **E7** | **E8** | **E9** | **E10** | **E11** | **E12** | **E13** | **E14** | **E15** | **E16** | **E17** |
| **P1** | 2,83 | 0,00 | 1,73 | 2,45 | 0,00 | 0,00 | 0,00 | 1,00 | 1,41 | 1,00 | 0,00 | 2,24 | 0,000 | 2,64 | 0,00 | 0,00 | 0,00 |
| **P2** | 1,00 | 0,00 | 3,00 | 1,73 | 3,46 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| **P3** | 5,20 | 0,00 | 2,00 | 1,73 | 0,00 | 1,00 | 3,16 | 0,00 | 1,41 | 0,00 | 0,00 | 2,65 | 1,00 | 0,00 | 0,00 | 0,00 | 1,00 |
| **P4** | 0,00 | 0,00 | 2,45 | 2,00 | 1,00 | 1,00 | 1,00 | 1,41 | 1,41 | 1,41 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| **P5** | 1,00 | 0,00 | 4,69 | 5,39 | 0,00 | 3,00 | 3,00 | 1,00 | 1,41 | 0,00 | 0,00 | 2,24 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| **P6** | 3,00 | 0,00 | 3,00 | 3,46 | 0,00 | 1,00 | 1,41 | 3,61 | 2,45 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| **P7** | 1,41 | 3,46 | 2,24 | 0,00 | 1,00 | 0,00 | 1,00 | 0,00 | 0,00 | 1,00 | 2,45 | 0,00 | 1,41 | 0,00 | 0,00 | 1,00 | 0,00 |
| **P8** | 1,73 | 1,00 | 1,41 | 2,00 | 0,00 | 0,00 | 0,00 | 0,00 | 2,24 | 2,45 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| **P9** | 4,69 | 4,12 | 2,65 | 2,00 | 1,41 | 2,24 | 0,00 | 0,00 | 0,00 | 1,41 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| **P10** | 3,87 | 1,00 | 2,00 | 2,00 | 0,00 | 3,32 | 3,32 | 1,73 | 1,41 | 1,73 | 0,00 | 1,000 | 2,24 | 0,00 | 1,41 | 0,00 | 0,00 |
| **P11** | 2,65 | 3,00 | 2,00 | 2,00 | 0,00 | 1,00 | 2,24 | 1,00 | 1,41 | 1,00 | 2,24 | 0,00 | 1,41 | 0,00 | 1,00 | 0,00 | 0,00 |

Em que: **E1** = *Casearia decandra* Jacq.; **E2** = *Anadenanthera peregrina* (L.) Speg.; **E3** = *Apuleia leiocarpa* (Vog.) J. F. Macbr.; **E4** = *Mabea fistulifera* Mart.; **E5** = *Anadenanthera macrocarpa* (Benth.) Brenan.; **E6** = *Platypodium elegans* Vog.; **E7** = *Machaerium floridum* (Mart. ex Benth.) Ducke; **E8** = *Copaifera langsdorffii* Desf.; **E9** = *Ocotea pretiosa* (Nees) Mez.; **E10** = *Cabralea cangerana* Saldanha; **E11** = *Piptadenia gonoacantha* (Mart.) J. F. Macbr.; **E12** = *Dalbergia nigra* (Vell.) Allemão ex Benth.; **E13** = *Luehea divaricata* Mart.; **E14** = *Cecropia hololeuca* Miq.; **E15** = *Melanoxylum brauna* Schott.; **E16** = *Cedrela fissilis* Vell.; e **E17** = *Croton floribundus* Spreng.

Para o cálculo da matriz de covariância, é fundamental que os dados estejam centralizados. A centralização é o processo de remover a média de cada variável (nesse caso, cada espécie) do conjunto de dados. Isso garante que cada espécie tenha uma média igual a zero, o que é uma premissa para o cálculo da covariância e para a interpretação correta das relações entre as variáveis (Johnson & Wichern, 2007). A centralização elimina o efeito da magnitude absoluta das abundâncias, focando nas variações relativas de cada espécie em torno de sua própria média.

A centralização de um valor **Xij**, correspondente à abundância da espécie **j** na parcela **i**, é realizada pela seguinte expressão:

$$X_{ij}^{c} = X_{ij} - \overline{X}_{j}$$

Em que: $\mathbf{X_{ij}^{c}}$ representa o valor centralizado da espécie j na parcela **i**; $\mathbf{X_{ij}}$ corresponde ao valor observado após a transformação pela raiz quadrada; e $\mathbf{\overline{X}_{j}}$ representa a média da espécie **j**, considerando todas as parcelas.

Para ilustrar o processo de centralização, utilizam-se os valores transformados da espécie E1:

Vetor transposto de valores transformados de **E1**:

$$(2{,}83\ 1{,}00\ 5{,}20\ 0{,}00,\ 1{,}00\ 3{,}00,\ 1{,}41\ 1{,}732\ 4{,}69\ 3{,}87\ 2{,}65)$$

Soma dos valores:

$$2{,}83 + 1{,}00 + 5{,}20 + 0{,}00 + 1{,}00 + 3{,}00 + 1{,}41 + 1{,}73 + 4{,}69 + 3{,}87 + 2{,}65 = 27{,}38$$

Média da espécie **E1**:

$$\overline{X}_j = \frac{27{,}38}{11} \approx 2{,}49$$

Assim, os valores centralizados são obtidos pela subtração da média de **E1** de cada observação original transformada (Tabela 15).

A covariância é uma medida estatística que quantifica o grau em que duas variáveis variam conjuntamente (Johnson & Wichern, 2007). Em ecologia, a covariância entre as abundâncias de duas espécies pode indicar interações ecológicas (competição, mutualismo), respostas semelhantes a gradientes ambientais ou, simplesmente, coocorrência devido a fatores espaciais ou históricos. A matriz de covariância é um componente essencial para o cálculo da distância de Mahalanobis, pois descreve a estrutura de interdependência entre todas as variáveis do conjunto de dados.

**Tabela 15.** Valores centralizados para espécie **E1** conforme a parcela

| Parcela | Valor Transformado | Média $\overline{X}_1$ | Desvio ($X_{i1}^c$) |
|---|---|---|---|
| P1 | 2,83 | 2,50 | 0,34 |
| P2 | 1,00 | 2,50 | -1,49 |
| P3 | 5,20 | 2,50 | 2,71 |
| P4 | 0,00 | 2,50 | -2,49 |
| P5 | 1,00 | 2,50 | -1,49 |
| P6 | 3,00 | 2,50 | 0,51 |
| P7 | 1,41 | 2,50 | -1,08 |
| P8 | 1,73 | 2,50 | -0,76 |
| P9 | 4,69 | 2,50 | 2,20 |
| P10 | 3,87 | 2,50 | 1,38 |
| P11 | 2,65 | 2,50 | 0,16 |

## Interpretação da Covariância

- Covariância Positiva: Indica que, quando a abundância de uma espécie aumenta, a abundância da outra espécie também tende a aumentar. Isso pode sugerir uma associação positiva ou uma resposta comum a fatores ambientais.
- Covariância Negativa: Sugere que, quando a abundância de uma espécie aumenta, a abundância da outra tende a diminuir. Isso pode indicar competição ou respostas opostas a gradientes ambientais.

- Covariância Próxima de Zero: Implica uma ausência de associação linear entre as abundâncias das duas espécies. É importante notar que uma covariância próxima de zero não significa ausência de qualquer tipo de associação, apenas ausência de associação linear.

A covariância entre duas espécies (variáveis) **X** e **Y** é calculada pela seguinte expressão:

$$cov(X,Y) = \frac{\sum_n^i (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Em que: $\mathbf{X_i}$ e $\mathbf{Y_i}$ representam os valores observados (centralizados) das espécies **X** e **Y** na parcela ***i***; $\overline{\mathbf{X}}$ e $\overline{\mathbf{Y}}$ correspondem às médias das espécies ***X*** e ***Y***, respectivamente, sendo iguais a zero após a centralização; **n** representa o número de parcelas (observações); e **n−1** corresponde ao grau de liberdade utilizado para obter um estimador não viciado da covariância populacional.

Para demonstrar o cálculo da covariância, utilizam-se os vetores centralizados das espécies **E1** e **E3**.

*Vetor centralizado transposto de **E1**:*

$$(0{,}34\ \ -1{,}49\ 2{,}71\ \ -2{,}49\ \ -1{,}49\ 0{,}51\ \ -1{,}08\ \ -0{,}76\ 2{,}20\ 1{,}38\ 0{,}16)$$

*Vetor transposto de valores transformados da espécie **E3**:*

$$(1{,}73\ 3{,}00\ 2{,}00\ 2{,}45\ 4{,}69\ 3{,}00\ 2{,}24\ 1{,}41\ 2{,}65\ 2{,}00\ 2{,}00)$$

*Média da espécie **E3**:*

$$\overline{X}_3 = \frac{1{,}73 + 3{,}00 + 2{,}00 + 2{,}45 + 4{,}69 + 3{,}00 + 2{,}24 + 1{,}41 + 2{,}65 + 2{,}00 + 2{,}00}{11} \approx 2{,}47$$

Assim, o vetor centralizado da espécie **E3** é obtido pela subtração da média $\overline{\boldsymbol{X}}_{\mathbf{3}}$ de cada valor transformado (Tabela 16).

**Tabela 16.** Valores centralizados para a espécie **E3** conforme a parcela

| Parcela | Valor Transformado | Média ($\overline{X}_3$) | Desvio ($X_{i3}^c$) |
|---|---|---|---|
| **P1** | 1,73 | 2,47 | -0,74 |
| **P2** | 3,00 | 2,47 | 0,53 |
| **P3** | 2,00 | 2,47 | -0,47 |
| **P4** | 2,45 | 2,47 | -0,02 |
| **P5** | 4,69 | 2,47 | 2,22 |
| **P6** | 3,00 | 2,47 | 0,53 |
| **P7** | 2,24 | 2,47 | -0,23 |
| **P8** | 1,41 | 2,47 | -1,06 |
| **P9** | 2,65 | 2,47 | 0,18 |
| **P10** | 2,00 | 2,47 | -0,47 |
| **P11** | 2,00 | 2,47 | -0,47 |

*Produto dos desvios*

De forma auxiliar, obtemos a Tabela 17.

**Tabela 17.** Produtos dos desvios de **E1** e **E3**, conforme parcela

| Parcela | Produtos dos desvios ($Xc_{i1} \times Xc_{i3}$) |
|---|---|
| **P1** | 0,34 × (-0,74) = -0,25 |
| **P2** | -1,49 × 0,53 = -0,79 |
| **P3** | 2,71 × (-0,47) = -1,27 |
| **P4** | -2,49 × (-0,02) = 0,05 |
| **P5** | -1,49 × 2,22 = -3,31 |
| **P6** | 0,51 × 0,53 = 0,271 |
| **P7** | -1,08 × (-0,23) = 0,25 |
| **P8** | -0,76 × (-1,06) = 0,80 |
| **P9** | 2,20 × 0,18 = 0,39 |
| **P10** | 1,38 × (-0,47) = -0,65 |
| **P11** | 0,16 × (-0,47) = -0,07 |

*Soma dos produtos dos desvios*

$$\sum_{n}^{i} X_{ci1} x X_{ci3} = -0{,}25 - 0{,}79 - 1{,}27 + 0{,}05 - 3{,}31 + 0{,}27 + 0{,}25 + 0{,}80 + 0{,}39 - 0{,}65 - 0{,}07 = -4{,}55$$

*Covariância entre **E1** e **E3***

$$cov(E_1, E_3) = \frac{\sum_{n}^{i} X_{ci1} x X_{ci3}}{n-1} = \frac{-4{,}55}{11-1} = -0{,}46$$

A variância é um caso particular da covariância, em que a relação é calculada entre uma variável e ela mesma. Ela representa a dispersão dos valores de uma variável em torno de sua média. A variância da espécie **X** é dada por:

$$var(X) = cov(X, X) = \frac{\sum_{n}^{i=1} \left(X_i - \overline{X}\right)^2}{n-1}$$

**Aplicação - var(E1)**

Utilizando os desvios centralizados da espécie **E1**, calculam-se os quadrados dos desvios $(\boldsymbol{X_{i1}^c})^2$ conforme a Tabela 18.

**Tabela 18.** Quadrados dos desvios para espécie **E1** $(X_{i1}^c)^2$, conforme a parcela.

| Parcela | Desvio ($Xc_{i1}$) | Quadrado ($(X_{i1}^c)^2$) |
|---|---|---|
| P1 | 0,34 | 0,13 |
| P2 | -1,49 | 2,23 |
| P3 | 2,71 | 7,34 |
| P4 | -2,49 | 6,20 |
| P5 | -1,49 | 2,22 |
| P6 | 0,51 | 0,26 |
| P7 | -1,08 | 1,16 |
| P8 | -0,76 | 0,57 |
| P9 | 2,20 | 4,84 |
| P10 | 1,38 | 1,92 |
| P11 | 0,16 | 0,03 |

*Soma dos quadrados dos desvios*

$$\sum_n^i (X_{ci1})^2 = 0{,}12 + 2{,}22 + 7{,}33 + 6{,}20 + 2{,}22 + 0{,}26 + 1{,}16 + 0{,}57 + 4{,}84 + 1{,}92 + 0{,}03 = 26{,}85$$

*Variância de **E1***

$$var(E_1) = \frac{\sum_n^i (X_{ci1})^2}{n-1} = \frac{26{,}85}{10-1} = 2{,}69$$

A matriz de covariância ($\Sigma$) é uma matriz quadrada e simétrica, na qual os elementos da diagonal principal correspondem às variâncias de cada espécie, enquanto os elementos fora da diagonal representam as covariâncias entre pares de espécies. Para esse exemplo, contendo 17 espécies, a matriz possui dimensão *17×17*.

A estrutura geral da matriz de covariância é dada por:

$$\Sigma = \begin{bmatrix} var(E1) & cov(E1,E2) & \cdots & cov(E1,E17) \\ cov(E2,E1) & var(E2) & \cdots & cov(E2,E17) \\ \vdots & \vdots & \ddots & \vdots \\ cov(E17,E1) & cov(E17,E2) & \cdots & var(E17) \end{bmatrix}$$

Um trecho da matriz de covariância calculada a partir dos dados transformados e centralizados é apresentado na Tabela 19.

**Tabela 19.** Partição da matriz de covariância ($\Sigma$), considerando as espécies **E1**, **E2**, **E3** e **E4**

| | E1 | E2 | E3 | E4 |
|---|---|---|---|---|
| E1 | 2,69 | 0,88 | -0,46 | -0,31 |
| E2 | 0,88 | 2,01 | 0,17 | -0,30 |
| E3 | -0,46 | 0,17 | 0,92 | 0.64 |
| E4 | -0,31 | -0,30 | 0,64 | 1,54 |

Um dos principais desafios na análise de dados ecológicos multivariados ocorre quando o número de variáveis (espécies, **p**) é maior que o número de observações (parcelas, **n**). No presente exemplo, existem **p** = 17 espécies e **n** = 11 parcelas, caracterizando uma situação de **p** > **n**.

Nessas condições, a matriz de covariância amostral clássica tende a se tornar mal condicionada ou singular. Uma matriz singular é aquela cujo determinante é igual a zero:

$$|\sum| = 0$$

A principal implicação prática dessa singularidade é a impossibilidade de calcular de forma estável a inversa da matriz de covariância:

$$\sum^{-1}$$

A inversa da matriz de covariância é fundamental para o cálculo da distância de Mahalanobis, pois permite padronizar os dados e considerar as correlações existentes entre as variáveis. Quando a matriz é singular ou numericamente instável, sua inversa não pode ser obtida adequadamente, comprometendo a aplicação da distância de Mahalanobis em dados ecológicos com alta dimensionalidade.

Além da alta dimensionalidade, os dados ecológicos apresentam características que intensificam os problemas associados à estimação clássica da matriz de covariância.

- Muitas espécies raras e excesso de zeros: a ocorrência frequente de espécies com baixas abundâncias ou ausentes em grande parte das parcelas resulta em variâncias muito pequenas. Isso pode gerar instabilidade na matriz de covariância, pois pequenas variações passam a exercer influência desproporcional nos cálculos.
- Forte correlação entre espécies: espécies que coexistem ou respondem de maneira semelhante aos fatores ambientais tendem a apresentar elevadas correlações. Embora a distância de Mahalanobis considere a correlação entre variáveis, estimativas imprecisas da covariância em situações de **p** > **n** podem produzir resultados distorcidos, comprometendo a ponderação adequada da redundância entre espécies.
- Baixa repetição amostral: quando o número de parcelas (**n**) é pequeno em relação ao número de espécies (**p**), existe pouca informação disponível para estimar de forma robusta as variâncias e covariâncias independentes da matriz de covariância. Isso aumenta a incerteza das estimativas e favorece a ocorrência de singularidade da matriz.

No presente exemplo, espécies como **E15**, **E16** e **E17** apresentam grande quantidade de valores iguais a zero ou abundâncias muito baixas. Como consequência, suas variâncias tendem

a ser extremamente pequenas, contribuindo para a instabilidade da matriz de covariância amostral.

O método *Shrinkage* (encolhimento ou contração) é uma técnica de regularização utilizada para melhorar a estimação da matriz de covariância, principalmente em situações de alta dimensionalidade (**p** > **n**) ou quando os dados apresentam elevado nível de ruído. A proposta do método consiste em combinar a matriz de covariância amostral, geralmente instável e sensível a pequenas variações, com uma matriz-alvo mais simples, estável e numericamente bem condicionada.

Essa combinação é controlada por um parâmetro de regularização (**λ**), responsável por determinar o grau de encolhimento da matriz de covariância amostral em direção à matriz-alvo. Quanto maior o valor de **λ**, maior será a influência da matriz-alvo na estimativa final da covariância. Dessa forma, o método *shrinkage* reduz problemas de singularidade e instabilidade numérica, produzindo estimativas mais robustas e adequadas para análises multivariadas, como a distância de Mahalanobis.

A estimação da matriz de covariância por *shrinkage* é realizada pela seguinte expressão:

$$\sum shrink = (1 - \lambda)\Sigma + \lambda T$$

Em que: **Σ*shrink*** representa a matriz de covariância regularizada pelo método *shrinkage*; Σ corresponde à matriz de covariância amostral clássica; **T** representa a matriz-alvo (*target matrix*), definida como uma estrutura de covariância mais estável e numericamente bem condicionada; e **λ** é o parâmetro de regularização, assumindo valores no intervalo:

$$0 \leq \lambda \leq 1$$

Quando **λ** está próximo de 0, a matriz resultante permanece mais semelhante à matriz de covariância amostral. Por outro lado, valores de **λ** próximos de 1 aumentam a influência da matriz-alvo, produzindo uma estimativa mais regularizada e estável.

Entre os diferentes métodos de *shrinkage*, o estimador de Ledoit–Wolf se destaca por sua eficiência na estimação de matrizes de covariância em cenários de alta dimensionalidade. Desenvolvido por Ledoit e Wolf (2003), esse método determina automaticamente um valor ótimo para o parâmetro de regularização **λ**, minimizando o erro quadrático médio entre a matriz de covariância estimada e a verdadeira matriz de covariância populacional.

O principal objetivo do método é encontrar um equilíbrio entre viés e variância na estimação da matriz de covariância. Para isso, o procedimento estima, simultaneamente, a intensidade do ruído presente nos dados e a distância entre a matriz de covariância amostral e a matriz-alvo utilizada no processo de regularização.

No método Ledoit–Wolf, a matriz-alvo mais frequentemente adotada é uma matriz diagonal, na qual os elementos da diagonal principal correspondem às variâncias das variáveis, enquanto todos os elementos fora da diagonal são iguais a zero. Essa estrutura assume independência entre as variáveis e contribui para produzir uma matriz de covariância mais estável, invertível e adequada para análises multivariadas em situações de $\mathbf{p} > \mathbf{n}$.

O método *shrinkage*, especialmente o estimador de Ledoit–Wolf, busca resolver o clássico compromisso entre viés e variância na estimação da matriz de covariância (Tabela 20).

**Tabela 20.** Situação e consequência entre viés e variância na estimação de covariância

| Situação | Consequência |
|---|---|
| Pouca regularização ($\boldsymbol{\lambda}$ baixo) | A matriz estimada permanece muito próxima da matriz amostral, podendo apresentar alta variância, sensibilidade ao ruído dos dados e instabilidade numérica, principalmente em situações de $\mathbf{p} > \mathbf{n}$. |
| Excesso de regularização ($\boldsymbol{\lambda}$ alto) | A matriz estimada se aproxima excessivamente da matriz-alvo, podendo introduzir elevado viés e subestimar a estrutura real de correlação entre as variáveis. |

O método Ledoit–Wolf procura determinar automaticamente o valor ótimo de $\boldsymbol{\lambda}$ que minimiza o erro quadrático médio da estimativa. Dessa forma, o método encontra um equilíbrio entre variância e viés, produzindo uma matriz de covariância mais robusta, estável e adequada para análises multivariadas em dados de alta dimensionalidade.

O algoritmo de Ledoit–Wolf estima o parâmetro de regularização $\lambda$ de forma adaptativa, levando em consideração as características dos dados analisados. Para isso, o método avalia: a variabilidade do ruído presente nos dados; a distância entre a matriz de covariância amostral e a matriz-alvo; e a estabilidade numérica necessária para garantir a inversão adequada da matriz de covariância.

*Interpretação do parâmetro $\lambda$*

a. Dados muito ruidosos ou situações em que $\mathbf{p} > \mathbf{n}$: o valor de $\boldsymbol{\lambda}$ tende a ser elevado ($\boldsymbol{\lambda}\uparrow$), indicando necessidade de maior regularização para estabilizar a estimação da covariância.

b. Dados mais estáveis ou cenários em que $\mathbf{n} > \mathbf{p}$: o valor de $\boldsymbol{\lambda}$ tende a ser reduzido ($\boldsymbol{\lambda}\downarrow$), sugerindo que a matriz de covariância amostral já fornece uma estimativa adequada, exigindo pouca regularização.

Suponha que, após a aplicação do método Ledoit–Wolf aos dados ecológicos transformados e centralizados, o valor ótimo estimado para o parâmetro de regularização seja:

$$\lambda = 0{,}4$$

Considerando a covariância clássica entre $\mathbf{E_1}$ e $\mathbf{E_3}$, previamente calculada:

$$\mathrm{cov}(E_1,E_3) = -0{,}46$$

E assumindo uma matriz-alvo **T**, em que os elementos fora da diagonal são iguais a zero ($T_{13}$ = 0), a covariância regularizada por *shrinkage* é calculada da seguinte forma:

$$covshrink\ (E1, E3) = (1 - \lambda)cov(E1, E3) + \ \lambda T13$$

Substituindo os valores:

$$covshrink\ (E1, E3)\ = (1 - 0{,}4)(-0{,}46) + 0{,}4(0)$$

$$covshrink\ (E1, E3) = 0{,}6(-0{,}46) = -0{,}27$$

**Resultado**

Após a regularização por *shrinkage*, a covariância entre $\mathbf{E_1}$ e $\mathbf{E_3}$ passou de −0,46 para −0,27. Esse resultado demonstra que o método reduz a intensidade das correlações extremas, produzindo estimativas mais conservadoras e numericamente estáveis. Esse ajuste é particularmente importante em dados ecológicos de alta dimensionalidade, pois evita que correlações espúrias ou excessivamente influenciadas pelo ruído afetem indevidamente o cálculo da distância de Mahalanobis.

O método *shrinkage* não apenas estabiliza as estimativas da matriz de covariância, mas também altera, de maneira importante, a geometria multivariada dos dados. A matriz de covariância é responsável por definir a forma e a orientação dos elipsóides de densidade de probabilidade no espaço multivariado. Em situações de alta dimensionalidade ($\boldsymbol{p} > \boldsymbol{n}$) ou na presença de dados ruidosos, a matriz de covariância amostral pode produzir elipsóides degenerados ou colapsados, caracterizados por eixos excessivamente longos ou extremamente curtos. Esse comportamento reflete a instabilidade numérica e as dificuldades na estimação da verdadeira estrutura de dispersão dos dados.

Com a aplicação do *shrinkage*, a matriz de covariância regularizada $\boldsymbol{\sum shrink}$ passa a produzir elipsóides multivariados mais estáveis e bem condicionados, com eixos de comprimentos finitos e orientações mais consistentes. Isso permite representar, de forma mais realista, a dispersão dos dados, evitando superestimações de variância em determinadas direções e subestimações em outras.

Como consequência, a distância de Mahalanobis, calculada a partir da matriz regularizada ($\boldsymbol{\sum shrink}$), torna-se mais robusta e confiável, refletindo, de maneira mais adequada, a dissimilaridade estatística entre as observações e incorporando a estrutura de correlação entre as variáveis de forma estável.

Após a aplicação do método *shrinkage*, ou seja, obtenção da matriz **($\boldsymbol{\sum shrink}$)**, o passo seguinte para o cálculo da distância de Mahalanobis consiste na inversão dessa matriz, obtendo-se $\boldsymbol{\sum shrink^{-1}}$.

Vale ressaltar que a regularização promovida pelo método *shrinkage* garante que Σ*shrink* seja uma matriz positiva definida e numericamente bem condicionada. Como consequência, sua inversão pode ser realizada de forma estável e confiável, mesmo em situações nas quais a matriz de covariância amostral clássica seja singular ou apresente forte instabilidade numérica.

Essa propriedade é fundamental para a aplicação da distância de Mahalanobis em dados ecológicos de alta dimensionalidade, pois permite incorporar adequadamente a estrutura de variâncias e correlações entre as espécies no cálculo das dissimilaridades multivariadas.

A inversa da matriz de covariância, também conhecida como matriz de precisão, desempenha um papel fundamental na distância de Mahalanobis. Ela atua como um ponderador, ajustando as distâncias entre os pontos no espaço multivariado. Especificamente, a inversa pondera:

a) Variâncias: Variáveis com maior variância (e, portanto, menor precisão) recebem um peso menor na distância, enquanto variáveis com menor variância (maior precisão) recebem um peso maior.
b) Correlações: A inversa da matriz de covariância remove o efeito das correlações entre as variáveis. Isso significa que espécies altamente correlacionadas recebem um peso menor na distância, pois sua contribuição para a dissimilaridade já é parcialmente capturada por outras espécies correlacionadas. Por outro lado, espécies mais independentes recebem um peso maior, pois fornecem informações únicas sobre a dissimilaridade (Mahalanobis, 1936; De Maesschalck, Jouan-Rimbaud & Massart, 2000).
c) Redundâncias entre Espécies: Ao considerar as correlações, a inversa da matriz de covariância efetivamente reduz a redundância de informação entre as espécies, garantindo que cada dimensão do espaço multivariado contribua de forma independente para a medida de distância.

Em resumo, a matriz inversa regularizada permite que a distância de Mahalanobis leve em conta a estrutura de dispersão e correlação dos dados de forma robusta, fornecendo uma medida de dissimilaridade estatística que é invariante à escala e à correlação das variáveis.

Assim, com a aplicação do método *shrinkage*, a distância de Mahalanobis ($D^2$) entre dois vetores de observações, $\mathbf{x_i}$ e $\mathbf{x_j}$ - neste caso, duas parcelas - é definida pela seguinte expressão:

$$D_{ij}^2 = (x_i - x_i)' \sum shrink^{-1} (x_i - x_i)$$

Em que: $\mathbf{x_i}$ representa o vetor de abundâncias transformadas e centralizadas da parcela ***i***; $\mathbf{x_j}$ representa o vetor de abundâncias transformadas e centralizadas da parcela ***j***; $(\mathbf{x_i} - \mathbf{x_j})$ corresponde ao vetor diferença entre as duas parcelas; $(\mathbf{x_i} - \mathbf{x_j})$' representa a transposta do vetor diferença; e $\sum \boldsymbol{shrink}^{-1}$ corresponde à inversa da matriz de covariância regularizada estimada pelo método *shrinkage*.

***Componentes da fórmula***

a. Vetor da parcela ***i***($x_i$): contém as abundâncias transformadas e centralizadas de todas as espécies na parcela **i**; Vetor da parcela ***j***($x_j$): contém as abundâncias transformadas e centralizadas de todas as espécies na parcela **j**; e inversa regularizada ($\sum \boldsymbol{shrink}^{-1}$): atua como uma matriz de precisão, ponderando, simultaneamente, as variâncias e covariâncias entre as espécies, tornando a distância estatisticamente robusta e adequada para dados multivariados de alta dimensionalidade.

Para ilustrar o cálculo da distância de Mahalanobis, considere a distância entre a Parcela 1 (P1) e a Parcela 2 (P2). Assume-se que os vetores já foram transformados pela raiz quadrada e, posteriormente, centralizados.

Vetores das parcelas:

$x1 = (0{,}34 \ -1{,}49 \ \ 2{,}71 \ -2{,}49 \ -1{,}49 \ \ 0{,}51 \ -1{,}08 \ -0{,}76 \ \ 2{,}20 \ \ 1{,}38 \ \ 0{,}16)$

$x2 = (1{,}00 \ 0{,}00 \ 3{,}00 \ 1{,}73 \ 3{,}46 \ 0{,}00 \ 0{,}00 \ 1{,}00 \ 0{,}00 \ 0{,}00 \ 0{,}00)$

Vetor diferença:

$$d = x_1 - x_2$$

Substituindo os valores:

$d = ((0{,}34 - 1{,}00) \ (-1{,}49 - 0{,}00) \ \ (2{,}71 - 3{,}00) \ (-2{,}49 - 1{,}73) \ (-1{,}49 - 3{,}46) \ (0{,}51 - 0{,}00) \ (-1{,}08 - 0{,}00) \ (-0{,}76 - 1{,}00) \ (2{,}20 - 0{,}000) \ (1{,}384 - 0{,}00) \ (0{,}16 - 0.00))$

Resultando em:

$d = (-0{,}66 \ -1{,}49 \ -0{,}29 \ -4{,}22 \ -4{,}95 \ \ 0{,}51 \ -1{,}075 \ -1{,}76 \ \ 2{,}20 \ \ 1{,}38 \ \ 0{,}16)$

O cálculo da distância de Mahalanobis envolve duas multiplicações matriciais:

1. **Primeira multiplicação**

$$v = \sum shrink^{-1} \ d$$

Nesse passo, o vetor diferença **d** é multiplicado pela inversa da matriz de covariância regularizada estimada. Essa operação padroniza o vetor diferença, considerando simultaneamente as variâncias e covariâncias entre as espécies.

**2. Segunda multiplicação**

$$D^2 = d^T v$$

Em seguida, a transposta do vetor diferença ($\mathbf{d^T}$) é multiplicada pelo vetor resultante **v**. O resultado é um valor escalar correspondente à distância de Mahalanobis ao quadrado entre as duas parcelas.

***Resultado final***

$$D^2(P_1, P_2) = 25{,}04$$

Esse valor representa a dissimilaridade multivariada entre as parcelas P1 e P2, considerando não apenas as diferenças de abundância entre espécies, mas também a estrutura de variâncias e correlações existente nos dados ecológicos.

Ao calcular a distância de Mahalanobis para todos os pares de parcelas, obtemos uma matriz de distâncias. Esta matriz é simétrica, com zeros na diagonal principal (a distância de uma parcela para si mesma é zero). Os valores fora da diagonal representam a dissimilaridade estatística entre cada par de parcelas, considerando a estrutura multivariada dos dados.

Na Tabela 21, apresentamos a matriz final de distâncias D² de Mahalanobis, calculada para todas as 11 parcelas, utilizando a matriz de covariância regularizada por *shrinkage*.

**Tabela 21.** Matriz de distâncias D² de Mahalanobis

| | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** | **P10** | **P11** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **P1** | 0,00 | 25,04 | 20,13 | 19,19 | 30,55 | 19,39 | 26,76 | 13,13 | 28,36 | 27,96 | 24,01 |
| **P2** | 25,04 | 0,00 | 30,64 | 11,61 | 32,42 | 25,18 | 21,37 | 24,94 | 26,14 | 34,44 | 27,32 |
| **P3** | 20,13 | 30,64 | 0,00 | 26,41 | 28,12 | 26,70 | 27,57 | 26,08 | 28,29 | 18,54 | 21,67 |
| **P4** | 19,19 | 11,61 | 26,41 | 0,00 | 21,04 | 15,68 | 17,86 | 9,30 | 26,07 | 20,04 | 16,46 |
| **P5** | 30,55 | 32,42 | 28,12 | 21,04 | 0,00 | 24,33 | 38,72 | 34,01 | 34,98 | 27,31 | 28,07 |
| **P6** | 19,39 | 25,18 | 26,70 | 15,68 | 24,33 | 0,00 | 32,57 | 22,46 | 30,64 | 22,92 | 19,12 |
| **P7** | 26,76 | 21,37 | 27,57 | 17,86 | 38,72 | 32,57 | 0,00 | 21,15 | 19,74 | 28,32 | 9,44 |
| **P8** | 13,13 | 24,94 | 26,08 | 9,30 | 34,01 | 22,46 | 21,15 | 0,00 | 23,25 | 25,64 | 17,45 |
| **P9** | 28,36 | 26,14 | 28,29 | 26,07 | 34,98 | 30,64 | 19,74 | 23,25 | 0,00 | 25,68 | 18,11 |
| **P10** | 27,96 | 34,44 | 18,54 | 20,04 | 27,31 | 22,92 | 28,32 | 25,64 | 25,68 | 0,00 | 15,31 |
| **P11** | 24,01 | 27,32 | 21,67 | 16,46 | 28,07 | 19,12 | 9,44 | 17,45 | 18,11 | 15,31 | 0,00 |

A matriz de distâncias **D² de Mahalanobis** fornece uma base robusta para a interpretação da dissimilaridade entre as parcelas em um contexto ecológico. A interpretação dos valores de D² é direta:

a) Valores Pequenos de **D²**: Indicam que as parcelas são estatisticamente semelhantes em sua composição de espécies, considerando a variância e covariância entre elas. Isso pode sugerir que as parcelas compartilham condições ambientais semelhantes, estão sujeitas a processos ecológicos parecidos ou representam o mesmo tipo de comunidade.
b) Valores Elevados de **D²**: Indicam que as parcelas são floristicamente (ou faunisticamente, dependendo do tipo de dado) distintas. Isso pode ser um indicativo de diferenças significativas nas condições ambientais, perturbações, estágios sucessionais distintos ou presença de comunidades ecológicas fundamentalmente diferentes.

**• Exemplo 1 - Parcelas Semelhantes**

$D^2(P4, P8) = 9{,}30$. Esse valor relativamente baixo sugere que as parcelas P4 e P8 possuem uma composição de espécies estatisticamente semelhante. Apesar de possíveis diferenças nas abundâncias absolutas, a estrutura multivariada das espécies nessas parcelas é bastante parecida, indicando que elas podem pertencer ao mesmo tipo de habitat ou comunidade.

**• Exemplo 2 — Parcelas Muito Diferentes**

$D^2(P5, P7) = 38{,}72$. Esse valor elevado indica que as parcelas P5 e P7 apresentam uma composição multivariada de espécies bastante distinta. Isso pode ser atribuído a diferenças marcantes em fatores ambientais, como tipo de solo, umidade, exposição à luz, ou processos ecológicos que levaram a comunidades com estruturas e composições de espécies muito diferentes. **passo no R**

**Passo a passo no R: Distância D² de Mahalanobis**

A Distância D² de Mahalanobis é uma medida de dissimilaridade que leva em conta a correlação entre as variáveis (espécies) e suas variâncias. Ela é útil para identificar o quão diferentes duas parcelas são, considerando a estrutura de covariância dos dados. Essa medida é particularmente robusta para dados multivariados.

**1. Instalando e Carregando o Pacote Necessário**

Para estimar a matriz de covariância regularizada, que é um passo importante para a Distância de Mahalanobis, utilizaremos o pacote *corpcor*.

```
# Instalar o pacote 'corpcor' (execute apenas se não tiver instalado)
# install.packages("corpcor")

# Carregar o pacote 'corpcor' para utilizar suas funções
library(corpcor)
```

### 2. Organizando e Transformando a Matriz de Abundância

Nossa matriz X contém as abundâncias das espécies, com as parcelas nas colunas e as espécies nas linhas. Para os cálculos subsequentes, é mais conveniente ter as parcelas nas linhas. Além disso, aplicaremos uma transformação pela raiz quadrada nos dados para reduzir o efeito de valores de abundância muito altos, tornando a análise mais robusta.

```
# Transpomos a matriz X para que as parcelas fiquem nas linhas e as espécies nas colunas
X_transp <- t(X)

# Aplicamos a transformação pela raiz quadrada para reduzir o efeito de abundâncias elevadas
X_transf <- sqrt(X_transp)
```

### 3. Estimando a Matriz de Covariância Regularizada

A matriz de covariância é um componente chave da Distância de Mahalanobis. Para garantir a estabilidade dos cálculos, especialmente com um número limitado de amostras, utilizaremos o método de *shrinkage* de Ledoit-Wolf, implementado na função *cov.shrink()*.

```
# Estimamos a matriz de covariância regularizada (sigma_shrink) a partir de X_transf
sigma_shrink <- cov.shrink(X_transf)
```

```
Estimating optimal shrinkage intensity lambda.var (variance vector): 0.5044
Estimating optimal shrinkage intensity lambda (correlation matrix): 0.7215
```

### 4. Invertendo a Matriz de Covariância

A inversa da matriz de covariância é um elemento essencial na fórmula da Distância de Mahalanobis. A função *solve()* do R é utilizada para esse propósito**.**

```
# Invertemos a matriz de covariância regularizada
sigma_inv <- solve(sigma_shrink)
```

### 5. Calculando a Matriz de Distâncias D² de Mahalanobis

Agora, calcularemos a Distância D² de Mahalanobis para todas as combinações de parcelas. A fórmula envolve a diferença entre os vetores de abundância transformados das parcelas e a inversa da matriz de covariância.

```
# Calculamos a matriz D² de Mahalanobis entre todas as parcelas
matriz_D2 <- outer(
  1:nrow(X_transf), # Itera sobre as linhas (parcelas) da matriz X_transf
  1:nrow(X_transf), # Itera novamente sobre as linhas (parcelas) da matriz X_transf
```

```
  Vectorize(function(i, j) {

    # Calcula a diferença entre os vetores de abundância transformados das parcelas i e j
    d <- X_transf[i, ] - X_transf[j, ]

    # Aplica a fórmula da Distância de Mahalanobis: d' * sigma_inv * d
    # t(d) é a transposta do vetor d
    dist_mahalanobis <- t(d) %*% sigma_inv %*% d

    return(dist_mahalanobis)
  })
)

# Atribuímos os nomes das parcelas às linhas e colunas da matriz resultante
dimnames(matriz_D2) <- list(
  rownames(X_transf),
  rownames(X_transf)
)

# Exibimos a matriz de distâncias D² de Mahalanobis, arredondada para duas casas decimais
print(round(matriz_D2, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 25.04 | 20.13 | 19.19 | 30.55 | 19.39 | 26.76 | 13.13 | 28.36 | 27.96 | 24.01 |
| P2 | 25.04 | 0.00 | 30.64 | 11.61 | 32.42 | 25.18 | 21.37 | 24.94 | 26.14 | 34.44 | 27.32 |
| P3 | 20.13 | 30.64 | 0.00 | 26.41 | 28.12 | 26.70 | 27.57 | 26.08 | 28.29 | 18.54 | 21.67 |
| P4 | 19.19 | 11.61 | 26.41 | 0.00 | 21.04 | 15.68 | 17.86 | 9.30 | 26.07 | 20.04 | 16.46 |
| P5 | 30.55 | 32.42 | 28.12 | 21.04 | 0.00 | 24.33 | 38.72 | 34.01 | 34.98 | 27.31 | 28.07 |
| P6 | 19.39 | 25.18 | 26.70 | 15.68 | 24.33 | 0.00 | 32.57 | 22.46 | 30.64 | 22.92 | 19.12 |
| P7 | 26.76 | 21.37 | 27.57 | 17.86 | 38.72 | 32.57 | 0.00 | 21.15 | 19.74 | 28.32 | 9.44 |
| P8 | 13.13 | 24.94 | 26.08 | 9.30 | 34.01 | 22.46 | 21.15 | 0.00 | 23.25 | 25.64 | 17.45 |
| P9 | 28.36 | 26.14 | 28.29 | 26.07 | 34.98 | 30.64 | 19.74 | 23.25 | 0.00 | 25.68 | 18.11 |
| P10 | 27.96 | 34.44 | 18.54 | 20.04 | 27.31 | 22.92 | 28.32 | 25.64 | 25.68 | 0.00 | 15.31 |
| P11 | 24.01 | 27.32 | 21.67 | 16.46 | 28.07 | 19.12 | 9.44 | 17.45 | 18.11 | 15.31 | 0.00 |

## 1.4. Relação entre Medidas de Distância e de Similaridade

O principal objetivo da análise de agrupamento é obter uma representação dos dados por meio da construção de agrupamentos que tenham relevância no mundo real, isto é, que não sejam puramente teóricos. Para tal, os algoritmos existentes utilizam a matriz de dados **X** (**n** x **p**) e conjuntos de pontos em $\mathbf{I_p}$, o que permite calcular os valores de uma função de agrupamento. Normalmente, esses valores são apresentados por uma matriz de similaridade ou de distância (**n** x **n**).

A diferença principal entre os dois tipos de medidas é que os coeficientes de similaridade assumem valores entre "**0** e **1**", enquanto as medidas de distância podem assumir qualquer valor positivo.

A transformação apropriada de uma medida de distância no correspondente coeficiente de similaridade pode ser feita, segundo Everitt et al. (2011), pela seguinte relação:

$$S_{ij} = \frac{1}{1 + d_{ij}}$$

Em que: $S_{ij}$ é o coeficiente de similaridade entre as parcelas **i** e **j**; e $\mathbf{d_{ij}}$ é a correspondente medida de distância estatística entre as parcelas **i** e **j**.

É importante ressaltar que, para determinar uma medida de distância a partir do coeficiente de similaridade, é necessário que a nova medida satisfaça aos axiomas do espaço métrico, principalmente ao axioma denominado de desigualdade triangular.

A seguinte expressão, apresentada por Gower (1967), é a fórmula conveniente de expressar o coeficiente de similaridade na forma de distância:

$$d_{ij} = \left(1 - S_{ij}\right)^{1/2}$$

Em que: $d_{ij}$ é a medida de distância e $S_{ij}$ é o correspondente coeficiente de similaridade.

## Aplicação

A partir da matriz de distância euclidiana e da aplicação da fórmula de Everitt et al. (2011), obtemos a seguinte matriz de similaridade:

$S_{12} = \frac{1}{1+17,8} = 0,05$ e $S_{1011} = \frac{1}{1+17,55} = 0,05$

$$\mathbf{S} = \begin{bmatrix}
1,00 & 0,05 & 0,04 & 0,07 & 0,03 & 0,05 & 0,05 & 0,08 & 0,04 & 0,05 & 0,06 \\
 & 1,00 & 0,03 & 0,08 & 0,03 & 0,04 & 0,05 & 0,06 & 0,03 & 0,04 & 0,05 \\
 & & 1,00 & 0,03 & 0,02 & 0,04 & 0,03 & 0,03 & 0,04 & 0,05 & 0,04 \\
 & & & 1,00 & 0,03 & 0,05 & 0,06 & 0,12 & 0,03 & 0,04 & 0,07 \\
 & & & & 1,00 & 0,03 & 0,03 & 0,03 & 0,02 & 0,03 & 0,03 \\
 & & & & & 1,00 & 0,04 & 0,05 & 0,03 & 0,04 & 0,05 \\
 & & & & & & 1,00 & 0,06 & 0,04 & 0,04 & 0,10 \\
 & & & & & & & 1,00 & 0,04 & 0,05 & 0,07 \\
 & & & & & & & & 1,00 & 0,04 & 0,05 \\
 & & & & & & & & & 1,00 & 0,05 \\
 & & & & & & & & & & 1,00
\end{bmatrix}$$

Já a partir da matriz de similaridade S, obtida por meio do cálculo do coeficiente de comunidade de Jaccard, e do uso da fórmula de Gower (1967), obtém-se a seguinte matriz de distância (D):

$d_{12} = (1 - S_{12})^{1/2} = (1 - 0,44)^{1/2} = 0,75$ e

$d_{1011} = (1 - S_{1011})^{1/2} = (1 - 0,85)^{1/2} = 0,39$

$$\mathbf{D} = \begin{bmatrix} 0,00 & 0,75 & 0,76 & 0,74 & 0,63 & 0,63 & 0,89 & 0,57 & 0,82 & 0,68 & 0,75 \\ & 0,00 & 0,85 & 0,75 & 0,75 & 0,75 & 0,85 & 0,82 & 0,75 & 0,83 & 0,83 \\ & & 0,00 & 0,76 & 0,55 & 0,67 & 0,84 & 0,82 & 0,83 & 0,62 & 0,71 \\ & & & 0,00 & 0,63 & 0,63 & 0,83 & 0,80 & 0,74 & 0,68 & 0,68 \\ & & & & 0,00 & 0,47 & 0,89 & 0,80 & 0,82 & 0,57 & 0,68 \\ & & & & & 0,00 & 0,89 & 0,71 & 0,82 & 0,68 & 0,68 \\ & & & & & & 0,00 & 0,82 & 0,67 & 0,77 & 0,71 \\ & & & & & & & 0,00 & 0,71 & 0,73 & 0,73 \\ & & & & & & & & 0,00 & 0,75 & 0,68 \\ & & & & & & & & & 0,00 & 0,39 \\ & & & & & & & & & & 0,00 \end{bmatrix}$$

Medidas de distância são números reais e positivos que satisfazem aos axiomas do espaço métrico, portanto, constituem um espaço métrico.

Um valor **próximo de zero** para medida de distância entre os elementos $\mathbf{X_i}$ e $\mathbf{X_j}$ indica proximidade ou alta similaridade. Os coeficientes de similaridade pertencem ao intervalo [0;1], sendo números reais positivos; um **valor próximo de um** indica proximidade ou alta similaridade.

Tanto as medidas de distância quanto os coeficientes são apresentados na forma de uma matriz triangular superior denominada, respectivamente, matriz de distância (**D**) e de similaridade (**S**).

**Passo a passo no R: Conversões entre Distância e Similaridade**

**1. Convertendo Distância para Similaridade (Fórmula de Everitt)**

Podemos transformar uma medida de distância (**d**) em uma medida de similaridade (**S**) usando a fórmula de Everitt: **S** = 1 / (1 + **d**). Quanto maior a distância, menor será a similaridade resultante.

```
# Exemplo de uma distância hipotética
d_exemplo <- 17.8

# Aplicamos a fórmula de Everitt para converter a distância em similaridade
s_everitt <- 1 / (1 + d_exemplo)

# Exibimos o resultado, arredondado para duas casas decimais
print(round(s_everitt, 2))
```

```
[1] 0.05
```

**2 Convertendo Similaridade para Distância (Gower, 1967)**

Para fazer o caminho inverso, ou seja, transformar uma medida de similaridade (S) em uma medida de distância (d), podemos usar a fórmula de Gower: d = sqrt(1 - S). Utilizaremos

a matriz_similaridade_jaccard (calculada na seção 1.3.1, item 3) como exemplo para esta conversão.

```
# Assumimos que a matriz_similaridade_jaccard já foi calculada.
# Se não, você precisaria calculá-la primeiro.
# Exemplo: matriz_similaridade_jaccard <- 1 - as.matrix(vegdist(t(X_binaria), method = "jaccard", binary = TRUE))

# Aplicamos a fórmula de Gower para converter a similaridade de Jaccard em distância
matriz_dist_gower_jaccard <- sqrt(
  1 - matriz_similaridade_jaccard
)

# Exibimos a matriz de distâncias resultante, arredondada para duas casas decimais
print(round(matriz_dist_gower_jaccard, 2))
```

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0.00 | 0.75 | 0.76 | 0.74 | 0.63 | 0.63 | 0.89 | 0.58 | 0.82 | 0.68 | 0.76 |
| P2 | 0.75 | 0.00 | 0.85 | 0.75 | 0.75 | 0.75 | 0.85 | 0.82 | 0.75 | 0.83 | 0.83 |
| P3 | 0.76 | 0.85 | 0.00 | 0.76 | 0.55 | 0.67 | 0.85 | 0.82 | 0.83 | 0.62 | 0.71 |
| P4 | 0.74 | 0.75 | 0.76 | 0.00 | 0.63 | 0.63 | 0.83 | 0.80 | 0.74 | 0.68 | 0.68 |
| P5 | 0.63 | 0.75 | 0.55 | 0.63 | 0.00 | 0.47 | 0.89 | 0.80 | 0.82 | 0.58 | 0.68 |
| P6 | 0.63 | 0.75 | 0.67 | 0.63 | 0.47 | 0.00 | 0.89 | 0.71 | 0.82 | 0.68 | 0.68 |
| P7 | 0.89 | 0.85 | 0.85 | 0.83 | 0.89 | 0.89 | 0.00 | 0.82 | 0.67 | 0.71 | 0.71 |
| P8 | 0.58 | 0.82 | 0.82 | 0.80 | 0.80 | 0.71 | 0.82 | 0.00 | 0.71 | 0.73 | 0.73 |
| P9 | 0.82 | 0.75 | 0.83 | 0.74 | 0.82 | 0.82 | 0.67 | 0.71 | 0.00 | 0.76 | 0.39 |
| P10 | 0.68 | 0.83 | 0.62 | 0.68 | 0.58 | 0.68 | 0.77 | 0.73 | 0.76 | 0.00 | 0.39 |
| P11 | 0.76 | 0.83 | 0.71 | 0.68 | 0.68 | 0.68 | 0.71 | 0.73 | 0.68 | 0.39 | 0.00 |

## 1.5. Seleção, Escala e Ponderação das Variáveis

Uma das questões básicas no estudo de análise de agrupamento diz respeito ao número e tipo de variáveis a serem avaliadas, pois a maior ou menor extensão de similaridade depende não somente das variáveis avaliadas, mas também do seu número. Não existem bases teóricas para determinar o número de variáveis a serem medidas, sendo essa questão inerente ao campo de aplicação do problema em estudo. Normalmente, variáveis que assumem praticamente o mesmo valor para todas as unidades amostrais são pouco discriminatórias, e sua inclusão pouco contribui para a determinação da estrutura do agrupamento. Por outro lado, a inclusão de variáveis com grande poder de discriminação, porém irrelevantes ao problema, pode mascarar os grupos e levar a resultados equivocados. Assim, as variáveis a serem preservadas em uma análise de agrupamento deverão ser apenas aquelas que representam a estrutura fundamental do sistema em estudo, devendo, ainda, ser suficientemente diversas para representar, no mínimo, as dimensões mais importantes do sistema (Adams; Wiersma, 1978; Bussab et al., 1990; Manly; Navarro Alberto, 2019).

Outros dois aspectos importantes são ressaltados por Bussab et al. (1990). O primeiro a ser considerado é a homogeneidade entre variáveis. Assim, ao se agruparem observações, é

necessário combinar todas as variáveis em um único índice de similaridade, de forma que a contribuição de cada variável dependa tanto da sua escala de mensuração quanto da escala das demais variáveis. Dessa forma, visando reduzir o efeito de escalas diferentes, surgiram várias propostas de relativização de variáveis. Entretanto, recomenda-se que a escala das variáveis seja definida por meio de transformações sugeridas pelo bom senso e pela área de conhecimento. O segundo aspecto se refere à ponderação das variáveis, já que elas não têm a mesma importância para um determinado problema. De modo geral, os pesos são atribuídos de forma subjetiva e a sua escolha também depende do contexto e do grau de conhecimento que o pesquisador tem do problema. Como a análise de agrupamento é uma técnica exploratória que busca, a partir dos dados, a formulação de hipóteses, o mais comum é atribuir o mesmo peso para todas as variáveis.

# CAPÍTULO 2

## Classificação dos Processos de Agrupamento

Com a realização da primeira etapa da análise de agrupamento, ou seja, a estimativa de uma medida de similaridade ou de distância entre as unidades amostrais, a etapa seguinte é a adoção de uma técnica de agrupamento para formação dos grupos. Para realização dessa tarefa, existem muitos métodos disponíveis, dos quais o pesquisador tem de decidir qual o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções. Os principais métodos são classificados em três grupos: de hierarquização (aglomerativos ou divisivos), de partição ou otimização (agrupamentos mutuamente exclusivos) e de modelos de misturas finitas (permitem agrupamentos sobrepostos) (Everitt; Dunn, 2001; Everitt et al., 2011).

### 2.1. Técnicas de Hierarquização

Na análise de agrupamento, todos os processos de hierarquização são similares, iniciando-se pela determinação dos valores da função de agrupamento (medida de similaridade ou de distância), apresentada em uma matriz de similaridade (**S**) ou de distância (**D**). Procede-se a uma série de fusões dos **n** elementos em **g** grupos ou agrupamentos, com base em um critério de agrupamento, repetindo-se o processo até que todos os **n** elementos sejam alocados em um único agrupamento.

Nos processos de hierarquização, utilizam-se critérios de agrupamento inteiramente baseados na matriz de valores da função de agrupamento. As diferenças básicas entre os diversos procedimentos estão na definição do critério, isto é, na utilização da medida de distância ou de similaridade entre grupos. Esses processos têm como ideias básicas a coesão interna dos objetos e isolamento externo entre grupos.

Os métodos hierárquicos são, também, divididos em métodos aglomerativos e divisivos. Entre os métodos aglomerativos, citam-se o do "**vizinho mais próximo**" ("**Single Linkage Method**"), do "**vizinho mais distante**" ("**Complete Linkage Method**"), da "**ligação média**" ("**Average Linkage**"), do "**centroide**", da "**variância mínima**" e o proposto por **Ward** (1963). Entre os métodos divisivos, o mais conhecido é o de **Edwards e Cavalli-Sforza** (1965)".

## 2.1.1. Algoritmos de Agrupamento Hierárquicos

O agrupamento hierárquico pode ser dividido em dois tipos principais: aglomerativo e divisivos, também conhecidos, respectivamente, como AGNES (*Agglomerative Nesting*) e DIANA (*Divise Analysis*).

Nos algoritmos aglomerativos, consideramos, inicialmente, que cada objeto é um grupo. No decorrer da análise, com base em critérios de similaridade, os grupos vão sendo formados até que um corte seja feito para finalizarmos o processo de agrupamento, ou seja, a definição do número de grupos adequados para a massa de dados analisada.

Nos algoritmos divisivos, consideramos, inicialmente, um grupo formado por todos os objetos. No decorrer da análise, conforme os critérios de similaridade, vamos dividindo o grande grupo até o número de grupos considerado ideal para a massa de dados em estudo.

## 2.1.1.1. Métodos Hierárquicos Aglomerativos

Os métodos hierárquicos aglomerativos são os mais comumente apresentados na literatura sobre análise de agrupamento (Johnson; Wichern, 2007; Malhotra, 2019; Mardia et al., 2023).

**a. Método da Ligação Simples (*Single Linkage*)**

Esse algoritmo, também denominado de método do elemento mais próximo ("**Neighbourhoods**"), é um dos mais simples, sendo de uso geral e de rápida aplicação.

O método da ligação simples é uma técnica de hierarquização aglomerativa e tem como uma de suas características não exigir que o número de agrupamentos seja fixado "*a priori"* (Orlóci, 1978; Gama, 2022; Mardia et al., 2023).

Seja $\mathbf{E} = \{E_1, E_2, ..., E_p\}$ um conjunto de elementos em que cada um é representado por um vetor $X_i$, para **i** = 1, 2, ..., **p** pontos do espaço **p**-dimensional ($\mathbf{I_p}$). No caso de análise de vegetação, cada dimensão do espaço corresponde a uma espécie diferente. Então, qualquer medida de distância estatística ou de similaridade pode ser empregada nesse algoritmo.

Suponha que tenham sido determinados todos os $\mathbf{n(n-1)/2}$ diferentes valores de $\mathbf{d_{ij}}$ ou $\mathbf{S_{ij}}$ (i = j = 1, 2, ... , n), representados na forma de uma matriz de distância ($\mathbf{D_1}$) ou de similaridade ($\mathbf{S_1}$).

De posse da matriz primária de dados **X** (**n** x **p**), o método de ligação simples pode ser resolvido na seguinte sequência de cálculos:

1. Com base na matriz de dados **X** (**n** x **p**), determinam-se os valores da função de agrupamento $\mathbf{d_{ij}}$ ou $\mathbf{S_{ij}}$, que devem ser representados na forma matricial ($\mathbf{D_1}$) ou ($\mathbf{S_1}$).

2. Localiza-se o valor mínimo de $\mathbf{d_{ij}} > 0$ ou o valor máximo de $\mathbf{S_{ij}} > 0$. Os elementos $\mathbf{E_i}$ e $\mathbf{E_j}$, correspondentes a esse valor, são reunidos em um mesmo grupo e, então, têm-se (**n-1**) agrupamentos remanescentes.

3. Com base na matriz de distância inicial ($\mathbf{D_1}$) ou de similaridade ($\mathbf{S_1}$), determina-se a distância ou similaridade entre o novo agrupamento, por meio da relação:

$d_{(i,j)l} = min(d_{il}, d_{jl}), l = 1, (n-2)$ para $l \neq i \neq j$

$S_{(i,j)l} = náx(S_{il}, S_{jl}), l = 1, (n-2)$ para $l \neq i \neq j$

E se constrói uma nova matriz de distância ($\mathbf{D_2}$) ou de similaridade ($\mathbf{S_2}$).

4. Localiza-se, em $\mathbf{D_2}$ ou em $\mathbf{S_2}$, o menor valor de $\mathbf{d_{ij}} > 0$ ou o valor máximo de $\mathbf{S_{ij}} > 0$. Em seguida, agrupam-se os elementos que deram origem a essa nova distância ou similaridade, formando-se novo agrupamento. Neste passo, têm-se **(n-2)** agrupamentos.

5. Compõe-se uma nova matriz de distâncias ou de similaridade, baseando-se na matriz de distância ou de similaridade anterior. Para isso, calcula-se a distância ou a similaridade entre o agrupamento formado na etapa anterior e os demais, considerando-se um elemento isolado de **E** como um agrupamento.

A distância entre dois agrupamentos **A** e **B** é dada por $d_{(\mathbf{A},\mathbf{B})} = \min d(X_v, X_r)$, $v = 1, n_A$ e $r = 1, n_B$, sendo $X_v$ e $X_r$ vetores ponto de espaço **p**-dimensional dos elementos de **A** e **B**, respectivamente. No caso de similaridade, $S_{(\mathbf{A},\mathbf{B})} = \text{máx } S(X_v, X_n)$. Retorna-se, a seguir, à etapa 4.

O processo é repetido até que todos **n** elementos de **E** tenham sido alocados em um só agrupamento.

É importante ressaltar que qualquer uma das medidas de distância ou de similaridade pode ser usada nesse algoritmo, porém, quando se trata de dados quantitativos, utiliza-se a distância euclidiana.

**Aplicação**

A partir da matriz de distância euclidiana ($\mathbf{D_1}$), calculada no item 1.3.2.1, letra a, será aplicado o método de ligações simples visando o agrupamento das 11 parcelas.

$\mathbf{D_1} =$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 12,5 | 33,9 | 17,3 | 18,5 | 11,3 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 12,0 | 34,4 | 21,8 | 18,3 | 16,2 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 29,6 | 41,8 | 27,0 | 30,9 | 27,9 | 22,4 | 17,9 | 24,2 |
| 4 | | | | 0,00 | 32,3 | 17,3 | 14,7 | **7,7** | 28,3 | 21,5 | 13,5 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 35,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 19,0 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 15,5 | 22,4 | 24,4 | 8,8 |
| 8 | | | | | | | | 0,00 | 26,7 | 21,2 | 13,3 |
| 9 | | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | | 0,00 |

Os seguintes passos são considerados:

Passo 1: Objetos mais similares – parcelas 4 e 8.

Distância entre os objetos: 7,7.

Cálculo das distâncias entre o grupo (4,8) e os demais objetos:

$d_{(\mathbf{4,8})1} = \min [d_{(1,4)}; d_{(1,8)}] = 11,3;$

$d_{(\mathbf{4,8})2} = \min [d_{(2,4)}; d_{(2,8)}] = 12,0;$

$d_{(\mathbf{4,8})3} = \min [d_{(3,4)}; d_{(3,8)}] = 27,9;$

$d_{(\mathbf{4,8})5} = \min [d_{(4,5)}; d_{(5,8)}] = 32,3;$

$d_{(\mathbf{4,8})6} = \min [d_{(4,6)}; d_{(6,8)}] = 17,3;$

$d_{(\mathbf{4,8})7} = \min [d_{(4,7)}; d_{(7,8)}] = 14,7;$

$d_{(\mathbf{4,8})9} = \min [d_{(4,9)}; d_{(8,9)}] = 26,7;$

$d_{(\mathbf{4,8})10} = \min [d_{(4,10)}; d_{(8,10)}] = 21,2;$ e

$d_{(\mathbf{4,8})11} = \min [d_{(4,11)}; d_{(8,11)}] = 13,3.$

Nova Matriz de Distância: $D_2$

$\mathbf{D_2} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 11,3 | 33,9 | 17,3 | 18,5 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 12,0 | 34,4 | 21,8 | 18,3 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 27,9 | 41,8 | 27,0 | 30,9 | 22,4 | 17,9 | 24,2 |
| (4,8) | | | | 0,00 | 32,3 | 17,3 | 14,7 | 26,7 | 21,2 | 13,3 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 22,4 | 24,4 | **8,8** |
| 9 | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | 0,00 |

Passo 2: Objetos mais similares – parcelas 7 e 11.

Distância entre os objetos: 8,8.

Cálculo das distâncias entre o grupo (7,11) e os demais objetos:

$d_{(7,\mathbf{11})1} = \min [d_{(1,7)}; d_{(1,11)}] = 14{,}7;$

$d_{(7,\mathbf{11})2} = \min [d_{(2,7)}; d_{(2,11)}] = 18{,}3;$

$d_{(7,\mathbf{11})3} = \min [d_{(3,7)}; d_{(3,11)}] = 24{,}2;$

$d_{(7,\mathbf{11})(4,8)} = \min [d_{(4,7,8)}; d_{(4,8,11)}] = 13{,}3;$

$d_{(7,\mathbf{11})5} = \min [d_{(5,7)}; d_{(5,11)}] = 34{,}7;$

$d_{(7,\mathbf{11})6} = \min [d_{(6,7)}; d_{(6,11)}] = 19{,}4;$

$d_{(7,\mathbf{11})9} = \min [d_{(7,9)}; d_{(9,11)}] = 19{,}2;$ e

$d_{(7,\mathbf{11})10} = \min [d_{(7,10)}; d_{(10,11)}] = 17{,}5.$

Nova Matriz de Distância: $D_3$

$\mathbf{D_3} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | **11,3** | 33,9 | 17,3 | 14,7 | 24,8 | 20,0 |
| 2 | | 0,00 | 31,7 | 12,0 | 34,4 | 21,8 | 18,3 | 29,4 | 25,6 |
| 3 | | | 0,00 | 27,9 | 41,8 | 27,0 | 24,2 | 22,4 | 17,9 |
| (4,8) | | | | 0,00 | 32,3 | 17,3 | 13,3 | 26,7 | 21,2 |
| 5 | | | | | 0,00 | 28,7 | 34,7 | 41,4 | 34,8 |
| 6 | | | | | | 0,00 | 19,4 | 27,6 | 21,5 |
| (7,11) | | | | | | | 0,00 | 19,2 | 17,5 |
| 9 | | | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | | | 0,00 |

Passo 3: Objetos mais similares – parcela 1 e grupo (4,8).

Distância entre os objetos: 11,3.

Cálculo das distâncias entre o grupo (1,4,8) e os demais objetos:

$d_{(\mathbf{1,4,8})2} = \min [d_{(1,2)}; d_{(2,4,8)}] = 12{,}0;$

$d_{(\mathbf{1,4,8})3} = \min [d_{(1,3)}; d_{(3,4,8)}] = 23{,}0;$

$d_{(\mathbf{1,4,8})5} = \min [d_{(1,5)}; d_{(4,5,8)}] = 28{,}7;$

$d_{(\mathbf{1,4,8})6} = \min [d_{(1,6)}; d_{(4,6,8)}] = 17{,}3;$

$d_{(\mathbf{1,4,8})(7\text{-}11)} = \min [d_{(1,7,11)}; d_{(4,7,811)}] = 13{,}3;$

$d_{(\mathbf{1,4,8})9} = \min [d_{(1,9)}; d_{(4,8,9)}] = 24{,}8;$ e

$d_{(\mathbf{1,4,8})10} = \min [d_{(1,10)}; d_{(4,8,10)}] = 20{,}0.$

Nova Matriz de Distância: $D_4$

$\mathbf{D_4} =$

| | (1,4,8) | 2 | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| (1,4,8) | 0,00 | **12,0** | 23,0 | 28,7 | 17,3 | 13,3 | 24,8 | 20,0 |
| 2 | | 0,00 | 31,7 | 34,4 | 21,8 | 18,3 | 29,4 | 25,6 |
| 3 | | | 0,00 | 41,8 | 27,0 | 24,2 | 22,4 | 17,9 |
| 5 | | | | 0,00 | 28,7 | 34,7 | 41,4 | 34,8 |
| 6 | | | | | 0,00 | 19,4 | 27,6 | 21,5 |
| (7,11) | | | | | | 0,00 | 19,2 | 17,5 |
| 9 | | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | | 0,00 |

Passo 4: Objetos mais similares – parcela 2 e grupo (1,4,8).

Distância entre os objetos: 12,0.

Cálculo das distâncias entre o grupo (1,2,4,8) e os demais objetos:

$d_{(\mathbf{1,2,4,8})3} = \min [d_{(2,3)}; d_{(1,3,4,8)}] = 23,0;$

$d_{(\mathbf{1,2,4,8})5} = \min [d_{(2,5)}; d_{(1,4,5,8)}] = 28,7;$

$d_{(\mathbf{1,2,4,8})6} = \min [d_{(2,6)}; d_{(1,4,6,8)}] = 17,3;$

$d_{(\mathbf{1,2,4,8})(7,11)} = \min [d_{(2,7,11)}; d_{(1,4,7,8,11)}] = 13,3;$

$d_{(\mathbf{1,2,4,8})9} = \min [d_{(2,9)}; d_{(1,4,8,9)}] = 24,8;$ e

$d_{(\mathbf{1,2,4,8})10} = \min [d_{(2,10)}; d_{(1,4,8,10)}] = 20,0.$

Nova Matriz de Distância: $D_5$

$$\mathbf{D_5} = \begin{array}{c} \\ (1,2,4,8) \\ 3 \\ 5 \\ 6 \\ (7,11) \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccccc} (1,2,4,8) & 3 & 5 & 6 & (7,11) & 9 & 10 \end{array} \\ \left[\begin{array}{ccccccc} 0,00 & 23,0 & 28,7 & 17,3 & \mathbf{13,3} & 24,8 & 20,0 \\ & 0,00 & 41,8 & 27,0 & 24,2 & 22,4 & 17,9 \\ & & 0,00 & 28,7 & 34,7 & 41,4 & 34,8 \\ & & & 0,00 & 19,4 & 27,6 & 21,5 \\ & & & & 0,00 & 19,2 & 17,5 \\ & & & & & 0,00 & 22,8 \\ & & & & & & 0,00 \end{array}\right] \end{array}$$

Passo 5: Objetos mais similares – grupos (7,11) e (1,2,4,8).

Distância entre grupos: 13,3.

Cálculo das distâncias entre o grupo (1,2,4,7,8,11) e os demais objetos:

$d_{(\mathbf{1,2,4,7,8,11})3} = \min [d_{(1,2,3,4,8)}; d_{(3,7,11)}] = 23,0;$

$d_{(\mathbf{1,2,4,7,8,11})5} = \min [d_{(1,2,4,5,8)}; d_{(5,7,11)}] = 28,7;$

$d_{(\mathbf{1,2,4,7,8,11})6} = \min [d_{(1,2,4,6,8)}; d_{(6,7,11)}] = 17,3;$

$d_{(\mathbf{1,2,4,7,8,11})9} = \min [d_{(1,2,4,8,9)}; d_{(7,9,11)}] = 19,2;$ e

$d_{(\mathbf{1,2,4,7,8,11})10} = \min [d_{(1,2,4,8,10)}; d_{(7,10,11)}] = 17,5.$

Nova Matriz de Distância: $D_6$

$$\mathbf{D_6} = \begin{array}{c} \\ (1,2,4,7,8,11) \\ 3 \\ 5 \\ 6 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccccc} (1,2,4,7,8,11) & 3 & 5 & 6 & 9 & 10 \end{array} \\ \left[\begin{array}{ccccccc} 0,00 & 23,0 & 28,7 & \mathbf{17,3} & 19,2 & 17,5 \\ & 0,00 & 41,8 & 27,0 & 22,4 & 17,9 \\ & & 0,00 & 28,7 & 41,4 & 34,8 \\ & & & 0,00 & 27,6 & 21,5 \\ & & & & 0,00 & 22,8 \\ & & & & & 0,00 \end{array}\right] \end{array}$$

Passo 6: Objetos mais similares – parcela 6 e grupo (1,2,4,7,8,11).

Distância entre os objetos: 17,3.

Cálculo das distâncias entre o grupo (1,2,4,6,7,8,11) e os demais objetos:

$d_{(\mathbf{1,2,4,6,7,8,11})3} = \min [d_{(3,6)}; d_{(1,2,3,4,7,8,11)}] = 23{,}0;$

$d_{(\mathbf{1,2,4,6,7,8,11})5} = \min [d_{(5,6)}; d_{(1,2,4,5,7,8,11)}] = 28{,}7;$

$d_{(\mathbf{1,2,4,6,7,8,11})9} = \min [d_{(6,9)}; d_{(1,2,4,7,8,9,11)}] = 19{,}2;$ e

$d_{(\mathbf{1,2,4,6,7,8,11})10} = \min [d_{(6,10)}; d_{(1,2,4,7,8,10,11)}] = 17{,}5.$

Nova Matriz de Distância: $D_7$

$$\mathbf{D_7} = \begin{array}{c} \\ (1,2,4,6,7,8,11) \\ 3 \\ 5 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccc} (1,2,4,6,7,8,11) & 3 & 5 & 9 & 10 \end{array} \\ \begin{bmatrix} 0{,}00 & 23{,}0 & 28{,}7 & 19{,}2 & \mathbf{17{,}5} \\ & 0{,}00 & 41{,}8 & 22{,}4 & 17{,}9 \\ & & 0{,}00 & 41{,}4 & 34{,}8 \\ & & & 0{,}00 & 22{,}8 \\ & & & & 0{,}00 \end{bmatrix} \end{array}$$

Passo 7: Objetos mais similares – parcela 10 e grupo (1,2,4,6,7,8,11).

Distância entre os objetos: 17,5.

Cálculo das distâncias entre o grupo (1,2,4,6,7,8,10,11) e os demais objetos:

$d_{(\mathbf{1,2,4,6,7,8,10,11})3} = \min [d_{(3,10)}; d_{(1,2,4,7,8,10,11)}] = 17{,}9;$

$d_{(\mathbf{1,2,4,6,7,8,10,11})5} = \min [d_{(5,10)}; d_{(1,2,4,7,8,10,11)}] = 28{,}7;$ e

$d_{(\mathbf{1,2,4,6,7,8,10,11})9} = \min [d_{(9,10)}; d_{(1,2,4,7,8,10,11)}] = 19{,}2.$

Nova Matriz de Distância: $D_8$

$$\mathbf{D_8} = \begin{array}{c} \\ (1,2,4,6,7,8,10,11) \\ 3 \\ 5 \\ 9 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2,4,6,7,8,10,11) & 3 & 5 & 9 \end{array} \\ \begin{bmatrix} 0{,}00 & \mathbf{17{,}9} & 28{,}7 & 19{,}2 \\ & 0{,}00 & 41{,}8 & 22{,}4 \\ & & 0{,}00 & 41{,}4 \\ & & & 0{,}00 \end{bmatrix} \end{array}$$

Passo 8: Objetos mais similares – parcela 3 e grupo (1,2,4,6,7,8,10,11).

Distância entre os objetos: 17,9.

Cálculo das distâncias entre o grupo (1,2,3,4,6,7,8,10,11) e os demais objetos:

$d_{(\mathbf{1,2,3,4,6,7,8,10,11})5} = \min [d_{(3,5)}; d_{(1,2,4,5,7,8,10,11)}] = 28{,}7;$ e

$d_{(\mathbf{1,2,3,4,6,7,8,10,11})9} = \min [d_{(3,9)}; d_{(1,2,4,7,8,9,10,11)}] = 19{,}2.$

Nova Matriz de Distância: $D_9$

$$\mathbf{D_9} = \begin{array}{c} \\ (1,2,3,4,6,7,8,10,11) \\ 5 \\ 9 \end{array} \begin{array}{c} \begin{array}{ccc} (1,2,3,4,6,7,8,10,11) & 5 & 9 \end{array} \\ \begin{bmatrix} 0{,}00 & 28{,}7 & \mathbf{19{,}2} \\ & 0{,}00 & 41{,}4 \\ & & 0{,}00 \end{bmatrix} \end{array}$$

Passo 9: Objetos mais similares – parcela 9 e grupo (1,2,3,4,6,7,8,10,11).

Distância entre os objetos: 19,2.

Cálculo das distâncias entre o grupo (1,2,3,4,6,7,8,9,10,11) e os demais objetos:

$d_{(1,2,3,4,6,7,8,9,10,11)5} = \min [d_{(5,9)}; d_{(1,2,3,4,5,7,8,9,10,11)}] = 28,7$.

Nova Matriz de Distância: $D_{10}$

$$\mathbf{D_{10}} = \begin{matrix} \\ (1,2,3,4,6,7,8,9,10,11) \\ 5 \end{matrix} \begin{matrix} \begin{bmatrix} (1,2,3,4,6,7,8,9,10,11) & 5 \\ 0,00 & \mathbf{28,7} \\ & 0,00 \end{bmatrix} \end{matrix}$$

Passo 10: Formação do grupo final: (1,2,3,4,5,6,7,8,9,10,11).

Distância entre os objetos: 28,7.

$d_{(1,2,3,4,6,7,8,9,10,11)5} = 28,7$.

Com base nos cálculos, o procedimento seguinte consiste em reunir os resultados na forma de um resumo (Tabela 22), os quais também podem ser apresentados em um gráfico em forma de um dendrograma (Figura 8, saída do R).

**Tabela 22.** Resumo da análise de agrupamento obtido pelo emprego do método de ligação simples, com base na distância euclidiana, com a sequência das fusões das parcelas

| Distância | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Euclidiana | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 7,7 | P4 | P8 | | | | | | | | | |
| 8,8 | P7 | P11 | | | | | | | | | |
| 11,3 | P1 | P4 | P8 | | | | | | | | |
| 12,0 | P1 | P4 | P8 | P2 | | | | | | | |
| 13,3 | P1 | P4 | P8 | P2 | P7 | P11 | | | | | |
| 17,3 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | | | | |
| 17,5 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P10 | | | |
| 17,9 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P10 | P3 | | |
| 19,2 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P10 | P3 | P9 | |
| 28,7 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P10 | P3 | P9 | P5 |

A análise do dendrograma (Figura 8) permite ao pesquisador verificar o grau de similaridade entre as parcelas e os grupos ou entre dois grupos distintos. O estabelecimento dos grupos é feito, em geral, de forma subjetiva, tendo-se, comumente, por base as mudanças acentuadas de níveis associados ao conhecimento prévio do pesquisador do material avaliado.

### Passo a passo no R: Agrupamento por Ligação Simples (*Single Linkage*)

### 1. Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust*(). O método "*single*" (ligação simples) agrupa as parcelas com base na menor distância entre quaisquer dois

pontos em grupos diferentes. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

É importante converter a matriz de distância para o formato *dist* que a função *hclust*() espera, utilizando *as.dist*().

```
# Execução do Agrupamento Hierárquico pelo método da Ligação Simples
# as.dist() converte a matriz para o formato de objeto de distância
agrupamento_ligacao_simples <- hclust(as.dist(matriz_euclidiana), method = "single")
```

**2. Detalhamento das Fusões (Passo a Passo)**

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância
resumo_fusoes <- data.frame(
  Passo = 1:nrow(agrupamento_ligacao_simples$merge),
  Nivel_Distancia = round(agrupamento_ligacao_simples$height, 1)
)

# Exibindo o resumo das fusões
print(resumo_fusoes)
```

| Passo | Fusão | Nivel_Distancia |
|---|---|---|
| 1 | 1 | 7.7 |
| 2 | 2 | 8.8 |
| 3 | 3 | 11.3 |
| 4 | 4 | 12.0 |
| 5 | 5 | 13.3 |
| 6 | 6 | 17.3 |
| 7 | 7 | 17.5 |
| 8 | 8 | 17.9 |
| 9 | 9 | 19.2 |
| 10 | 10 | 28.7 |

**3. Visualização do Dendrograma**

No dendrograma (Figura 8), é possível mostrar visualmente como as parcelas foram se unindo e formando grupos em diferentes níveis de distância. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes.

```
# Gerando o gráfico do Dendrograma (Figura 8)
plot(agrupamento_ligacao_simples,
     main = "Dendrograma - Método da Ligação Simples",
     xlab = "Parcelas",
     ylab = "Distância Euclidiana",
     sub = "Baseado na Distância Euclidiana",
     hang = -1) # hang = -1 alinha os rótulos na parte inferior

# Adicionando a linha de corte (Linha Fenon) em um nível de distância específico (ex: 18.06)
```

```
abline(h = 18.06, col = "orange", lty = 2, lwd = 2)

# Destacando os grupos principais com retângulos coloridos (ex: 4 grupos)
rect.hclust(agrupamento_ligacao_simples, k = 4, border = c("blue", "green", "purple", "red"))
```

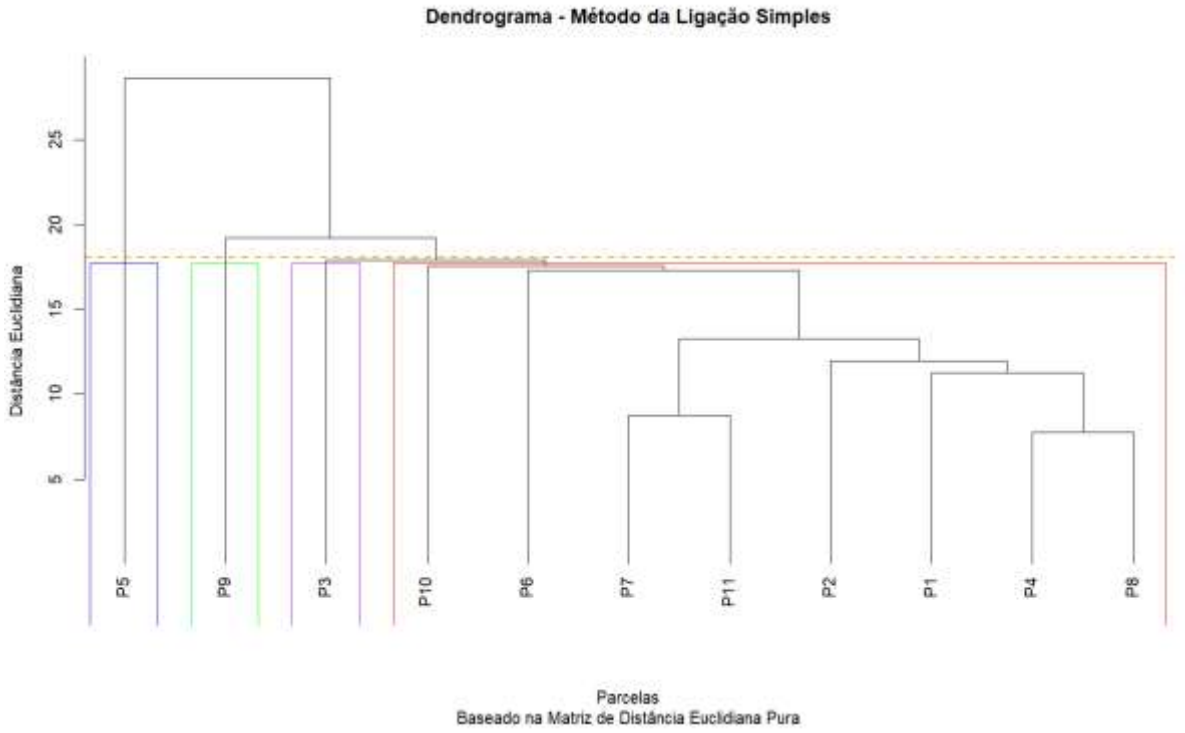



**Figura 8.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método de ligação simples, com base na distância euclidiana (d), considerando a linha fenon d = 18,06 e a formação de quatro grupos.

## b. Método da Ligação Completa (*Complete Linkage*)

Esse método também é denominado de método do elemento mais distante, sendo uma das técnicas de hierarquização aglomerativa de maior aplicação na análise de agrupamento (Gama, 2022). Como no método de ligação simples, aqui também não é exigida a fixação, "*a priori*", do número de agrupamentos.

Dados **n** elementos e admitindo-se conhecidos os **n(n-1)/2** valores de uma função de agrupamentos, $\mathbf{S_{ij}}$ ou $\mathbf{d_{ij}}$, $\mathbf{i = j = 1, 2, ... , n}$, apresentados na forma de uma matriz **S** ou **D**, esse método pode ser sintetizado nas seguintes etapas (Gama, 2022; Mardia et al., 2023):

1. Determina-se, com base na matriz de dados **X**, o conjunto de valores de uma função de agrupamento. Esses valores constituem os coeficientes de similaridade (**S**) ou a medida de distância estatística (**D**) que formam a matriz $\mathbf{S_1}$ ou $\mathbf{D_1}$.

2. Determina-se o valor mínimo de $\mathbf{d_{ij}}$ ou o valor máximo de $\mathbf{S_{ij}} > 0$, sendo, então, os elementos $\mathbf{E_i}$ e $\mathbf{E_j}$ reunidos em um primeiro grupo. Pressupõe-se que os (**n-2**) elementos restantes constituem, cada um, um agrupamento distinto.

3. Com base na matriz $\mathbf{D_1}$ ou $\mathbf{S_1}$, determina-se a distância entre cada um dos (**n-2**) elementos e o novo agrupamento formado pelos elementos ($\mathbf{E_i}$, $\mathbf{E_j}$). Essa distância é calculada pela relação:

$$\mathbf{d_{(i,j)\,l} = máx\ (d_{i1}, d_{jl}),\ l = 1, (n\text{-}2)}$$
$$\mathbf{l \neq i \neq j}$$
$$\mathbf{S_{(i,j)\,l} = min\ (S_{i1}, S_{jl}),\ \ l = 1, (n\text{-}2)}$$
$$\mathbf{l \neq i \neq j}$$

Sendo essas distâncias, ou coeficientes de similaridade, reunidas em uma matriz $\mathbf{D_2}$ ou $\mathbf{S_2}$.

4. Determina-se, com base na matriz $\mathbf{D_2}$ ou $\mathbf{S_2}$, o maior valor não nulo $\mathbf{d_{ij}}$ ou o menor valor $\mathbf{S_{ij} > 0}$, agrupando-se os elementos correspondentes, dando origem a um novo agrupamento, obtendo-se, assim, (**n-2**) agrupamentos.

5. Compõe-se um novo conjunto de valores de distâncias ou de similaridade com base na matriz **D** ou $\mathbf{S_2}$ (interação anterior), estabelecendo as relações entre o novo agrupamento e os demais.

Nesse método, é usual o emprego da distância euclidiana como função de agrupamento, embora qualquer uma das medidas de distância ou de similaridade possa ser empregada.

O critério da distância máxima ou mínima similaridade indica os elementos, ou grupos que devem ser reunidos para formar um novo agrupamento.

**Aplicação**

A partir da matriz de distância euclidiana $\mathbf{D_1}$, será aplicado o método de ligações completa para o agrupamento das 11 parcelas.

$\mathbf{D_1} =$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 12,5 | 33,9 | 17,3 | 18,5 | 11,3 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 12,0 | 34,4 | 21,8 | 18,3 | 16,2 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 29,6 | 41,8 | 27,0 | 30,9 | 27,9 | 22,4 | 17,9 | 24,2 |
| 4 | | | | 0,00 | 32,3 | 17,3 | 14,7 | 7,7 | 28,3 | 21,5 | 13,5 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 35,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 19,0 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 15,5 | 22,4 | 24,4 | 8,8 |
| 8 | | | | | | | | 0,00 | 26,7 | 21,2 | 13,3 |
| 9 | | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | | 0,00 |

Os seguintes passos são considerados:

Passo 1: Objetos mais similares – parcelas 4 e 8.

Distância os entre objetos: 7,7.

Cálculo das distâncias entre o grupo (4,8) e os demais objetos:

$d_{(\mathbf{4,8})1} = máx\ [d_{(1,4)}; d_{(1,8)}] = 12,5;$

$d_{(\mathbf{4,8})2} = máx\ [d_{(2,4)}; d_{(2,8)}] = 16,2;$

$d_{(\mathbf{4,8})3} = máx\ [d_{(3,4)}; d_{(3,8)}] = 29,6;$

$d_{(\mathbf{4,8})5} = máx\ [d_{(4,5)}; d_{(5,8)}] = 35,6;$

$d_{(\mathbf{4,8})6} = máx\ [d_{(4,6)}; d_{(6,8)}] = 19,0;$

$d_{(\mathbf{4,8})7} = máx\ [d_{(4,7)}; d_{(7,8)}] = 15,5;$

$d_{(\mathbf{4,8})9} = máx\ [d_{(4,9)}; d_{(8,9)}] = 28,3;$

$d_{(\mathbf{4,8})10} = máx\ [d_{(4,10)}; d_{(8,10)}] = 21,5;$ e

$d_{(\mathbf{4,8})11} = máx\ [d_{(4,9)}; d_{(8,9)}] = 13,5.$

Nova Matriz de Distância: $D_2$

$\mathbf{D_2} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 12,5 | 33,9 | 17,3 | 18,5 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 16,2 | 34,4 | 21,8 | 18,3 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 29,6 | 41,8 | 27,0 | 30,9 | 22,4 | 17,9 | 24,2 |
| (4,8) | | | | 0,00 | 35,6 | 19,0 | 15,5 | 28,3 | 21,5 | 13,5 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 22,4 | 24,4 | **8,8** |
| 9 | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | 0,00 |

Passo 2: Objetos mais similares – parcelas 7 e 11.

Distância entre os objetos: 8,8.

Cálculo das distâncias entre o grupo (7,11) e os demais objetos:

$d_{(\mathbf{7,11})1} = máx\ [d_{(1,7)}; d_{(1,11)}] = 18,5;$

$d_{(\mathbf{7,11})2} = máx\ [d_{(2,7)}; d_{(2,11)}] = 18,7;$

$d_{(\mathbf{7,11})3} = máx\ [d_{(3,7)}; d_{(3,11)}] = 30,9;$

$d_{(\mathbf{7,11})(4,8)} = máx\ [d_{(4,7,8)}; d_{(4,8,11)}] = 15,5;$

$d_{(\mathbf{7,11})5} = máx\ [d_{(5,7)}; d_{(5,11)}] = 38,6;$

$d_{(\mathbf{7,11})6} = máx\ [d_{(6,7)}; d_{(6,11)}] = 24,6;$

$d_{(\mathbf{7,11})9} = máx\ [d_{(7,9)}; d_{(9,11)}] = 22,8;$ e

$d_{(\mathbf{7,11})10} = máx\ [d_{(7,10)}; d_{(10,11)}] = 24,4.$

Nova Matriz de Distância: $D_3$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | **12,5** | 33,9 | 17,3 | 18,5 | 24,8 | 20,0 |
| 2 | | 0,00 | 31,7 | 16,2 | 34,4 | 21,8 | 18,7 | 29,4 | 25,6 |

$\mathbf{D_3}$ =

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | 0,00 | 29,6 | 41,8 | 27,0 | 30,9 | 22,4 | 17,9 |
| (4,8) | | 0,00 | 35,6 | 19,0 | 15,5 | 28,3 | 21,5 |
| 5 | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 |
| 6 | | | | 0,00 | 24,6 | 27,6 | 21,5 |
| (7,11) | | | | | 0,00 | 22,8 | 24,4 |
| 9 | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | 0,00 |

Passo 3: Objetos mais similares – parcela 1 e grupo (4,8).

Distância entre os objetos: 12,5.

Cálculo das distâncias entre o grupo (1,4,8) e os demais objetos:

$d_{(\mathbf{1,4,8})2} = \text{máx}\ [d_{(1,2)}; d_{(2,4,8)}] = 17,8;$

$d_{(\mathbf{1,4,8})3} = \text{máx}\ [d_{(1,3)}; d_{(3,4,8)}] = 29,6;$

$d_{(\mathbf{1,4,8})5} = \text{máx}\ [d_{(1,5)}; d_{(4,5,8)}] = 35,6;$

$d_{(\mathbf{1,4,8})6} = \text{máx}\ [d_{(1,6)}; d_{(4,6,8)}] = 19,0;$

$d_{(\mathbf{1,4,8})(7\text{-}11)} = \text{máx}\ [d_{(1,7,11)}; d_{(4,7,811)}] = 18,5;$

$d_{(\mathbf{1,4,8})9} = \text{máx}\ [d_{(1,9)}; d_{(4,8,9)}] = 28,3;$ e

$d_{(\mathbf{1,4,8})10} = \text{máx}\ [d_{(1,10)}; d_{(4,8,10)}] = 21,5.$

Nova Matriz de Distância: $D_4$

$\mathbf{D_4}$ =

| | (1,4,8) | 2 | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| (1,4,8) | 0,00 | **17,8** | 29,6 | 35,6 | 19,0 | 18,5 | 28,3 | 21,5 |
| 2 | | 0,00 | 31,7 | 34,4 | 21,8 | 18,7 | 29,4 | 25,6 |
| 3 | | | 0,00 | 41,8 | 27,0 | 30,9 | 22,4 | 17,9 |
| 5 | | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 |
| 6 | | | | | 0,00 | 24,6 | 27,6 | 21,5 |
| (7,11) | | | | | | 0,00 | 22,8 | 24,4 |
| 9 | | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | | 0,00 |

Passo 4: Objetos mais similares – parcela 2 e grupo (1,4,8).

Distância entre os objetos: 17,8.

Cálculo das distâncias entre o grupo (1,2,4,8) e os demais objetos:

$d_{(\mathbf{1,2,4,8})3} = \text{máx}\ [d_{(2,3)}; d_{(1,3,4,8)}] = 31,7;$

$d_{(\mathbf{1,2,4,8})5} = \text{máx}\ [d_{(2,5)}; d_{(1,4,5,8)}] = 35,6;$

$d_{(\mathbf{1,2,4,8})6} = \text{máx}\ [d_{(2,6)}; d_{(1,4,6,8)}] = 21,8;$

$d_{(\mathbf{1,2,4,8})(7,11)} = \text{máx}\ [d_{(2,7,11)}; d_{(1,4,7,8,11)}] = 18,7;$

$d_{(\mathbf{1,2,4,8})9} = \text{máx}\ [d_{(2,9)}; d_{(1,4,8,9)}] = 29,4;$ e

$d_{(\mathbf{1,2,4,8})10} = \text{máx}\ [d_{(2,10)}; d_{(1,4,8,10)}] = 25,6.$

Nova Matriz de Distância: $D_5$

$\mathbf{D_5}$ =

| | (1,2,4,8) | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|
| (1,2,4,8) | 0,00 | 31,7 | 35,6 | 21,8 | 18,7 | 29,4 | 25,6 |
| 3 | | 0,00 | 41,8 | 27,0 | 30,9 | 22,4 | **17,9** |
| 5 | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 |
| 6 | | | | 0,00 | 24,6 | 27,6 | 21,5 |
| (7,11) | | | | | 0,00 | 22,8 | 24,4 |
| 9 | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | 0,00 |

Passo 5: Objetos mais similares – parcelas 3 e 10.

Distância entre grupos: 17,9.

Cálculo das distâncias entre o grupo (3,10) e os demais objetos:

$d_{(\mathbf{3,10})(1,2,4,8)}$ = máx [$d_{(1,2,3,4,8)}$; $d_{(1,2,4,8,10)}$] = 31,7;

$d_{(\mathbf{3,10})5}$ = máx [$d_{(3,5)}$; $d_{(5,10)}$] = 41,8;

$d_{(\mathbf{3,10})6}$ = máx [$d_{(3,6)}$; $d_{(6,10)}$] = 27,0;

$d_{(\mathbf{3,10})(7,11)}$ = máx [$d_{(3,7,11)}$; $d_{(7,10,11)}$] = 30,9; e

$d_{(\mathbf{3,10})9}$ = máx [$d_{(3,9)}$; $d_{(9,10)}$] = 22,8.

Nova Matriz de Distância: $D_6$

$\mathbf{D_6}$ =

| | (1,2,4,8) | (3,10) | 5 | 6 | (7,11) | 9 |
|---|---|---|---|---|---|---|
| (1,2,4,8) | 0,00 | 31,7 | 35,6 | 21,8 | **18,7** | 29,4 |
| (3,10) | | 0,00 | 41,8 | 27,0 | 30,9 | 22,8 |
| 5 | | | 0,00 | 28,7 | 38,6 | 41,4 |
| 6 | | | | 0,00 | 24,6 | 27,6 |
| (7,11) | | | | | 0,00 | 22,8 |
| 9 | | | | | | 0,00 |

Passo 6: Objetos mais similares: grupos (1,2,4,8) e (7,11).

Distância entre os objetos: 18,7.

Cálculo das distâncias entre o grupo (1,2,4,7,8,11) e os demais objetos:

$d_{(\mathbf{1,2,4,7,8,11})(3,10)}$ = máx [$d_{(1,2,3,4,8)}$ ; $d_{(3,\ 7,10,11)}$] = 31,7;

$d_{(\mathbf{1,2,4,7,8,11})5}$ = máx [$d_{(1,2,4,5,8)}$ ; $d_{(5,7,11)}$] = 38,6;

$d_{(\mathbf{1,2,4,7,8,11})6}$ = máx [$d_{(1,2,4,6,8)}$; $d_{(6,7,11)}$] = 24,6; e

$d_{(\mathbf{1,2,4,7,8,11})9}$ = máx [$d_{(1,2,4,8,9)}$; $d_{(7,9,11)}$] = 29,4.

Nova Matriz de Distância: $D_7$

$\mathbf{D_7}$ =

| | (1,2,4,7,8,11) | (3,10) | 5 | 6 | 9 |
|---|---|---|---|---|---|
| (1,2,4,7,8,11) | 0,00 | 31,7 | 38,6 | 24,6 | 29,4 |
| (3,10) | | 0,00 | 41,8 | 27,0 | **22,8** |
| 5 | | | 0,00 | 28,7 | 41,4 |
| 6 | | | | 0,00 | 27,6 |
| 9 | | | | | 0,00 |

Passo 7: Objetos mais similares – parcela 9 e grupo (3,10).

Distância entre os objetos: 22,8.

Cálculo das distâncias entre o grupo (3,9,10) e os demais objetos:

$d_{(\mathbf{3,9,10})(1,2,4,7,8,11)}$ = máx [$d_{(1,2,3,4,7,8,10,11)}$; $d_{(1,2,4,7,8,9,11)}$] = 31,7;

$d_{(\mathbf{3,9,10})5}$ = máx [$d_{(3,5,10)}$; $d_{(5,9)}$] = 41,8; e

$d_{(\mathbf{3,9,10})6}$ = máx [$d_{(3,6,,10)}$; $d_{(6,9)}$] = 27,6.

Nova Matriz de Distância: $D_8$

$$\mathbf{D_8} = \begin{array}{c} \\ (1,2,4,7,8,11) \\ (3,9,10) \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2,4,7,8,11) & (3,9,10) & 5 & 6 \end{array} \\ \begin{bmatrix} 0{,}00 & 31{,}7 & 38{,}6 & \mathbf{24{,}6} \\ & 0{,}00 & 41{,}8 & 27{,}6 \\ & & 0{,}00 & 28{,}7 \\ & & & 0{,}00 \end{bmatrix} \end{array}$$

Passo 8: Objetos mais similares – parcela 6 e grupo (1,2,4,7,8,11).

Distância entre os objetos: 24,6.

Cálculo das distâncias entre o grupo (1,2,4,7,8,11) e os demais objetos:

$d_{(\mathbf{1,2,4,6,7,8,11})(3,9,10)}$= máx [$d_{(1,2,3,4,7,8,11)}$; $d_{(3,6,9,10)}$] = 31,7; e

$d_{(\mathbf{1,2,4,6,7,8,11})5}$ = máx [$d_{(1,2,4,5,7,8,11)}$; $d_{(5,6)}$] = 38,6.

Nova Matriz de Distância: $D_9$

$$\mathbf{D_9} = \begin{array}{c} \\ (1,2,4,6,7,8,11) \\ (3,9,10) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} (1,2,4,6,7,8,11) & (3,9,10) & 5 \end{array} \\ \begin{bmatrix} 0{,}00 & \mathbf{31{,}7} & 38{,}6 \\ & 0{,}00 & 41{,}8 \\ & & 0{,}00 \end{bmatrix} \end{array}$$

Passo 9: Objetos mais similares – grupos (3,9,10) e (1,2,4,6,7,8,11).

Distância entre os objetos: 31,7.

Cálculo das distâncias entre o grupo (1,2,3,4,6,7,8,9,10,11) e os demais objetos:

$d_{(\mathbf{1,2,3,4,6,7,8,9,10,11})5}$ = máx [$d_{(1,2,4,5,6,7,8,11)}$; $d_{(3,5,9,10)}$] = 41,8.

Nova Matriz de Distância: $D_{10}$

$$\mathbf{D_{10}} = \begin{array}{c} \\ (1,2,3,4,6,7,8,10,11) \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} (1,2,3,4,6,7,8,9,10,11) & 5 \end{array} \\ \begin{bmatrix} 0{,}00 & 41{,}8 \\ & 0{,}00 \end{bmatrix} \end{array}$$

Passo 10: Formação do grupo final: (1,2,3,4,5,6,7,8,9,10,11).

Distância entre os objetos: 41,8.

$d_{(\mathbf{1,2,3,4,6,7,8,9,10,11})5}$ = 41,8.

Tendo completado a fusão de todas as parcelas em um único agrupamento, os resultados foram reunidos na Tabela 23 e, a partir dela, foi obtido o dendrograma apresentado na Figura 9 (saída do R).

**Tabela 23.** Resumo da análise de agrupamento obtido pelo emprego do **método de ligação completa,** com base na distância euclidiana, mostrando a sequência das fusões das parcelas

| Distância | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Euclidiana | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 7,7 | P4 | P8 | | | | | | | | | |
| 8,8 | P7 | P11 | | | | | | | | | |
| 12,5 | P1 | P4 | P8 | | | | | | | | |
| 17,8 | P1 | P4 | P8 | P2 | | | | | | | |
| 17,9 | P3 | P10 | | | | | | | | | |
| 18,7 | P1 | P4 | P8 | P2 | P7 | P11 | | | | | |
| 22,8 | P3 | P10 | P9 | | | | | | | | |
| 24,6 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | | | | |
| 31,7 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P3 | P10 | P9 | |
| 41,8 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P3 | P10 | P9 | P5 |

## Passo a passo no R: Agrupamento por Ligação Completa (*Complete Linkage*)

### 1. Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust*(). O método "*complete*" (ligação completa) agrupa as parcelas com base na maior distância entre quaisquer dois pontos em grupos diferentes. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

É importante converter a matriz de distância para o formato *dist* que a função *hclust*() espera, utilizando *as.dist*().

```
# Execução do Agrupamento Hierárquico pelo método da Ligação Completa
# as.dist() converte a matriz para o formato de objeto de distância
agrupamento_ligacao_completa <- hclust(as.dist(matriz_euclidiana), method = "complete")
```

### 2. Detalhamento das Fusões (Passo a Passo)

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância
resumo_completo <- data.frame(
  Passo = 1:nrow(agrupamento_ligacao_completa$merge),
  Distancia = round(agrupamento_ligacao_completa$height, 1)
)

# Exibindo o resumo das fusões
print(resumo_completo)
```

| Passo | Fusão | Distância |
|---|---|---|
| 1 | 1 | 7.7 |
| 2 | 2 | 8.8 |
| 3 | 3 | 12.5 |
| 4 | 4 | 17.8 |
| 5 | 5 | 17.9 |
| 6 | 6 | 18.7 |
| 7 | 7 | 22.8 |
| 8 | 8 | 24.6 |
| 9 | 9 | 31.6 |
| 10 | 10 | 41.8 |

### 3. Visualização do Dendrograma

No dendrograma (Figura 9), mostramos visualmente como as parcelas foram se unindo e formando grupos em diferentes níveis de distância. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes, além de personalizar os eixos para melhor visualização.

```
# Configurando a margem para caber o título inferior e os rótulos
par(mar=c(6, 4, 4, 2))

# Gerando o gráfico do Dendrograma (Figura 9)
plot(agrupamento_ligacao_completa,
     main = "Dendrograma - Método da Ligação Completa",
     xlab = "Sequência de Agrupamento de Parcelas",
     ylab = "Distância Euclidiana",
     sub = "Baseado na Distância Euclidiana",
     hang = -1, # hang = -1 alinha os rótulos na parte inferior
     axes = FALSE) # Remove os eixos padrão para customizar

# Adicionando o eixo Y lateral
axis(2, at = seq(0, 40, by = 10))

# Adicionando a linha de corte (Linha Fenon) em um nível de distância específico (ex: 18.5)
abline(h = 18.5, col = "orange", lwd = 2)

# Destacando os grupos principais com retângulos coloridos (ex: 4 grupos)
rect.hclust(agrupamento_ligacao_completa, k = 4, border = c("red", "green", "blue", "cyan"))
```

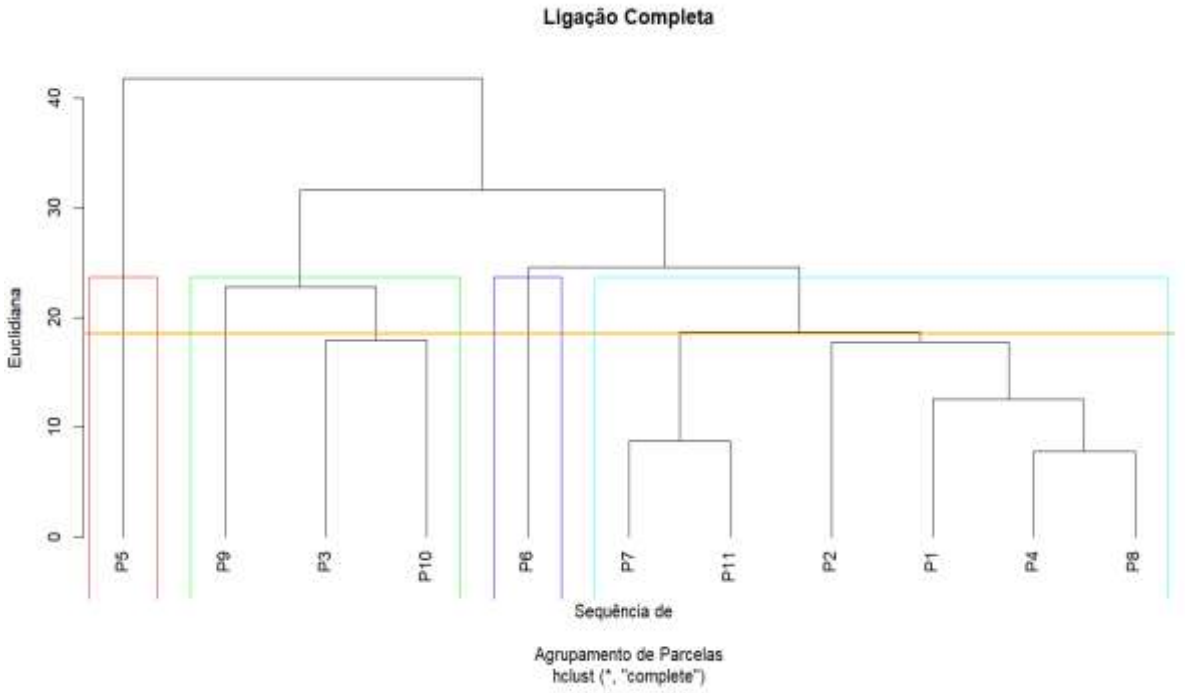



**Figura 9.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método de ligação completa, com base na distância euclidiana, considerando a linha fenon d = 18,5 e a formação de quatro grupos.

## c. Método do Centroide (UPGMC)

O método do centroide, conhecido como UPGMC (*Unweighted Pair-Group Method using the Centroid Approach*), foi proposto por Sokal e Michener (1958) e teve como origem a caracterização da matriz de dados como pontos do espaço euclidiano ($\mathbf{I_p}$) (Everitt et al., 2011). Cada agrupamento é considerado um simples ponto, representado pelo seu centro de massa, chamado centroide. Esse método utiliza uma função de agrupamento para medir a distância entre os centros de massa dos dados.

A cada passo dessa técnica, há a redefinição da matriz de dados, em que cada agrupamento é representado pelo vetor médio das **p** variáveis envolvidas. Na realidade, uma nova matriz de distâncias é determinada a cada interação.

Seja **X (n** x **p)** a matriz de dados básicos, em que cada linha representa o conjunto de valores observados para cada espécie, e cada coluna representa uma parcela de área fixa ou variável.

Esse algoritmo pode ser desenvolvido nas seguintes etapas:

1. Com base na matriz de dados e em uma função de agrupamentos, calculam-se os índices de similaridade ($\mathbf{S_{ij}}$) ou as distâncias ($\mathbf{d_{ij}}$) entre as parcelas (**i**,**j**), **i** = 1, 2, ... , **n**. Esses dados são reunidos em uma matriz de similaridades (**S**) ou de distâncias (**D**).

2. A primeira operação em **S** ou **D** consiste em procurar pelo maior valor de $\mathbf{S_{ij}}$ ou menor valor de $\mathbf{d_{ij}}$, excluindo os valores de $\mathbf{S_{ij}}$ ou $\mathbf{d_{ij}}$ em que **i = j**, ou seja, excluir os elementos da diagonal principal. Os elementos (parcelas) $\mathbf{X_i}$ e $\mathbf{X_j}$, cuja similaridade é maior, são reunidos em um mesmo agrupamento.

3. Em seguida, uma nova matriz de distância é calculada, identificam-se os elementos do agrupamento mais próximo e constrói-se um novo agrupamento. Segundo Lance e Williams (1967), a distância pode ser obtida por meio da seguinte expressão:

$$d^2_{k(i,j)} = \left(\frac{n_i}{n_i + n_j}\right) d^2_{ki} + \left(\frac{n_j}{n_i + n_j}\right) d^2_{kj} - \left(\frac{n_i n_j}{(n_i + n_j)^2}\right) d^2_{ij}$$

Em que: $d^2_{k(i,j)}$, $d^2_{ki}$, $d^2_{kj}$ e $d^2_{ij}$ = distâncias euclidianas quadradas entre os elementos k e agrupamento ij, k e i, k e j, e i e j, respectivamente; e $n_i$, $n_j$ e $n_k$ = número de elementos nos agrupamentos i, j e k, respectivamente.

4. Verifica-se se o número de agrupamentos determinados é igual ao valor fixado, **g ≤ n**. Se a condição for verdadeira, termina-se o processo, caso contrário, retorna-se ao item 2.

Uma característica importante desse algoritmo é que a distância entre agrupamentos é determinada pela distância entre os pontos representativos dos seus respectivos centros de massa (**centroide**).

**Aplicação**

Consideremos a matriz de distâncias euclidianas quadradas $\mathbf{D^2_1}$:

$\mathbf{D^2_1} =$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,0 | 317,0 | 529,0 | 156,0 | 1151,0 | 300,0 | 342,0 | 127,0 | 615,0 | 399,0 | 217,0 |
| 2 | | 0,0 | 1002,0 | 143,0 | 1180,0 | 475,0 | 335,0 | 262,0 | 866,0 | 656,0 | 348,0 |
| 3 | | | 0,00 | 875,0 | 1748,0 | 731,0 | 955,0 | 780,0 | 504,0 | 322,0 | 586,0 |
| 4 | | | | 0,0 | 1041,0 | 298,0 | 216,0 | **59,0** | 801,0 | 463,0 | 183,0 |
| 5 | | | | | 0,0 | 821,0 | 1493,0 | 1264,0 | 1716,0 | 1212,0 | 1202,0 |
| 6 | | | | | | 0,0 | 604,0 | 361,0 | 761,0 | 463,0 | 375,0 |
| 7 | | | | | | | 0,0 | 241,0 | 503,0 | 597,0 | 77,0 |
| 8 | | | | | | | | 0,0 | 714,0 | 448,0 | 176,0 |
| 9 | | | | | | | | | 0,0 | 520,0 | 370,0 |
| 10 | | | | | | | | | | 0,0 | 308,0 |
| 11 | | | | | | | | | | | 0,0 |

| N = | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|

Em que: N = número de objetos que formam o grupo.

Os seguintes passos serão considerados:

Passo 1: Objetos mais similares – parcelas 4 e 8.

Distância os entre objetos: 59,0.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**)

$$d^2_{1(4,8)} = \left(\frac{1}{1+1}\right)x156{,}0 + \left(\frac{1}{1+1}\right)x217{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 126{,}8$$

$$d^2_{2(4,8)} = \left(\frac{1}{1+1}\right)x143{,}0 + \left(\frac{1}{1+1}\right)x262{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 187{,}8$$

$$d^2_{3(4,8)} = \left(\frac{1}{1+1}\right)x875{,}0 + \left(\frac{1}{1+1}\right)x780{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 812{,}8$$

$$d^2_{5(4,8)} = \left(\frac{1}{1+1}\right)x1041{,}0 + \left(\frac{1}{1+1}\right)x1264{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 1137{,}8$$

$$d^2_{6(4,8)} = \left(\frac{1}{1+1}\right)x298{,}0 + \left(\frac{1}{1+1}\right)x361{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 314{,}8$$

$$d^2_{7(4,8)} = \left(\frac{1}{1+1}\right)x216{,}0 + \left(\frac{1}{1+1}\right)x241{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 213{,}8$$

$$d^2_{9(4,8)} = \left(\frac{1}{1+1}\right)x801{,}0 + \left(\frac{1}{1+1}\right)x714{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 742{,}8$$

$$d^2_{10(4,8)} = \left(\frac{1}{1+1}\right)x463{,}0 + \left(\frac{1}{1+1}\right)x448{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 440{,}8$$

$$d^2_{11(4,8)} = \left(\frac{1}{1+1}\right)x183{,}0 + \left(\frac{1}{1+1}\right)x176{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x59{,}0 = 164{,}8$$

Nova Matriz de Distância: $D^2_2$

$\mathbf{D^2_2} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,0 | 317,0 | 529,0 | 126,8 | 1151,0 | 300,0 | 342,0 | 615,0 | 399,0 | 217,0 |
| 2 | | 0,0 | 1002,0 | 187,8 | 1180,0 | 475,0 | 335,0 | 866,0 | 656,0 | 348,0 |
| 3 | | | 0,0 | 812,8 | 1748,0 | 731,0 | 955,0 | 504,0 | 322,0 | 586,0 |
| (4,8) | | | | 0,0 | 1137,8 | 314,8 | 213,8 | 742,8 | 440,8 | 164,8 |
| 5 | | | | | 0,0 | 821,0 | 1493,0 | 1716,0 | 1212,0 | 1202,0 |
| 6 | | | | | | 0,0 | 604,0 | 761,0 | 463,0 | 375,0 |
| 7 | | | | | | | 0,0 | 503,0 | 597,0 | **77,0** |
| 9 | | | | | | | | 0,0 | 520,0 | 370,0 |
| 10 | | | | | | | | | 0,0 | 308,0 |
| 11 | | | | | | | | | | 0,0 |

| N = | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|

Passo 2: Objetos mais similares – parcelas 7 e 11.

Distância entre os objetos: 77,0.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{1(7,11)} = \left(\frac{1}{1+1}\right)x342{,}0 + \left(\frac{1}{1+1}\right)x217{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 260{,}3$$

$$d^2_{2(7,11)} = \left(\frac{1}{1+1}\right)x335{,}0 + \left(\frac{1}{1+1}\right)x348{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 322{,}0$$

$$d^2_{3(7,11)} = \left(\frac{1}{1+1}\right)x955{,}0 + \left(\frac{1}{1+1}\right)x586{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 751{,}3$$

$$d^2_{(4,8)(7,11)} = \left(\frac{1}{1+1}\right)x213{,}8 + \left(\frac{1}{1+1}\right)x164{,}8 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 170{,}0$$

$$d^2_{5(7,11)} = \left(\frac{1}{1+1}\right)x1493{,}0 + \left(\frac{1}{1+1}\right)x1202{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 1328{,}3$$

$$d^2_{6(7,11)} = \left(\frac{1}{1+1}\right)x604{,}0 + \left(\frac{1}{1+1}\right)x375{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 470{,}3$$

$$d^2_{9(7,11)} = \left(\frac{1}{1+1}\right)x503{,}0 + \left(\frac{1}{1+1}\right)x370{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 417{,}3$$

$$d^2_{10(7,11)} = \left(\frac{1}{1+1}\right)x597{,}0 + \left(\frac{1}{1+1}\right)x308{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x77{,}0 = 433{,}3$$

Nova Matriz de Distância: $D^2_3$

$\mathbf{D^2_3} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,0 | 317,0 | 529,0 | **126,8** | 1151,0 | 300,0 | 260,3 | 615,0 | 399,0 |
| 2 | | 0,0 | 1002,0 | 187,8 | 1180,0 | 475,0 | 322,3 | 866,0 | 656,0 |
| 3 | | | 0,0 | 812,8 | 1748,0 | 731,0 | 751,3 | 504,0 | 322,0 |
| (4,8) | | | | 0,0 | 1137,8 | 314,8 | 170,0 | 742,8 | 440,8 |
| 5 | | | | | 0,0 | 821,0 | 1328,3 | 1716,0 | 1212,0 |
| 6 | | | | | | 0,0 | 470,3 | 761,0 | 463,0 |
| (7,11) | | | | | | | 0,0 | 417,3 | 433,3 |
| 9 | | | | | | | | 0,0 | 520,0 |
| 10 | | | | | | | | | 0,0 |

| N = | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|

Passo 3: Objetos mais similares – parcela 1 e grupo (4,8).

Distância entre os objetos: 126,8.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{2(1,4,8)} = \left(\frac{1}{1+2}\right)x317{,}0 + \left(\frac{2}{1+2}\right)x187{,}8 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 202{,}7$$

$$d^2_{3(1,4,8)} = \left(\frac{1}{1+2}\right)x529{,}0 + \left(\frac{2}{1+2}\right)x812{,}8 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 690{,}0$$

$$d^2_{5(1,4,8)} = \left(\frac{1}{1+2}\right)x1151{,}0 + \left(\frac{2}{1+2}\right)x1137{,}8 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 1114{,}0$$

$$d^2_{6(1,4,8)} = \left(\frac{1}{1+2}\right)x300{,}0 + \left(\frac{2}{1+2}\right)x314{,}8 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 281{,}7$$

$$d^2_{(7,11)(1,4,8)} = \left(\frac{1}{1+2}\right)x260{,}3 + \left(\frac{2}{1+2}\right)x170{,}0 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 171{,}9$$

$$d^2_{9(1,4,8)} = \left(\frac{1}{1+2}\right)x615{,}0 + \left(\frac{2}{1+2}\right)x742{,}8 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 672{,}0$$

$$d^2_{10(1,4,8)} = \left(\frac{1}{1+2}\right)x399{,}0 + \left(\frac{2}{1+2}\right)x440{,}8 - \left(\frac{1x2}{(1+2)^2}\right)x126{,}8 = 398{,}7$$

Nova Matriz de Distância: $D_{24}$

$\mathbf{D^2_4} =$

| | (1,4,8) | 2 | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| (1,4,8) | 0,0 | 202,7 | 690,0 | 1114,0 | 281,7 | **171,9** | 672,0 | 398,7 |
| 2 | | 0,0 | 1002,0 | 1180,0 | 475,0 | 322,3 | 866,0 | 656,0 |
| 3 | | | 0,0 | 1748,0 | 731,0 | 751,3 | 504,0 | 322,0 |
| 5 | | | | 0,0 | 821,0 | 1328,3 | 1716,0 | 1212,0 |
| 6 | | | | | 0,0 | 470,3 | 761,0 | 463,0 |
| (7,11) | | | | | | 0,0 | 417,3 | 433,3 |
| 9 | | | | | | | 0,0 | 520,0 |
| 10 | | | | | | | | 0,0 |

| N = | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|

Passo 4: Objetos mais similares – grupos (1,4,8) e (7,11).

Distância entre os objetos: 171,9.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{2(1,4,7,8,11)} = \left(\frac{3}{3+2}\right)x202{,}7 + \left(\frac{2}{3+2}\right)x322{,}3 - \left(\frac{3x2}{(3+2)^2}\right)x171{,}9 = 209{,}2$$

$$d^2_{3(1,4,7,8,11)} = \left(\frac{3}{3+2}\right)x690{,}0 + \left(\frac{2}{3+2}\right)x751{,}3 - \left(\frac{3x2}{(3+2)^2}\right)x171{,}9 = 673{,}2$$

$$d^2_{5(1,4,7,8,11)} = \left(\frac{3}{3+2}\right)x1114{,}0 + \left(\frac{2}{3+2}\right)x1328{,}3 - \left(\frac{3x2}{(3+2)^2}\right)x171{,}9 = 1158{,}4$$

$$d^2_{6(1,4,7,8,11)} = \left(\frac{3}{3+2}\right)x281{,}7 + \left(\frac{2}{3+2}\right)x470{,}3 - \left(\frac{3x2}{(3+2)^2}\right)x171{,}9 = 315{,}8$$

$$d^2_{9(1,4,7,8,11)} = \left(\frac{3}{3+2}\right)x672{,}0 + \left(\frac{2}{3+2}\right)x417{,}3 - \left(\frac{3x2}{(3+2)^2}\right)x171{,}9 = 528{,}8$$

$$d^2_{10(1,4,7,8,11)} = \left(\frac{3}{3+2}\right)x398{,}7 + \left(\frac{2}{3+2}\right)x433{,}3 - \left(\frac{3x2}{(3+2)^2}\right)x171{,}9 = 371{,}2$$

Nova Matriz de Distância: $D^2_5$

$\mathbf{D^2_5}$ =

| | (1,4,7,8,11) | 2 | 3 | 5 | 6 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| (1,4,7,8,11) | 0,0 | **209,2** | 673,2 | 1158,4 | 315,8 | 528,8 | 371,2 |
| 2 | | 0,0 | 1002,0 | 1180,0 | 475,0 | 866,0 | 656,0 |
| 3 | | | 0,0 | 1748,0 | 731,0 | 504,0 | 322,0 |
| 5 | | | | 0,0 | 821,0 | 1716,0 | 1212,0 |
| 6 | | | | | 0,0 | 761,0 | 463,0 |
| 9 | | | | | | 0,0 | 520,0 |
| 10 | | | | | | | 0,0 |

N = [ 5 1 1 1 1 1 1 ]

Passo 5: Objetos mais similares – grupo (1,4,7,8,11) e parcela 2.

Distância entre os objetos: 209,2.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{3(1,2,4,7,8,11)} = \left(\frac{5}{5+1}\right)x673{,}2 + \left(\frac{1}{5+1}\right)x1002{,}0 - \left(\frac{5x1}{(5+1)^2}\right)x209{,}2 = 699{,}0$$

$$d^2_{5(1,2,4,7,8,11)} = \left(\frac{5}{5+1}\right)x1158{,}4 + \left(\frac{1}{5+1}\right)x1180{,}0 - \left(\frac{5x1}{(5+1)^2}\right)x209{,}2 = 1133{,}0$$

$$d^2_{6(1,2,4,7,8,11)} = \left(\frac{5}{5+1}\right)x315{,}8 + \left(\frac{1}{5+1}\right)x475{,}0 - \left(\frac{5x1}{(5+1)^2}\right)x = 313{,}3$$

$$d^2_{9(1,2,4,7,8,11)} = \left(\frac{5}{5+1}\right)x528{,}8 + \left(\frac{1}{5+1}\right)x866{,}0 - \left(\frac{5x1}{(5+1)^2}\right)x = 556{,}0$$

$$d^2_{10(1,2,4,7,8,11)} = \left(\frac{5}{5+1}\right)x371{,}2 + \left(\frac{1}{5+1}\right)x656{,}0 - \left(\frac{5x1}{(5+1)^2}\right)x = 389{,}6$$

Nova Matriz de Distância: $D^2_6$

$\mathbf{D^2_6}$ =

| | (1,2,4,7,8,11) | 3 | 5 | 6 | 9 | 10 |
|---|---|---|---|---|---|---|
| (1,2,4,7,8,11) | 0,0 | 699,0 | 1133,0 | **313,3** | 556,0 | 389,6 |
| 3 | | 0,0 | 1748,0 | 731,0 | 504,0 | 322,0 |
| 5 | | | 0,0 | 821,0 | 1716,0 | 1212,0 |
| 6 | | | | 0,0 | 761,0 | 463,0 |
| 9 | | | | | 0,0 | 520,0 |
| 10 | | | | | | 0,0 |

N = [ 6 1 1 1 1 1 ]

Passo 6: Objetos mais similares – grupo (1,2,4,7,8,11) e parcela 6.

Distância entre os objetos: 313,3.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{3(1,2,4,6,7,8,11)} = \left(\frac{6}{6+1}\right)x699{,}0 + \left(\frac{1}{6+1}\right)x731{,}0 - \left(\frac{6x1}{(6+1)^2}\right)x313{,}3 = 665{,}2$$

$$d^2_{5(1,2,4,6,7,8,11)} = \left(\frac{6}{6+1}\right)x1133{,}0 + \left(\frac{1}{6+1}\right)x821{,}0 - \left(\frac{6x1}{(6+1)^2}\right)x313{,}3 = 1050{,}0$$

$$d^2_{9(1,2,4,6,7,8,11)} = \left(\frac{6}{6+1}\right)x556{,}0 + \left(\frac{1}{6+1}\right)x761{,}0 - \left(\frac{6x1}{(6+1)^2}\right)x313{,}3 = 546{,}9$$

$$d^2_{10(1,2,4,6,7,8,11)} = \left(\frac{6}{6+1}\right)x389{,}6 + \left(\frac{1}{6+1}\right)x463{,}0 - \left(\frac{6x1}{(6+1)^2}\right)x313{,}3 = 361{,}8$$

Nova Matriz de Distância: $D^2_7$

$$\mathbf{D^2_7} = \begin{array}{c} \\ (1,2,4,6,7,8,11) \\ 3 \\ 5 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccc} (1,2,4,6,7,8,11) & 3 & 5 & 9 & 10 \end{array} \\ \left[\begin{array}{ccccc} 0{,}0 & 665{,}2 & 1050{,}0 & 546{,}9 & 361{,}8 \\ & 0{,}0 & 1748{,}0 & 504{,}0 & \mathbf{322{,}0} \\ & & 0{,}0 & 1716{,}0 & 1212{,}0 \\ & & & 0{,}0 & 520{,}0 \\ & & & & 0{,}0 \end{array}\right] \end{array}$$

$$N = \left[\begin{array}{ccccc} 7 & 1 & 1 & 1 & 1 \end{array}\right]$$

Passo 7: Objetos mais similares – parcela 3 e parcela 10.

Distância entre os objetos: 322,0.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{(1,2,4,6,7,8,11)(3,10)} = \left(\frac{1}{1+1}\right)x665{,}2 + \left(\frac{1}{1+1}\right)x361{,}8 - \left(\frac{1x1}{(1+1)^2}\right)x322{,}0 = 433{,}0$$

$$d^2_{5(3,10)} = \left(\frac{1}{1+1}\right)x1748{,}0 + \left(\frac{1}{1+1}\right)x1212{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x322{,}0 = 1399{,}5$$

$$d^2_{9(3,10)} = \left(\frac{1}{1+1}\right)x504{,}0 + \left(\frac{1}{1+1}\right)x520{,}0 - \left(\frac{1x1}{(1+1)^2}\right)x322{,}0 = 431{,}5$$

Nova Matriz de Distância: $D^2_8$

| $\mathbf{D^2_8}$ = | | (1,2,4,6,7,8,11) | (3,10) | 5 | 9 |
|---|---|---|---|---|---|
| | (1,2,4,6,7,8,11) | 0,0 | 433,0 | 1050,0 | 546,9 |
| | (3,10) | | 0,0 | 1399,5 | **431,5** |
| | 5 | | | 0,0 | 1716,0 |
| | 9 | | | | 0,0 |
| N | = | 7 | 2 | 1 | 1 |

Passo 8: Objetos mais similares – grupo (3, 10) e parcela 9.

Distância entre os objetos: 431,5.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**):

$$d^2_{(1,2,4,6,7,8,11)(3,9,10)} = \left(\frac{2}{2+1}\right)x433{,}0 + \left(\frac{1}{2+1}\right)x546{,}9 - \left(\frac{2x1}{(2+1)^2}\right)x431{,}5 = 375{,}1$$

$$d^2_{5(3,9,10)} = \left(\frac{2}{2+1}\right)x1399{,}5 + \left(\frac{1}{2+1}\right)x1716{,}0 - \left(\frac{2x1}{(2+1)^2}\right)x433{,}0 = 1409{,}1$$

Nova Matriz de Distância: $D^2_9$

| $\mathbf{D^2_9}$ = | | (1,2,4,6,7,8,11) | (3,9,10) | 5 |
|---|---|---|---|---|
| | (1,2, 4,6,7,8,11) | 0,00 | **375,1** | 1050,0 |
| | (3,9,10) | | 0,00 | 1409,1 |
| | 5 | | | 0,00 |
| N | = | 7 | 3 | 1 |

Passo 9: Objetos mais similares – grupos (1,2,4,6,7,8,11) e (3,9,10).

Distância entre os objetos: 375,1.

Cálculo das distâncias entre os pontos representativos dos seus respectivos centros de massa (**centroide**).

$$d^2_{5(1,2,3,4,6,7,8,9,10,11)} = \left(\frac{7}{7+3}\right)x1050{,}0 + \left(\frac{3}{7+3}\right)x1409{,}1 - \left(\frac{7x3}{(7+3)^2}\right)x375{,}1 = 1069{,}5$$

Passo 10: Objetos mais similares – grupo (1,2,3,4,6,7,8,9,10,11) e parcela 5.

Distância entre os objetos: 1069,5.

Os resultados estão resumidos na Tabela 24, e o dendrograma, na Figura 10 (saída do R).

**Tabela 24.** Resumo da análise de agrupamento obtido pelo emprego do método de centroide, com base na distância euclidiana, mostrando a sequência das fusões das parcelas

| Passo | Distância Euclidiana | | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Quadrada | Simples | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 1 | 59,0 | 7,7 | P4 | P8 | | | | | | | | | |
| 2 | 77,0 | 8,8 | P7 | P11 | | | | | | | | | |
| 3 | 126,8 | 11,3 | P4 | P8 | P1 | | | | | | | | |
| 4 | 171,9 | 13,1 | P4 | P8 | P1 | P7 | P11 | | | | | | |
| 5 | 209,2 | 14,5 | P4 | P8 | P1 | P7 | P11 | P2 | | | | | |
| 6 | 313,3 | 17,7 | P4 | P8 | P1 | P7 | P11 | P2 | P6 | | | | |
| 7 | 322,0 | 17,9 | P3 | P10 | | | | | | | | | |
| 8 | 431,5 | 20,8 | P3 | P10 | P9 | | | | | | | | |
| 9 | 375,1 | 19,4 | P4 | P8 | P1 | P7 | P11 | P2 | P6 | P3 | P10 | P9 | |
| 10 | 1069,5 | 32,8 | P4 | P8 | P1 | P7 | P11 | P2 | P6 | P3 | P10 | P9 | P5 |

## Passo a passo no R: Agrupamento por Centroide (UPGMC)

### 1. Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust*(). O método "centroid" (centroide) agrupa as parcelas com base na distância entre os centros de massa dos grupos. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

**Nota Técnica Importante:** No R, o método do centroide (*method = "centroid"*) na função *hclust()* exige que as distâncias sejam elevadas ao quadrado (^2) antes de serem passadas para a função. Isso é necessário para manter a consistência geométrica do método. Após o agrupamento, ajustamos a altura (*height*) do objeto *hclust*, aplicando a raiz quadrada, para que os níveis de distância no dendrograma sejam comparáveis com as distâncias originais.

```
# Execução do Agrupamento Hierárquico pelo método do Centroide
# IMPORTANTE: O método centroid requer distâncias ao quadrado
agrupamento_centroide <- hclust(as.dist(matriz_euclidiana)^2, method = "centroid")

# Ajustamos a altura (height) para a raiz quadrada para que os níveis de distância sejam comparáveis
agrupamento_centroide$height <- sqrt(agrupamento_centroide$height)
```

### 2. Detalhamento das Fusões (Passo a Passo)

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância
resumo_centroide <- data.frame(
  Passo = 1:nrow(agrupamento_centroide$merge),
  Distancia = round(agrupamento_centroide$height, 1)
)
```

```
# Exibindo o resumo das fusões
print(resumo_centroide)
```

| Passo | Fusão | Distância |
|---|---|---|
| 1 | 1 | 7.7 |
| 2 | 2 | 8.8 |
| 3 | 3 | 11.3 |
| 4 | 4 | 13.1 |
| 5 | 5 | 14.5 |
| 6 | 6 | 17.7 |
| 7 | 7 | 17.9 |
| 8 | 8 | 20.8 |
| 9 | 9 | 19.4 |
| 10 | 10 | 32.8 |

### 3. Visualização do Dendrograma

A Figura 10 é a representação gráfica do agrupamento hierárquico, mostrando visualmente como foram formados os grupos. No método do centroide, é possível que ocorram inversões, o que pode causar cruzamentos visuais nos “galhos” do dendrograma. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes, além de personalizar os eixos para melhor visualização.

```
# Configurando a margem para caber o título inferior e os rótulos
par(mar=c(6, 4, 4, 2))

# Gerando o gráfico do Dendrograma (Figura 10)
plot(agrupamento_centroide,
     main = "Dendrograma - Método do Centroide (UPGMC)",
     xlab = "Sequência de Agrupamento de Parcelas",
     ylab = "Distância Euclidiana",
     sub = "Baseado na Distância Euclidiana",
     hang = -1, # hang = -1 alinha os rótulos na parte inferior
     axes = FALSE) # Remove os eixos padrão para customizar

# Adicionando o eixo Y lateral
axis(2, at = seq(0, 30, by = 5))

# Adicionando a linha de corte (Linha Fenon) em um nível de distância específico (ex: 18.2)
abline(h = 18.2, col = "orange", lwd = 2)

# Destacando os grupos principais com retângulos coloridos (ex: 4 grupos)
rect.hclust(agrupamento_centroide, k = 4, border = c("blue", "green", "purple", "red"))
```

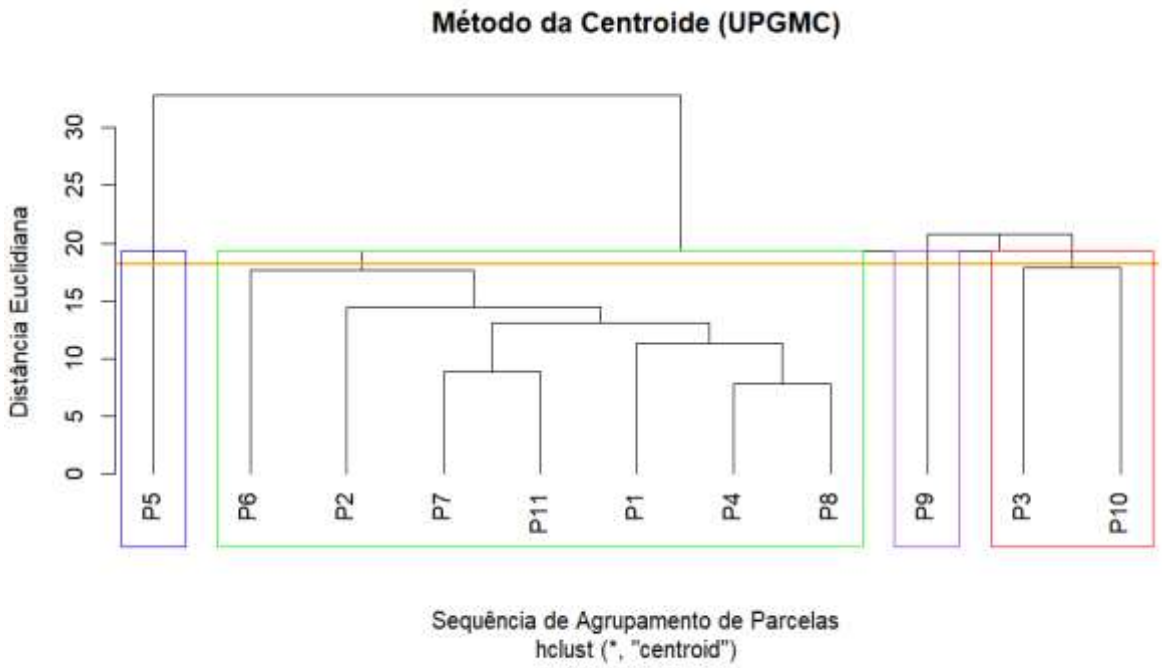



**Figura 10.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método do centroide (UPGMC), com base na distância euclidiana, considerando a linha fenon d = 18,2 e a formação de quatro grupos.

## d. Método da Mediana (WPGMC)

Esse algoritmo é um caso particular do método do centroide (Orlóci, 1978; Gama, 2022), conhecido como WPGMC (*Weighted Pair-Group Method using the Centroid Approach*). A determinação da distância entre dois agrupamentos por meio do cálculo do centro de massa não considera o número de elementos em cada um dos agrupamentos. Assim, o vetor médio que representa o novo agrupamento pode, eventualmente, ficar situado entre os elementos do agrupamento com maior número de elementos. Para contornar esse problema, Gower (1967) desenvolveu um procedimento de cálculo que pondera a medida de distância pelo número de elementos de cada agrupamento.

A rigor, os dois métodos são um só, não havendo razões para classificá-los como métodos distintos.

Esse algoritmo pode ser desenvolvido nas seguintes etapas:

1. Com base na matriz de dados e em uma função de agrupamentos, calculam-se os índices de similaridade ($\mathbf{S_{ij}}$) ou as distâncias ($\mathbf{d_{ij}}$) entre as parcelas (**i**,**j**), **i** = 1, 2, .. , **n**. Esses dados são reunidos em uma matriz de similaridades (**S**) ou de distâncias (**D**).

2. A primeira operação em **S** ou **D** consiste em procurar pelo maior valor de $\mathbf{S_{ij}}$ ou menor valor de $\mathbf{d_{ij}}$, excluindo os valores de $\mathbf{S_{ij}}$ ou $\mathbf{d_{ij}}$ em que **i** = **j**, ou seja, excluir os elementos da

diagonal principal. Os elementos (parcelas) $\mathbf{X_i}$ e $\mathbf{X_j}$, cuja similaridade é maior, são reunidos em um mesmo agrupamento.

3. Em seguida, uma nova matriz de distância é calculada, identificam-se os elementos do agrupamento mais próximo e constrói-se um novo agrupamento. Segundo Lance e Williams (1967), a distância pode ser obtida por meio da seguinte expressão:

$$d_{k(ij)} = \left(\frac{1}{2}\right) d_{ki} + \left(\frac{1}{2}\right) d_{kj} - \left(\frac{1}{4}\right) d_{ij}$$

Em que: $\mathbf{d_{k(ij)}}$, $\mathbf{d_{ki}}$, $\mathbf{d_{kj}}$ e $\mathbf{d_{ij}}$ = distâncias euclidianas entre os elementos **k** e agrupamento **ij**, **k** e **i**, **k** e **j**, e **i** e **j**, respectivamente; e $\mathbf{n_i}$, $\mathbf{n_j}$ e $\mathbf{n_k}$ = número de elementos nos agrupamentos **i**, **j** e **k**, respectivamente.

4. Verifica-se se o número de agrupamentos determinados é igual ao valor fixado, $\mathbf{g \leq n}$. Se a condição for verdadeira, termina-se o processo, caso contrário, retorna-se ao item 2.

**Aplicação:**

Consideremos a matriz de distâncias euclidianas quadradas $D^2_1$:

$\mathbf{D^2_1} =$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,0 | 317,0 | 529,0 | 156,0 | 1151,0 | 300,0 | 342,0 | 127,0 | 615,0 | 399,0 | 217,0 |
| 2 | | 0,0 | 1002,0 | 143,0 | 1180,0 | 475,0 | 335,0 | 262,0 | 866,0 | 656,0 | 348,0 |
| 3 | | | 0,00 | 875,0 | 1748,0 | 731,0 | 955,0 | 780,0 | 504,0 | 322,0 | 586,0 |
| 4 | | | | 0,0 | 1041,0 | 298,0 | 216,0 | **59,0** | 801,0 | 463,0 | 183,0 |
| 5 | | | | | 0,0 | 821,0 | 1493,0 | 1264,0 | 1716,0 | 1212,0 | 1202,0 |
| 6 | | | | | | 0,0 | 604,0 | 361,0 | 761,0 | 463,0 | 375,0 |
| 7 | | | | | | | 0,0 | 241,0 | 503,0 | 597,0 | 77,0 |
| 8 | | | | | | | | 0,0 | 714,0 | 448,0 | 176,0 |
| 9 | | | | | | | | | 0,0 | 520,0 | 370,0 |
| 10 | | | | | | | | | | 0,0 | 308,0 |
| 11 | | | | | | | | | | | 0,0 |

Os seguintes passos serão considerados:

Passo 1: Objetos mais similares – parcelas 4 e 8.

Distância entre os objetos: 59.

Cálculo das distâncias entre os outros elementos e o grupo (4,8):

$$d_{1(4,8)} = \left(\frac{1}{2}\right) x156{,}0 + \left(\frac{1}{2}\right) x127{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 126{,}8$$

$$d_{2(4,8)} = \left(\frac{1}{2}\right) x143{,}0 + \left(\frac{1}{2}\right) x262{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 187{,}8$$

$$d_{3(4,8)} = \left(\frac{1}{2}\right) x875{,}0 + \left(\frac{1}{2}\right) x780{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 812{,}8$$

$$d_{5(4,8)} = \left(\frac{1}{2}\right) x1041{,}0 + \left(\frac{1}{2}\right) x1264{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 1137{,}8$$

$$d_{6(4,8)} = \left(\frac{1}{2}\right) x298{,}0 + \left(\frac{1}{2}\right) x361{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 314{,}8$$

$$d_{7(4,8)} = \left(\frac{1}{2}\right) x216{,}0 + \left(\frac{1}{2}\right) x241{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 213{,}8$$

$$d_{9(4,8)} = \left(\frac{1}{2}\right) x801{,}0 + \left(\frac{1}{2}\right) x714{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 742{,}8$$

$$d_{10(4,8)} = \left(\frac{1}{2}\right) x463{,}0 + \left(\frac{1}{2}\right) x448{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 440{,}8$$

$$d_{11(4,8)} = \left(\frac{1}{2}\right) x183{,}0 + \left(\frac{1}{2}\right) x176{,}0 - \left(\frac{1}{4}\right) x59{,}0 = 164{,}8$$

Nova Matriz de Distância: $D^2_2$

$\mathbf{D^2_2} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,0 | 317,0 | 529,0 | 126,8 | 1151,0 | 300,0 | 342,0 | 615,0 | 399,0 | 217,0 |
| 2 | | 0,0 | 1002,0 | 187,8 | 1180,0 | 475,0 | 335,0 | 866,0 | 656,0 | 348,0 |
| 3 | | | 0,00 | 812,8 | 1748,0 | 731,0 | 955,0 | 504,0 | 322,0 | 586,0 |
| (4,8) | | | | 0,0 | 1137,8 | 314,8 | 213,8 | 742,8 | 440,8 | 164,8 |
| 5 | | | | | 0,0 | 821,0 | 1493,0 | 1716,0 | 1212,0 | 1202,0 |
| 6 | | | | | | 0,0 | 604,0 | 761,0 | 463,0 | 375,0 |
| 7 | | | | | | | 0,0 | 503,0 | 597,0 | **77,0** |
| 9 | | | | | | | | 0,0 | 520,0 | 370,0 |
| 10 | | | | | | | | | 0,0 | 308,0 |
| 11 | | | | | | | | | | 0,0 |

Passo 2: Objetos mais similares – parcela 7 e parcela 11.

Distância entre os objetos: 77,0.

Cálculo das distâncias entre os outros elementos e o grupo (7,11):

$$d_{1(7,11)} = \left(\frac{1}{2}\right) x342{,}0 + \left(\frac{1}{2}\right) x217{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 260{,}3$$

$$d_{2(7,11)} = \left(\frac{1}{2}\right) x335{,}0 + \left(\frac{1}{2}\right) x348{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 322{,}3$$

$$d_{3(7,11)} = \left(\frac{1}{2}\right) x955{,}0 + \left(\frac{1}{2}\right) x586{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 751{,}3$$

$$d_{(4,8)(7,11)} = \left(\frac{1}{2}\right) x213{,}8 + \left(\frac{1}{2}\right) x164{,}8 - \left(\frac{1}{4}\right) x77{,}0 = 170{,}0$$

$$d_{5(7,11)} = \left(\frac{1}{2}\right) x1493{,}0 + \left(\frac{1}{2}\right) x1202{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 1328{,}3$$

$$d_{6(7,11)} = \left(\frac{1}{2}\right) x604{,}0 + \left(\frac{1}{2}\right) x375{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 470{,}3$$

$$d_{9(7,11)} = \left(\frac{1}{2}\right) x503{,}0 + \left(\frac{1}{2}\right) x370{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 417{,}3$$

$$d_{10(7,11)} = \left(\frac{1}{2}\right) x597{,}0 + \left(\frac{1}{2}\right) x308{,}0 - \left(\frac{1}{4}\right) x77{,}0 = 433{,}3$$

Nova Matriz de Distância: $D^2_3$

$\mathbf{D^2_3} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,0 | 317,0 | 529,0 | **126,8** | 1151,0 | 300,0 | 260,3 | 615,0 | 399,0 |
| 2 | | 0,0 | 1002,0 | 187,8 | 1180,0 | 475,0 | 322,3 | 866,0 | 656,0 |
| 3 | | | 0,00 | 812,8 | 1748,0 | 731,0 | 751,3 | 504,0 | 322,0 |
| (4,8) | | | | 0,0 | 1137,8 | 314,8 | 170,0 | 742,8 | 440,8 |
| 5 | | | | | 0,0 | 821,0 | 1328,3 | 1716,0 | 1212,0 |
| 6 | | | | | | 0,0 | 470,3 | 761,0 | 463,0 |
| (7,11) | | | | | | | 0,0 | 417,3 | 433,3 |
| 9 | | | | | | | | 0,0 | 520,0 |
| 10 | | | | | | | | | 0,0 |

Passo 3: Objetos mais similares – parcela 1 e grupo (4,8).

Distância entre os objetos: 126,8.

Cálculo das distâncias entre os outros elementos e o grupo (1,4,8):

$$d_{2(1,4,8)} = \left(\frac{1}{2}\right) x317{,}0 + \left(\frac{1}{2}\right) x187{,}8 - \left(\frac{1}{4}\right) x126{,}8 = 220{,}7$$

$$d_{3(1,4,8)} = \left(\frac{1}{2}\right) x529{,}0 + \left(\frac{1}{2}\right) x812{,}8 - \left(\frac{1}{4}\right) x126{,}8 = 639{,}2$$

$$d_{5(1,4,8)} = \left(\frac{1}{2}\right) x1151{,}0 + \left(\frac{1}{2}\right) x1137{,}8 - \left(\frac{1}{4}\right) x126{,}8 = 1112{,}7$$

$$d_{6(1,4,8)} = \left(\frac{1}{2}\right) x300{,}0 + \left(\frac{1}{2}\right) x314{,}8 - \left(\frac{1}{4}\right) x126{,}8 = 275{,}7$$

$$d_{(7,11)(1,4,8)} = \left(\frac{1}{2}\right) x260{,}3 + \left(\frac{1}{2}\right) x170{,}0 - \left(\frac{1}{4}\right) x126{,}8 = 183{,}4$$

$$d_{9(1,4,8)} = \left(\frac{1}{2}\right) x615{,}0 + \left(\frac{1}{2}\right) x742{,}8 - \left(\frac{1}{4}\right) x126{,}8 = 647{,}2$$

$$d_{10(1,4,8)} = \left(\frac{1}{2}\right) x399{,}0 + \left(\frac{1}{2}\right) x440{,}8 - \left(\frac{1}{4}\right) x126{,}8 = 388{,}2$$

Nova Matriz de Distância: $D^2_4$

$\mathbf{D^2_4} =$

| | (1,4,8) | 2 | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| (1,4,8) | 0,0 | 220,7 | 639,2 | 1112,7 | 275,7 | **183,4** | 647,2 | 388,2 |
| 2 | | 0,0 | 1002,0 | 1180,0 | 475,0 | 322,3 | 866,0 | 656,0 |
| 3 | | | 0,00 | 1748,0 | 731,0 | 751,3 | 504,0 | 322,0 |
| 5 | | | | 0,0 | 821,0 | 1328,3 | 1716,0 | 1212,0 |
| 6 | | | | | 0,0 | 470,3 | 761,0 | 463,0 |
| (7,11) | | | | | | 0,0 | 417,3 | 433,3 |
| 9 | | | | | | | 0,0 | 520,0 |
| 10 | | | | | | | | 0,0 |

Passo 4: Objetos mais similares – grupos (1,4,8) e (7,11).

Distância entre os objetos: 183,4.

Cálculo das distâncias entre os demais elementos e o grupo (1,4,7,8,11):

$$d_{2(1,4,7,8,11)} = \left(\frac{1}{2}\right) x220{,}7 + \left(\frac{1}{2}\right) x322{,}3 - \left(\frac{1}{4}\right) x183{,}4 = 225{,}6$$

$$d_{3(1,4,7,8,11)} = \left(\frac{1}{2}\right) x639{,}2 + \left(\frac{1}{2}\right) x751{,}3 - \left(\frac{1}{4}\right) x183{,}4 = 649{,}4$$

$$d_{5(1,4,7,8,11)} = \left(\frac{1}{2}\right) x1112{,}7 + \left(\frac{1}{2}\right) x1328{,}3 - \left(\frac{1}{4}\right) x183{,}4 = 1174{,}6$$

$$d_{6(1,4,7,8,11)} = \left(\frac{1}{2}\right) x275{,}7 + \left(\frac{1}{2}\right) x470{,}3 - \left(\frac{1}{4}\right) x183{,}4 = 327{,}1$$

$$d_{9(1,4,7,8,11)} = \left(\frac{1}{2}\right) x647{,}2 + \left(\frac{1}{2}\right) x417{,}3 - \left(\frac{1}{4}\right) x183{,}4 = 486{,}4$$

$$d_{10(1,4,7,8,11)} = \left(\frac{1}{2}\right) x388{,}2 + \left(\frac{1}{2}\right) x433{,}3 - \left(\frac{1}{4}\right) x183{,}4 = 369{,}4$$

Nova Matriz de Distância: $D^2_5$

$\mathbf{D^2_5} =$

| | (1,4,7,8,11) | 2 | 3 | 5 | 6 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| (1,4,7,8,11) | 0,0 | **225,6** | 649,4 | 1174,6 | 327,1 | 486,4 | 369,4 |
| 2 | | 0,0 | 1002,0 | 1180,0 | 475,0 | 866,0 | 656,0 |
| 3 | | | 0,00 | 1748,0 | 731,0 | 504,0 | 322,0 |
| 5 | | | | 0,0 | 821,0 | 1716,0 | 1212,0 |
| 6 | | | | | 0,0 | 761,0 | 463,0 |
| 9 | | | | | | 0,0 | 520,0 |
| 10 | | | | | | | 0,0 |

Passo 5: Objetos mais similares – grupo (1,4,7,8,11) e parcela 2.

Distância entre os objetos: 225,6.

Cálculo das distâncias entre os demais elementos e o grupo (1,2,4,7,8,11):

$$d_{3(1,2,4,7,8,11)} = \left(\frac{1}{2}\right) x649{,}5 + \left(\frac{1}{2}\right) x1002{,}0 - \left(\frac{1}{4}\right) x225{,}6 = 769{,}3$$

$$d_{5(1,2,4,7,8,11)} = \left(\frac{1}{2}\right) x1174{,}6 + \left(\frac{1}{2}\right) x1180{,}0 - \left(\frac{1}{4}\right) x225{,}6 = 1120{,}9$$

$$d_{6(1,2,4,7,8,11)} = \left(\frac{1}{2}\right) x327{,}1 + \left(\frac{1}{2}\right) x475{,}0 - \left(\frac{1}{4}\right) x225{,}6 = 344{,}7$$

$$d_{9(1,2,4,7,8,11)} = \left(\frac{1}{2}\right) x486{,}4 + \left(\frac{1}{2}\right) x866{,}0 - \left(\frac{1}{4}\right) x225{,}6 = 619{,}8$$

$$d_{10(1,2,4,7,8,11)} = \left(\frac{1}{2}\right) x369{,}4 + \left(\frac{1}{2}\right) x656{,}0 - \left(\frac{1}{4}\right) x225{,}6 = 454{,}0$$

Nova Matriz de Distância: $D^2_6$

$\mathbf{D^2_6} =$

| | (1,2,4,7,8,11) | 3 | 5 | 6 | 9 | 10 |
|---|---|---|---|---|---|---|
| (1,2,4,7,8,11) | 0,0 | 769,3 | 1120,9 | 344,7 | 619,8 | 454,0 |
| 3 | | 0,00 | 1748,0 | 731,0 | 504,0 | **322,0** |
| 5 | | | 0,0 | 821,0 | 1716,0 | 1212,0 |
| 6 | | | | 0,0 | 761,0 | 463,0 |
| 9 | | | | | 0,0 | 520,0 |
| 10 | | | | | | 0,0 |

Passo 6: Objetos mais similares – parcelas 3 e 10.

Distância entre os objetos: 322,0.

Cálculo das distâncias entre os demais elementos e o grupo (3,10):

$$d_{(1,2,4,7,8,11)(3,10)} = \left(\frac{1}{2}\right) x769{,}3 + \left(\frac{1}{2}\right) x454{,}0 - \left(\frac{1}{4}\right) x322{,}0 = 531{,}2$$

$$d_{5(3,10)} = \left(\frac{1}{2}\right) x1748{,}0 + \left(\frac{1}{2}\right) x1212{,}0 - \left(\frac{1}{4}\right) x322{,}0 = 1399{,}5$$

$$d_{6(3,10)} = \left(\frac{1}{2}\right) x731{,}0 + \left(\frac{1}{2}\right) x463{,}0 - \left(\frac{1}{4}\right) x322{,}0 = 516{,}5$$

$$d_{9(3,10)} = \left(\frac{1}{2}\right) x504{,}0 + \left(\frac{1}{2}\right) x520{,}0 - \left(\frac{1}{4}\right) x322{,}0 = 431{,}5$$

Nova Matriz de Distância: $D^2_7$

$$\mathbf{D^2_7} = \begin{array}{c|ccccc} & (1,2,4,7,8,11) & (3,10) & 5 & 6 & 9 \\ (1,2,4,7,8,11) & 0,0 & 531,2 & 1120,9 & \mathbf{344,7} & 619,8 \\ (3,10) & & 0,00 & 1399,5 & 516,5 & 431,5 \\ 5 & & & 0,0 & 821,0 & 1716,0 \\ 6 & & & & 0,0 & 761,0 \\ 9 & & & & & 0,0 \end{array}$$

Passo 7: Objetos mais similares – grupo (1,2,4,7,8,11) e parcela 6.

Distância entre os objetos: 344,7.

Cálculo das distâncias entre os demais elementos e o grupo (1,2,4,6,7,8,10,11):

$$d_{(3,10)(1,2,4,6,7,8,11)} = \left(\frac{1}{2}\right) x531,2 + \left(\frac{1}{2}\right) x516,5 - \left(\frac{1}{4}\right) x344,7 = 437,7$$

$$d_{5(1,2,4,6,7,8,11)} = \left(\frac{1}{2}\right) x1120,9 + \left(\frac{1}{2}\right) x821,0 - \left(\frac{1}{4}\right) x344,7 = 884,8$$

$$d_{9(1,2,4,6,7,8,11)} = \left(\frac{1}{2}\right) x619,8 + \left(\frac{1}{2}\right) x761,0 - \left(\frac{1}{4}\right) x344,7 = 604,2$$

Nova Matriz de Distância: $D^2_8$

$$\mathbf{D^2_8} = \begin{array}{c|cccc} & (1,2,4,6,7,8,11) & (3,10) & 5 & 9 \\ (1,2,4,6,7,8,11) & 0,0 & 437,7 & 884,8 & 604,2 \\ (3,10) & & 0,00 & 1399,5 & \mathbf{431,5} \\ 5 & & & 0,0 & 1716,0 \\ 9 & & & & 0,0 \end{array}$$

Passo 8: Objetos mais similares – grupo (3,10) e parcela 9.

Distância entre os objetos: 431,5.

Cálculo das distâncias entre os demais elementos e o grupo (3,9,10):

$$d_{(1,2,4,6,7,8,11)(3,9,10)} = \left(\frac{1}{2}\right) x437,7 + \left(\frac{1}{2}\right) x604,2 - \left(\frac{1}{4}\right) x431,5 = 413,1$$

$$d_{5(3,9,10)} = \left(\frac{1}{2}\right) x1399,5 + \left(\frac{1}{2}\right) x1716,0 - \left(\frac{1}{4}\right) x431,5 = 1449,9$$

Nova Matriz de Distância: $D^2_9$

$$\mathbf{D^2_9} = \begin{array}{c|ccc} & (1,2,4,6,7,8,11) & (3,9,10) & 5 \\ (1,2,4,6,7,8,11) & 0,0 & \mathbf{413,1} & 884,8 \\ (3,9,10) & & 0,00 & 1449,9 \\ 5 & & & 0,0 \end{array}$$

Passo 9: Objetos mais similares – grupo (1,2,4,6,7,8,11) e parcelas (3,9,10).

Distância entre os objetos: 413,1.

Cálculo das distâncias entre a parcela 5 e o grupo (1,2,3,4,6,7,8,9,10,11):

$$d_{5(1,2,3,4,6,7,8,9,10,11)} = \left(\frac{1}{2}\right) x884{,}8 + \left(\frac{1}{2}\right) x1449{,}9 - \left(\frac{1}{4}\right) x413{,}1 = 1604{,}1$$

Com base nos cálculos, os resultados estão resumidos na Tabela 25 e apresentados como um dendrograma na Figura 11 (saída do R).

**Tabela 25.** Resumo da análise de agrupamento obtido pelo emprego do **método da mediana**, com base na distância euclidiana, mostrando a sequência das fusões das parcelas

| Passo | Distância Euclidiana | | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Quadrada | Simples | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 1 | 59,0 | 7,7 | P4 | P8 | | | | | | | | | |
| 2 | 77,0 | 8,8 | P7 | P11 | | | | | | | | | |
| 3 | 126,8 | 11,3 | P4 | P8 | P1 | | | | | | | | |
| 4 | 183,4 | 13,5 | P4 | P8 | P1 | P7 | P11 | | | | | | |
| 5 | 225,6 | 15 | P4 | P8 | P1 | P7 | P11 | P2 | | | | | |
| 6 | 322,0 | 17,9 | P3 | P10 | | | | | | | | | |
| 7 | 344,7 | 18,6 | P4 | P8 | P1 | P7 | P11 | P2 | P6 | | | | |
| 8 | 431,5 | 20,8 | P3 | P10 | P9 | | | | | | | | |
| 9 | 413,1 | 20,3 | P4 | P8 | P1 | P7 | P11 | P2 | P6 | P3 | P10 | P9 | |
| 10 | 1064,1 | 32,6 | P4 | P8 | P1 | P7 | P11 | P2 | P6 | P3 | P10 | P9 | P5 |

## Passo a passo no R: Agrupamento por Mediana (WPGMC)

### 1. Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust*(). O método "*median*" (mediana) agrupa as parcelas com base na distância entre as medianas dos grupos. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

**Nota Técnica Importante:** Assim como no método do centroide, o R exige que as distâncias sejam elevadas ao quadrado (^2) antes de serem passadas para a função *hclust*() com o método "*median*". Isso é necessário para manter a consistência geométrica do método. Após o agrupamento, ajustamos a altura (*height*) do objeto *hclust*, aplicando a raiz quadrada, para que os níveis de distância no dendrograma sejam comparáveis com as distâncias originais e com a Linha Fenon.

```
# Execução do Agrupamento Hierárquico pelo método da Mediana
# IMPORTANTE: O método median requer distâncias ao quadrado
agrupamento_mediana <- hclust(as.dist(matriz_euclidiana)^2, method = "median")

# Ajustamos a altura (height) para a raiz quadrada para que os níveis de distância sejam comparáveis
agrupamento_mediana$height <- sqrt(agrupamento_mediana$height)
```

## 2. Detalhamento das Fusões (Passo a Passo)

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância
resumo_mediana <- data.frame(
  Passo = 1:nrow(agrupamento_mediana$merge),
  Distancia = round(agrupamento_mediana$height, 1)
)

# Exibindo o resumo das fusões
print(resumo_mediana)
```

| Passo | Distância |
|---|---|
| 1 | 7.7 |
| 2 | 8.8 |
| 3 | 11.3 |
| 4 | 13.5 |
| 5 | 15.0 |
| 6 | 17.9 |
| 7 | 18.6 |
| 8 | 20.8 |
| 9 | 20.3 |
| 10 | 32.6 |

## 3. Visualização do Dendrograma

O dendrograma é a representação gráfica do agrupamento hierárquico (Figura 11). Ele mostra visualmente como as parcelas foram se unindo e formando grupos em diferentes níveis de distância. No método da mediana, é possível que ocorram inversões, o que pode causar cruzamentos visuais nos “galhos” do dendrograma. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes, além de personalizar os eixos para melhor visualização.

```
# Configurando a margem para caber o título inferior
par(mar=c(6, 4, 4, 2))

plot(agrupamento_mediana,
     main = "Método da Mediana (WPGMC)",
     xlab = "Sequência de Agrupamento de Parcelas\nhclust (*, \"median\")",
     ylab = "Distância Euclidiana", sub = "", hang = -1, axes = FALSE)

# Adicionando o eixo Y lateral
axis(2, at = seq(0, 40, by = 10))

# Adicionando a linha de corte (em vermelho) em d = 18.0
```

```
abline(h = 18.0, col = "red", lty = 2, lwd = 2)

# Adicionando os retângulos coloridos para os 4 grupos principais
rect.hclust(agrupamento_mediana, k = 4, border = c("blue", "green", "orange", "purple"))
```

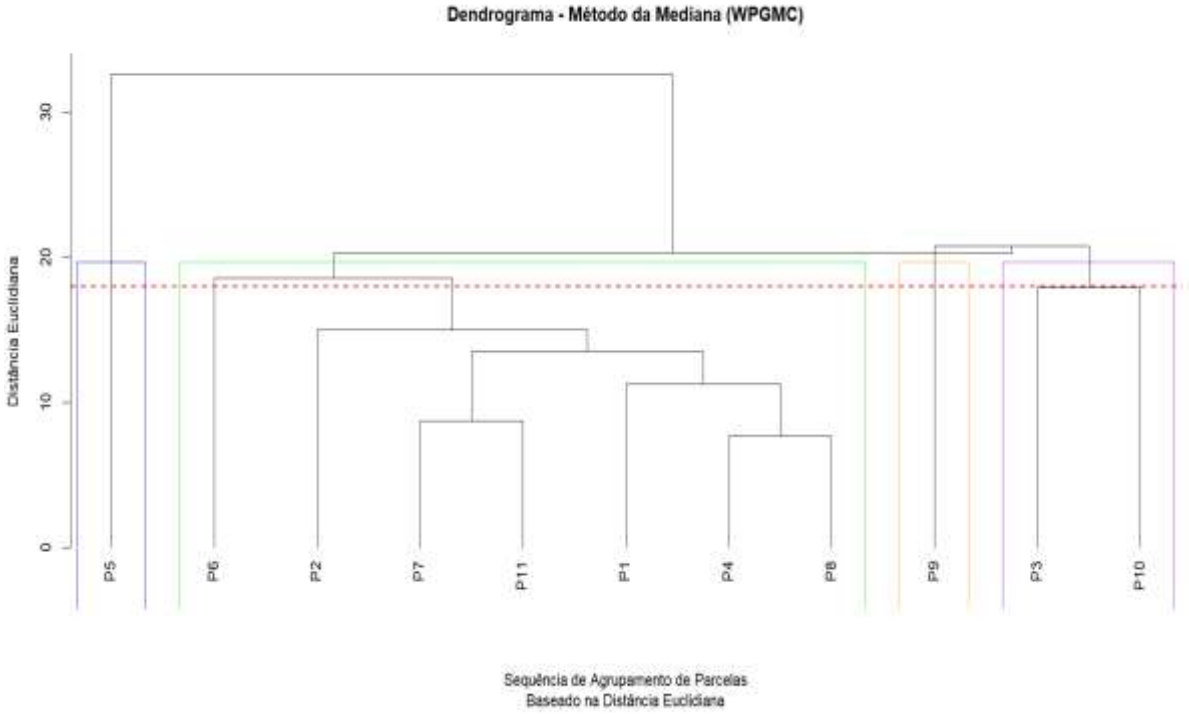



**Figura 11.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método da mediana, com base na distância euclidiana, considerando a linha fenon d = 18,0 e a formação de quatro grupos.

### e. Método da Ligação Média entre Grupos não Ponderado (UPGMA)

Esse método, mais conhecido como UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*), define a distância entre dois grupos como a média das distâncias entre todos os pares de elementos, sendo um em cada grupo (Sokal; Michener, 1958). Esse procedimento pode ser utilizado tanto para medidas de similaridade como de distância, contanto que o conceito de uma medida média seja aceitável (Everitt et al., 2011).

Os grupos são reunidos em um novo grupo quando a média das distâncias entre seus elementos é mínima.

O método pode ser resumido nos seguintes passos:

1. Determina-se a matriz de distâncias inicial.

2. Localizam-se os dois elementos que apresentam a menor distância, reunindo-os em um único grupo.

3. Calcula-se a distância entre os diversos pares de grupos como sendo a média das distâncias entre todos os pares de seus elementos, sendo um elemento de cada um dos grupos. Esaa distância pode ser obtida por meio da seguinte expressão (Lance; Williams, 1967):

$$\boldsymbol{d^*_{(i,j)k}} = \left[\frac{N_i}{N_i + N_j}\right] d_{ik} + \left[\frac{N_j}{N_i + N_j}\right] d_{jk}$$

4. Os dois grupos que apresentam menor distância média serão reunidos em um único grupo.

5. Se o número de grupos obtidos é igual a um número **g** ≤ **n**, o processo termina, caso contrário, retorna-se ao passo 3.

**Aplicação**

Considerando a matriz de distância euclidiana ($D_1$), devem-se seguir os seguintes passos:

$\mathbf{D_1} =$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 12,5 | 33,9 | 17,3 | 18,5 | 11,3 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 12,0 | 34,4 | 21,8 | 18,3 | 16,2 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 29,6 | 41,8 | 27,0 | 30,9 | 27,9 | 22,4 | 17,9 | 24,2 |
| 4 | | | | 0,00 | 32,3 | 17,3 | 14,7 | **7,7** | 28,3 | 21,5 | 13,5 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 35,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 19,0 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 15,5 | 22,4 | 24,4 | 8,8 |
| 8 | | | | | | | | 0,00 | 26,7 | 21,2 | 13,3 |
| 9 | | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | | 0,00 |

N = [ 1 1 1 1 1 1 1 1 1 1 1 ]

Passo 1: Procura-se, em $D_1$, a menor distância euclidiana. Este valor corresponde a $d_{(4,8)} = 7,7$.

Cálculo da matriz de dissimilaridade reduzida $D_2$:

$$d^*_{(4,8)k} = \left[\frac{N_4}{N_4 + N_8}\right] d_{4k} + \left[\frac{N_8}{N_4 + N_8}\right] d_{8k}$$

$$d^*_{1(4,8)} = \left[\frac{1}{1+1}\right] x12,5 + \left[\frac{1}{1+1}\right] x11,3 = 11,9$$

$$d^*_{2(4,8)} = \left[\frac{1}{1+1}\right] x12,0 + \left[\frac{1}{1+1}\right] x16,2 = 14,1$$

$$d^*_{3(4,8)} = \left[\frac{1}{1+1}\right] x29,6 + \left[\frac{1}{1+1}\right] x27,9 = 28,8$$

$$d^*_{5(4,8)} = \left[\frac{1}{1+1}\right] x32{,}3 + \left[\frac{1}{1+1}\right] x35{,}6 = 33{,}9$$

$$d^*_{6(4,8)} = \left[\frac{1}{1+1}\right] x17{,}3 + \left[\frac{1}{1+1}\right] x19{,}0 = 18{,}1$$

$$d^*_{7(4,8)} = \left[\frac{1}{1+1}\right] x14{,}7 + \left[\frac{1}{1+1}\right] x15{,}5 = 15{,}1$$

$$d^*_{9(4,8)} = \left[\frac{1}{1+1}\right] x28{,}3 + \left[\frac{1}{1+1}\right] x26{,}7 = 27{,}5$$

$$d^*_{10(4,8)} = \left[\frac{1}{1+1}\right] x21{,}5 + \left[\frac{1}{1+1}\right] x21{,}2 = 21{,}3$$

$$d^*_{11(4,8)} = \left[\frac{1}{1+1}\right] x13{,}5 + \left[\frac{1}{1+1}\right] x13{,}3 = 13{,}4$$

$\mathbf{D_2} =$

| | 1 | 2 | 3 | (4,8) | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 11,9 | 33,9 | 17,3 | 18,5 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 14,1 | 34,4 | 21,8 | 18,3 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 28,8 | 41,8 | 27,0 | 30,9 | 22,4 | 17,9 | 24,2 |
| (4,8) | | | | 0,00 | 33,9 | 18,1 | 15,1 | 27,5 | 21,3 | 13,4 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 22,4 | 24,4 | **8,8** |
| 9 | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | 0,00 |

| N = | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|

Passo 2: Procura-se, em $D_2$, a menor distância euclidiana. Este valor corresponde a $d_{(7,11)} = 8{,}8$.

Cálculo da matriz de dissimilaridade reduzida $D_3$:

$$d^*_{j(7,11)} = \left[\frac{N_7}{N_7 + N_{11}}\right] d_{j7} + \left[\frac{N_{11}}{N_7 + N_{11}}\right] d_{j11}$$

$$d^*_{1(7,11)} = \left[\frac{1}{1+1}\right] x18{,}5 + \left[\frac{1}{1+1}\right] x14{,}7 = 16{,}6$$

$$d^*_{2(7,11)} = \left[\frac{1}{1+1}\right] x18{,}3 + \left[\frac{1}{1+1}\right] x18{,}7 = 18{,}5$$

$$d^*_{3(7,11)} = \left[\frac{1}{1+1}\right] x30{,}9 + \left[\frac{1}{1+1}\right] x24{,}2 = 27{,}6$$

$$d^*_{(4,8)(7,11)} = \left[\frac{1}{1+1}\right] x15{,}1 + \left[\frac{1}{1+1}\right] x13{,}4 = 14{,}3$$

$$d^*_{5(7,11)} = \left[\frac{1}{1+1}\right] x38{,}6 + \left[\frac{1}{1+1}\right] x34{,}7 = 36{,}7$$

$$d^*_{6(7,11)} = \left[\frac{1}{1+1}\right] x24{,}6 + \left[\frac{1}{1+1}\right] x19{,}4 = 22{,}0$$

$$d^*_{9(7,11)} = \left[\frac{1}{1+1}\right] x22{,}4 + \left[\frac{1}{1+1}\right] x19{,}2 = 20{,}8$$

$$d^*_{10(7,11)} = \left[\frac{1}{1+1}\right] x24{,}4 + \left[\frac{1}{1+1}\right] x17{,}5 = 21{,}0$$

**$D_3$ =**

| | 1 | 2 | 3 | (4,8) | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | **11,9** | 33,9 | 17,3 | 16,6 | 24,8 | 20,0 |
| 2 | | 0,00 | 31,7 | 14,1 | 34,4 | 21,8 | 18,5 | 29,4 | 25,6 |
| 3 | | | 0,00 | 28,8 | 41,8 | 27,0 | 27,6 | 22,4 | 17,9 |
| (4,8) | | | | 0,00 | 33,9 | 18,1 | 14,3 | 27,5 | 21,3 |
| 5 | | | | | 0,00 | 28,7 | 36,7 | 41,4 | 34,8 |
| 6 | | | | | | 0,00 | 22,0 | 27,6 | 21,5 |
| (7,11) | | | | | | | 0,00 | 20,8 | 21,0 |
| 9 | | | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | | | 0,00 |

| N = | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|

Passo 3: Procura-se, em $D_3$, a menor distância euclidiana. Este valor corresponde a $d_{(1,4,8)}$ = 11,9.

Cálculo da matriz de dissimilaridade reduzida $D_4$:

$$d^*_{j(1,4,8)} = \left[\frac{N_1}{N_1 + N_{(4,8)}}\right] d_{j1} + \left[\frac{N_{(4,8)}}{N_1 + N_{(4,8)}}\right] d_{j(4,8)}$$

$$d^*_{2(1,4,8)} = \left[\frac{1}{1+2}\right] x17{,}8 + \left[\frac{2}{1+2}\right] x14{,}1 = 15{,}3$$

$$d^*_{3(1,4,8)} = \left[\frac{1}{1+2}\right] x23{,}0 + \left[\frac{2}{1+2}\right] x28{,}8 = 26{,}8$$

$$d^*_{5(1,4,8)} = \left[\frac{1}{1+2}\right] x33{,}9 + \left[\frac{2}{1+2}\right] x34{,}0 = 33{,}9$$

$$d^*_{6(1,4,8)} = \left[\frac{1}{1+2}\right] x17{,}3 + \left[\frac{2}{1+2}\right] x18{,}2 = 17{,}9$$

$$d^*_{(7,11)(1,4,8)} = \left[\frac{1}{1+2}\right] x16{,}6 + \left[\frac{2}{1+2}\right] x14{,}3 = 15{,}1$$

$$d^*_{9(1,4,8)} = \left[\frac{1}{1+2}\right] x24{,}8 + \left[\frac{2}{1+2}\right] x27{,}5 = 26{,}6$$

$$d^*_{10(1,4,8)} = \left[\frac{1}{1+2}\right] x20{,}0 + \left[\frac{2}{1+2}\right] x21{,}4 = 20{,}9$$

$\mathbf{D_4}=$

| | (1,4,8) | 2 | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| (1,4,8) | 0,00 | 15,3 | 26,8 | 33,9 | 17,9 | **15,1** | 26,6 | 20,9 |
| 2 | | 0,00 | 31,7 | 34,4 | 21,8 | 18,5 | 29,4 | 25,6 |
| 3 | | | 0,00 | 41,8 | 27,0 | 27,6 | 22,4 | 17,9 |
| 5 | | | | 0,00 | 28,7 | 36,7 | 41,4 | 34,8 |
| 6 | | | | | 0,00 | 22,0 | 27,6 | 21,5 |
| (7,11) | | | | | | 0,00 | 20,8 | 21,0 |
| 9 | | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | | 0,00 |

N = [ 3 1 1 1 1 2 1 1 ]

Passo 4: Procura-se, em $D_4$, a menor distância euclidiana. Este valor corresponde a $d_{(1,4,7,8,11)}$ = 15,1.

Cálculo da matriz de dissimilaridade reduzida $D_5$:

$$d^*_{j(1,4,7,8,11)} = \left[\frac{N_{(1,4,8)}}{N_{(1,4,8)} + N_{(4,8)}}\right] d_{j(1,4,8)} + \left[\frac{N_{(7,11)}}{N_{(1,4,8)} + N_{(7,11)}}\right] d_{j(7,11)}$$

$$d^*_{2(1,4,7,8,11)} = \left[\frac{3}{3+2}\right] x15{,}3 + \left[\frac{2}{3+2}\right] x18{,}5 = 16{,}6$$

$$d^*_{3(1,4,7,8,11)} = \left[\frac{3}{3+2}\right] x26{,}9 + \left[\frac{2}{3+2}\right] x27{,}6 = 27{,}1$$

$$d^*_{5(1,4,7,8,11)} = \left[\frac{3}{3+2}\right] x34{,}0 + \left[\frac{2}{3+2}\right] x36{,}7 = 35{,}0$$

$$d^*_{6(1,4,7,8,11)} = \left[\frac{3}{3+2}\right] x17{,}9 + \left[\frac{2}{3+2}\right] x22{,}0 = 19{,}5$$

$$d^*_{9(1,4,7,8,11)} = \left[\frac{3}{3+2}\right] x26{,}6 + \left[\frac{2}{3+2}\right] x20{,}8 = 24{,}3$$

$$d^*_{10(1,4,7,8,11)} = \left[\frac{3}{3+2}\right] x20{,}9 + \left[\frac{2}{3+2}\right] x21{,}0 = 20{,}9$$

| | | (1,4,7,8,11) | 2 | 3 | 5 | 6 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| | (1,4,7,8,11) | 0,00 | **16,6** | 27,1 | 35,0 | 19,5 | 24,3 | 20,9 |
| | 2 | | 0,00 | 31,7 | 34,4 | 21,8 | 29,4 | 25,6 |
| | 3 | | | 0,00 | 41,8 | 27,0 | 22,4 | 17,9 |
| | 5 | | | | 0,00 | 28,7 | 41,4 | 34,8 |
| **$D_5$ =** | 6 | | | | | 0,00 | 27,6 | 21,5 |
| | 9 | | | | | | 0,00 | 22,8 |
| | 10 | | | | | | | 0,00 |
| N | = | 5 | 1 | 1 | 1 | 1 | 1 | 1 |

Passo 5: Procura-se, em $D_5$, a menor distância euclidiana. Este valor corresponde a $d_{(1,2,4,7,8,11)}$ = 16,6.

Cálculo da matriz de dissimilaridade reduzida $D_6$:

$$d^*_{j(1,,2,4,7,8,11)} = \left[\frac{N_2}{N_2 + N_{(1,4,7,8,11)}}\right] d_{j2} + \left[\frac{N_{(1,4,7,8,11)}}{N_2 + N_{(1,4,7,8,11)}}\right] d_{j(1,4,7,8,11)}$$

$$d^*_{3(1,2,4,7,8,11)} = \left[\frac{1}{1+5}\right] x31,7 + \left[\frac{5}{1+5}\right] x27,2 = 27,9$$

$$d^*_{5(1,2,4,7,8,11)} = \left[\frac{1}{1+5}\right] x34,0 + \left[\frac{5}{1+5}\right] x35,1 = 34,9$$

$$d^*_{6(1,2,4,7,8,11)} = \left[\frac{1}{1+5}\right] x21,8 + \left[\frac{5}{1+5}\right] x19,5 = 19,9$$

$$d^*_{9(1,2,4,7,8,11)} = \left[\frac{1}{1+5}\right] x29,4 + \left[\frac{5}{1+5}\right] x24,3 = 25,2$$

$$d^*_{10(1,2,4,7,8,11)} = \left[\frac{1}{1+5}\right] x25,6 + \left[\frac{5}{1+5}\right] x20,3 = 21,7$$

| | | (1,2,4,7,8,11) | 3 | 5 | 6 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| | (1,2,4,7,8,11) | 0,00 | 27,9 | 34,9 | 19,9 | 25,2 | 21,7 |
| | 3 | | 0,00 | 41,8 | 27,0 | 22,4 | **17,9** |
| | 5 | | | 0,00 | 28,7 | 41,4 | 34,8 |
| **$D_6$ =** | 6 | | | | 0,00 | 27,6 | 21,5 |
| | 9 | | | | | 0,00 | 22,8 |
| | 10 | | | | | | 0,00 |
| N | = | 6 | 1 | 1 | 1 | 1 | 1 |

Passo 6: Procura-se, em $D_6$, a menor distância euclidiana. Este valor corresponde a $d_{(3,10)}$ = 17,9.

Cálculo da matriz de dissimilaridade reduzida:

$$d^*_{j(3,10)} = \left[\frac{N_3}{N_3 + N_{10}}\right] d_{j3} + \left[\frac{N_{10}}{N_3 + N_{10}}\right] d_{j10}$$

$$d^*_{(1,2,4,7,8,11)(3,10)} = \left[\frac{1}{1+1}\right] x28{,}0 + \left[\frac{1}{1+1}\right] x21{,}2 = 24{,}8$$

$$d^*_{5(3,10)} = \left[\frac{1}{1+1}\right] x41{,}8 + \left[\frac{1}{1+1}\right] x34{,}8 = 38{,}3$$

$$d^*_{6(3,10)} = \left[\frac{1}{1+1}\right] x27{,}0 + \left[\frac{1}{1+1}\right] x21{,}5 = 24{,}3$$

$$d^*_{9(3,10)} = \left[\frac{1}{1+1}\right] x22{,}4 + \left[\frac{1}{1+1}\right] x22{,}8 = 22{,}6$$

$\mathbf{D_7}$ =

| | (1,2,4,7,8,11) | (3,10) | 5 | 6 | 9 |
|---|---|---|---|---|---|
| (1,2,4,7,8,11) | 0,00 | 24,8 | 34,9 | **19,9** | 25,2 |
| (3,10) | | 0,00 | 38,3 | 24,3 | 22,6 |
| 5 | | | 0,00 | 28,7 | 41,4 |
| 6 | | | | 0,00 | 27,6 |
| 9 | | | | | 0,00 |

N = [ 6 2 1 1 1 ]

Passo 7: Procura-se, em $D_7$, a menor distância euclidiana. Este valor corresponde a $d_{(1,2,4,6,7,8,11)} = 19{,}9$.

Cálculo da matriz de dissimilaridade reduzida:

$$d^*_{j(1,2,4,6,7,8,11)} = \left[\frac{N_{(1,4,7,8,11)}}{N_{(1,4,7,8,11)} + N_6}\right] d_{j6} + \left[\frac{N_6}{N_{(1,4,7,8,11)} + N_6}\right] d_{j(1,4,7,8,11)}$$

$$d^*_{(3,10)(1,2,4,6,7,8,11)} = \left[\frac{6}{1+6}\right] x24{,}8 + \left[\frac{1}{1+6}\right] x24{,}3 = 24{,}7$$

$$d^*_{5(1,2,4,6,7,8,11)} = \left[\frac{6}{1+6}\right] x34{,}9 + \left[\frac{1}{1+6}\right] x28{,}7 = 34{,}0$$

$$d^*_{9(1,2,4,6,7,8,11)} = \left[\frac{6}{1+6}\right] x25{,}2 + \left[\frac{1}{1+6}\right] x27{,}6 = 25{,}5$$

$\mathbf{D_8}$ =

| | (1,2,4,6,7,8,11) | (3,10) | 5 | 9 |
|---|---|---|---|---|
| (1,2,4,6,7,8,11) | 0,00 | 24,7 | 34,0 | 25,5 |
| (3,10) | | 0,00 | 38,3 | **22,6** |
| 5 | | | 0,00 | 41,4 |
| 9 | | | | 0,00 |

N = [ 7 2 1 1 ]

Passo 8: Procura-se, em $D_8$, a menor distância euclidiana. Este valor corresponde a $d_{(3,9,10)}$ = 22,6.

Cálculo da matriz de dissimilaridade reduzida:

$$d^*_{j(3,9,10)} = \left[\frac{N_{(3,10)}}{N_{(3,10)} + N_9}\right] d_{j(3,10)} + \left[\frac{N_9}{N_{(3,10)} + N_9}\right] d_{j9}$$

$$d^*_{(1,2,4,6,7,8,11)(3,9,10)} = \left[\frac{2}{2+1}\right] x24{,}7 + \left[\frac{1}{2+1}\right] x25{,}5 = 25{,}0$$

$$d^*_{5(3,9,10)} = \left[\frac{2}{2+1}\right] x38{,}3 + \left[\frac{1}{2+1}\right] x41{,}4 = 39{,}3$$

| **$D_9$** = | | (1,2,4,6,7,8,11) | (3,9,10) | 5 |
|---|---|---|---|---|
| | (1,2,4,6,7,8,11) | 0,00 | **25,0** | 34,0 |
| | (3,9,10) | | 0,00 | 39,3 |
| | 5 | | | 0,00 |
| N | = | 7 | 3 | 1 |

Passo 9: Procura-se, em $D_8$, a menor distância euclidiana. Este valor corresponde a $d_{(1,2,3,4,6,7,8,9,10,11)}$ = 25,0.

Cálculo da matriz de dissimilaridade reduzida:

$$d^*_{j(1,2,3,4,6,7,8,9,10,11)} = \left[\frac{N_{(1,2,4,6,7,8,11)}}{N_{(1,2,4,6,7,8,11)}+N_{(3,9,10)}}\right] d_{j(1,2,4,6,7,8,11)} + \left[\frac{N_{(3,9,10)}}{N_{(1,2,4,6,7,8,11)}+N_{(3,9,10)}}\right] d_{j(3,9,10)}$$

$$d^*_{5(1,2,3,4,6,7,8,9,10,11)} = \left[\frac{7}{7+3}\right] x34{,}0 + \left[\frac{3}{7+3}\right] x39{,}3 = 35{,}6$$

| **$D_{10}$** = | | (1,2,3,4,6,7,8,9,10,11) | 5 |
|---|---|---|---|
| | (1,2,3,4,6,7,8,9,10,11) | 0,00 | **35,6** |
| | 5 | | 0,00 |
| N | = | 10 | 1 |

Passo 10: A fusão final ocorre para $d_{(1,2,3,4,5,6,7,8,9,10,11)}$ = 35,6

Com a fusão de todos os elementos em um único agrupamento, os resultados foram resumidos na Tabela 26 e na Figura 12 (saída do R).

**Tabela 26.** Resumo da análise de agrupamento obtido pelo emprego do método das médias das distâncias, com base na distância euclidiana, mostrando a sequência das fusões das parcelas

| Distância | Objeto | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 7,7 | P4 | P8 | | | | | | | | | |
| 8,8 | P7 | P11 | | | | | | | | | |
| 11,9 | P1 | P4 | P8 | | | | | | | | |
| 15,1 | P1 | P4 | P8 | P7 | P11 | | | | | | |
| 16,6 | P1 | P4 | P8 | P7 | P11 | P2 | | | | | |
| 17,9 | P3 | P10 | | | | | | | | | |
| 19,9 | P1 | P4 | P8 | P7 | P11 | P2 | P6 | | | | |
| 22,6 | P3 | P10 | P9 | | | | | | | | |
| 25,0 | P1 | P4 | P8 | P7 | P11 | P2 | P6 | P3 | P10 | P9 | |
| 35,6 | P1 | P4 | P8 | P7 | P11 | P2 | P6 | P3 | P10 | P9 | P5 |

## Passo a passo no R: Agrupamento por Ligação Média (UPGMA)

### 1 Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust*(). O método "*average*" (média aritmética não ponderada ou UPGMA) agrupa as parcelas com base na distância média entre todos os pares de pontos em grupos diferentes. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

É importante converter a matriz de distância para o formato *dist* que a função *hclust*() espera, utilizando *as.dist*().

```
# Execução do Agrupamento Hierárquico pelo método UPGMA (Ligação Média)
# as.dist() converte a matriz para o formato de objeto de distância
agrupamento_upgma <- hclust(as.dist(matriz_euclidiana), method = "average")
```

### 2. Detalhamento das Fusões (Passo a Passo)

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância
resumo_upgma <- data.frame(
  Passo = 1:nrow(agrupamento_upgma$merge),
  Distancia_Fusao = round(agrupamento_upgma$height, 1)
)

# Exibindo o resumo das fusões
print(resumo_upgma)
```

| Passo | Fusão | Distancia_Fusão |
|---|---|---|
| 1 | 1 | 7.7 |
| 2 | 2 | 8.8 |
| 3 | 3 | 11.9 |
| 4 | 4 | 15.1 |
| 5 | 5 | 16.6 |
| 6 | 6 | 17.9 |
| 7 | 7 | 19.9 |
| 8 | 8 | 22.6 |
| 9 | 9 | 25.0 |
| 10 | 10 | 35.6 |

### 3. Visualização do Dendrograma

Na Figura 12, observa-se visualmente como as parcelas foram se unindo e formando grupos em diferentes níveis de distância. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes, além de personalizar os eixos para melhor visualização.

```
# Configurando a margem para caber o título inferior e os rótulos
par(mar=c(6, 4, 4, 2))

# Gerando o gráfico do Dendrograma (Figura 12)
plot(agrupamento_upgma,
     main = "Dendrograma - Método UPGMA (Ligação Média)",
     xlab = "Sequência de Agrupamento de Parcelas",
     ylab = "Distância Euclidiana",
     sub = "Baseado na Distância Euclidiana",
     hang = -1, # hang = -1 alinha os rótulos na parte inferior
     axes = FALSE) # Remove os eixos padrão para customizar

# Adicionando o eixo Y lateral
axis(2, at = seq(0, 40, by = 10))

# Adicionando a linha de corte (Linha Fenon) em um nível de distância específico (ex: 23.0)
abline(h = 23, col = "orange", lty = 2, lwd = 2)

# Destacando os grupos principais com retângulos coloridos (ex: 3 grupos)
rect.hclust(agrupamento_upgma, k = 3, border = c("blue", "green", "purple"))
```

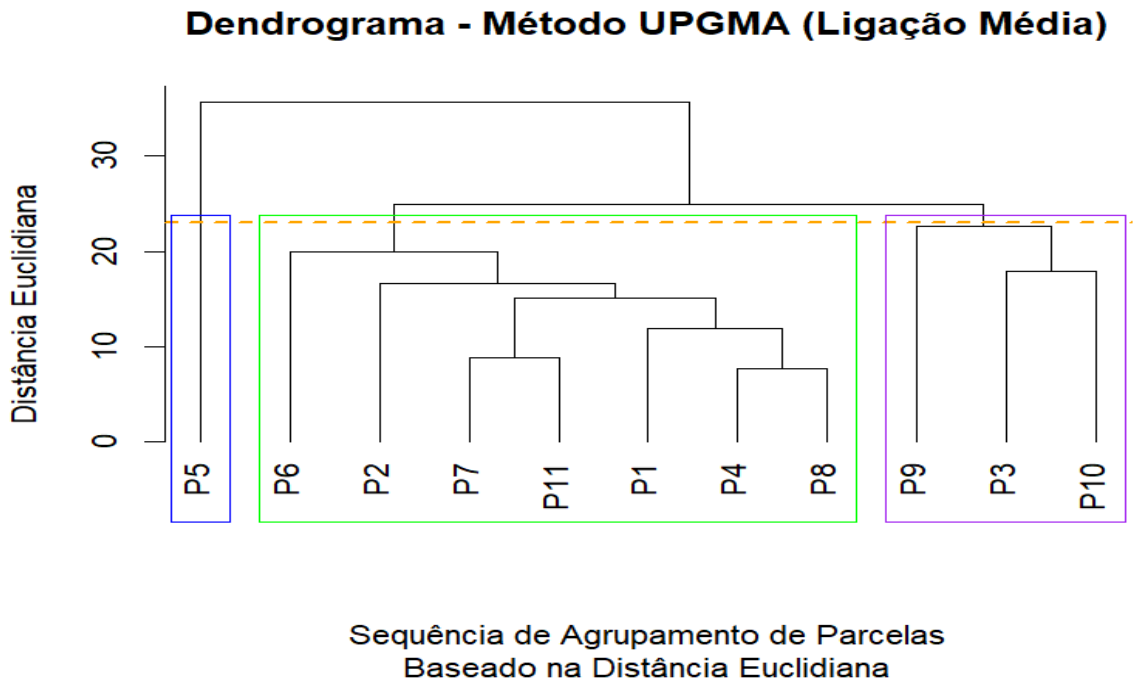



**Figura 12.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método da ligação média (UPGMA), com base na distância euclidiana, considerando a linha fenon d = 23,0 e a formação de três grupos

## f. Método de Ward

Um processo geral de classificação foi proposto por Ward (1963), em que **n** elementos são progressivamente reunidos dentro de grupos por meio da minimização de uma função-objetivo para cada (**n**-2) passos de fusão.

Inicialmente, esse algoritmo admite que cada um dos elementos constitui um único agrupamento. Considerando a primeira reunião de elementos em um novo agrupamento, a soma dos desvios dos pontos representativos de seus elementos, em relação à média do agrupamento, é calculada e dá uma indicação de homogeneidade do agrupamento formado. Essa medida fornece a "**perda de informação**" que se produz ao reunir os elementos de **E** em um agrupamento (Gama, 2022).

Quando os indivíduos são pontos de um espaço euclidiano ($\mathbf{I_p}$), a qualidade de uma partição é definida por sua inércia intraclasse ou interclasse. Quando se parte de **K+1** classes para **K** classes, ou seja, agrupando duas classes em uma só, a inércia interclasse só pode diminuir. Essa inércia é a média da soma dos quadrados das distâncias entre os centros de gravidade de cada classe e o centro de gravidade total (Bouroche; Saporta, 1982).

A reunião de elementos em grupos é feita pela análise dos valores da função de agrupamento, reunindo-se os elementos mais próximos, isto é, aqueles que apresentam **Min ($\mathbf{d_{ij}}$)** (Gama, 2022).

O algoritmo de Ward pode ser resumido nas seguintes etapas:

1. Determina-se a matriz de distâncias. Localizam-se os dois agrupamentos para os quais $\mathbf{d_{ij}}$ é mínimo.

2. Reúnem-se esses agrupamentos, formando um novo agrupamento, e verifica-se se o número de agrupamentos (**g**) já foi alcançado, se não, segue-se à etapa 3, caso contrário, termina-se a análise.

3. Calcula-se o valor do aumento a ser obtido da soma de quadrados pela reunião de qualquer dos agrupamentos: $\mathbf{I} = (1/2).\mathbf{d_{pq}}$.

4. Determinam-se os dois agrupamentos que apresentam um menor incremento na matriz **D**, isto é, **Min ($\mathbf{I_{ij}}$)**, e volta-se à etapa 2.

Vale ressaltar que o **método de Ward** pode ser formulado de diferentes maneiras, levando a diferentes variantes, conhecidas como **Ward.D** e **Ward.D2**. A diferença entre as duas variantes está na medida de distância. No **Ward.D**, é considerada a dissimilaridade das entradas em termos das distâncias euclidianas, enquanto no **método de Ward.D2**, está em termos das distâncias euclidianas quadrada. Para mais esclarecimentos, sugere-se consultar em Murtagh e Legendre (2014).

### f.1. Método de Ward.D

No **método War.D**, usa-se a soma das distâncias quadradas como critério para minimizar ao realizar agrupamento de objetos, o que equivale a minimizar o aumento na soma dos desvios quadrados da média do grupo de objetos considerados no agrupamento.

Esse método tem como função de agrupamentos a distância euclidiana, e o critério de agrupamento é dado pelo valor do incremento, que se obtém da soma de quadrados do erro.

**Observação**:

$d_{pk}^2 = \left(X_p - X_k\right)^2$ é a distância entre as médias dos elementos de $\mathbf{G_p}$ e $\mathbf{G_k}$, sendo $\mathbf{G_p}$ e $\mathbf{G_k}$, respectivamente, os grupos **p** e **k**.

$I_{pk} = \left(\frac{N_p N_k}{N_p + N_k}\right) d_{pk}^2$, em que a reunião dos agrupamentos $\mathbf{G_p}$ e $\mathbf{G_k}$ será feita se $\mathbf{I_{pk}}$ = mínimo.

Admitindo-se que, para o agrupamento $\mathbf{G_p} \cup \mathbf{G_k} = \mathbf{G_r}$, o incremento na soma das médias do erro é dado por:

$I_{pk} = \left(\frac{N_p N_k}{N_p + N_k}\right) d_{tr}^2$, em que: $d_{tr}^2 = (\bar{X}_t - \bar{X}_r)^2$, podendo ser escrito por:

$$d_{tr} = \left(\frac{n_p}{n_r}\right) d_{tp}^2 + \left(\frac{n_k}{n_r}\right) d_{kp}^2 - \left(\frac{n_p n_k}{n_k^2}\right) d_{pk}^2$$

Substituindo cada distância em função do número de elementos no agrupamento, tem-se:

$$I_{tr} = \left(\frac{1}{n_t + n_r}\right)\left[(n_t + n_p)I_{tp} + (n_t + n_k)I_{tk} - n_r I_{pk}\right]$$

Ou ainda, considerando $d_{tp}^2 = 2I_{tp}$, quando os grupos têm somente um elemento, teremos:

$$d_{tr} = \left(\frac{1}{n_t + n_r}\right)\left[(n_t + n_p)d_{tp}^2 + (n_t + n_k)d_{tk}^2 - n_t d_{pk}^2\right]$$

**Aplicação**

Considerando a matriz de distância euclidiana ($D_1$):

**$D_1$ =**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 12,5 | 33,9 | 17,3 | 18,5 | 11,3 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 12,0 | 34,4 | 21,8 | 18,3 | 16,2 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 29,6 | 41,8 | 27,0 | 30,9 | 27,9 | 22,4 | 17,9 | 24,2 |
| 4 | | | | 0,00 | 32,3 | 17,3 | 14,7 | **7,7** | 28,3 | 21,5 | 13,5 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 35,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 19,0 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 15,5 | 22,4 | 24,4 | 8,8 |
| 8 | | | | | | | | 0,00 | 26,7 | 21,2 | 13,3 |
| 9 | | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | | 0,00 |

A partir da matriz $D_1$, os seguintes passos são considerados:

Passo 1: Localizam-se os dois grupamentos que apresentam a menor distância. Neste caso, $d_{(4,8)} = 7,7$.

Cálculo do incremento da soma de quadrados: $I_{(4,8)} = (1/2).7,7 = 3,85$.

Cálculo das distâncias euclidianas entre o grupo (4,8) e os demais objetos:

$$d_{1(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x12,5 + (1+1)x11,3 - 1x7,7] = 13,3$$

$$d_{2(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x12,0 + (1+1)x16,2 - 1x7,7] = 16,2$$

$$d_{3(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x29,6 + (1+1)x27,9 - 1x7,7] = 35,8$$

$$d_{5(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x32,3 + (1+1)x35,6 - 1x7,7] = 42,7$$

$$d_{6(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x17{,}3 + (1+1)x19{,}0 - 1x7{,}7] = 21{,}6$$

$$d_{7(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x14{,}7 + (1+1)x15{,}5 - 1x7{,}7] = 17{,}6$$

$$d_{9(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x28{,}3 + (1+1)x26{,}7 - 1x7{,}7] = 34{,}1$$

$$d_{10(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x21{,}5 + (1+1)x21{,}2 - 1x7{,}7] = 25{,}9$$

$$d_{11(4,8)} = \left(\frac{1}{1+2}\right) x[(1+1)x13{,}5 + (1+1)x13{,}3 - 1x7{,}7] = 15{,}3$$

Nova Matriz $D_2$:

| $\mathbf{D_2}$ = | 1 | 2 | 3 | (4,8) | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,00 | 17,8 | 23,0 | 13,3 | 33,9 | 17,3 | 18,5 | 24,8 | 20,0 | 14,7 |
| 2 | | 0,00 | 31,7 | 16,2 | 34,4 | 21,8 | 18,3 | 29,4 | 25,6 | 18,7 |
| 3 | | | 0,00 | 35,8 | 41,8 | 27,0 | 30,9 | 22,4 | 17,9 | 24,2 |
| (4,8) | | | | 0,00 | 42,7 | 21,6 | 17,6 | 34,1 | 25,9 | 15,3 |
| 5 | | | | | 0,00 | 28,7 | 38,6 | 41,4 | 34,8 | 34,7 |
| 6 | | | | | | 0,00 | 24,6 | 27,6 | 21,5 | 19,4 |
| 7 | | | | | | | 0,00 | 22,4 | 24,4 | **8,8** |
| 9 | | | | | | | | 0,00 | 22,8 | 19,2 |
| 10 | | | | | | | | | 0,00 | 17,5 |
| 11 | | | | | | | | | | 0,00 |

| N = | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|

Passo 2: Os elementos 7 e 11 apresentam a menor distância em $D_2$ ($d_{(7,11)} = 8{,}8$).

Cálculo do incremento da soma de quadrados: $I_{(4,8)} = (1/2).8{,}8 = 4{,}4$.

Cálculo das distâncias euclidianas entre o grupo (7,11) e os demais objetos:

$$d_{1(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x18{,}5 + (1+1)x14{,}7 - 1x8{,}8] = 19{,}2$$

$$d_{2(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x18{,}3 + (1+1)x18{,}7 - 1x8{,}8] = 21{,}7$$

$$d_{3(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x30{,}9 + (1+1)x24{,}2 - 1x8{,}8] = 33{,}8$$

$$d_{(4,8)(7,11)} = \left(\frac{1}{2+2}\right) x[(2+1)x17{,}6 + (2+1)x15{,}3 - 2x8{,}8] = 20{,}3$$

$$d_{5(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x38{,}6 + (1+1)x34{,}7 - 1x8{,}8] = 45{,}9$$

$$d_{6(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x24{,}6 + (1+1)x19{,}4 - 1x8{,}8] = 26{,}4$$

$$d_{9(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x22{,}4 + (1+1)x19{,}2 - 1x8{,}8] = 24{,}8$$

$$d_{10(7,11)} = \left(\frac{1}{1+2}\right) x[(1+1)x24{,}4 + (1+1)x17{,}5 - 1x8{,}8] = 25{,}0$$

Nova Matriz $D_3$:

| | | 1 | 2 | 3 | (4,8) | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0,00 | 17,8 | 23,0 | **13,3** | 33,9 | 17,3 | 19,2 | 24,8 | 20,0 |
| | 2 | | 0,00 | 31,7 | 16,2 | 34,4 | 21,8 | 21,7 | 29,4 | 25,6 |
| | 3 | | | 0,00 | 35,8 | 41,8 | 27,0 | 33,8 | 22,4 | 17,9 |
| | (4,8) | | | | 0,00 | 42,7 | 21,6 | 20,3 | 34,1 | 25,9 |
| | 5 | | | | | 0,00 | 28,7 | 45,9 | 41,4 | 34,8 |
| **$D_3$ =** | 6 | | | | | | 0,00 | 26,4 | 27,6 | 21,5 |
| | (7,11) | | | | | | | 0,00 | 24,8 | 25,0 |
| | 9 | | | | | | | | 0,00 | 22,8 |
| | 10 | | | | | | | | | 0,00 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| N | = | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |

Passo 3: A menor distância em $D_3$ é $d_{(1,4,8)}$ = 13,3, assim, agrupam-se os elementos 1 e (4,8).

Cálculo do incremento da soma de quadrados: $I_{(1,4,8)} = ((1x2)/(1+2))x13{,}3 = 6{,}65$.

Cálculo das distâncias euclidianas entre o grupo (1,4,8) e os demais objetos:

$$d_{2(1,4,8)} = \left(\frac{1}{1+3}\right) x[(1+1)x17{,}8 + (1+2)x16{,}2 - 1x13{,}3] = 17{,}7$$

$$d_{2(1,4,8)} = \left(\frac{1}{1+3}\right) x[(1+1)x23{,}0 + (1+2)x35{,}8 - 1x13{,}3] = 35{,}0$$

$$d_{5(1,4,8)} = \left(\frac{1}{1+3}\right) x[(1+1)x33{,}9 + (1+2)x42{,}7 - 1x13{,}3] = 45{,}7$$

$$d_{6(1,4,8)} = \left(\frac{1}{1+3}\right) x[(1+1)x17{,}3 + (1+2)x21{,}6 - 1x13{,}3] = 21{,}5$$

$$d_{(7,11)(1,4,8)} = \left(\frac{1}{2+3}\right) x[(2+1)x19{,}2 + (2+2)x20{,}3 - 2x13{,}3] = 22{,}4$$

$$d_{9(1,4,8)} = \left(\frac{1}{1+3}\right) x[(1+1)x24{,}8 + (1+2)x34{,}1 - 1x13{,}3] = 34{,}7$$

$$d_{10(1,4,8)} = \left(\frac{1}{1+3}\right) x[(1+1)x20{,}0 + (1+2)x25{,}9 - 1x13{,}3] = 26{,}1$$

Nova Matriz $D_4$:

| $\mathbf{D_4}$ = | (1,4,8) | 2 | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| (1,4,8) | 0,00 | **17,7** | 35,0 | 45,7 | 21,5 | 22,4 | 34,7 | 26,1 |
| 2 | | 0,00 | 31,7 | 34,4 | 21,8 | 21,7 | 29,4 | 25,6 |
| 3 | | | 0,00 | 41,8 | 27,0 | 33,8 | 22,4 | 17,9 |
| 5 | | | | 0,00 | 28,7 | 45,9 | 41,4 | 34,8 |
| 6 | | | | | 0,00 | 26,4 | 27,6 | 21,5 |
| (7,11) | | | | | | 0,00 | 24,8 | 25,0 |
| 9 | | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | | 0,00 |
| N = | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |

Passo 4: Na matriz $D_4$, os elementos (1,4,8) e 2 apresentam a menor distância, ou seja, $d_{(1,2,4,8)}$ = 17,7. Logo, o incremento da soma de quadrados será I = (1/2)$x$17,7 = 8,85.

Cálculo das distâncias euclidianas entre o grupo (1,2,4,8) e os demais objetos:

$$d_{3(1,2,4,8)} = \left(\frac{1}{1+4}\right) x[(1+3)x35{,}0 + (1+1)x31{,}7 - 1x17{,}7] = 37{,}1$$

$$d_{5(1,2,4,8)} = \left(\frac{1}{1+4}\right) x[(1+3)x45{,}7 + (1+1)x34{,}4 - 1x17{,}7] = 46{,}8$$

$$d_{6(1,2,4,8)} = \left(\frac{1}{1+4}\right) x[(1+3)x21{,}5 + (1+1)x21{,}8 - 1x17{,}7] = 22{,}4$$

$$d_{(7,11)(1,2,4,8)} = \left(\frac{1}{2+4}\right) x[(2+3)x22{,}4 + (2+1)x21{,}7 - 2x17{,}7] = 23{,}6$$

$$d_{9(1,2,4,8)} = \left(\frac{1}{1+4}\right) x[(1+3)x34{,}7 + (1+1)x29{,}4 - 1x17{,}7] = 36{,}0$$

$$d_{10(1,2,4,8)} = \left(\frac{1}{1+4}\right) x[(1+3)x26{,}1 + (1+1)x25{,}6 - 1x17{,}7] = 27{,}6$$

Nova Matriz $D_5$:

| $\mathbf{D_5}$ = | (1,2,4,8) | 3 | 5 | 6 | (7,11) | 9 | 10 |
|---|---|---|---|---|---|---|---|
| (1,2,4,8) | 0,00 | 37,1 | 46,8 | 22,4 | 23,6 | 36,0 | 27,6 |
| 3 | | 0,00 | 41,8 | 27,0 | 33,8 | 22,4 | **17,9** |
| 5 | | | 0,00 | 28,7 | 45,9 | 41,4 | 34,8 |
| 6 | | | | 0,00 | 26,4 | 27,6 | 21,5 |
| (7,11) | | | | | 0,00 | 24,8 | 25,0 |
| 9 | | | | | | 0,00 | 22,8 |
| 10 | | | | | | | 0,00 |
| N = | 3 | 1 | 2 | 1 | 1 | 2 | 1 |

Passo 5: Os dois agrupamentos que apresentam uma menor distância na matriz $D_5$ são os elementos 3 e 10 ($d_{(3,10)}$ = 17,9). Neste caso, I = (1/2)*x*17,9 = 8,95.

Cálculo das distâncias euclidianas entre o grupo (3,10) e os demais objetos:

$$d_{(1,2,4,8)(3,10)} = \left(\frac{1}{4+2}\right) x[(4+1)x37{,}1 + (4+1)x27{,}6 - 4x17{,}9] = 42{,}0$$

$$d_{5(3,10)} = \left(\frac{1}{1+2}\right) x[(1+1)x41{,}8 + (1+1)x34{,}8 - 1x17{,}9] = 45{,}1$$

$$d_{6(3,10)} = \left(\frac{1}{1+2}\right) x[(1+1)x27{,}0 + (1+1)x21{,}5 - 1x17{,}9] = 26{,}4$$

$$d_{(7,11)(3,10)} = \left(\frac{1}{2+2}\right) x[(2+1)x33{,}8 + (2+1)x25{,}0 - 2x17{,}9] = 35{,}2$$

$$d_{9(3,10)} = \left(\frac{1}{1+2}\right) x[(1+1)x22{,}4 + (1+1)x22{,}8 - 1x17{,}9] = 24{,}2$$

Nova Matriz $D_6$:

| | | (1,2,4,8) | (3,10) | 5 | 6 | (7,11) | 9 |
|---|---|---|---|---|---|---|---|
| | (1,2,4,8) | 0,00 | 42,0 | 46,8 | **22,4** | 23,6 | 36,0 |
| | (3,10) | | 0,00 | 45,1 | 26,4 | 35,2 | 24,2 |
| | 5 | | | 0,00 | 28,7 | 45,9 | 41,4 |
| **D6 =** | 6 | | | | 0,00 | 26,4 | 27,6 |
| | (7,11) | | | | | 0,00 | 24,8 |
| | 9 | | | | | | 0,00 |
| N | = | 4 | 2 | 1 | 1 | 2 | 1 |

Passo 6: Na matriz $D_6$, a menor distância entre objetos é $d_{(1,2,4,6,8)}$ = 22,4. Assim, agrupam-se os elementos 6 e (1,2,4,8). Logo, I = (1/2)*x*22,4 = 11,2.

Cálculo das distâncias euclidianas entre o grupo (1,2,4,6,8) e os demais objetos:

$$d_{(3,10)(1,2,4,6,8)} = \left(\frac{1}{2+5}\right) x[(2+4)x42{,}0 + (2+1)x26{,}4 - 2x22{,}4] = 40{,}9$$

$$d_{5(1,2,4,6,8)} = \left(\frac{1}{1+5}\right) x[(1+4)x46{,}8 + (1+1)x28{,}7 - 1x22{,}4] = 44{,}8$$

$$d_{(7,11)(1,2,4,6,8)} = \left(\frac{1}{2+5}\right) x[(2+4)x23{,}6 + (2+1)x26{,}4 - 2x22{,}4] = 25{,}1$$

$$d_{9(1,2,4,6,8)} = \left(\frac{1}{1+5}\right) x[(1+4)x36{,}0 + (1+1)x27{,}6 - 1x22{,}4] = 35{,}5$$

Nova Matriz $D_7$:

| | | (1,2,4,6,8) | (3,10) | 5 | (7,11) | 9 |
|---|---|---|---|---|---|---|
| | (1,2,4,6,8) | 0,00 | 40,9 | 44,8 | 25,1 | 35,5 |
| | (3,10) | | 0,00 | 45,1 | 35,2 | **24,2** |
| | 5 | | | 0,00 | 45,9 | 41,4 |
| **$D_7$ =** | (7,11) | | | | 0,00 | 24,8 |
| | 9 | | | | | 0,00 |
| N | = | 5 | 2 | 1 | 2 | 1 |

Passo 7: A menor distância entre os objetos em $D_7$ é $d_{(3,9,10)}$ = 24,2.

Logo, I = (1/2)*x*24,2 = 12,1.

Cálculo das distâncias euclidianas entre o grupo (3,9,10) e os demais objetos:

$$d_{(1,2,4,6,8)(3,9,10)} = \left(\frac{1}{5+3}\right) x[(5+2)x40{,}9 + (5+1)x35{,}5 - 4x24{,}2] = 47{,}3$$

$$d_{5(3,9,10)} = \left(\frac{1}{1+3}\right) x[(1+2)x45{,}1 + (1+1)x41{,}4 - 1x24{,}2] = 48{,}5$$

$$d_{(7,11)(3,9,10)} = \left(\frac{1}{2+3}\right) x[(2+2)x35{,}2 + (2+1)x24{,}8 - 2x24{,}2] = 33{,}4$$

Nova Matriz $D_8$:

| | | (1,2,4,8) | (3,9,10) | 5 | (7,11) |
|---|---|---|---|---|---|
| | (1,2,6,4,8) | 0,00 | 47,3 | 44,8 | **25,1** |
| | (3,9,10) | | 0,00 | 48,5 | 33,4 |
| **$D_8$ =** | 5 | | | 0,00 | 45,9 |
| | (7,11) | | | | 0 |
| N | = | 5 | 3 | 1 | 2 |

Passo 8: Na matriz $D_8$, os dois agrupamentos que apresentam menor distância são (1,2,4,6,8) e (7,11). Distância entre os objetos: $d_{(1,2,4,6,7,8,11)}$ = 25,1.

Neste caso, I = (1/2)*x*25,1 = 12,55.

Cálculo das distâncias euclidianas entre o grupo (1,2,4,6,7,8,11) e os demais objetos:

$$d_{(3,9,10)(1,2,4,6,7,8,11)} = \left(\frac{1}{3+7}\right) x[(3+5)x47{,}3 + (3+2)x33{,}4 - 3x25{,}1] = 47{,}0$$

$$d_{5(1,2,4,6,7,8,11)} = \left(\frac{1}{1+7}\right) x[(1+5)x44{,}8 + (1+2)x45{,}9 - 1x25{,}1] = 47{,}7$$

Nova Matriz $D_9$:

$$\mathbf{D_9} = \begin{array}{c} \\ (1,2,4,6,7,8,11) \\ (3,9,10) \\ 5 \end{array} \begin{bmatrix} (1,2,4,6,7,8,11) & (3,9,10) & 5 \\ 0{,}00 & \mathbf{47{,}0} & 47{,}7 \\ & 0{,}00 & 48{,}5 \\ & & 0{,}00 \end{bmatrix}$$

$$N = \begin{bmatrix} 7 & 3 & 1 \end{bmatrix}$$

Passo 9: Em $D_9$, a menor distância entre os objetos é $d_{(1,2,3,4,6,7,8,9,10)} = 47$.

I = (1/2)*x*47 = 23,5.

Cálculo das distâncias euclidianas entre o grupo (1, 2, 3, 4, 6, 7, 8, 9, 10, 11) e 5:

$$d_{5(1,2,4,6,7,8,9,10,11)} = \left(\frac{1}{1+10}\right) x[(1+7)x47{,}7 + (1+3)x48{,}5 - 1x47{,}0] = 48{,}0$$

Assim, a fusão de todos os elementos é obtida para a distância igual a 48,0.

Os resultados estão resumidos na Tabela 27 e ilustrados por um dendrograma (Figura 13, saída do R).

**Tabela 27**. Resumo da análise de agrupamento obtido pelo emprego do **método Ward.D**, mostrando a sequência das fusões das parcelas, com base na distância euclidiana

| Distância | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 7,7 | P4 | P8 | | | | | | | | | |
| 8,8 | P7 | P11 | | | | | | | | | |
| 13,3 | P1 | P4 | P8 | | | | | | | | |
| 17,7 | P1 | P4 | P8 | P2 | | | | | | | |
| 17,9 | P3 | P10 | | | | | | | | | |
| 22,4 | P1 | P4 | P8 | P2 | P6 | | | | | | |
| 24,2 | P3 | P10 | P9 | | | | | | | | |
| 25,2 | P1 | P4 | P8 | P2 | P6 | P7 | P11 | | | | |
| 47,0 | P1 | P4 | P8 | P2 | P6 | P7 | P11 | P3 | P10 | P9 | |
| 48,0 | P1 | P4 | P8 | P2 | P6 | P7 | P11 | P3 | P10 | P9 | P5 |

## Passo a passo no R: Agrupamento por Método de Ward.D

### 1. Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust*(). O método "*ward.D*" (ou "*Ward D*") busca minimizar a variância total dentro de cada grupo, tendendo a formar grupos mais compactos e esféricos. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

É importante converter a matriz de distância para o formato *dist* que a função *hclust*() espera, utilizando *as.dist*().

```
# Execução do Agrupamento Hierárquico pelo método de Ward D
# as.dist() converte a matriz para o formato de objeto de distância
agrupamento_ward_D <- hclust(as.dist(matriz_euclidiana), method = "ward.D")
```

**2. Detalhamento das Fusões (Passo a Passo)**

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância
resumo_ward_D <- data.frame(
  Passo = 1:nrow(agrupamento_ward_D$merge),
  Distancia_Fusao = round(agrupamento_ward_D$height, 1)
)

# Exibindo o resumo das fusões
print(resumo_ward_D)
```

| Passo | Fusão | Distancia_Fusao |
|---|---|---|
| 1 | 1 | 7.7 |
| 2 | 2 | 8.8 |
| 3 | 3 | 13.3 |
| 4 | 4 | 17.7 |
| 5 | 5 | 17.9 |
| 6 | 6 | 22.4 |
| 7 | 7 | 24.2 |
| 8 | 8 | 25.2 |
| 9 | 9 | 46.9 |
| 10 | 10 | 48.0 |

**3. Visualização do Dendrograma**

O dendrograma é a representação gráfica do agrupamento hierárquico (Figura 13). Nele, mostra-se visualmente como as parcelas foram se unindo e formando grupos em diferentes níveis de distância. O método de *Ward D* tende a produzir dendrogramas com grupos bem definidos. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes, além de personalizar os eixos para melhor visualização.

```
# Configurando a margem para caber o título e legendas
par(mar=c(6, 4, 4, 2))

# Gerando o gráfico do Dendrograma (Figura 13)
plot(agrupamento_ward_D,
     main = "Dendrograma - Método de Ward D",
     xlab = "Parcelas",
     ylab = "Distância Euclidiana",
     sub = "Implementação conforme Murtagh e Legendre (2014)",
     hang = -1, # hang = -1 alinha os rótulos na parte inferior
     axes = FALSE) # Remove os eixos padrão para customizar
```

```
# Adicionando o eixo Y lateral
axis(2, at = seq(0, 50, by = 10))

# Adicionando a linha de corte (Linha Fenon) em um nível de distância específico (ex: 25.0)
abline(h = 25, col = "red", lty = 2, lwd = 2)

# Destacando os grupos principais com retângulos coloridos (ex: 3 grupos)
rect.hclust(agrupamento_ward_D, k = 3, border = c("blue", "green", "purple"))
```

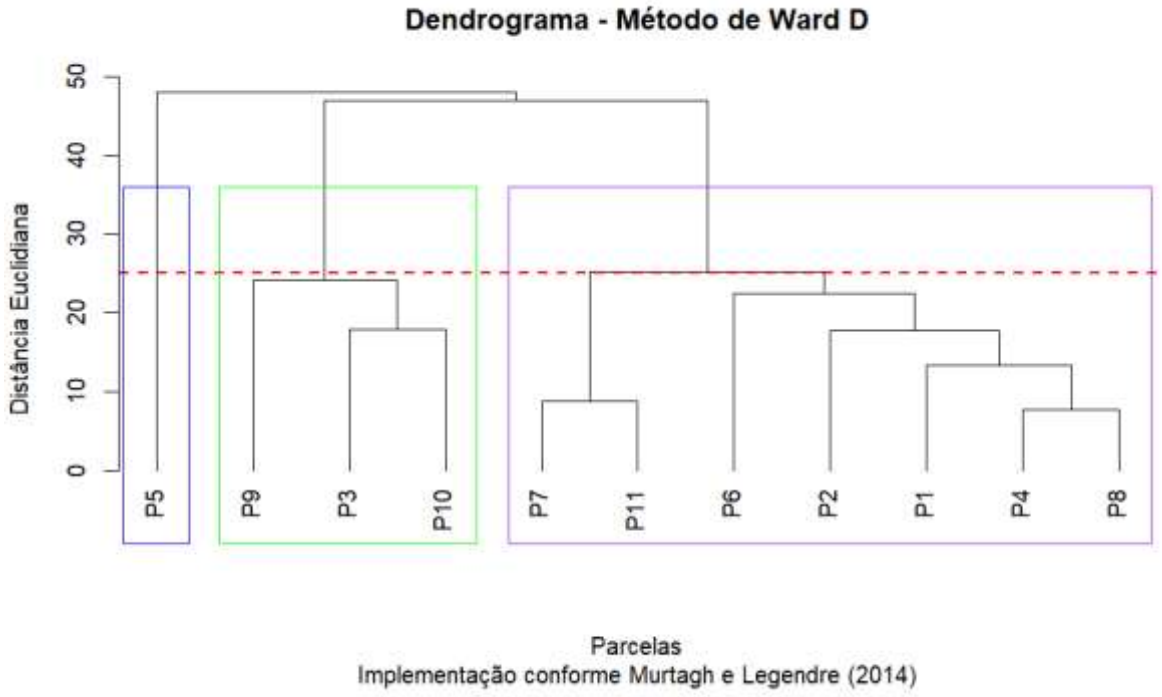



**Figura 13.** Dendrograma representando as sequências das fusões das parcelas obtido pelo emprego do método de Ward.D, com base na distância euclidiana, considerando a linha fenon d = 25,0 e a formação de três grupos.

### f.2. Método de Ward.D2

No método de Ward.D2, usa-se a soma dos quadrados das diferenças do centroide como critério para minimizar ao agrupar objetos, o que equivale a minimizar o aumento na soma dos desvios quadrados do centroide do grupo de objetos considerados no agrupamento.

Esse método tem como função de agrupamentos a **distância euclidiana quadrada**, e o critério de agrupamento é dado pelo valor do incremento.

### Aplicação

Considerando a matriz de distância euclidiana quadrada ($D_1$):

$\mathbf{D_1}$ =

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 317 | 529 | 156 | 1151 | 300 | 342 | 127 | 615 | 399 | 217 |
| 2 | | 0 | 1002 | 143 | 1180 | 475 | 335 | 262 | 866 | 656 | 348 |
| 3 | | | 0 | 875 | 1748 | 731 | 955 | 780 | 504 | 322 | 586 |
| 4 | | | | 0 | 1041 | 298 | 216 | **59** | 801 | 463 | 183 |
| 5 | | | | | 0 | 821 | 1493 | 1264 | 1716 | 1212 | 1202 |
| 6 | | | | | | 0 | 604 | 361 | 761 | 463 | 375 |
| 7 | | | | | | | 0 | 241 | 503 | 597 | 77 |
| 8 | | | | | | | | 0 | 714 | 448 | 176 |
| 9 | | | | | | | | | 0 | 520 | 370 |
| 10 | | | | | | | | | | 0 | 308 |
| 11 | | | | | | | | | | | 0 |

Passo 1:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(4,8)} = 59$.

b) Cálculo do incremento da soma de quadrados devido ao grupo formado ($I_{pk}$):

$$I_{pk} = \left(\frac{N_p x N_k}{N_p + N_k}\right) x d^2_{pk}$$

$$I_{48} = \left(\frac{N_4 x N_8}{N_4 + N_8}\right) x d^2_{48} = \left(\frac{1x1}{1+1}\right) x59 = 29{,}5$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (4,8) e os demais objetos:

$$d^2_{z(p,k)} = \frac{(N_z+N_p)xd^2_{zp}+(N_z+N_k)xd^2_{zk}-N_z xd^2_{pk}}{N_z+N_p+N_k}$$

$$d^2_{1(4,8)} = \frac{(N_1+N_4)xd^2_{24}+(N_1+N_8)xd^2_{28}-N_1 xd^2_{48}}{N_1+N_4+N_8} = \frac{(1+1)x156+(1+1)x127-1x59}{1+1+1} = 169{,}00$$

$$d_{2(4,8)} = \frac{(1+1)x143+(1+1)x262-1x59}{1+1+1} = 250{,}33$$

$$d_{3(4,8)} = \frac{(1+1)x875+(1+1)x780-1x59}{1+1+1} = 1083{,}67$$

$$d_{5(4,8)} = \frac{(1+1)x1041+(1+1)x1264-1x59}{1+1+1} = 1517{,}00$$

$$d_{6(4,8)} = \frac{(1+1)x298+(1+1)x361-1x59}{1+1+1} = 419{,}67$$

$$d_{7(4,8)} = \frac{(1+1)x216+(1+1)x241-1x59}{1+1+1} = 285{,}00$$

$$d_{9(4,8)} = \frac{(1+1)x801+(1+1)x714-1x59}{1+1+1} = 990{,}33$$

$$d_{10(4,8)} = \frac{(1+1)x463+(1+1)x448-1x59}{1+1+1} = 587{,}67$$

$$d_{11(4,8)} = \frac{(1+1)x183+(1+1)x176-1x59}{1+1+1} = 219{,}67$$

Nova Matriz $D_2$:

**$D_2$ =**

| | 1 | 2 | 3 | 4,8 | 5 | 6 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 317 | 529 | 169 | 1151 | 300 | 342 | 615 | 399 | 217 |
| 2 | | 0 | 1002 | 250,33 | 1180 | 475 | 335 | 866 | 656 | 348 |
| 3 | | | 0 | 1083,67 | 1748 | 731 | 955 | 504 | 322 | 586 |
| 4,8 | | | | 0 | 1517 | 419,67 | 285 | 990,33 | 587,67 | 219,67 |
| 5 | | | | | 0 | 821 | 1493 | 1716 | 1212 | 1202 |
| 6 | | | | | | 0 | 604 | 761 | 463 | 375 |
| 7 | | | | | | | 0 | 503 | 597 | 77 |
| 9 | | | | | | | | 0 | 520 | 370 |
| 10 | | | | | | | | | 0 | 308 |
| 11 | | | | | | | | | | 0 |
| N | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |

Passo 2:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(711)}$ = 77.

b) Cálculo do incremento da soma de quadrados:

$$I_{711} = \left(\frac{N_7 x N_{11}}{N_7 + N_{11}}\right) x d^2_{711} = \left(\frac{1x1}{1+1}\right) x77 = 38{,}5$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (7,11) e os demais objetos:

$$d^2_{1(7,11)} = \frac{(N_1 + N_7)x d^2_{17} + (N_1 + N_{11})x d^2_{111} - N_1 x d^2_{711}}{N_1 + N_7 + N_{11}}$$

$$d^2_{1(7,11)} = \frac{(1+1)x342 + (1+1)x217 - 1x77}{1+1+1} = 347{,}00$$

$$d^2_{2(7,11)} = \frac{(1+1)x335 + (1+1)x348 - 1x77}{1+1+1} = 429{,}67$$

$$d^2_{3(7,11)} = \frac{(1+1)x955 + (1+1)x586 - 1x77}{1+1+1} = 1001{,}67$$

$$d^2_{(4,8)(7,11)} = \frac{(2+1)x285 + (2+1)x219{,}67 - 2x77}{2+1+1} = 340{,}00$$

$$d^2_{5(7,11)} = \frac{(1+1)x1493 + (1+1)x1202 - 1x77}{1+1+1} = 1771{,}00$$

$$d^2_{6(7,11)} = \frac{(1+1)x604 + (1+1)x375 - 1x77}{1+1+1} = 627{,}00$$

$$d^2_{9(7,11)} = \frac{(1+1)x503 + (1+1)x370 - 1x77}{1+1+1} = 556{,}33$$

$$d^2_{10(7,11)} = \frac{(1+1)x597 + (1+1)x308 - 1x77}{1+1+1} = 577{,}67$$

Nova Matriz $D_3$:

**$D_3$ =**

| | 1 | 2 | 3 | 4,8 | 5 | 6 | 7,11 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 317 | 529 | **169** | 1151 | 300 | 347 | 615 | 399 |
| 2 | | 0 | 1002 | 250,33 | 1180 | 475 | 429,67 | 866 | 656 |
| 3 | | | 0 | 1083,67 | 1748 | 731 | 1001,67 | 504 | 322 |
| 4,8 | | | | 0 | 1517 | 419,67 | 340 | 990,33 | 587,67 |
| 5 | | | | | 0 | 821 | 1771 | 1716 | 1212 |
| 6 | | | | | | 0 | 627 | 761 | 463 |
| 7,11 | | | | | | | 0 | 556,33 | 577,67 |
| 9 | | | | | | | | 0 | 520 |
| 10 | | | | | | | | | 0 |
| N | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |

Passo 3:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(1,4,8)}$ = 169.

b) Cálculo do incremento da soma de quadrados:

$$I_{1,4,8} = \left(\frac{N_1 x N_{48}}{N_1 + N_{48}}\right) x d^2_{148} = \left(\frac{1x2}{1+2}\right) x169 = 112{,}67$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (1,4,8) e os demais objetos:

$$d^2_{2(1,4,8)} = \frac{(N_2 + N_1) x d^2_{12} + (N_2 + N_{48}) x d^2_{248} - N_2 x d^2_{148}}{N_2 + N_1 + N_{48}}$$

$$d^2_{2(1,4,8)} = \frac{(1+1)x317 + (1+2)x250{,}33 - 1x169}{1+1+2} = 304{,}00$$

$$d^2_{3(1,4,8)} = \frac{(1+1)x529 + (1+2)x1083{,}67 - 1x169}{1+1+2} = 1035{,}00$$

$$d^2_{5(1,4,8)} = \frac{(1+1)x1151 + (1+2)x1517 - 1x169}{1+1+2} = 1671{,}00$$

$$d^2_{6(1,4,8)} = \frac{(1+1)x300 + (1+2)x419{,}67 - 1x169}{1+1+2} = 422{,}50$$

$$d^2_{(7,11)(1,4,8)} = \frac{(1+2)x347 + (2+2)x340 - 2x169}{1+2+2} = 412{,}60$$

$$d^2_{9(1,4,8)} = \frac{(1+1)x615 + (1+2)x990{,}33 - 1x169}{1+1+2} = 1008{,}00$$

$$d^2_{10(1,4,8)} = \frac{(1+1)x399 + (1+2)x587{,}67 - 1x169}{1+1+2} = 598{,}00$$

Nova Matriz $D_4$:

**$D_4$ =**

| | 1,4,8 | 2 | 3 | 5 | 6 | 7,11 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 1,4,8 | 0 | **304** | 1035 | 1671 | 422,5 | 412,6 | 1008 | 598 |
| 2 | | 0 | 1002 | 1180 | 475 | 429,67 | 866 | 656 |
| 3 | | | 0 | 1748 | 731 | 1001,67 | 504 | 322 |
| 5 | | | | 0 | 821 | 1771 | 1716 | 1212 |
| 6 | | | | | 0 | 627 | 761 | 463 |
| 7,11 | | | | | | 0 | 556,33 | 577,67 |
| 9 | | | | | | | 0 | 520 |
| 10 | | | | | | | | 0 |
| N | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |

Passo 4:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(2,1,4,8)} = 304$.

b) Cálculo do incremento da soma de quadrados:

$$I_{2,1,4,8} = \left(\frac{N_2 x N_{148}}{N_2 + N_{148}}\right) x d^2_{1248} = \left(\frac{1x3}{1+3}\right) x304 = 228{,}00$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (2,1,4,8) e os demais objetos:

$$d^2_{3(1,2,4,8)} = \frac{(N_3 + N_2) x d^2_{23} + (N_3 + N_{148}) x d^2_{148} - N_3 x d^2_{1248}}{N_3 + N_2 + N_{148}}$$

$$d^2_{3(2,1,4,8)} = \frac{(1+1)x1002 + (1+3)x1035 - 1x304}{1+1+3} = 1168{,}00$$

$$d^2_{5(2,1,4,8)} = \frac{(1+1)x1180 + (1+3)x1671 - 1x304}{1+1+3} = 1748{,}00$$

$$d^2_{6(2,1,4,8)} = \frac{(1+1)x475 + (1+3)x422{,}5 - 1x304}{1+1+3} = 467{,}20$$

$$d^2_{(7,11)(2,1,4,8)} = \frac{(2+1)x429{,}67 + (2+3)x412{,}6 - 2x304}{2+1+3} = 457{,}33$$

$$d^2_{9(2,1,4,8)} = \frac{(1+1)x866 + (1+3)x1008 - 1x304}{1+1+3} = 1092{,}00$$

$$d^2_{10(2,1,4,8)} = \frac{(1+1)x656 + (1+3)x598 - 1x304}{1+1+3} = 680{,}00$$

Nova Matriz $D_5$:

**$D_5$ =**

| | 1,2,4,8 | 3 | 5 | 6 | 7,11 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| 1,2,4,8 | 0 | 1168 | 1748 | 467,2 | 457,33 | 1092 | 680 |
| 3 | | 0 | 1748 | 731 | 1001,67 | 504 | **322** |
| 5 | | | 0 | 821 | 1771 | 1716 | 1212 |
| 6 | | | | 0 | 627 | 761 | 463 |
| 7,11 | | | | | 0 | 556,33 | 577,67 |
| 9 | | | | | | 0 | 520 |
| 10 | | | | | | | 0 |
| N | 4 | 1 | 1 | 1 | 2 | 1 | 1 |

Passo 5:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(3,10)}$ = 322.

b) Cálculo do incremento da soma de quadrados:

$$I_{3,10} == \left(\frac{N_3 x N_{10}}{N_3 + N_{10}}\right) x d^2_{310} = \left(\frac{1x1}{1+1}\right) x322 = 161{,}00$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (3,10) e os demais objetos:

$$d^2_{(1,2,4,8)(3,10)} = \frac{(N_{1248} + N_3) x d^2_{12348} + (N_{1248} + N_{10}) x d^2_{124810} - N_{1248} x d^2_{310}}{N_3 + N_{1248} + N_{10}}$$

$$d^2_{(2,1,4,8)(3,10)} = \frac{(4+1)x1168 + (4+1)x680 - 4x322}{4+1+1} = 1325{,}33$$

$$d^2_{5(3,10)} = \frac{(1+1)x1748 + (1+1)x1212 - 1x322}{1+1+1} = 1866{,}00$$

$$d^2_{6(3,10)} = \frac{(1+1)x731 + (1+1)x463 - 1x322}{1+1+1} = 688{,}67$$

$$d^2_{(7,11)(3,10)} = \frac{(2+1)x1001{,}67 + (2+1)x577{,}67 - 2x322}{2+1+1} = 1023{,}50$$

$$d^2_{9(3,10)} = \frac{(1+1)x504 + (1+1)x520 - 1x322}{1+1+1} = 575{,}33$$

Nova Matriz $D_6$:

| | | 1,2,4,8 | 3,10 | 5 | 6 | 7,11 | 9 |
|---|---|---|---|---|---|---|---|
| | 1,2,4,8 | 0 | 1325,33 | 1748 | 467,2 | **457,33** | 1092 |
| | 3,10 | | 0 | 1866 | 688,67 | 1023,5 | 575,33 |
| **$D_6$** = | 5 | | | 0 | 821 | 1771 | 1716 |
| | 6 | | | | 0 | 627 | 761 |
| | 7,11 | | | | | 0 | 556,33 |
| | 9 | | | | | | 0 |
| | N | 4 | 2 | 1 | 1 | 2 | 1 |

Passo 6:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(1,2,4,8)(7,11)} = 457,33$.

b) Cálculo do incremento da soma de quadrados:

$$I_{(1,2,4,8)(7,11)}\left(\frac{N_{1248}xN_{711}}{N_{1248}+N_{711}}\right)xd^2_{1247811} = \left(\frac{4x2}{4+2}\right)x457,33 = 609,78$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (1,2,4,8,7,11) e os demais objetos:

$$d^2_{(1,2,4,7,8,11)(3,10)} = \frac{(N_{1248}+N_{310})xd^2_{1234810} + (N_{711}+N_{310})xd^2_{371110} - N_{310}xd^2_{1247811}}{N_{1248}+N_{711}+N_{310}}$$

$$d^2_{(2,1,4,8,7,11)(3,10)} = \frac{(4+2)x1325,33 + (2+2)x1023,50 - 2x457,33}{4+2+2} = 1391,42$$

$$d^2_{(2,1,4,8,7,11)5} = \frac{(4+1)x1748 + (2+1)x1866 - 1x457,33}{4+2+1} = 1942,24$$

$$d^2_{(2,1,4,8,7,11)6} = \frac{(4+1)x467,20 + (2+1)x627 - 1x457,33}{4+2+1} = 537,10$$

$$d^2_{(2,1,4,8,7,11)9} = \frac{(4+1)x1092 + (2+1)x556,33 - 1x457,33}{4+2+1} = 953,09$$

Nova Matriz $D_7$:

| | | 1,2,4,7,8,11 | 3,10 | 5 | 6 | 9 |
|---|---|---|---|---|---|---|
| | 1,2,4,7,8,11 | 0 | 1391,42 | 1942,24 | **537,10** | 953,09 |
| | 3,10 | | 0 | 1866 | 688,67 | 575,33 |
| **$D_7$** = | 5 | | | 0 | 821 | 1716 |
| | 6 | | | | 0 | 761 |
| | 9 | | | | | 0 |
| | N | 6 | 2 | 1 | 1 | 1 |

Passo 7:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(1,2,4,7,8,11)(6)} = 537,10$.

b) Cálculo do incremento da soma de quadrados:

$$I_{1,2,4,6,7,8,11} = \left(\frac{N_{1247811} x N_6}{N_{1247811} + N_6}\right) x d^2_{1247811} = \left(\frac{6x1}{6+1}\right) x537,10 = 460,37$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (1,2,4,8,7,11,6) e os demais objetos:

$$d^2_{(1,2,4,6,7,8,11)(3,10)} = \frac{(N_{1247811} + N_6) x d^2_{1247811} + (N_6 + N_{310}) x d^2_{3610} - N_6 x d^2_{12467811}}{N_{1247811} + N_{310} + N_6}$$

$$d^2_{(1,2,4,6,7,8,11)(3,10)} = \frac{(6+2)x1391,42 + (1+2)x688,67 - 2x537,10}{6+2+1} = 1347,02$$

$$d^2_{(1,2,4,6,7,8,11)5} = \frac{(6+1)x1942,24 + (1+1)x821 - 1x537,10}{6+1+1} = 1837,57$$

$$d^2_{(1,2,4,6,7,8,11)9} = \frac{(6+1)x953,09 + (1+1)x761 - 1x537,10}{6+1+1} = 957,07$$

Nova Matriz $D_8$:

| | | 1,2,4,6,7,8,11 | 3,10 | 5 | 9 |
|---|---|---|---|---|---|
| | 1,2,4,6,7,8,11 | 0 | 1347,02 | 1837,57 | 957,07 |
| | 3,10 | | 0 | 1866 | **575,33** |
| **D8** = | 5 | | | 0 | 1716 |
| | 9 | | | | 0 |
| | N | 7 | 2 | 1 | 1 |

Passo 8:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(3,10)9} = 575,33$.

b) Cálculo do incremento da soma de quadrados:

$$I_{3,9,10} = \left(\frac{N_{310} x N_9}{N_{310} + N_9}\right) x d^2_{310} = \left(\frac{2x1}{2+1}\right) x575,33 = 383,56$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (3,10,9) e os demais objetos:

$$d^2_{(1,2,4,6,7,8,11)3910} = \frac{(N_{12467811}+N_9) x d^2_{12647811} + (N_{310}+N_{12467811}) x d^2_{3910} - N_{12467811} x d^2_{3910}}{N_{12467811}+N_{310}+N_9}$$

$$d^2_{(1,2,4,6,7,8,11)3910} = \frac{(7+1)x1347{,}02 + (2+7)x957{,}07 - 7x575{,}33}{7+2+1} = 1575{,}24$$

$$d^2_{(3,10,9)5} = \frac{(2+1)x1866 + (1+1)x1716 - 1x575{,}33}{2+1+1} = 2113{,}67$$

Nova Matriz $D_9$:

| | | 1,2,4,6,7,8,11 | 3,9,10 | 5 |
|---|---|---|---|---|
| | 1,2,4,6,7,8,11 | 0 | **1575,24** | 1837,57 |
| | 3,9,10 | | 0 | 2113,67 |
| **D9** = | 5 | | | 0 |
| | N | 7 | 3 | 1 |

Passo 9:

a) Localizam-se os dois agrupamentos que apresentam a menor distância. Neste caso, $d^2_{(1,2,4,6,7,8,11)(3,9,10)} = 1575{,}24$.

b) Cálculo do incremento da soma de quadrados:

$$I_{1,2,3,4,6,7,8,9,10,11} = \left(\frac{N_{12467811} x N_{3910}}{N_{12467811} + N_{3910}}\right) x d^2_{310} = \left(\frac{7x3}{7+3}\right) x1701{,}04 = 3308{,}00$$

c) Cálculo das distâncias euclidianas quadradas entre o grupo (1,2,4,6,7,8,11,3,10,9) e o objeto 5.

$$d^2_{(1,2,4,6,7,8,11,3,10,9)5} = \frac{(7+1)x1837{,}57 + (3+1)x2113{,}67 - 1x1575{,}24}{7+3+1} = 1961{,}82$$

Nova Matriz $D_{10}$:

| | | 1,2,3,4,6,7,8,9,10,11 | 5 |
|---|---|---|---|
| | 1,2,3,4,6,7,8,9,10,11 | 0 | **1961,82** |
| **D10** = | 5 | | 0 |
| | N | 10 | 1 |

Os resultados estão resumidos na Tabela 28 e ilustrados por um dendrograma (Figura 14).

**Tabela 28**. Resumo da análise de agrupamento obtido pelo emprego do método Ward D2, mostrando a sequência das fusões das parcelas

| Distância Euclidiana | | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quadrada | Simples | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 59 | 7,68 | P4 | P8 | | | | | | | | | |
| 77 | 8,77 | P7 | P11 | | | | | | | | | |
| 169 | 13,00 | P1 | P4 | P8 | | | | | | | | |
| 304 | 17,44 | P1 | P4 | P8 | P2 | | | | | | | |
| 322 | 17,94 | P3 | P10 | | | | | | | | | |
| 457,33 | 21,39 | P1 | P4 | P8 | P2 | P7 | P11 | | | | | |
| 537,10 | 23,18 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | | | | |
| 575,33 | 23,99 | P3 | P10 | P9 | | | | | | | | |
| 1575,24 | 41,24 | P1 | P4 | P8 | P2 | P7 | P11 | P6 | P3 | P10 | P9 | |
| 1961,82 | 44,16 | P1 | P4 | P8 | P2 | P6 | P7 | P11 | P3 | P10 | P9 | P5 |

## Passo a passo no R: Agrupamento por Método de Ward.D2

### 1. Executando o Agrupamento Hierárquico

Para realizar o agrupamento hierárquico, utilizaremos a função *hclust()*. O método "*ward.D2*" é uma das implementações do método de Ward, que busca minimizar a variância total dentro de cada grupo. Ele é baseado na distância euclidiana ao quadrado, o que o torna robusto para diversas aplicações. Como base, utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60).

É importante converter a matriz de distância para o formato *dist* que a função *hclust()* espera, utilizando *as.dist()*.

```
# Execução do Agrupamento Hierárquico pelo método de Ward D2
# as.dist() converte a matriz para o formato de objeto de distância
agrupamento_ward_d2 <- hclust(as.dist(matriz_euclidiana), method = "ward.D2")
```

### 2. Detalhamento das Fusões (Passo a Passo)

O R registra cada fusão (união de parcelas ou grupos) que ocorre durante o processo de agrupamento. O critério de agrupamento no Ward D2 é o valor do incremento (I), que pode ser visualizado em duas escalas: a distância euclidiana ao quadrado (Dist_Quadrada) e a distância euclidiana simples (Dist_Simples). Podemos criar um resumo para visualizar os passos de união e os níveis de distância em que ocorreram, o que é útil para entender a formação dos grupos.

```
# Criando um resumo das fusões e dos níveis de distância em duas escalas
resumo_ward_d2 <- data.frame(
  Passo = 1:nrow(agrupamento_ward_d2$merge),
  Dist_Quadrada = round(agrupamento_ward_d2$height^2, 2), # Distância ao quadrado
  Dist_Simples = round(agrupamento_ward_d2$height, 2)    # Distância simples
)
```

```
# Exibindo o resumo das fusões
print(resumo_ward_d2)
```

| Passo | Distância Quadrada | Distância Simples |
|---|---|---|
| 1 | 59.00 | 7.68 |
| 2 | 77.00 | 8.77 |
| 3 | 169.00 | 13.00 |
| 4 | 304.00 | 17.44 |
| 5 | 322.00 | 17.94 |
| 6 | 457.33 | 21.39 |
| 7 | 537.10 | 23.18 |
| 8 | 575.33 | 23.99 |
| 9 | 1575.24 | 39.69 |
| 10 | 1961.82 | 44.29 |

### 3. Visualização do Dendrograma

Por meio do dendrograma (Figura 14), observa-se como as parcelas foram unidas e a formação dos grupos em diferentes níveis de distância. O método de Ward D2 costuma apresentar ramos mais longos, facilitando a identificação visual dos grupos. Podemos adicionar uma "Linha Fenon" para indicar um corte na distância e destacar os grupos resultantes, além de personalizar os eixos para melhor visualização.

```
# Configurando a margem para caber o título e legendas
par(mar=c(6, 4, 4, 2))

# Gerando o gráfico do Dendrograma (Figura 14)
plot(agrupamento_ward_d2,
     main = "Dendrograma - Método de Ward D2",
     xlab = "Sequência de Agrupamento de Parcelas",
     ylab = "Distância Euclidiana",
     sub = "Baseado na Distância Euclidiana",
     hang = -1, # hang = -1 alinha os rótulos na parte inferior
     axes = FALSE) # Remove os eixos padrão para customizar

# Adicionando o eixo Y lateral
axis(2, at = seq(0, 50, by = 10))

# Adicionando a linha de corte (Linha Fenon) em um nível de distância específico (ex: 24.2)
abline(h = 24.2, col = "orange", lwd = 2)

# Destacando os grupos principais com retângulos coloridos (ex: 4 grupos)
rect.hclust(agrupamento_ward_d2, k = 4, border = c("red", "green", "blue", "cyan"))
```

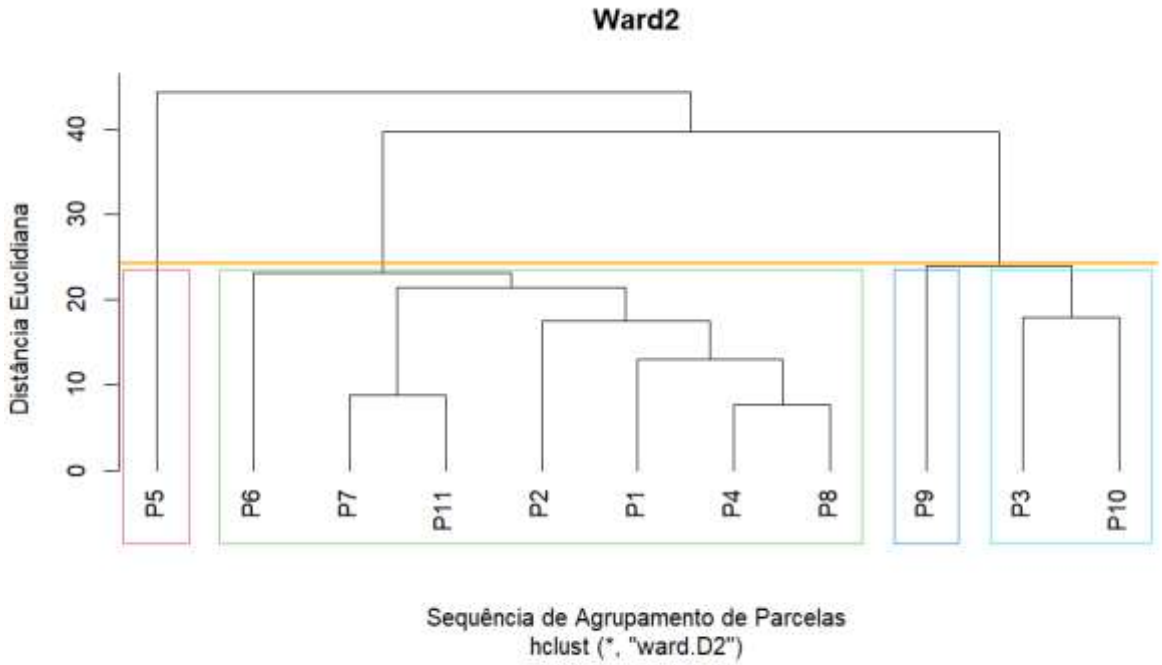



**Figura 14.** Dendrograma representando as sequências das fusões das parcelas obtido pelo emprego do método de Ward.D2, com base na distância euclidiana, considerando a linha fenon d = 24,2 e a formação de quatro grupos.

## 2.1.1.2. Método Hierárquico Divisivo

Os **métodos divisivos** operam na direção oposta ao sentido dos **métodos aglomerativos**, começando com um grande grupo, o qual é partido em dois, que são, novamente, divididos e assim sucessivamente, até que cada indivíduo isolado se constituia em um grupo. Dessa forma, esses métodos são exigentes computacionalmente se todas as $2^{k-1} - 1$ divisões possíveis em dois subgrupos de um grupo de **k** objetos forem consideradas em cada estágio. No entanto, para os dados que consistem de **p** variáveis binárias, métodos relativamente simples e computacionalmente eficientes, conhecidos como métodos divisivos monotéticos, estão disponíveis. Esses métodos geralmente dividem os grupos de acordo com a presença ou ausência de cada uma das p variáveis, de modo que cada cluster de estágio contenha membros com certos atributos, todos presentes ou todos ausentes. Os dados para esses métodos, portanto, precisam estar na forma de um modo de dois matriz (binária).

Embora menos comumente usados do que os aglomerativo, os métodos divisivos apresentam a vantagem de que a maioria dos usuários está interessada na estrutura principal de seus dados, e isso é revelado desde o início de um método divisivo (Kaufman; Rousseeuw, 2005).

Entre os métodos divisivos, o mais conhecido é o de Edwards e Cavalli-Sforza (1965):

### a) Método de Edwards e Cavalli-Sforza

No **método de Edwards e Cavalli-Sforza**, inicialmente, consideramos que todos os objetos são pertencentes a um único grupo. Assim, dividimos esse grupo em dois e, sucessivamente, cada novo grupo será dividido em outros dois, até que cada grupo contenha apenas um único objeto. Logo, o método se baseia em dois passos:

i) A partir de distâncias quadráticas ($\boldsymbol{d_{ij}^2}$), comparam-se todas as possíveis partições dos indivíduos em dois grupos, escolhendo-se aquela que resulte na menor soma de quadrados dentro dos grupos e, equivalentemente, na maior soma de quadrados entre os grupos; e

ii) O processo continua dentro de cada um dos dois grupos, baseando-se no mesmo critério, até chegar ao final com grupos individuais.

O resultado da aplicação do método é apresentado por meio de um diagrama de árvore. No exemplo das 11 parcelas da Tabela 1 (página 6), temos $(2^{n-1} - 1) = (2^{11-1} - 1) = 1023$ partições possíveis de parcelas em grupos não vazios e disjuntos. Uma forma de atenuar o problema das partições possíveis é ordenar as médias dos objetos, assim, o número de partições passa a ser obtido por $\boldsymbol{n - 1}$ (Ramalho et al., 2012). Como exemplo, podemos ter o diagrama apresentado na Figura 15 como uma das partições. Logo, evidencia-se que o Método de Edwards e Cavalli-Sforza é bastante simples, porém trabalhoso, fazendo-se necessário o uso de recursos computacionais.

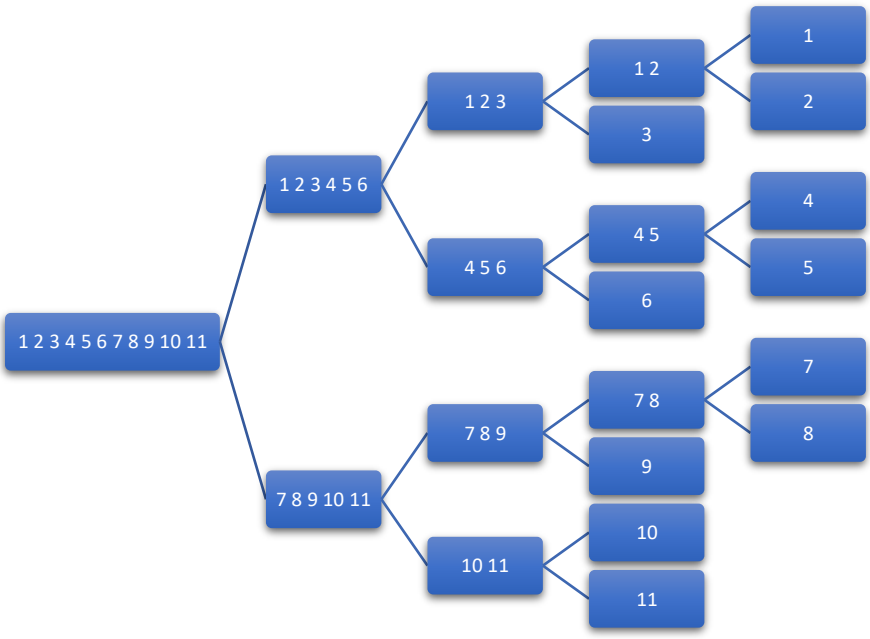



**Figura 15.** Diagrama de árvore conforme às 11 parcelas.

O problema principal se resume a obter um critério para dividir um grupo em dois grupos o mais distintos possíveis entre si. E, por meio da aplicação sucessiva desse critério, teremos a análise completa.

## 1. Critério para divisão de um grupo

Quando os objetos estão classificados em dois grupos, sabemos, da análise de variância, que a soma dos quadrados das distâncias de cada um à média geral (**SQT**) pode ser decomposta em duas partes (Pimentel, 1987):

1) **SQW**: soma de quadrados dentro dos grupos, a qual é subdividida em:

**$SQW_1$**: soma dos quadrados das distâncias de cada objeto do primeiro grupo à média deste grupo (soma de quadrados dentro do primeiro grupo); e

**$SQW_2$**: soma dos quadrados das distâncias de cada indivíduo do segundo grupo à média deste grupo (soma de quadrados dentro do segundo grupo).

2) **SQB**: soma dos quadrados das distâncias da média de cada grupo à média geral, ponderada pelo tamanho do grupo (soma de quadrados entre os grupos).

Um critério natural para obtermos os dois grupos mais distintos entre si é escolhermos a partição que maximiza a soma de quadrados entre grupos (**SQB**), cujo valor será, também, uma medida de "importância" da divisão. Os grupos obtidos continuarão a ser divididos até o final do processo, quando cada grupo conterá apenas um objeto, a soma de quadrados dentro de cada grupo será nula e o somatório das **SQB**'s associadas a cada partição feita será igual à **SQT** inicial (Pimentel, 1987).

Dessa forma, podemos obter as somas de quadrados por meio das seguintes expressões:

$$SQT = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ todos\ objetos}{g}$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ \ grupo\ 1}{k_1}$$

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ grupo\ 2}{k_2}$$

$$SQW = SQW_1 + SQW_2$$

$$SQB = SQT - SQW$$

Em que: **g** = número total de objetos; $k_1$ = número de objetos do grupo 1; e **$k_2$** = número de objetos do grupo 2.

A variância de um grupo pode ser tomada como uma medida da sua homogeneidade interna e é calculada dividindo-se a sua **SQT** pelo número de objetos que o compõem. Vale salientar que a $\mathbf{SQW_1}$ e $\mathbf{SQW_2}$ serão as **SQT** dos respectivos grupos na próxima divisão.

Do exposto, conclui-se que, em cada divisão, teremos como objetivo maximizar a **SQB** e, consequentemente, minimizar a **SQW**.

Vale ressaltar que o objetivo da análise de agrupamento é obter grupos homogêneos entre si. Logo, deve-se buscar um critério para determinar a melhor partição de um grupo e outro critério para decidir quando um grupo não deve ser mais dividido.

## 2. Critério para obtenção de grupos

Inicialmente, a partir do grupo formado por todos os objetos, as suas médias podem ser divididas em dois grupos distintos e homogêneos ($\mathbf{g} = \mathbf{k_1}$ e $\mathbf{g} = \mathbf{k_2}$). Sendo assim, podemos formular as hipóteses:

$\mathbf{H_0}$: $\boldsymbol{\mu_i = \mu}$ ($\mathbf{i} = 1, 2, ..., \mathbf{g}$), ou seja, os dois grupos são idênticos a determinada probabilidade $\boldsymbol{\alpha}$.

$H_1$: $\boldsymbol{\mu_i = \mu_1}$; $\boldsymbol{\mu_i = \mu_2}$ e $\boldsymbol{\mu_i = \mu_1}$ *e* $\boldsymbol{\mu_2}$ , ou seja, os dois grupos não são idênticos a $\boldsymbol{\alpha}$.

Em que: $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ representam as médias desconhecidas dos dois grupos.

Para testar a significância de que os **n** objetos podem ser divididos em dois grupos que maximizem a soma de quadrados entre grupos, premissa do **Método de Edwards e Cavalli-Sforza** (1965), pode-se utilizar o **procedimento de Scott-Knott** (1974) como critério, o qual se baseia na **razão de verossimilhança** sob $\mathbf{H}_1$.

Uma vez ordenadas as médias, procede-se do seguinte modo, fazendo inicialmente o número de grupos = número de objetos (Ramalho et al., 2012).

a) Determinar a partição entre dois grupos que maximizem a soma de quadrados entre grupos (**SQB**).

Os dois grupos deverão ser identificados por meio da inspeção das somas de quadrados das g-1 partições possíveis, sendo g igual ao número de objetos envolvidos no grupo de médias considerado.

b) Determinar o valor da estatística $\boldsymbol{l} = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{SQB}{\hat{\sigma}_0^2}\right)$

Em que: $\boldsymbol{\hat{\sigma}_0^2}$ é o estimador de máxima verossimilhança de $\boldsymbol{\sigma_{\bar{Y}}^2}$. Seja $S_{\bar{Y}}^2 = \frac{QMR}{r}$ o estimador não viesado de $\boldsymbol{\sigma_{\bar{Y}}^2}$ e $\boldsymbol{v}$ igual aos graus de liberdade associados a estimador.

$$\hat{\sigma}_0^2 = \frac{1}{g+v}\left[\sum_{i=1}^{g}\left(\bar{Y}_{(i)} - \bar{Y}\right)^2 + v\,S_{\bar{Y}}^2\right] \rightarrow \hat{\sigma}_0^2 = \frac{1}{g+v}\left[SQT + vS_{\bar{Y}}^2\right]$$

c) Se $\boldsymbol{l \geq c^2_{(\alpha;g/(\pi-2))}}$, rejeita-se a hipótese de que os dois grupos são idênticos ($\mathbf{H_0}$) em favor da hipótese alternativa ($\mathbf{H_a}$) de que os dois grupos diferem a α probabilidade e $\boldsymbol{g/(\pi-2)}$ graus de liberdade.

d) No caso de rejeitar $\mathbf{H_0}$, os dois grupos serão independentemente submetidos aos passos (a) a (c), fazendo $\boldsymbol{g = k_1}$ e $\boldsymbol{g = k_2}$. O processo em cada subgrupo se encerra ao aceitar $H_0$ no passo (c) ou se cada grupo contiver apenas uma média.

**Aplicação**

Consideremos que temos um experimento inteiramente casualizado com 11 tratamentos (parcelas) e 17 observações em cada um deles (Tabela 29). Logo, o modelo estatístico é:

$$y_{ij} = \mu + t_i + \varepsilon_{ij}$$

Em que: $\boldsymbol{y_{ij}}$= **j**-ésima observação (**j** = 1, 2, ..., 17) no **i**-ésimo tratamento (parcela) (**i** = 1, 2, ..., **i**); $\boldsymbol{\mu}$ = média geral; $\boldsymbol{t_i}$= efeito do **i**-ésimo tratamento; e $\boldsymbol{\varepsilon_{ij}}$ = erro aleatório associado à **j**-ésima observação no **i**-ésimo tratamento.

**Tabela 29.** Densidade de 17 espécies da Mata da Silvicultura, em parcelas de 20 x 50 m, município de Viçosa-MG

| Espécie | Tratamentos (Parcelas) | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | |
| 1 | 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 | |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 | |
| 3 | 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 | |
| 4 | 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 | |
| 5 | 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | |
| 6 | 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 | |
| 7 | 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 | |
| 8 | 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 | |
| 9 | 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 | |
| 10 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 | |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 | |
| 12 | 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 | |
| 13 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 | |
| 14 | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 17 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Total Amostral | 33 | 26 | 56 | 19 | 78 | 53 | 31 | 22 | 60 | 62 | 42 | 482 |
| Média Amostral | 1,94 | 1,53 | 3,29 | 1,12 | 4,59 | 3,12 | 1,82 | 1,29 | 3,53 | 3,65 | 2,47 | 2,58 |
| Tamanho da Amostra | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 187 |

Vale salientar que, para empregar a análise de variância adequadamente, é necessário testar as pressuposições básicas da variável resposta: distribuição normal e homoscedasticidade. No presente exemplo, não realizamos esses testes, pois objetivamos apenas ilustrar como seria a aplicação do Método de Edwards e Cavalli-Sforza. Porém, a rigor, devem ser realizados e,

caso não haja atendimento às referidas pressuposições, o pesquisador deverá recorrer às técnicas de transformações de variáveis, como, por exemplo, a de Box e Cox (1964).

1º Passo: Análise de Variância

**Cálculos**

$$GLTotal = IJ - 1 = 11x17 - 1 = 481$$

$$GLTratamentos = I - 1 = 11 - 1 = 10$$

$$GLResíduo = GLTotal - GLTratamentos = 481 - 10 = 471$$

$$SQTotal = \sum_{i=1,j=1}^{I,J} Y_{ij}^2 - \frac{\left(\sum_{i=1,j=1}^{I,J} Y_{ij}\right)^2}{IxJ} = 5394 - \frac{(482)^2}{187} = 4151{,}63$$

$$SQTratamentos = \left(\frac{\sum_{i=1}^{I} Y_{i.}^2}{J}\right) - \frac{\left(\sum_{i=1,j=1}^{I,J} Y_{ij}\right)^2}{IxJ} = 1459{,}29 - \frac{(482)^2}{187} = 216{,}92$$

$$SQResíduo = SQTotal - SQTratamentos = 3934{,}71$$

$$QRTratamentos = \frac{SQTratamentos}{GLTratamentos} = \frac{216{,}92}{10} = 21{,}69$$

$$QMResíduo = \frac{SQResíduo}{GLResíduo} = \frac{3934{,}71}{471} = 8{,}35$$

$$F = \frac{QMTratamentos}{QMResíduos} = \frac{21{,}69}{8{,}35} = 2{,}60$$

Assim, de acordo com a Tabela 30, podemos afirmar que existem parcelas que diferem significativamente entre si ($p < 0{,}05$). O que indica que seria adequado aplicarmos a técnica de Scott e Knott para obtermos grupos de parcelas.

**Tabela 30.** Análise de variância considerando um delineamento inteiramente casualizado.

| FV | GL | SQ | QM | F | p |
|---|---|---|---|---|---|
| **Tratamentos (Parcelas)** | 10 | 216,92 | 21,69 | 2,60 | 0,0045 |
| **Resíduo** | 471 | 3934,71 | 8,35 | | |
| **Total** | 481 | 4151,63 | | | |

2º Passo: Obtenção dos Grupos

**Médias das Parcelas**

$\bar{Y}_{i.} = \frac{\sum_{i=1}^{I} Y_{i.}}{J}$

$\bar{Y}_1 = \frac{33}{17} = 1{,}94$; $\bar{Y}_2 = \frac{26}{17} = 1{,}53$; $\bar{Y}_3 = \frac{56}{17} = 3{,}29$; $\bar{Y}_4 = \frac{19}{17} = 1{,}12$; $\bar{Y}_5 = \frac{78}{17} = 4{,}59$; $\bar{Y}_6 = \frac{53}{17} = 3{,}12$; $\bar{Y}_7 = \frac{31}{17} = 1{,}82$; $\bar{Y}_8 = \frac{22}{17} = 1{,}29$; $\bar{Y}_9 = \frac{60}{17} = 3{,}53$; $\bar{Y}_{10} = \frac{62}{17} = 3{,}65$; e $\bar{Y}_{11} = \frac{42}{17} = 2{,}47$

**Estimativa da variância da média**

$$S_{\bar{Y}}^2 = \frac{QMResíduo}{n} = \frac{8{,}35}{17} = 0{,}49$$

**Matriz dos Quadrados das Distâncias entre as Médias**

Por meio da recomendação de Ramalho et al. (2012), ordenamos as médias dos objetos (Tabela), portanto, o número de partições passa a ser obtido por $11 - 1 = 10$. Assim, resolvemos realizar o agrupamento partindo da sequência **P4, P8, P7, P1, P11, P6, P3, P9, P10** e **P5.**

E, a partir da Tabela 31, calculamos as distâncias euclidianas quadradas médias.

**Tabela 31.** Ordenação das médias de indivíduos por parcelas

| Parcela | 4 | 8 | 2 | 7 | 1 | 11 | 6 | 3 | 9 | 10 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Média de Indivíduos | 1,12 | 1,29 | 1,53 | 1,82 | 1,94 | 2,47 | 3,12 | 3,29 | 3,53 | 3,65 | 4,59 |

$d_{ii'}^2 = (\bar{Y}_i - \bar{Y}_{i'})^2$ com i = 4, 8, ..., 10, i' = 8, 2, ..., 5 e i ≠ i'.

$$d_{48}^2 = (\bar{Y}_4 - \bar{Y}_8)^2 = (1{,}12 - 1{,}29)^2 = 0{,}03$$
$$d_{42}^2 = (\bar{Y}_4 - \bar{Y}_2)^2 = (1{,}12 - 1{,}53)^2 = 0{,}17$$
$$d_{47}^2 = (\bar{Y}_4 - \bar{Y}_7)^2 = (1{,}12 - 1{,}82)^2 = 0{,}50$$
$$d_{41}^2 = (\bar{Y}_4 - \bar{Y}_1)^2 = (1{,}12 - 1{,}94)^2 = 0{,}68$$
$$d_{411}^2 = (\bar{Y}_4 - \bar{Y}_{11})^2 = (1{,}12 - 2{,}47)^2 = 1{,}83$$
$$d_{46}^2 = (\bar{Y}_4 - \bar{Y}_6)^2 = (1{,}12 - 3{,}12)^2 = 4{,}00$$
$$d_{43}^2 = (\bar{Y}_4 - \bar{Y}_3)^2 = (1{,}12 - 3{,}29)^2 = 4{,}74$$
$$d_{49}^2 = (\bar{Y}_4 - \bar{Y}_9)^2 = (1{,}12 - 3{,}53)^2 = 5{,}82$$
$$d_{410}^2 = (\bar{Y}_4 - \bar{Y}_{10})^2 = (1{,}12 - 3{,}65)^2 = 6{,}40$$
$$d_{45}^2 = (\bar{Y}_4 - \bar{Y}_5)^2 = (1{,}12 - 4{,}59)^2 = 12{,}04$$
$$d_{82}^2 = (\bar{Y}_8 - \bar{Y}_2)^2 = (1{,}29 - 1{,}53)^2 = 0{,}06$$
$$d_{87}^2 = (\bar{Y}_8 - \bar{Y}_7)^2 = (1{,}29 - 1{,}82)^2 = 0{,}28$$
$$d_{81}^2 = (\bar{Y}_8 - \bar{Y}_1)^2 = (1{,}29 - 1{,}94)^2 = 0{,}42$$
$$d_{811}^2 = (\bar{Y}_8 - \bar{Y}_{11})^2 = (1{,}29 - 2{,}47)^2 = 1{,}38$$
$$d_{86}^2 = (\bar{Y}_8 - \bar{Y}_6)^2 = (1{,}29 - 3{,}12)^2 = 3{,}33$$
$$d_{83}^2 = (\bar{Y}_8 - \bar{Y}_3)^2 = (1{,}29 - 3{,}29)^2 = 4{,}00$$
$$d_{89}^2 = (\bar{Y}_8 - \bar{Y}_9)^2 = (1{,}29 - 3{,}53)^2 = 5{,}00$$
$$d_{810}^2 = (\bar{Y}_8 - \bar{Y}_{10})^2 = (1{,}29 - 3{,}65)^2 = 5{,}54$$
$$d_{85}^2 = (\bar{Y}_8 - \bar{Y}_5)^2 = (1{,}29 - 4{,}59)^2 = 10{,}85$$

$$d_{27}^2 = (\bar{Y}_2 - \bar{Y}_7)^2 = (1{,}53 - 1{,}82)^2 = 0{,}09$$
$$d_{21}^2 = (\bar{Y}_2 - \bar{Y}_1)^2 = (1{,}53 - 1{,}94)^2 = 0{,}17$$
$$d_{211}^2 = (\bar{Y}_2 - \bar{Y}_{11})^2 = (1{,}53 - 2{,}47)^2 = 0{,}89$$
$$d_{26}^2 = (\bar{Y}_2 - \bar{Y}_6)^2 = (1{,}53 - 3{,}12)^2 = 2{,}52$$
$$d_{23}^2 = (\bar{Y}_2 - \bar{Y}_3)^2 = (1{,}53 - 3{,}29)^2 = 3{,}11$$
$$d_{29}^2 = (\bar{Y}_2 - \bar{Y}_9)^2 = (1{,}53 - 3{,}53)^2 = 4{,}00$$
$$d_{210}^2 = (\bar{Y}_2 - \bar{Y}_{10})^2 = (1{,}53 - 3{,}65)^2 = 4{,}48$$
$$d_{25}^2 = (\bar{Y}_2 - \bar{Y}_5)^2 = (1{,}53 - 4{,}59)^2 = 9{,}36$$
$$d_{71}^2 = (\bar{Y}_7 - \bar{Y}_1)^2 = (1{,}82 - 1{,}94)^2 = 0{,}01$$
$$d_{711}^2 = (\bar{Y}_7 - \bar{Y}_{11})^2 = (1{,}82 - 2{,}47)^2 = 0{,}42$$
$$d_{76}^2 = (\bar{Y}_7 - \bar{Y}_6)^2 = (1{,}82 - 3{,}12)^2 = 1{,}67$$
$$d_{73}^2 = (\bar{Y}_7 - \bar{Y}_3)^2 = (1{,}82 - 3{,}29)^2 = 2{,}16$$
$$d_{79}^2 = (\bar{Y}_7 - \bar{Y}_9)^2 = (1{,}82 - 3{,}53)^2 = 2{,}91$$
$$d_{710}^2 = (\bar{Y}_7 - \bar{Y}_{10})^2 = (1{,}82 - 3{,}65)^2 = 3{,}33$$
$$d_{75}^2 = (\bar{Y}_7 - \bar{Y}_5)^2 = (1{,}82 - 4{,}59)^2 = 7{,}64$$
$$d_{111}^2 = (\bar{Y}_1 - \bar{Y}_{11})^2 = (1{,}94 - 2{,}47)^2 = 0{,}28$$
$$d_{16}^2 = (\bar{Y}_1 - \bar{Y}_6)^2 = (1{,}94 - 3{,}12)^2 = 1{,}38$$
$$d_{13}^2 = (\bar{Y}_1 - \bar{Y}_3)^2 = (1{,}94 - 3{,}29)^2 = 1{,}83$$
$$d_{19}^2 = (\bar{Y}_1 - \bar{Y}_9)^2 = (1{,}94 - 3{,}53)^2 = 2{,}52$$
$$d_{110}^2 = (\bar{Y}_1 - \bar{Y}_{10})^2 = (1{,}94 - 3{,}65)^2 = 2{,}91$$
$$d_{15}^2 = (\bar{Y}_1 - \bar{Y}_5)^2 = (1{,}94 - 4{,}59)^2 = 7{,}01$$
$$d_{116}^2 = (\bar{Y}_{11} - \bar{Y}_6)^2 = (2{,}47 - 3{,}12)^2 = 0{,}42$$
$$d_{113}^2 = (\bar{Y}_{11} - \bar{Y}_3)^2 = (2{,}47 - 3{,}29)^2 = 0{,}68$$
$$d_{119}^2 = (\bar{Y}_{11} - \bar{Y}_9)^2 = (2{,}47 - 3{,}53)^2 = 1{,}12$$
$$d_{1110}^2 = (\bar{Y}_{11} - \bar{Y}_{10})^2 = (2{,}47 - 3{,}65)^2 = 1{,}38$$
$$d_{115}^2 = (\bar{Y}_{11} - \bar{Y}_5)^2 = (2{,}47 - 4{,}59)^2 = 4{,}48$$
$$d_{63}^2 = (\bar{Y}_6 - \bar{Y}_3)^2 = (3{,}12 - 3{,}29)^2 = 0{,}03$$
$$d_{69}^2 = (\bar{Y}_6 - \bar{Y}_9)^2 = (3{,}12 - 3{,}53)^2 = 0{,}17$$
$$d_{610}^2 = (\bar{Y}_6 - \bar{Y}_{10})^2 = (3{,}12 - 3{,}65)^2 = 0{,}28$$
$$d_{65}^2 = (\bar{Y}_6 - \bar{Y}_5)^2 = (3{,}12 - 4{,}59)^2 = 2{,}16$$
$$d_{39}^2 = (\bar{Y}_3 - \bar{Y}_9)^2 = (3{,}29 - 3{,}53)^2 = 0{,}06$$
$$d_{310}^2 = (\bar{Y}_3 - \bar{Y}_{10})^2 = (3{,}29 - 3{,}65)^2 = 0{,}12$$
$$d_{35}^2 = (\bar{Y}_3 - \bar{Y}_5)^2 = (3{,}29 - 4{,}59)^2 = 1{,}67$$

$$d_{910}^2 = (\bar{Y}_9 - \bar{Y}_{10})^2 = (3{,}53 - 3{,}65)^2 = 0{,}01$$
$$d_{95}^2 = (\bar{Y}_9 - \bar{Y}_5)^2 = (3{,}53 - 4{,}59)^2 = 1{,}12$$
$$d_{105}^2 = (\bar{Y}_{10} - \bar{Y}_5)^2 = (3{,}65 - 4{,}59)^2 = 0{,}89$$

Obtendo-se, assim, a matriz $D_{ii'}^2$:

$D_{ii'}^2 =$

| | P4 | P8 | P2 | P7 | P1 | P11 | P6 | P3 | P9 | P10 | P5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P4 | 0,00 | 0,03 | 0,17 | 0,50 | 0,68 | 1,83 | 4,00 | 4,74 | 5,82 | 6,40 | 12,04 |
| P8 | | 0,00 | 0,06 | 0,28 | 0,42 | 1,38 | 3,33 | 4,00 | 5,00 | 5,54 | 10,85 |
| P2 | | | 0,00 | 0,09 | 0,17 | 0,89 | 2,52 | 3,11 | 4,00 | 4,48 | 9,36 |
| P7 | | | | 0,00 | 0,01 | 0,42 | 1,67 | 2,16 | 2,91 | 3,33 | 7,64 |
| P1 | | | | | 0,00 | 0,28 | 1,38 | 1,83 | 2,52 | 2,91 | 7,01 |
| P11 | | | | | | 0,00 | 0,42 | 0,68 | 1,12 | 1,38 | 4,48 |
| P6 | | | | | | | 0,00 | 0,03 | 0,17 | 0,28 | 2,16 |
| P3 | | | | | | | | 0,00 | 0,06 | 0,12 | 1,67 |
| P9 | | | | | | | | | 0,00 | 0,01 | 1,12 |
| P10 | | | | | | | | | | 0,00 | 0,89 |
| P5 | | | | | | | | | | | 0,00 |

Cálculos

**Partição do Grupo P4 P8 P7 P1 P11 P6 P3 P9 P10 P5**

Seja k o número de médias do grupo a ser particionado, e $k_1$ e $k_2$, o dos seus respectivos subgrupos. Assim, teremos:

$$SQT = \frac{Soma\ das\ distâncias\ quadradas\ entre\ todos\ os\ objetos}{g}$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1}$$

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} SQW = SQW_1 + SQW_2$$

$$SQB = SQT - SQW$$

$$SQT = \frac{Soma\ da\ distâncias\ entre\ todos\ os\ objetos}{g} =$$

$$\frac{\begin{pmatrix} d_{4,8}^2+\cdots+d_{4,5}^2+d_{2,8}^2+\cdots+d_{8,5}^2+d_{2,7}^2+\cdots+d_{2,5}^2+d_{7,1}^2+\cdots+d_{7,5}^2+d_{1,11}^2+\cdots+d_{1,5}^2 \\ +d_{11,6}^2+\cdots+d_{11,5}^2+d_{6,3}^2+\cdots+d_{6,5}^2+d_{3,9}^2+\cdots+d_{3,5}^2+d_{9,10}^2+d_{9,5}^2+d_{10,5}^2 \end{pmatrix}}{g} = \frac{(0{,}03+0{,}17+\cdots+0{,}89)}{11} = \frac{140{,}360}{11} =$$

$12{,}7600$

$$\hat{\sigma}_0^2 = \left(\frac{1}{g+v}\right)\left[SQT + vS_{\bar{Y}}^2\right] \rightarrow \hat{\sigma}_0^2 = \left(\frac{1}{11+471}\right)[12{,}7600 + 471x0{,}49] = 0{,}51$$

1ª Partição: 4 | 8 2 7 1 11 6 3 9 10 5 ($k_1 = 1$ e $k_2 = 10$), ou seja, grupo 1 – objeto 4, contra grupo 2 – objetos 8 2 7 1 11 6 3 9 10 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\begin{pmatrix} d_{8,2}^2+d_{8,7}^2+d_{8,1}^2+d_{8,11}^2+d_{8,6}^2+d_{8,3}^2+d_{8,9}^2+d_{8,10}^2+d_{8,5}^2+d_{2,7}^2+d_{2,1}^2+d_{2,11}^2+d_{2,6}^2 \\ +d_{2,3}^2+d_{2,9}^2+d_{2,10}^2+d_{2,5}^2+d_{7,1}^2+d_{7,11}^2+d_{7,6}^2+d_{7,3}^2+d_{7,9}^2+d_{7,10}^2+d_{7,5}^2+d_{1,11}^2 \\ +d_{1,6}^2+d_{1,3}^2+d_{1,9}^2+d_{1,10}^2+d_{1,5}^2+d_{11,6}^2+d_{11,3}^2+d_{11,9}^2+d_{11,10}^2+d_{11,5}^2+d_{6,3}^2+d_{6,9}^2 \\ +d_{6,10}^2+d_{6,5}^2+d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+d_{9,10}^2+d_{9,5}^2+d_{10,5}^2 \end{pmatrix}}{k_2} =$$

$$\frac{\begin{pmatrix} 0{,}06+0{,}28+0{,}42+1{,}38+3{,}33+4{,}00+5{,}00+5{,}54+10{,}85+0{,}09+0{,}17 \\ +0{,}89+2{,}52+3{,}11+4{,}00+4{,}48+9{,}36+0{,}01+0{,}42+1{,}67+2{,}16+2{,}91 \\ +3{,}33+7{,}64+0{,}28+1{,}38+1{,}83+2{,}52+2{,}91+7{,}01+0{,}42+0{,}68+1{,}12 \\ +1{,}38+4{,}48+0{,}03+0{,}17+0{,}28+2{,}16+0{,}06+0{,}12+1{,}67+0{,}01+1{,}12+0{,}89 \end{pmatrix}}{10} = \frac{104{,}156}{10} = 10{,}4156$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} = \frac{d_{4,4}^2}{k_1} = \frac{0{,}00}{1} = 0{,}00$$

$$SQW = SQW_1 + SQW_2 = 0{,}00 + 10{,}4156 = 10{,}4156$$

$$SQB = SQT - SQW = 12{,}7600 - 10{,}4156 = 2{,}3444$$

2ª Partição: 4 8 | 2 7 1 11 6 3 9 10 5 ($k_1 = 2$ e $k_2 = 9$), ou seja, grupo 1 – objetos 4 e 8, contra grupo 2 – objetos 2 7 1 11 6 3 9 10 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\begin{pmatrix} d_{2,7}^2+d_{2,1}^2+d_{2,11}^2+d_{2,6}^2+d_{2,3}^2+d_{2,9}^2+d_{2,10}^2+d_{2,5}^2+d_{7,1}^2+d_{7,11}^2+d_{7,6}^2+d_{7,3}^2 \\ +d_{7,9}^2+d_{7,10}^2+d_{7,5}^2+d_{1,11}^2+d_{1,6}^2+d_{1,3}^2+d_{1,9}^2+d_{1,10}^2+d_{1,5}^2+d_{11,6}^2+d_{11,3}^2+d_{11,9}^2 \\ +d_{11,10}^2+d_{11,5}^2+d_{6,3}^2+d_{6,9}^2+d_{6,10}^2+d_{6,5}^2+d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+d_{9,10}^2+d_{9,5}^2+d_{10,5}^2 \end{pmatrix}}{k_2} =$$

$$\frac{\begin{pmatrix} 0{,}09+0{,}17+0{,}89+2{,}52+3{,}11+4{,}00+4{,}48+9{,}36+0{,}01+0{,}42+1{,}67 \\ +2{,}16+2{,}91+3{,}33+7{,}64+0{,}28+1{,}38+1{,}83+2{,}52+2{,}91+7{,}01 \\ +0{,}42+0{,}68+1{,}12+1{,}38+4{,}48+0{,}03+0{,}17+0{,}28+2{,}16+0{,}06 \\ +0{,}12+1{,}67+0{,}01+1{,}12+0{,}89 \end{pmatrix}}{9} = \frac{73{,}308}{9} = 8{,}1453$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_1} = \frac{d_{4,8}^2}{k_1} = \frac{0{,}03}{2} = 0{,}0156$$

$$SQW = SQW_1 + SQW_2 = 0{,}0156 + 8{,}1453 = 8{,}1609$$

$$SQB = SQT - SQW = 12{,}7600 - 8{,}1609 = 4{,}5991$$

3ª Partição: 4 8 2 | 7 1 11 6 3 9 10 5 ($k_1 = 3$ e $k_2 = 8$), ou seja, grupo 1 – objetos 4, 8 e 2, contra grupo 2 – objetos 7 1 11 6 3 9 10 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\begin{pmatrix} d_{7,1}^2+d_{7,11}^2+d_{7,6}^2+d_{7,3}^2+d_{7,9}^2+d_{7,10}^2+d_{7,5}^2+d_{1,11}^2+d_{1,6}^2+d_{1,3}^2+d_{1,9}^2+d_{1,10}^2+d_{1,5}^2 \\ +d_{11,6}^2+d_{11,3}^2+d_{11,9}^2+d_{11,10}^2+d_{11,5}^2+d_{6,3}^2+d_{6,9}^2 \\ +d_{6,10}^2+d_{6,5}^2+d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+d_{9,10}^2+d_{9,5}^2+d_{10,5}^2 \end{pmatrix}}{k_2} ==$$

$$\frac{\begin{pmatrix} 0{,}01+0{,}42+1{,}67+2{,}16+2{,}91+3{,}33+7{,}64+0{,}28+1{,}38+1{,}83+2{,}52 \\ +2{,}91+7{,}01+0{,}42+0{,}68+1{,}12+1{,}38+4{,}48+0{,}03+0{,}17 \\ +0{,}28+2{,}16+0{,}06+0{,}12+1{,}67+0{,}01+1{,}12+0{,}89 \end{pmatrix}}{8} = \frac{48{,}867}{8} = 6{,}0861$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} = \frac{d_{4,8}^2+d_{4,2}^2+d_{8,2}^2}{k_1} =$$

$$\frac{(0{,}03++0{,}17+0{,}06)}{3} = \frac{0{,}2561}{3} = 0{,}0854$$

$$SQW = SQW_1 + SQW_2 = 0{,}0854 + 6{,}0861 = 6{,}1714$$

$$SQB = SQT - SQW = 12{,}7600 - 6{,}1714 = 6{,}5886$$

4ª Partição: 4 8 2 7 | 1 11 6 3 9 10 5 ($k_1 = 4$ e $k_2 = 7$), ou seja, grupo 1 – objetos 4, 8, 2 e 7, contra grupo 2 – objetos 1 11 6 3 9 10 e 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\begin{pmatrix} d_{1,11}^2+d_{1,6}^2+d_{1,3}^2+d_{1,9}^2+d_{1,10}^2+d_{1,5}^2+d_{11,6}^2+d_{11,3}^2+d_{11,9}^2+d_{11,10}^2+d_{11,5}^2+d_{6,3}^2 \\ +d_{6,9}^2+d_{6,10}^2+d_{6,5}^2++d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+d_{9,10}^2+d_{9,5}^2+d_{10,5}^2 \end{pmatrix}}{k_2} ==$$

$$\frac{\begin{pmatrix} 0{,}28+1{,}38+1{,}83+2{,}52+2{,}91+7{,}01+0{,}42+0{,}68+1{,}12+1{,}38+4{,}48+0{,}03+0{,}17+ \\ 0{,}28+2{,}16+0{,}06+0{,}12+1{,}67+0{,}01+1{,}12+0{,}89 \end{pmatrix}}{7} = \frac{30{,}5398}{7} = 4{,}3628$$

$$SQW_1 = \frac{\sum_{i=1}^{1}\sum_{j=1}^{1}(d_{ij}^2)}{k_1} = \frac{d_{2,4}^2+d_{2,7}^2+d_{2,8}^2+d_{4,7}^2+d_{4,8}^2+d_{7,8}^2}{k_1} = \frac{(0{,}03+0{,}17+0{,}06+0{,}50+0{,}28+0{,}99)}{4} = \frac{1{,}1211}{4} =$$

$0{,}2803$

$$SQW = SQW_1 + SQW_2 = 0{,}2803 + 4{,}3628 = 4{,}6431$$

$$SQB = SQT - SQW = 12{,}7600 - 4{,}6431 = 8{,}1169$$

5ª Partição: 4 8 2 7 1 | 11 6 3 9 10 5 ($k_1 = 5$ e $k_2 = 6$), ou seja, grupo 1 – objetos 4, 8, 2, 7 e 1, contra grupo 2 – objetos 11 6 3 9 10 e 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\begin{pmatrix} d_{11,6}^2+d_{11,3}^2+d_{11,9}^2+d_{11,10}^2+d_{11,5}^2+d_{6,3}^2+d_{6,9}^2+d_{6,10}^2+d_{6,5}^2+d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+ \\ +d_{9,10}^2+d_{9,5}^2+d_{10,5}^2 \end{pmatrix}}{k_2} =$$

$$\frac{\begin{pmatrix} 0{,}42+0{,}68+1{,}12+1{,}38+4{,}48+0{,}03+0{,}17+ \\ 0{,}28+2{,}16+0{,}06+0{,}12+1{,}67+0{,}01+1{,}12+0{,}89 \end{pmatrix}}{6} = \frac{14{,}6055}{6} = 2{,}4343$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} =$$

$$\frac{\left(d_{4,8}^2+d_{4,2}^2+d_{4,7}^2+d_{4,1}^2+d_{8,2}^2+d_{8,7}^2+d_{8,1}^2+d_{2,7}^2+d_{2,1}^2+d_{7,1}^2\right)}{k_1} =$$

$$\frac{(0{,}03+0{,}17+0{,}50+0{,}68+0{,}06+0{,}28+0{,}42+0{,}09+0{,}17+0{,}01)}{5} = \frac{2{,}4014}{5} = 0{,}4803$$

$$SQW = SQW_1 + SQW_2 = 0{,}4803 + 2{,}4343 = 2{,}9145$$

$$SQB = SQT - SQW = 12{,}7600 - 2{,}9145 = 9{,}8455$$

6ª Partição: 4 8 2 7 1 11 | 6 3 9 10 5 ($k_1 = 6$ e $k_2 = 5$), ou seja, grupo 1 – objetos 4, 8, 2, 7, 1 e 11, contra grupo 2 – objetos 6 3 9 10 e 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\left(d_{6,3}^2+d_{6,9}^2+d_{6,10}^2+d_{6,5}^2+d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+d_{9,10}^2+d_{9,5}^2+d_{10,5}^2\right)}{k_2} =$$

$$\frac{(0{,}03+0{,}17+0{,}28+2{,}16+0{,}06+0{,}12+1{,}67+0{,}01+1{,}12+0{,}89)}{5} = \frac{6{,}5190}{5} = 1{,}3038$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1}$$

$$= \frac{\begin{pmatrix} d_{4,8}^2 + d_{4,2}^2 + d_{4,7}^2 + d_{4,1}^2 + d_{4,11}^2 + d_{8,2}^2 + d_{8,7}^2 + d_{8,1}^2 + d_{8,11}^2 + d_{2,7}^2 + d_{2,1}^2 + d_{2,11}^2 \\ + d_{7,1}^2 + d_{7,11}^2 + d_{1,11}^2 \end{pmatrix}}{k_1}$$

$$= \frac{\begin{pmatrix} 0{,}03 + 0{,}17 + 0{,}50 + 0{,}68 + 1{,}83 + 0{,}06 + 0{,}28 + 0{,}42 + 1{,}38 + 0{,}09 + 0{,}17 + 0{,}89 + 0{,}01 \\ + 0{,}42 + 0{,}28 \end{pmatrix}}{6}$$

$$= \frac{7{,}2007}{6} = 1{,}2001$$

$$SQW = SQW_1 + SQW_2 = 1{,}2001 + 1{,}3038 = 2{,}5039$$

$$SQB = SQT - SQW = 12{,}7600 - 2{,}5039 = 10{,}2561$$

7ª Partição: 4 8 2 7 1 11 6 | 3 9 10 5 ($k_1 = 7$ e $k_2 = 4$), ou seja, grupo 1 – objetos 4, 8, 2, 7, 1, 11 e 6, contra grupo 2 – objetos 3 9 10 e 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_1} =$$

$$\frac{\left(d_{3,9}^2+d_{3,10}^2+d_{3,5}^2+d_{9,10}^2+d_{9,,5}^2+d_{10,5}^2\right)}{k_1} = \frac{(0{,}06+0{,}12+1{,}67+0{,}01+1{,}12+0{,}89)}{4} = \frac{3{,}8754}{4} = 0{,}9689$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} =$$

$$\frac{\begin{matrix} d_{4,8}^2+d_{4,2}^2+d_{4,7}^2+d_{4,1}^2+d_{4,11}^2+d_{4,6}^2+d_{8,2}^2+d_{8,7}^2+d_{8,1}^2+d_{8,11}^2+d_{8,6}^2 \\ +d_{2,7}^2+d_{2,1}^2+d_{2,11}^2+d_{2,6}^2+d_{7,1}^2+d_{7,11}^2+d_{7,6}^2+d_{1,11}^2+d_{1,6}^2+d_{6,11}^2 \end{matrix}}{k_1} =$$

$$\frac{\begin{pmatrix} 0{,}03+0{,}17+0{,}50+0{,}68+1{,}83+4{,}00+0{,}06+0{,}28+0{,}42+1{,}38+3{,}33+0{,}09 \\ +0{,}17+0{,}89+2{,}52+0{,}01+0{,}42+1{,}67+0{,}28+1{,}38+0{,}42 \end{pmatrix}}{7} = \frac{20{,}5260}{7} = 2{,}9323$$

$$SQW = SQW_1 + SQW_2 = 2{,}9323 + 0{,}9689 = 3{,}9011$$

$$SQB = SQT - SQW = 12{,}7600 - 3{,}9011 = 8{,}8589$$

8ª Partição: 4 8 2 7 1 11 6 3 | 9 10 5 ($k_1 = 8$ e $k_2 = 3$), ou seja, grupo 1 – objetos 4, 8, 2, 7, 1, 11, 6 e 3, contra grupo 2 – objetos 9 10 e 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{(d_{9,10}^2+d_{9,5}^2+d_{10,5}^2)}{k_2} =$$

$$\frac{(0{,}01+1{,}12+0{,}89)}{3} = \frac{2{,}0208}{3} = 0{,}6736$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} =$$

$$\frac{\begin{pmatrix} d_{4,8}^2+d_{4,2}^2+d_{4,7}^2+d_{4,1}^2+d_{4,11}^2+d_{4,6}^2+d_{4,3}^2+d_{8,2}^2+d_{8,7}^2+d_{8,1}^2+d_{8,11}^2+d_{8,6}^2 \\ +d_{8,3}^2+d_{2,7}^2+d_{2,1}^2+d_{2,11}^2+d_{2,6}^2+d_{2,3}^2+d_{7,1}^2+d_{7,11}^2+d_{7,6}^2+d_{7,3}^2+d_{1,11}^2+d_{1,6}^2 \\ +d_{1,3}^2+d_{11,6}^2+d_{11,3}^2+d_{6,3}^2 \end{pmatrix}}{k_1} =$$

$$\frac{\begin{pmatrix} 0{,}03+0{,}17+0{,}50+0{,}68+1{,}83+4{,}00+4{,}74+0{,}06+0{,}28+0{,}42+1{,}38+3{,}33 \\ +4{,}00+0{,}09+0{,}17+0{,}89+2{,}52+3{,}11+0{,}01+0{,}42+1{,}67+2{,}16 \\ +0{,}28+1{,}38+1{,}83+0{,}42+0{,}68+0{,}03 \end{pmatrix}}{8} = \frac{37{,}0796}{8} = 4{,}6349$$

$$SQW = SQW_1 + SQW_2 = 4{,}6349 + 0{,}6736 = 5{,}3085$$

$$SQB = SQT - SQW = 12{,}7600 - 5{,}3085 = 7{,}4515$$

9ª Partição: 4 8 2 7 1 11 6 3 9 | 10 5 ($k_1 = 9$ e $k_2 = 2$), ou seja, grupo 1 – objetos 4, 8, 2, 7, 1, 11, 6, 3 e 9, contra grupo 2 – objetos 10 e 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{(d_{10,5}^2)}{k_2} = \frac{0{,}8858}{2} = 0{,}4429$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ grupo\ 1}{k_1} =$$

$$\frac{\begin{pmatrix} d_{4,8}^2+d_{4,2}^2+d_{4,7}^2+d_{4,1}^2+d_{4,11}^2+d_{4,6}^2+d_{4,3}^2+d_{4,9}^2+d_{8,2}^2+d_{8,7}^2+d_{8,1}^2+d_{8,11}^2+d_{8,6}^2 \\ +d_{8,3}^2+d_{8,9}^2+d_{2,7}^2+d_{2,1}^2+d_{2,11}^2+d_{2,6}^2+d_{2,3}^2+d_{2,9}^2+d_{7,1}^2+d_{7,11}^2+d_{7,6}^2+d_{7,3}^2+d_{1,11}^2+d_{1,6}^2 \\ +d_{1,3}^2+d_{1,9}^2+d_{11,6}^2+d_{11,3}^2+d_{11,9}^2+d_{6,3}^2+d_{6,9}^2+d_{3,9}^2 \end{pmatrix}}{k_1} =$$

$$\frac{\begin{pmatrix} 0{,}03+0{,}17+0{,}50+0{,}68+1{,}83+4{,}00+4{,}74+5{,}82+0{,}06+0{,}28+0{,}42 \\ +1{,}38+3{,}33+4{,}00+5{,}00+0{,}09+0{,}17+0{,}89+2{,}52+3{,}11+4{,}00+0{,}01+0{,}42 \\ +1{,}67+2{,}16+2{,}91+0{,}28+1{,}38+1{,}83+2{,}52+0{,}42+0{,}68+1{,}12+0{,}03+0{,}17+0{,}06 \end{pmatrix}}{9} = \frac{58{,}6713}{9} = 6{,}5190$$

$$SQW = SQW_1 + SQW_2 = 6{,}5190 + 0{,}4429 = 6{,}9619$$

$$SQB = SQT - SQW = 12{,}7600 - 6{,}9619 = 5{,}7980$$

10ª Partição: 4 8 2 7 1 11 6 3 9 10 | 5 ($k_1 = 10$ e $k_2 = 1$), ou seja, grupo 1 – objetos 4, 8, 2, 7, 1, 11, 6, 3, 9 e 10, contra grupo 2 – objeto 5:

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{(d^2_{5,5})}{k_2} = \frac{0}{1} = 0$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ grupo\ 2}{k_1} =$$

$$\frac{\begin{pmatrix} d^2_{4,8}+d^2_{4,2}+d^2_{4,7}+d^2_{4,1}+d^2_{4,11}+d^2_{4,6}+d^2_{4,3}+d^2_{4,9}+d^2_{4,10}+d^2_{8,2}+d^2_{8,7} \\ d^2_{8,1}+d^2_{8,11}+d^2_{8,6}+d^2_{8,3}+d^2_{8,9}+d^2_{8,10}+d^2_{2,7}+d^2_{2,1}+d^2_{2,11}+d^2_{2,6}+d^2_{2,3}+d^2_{2,9} \\ +d^2_{2,10}+d^2_{7,1}+d^2_{7,11}+d^2_{7,6}+d^2_{7,3}+d^2_{7,10}+d^2_{1,11}+d^2_{1,6}+d^2_{1,3}+d^2_{1,9}+d^2_{1,10}+d^2_{11,6} \\ +d^2_{11,3}+d^2_{11,9}d^2_{11,10}+d^2_{6,3}+d^2_{6,9}+d^2_{6,10}+d^2_{3,9}+d^2_{3,10}+d^2_{10,9} \end{pmatrix}}{k_1} =$$

$$\frac{\begin{pmatrix} 0{,}03+0{,}17+0{,}50+0{,}68+1{,}83+4{,}00+4{,}74+5{,}82+6{,}40+0{,}06+0{,}28 \\ +0{,}42+1{,}38+3{,}33+4{,}00+5{,}00+5{,}54+0{,}09+0{,}17+0{,}89+2{,}52+3{,}11+4{,}00 \\ +4{,}48+0{,}01+0{,}42+1{,}67+2{,}16+2{,}91+3{,}33+0{,}28+1{,}38+1{,}83+2{,}52+2{,}91 \\ +0{,}42+0{,}68+1{,}12+1{,}38+0{,}03+0{,}17+0{,}28+0{,}06+0{,}12+0{,}01 \end{pmatrix}}{10} = \frac{83{,}1280}{10} = 8{,}3128$$

$$SQW = SQW_1 + SQW_2 = 8{,}3128 + 0 = 8{,}3128$$

$$SQB = SQT - SQW = 12{,}7600 - 8{,}3128 = 4{,}4472$$

Na Tabela 32, observa-se que o valor máximo da SQB é igual a 10,2561 para a partição 4 8 2 7 1 11 | 6 3 9 10 5. Logo:

$$l = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{SQB}{\hat{\sigma}^2_0}\right) = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{10{,}2561}{0{,}51}\right) = 25{,}8527$$

$$c^2_{(\alpha;g/(\pi-2))} \rightarrow c^2_{(0{,}05;9{,}636)} = 16{,}92$$

**Tabela 32.** Partições do grupo 4 8 2 7 1 11 6 3 9 10 5

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| 4 \| 8 2 7 1 11 6 3 9 10 5 | 12,7600 | 0,0000 | 10,4156 | 10,4156 | 2,3444 |
| 4 8 \| 2 7 1 11 6 3 9 10 5 | 12,7600 | 0,0156 | 8,1453 | 8,1609 | 4,5991 |
| 4 8 2 \| 7 1 11 6 3 9 10 5 | 12,7600 | 0,0854 | 6,0861 | 6,1714 | 6,5886 |
| 4 8 2 7 \| 1 11 6 3 9 10 5 | 12,7600 | 0,2803 | 4,3628 | 4,6431 | 8,1169 |
| 4 8 2 7 1 \| 11 6 3 9 10 5 | 12,7600 | 0,4803 | 2,4343 | 2,9145 | 9,8455 |
| 4 8 2 7 1 11 \| 6 3 9 10 5 | 12,7600 | 1,2001 | 1,3038 | 2,5039 | 10,2561 |
| 4 8 2 7 1 11 6 \| 3 9 10 5 | 12,7600 | 2,9081 | 1,5138 | 4,4219 | 8,3381 |
| 4 8 2 7 1 11 6 3 \| 9 10 5 | 12,7600 | 4,5502 | 0,6736 | 5,2238 | 7,5362 |
| 4 8 2 7 1 11 6 3 9 \| 10 5 | 12,7600 | 6,5190 | 0,4498 | 6,9689 | 5,7911 |
| 4 8 2 7 1 11 6 3 9 10 \| 5 | 12,7600 | 8,3128 | 0,0000 | 8,3128 | 4,4472 |

Como $l \geq c^2_{(0,05;9,636)}$, rejeita-se a hipótese de que os dois grupos são idênticos ($H_0$), ou seja, ao nível de significância 0,05, os grupos de médias 4 8 2 7 1 11 e 6 3 9 10 5 diferem entre si.

Assim, os dois grupos serão independentemente submetidos aos passos (a) a (c), fazendo $g = k_1$ e $g = k_2$. O processo em cada subgrupo se encerra ao aceitar $H_0$ no passo (c) ou se cada grupo contiver apenas uma média.

Partição do Grupo 4 8 2 7 1 11 (g = 6):

Inicialmente, obtivemos a Matriz dos Quadrados das Distâncias entre as Médias dos objetos do subgrupo, ou seja:

$D^2_{ii'}$ =

| | P4 | P8 | P2 | P7 | P1 | P11 |
|---|---|---|---|---|---|---|
| P4 | 0,00 | 0,03 | 0,17 | 0,50 | 0,68 | 1,83 |
| P8 | | 0,00 | 0,06 | 0,28 | 0,42 | 1,38 |
| P2 | | | 0,00 | 0,09 | 0,17 | 0,89 |
| P7 | | | | 0,00 | 0,01 | 0,42 |
| P1 | | | | | 0,00 | 0,28 |
| P11 | | | | | | 0,00 |

$$SQT = \frac{Soma\ da\ distâncias\ entre\ todos\ os\ objetos}{g} = \frac{\left(\begin{matrix} d^2_{4,8}+d^2_{4,2}+d^2_{4,7}+d^2_{4,1}+d^2_{4,11}+d^2_{8,2}+d^2_{8,7}+d^2_{8,1}+d^2_{8,11}+ \\ d^2_{2,7}+d^2_{2,1}+d^2_{2,11}+d^2_{7,1}+d^2_{7,11}+d^2_{1,11} \end{matrix}\right)}{6} = \frac{\left(\begin{matrix} 0,06+0,17+0,50+0,68+1,83+0,06+0,28+0,42+1,38 \\ +0,09+0,17+0,89+0,01+0,42+0,28 \end{matrix}\right)}{6} = \frac{7,2007}{6} = 1,2001$$

$$\hat{\sigma}^2_0 = \frac{1}{g+v}\left[SQT + vS^2_{\bar{Y}}\right] \rightarrow \hat{\sigma}^2_0 = \frac{1}{6+471}\left[1,2001 + 471x0,49\right] = 0,49$$

1ª Partição: 4 | 8 2 7 1 11 ($k_1 = 1$ e $k_2 = 5$):

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{\left(d^2_{8,2}+d^2_{8,7}+d^2_{8,1}+d^2_{8,11}+d^2_{2,7}+d^2_{2,1}+d^2_{2,11}+d^2_{7,1}+d^2_{7,11}+d^2_{1,11}\right)}{k_2} = \frac{(0,06+0,28+0,42+1,38+0,09+0,17+0,89+0,01+0,42+0,28)}{5} = \frac{3,9931}{5} = 0,7986$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ dentre\ os\ objetos\ do\ grupo\ 1}{k_1} = \frac{d^2_{4,4}}{1} = \frac{0}{1} = 0$$

$$SQW = SQW_1 + SQW_2 = 0,000 + 0,7986 = 0,7986$$

$$SQB = SQT - SQW = 1,2001 - 0,7986 = 0,4015$$

2ª Partição: 4 8 | 2 7 1 11 ($k_1 = 2$ e $k_2 = 4$):

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} =$$

$$\frac{\left(d^2_{2,7}+d^2_{2,1}+d^2_{2,11}+d^2_{7,1}+d^2_{7,11}+d^2_{1,11}\right)}{k_2} = \frac{(0{,}09+0{,}17+0{,}89+0{,}01+0{,}42+0{,}28)}{4} = \frac{1{,}8547}{4} = 0{,}4637$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} = \frac{d^2_{1,8}}{k_1} = \frac{0{,}03}{2} = 0{,}0156$$

$$SQW = SQW_1 + SQW_2 = 0{,}0156 + 0{,}4637 = 0{,}4792$$

$$SQB = SQT - SQW = 1{,}2001 - 0{,}4792 = 0{,}7209$$

3ª Partição: 4 8 2 | 7 1 11 ($k_1 = 3$ e $k_2 = 3$):

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{\left(d^2_{7,1}+d^2_{7,11}+d^2_{1,11}\right)}{k_2} =$$

$$\frac{(0{,}01+0{,}42+0{,}28)}{3} = \frac{0{,}7128}{3} = 0{,}2376$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} = \frac{\left(d^2_{4,8}+d^2_{4,2}+d^2_{8,2}\right)}{k_1} =$$

$$\frac{(0{,}03+0{,}17+0{,}06)}{3} = \frac{0{,}2561}{3} = 0{,}0854$$

$$SQW = SQW_1 + SQW_2 = 0{,}0854 + 0{,}2376 = 0{,}3230$$

$$SQB = SQT - SQW = 1{,}2001 - 0{,}3230 = 0{,}8772$$

4ª Partição: 4 8 2 7 | 1 11 ($k_1 = 4$ e $k_2 = 2$):

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{\left(d^2_{1,11}\right)}{2} = \frac{0{,}2803}{2} = 0{,}1401$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} =$$

$$\frac{\left(d^2_{4,8}+d^2_{4,2}+d^2_{4,7}+d^2_{8,2}+d^2_{8,7}+d^2_{2,7}\right)}{k_1} = \frac{(0{,}03+0{,}17+0{,}50+0{,}06+0{,}28+0{,}09)}{4} = \frac{1{,}1211}{4} = 0{,}2803$$

$$SQW = SQW_1 + SQW_2 = 0{,}2803 + 0{,}1401 = 0{,}4204$$

$$SQB = SQT - SQW = 1{,}2001 - 0{,}4204 = 0{,}7797$$

5ª Partição: 4 8 2 7 1 | 11 ($k_1 = 5$ e $k_2 = 1$):

$$SQW_2 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 2}{k_2} = \frac{d^2_{11,11}}{1} = \frac{0}{1} = 0$$

$$SQW_1 = \frac{Soma\ das\ distâncias\ quadradas\ entre\ os\ objetos\ do\ grupo\ 1}{k_1} =$$

$$\frac{\left(d_{4,8}^2+d_{4,2}^2+d_{4,7}^2+d_{4,1}^2+d_{8,2}^2+d_{8,7}^2+d_{8,1}^2+d_{2,7}^2+d_{2,1}^2+d_{7,1}^2\right)}{k_1} =$$

$$\frac{\begin{pmatrix}0,03+0,17+050+0,68+0,06+0,28+0,42+1,38+0,09+0,17\\+0,89+0,01+0,42+0,28\end{pmatrix}}{5} = \frac{2,4014}{5} = 0,4803$$

$$SQW = SQW_1 + SQW_2 = 0,4803 + 0 = 0,4803$$

$$SQB = SQT - SQW = 1,2001 - 0,4803 = 0,7198$$

Na Tabela 33, observa-se que o valor máximo da SQB é igual a 0,8772 para a partição 4 8 2 | 7 1 11. Logo:

$$l = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{SQB}{\hat{\sigma}_0^2}\right) = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{0,8772}{0,49}\right) = 5,89$$

$$c^2_{(\alpha;g/(\pi-2))} \rightarrow c^2_{(0,05;5,256)} = 11,07$$

Como $l < c^2_{(\alpha;g/(\pi-2))}$, aceita-se a hipótese de que os dois grupos são idênticos ($H_0$), ou seja, ao nível de significância 0,05, os grupos de médias 4 8 2 e 7 1 11 não diferem entre si.

Assim, encerra-se o processo ao aceitar $H_0$ nesse passo.

**Tabela 33.** Partições do grupo 4 8 2 7 1 11

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| 4 \| 8 2 7 1 11 | 1,2001 | 0,0000 | 0,7986 | 0,7986 | 0,4015 |
| 4 8 \| 2 7 1 11 | 1,2001 | 0,0156 | 0,7050 | 0,7206 | 0,4795 |
| 4 8 2 \| 7 1 11 | 1,2001 | 0,0854 | 0,2376 | 0,3230 | 0,8772 |
| 4 8 2 7 \| 1 11 | 1,2001 | 0,2803 | 0,1401 | 0,4204 | 0,7797 |
| 4 8 2 7 1 \| 11 | 1,2001 | 0,4805 | 0,0000 | 0,4805 | 0,7198 |

Partição do grupo 4 8 2:

Na Tabela 34, observa-se que o valor máximo da SQB é igual a 0,0577 para a partição 4 | 8 2. Logo:

$$l = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{SQB}{\hat{\sigma}_0^2}\right) = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{0,0698}{0,49}\right) = 0,20$$

$$c^2_{(\alpha;g/(\pi-2))} \rightarrow c^2_{(0,05;2,628)} = 5,99$$

Como $l < c^2_{(\alpha;g/(\pi-2))}$, aceita-se a hipótese de que os dois grupos são idênticos ($H_0$), ou seja, ao nível de significância 0,05, os grupos de médias 4 e 8 2 são idênticos entre si. Finalizando-se o processo de agrupamento.

**Tabela 34.** Partições do grupo 4 8 2

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| 4 \| 8 2 | 0,0854 | 0,0000 | 0,0277 | 0,0277 | 0,0577 |
| 4 8 \| 2 | 0,0854 | 0,0156 | 0,0000 | 0,0156 | 0,0698 |

Na Tabela 35, observa-se que o valor máximo da SQB é igual a 0,2307 para a partição 4 | 8 2. Logo:

$$l = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{SQB}{\hat{\sigma}_0^2}\right) = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{0{,}0698}{0{,}49}\right) = 0{,}20$$

$$c^2_{(\alpha;g/(\pi-2))} \rightarrow c^2_{(0{,}05;2{,}628)} = 5{,}99$$

Como $l < c^2_{(\alpha;g/(\pi-2))}$, aceita-se a hipótese de que os dois grupos são idênticos ($H_0$), ou seja, ao nível de significância 0,05, os grupos de médias 7 e 1 11 são idênticos entre si e, por conseguinte, finaliza-se o processo de agrupamento.

**Tabela 35.** Partições do grupo 7 1 11

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| 7 \| 1 11 | 0,2376 | 0 | 0,1401 | 0,1401 | 0,0975 |
| 7 1 \| 11 | 0,2376 | 0,0069 | 0 | 0,0069 | 0,2307 |

Partição do Grupo 6 3 9 10 5:

Na Tabela 36, observa-se que o valor máximo da SQB é igual a 1,1351 para a partição 6 3 9 10 | 5. Logo:

$$l = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{SQB}{\hat{\sigma}_0^2}\right) = \left(\frac{\pi}{2(\pi-2)}\right)\left(\frac{1{,}1351}{0{,}49}\right) = 3{,}19$$

$$c^2_{(\alpha;g/(\pi-2))} \rightarrow c^2_{(0{,}05;4.38)} = 9{,}49$$

**Tabela 36.** Partições do grupo 6 3 9 10 5

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| 6 \| 3 9 10 5 | 1,3038 | 0,0000 | 0,9689 | 0,9689 | 0,3349 |
| 6 3 \| 9 10 5 | 1,3038 | 0,0156 | 0,6736 | 0,6892 | 0,6146 |
| 6 3 9 \| 10 5 | 1,3038 | 0,0854 | 0,4429 | 0,5283 | 0,7755 |
| 6 3 9 10 \| 5 | 1,3038 | 0,1687 | 0,0000 | 0,1687 | 1,1351 |

Como $l < c^2_{(\alpha;g/(\pi-2))}$, aceita-se a hipótese de que os dois grupos são idênticos ($H_0$), ou seja, ao nível de significância 0,05, não há evidências para que o grupo 6 3 9 10 5 seja dividido. Finalizando-se o processo de agrupamento.

Dessa forma, chegamos aos grupos conforme o diagrama de árvore ilustrado na Figura 16 e na Tabela 37.

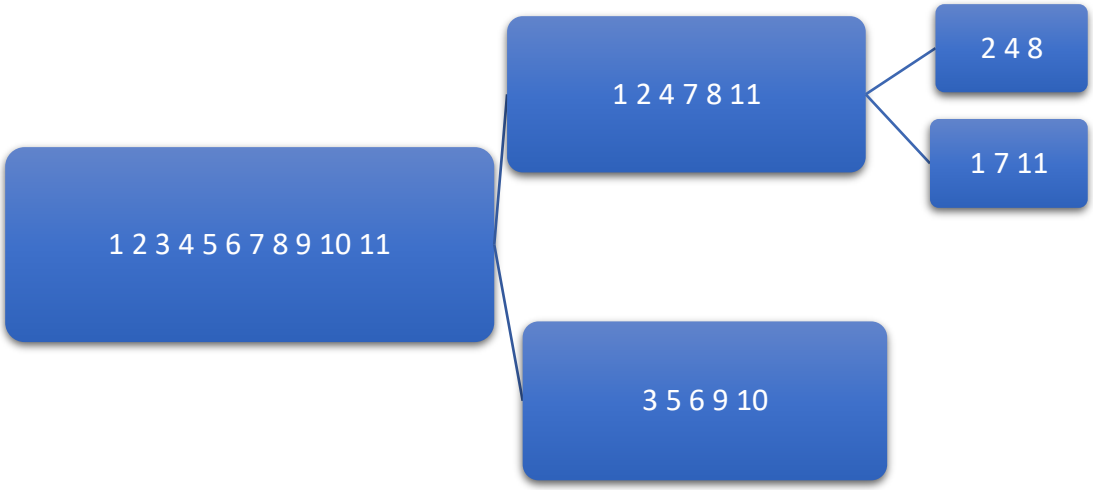



**Figura 16.** Grupos resultantes das partições significativas, por meio do método de Edwards e Cavalli-Sforza (1965).

**Tabela 37.** Médias dos grupos conforme a partição pelo método de Edwards e Cavalli-Sforza (1965)

| Grupo | Objetos | Média do Grupo |
|---|---|---|
| 1 | 2 4 8 | 1,31 |
| 2 | 1 7 11 | 2,08 |
| 3 | 3 5 6 9 10 | 3,64 |

## Passo a Passo no R: Método de Edwards e Cavalli-Sforza

O Método de Edwards e Cavalli-Sforza é uma técnica de agrupamento que utiliza a Análise de Variância (ANOVA) para identificar a melhor forma de dividir as parcelas em grupos, maximizando a variância entre eles. Esse método é especialmente útil para entender a estrutura de agrupamento em dados de abundância.

### 1. Análise de variância (Manual)

Para aplicar o método de Edwards e Cavalli-Sforza, precisamos calcular os componentes da ANOVA manualmente a partir da matriz X. Isso nos ajudará a entender a variabilidade dos dados e a identificar as melhores partições.

```
# Número de parcelas e espécies
I <- ncol(X)   # número de parcelas
J <- nrow(X)   # número de espécies

# Total geral de abundância
T_geral <- sum(as.matrix(X))

# Soma de Quadrados Total
SQTotal <- sum(as.matrix(X)^2) - T_geral^2/(I*J)

# Soma de Quadrados Tratamentos (Parcelas)
Totais_Parcelas <- colSums(X)
```

```
SQTrat <- sum(Totais_Parcelas^2)/J - T_geral^2/(I*J)

# Soma de Quadrados Resíduo
SQRes <- SQTotal - SQTrat

# Graus de liberdade
GLTrat <- I - 1
GLTotal <- T_geral - 1      # Conforme o método do livro
GLRes <- GLTotal - GLTrat

# Quadrados Médios
QMTrat <- SQTrat/GLTrat
QMRes <- SQRes/GLRes

# Estatística F
Fcalc <- QMTrat/QMRes

# p-valor
pvalor <- pf(Fcalc, df1 = GLTrat, df2 = GLRes, lower.tail = FALSE)

# Quadro da ANOVA
anova_manual <- data.frame(
  Fonte_de_Variacao = c("Tratamentos", "Resíduo", "Total"),
  GL = c(GLTrat, GLRes, GLTotal),
  SQ = c(round(SQTrat,2), round(SQRes,2), round(SQTotal,2)),
  QM = c(round(QMTrat,2), round(QMRes,2), NA),
  F = c(round(Fcalc,2), NA, NA),
  p = c(round(pvalor,4), NA, NA)
)

print(anova_manual)
```

| Fonte de Variação | GL | SQ | QM | F | p |
|---|---|---|---|---|---|
| Tratamentos | 10 | 216.92 | 21.69 | 2.60 | 0.0045 |
| Resíduo | 471 | 3934.71 | 8.35 | — | — |
| Total | 481 | 4151.63 | — | — | — |

## 2. Cálculo e ordenação das médias

Com base nos resultados da ANOVA, calculamos as médias de abundância por parcela e a variância da média (Sy2), que será utilizada para avaliar a estabilidade dos grupos.

```
# Médias das parcelas
medias <- colSums(X) / J

# Ordenar as médias em ordem crescente para o particionamento
medias_ordenadas <- sort(medias)
valores_ordenados <- as.numeric(medias_ordenadas)

# Estimativa da variância da média
# n é o número de observações por tratamento (espécies)
```

```
n <- J
Sy2 <- QMRes / n

# Exibir a Variância da média (Sy2)
print(Sy2)
```

0.4914083

### 3. Partições do Grupo Global

O objetivo, aqui, é encontrar a melhor partição do conjunto total de parcelas. Testamos todas as divisões possíveis entre as parcelas ordenadas e escolhemos aquela que resulta na maior Soma de Quadrados Entre Grupos (**SQB**).

```
# Soma de Quadrados Total das Médias (SQT)
SQT <- sum((medias_ordenadas - mean(medias_ordenadas))^2)

# Função para calcular todas as partições possíveis
calcular_particoes <- function(medias_grupo){
  k <- length(medias_grupo)
  resultados <- data.frame()

  for(i in 1:(k-1)){
    grupo1 <- medias_grupo[1:i]
    grupo2 <- medias_grupo[(i+1):k]

    SQW1 <- sum((grupo1 - mean(grupo1))^2)
    SQW2 <- sum((grupo2 - mean(grupo2))^2)
    SQW  <- SQW1 + SQW2
    SQB  <- SQT - SQW

    resultados <- rbind(resultados, data.frame(
      Particao = paste(paste(names(grupo1), collapse = " "), "|", paste(names(grupo2), collapse
= " ")),
      SQT = SQT, SQW1 = SQW1, SQW2 = SQW2, SQW = SQW, SQB = SQB
    ))
  }
  return(resultados)
}

# Gerando a tabela de partições para o grupo global
tabela_global <- calcular_particoes(medias_ordenadas)
tabela_global[,2:6] <- round(tabela_global[,2:6], 4)
print(tabela_global)
```

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| P4 \| P8 P2 P7 P1 P11 P6 P3 P9 P10 P5 | 12.76 | 0.0000 | 10.4156 | 10.4156 | 2.3444 |
| P4 P8 \| P2 P7 P1 P11 P6 P3 P9 P10 P5 | 12.76 | 0.0156 | 8.1453 | 8.1609 | 4.5991 |
| P4 P8 P2 \| P7 P1 P11 P6 P3 P9 P10 P5 | 12.76 | 0.0854 | 6.0861 | 6.1714 | 6.5886 |
| P4 P8 P2 P7 \| P1 P11 P6 P3 P9 P10 P5 | 12.76 | 0.2803 | 4.3628 | 4.6431 | 8.1169 |
| P4 P8 P2 P7 P1 \| P11 P6 P3 P9 P10 P5 | 12.76 | 0.4803 | 2.4343 | 2.9145 | 9.8455 |
| P4 P8 P2 P7 P1 P11 \| P6 P3 P9 P10 P5 | 12.76 | 1.2001 | 1.3038 | 2.5039 | 10.2561 |

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| P4 P8 P2 P7 P1 P11 P6 \| P3 P9 P10 P5 | 12.76 | 2.9323 | 0.9689 | 3.9011 | 8.8589 |
| P4 P8 P2 P7 P1 P11 P6 P3 \| P9 P10 P5 | 12.76 | 4.6349 | 0.6736 | 5.3085 | 7.4515 |
| P4 P8 P2 P7 P1 P11 P6 P3 P9 \| P10 P5 | 12.76 | 6.5190 | 0.4429 | 6.9619 | 5.7980 |
| P4 P8 P2 P7 P1 P11 P6 P3 P9 P10 \| P5 | 12.76 | 8.3128 | 0.0000 | 8.3128 | 4.4472 |

```
# Identificando a melhor partição (Maior SQB)
melhor_global <- tabela_global[which.max(tabela_global$SQB),]
print(melhor_global)
```

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| P4 P8 P2 P7 P1 P11 \| P6 P3 P9 P10 P5 | 12.76 | 1.2001 | 1.3038 | 2.5039 | 10.2561 |

## 4. Particionamento Recursivo dos Subgrupos

O processo continua dividindo os subgrupos formados na etapa anterior. Abaixo, aplicamos a mesma lógica para os subgrupos resultantes da primeira divisão.

```
# Função auxiliar para particionar subgrupos específicos
particionar <- function(grupo){
  SQT_g <- sum((grupo - mean(grupo))^2)
  k <- length(grupo)
  tabela <- data.frame()

  for(i in 1:(k-1)){
    g1 <- grupo[1:i]; g2 <- grupo[(i+1):k]
    SQW1 <- sum((g1-mean(g1))^2); SQW2 <- sum((g2-mean(g2))^2)
    SQW <- SQW1+SQW2; SQB <- SQT_g-SQW

    tabela <- rbind(tabela, data.frame(
      Particao=paste(paste(names(g1),collapse=" "), "|", paste(names(g2),collapse=" ")),
      SQT=SQT_g, SQW1=SQW1, SQW2=SQW2, SQW=SQW, SQB=SQB
    ))
  }
  tabela[,2:6] <- round(tabela[,2:6], 4)
  return(tabela)
}

# Exemplo de partição do Grupo 1 (Parcelas com menores médias)
grupo1 <- medias_ordenadas[c("P4","P8","P2","P7","P1","P11")]
tabela_g1 <- particionar(grupo1)
print(tabela_g1)
```

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| P4 \| P8 P2 P7 P1 P11 | 1.2001 | 0.0000 | 0.7986 | 0.7986 | 0.4015 |
| P4 P8 \| P2 P7 P1 P11 | 1.2001 | 0.0156 | 0.4637 | 0.4792 | 0.7209 |
| P4 P8 P2 \| P7 P1 P11 | 1.2001 | 0.0854 | 0.2376 | 0.3230 | 0.8772 |
| P4 P8 P2 P7 \| P1 P11 | 1.2001 | 0.2803 | 0.1401 | 0.4204 | 0.7797 |
| P4 P8 P2 P7 P1 \| P11 | 1.2001 | 0.4803 | 0.0000 | 0.4803 | 0.7198 |

```
# Exemplo de partição do Grupo 2 (Parcelas com maiores médias)
grupo2 <- medias_ordenadas[c("P6","P3","P9","P10","P5")]
tabela_g2 <- particionar(grupo2)
print(tabela_g2)
```

| Partição | SQT | SQW1 | SQW2 | SQW | SQB |
|---|---|---|---|---|---|
| P6 \| P3 P9 P10 P5 | 1.3038 | 0.0000 | 0.9689 | 0.9689 | 0.3349 |
| P6 P3 \| P9 P10 P5 | 1.3038 | 0.0156 | 0.6736 | 0.6892 | 0.6146 |
| P6 P3 P9 \| P10 P5 | 1.3038 | 0.0854 | 0.4429 | 0.5283 | 0.7755 |
| P6 P3 P9 P10 \| P5 | 1.3038 | 0.1687 | 0.0000 | 0.1687 | 1.1351 |

## 5. Grupos Finais e Diagrama de Árvore

Ao final do processo, consolidamos os grupos e visualizamos a estrutura de partição por meio de um diagrama (Figura 17).

```
# Consolidação dos grupos finais
grupo_A <- c("P2","P4","P8")
grupo_B <- c("P1","P7","P11")
grupo_C <- c("P3","P5","P6","P9","P10")

resultado_final <- data.frame(
  Grupo = c(1, 2, 3),
  Parcelas = c(paste(grupo_A, collapse=" "), paste(grupo_B, collapse=" "), paste(grupo_C, collapse=" ")),
  Media = c(mean(medias[grupo_A]), mean(medias[grupo_B]), mean(medias[grupo_C]))
)

print(round(resultado_final$Media, 2))
```

```
1.31 2.08 3.64
```

```
# Diagrama de Árvore (Visualização da Hierarquia de Partição)
plot(0, type="n", xlim=c(0,100), ylim=c(0,100), axes=FALSE, xlab="", ylab="",
     main="Diagrama de Partição - Edwards e Cavalli-Sforza")

# Nó inicial
rect(20,80,80,92, col="lightblue")
text(50,86, "P1 a P11", cex=1.1)

# Ramos e Subgrupos
segments(50,80, 35,60, lwd=2); segments(50,80, 65,60, lwd=2)
rect(10,45,50,60, col="lightblue"); text(30,52, "P1, P2, P4, P7, P8, P11")
rect(60,45,90,60, col="lightblue"); text(75,52, "P3, P5, P6, P9, P10")

# Divisão final do grupo esquerdo
segments(30,45, 20,25, lwd=2); segments(30,45, 40,25, lwd=2)
rect(8,10,28,25, col="lightblue"); text(18,17, "P2, P4, P8")
rect(32,10,52,25, col="lightblue"); text(42,17, "P1, P7, P11")
```

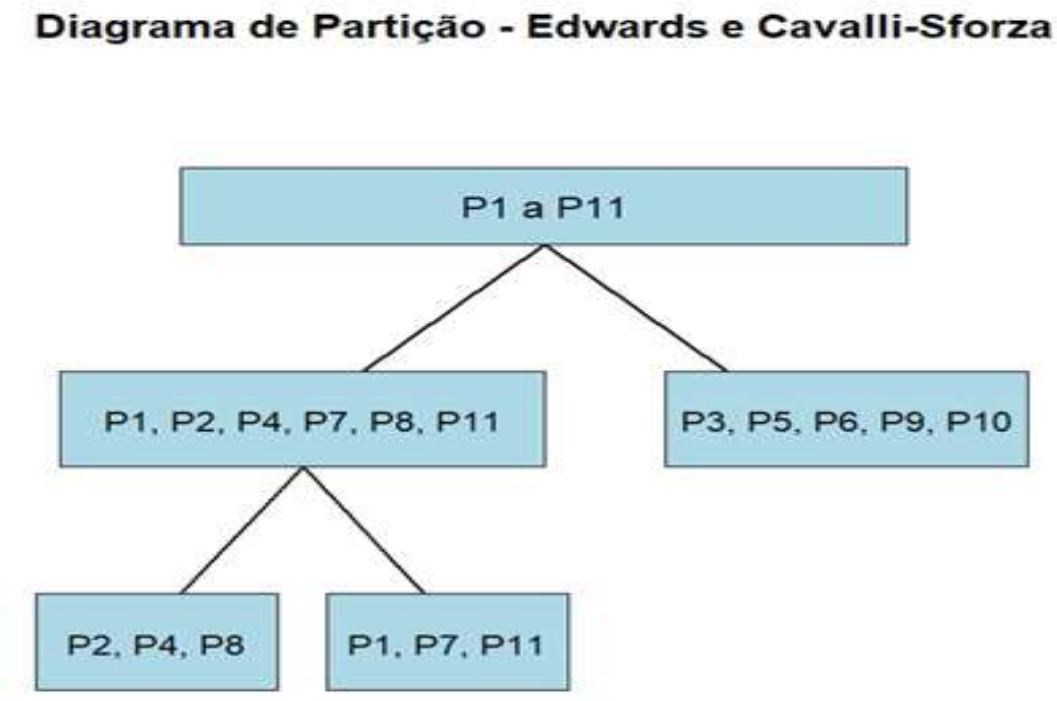



**Figura 17.** Grupos resultantes das partições significativas, com o método de Edwards e Cavalli-Sforza (1965), por meio do R.

## 2.2. Técnicas de Partição ou Otimização

Essas técnicas diferem das técnicas de hierarquização por permitirem uma realocação dos elementos dos agrupamentos, objetivando obter uma otimização do critério de agrupamento empregado. Elas são caracterizadas pela partição do conjunto **E** de elementos em agrupamentos, o que traz como vantagem permitir que uma partição inicial seja corrigida a cada interação do processo.

De um grupo inicial **I** de **n** indivíduos, formam-se **k** subgrupos, os quais são identificados por $\mathbf{I_i}$ (i = 1, 2, ... , k) e apresentam as seguintes propriedades:

a) $\mathbf{k \leq n}$;

b) $\mathbf{I_i \neq \Phi}$ (não existe subgrupo vazio);

c) $\mathbf{I_i \cap I_j = \Phi}$, para $\mathbf{i \neq h}$ (os subgrupos são mutuamente exclusivos); e

d) $\mathbf{I_1 \cup I_2 \cup ... \cup I_k = I}$ (a reunião de todos os subgrupos restabelece o grupo inicial).

As diferenças entre as técnicas de otimização decorrem dos métodos empregados para uma partição inicial e do critério de agrupamento utilizado para a otimização. Assim, um grande número dessas técnicas tem como critério de agrupamento a minimização do traço da matriz de dispersão dentro dos grupos, a minimização do determinante, entre outros. No entanto, um

método amplamente utilizado no melhoramento de plantas e com boa frequência na literatura é o método de Tocher, citado por Rao (1952).

Os métodos de otimização buscam a partição do conjunto de indivíduos que otimiza (maximiza ou minimiza) alguma medida predefinida. No processo de partição ótima, tomam-se, inicialmente, alguns grupos centrais de forma arbitrária e os indivíduos restantes são alocados aos grupos mais próximos (Manly; Navarro Alberto, 2019). Novos grupos são formados e grupos próximos são fundidos, movendo-se indivíduos entre grupos, iterativamente, até a otimização do critério para um número predeterminado de grupos.

Dentre os métodos de otimização, destacamos o de Tocher e o k-médias.

### a. Método de Tocher

Esse é um método de otimização proposto por Rao (1952), que adota o critério de que a média das medidas de dissimilaridade intragrupo (distância euclidiana, distância euclidiana média, $D^2$ de Mahalanobis etc.) deve ser menor do que as distâncias médias intergrupos.

Inicialmente, a partir de uma matriz de distância, é identificado o par de indivíduos mais similar, o qual constituu o grupo inicial. Em seguida, é avaliada a possibilidade de inclusão de novos indivíduos no grupo inicial, adotando-se o critério anteriormente citado.

A inclusão de um novo indivíduo provoca um aumento no valor médio da distância intragrupo. Assim, decide-se pela inclusão de um determinado indivíduo em um grupo se o acréscimo resultante no valor médio da distância intragrupo não ultrapassar um nível máximo predefinido, que pode ser estabelecido de forma arbitrária, ou adota-se o valor máximo da medida de dissimilaridade (28,7) encontrado no conjunto das menores distâncias que envolvem cada indivíduo (Tabela 38).

### Aplicação

Passo 1: Determinação do limite de acréscimo ($\alpha$) na média da distância euclidiana intragrupo para a formação ou inclusão de um novo elemento no grupo (Tabela 38). Assim, $\alpha = 28,7$.

**Tabela 38.** Distância euclidiana mínima entre as parcelas

| Parcela | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D mínima | 11,3 | 12,0 | 17,9 | 7,70 | 28,7 | 17,3 | 8,8 | 7,7 | 19,2 | 17,5 | 13,3 |

Passo 2: O grupo I é formado inicialmente pelas parcelas 4 e 8, cuja medida de dissimilaridade é a menor distância. A distância entre elas é de 7,7, menor que o mínimo estabelecido ($\alpha$ = 28,7).

Cálculo das medidas de dissimilaridades entre o grupo (4,8) e as demais parcelas:

$$d_{(4,8)1} = 12{,}5 + 11{,}3 = 23{,}8$$

$$d_{(4,8)2} = 12{,}0 + 16{,}2 = 28{,}2$$

$$d_{(4,8)3} = 29{,}6 + 27{,}9 = 57{,}5$$

$$d_{(4,8)5} = 32{,}3 + 35{,}6 = 67{,}9$$

$$d_{(4,8)6} = 17{,}3 + 19{,}0 = 36{,}3$$

$$d_{(4,8)7} = 14{,}7 + 15{,}5 = 30{,}2$$

$$d_{(4,8)9} = 28{,}3 + 26{,}7 = 55{,}0$$

$$d_{(4,8)10} = 21{,}5 + 21{,}2 = 42{,}7$$

$$d_{(4,8)11} = 13{,}5 + 13{,}3 = 26{,}8$$

A parcela 1 é mais similar ao grupo (4,8), pois $d_{(4,8)1} = 23{,}8$.

Cálculo do acréscimo médio no valor de D no grupo inicial, pela inclusão da parcela 1: $\frac{D_{(grupo)i}}{n} = \frac{D_{(4,8)1}}{2} = \frac{23{,}8}{2} = 11{,}9 < \alpha = 28{,}7$, logo, pode-se incluir a parcela 1.

Passo 3: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,4,8):

$$d_{(1,4,8)2} = 28{,}1 + 17{,}8 = 45{,}9$$

$$d_{(1,4,8)3} = 57{,}5 + 23{,}0 = 80{,}5$$

$$d_{(1,4,8)5} = 67{,}9 + 33{,}9 = 101{,}8$$

$$d_{(1,4,8)6} = 36{,}3 + 17{,}3 = 53{,}6$$

$$d_{(1,4,8)7} = 30{,}2 + 18{,}5 = 48{,}7$$

$$d_{(1,4,8)9} = 55{,}0 + 24{,}8 = 79{,}8$$

$$d_{(1,4,8)10} = 42{,}7 + 20{,}0 = 62{,}7$$

$$d_{(1,4,8)11} = 26{,}8 + 14{,}7 = 41{,}5$$

A parcela 11 é mais similar ao grupo (1,4,8), pois $d_{(1,4,8)11} = 41{,}5$.

Cálculo do acréscimo médio no valor de D no grupo (1,4,8) pela inclusão da parcela 11: $\frac{D_{(grupo)i}}{n} = \frac{D_{(1,4,8)11}}{3} = \frac{41{,}5}{3} = 13{,}83 \quad < \alpha = 28{,}7$, logo, incluir-se a parcela 11.

Passo 4: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,4,8,11):

$$d_{(1,4,8,11)2} = 45{,}9 + 18{,}7 = 64{,}6$$

$$d_{(1,4,8,11)3} = 80{,}5 + 24{,}2 = 104{,}7$$

$$d_{(1,4,8,11)5} = 101{,}7 + 34{,}7 = 136{,}4$$

$$d_{(1,4,8,11)6} = 53{,}6 + 19{,}4 = 73{,}0$$

$$d_{(1,4,8,11)7} = 48{,}7 + 8{,}8 = 57{,}5$$

$$d_{(1,4,8,11)9} = 79{,}8 + 19{,}2 = 99{,}0$$

$$d_{(1,4,8,11)10} = 62{,}7 + 17{,}5 = 80{,}2$$

A parcela 7 é mais similar ao grupo (1,4,8,11), pois $d_{(1,4,8,11)7} = 57{,}5$.

Cálculo do acréscimo médio no valor de D no grupo (1,4,8,11) pela inclusão da parcela 7:

$\frac{D_{(grupo)i}}{n} = \frac{D_{(1,4,8,11)7}}{4} = \frac{57{,}5}{4} = 14{,}3 < \alpha = 28{,}7$, logo, é permitida a inclusão da parcela 7.

Passo 5: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,4,7,8,11):

$$d_{(1,4,7,8,11)2} = 64{,}6 + 18{,}3 = 82{,}9$$

$$d_{(1,4,7,8,11)3} = 104{,}7 + 30{,}9 = 135{,}6$$

$$d_{(1,4,7,8,11)5} = 136{,}4 + 38{,}6 = 175{,}0$$

$$d_{(1,4,7,8,11)6} = 73{,}0 + 24{,}6 = 97{,}6$$

$$d_{(1,4,7,8,11)9} = 99{,}0 + 22{,}4 = 121{,}4$$

$$d_{(1,4,7,8,11)10} = 80{,}2 + 24{,}4 = 104{,}6$$

A parcela 2 é mais similar ao grupo (1,4,7,8,11), pois $d_{(1,4,7,8,11)2} = 82{,}9$.

Cálculo do acréscimo médio no valor de D no grupo (1,4,7,8,11) pela inclusão da parcela 2:

$\frac{D_{(grupo)i}}{n} = \frac{D_{(1,4,7,8,11)2}}{5} = \frac{82{,}9}{5} = 16{,}58 < \alpha = 28{,}7$, logo, é incluída a parcela 2.

Passo 6: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,2,4,7,8,11):

$$d_{(1,2,4,7,8,11)3} = 135{,}6 + 31{,}7 = 167{,}3$$

$$d_{(1,2,4,7,8,11)5} = 175{,}0 + 34{,}4 = 209{,}4$$

$$d_{(1,2,4,7,8,11)6} = 97{,}6 + 21{,}8 = 119{,}4$$

$$d_{(1,2,4,7,8,11)9} = 121{,}4 + 29{,}4 = 150{,}8$$

$$d_{(1,2,4,7,8,11)10} = 104{,}6 + 25{,}6 = 130{,}2$$

A parcela 6 é mais similar ao grupo (1,2,4,7,8,11), pois $d_{(1,2,4,7,8,11)6} = 119{,}4$.

Cálculo do acréscimo médio no valor de D no grupo (1,2,4,7,8,11) pela inclusão da parcela 6:

$\frac{D_{(grupo)i}}{n} = \frac{D_{(1,2,4,7,8,11)2}}{6} = \frac{119{,}4}{6} = 19{,}9 < \alpha = 28{,}7$, logo, é incluída a parcela 6.

Passo 7: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,2,4,6,7,8,11):

$$d_{(1,2,4,6,7,8,11)3} = 167{,}3 + 27{,}0 = 194{,}3$$

$$d_{(1,2,4,6,7,8,11)5} = 209{,}4 + 28{,}7 = 238{,}1$$

$$d_{(1,2,4,6,7,8,11)9} = 150{,}8 + 27{,}6 = 178{,}4$$

$$d_{(1,2,4,6,7,8,11)10} = 130{,}2 + 21{,}5 = 151{,}7$$

A parcela 10 é mais similar ao grupo (1,2,4,6,7,8,11), pois $d_{(1,2,4,6,7,8,11)10} = 151{,}7$.

Cálculo do acréscimo médio no valor de D no grupo (1,2,4,6,7,8,11) pela inclusão da parcela 10:

$\frac{D_{(grupo)i}}{n} = \frac{D_{(1,2,4,6,7,8,11)10}}{7} = \frac{151{,}7}{7} = 21{,}67 < \alpha = 28{,}7$, logo, é permitida a inclusão da parcela 10.

Passo 8: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,2,4,6,7,8,10,11):

$$d_{(1,2,4,6,7,8,10,11)3} = 194{,}3 + 17{,}9 = 212{,}2$$

$$d_{(1,2,4,6,7,8,10,11)5} = 238{,}1 + 34{,}8 = 272{,}9$$

$$d_{(1,2,4,6,7,8,10,11)9} = 178{,}4 + 22{,}8 = 201{,}2$$

A parcela 9 é mais similar ao grupo (1,2,4,6,7,8,10,11), pois $d_{(1,2,4,6,7,8,10,11)9} = 201{,}2$.

Cálculo do acréscimo médio no valor de D no grupo (1,2,4,6,7,8,10,11) pela inclusão da parcela 9:

$\frac{D(grupo)i}{n} = \frac{D_{(1,2,4,6,7,8,10,11)9}}{8} = \frac{201,2}{8} = 25,1$ < α = 28,7, logo, é permitida a inclusão da parcela 9.

Passo 9: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,2,4,6,7,8,9,10,11):

$$d_{(1,2,4,6,7,8,9,10,11)3} = 212,2 + 22,4 = 234,6$$

$$d_{(1,2,4,6,7,8,9,10,11)5} = 272,9 + 41,4 = 314,3$$

A parcela 3 é mais similar ao grupo (1,2,4,6,7,8,9,10,11), pois $d_{(1,2,4,6,7,8,9,10,11)3}$ = 234,6.

Cálculo do acréscimo médio no valor de D no grupo (1,2,4,6,7,8,9,10,11), pela inclusão da parcela 3:

$\frac{D(grupo)i}{n} = \frac{D_{(1,2,4,6,7,8,9,10,11)3}}{9} = \frac{234,6}{9} = 26,07$ < α = 28,7, logo, é permitida a inclusão da parcela 3.

Passo 10: Cálculo das novas distâncias das parcelas não incluídas em relação ao grupo em formação (1,2,3,4,6,7,8,9,10,11):

$$d_{(1,2,3,4,6,7,8,9,10,11)5} = 314,3 + 41,8 = 356,1$$

Cálculo do acréscimo médio no valor de D no grupo (1,2,3,4,6,7,8,9,10,11) pela inclusão da parcela 5:

$\frac{D(grupo)i}{n} = \frac{D_{(1,2,3,4,6,7,8,9,10,11)10}}{10} = \frac{356,1}{10} = 35,61$ > α = 28,7, logo, não é permitida a inclusão da parcela 5. Implicando que a parcela 5 é o grupo II.

Os resultados estão resumidos na Tabela 39 e podem ser representados por meio de um dendrograma, conforme apresentados nos métodos de agrupamento precedentes.

**Tabela 39.** Resumo da análise de agrupamento obtido pelo emprego do método Tocher, mostrando a sequência das fusões das parcelas

| Distância | Objetos | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 7,7 | P4 | P8 | | | | | | | | | |
| 11,9 | P1 | P4 | P8 | | | | | | | | |
| 13,8 | P1 | P4 | P8 | P11 | | | | | | | |
| 14,4 | P1 | P4 | P8 | P11 | P7 | | | | | | |
| 16,6 | P1 | P4 | P8 | P11 | P7 | P2 | | | | | |
| 19,9 | P1 | P4 | P8 | P11 | P7 | P2 | P6 | | | | |
| 21,8 | P1 | P4 | P8 | P11 | P7 | P2 | P6 | P10 | | | |
| 25,2 | P1 | P4 | P8 | P11 | P7 | P2 | P6 | P10 | P9 | | |
| 26,1 | P1 | P4 | P8 | P11 | P7 | P2 | P6 | P10 | P9 | P3 | |
| **** | P5 | | | | | | | | | | |

### Passo a passo no R: Método de Tocher

### 1. Instalando e Carregando o Pacote Necessário

Para utilizar o Método de Tocher, precisamos do pacote *biotools*. Se você ainda não o tem instalado, pode instalá-lo e, depois, carregá-lo.

```
# Instalar o pacote 'biotools' (execute apenas se necessário)
# install.packages("biotools")

# Carregar o pacote 'biotools'
library(biotools)
```

### 2. Executando o Agrupamento de Tocher

Utilizaremos a função *tocher*() do pacote *biotools*. Como entrada, ela espera uma matriz de distâncias. Usaremos a matriz_euclidiana que calculamos anteriormente (página 60), convertendo-a para o formato *as.dist*() que a função *toche*r() entende.

```
# Executando o agrupamento de Tocher
# as.dist() converte a matriz para o formato de objeto de distância
res_tocher <- tocher(as.dist(matriz_euclidiana))

# Exibindo os resultados do agrupamento
print(res_tocher)
```

```
        Tocher's Clustering
Call: tocher.dist(d = as.dist(matriz_euclidiana))
Cluster algorithm: original
Number of objects: 11
Number of clusters: 2
Most contrasting clusters: cluster 1 and cluster 2, with
   average intercluster distance: 35.61047
$`cluster 1`
 [1] P4  P8  P1  P11 P7  P2  P6  P10 P9  P3
$`cluster 2`
[1] P5
```

### b. Método Tocher modificado

O **método de Tocher modificado** foi sugerido por Vasconcelos et al. (2007), os quais alegaram que o Tocher original apresenta uma ineficiência em casos de objetos com grande dissimilaridade, formando grupos com apenas um indivíduo, dada a influência das distâncias dos indivíduos agrupados anteriormente. Segundo os autores, no Tocher modificado, o processo de agrupamento é sequencial e não simultâneo, não existindo influência dos indivíduos já agrupados, logo, mostra-se mais eficaz.

Esse método é iniciado da mesma forma que o original, ou seja, determinando-se o nível máximo para a formação ou a inclusão de um novo objeto. Assim, teremos $\alpha_1 = 28,7$ como o limite de acréscimo na média da distância intragrupo, ou seja, para a formação ou a inclusão de acesso a um grupo. Logo, o primeiro grupo será formado pelos objetos com menores distâncias.

Novamente, são obtidas as distâncias dos objetos ainda não agrupados em relação ao grupo em formação e, da mesma forma, realiza-se um novo teste a fim de incluir um outro objeto ao grupo em formação. Na inclusão de um novo objeto ao grupo inicial, quando se obtém um valor de acréscimo da distância dentro do grupo superior ao valor de $\alpha_1$, o objeto não entra no grupo e fecha-se, portanto, o Grupo 1. Dessa forma, prossegue-se a análise apenas como os objetos ainda não agrupados, ou seja, utiliza-se uma nova matriz de dissimilaridade formada apenas com os objetos ainda não agrupados. Logo, para a formação de um novo grupo, estabelece-se um $\alpha_2$, que será o máximo da medida de dissimilaridade entre as menores distâncias, encontrado na nova matriz, propiciando um novo valor do critério de agrupamento após a formação de cada novo grupo, e não um critério único, conforme o método de Tocher original (Vasconcelos et al., 2007).

No presente estudo de caso, os métodos de Tocher original (Rao, 1952) e modificado (Vasconcelos et al., 2007) não diferiram.

**Passo a passo no R: Método de Tocher Modificado**

**1. Instalando e Carregando o Pacote Necessário**

Para utilizar o Método de Tocher Modificado, precisamos do pacote *MultivariateAnalysis*. Se você ainda não o tem instalado, pode instalá-lo e, depois, carregá-lo.

```
# Instalar o pacote 'MultivariateAnalysis' (execute apenas se necessário)
# install.packages("MultivariateAnalysis")

# Carregar o pacote 'MultivariateAnalysis'
library(MultivariateAnalysis)
```

**2. Preparando a Matriz de Distância**

O método de Tocher Modificado, assim como o original, opera sobre uma matriz de distâncias. Utilizaremos a matriz_euclidiana que calculamos anteriormente (página 60), convertendo-a para o formato *dist* que a função *Tocher*() espera.

```
# Gerar ou carregar sua matriz de dissimilaridade (ex: Distância Euclidiana)
# Usaremos a matriz_euclidiana calculada em seções anteriores
matriz_euclidiana <- as.matrix(dist(t(X)))
matriz_dist_euclidiana <- as.dist(matriz_euclidiana)
```

### 3. Executando o Agrupamento de Tocher Modificado

Para executar o método modificado, utilizamos a função *Tocher*() do pacote *MultivariateAnalysis*, especificando o argumento *Metodo = "sequential"*. Isso garante que a versão modificada do algoritmo seja aplicada.

```
# Rodar o Agrupamento de Tocher Modificado
resultado_tocher_modificado <- Tocher(matriz_dist_euclidiana, Metodo = "sequential")

# Visualizar os grupos formados
print(resultado_tocher_modificado)
```

```
Agrupamento Tocher
Cluster1:
P4 P8 P1 P11 P7 P2 P6 P10 P9 P3
Cluster2:
P5
Distancia intra e intercluster:
         Cluster1 Cluster2
Cluster1 35.67402 70.22256
Cluster2 70.22256  0.00000
Correlacao Cofenetica: 0.7983374
pvalor: 0.086 baseado no teste Mantel
Hipotese alternativa: A correlacao e maior que 0
```

### c. Método de k-médias

O método k-médias é um dos mais populares e simples algoritmos dentre as técnicas de agrupamento não hierárquicas (Jain, 2010; Matte; Nicoletti, 2019). No método, segue-se um procedimento simples e fácil de classificar os objetos com um determinado número de conjuntos, em que cada grupo possui um ponto central (centroide) e os pontos a uma distância mínima **d** dele são incluídos no grupo (Lloyd, 1982).

Passos:

1. Particionamento dos n objetos em k grupos iniciais de forma aleatória; e
2. Cálculo dos centroides ($\bar{\boldsymbol{X}}_{\boldsymbol{i}}$) de cada grupo:

$$\bar{\boldsymbol{X}}_{\boldsymbol{i}} = \frac{\sum_{\boldsymbol{i-1}}^{\boldsymbol{n}} \boldsymbol{X}_{\boldsymbol{i}}}{\boldsymbol{N}}$$

Em que: $\bar{\boldsymbol{X}}_{\boldsymbol{i}}$ = média do grupo; $\boldsymbol{X}_{\boldsymbol{i}}$ = valor da variável do i-ésimo grupo; e N = número de objetos no grupo.

3. Associação de cada objeto ao grupo cujo centroide está mais próximo, recalculando-se o centroide para esse grupo que recebeu o novo objeto e para o grupo que o perdeu.

4. Repetir o passo 3 até que se atinja um critério de convergência: a) até que nenhum ponto mude de agrupamento; b) até que os centroides não sejam alterados; e c) até alcançar um limite para um número de iterações.

Vale salientar que: o k-médias deve ser aplicado para grandes bancos de dados; é sensível a condições de inicialização; pode-se utilizar outras maneiras para o cálculo dos centroides iniciais; converge para um mínimo local; e a definição do número de centroides é muito subjetiva.

Esse método apresenta como vantagens o fato de os objetos serem automaticamente atribuídos a um grupo e a possibilidade de variação da localização inicial do centroide do grupo, o que permite estabelecer condições iniciais de dependência. Em relação às desvantagens, antes de o algoritmo ser iniciado, tem-se que escolher o número de grupos e todos os objetos de informação são forçados a pertencer a um grupo.

Para Rencher (2002), o **k-médias** pode ser usado como uma possível melhoria nas técnicas hierárquicas, agrupando-se os objetos por meio de um método hierárquico, definindo-se, assim, o número de grupos e, depois, usando os centroides desses grupos como sementes para uma abordagem **k-médias**, o que permitirá que os pontos sejam realocados de um grupo para outro.

## Aplicação

Considerando-se as densidades por parcela (Tabela 40) e os grupos resultantes das partições significativas por meio do **método de Edwards e Cavalli-Sforza** (Figura 16; página 181), ou seja, k = 3, teremos a inicialização do processo (1º Passo).

**Tabela 40.** Densidade (n) de espécies (Xi) por parcela, em Mata da Silvicultura no município de Viçosa-MG

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 0 | 3 | 6 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 5 | 0 | 7 | 0 | 0 | 0 |
| 2 | 1 | 0 | 9 | 3 | 12 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 27 | 0 | 4 | 3 | 0 | 1 | 10 | 0 | 2 | 0 | 0 | 7 | 1 | 0 | 0 | 0 | 1 |
| 4 | 0 | 0 | 6 | 4 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 22 | 29 | 0 | 9 | 9 | 1 | 2 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| 6 | 9 | 0 | 9 | 12 | 0 | 1 | 2 | 13 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 7 | 2 | 12 | 5 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 6 | 0 | 2 | 0 | 0 | 1 | 0 |
| 8 | 3 | 1 | 2 | 4 | 0 | 0 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 9 | 22 | 17 | 7 | 4 | 2 | 5 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 15 | 1 | 4 | 4 | 0 | 11 | 11 | 3 | 2 | 3 | 0 | 1 | 5 | 0 | 2 | 0 | 0 |
| 11 | 7 | 9 | 4 | 4 | 0 | 1 | 5 | 1 | 2 | 1 | 5 | 0 | 2 | 0 | 1 | 0 | 0 |

No segundo passo, realizamos os cálculos dos centróides (Tabela 41).

**Grupo 1**

$$\bar{X}_1 = \frac{1+0+3}{3} = 1,33;\ \bar{X}_2 = \frac{0+0+1}{3} = 0,33;\ \ldots;\ \bar{X}_{17} = \frac{0+0+0}{3} = 0,00$$

**Grupo 2**

$$\bar{X}_1 = \frac{8+2+7}{3} = 5,67;\ \bar{X}_2 = \frac{0+12+9}{3} = 7,00;\ \ldots;\ \bar{X}_{17} = \frac{0+0+0}{3} = 0,00$$

**Grupo 3**

$$\bar{X}_1 = \frac{27+1+9+22+15}{5} = 14,80;\ \bar{X}_2 = \frac{0+0+0+17+1}{5} = 3,60;\ \ldots;\ \bar{X}_{17} = \frac{1+0+0+0+0}{5} = 0,20$$

**Tabela 41.** Centróides dos K=3 grupos do **método de Edwards e Cavalli-Sforza**

| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

Grupo 1: 2, 4 e 8; Grupo 2: 1, 7 e 11; e Grupo 3: 3, 5, 6, 9 e 10.

No 3º passo, realiza-se o cálculo das distâncias de cada parcela em relação aos centróides. Neste caso, utilizaremos a distância euclidiana ao quadrado.

a) Distância da Parcela 2 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 0 | 9 | 3 | 12 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{2(2,4,8)} = (1-1,33)^2 + (0-0,33)^2 + \cdots + (0-0)^2 = 83,44$$

$$d^2_{2(1,7,11)} = (1-5,67)^2 + (0-7,00)^2 + \cdots + (0-0)^2 = 262,67$$

$$d^2_{2(3,5,6,9,10)} = (1-14,80)^2 + (0-3,60)^2 + \cdots + (0-0,20)^2 = 483,88$$

b) Distância da Parcela 4 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 0 | 0 | 6 | 4 | 1 | 1 | 1 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{4(2,4,8)} = (0-1,33)^2 + (0-0,33)^2 + \cdots + (0-0)^2 = 15,78$$

$$d^2_{4(1,7,11)} = (0-5{,}67)^2 + (0-7{,}00)^2 + \cdots + (0-0)^2 = 114{,}33$$

$$d^2_{4(3,5,6,9,10)} = (0-14{,}80)^2 + (0-3{,}60)^2 + \cdots + (0-0{,}20)^2 = 343{,}68$$

c) Distância da Parcela 8 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 3 | 1 | 2 | 4 | 0 | 0 | 0 | 0 | 5 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{8(2,4,8)} = (3-1{,}33)^2 + (1-0{,}33)^2 + \cdots + (0-0)^2 = 55{,}44$$

$$d^2_{8(1,7,11)} = (3-5{,}67)^2 + (1-7{,}00)^2 + \cdots + (0-0)^2 = 110{,}67$$

$$d^2_{8(3,5,6,9,10)} = (3-14{,}80)^2 + (1-3{,}60)^2 + \cdots + (0-0{,}20)^2 = 361{,}48$$

d) Distância da Parcela 1 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 0 | 3 | 6 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 5 | 0 | 7 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{1(2,4,8)} = (8-1{,}33)^2 + (0-0{,}33)^2 + \cdots + (0-0)^2 = 148{,}44$$

$$d^2_{1(1,7,11)} = (8-5{,}67)^2 + (0-7{,}00)^2 + \cdots + (0-0)^2 = 115{,}67$$

$$d^2_{1(3,5,6,9,10)} = (8-14{,}80)^2 + (0-3{,}60)^2 + \cdots + (0-0{,}20)^2 = 246{,}88$$

e) Distância da Parcela 7 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 2 | 12 | 5 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 6 | 0 | 2 | 0 | 0 | 1 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{7(2,4,8)} = (2-1{,}33)^2 + (12-0{,}33)^2 + \cdots + (0-0)^2 = 212{,}44$$

$$d^2_{7(1,7,11)} = (2-5{,}67)^2 + (12-7{,}00)^2 + \cdots + (0-0)^2 = 69{,}00$$

$$d^2_{7(3,5,6,9,10)} = (2-14{,}80)^2 + (12-3{,}60)^2 + \cdots + (0-0{,}20)^2 = 478{,}48$$

f) Distância da Parcela 11 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 7 | 9 | 4 | 4 | 0 | 1 | 5 | 1 | 2 | 1 | 5 | 0 | 2 | 0 | 1 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{11(2,4,8)} = (7 - 1{,}33)^2 + (9 - 0{,}33)^2 + \cdots + (0 - 0)^2 = 184{,}11$$

$$d^2_{11(1,7,11)} = (7 - 5{,}67)^2 + (9 - 7{,}00)^2 + \cdots + (0 - 0)^2 = 27{,}33$$

$$d^2_{11(3,5,6,9,10)} = (7 - 14{,}80)^2 + (9 - 3{,}60)^2 + \cdots + (0 - 0{,}20)^2 = 216{,}28$$

g) Distância da Parcela 3 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 27 | 0 | 4 | 3 | 0 | 1 | 10 | 0 | 2 | 0 | 0 | 7 | 1 | 0 | 0 | 0 | 1 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{3(2,4,8)} = (27 - 1{,}33)^2 + (0 - 0{,}33)^2 + \cdots + (1 - 0)^2 = 834{,}11$$

$$d^2_{3(1,7,11)} = (27 - 5{,}67)^2 + (0 - 7{,}00)^2 + \cdots + (1 - 0)^2 = 619{,}33$$

$$d^2_{3(3,5,6,9,10)} = (27 - 14{,}80)^2 + (0 - 3{,}60)^2 + \cdots + (1 - 0{,}20)^2 = 309{,}08$$

h) Distância da Parcela 5 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 0 | 22 | 29 | 0 | 9 | 9 | 1 | 2 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{5(2,4,8)} = (1 - 1{,}33)^2 + (0 - 0{,}33)^2 + \cdots + (1 - 0)^2 = 1110{,}11$$

$$d^2_{5(1,7,11)} = (1 - 5{,}67)^2 + (0 - 7{,}00)^2 + \cdots + (1 - 0)^2 = 1211{,}33$$

$$d^2_{5(3,5,6,9,10)} = (1 - 14{,}80)^2 + (0 - 3{,}60)^2 + \cdots + (1 - 0{,}20)^2 = 747{,}48$$

i) Distância da Parcela 6 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 9 | 0 | 9 | 12 | 0 | 1 | 2 | 13 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{6(2,4,8)} = (9 - 1{,}33)^2 + (0 - 0{,}33)^2 + \cdots + (0 - 0)^2 = 326{,}44$$

$$d^2_{6(1,7,11)} = (9 - 5{,}67)^2 + (0 - 7{,}00)^2 + \cdots + (0 - 0)^2 = 355{,}67$$

$$d^2_{6(3,5,6,9,10)} = (9 - 14{,}80)^2 + (0 - 3{,}60)^2 + \cdots + (0 - 0{,}20)^2 = 203{,}28$$

j) Distância da Parcela 9 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 22 | 17 | 7 | 4 | 2 | 5 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{9(2,4,8)} = (22 - 1{,}33)^2 + (17 - 0{,}33)^2 + \cdots + (1 - 0)^2 = 742{,}11$$

$$d^2_{9(1,7,11)} = (22 - 5{,}67)^2 + (17 - 7{,}00)^2 + \cdots + (1 - 0)^2 = 425{,}33$$

$$d^2_{9(3,5,6,9,10)} = (22 - 14{,}80)^2 + (17 - 3{,}60)^2 + \cdots + (1 - 0{,}20)^2 = 348{,}28$$

k) Distância da Parcela 10 para os grupos 1 (2, 4 e 8), 2 (1, 7 e 11) e 3 (3, 5, 6, 9, 10):

| Parcela | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 15 | 1 | 4 | 4 | 0 | 11 | 11 | 3 | 2 | 3 | 0 | 1 | 5 | 0 | 2 | 0 | 0 |
| Grupo | $\bar{X}_1$ | $\bar{X}_2$ | $\bar{X}_3$ | $\bar{X}_4$ | $\bar{X}_5$ | $\bar{X}_6$ | $\bar{X}_7$ | $\bar{X}_8$ | $\bar{X}_9$ | $\bar{X}_{10}$ | $\bar{X}_{11}$ | $\bar{X}_{12}$ | $\bar{X}_{13}$ | $\bar{X}_{14}$ | $\bar{X}_{15}$ | $\bar{X}_{16}$ | $\bar{X}_{17}$ |
| 1 | 1,33 | 0,33 | 5,67 | 3,67 | 4,33 | 0,33 | 0,33 | 1,00 | 2,33 | 2,67 | 0,00 | 0,00 | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 |
| 2 | 5,67 | 7,00 | 4,00 | 3,33 | 0,33 | 0,33 | 2,00 | 0,67 | 1,33 | 1,00 | 3,67 | 1,67 | 1,33 | 2,33 | 0,33 | 0,33 | 0,00 |
| 3 | 14,80 | 3,60 | 9,20 | 10,40 | 0,40 | 5,40 | 6,40 | 3,40 | 2,40 | 1,00 | 0,20 | 2,60 | 1,20 | 0,20 | 0,40 | 0,00 | 0,20 |

$$d^2_{10(2,4,8)} = (15 - 1{,}33)^2 + (1 - 0{,}33)^2 + \cdots + (0 - 0)^2 = 470{,}78$$

$$d^2_{10(1,7,11)} = (15 - 5{,}67)^2 + (1 - 7{,}00)^2 + \cdots + (0 - 0)^2 = 364{,}00$$

$$d^2_{10(3,5,6,9,10)} = (15 - 14{,}80)^2 + (1 - 3{,}60)^2 + \cdots + (0 - 0{,}20)^2 = 151{,}48$$

Resumindo, observa-se, na Tabela 42, que os grupos formados são os mesmos formados no **método Edwards e Cavalli-Sforza**, uma vez que as menores distâncias das parcelas são as de seus grupos da inicialização.

**Tabela 42.** Grupos formados por meio do k-médias (k = 3)

| Grupos | Distância euclidiana ao quadrado ao centroide | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Parcelas | | | | | | | | | | |
| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
| 1 | 148,44 | **83,44** | 834,11 | **15,78** | 1110,11 | 326,44 | 212,44 | **55,44** | 742,11 | 470,78 | 184,11 |
| 2 | **115,67** | 262,67 | 619,33 | 114,33 | 1211,33 | 355,67 | **69,00** | 110,67 | 425,33 | 364,00 | **27,33** |
| 3 | 246,88 | 483,88 | **309,08** | 343,68 | **747,48** | **203,28** | 478,48 | 361,48 | **348,28** | **151,48** | 216,28 |

## Passo a passo no R: Agrupamento por k-médias

### 1. Preparação dos Dados para o k-médias

Para a função *kmeans*(), é fundamental que as observações (nossas parcelas) estejam nas linhas e as variáveis (nossas espécies) estejam nas colunas. Como nossa matriz X original tem as espécies nas linhas e as parcelas nas colunas, precisamos transpô-la.

```
# Transpondo a matriz X para que as parcelas fiquem nas linhas
X_transposto <- t(X)
```

### 2. Definição dos Centroides Iniciais

Podemos fornecer ao algoritmo k-médias centroides iniciais, o que pode ajudar a guiar o agrupamento. Utilizaremos os grupos identificados em análises anteriores (por exemplo, os do método de Edwards e Cavalli-Sforza) para calcular os centroides médios de cada grupo.

```
# Definindo os grupos iniciais de parcelas
grupo_inicial_1 <- c("P2","P4","P8")
grupo_inicial_2 <- c("P1","P7","P11")
grupo_inicial_3 <- c("P3","P5","P6","P9","P10")

# Calculando os centroides iniciais com base nas médias de cada grupo
centros_iniciais <- rbind(
  G1 = colMeans(X_transposto[grupo_inicial_1, ]),
  G2 = colMeans(X_transposto[grupo_inicial_2, ]),
  G3 = colMeans(X_transposto[grupo_inicial_3, ])
)

# Visualizando os centroides iniciais
print(round(centros_iniciais, 2))
```

| Grupo | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 | E17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G1 | 1.33 | 0.33 | 5.67 | 3.67 | 4.33 | 0.33 | 0.33 | 1.00 | 2.33 | 2.67 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | 0.00 |
| G2 | 5.67 | 7.00 | 4.00 | 3.33 | 0.33 | 0.33 | 2.00 | 0.67 | 1.33 | 1.00 | 3.67 | 1.67 | 1.33 | 2.33 | 0.33 | 0.33 | 0.00 |
| G3 | 14.80 | 3.60 | 9.20 | 10.40 | 0.40 | 5.40 | 6.40 | 3.40 | 2.40 | 1.00 | 0.20 | 2.60 | 1.20 | 0.20 | 0.40 | 0.00 | 0.20 |

## 3. Aplicação do Algoritmo k-médias

Agora, aplicamos a função *kmeans*() à nossa matriz transposta (X_transposto), utilizando os centroides iniciais que definimos. O algoritmo irá iterar para ajustar os grupos e seus centroides até que a variabilidade dentro dos grupos seja minimizada.

```
# Aplicação do algoritmo k-médias
res_kmeans <- kmeans(
  x = X_transposto,          # Matriz de dados com parcelas nas linhas
  centers = centros_iniciais, # Centroides iniciais
  algorithm = "Hartigan-Wong", # Algoritmo a ser usado
  iter.max = 100             # Número máximo de iterações
)
```

## 4. Análise dos Grupos Finais

Após a execução do algoritmo, podemos analisar os resultados para ver a alocação final de cada parcela em seu grupo, os centroides dos grupos resultantes e as medidas de variabilidade.

```
# Visualizando a alocação de cada parcela em seu grupo
print(res_kmeans$cluster)
```

| P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 3 | 2 | 1 | 2 | 2 | 2 | 3 | 3 | 2 |

```
# Visualizando os centroides finais dos grupos
print(round(res_kmeans$centers, 2))
```

| Grupo | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 | E15 | E16 | E17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G1 | 1.00 | 0.00 | 22.00 | 29.00 | 0.00 | 9.00 | 9.00 | 1.00 | 2.00 | 0.00 | 0.00 | 5.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| G2 | 4.29 | 3.14 | 5.43 | 4.71 | 2.00 | 0.43 | 1.29 | 2.57 | 2.43 | 1.57 | 1.57 | 0.71 | 0.57 | 1.29 | 0.14 | 0.14 | 0.00 |
| G3 | 21.33 | 6.00 | 5.00 | 3.67 | 0.67 | 5.67 | 7.00 | 1.00 | 1.33 | 1.67 | 0.33 | 2.67 | 2.00 | 0.00 | 0.67 | 0.00 | 0.33 |

```
# Variabilidade dentro de cada grupo (within-cluster sum of squares)
print(res_kmeans$withinss)
```

```
0.0000 801.7143 448.6667
```

```
# Variabilidade entre os grupos (between-cluster sum of squares)
print(res_kmeans$betweenss)
```

```
1768.528
```

```
# Variabilidade total dentro dos grupos (total within-cluster sum of squares)
print(res_kmeans$tot.withinss)
```

```
1250.381
```

```
# Organizando os grupos finais em uma lista para melhor visualização
grupos_finais_kmeans <- data.frame(
  Parcela = names(res_kmeans$cluster),
  Grupo   = res_kmeans$cluster
)
```

```
print(split(grupos_finais_kmeans$Parcela, grupos_finais_kmeans$Grupo))
$`1`
[1] "P5"

$`2`
[1] "P1"  "P2"  "P4"  "P6"  "P7"  "P8"  "P11"

$`3`
[1] "P3"  "P9"  "P10"
```

## 2.3. Técnicas de Modelos de Densidade de Mistura Finita

Na maioria dos métodos hierárquicos e de otimização, as suposições sobre a estrutura de classes não são explicitamente declaradas, logo, soluções contraditórias podem ser encontradas para os grupos e o número de grupos. Portanto, os procedimentos para uma solução final, em geral, são informais e subjetivos (Everitt et al., 2011).

Para solucionar esse problema, uma alternativa é postular um modelo estatístico formal para a população da qual os dados são amostrados, que assume que essa população consiste em um número de subpopulações (os grupos), em cada uma das quais as variáveis têm distintas funções de densidade de probabilidade multivariada, conhecida como densidade de mistura finita para a população na totalidade (Everitt et al., 2011).

Os algoritmos que se baseiam nas técnicas de densidade têm por princípio básico a determinação de regiões de alta densidade de pontos. O conhecimento de alta densidade é bastante próximo daquele de *moda*, isto é, a ocorrência do evento de maior frequência. Portanto, os algoritmos só se aplicam a conjuntos cuja representação de seus elementos no espaço $\mathbf{I_p}$ é *bimodal* ou *polimodal*.

Todos os métodos de densidade têm suas bases no algoritmo da hierarquização, denominado de método de ligação simples. Outra particularidade desse método é que se admite, como princípio, que as **p** variáveis sejam quantitativas.

Entre os exemplos de algoritmos de agrupamento baseados em densidade, podemos citar: DBSCAN (Ester et al., 1996), DENCLUE (Hinneburg; Keim, 1998; Hinneburg; Gabriel, 2007) e OPTICS (Ankerst et al., 1999). Esses algoritmos definem os grupos como regiões de altas densidades separadas por regiões de baixas densidades. No entanto, eles têm dificuldade em encontrar grupos com densidades amplamente variadas porque um limite de densidade global é usado para identificar regiões de alta densidade (Zhu et al., 2021). Assim, pode-se dizer que não há agrupamento completo. Além disso, esses métodos requerem algum tipo de parâmetro, como número de grupos a serem pesquisados ou um ponto inicial para iniciar o processo (Güngör; Özmen, 2017).

# Capítulo 3

## Validação do Método de Agrupamento

A indicação de qual método de agrupamento é mais apropriado para determinado tipo de dados ainda é uma das questões controversas na Análise de Agrupamento. Isso porque o resultado pode ser interpretado de diversas formas. Assim, nenhuma solução deve ser aceita sem a avaliação de sua confiabilidade e validação (Malhotra, 2019).

Várias são as proposições para identificar a validação da estrutura obtida a partir de um método de agrupamento (Jain; Dubes, 1988; Faceli et al., 2005), tais como:

a) validações externas – avaliam os agrupamentos de acordo com uma estrutura pré-especificada, imposta ao conjunto de dados que reflete a intuição do pesquisador sobre a estrutura presente nos dados;

b) validações internas – analisam as informações contidas nos grupos obtidos com base apenas nos dados originais (matriz de dissimilaridade); e

c) validações relativas – Comparam diversos agrupamentos para decidir qual deles é o melhor em algum aspecto.

Dentre as técnicas de validação interna, citam-se: coesão, acoplamento, coeficiente de correlação cofenética e coeficiente de silhueta. Dentre essas, a mais empregada é o **Coeficiente de Correlação Cofenética**, proposto por Sokal e Rohlf (1962).

- **Coeficiente de correlação cofenética**

Por meio do **Coeficiente de Correlação Cofenética** ($\mathbf{r_{cof}}$), é realizada a verificação das proximidades das matrizes de dissimilaridade. A matriz original (**D**) é a **matriz fenética**, e a matriz final do processo de agrupamento (**C**) é a **matriz cofenética,** sendo $\mathbf{r_{cof}}$ calculado por meio da seguinte expressão:

$$r_{cof} = sendo\ \frac{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(C_{jj'}-\bar{C}\right)\left(D_{jj'}-\bar{D}\right)}{\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(C_{jj'}-\bar{C}\right)^2}\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(D_{jj'}-\bar{D}\right)^2}}$$

Com: $\bar{C} = \left(\frac{2}{n(n-1)}\right)\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n} C_{jj'}$ e $\bar{D} = \left(\frac{2}{n(n-1)}\right)\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n} D_{jj'}$

Em que: $\boldsymbol{r_{cof}}$ = coeficiente de correlação cofenética; $\boldsymbol{C_{jj'}}$ = distância euclidiana na matriz **C** entre o **j** e o **j'**-ésimo objeto (**j'** = **j** + 1); $\boldsymbol{\bar{C}}$ = média geral das distâncias euclidianas na matriz **C**, considerando apenas **j** e **j'** (**j'** = **j** + 1); $D_{jj'}$ = distância euclidiana na matriz **D** entre o **j** e o

**j'**-ésimo objeto (**j'** = **j** + 1); e $\bar{\boldsymbol{D}}$ = média geral das distâncias euclidianas na matriz **D**, considerando apenas **j** e **j'** (**j'** = **j** + 1).

Quanto maior o valor de $\boldsymbol{r_{cof}}$, menor será a distorção provocada pelo método de agrupamento, logo, o agrupamento que corresponde exatamente aos coeficientes da matriz fenética tem $\boldsymbol{r_{cof}} = 1{,}00$ (Legendre; Legendre, 2012; Kassambara, 2017). Na prática, dendrogramas com $\boldsymbol{r_{cof}} < 0{,}70$ indicam a inadequação do método de agrupamento adotado para resumir a informação do conjunto de dados analisado (Rohlf, 1970).

Uma alta correlação entre a matriz original (D) e a cofenética (**C**) é sinal de escassa distorção e, geralmente, os valores oscilam entre 0,60 e 0,95 (Palacio et al., 2020). Já valores acima de 0,75 (Kassambara, 2017) ou de 0,80 (Sneath & Sokal, 1973) indicam uma boa representação da matriz original de similaridade (**D**) por parte do dendrograma.

Uma outra forma de analisar o coeficiente de correlação cofenética é utilizar o diagrama de Shepard, que permite comparar a matriz de distância original e a cofenética por meio de um gráfico de dispersão (Legendre; Legendre, 1998; Borcard; Gillet; Legendre, 2011). Esse diagrama representa as proximidades (eixo x – matriz de dissimilaridade original) em relação às disparidades (eixo y – matriz cofenética), que detecta "*outliers*". Um ajuste perfeito indicaria que todos os pontos estariam alinhados em uma linha monotônica crescente, representada, neste gráfico, pela linha vermelha (Figura 18, saída R). Os possíveis "*outliers*" são os pontos mais distantes dessa linha vermelha.

Algumas das desvantagens da correlação cofenética são a ausência de averiguação por um teste de significância (Borcard; Gillet; Legendre, 2011) e a sensibilidade ao tamanho dos grupos formados (Farris, 1969).

## Aplicação

Considerando a matriz de distância euclidiana **D** como a matriz original de dissimilaridade, e C como a matriz de dissimilaridade após a aplicação do **método de ligação simples** (Figura 8; página 105):

$$\mathbf{D} = \begin{bmatrix} 0,00 & 17,8 & 23,0 & 12,5 & 33,9 & 17,3 & 18,5 & 11,3 & 24,8 & 20,0 & 14,7 \\ & 0,00 & 31,7 & 12,0 & 34,4 & 21,8 & 18,3 & 16,2 & 29,4 & 25,6 & 18,7 \\ & & 0,00 & 29,6 & 41,8 & 27,0 & 30,9 & 27,9 & 22,4 & 17,9 & 24,2 \\ & & & 0,00 & 32,3 & 17,3 & 14,7 & 7,7 & 28,3 & 21,5 & 13,5 \\ & & & & 0,00 & 28,7 & 38,6 & 35,6 & 41,4 & 34,8 & 34,7 \\ & & & & & 0,00 & 24,6 & 19,0 & 27,6 & 21,5 & 19,4 \\ & & & & & & 0,00 & 15,5 & 22,4 & 24,4 & 8,8 \\ & & & & & & & 0,00 & 26,7 & 21,2 & 13,3 \\ & & & & & & & & 0,00 & 22,8 & 19,2 \\ & & & & & & & & & 0,00 & 17,5 \\ & & & & & & & & & & 0,00 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0,00 & 12,0 & 17,9 & 11,3 & 28,7 & 17,3 & 13,3 & 11,3 & 19,2 & 17,5 & 13,3 \\ & 0,00 & 17,9 & 12,0 & 28,7 & 17,3 & 13,3 & 12,0 & 19,2 & 17,5 & 13,3 \\ & & 0,00 & 17,9 & 28,7 & 17,9 & 17,9 & 17,9 & 19,2 & 17,9 & 17,9 \\ & & & 0,00 & 28,7 & 17,3 & 13,3 & 7,7 & 19,2 & 17,5 & 13,3 \\ & & & & 0,00 & 28,7 & 28,7 & 28,7 & 28,7 & 28,7 & 28,7 \\ & & & & & 0,00 & 17,3 & 17,3 & 19,2 & 17,5 & 17,3 \\ & & & & & & 0,00 & 13,3 & 19,2 & 17,5 & 8,8 \\ & & & & & & & 0,00 & 19,2 & 17,5 & 13,3 \\ & & & & & & & & 0,00 & 19,2 & 19,2 \\ & & & & & & & & & 0,00 & 17,5 \\ & & & & & & & & & & 0,00 \end{bmatrix}$$

$$\bar{C} = \left(\frac{2}{11(11-1)}\right)(12{,}0 + 17{,}9 + 11{,}3 + 28{,}7 + 17{,}3 + 13{,}3 + 11{,}3 + 19{,}2 + 17{,}5 + 13{,}3 + 17{,}9 + 12{,}0 + 28{,}7 + 17{,}3 + 13{,}3 + 12{,}0 + 19{,}2 + 17{,}5 + 13{,}3 + 17{,}9 + 28{,}7 + 17{,}9 + 17{,}9 + 17{,}9 + 19{,}2 + 17{,}9 + 17{,}9 + 28{,}7 + 17{,}3 + 13{,}3 + 7{,}7 + 19{,}2 + 17{,}5 + 13{,}3 + 28{,}7 + 28{,}7 + 28{,}7 + 28{,}7 + 28{,}7 + 28{,}7 + 17{,}3 + 17{,}3 + 19{,}2 + 17{,}5 + 17{,}3 + 13{,}3 + 19{,}2 + 17{,}5 + 8{,}8 + 19{,}2 + 17{,}5 + 13{,}3 + 19{,}2 + 19{,}2 + 17{,}5) = 18{,}378182$$

$$\bar{D} = \left(\frac{2}{11(11-1)}\right)(17{,}8 + 23{,}0 + 12{,}5 + 33{,}9 + 17{,}3 + 18{,}5 + 11{,}3 + 24{,}8 + 20{,}0 + 14{,}7 + 31{,}7 + 12{,}0 + 34{,}4 + 21{,}8 + 18{,}3 + 16{,}2 + 29{,}4 + 25{,}6 + 18{,}7 + 29{,}6 + 41{,}8 + 27{,}0 + 30{,}9 + 27{,}9 + 22{,}4 + 17{,}9 + 24{,}2 + 32{,}3 + 17{,}3 + 14{,}7 + 7{,}7 + 28{,}3 + 21{,}5 + 13{,}5 + 28{,}7 + 38{,}6 + 35{,}6 + 41{,}4 + 34{,}8 + 34{,}7 + 24{,}6 + 19{,}0 + 27{,}6 + 21{,}5 + 19{,}4 + 15{,}5 + 22{,}4 + 24{,}4 + 8{,}8 + 26{,}7 + 21{,}2 + 13{,}3 + 22{,}8 + 19{,}2 + 17{,}5) = 23{,}210909$$

$$\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(C_{jj'} - \bar{C}\right)\left(D_{jj'} - \bar{D}\right) = [(12 - 18{,}378182)(17{,}8 - 23{,}210909) + (17{,}9 - 18{,}378182)(23{,}0 - 23{,}210909) + \cdots + (17{,}5 - 18{,}378182)(17{,}5 - 23{,}210909)] = 2222{,}653091$$

$\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(C_{jj'}-\bar{C}\right)^2 = [(12-18{,}378182)^2+(17{,}9-18{,}378182)^2+\cdots+(17{,}5-18{,}378182)^2] = 1719{,}993818$

$\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(D_{jj'}-\bar{D}\right)^2 = [(17{,}8-23{,}210909)^2+(23{,}0-23{,}210909)^2+\cdots+(17{,}5-23{,}210909)^2] = 3577{,}013455$

$r_{cof} = \frac{2222{,}653091}{\sqrt{1719{,}993818}\sqrt{3577{,}013455}} = 0{,}8961$ ou 89,61%.

Vale ressaltar que os procedimentos formais de avaliação da confiabilidade e de validação de soluções em análise de agrupamento são complexos e nem sempre totalmente defensáveis (Malhotra, 2019).

**Passo a passo no R: Correlação Cofenética**

**1. Correlação Cofenética para o Método da Ligação Simples**

Para realizar esse cálculo, é necessário ter a matriz de dados X (com as parcelas nas linhas e as espécies nas colunas, ou transposta conforme a necessidade da função *dist*()).

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método da ligação simples
hc_single <- hclust(D, method = "single")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_single <- cophenetic(hc_single)

# 4. Calcular o coeficiente de correlação cofenética
r_cof_single <- cor(D, C_single)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_single, 4)
```

```
0.897
```

```
# 6. Gerar o Diagrama de Shepard (Figura 18)
plot(D, C_single,
     main = "Diagrama de Shepard (Ligação Simples)",
     xlab = "Distâncias Originais",
     ylab = "Distâncias Cofenéticas",
     pch = 19, col = "blue")
abline(lm(C_single ~ D), col = "red", lwd = 2)
```

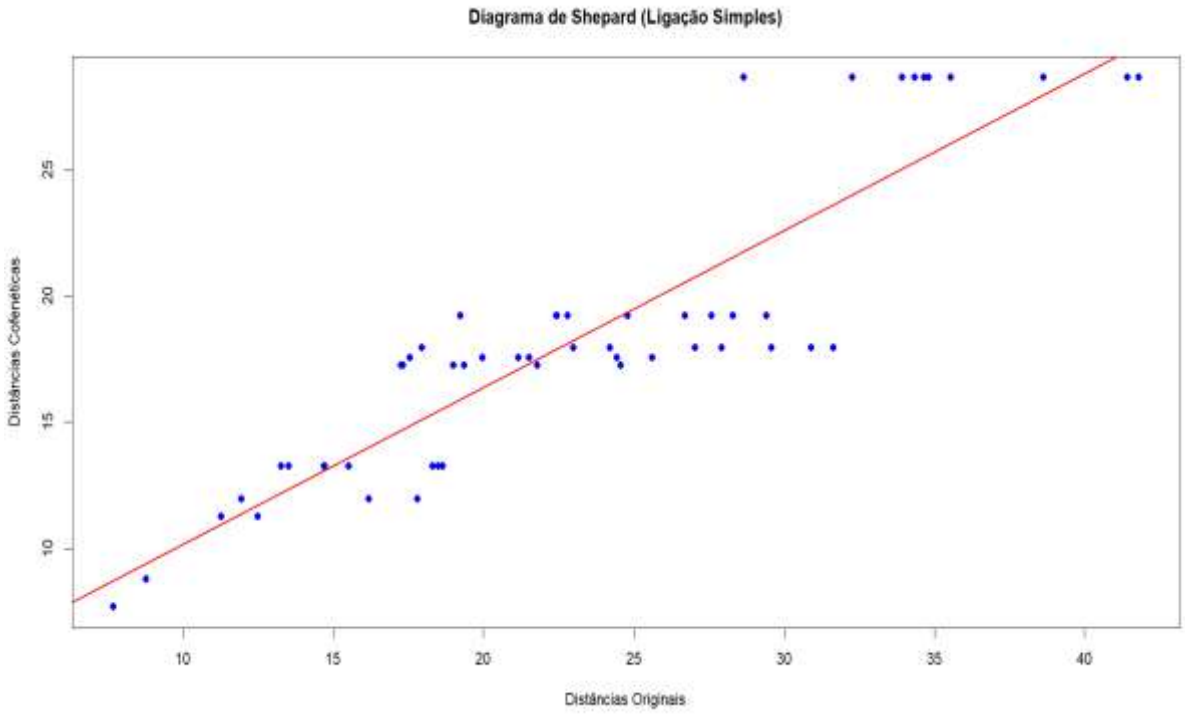



**Figura 18**. Diagrama de Shepard para o método da Ligação Simples obtido com a aplicação do R.

## 2. Correlação Cofenética para o Método da Ligação Completa

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método da ligação completa
hc_complete <- hclust(D, method = "complete")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_complete <- cophenetic(hc_complete)

# 4. Calcular o coeficiente de correlação cofenética
r_cof_complete <- cor(D, C_complete)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_complete, 4)
```

0.9054

## 3. Correlação Cofenética para o Método do Centroide

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método do centroide
hc_centroid <- hclust(D, method = "centroid")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_centroid <- cophenetic(hc_centroid)
```

```
# 4. Calcular o coeficiente de correlação cofenética
r_cof_centroid <- cor(D, C_centroid)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_centroid, 4)
```

0.9108

### 4. Correlação Cofenética para o Método da Mediana

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método da mediana
hc_median <- hclust(D, method = "median")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_median <- cophenetic(hc_median)

# 4. Calcular o coeficiente de correlação cofenética
r_cof_median <- cor(D, C_median)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_median, 4)
```

0.8599

### 5. Correlação Cofenética para o Método UPGMA (Ligação Média)

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método UPGMA (average)
hc_upgma <- hclust(D, method = "average")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_upgma <- cophenetic(hc_upgma)

# 4. Calcular o coeficiente de correlação cofenética
r_cof_upgma <- cor(D, C_upgma)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_upgma, 4)
```

0.9194

### 6. Correlação Cofenética para o Método de Ward.D

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método Ward.D
hc_ward_d <- hclust(D, method = "ward.D")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_ward_d <- cophenetic(hc_ward_d)
```

```
# 4. Calcular o coeficiente de correlação cofenética
r_cof_ward_d <- cor(D, C_ward_d)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_ward_d, 4)
```

```
0.7813
```

## 7. Correlação Cofenética para o Método de Ward.D2

```
# 1. Calcular a matriz de distâncias originais (usando a matriz X)
D <- dist(t(X))

# 2. Realizar o agrupamento hierárquico pelo método Ward.D2
hc_ward_d2 <- hclust(D, method = "ward.D2")

# 3. Obter a matriz de distâncias cofenéticas do dendrograma
C_ward_d2 <- cophenetic(hc_ward_d2)

# 4. Calcular o coeficiente de correlação cofenética
r_cof_ward_d2 <- cor(D, C_ward_d2)

# 5. Exibir o coeficiente de correlação cofenética
round(r_cof_ward_d2, 4)
```

```
0.8389
```

# Capítulo 4

## Determinação do Número de Grupos

Um problema fundamental e não totalmente resolvido na análise de agrupamento é determinar o número "verdadeiro" de grupos em um conjunto de dados (Sugar; James, 2003), visto que não há regras difíceis nem fáceis (Malhotra, 2019). Os procedimentos não hierárquicos geralmente exigem que o usuário especifique o número de grupos antes de qualquer agrupamento ser realizado. Já os métodos hierárquicos produzem rotineiramente uma série de soluções que variam de n grupos até uma solução com apenas um grupo presente (Charrad et al., 2014).

O número de grupos pode ser definido pelo conhecimento que se tem sobre os dados, pela conveniência do pesquisador ou por simplicidade, além da inspeção visual das ramificações do dendrograma ou recorrendo a algum procedimento estatístico (Cruz et al., 2020).

Na literatura e nos diversos métodos apresentados aqui, pode-se observar a possibilidade de formação de vários grupos, que, conforme o exame do dendrograma no método hierárquico aplicado, depende da visão do pesquisador quanto ao nível de similaridade para considerar grupos mais próximos. Por exemplo, considerando as linhas vermelhas na Figura 19, pode-se formar 9, 6 e 2 grupos. A linha considerada para a formação de grupos em um dendrograma é conhecida como *Linha Fenon*, a qual deve ser estabelecida em pontos em que há mudança abrupta da ramificação do dendrograma. Normalmente, busca-se traçar a *linha fenon* em termos relativos à maior distância de ligação, na Figura 19, temos 28,7 como a distância euclidiana máxima de ligação, logo, caso consideremos a *linha fenon* a uma distância igual a 20, o nível de similaridade seria de até 69,68% e haveria a formação de dois grupos. Já para uma distância igual a 15, o nível de similaridade seria de até 52,26% e haviria a formação de seis grupos. E, finalmente, para uma distância igual a 10, o nível de similaridade seria de até 34,84% e haveria a formação de nove grupos. Logo, o número de grupos dependerá do rigor adotado pelo pesquisador. Para Bouroche e Saporta (1982), um critério eficiente é realizar o *corte* no dendrograma em um nível igual à metade da maior distância de agrupamento (correspondente a uma linha em uma distância igual a 14,35 na Figura 19 e na formação de seis grupos), ou seja, nível de similaridade igual a 50,0%, no entanto, confere subjetividade considerável.

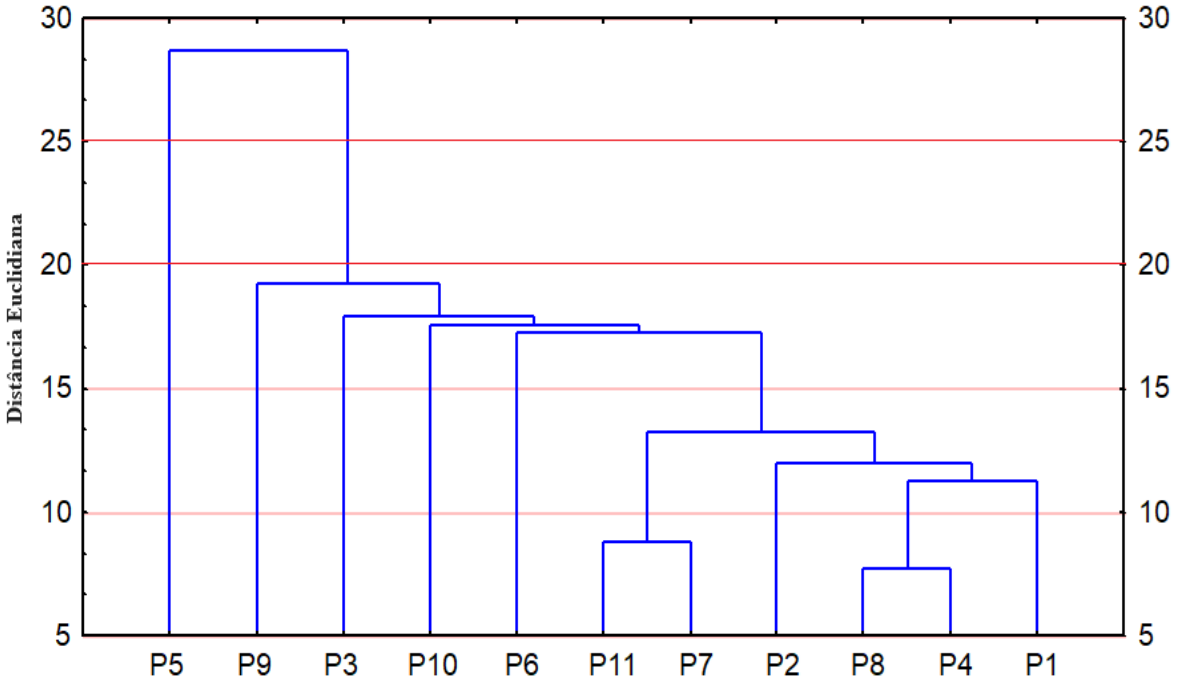



**Figura 19.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método de ligação simples, com base na distância euclidiana (d) e com linhas fenon traçadas em d = 10, 15, 20 e 25.

Outra forma seria criar uma curva de progressão de distâncias percorridas ao longo dos passos do processo de agrupamento (Pereira, 1999). No caso do **método de ligação simples**, verifica-se que, nos 2º, 5º e 9º passos, há uma inclinação diferente daqueles que os sucedem, correspondendo a 2, 2,8 e 6,8 vezes, respectivamente (Tabela 43). Para Hair et al. (2009), a *regra da parada* é baseada na avaliação de variações de heterogeneidade entre soluções de agrupamentos, em que, quando ocorrem grandes aumentos de heterogeneidade, o pesquisador escolhe a solução anterior, pois a combinação une agrupamentos muito diferentes. Pode-se, assim, entender todo o **trajeto de distâncias**, em que se definem 4 grupos (Figura 20): **I** (Parcelas 4, 7, 8, e 11); o **II** (1 e 2); **III** (3,6,9 e 10) e **IV** (5).

**Tabela 43**. Avaliação de variações de heterogeneidade entre soluções de agrupamentos considerando o método de ligação simples e distância euclidiana

| Passo | Distância | Parcelas | Regra de Parada<br>Coeficiente de Aglomeração<br>Aumento percentual para o próximo passo |
|---|---|---|---|
| 1 | 7,7 | 4 e 8 | 14,29 |
| 2 | 8,8 | 7 e 11 | 28,41 |
| 3 | 11,3 | 1, 4 e 8 | 6,19 |
| 4 | 12 | 1, 4, 8 e 2 | 10,83 |
| 5 | 13,3 | 1, 4, 8, 2, 7 e 11 | 30,08 |
| 6 | 17,3 | 1, 4, 8, 2, 7, 11 e 6 | 1,16 |
| 7 | 17,5 | 1, 4, 8, 2, 7, 11, 6 e 10 | 2,29 |
| 8 | 17,9 | 1, 4, 8, 2, 7, 11, 6, 10 e 3 | 7,26 |
| 9 | 19,2 | 1, 4, 8, 2, 7, 11, 6, 10, 3 e 9 | 49,48 |
| 10 | 28,7 | 1, 4, 8, 2, 7, 11, 6, 10, 3, 9 e 5 | - |

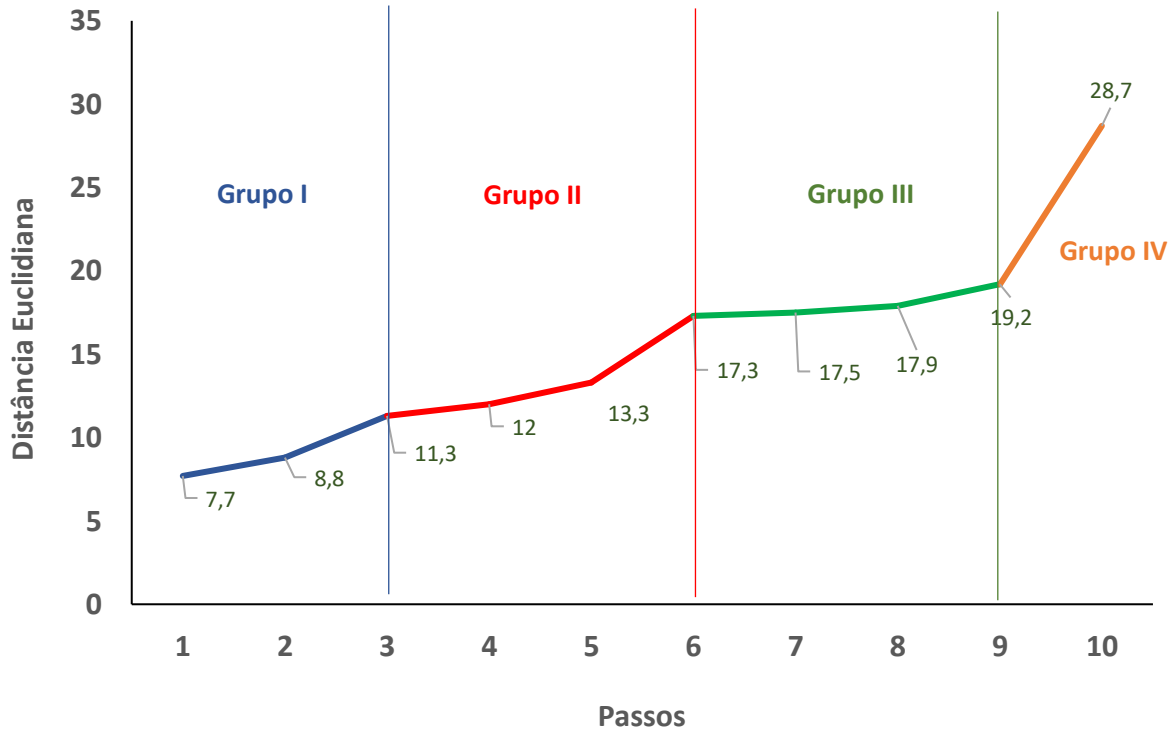



**Figura 20.** Trajeto das distâncias euclidianas de aglomeração conforme os passos de agrupamentos pelo método da ligação simples11.

Outro método alternativo é a comparação gráfica do número de grupos com o coeficiente de fusão, isto é, o valor numérico (distância ou similaridade) para o qual vários casos se unem para formar um grupo (Reis, 2001). Assim, quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão, essa partição poderá ser aceita como ótima. Tomando novamente o **método de ligação simples** e **distância euclidiana** como exemplo, obtêm-se quatro grupos, com base nas variações de 22,12%, 23,12% e 33,10% (Tabela 44 e Figura 21).

**Tabela 44.** Coeficientes de fusão para formar os agrupamentos, método de ligação simples e distância euclidiana

| Número de Grupos | Distância | Aumento % da distância |
|---|---|---|
| 1 | 28,7 | 33,10 |
| 2 | 19,2 | 6,77 |
| 3 | 17,9 | 2,23 |
| 4 | 17,5 | 1,14 |
| 5 | 17,3 | 23,12 |
| 6 | 13,3 | 9,77 |
| 7 | 12 | 5,83 |
| 8 | 11,3 | 22,12 |
| 9 | 8,8 | 12,5 |
| 10 | 7,7 | - |

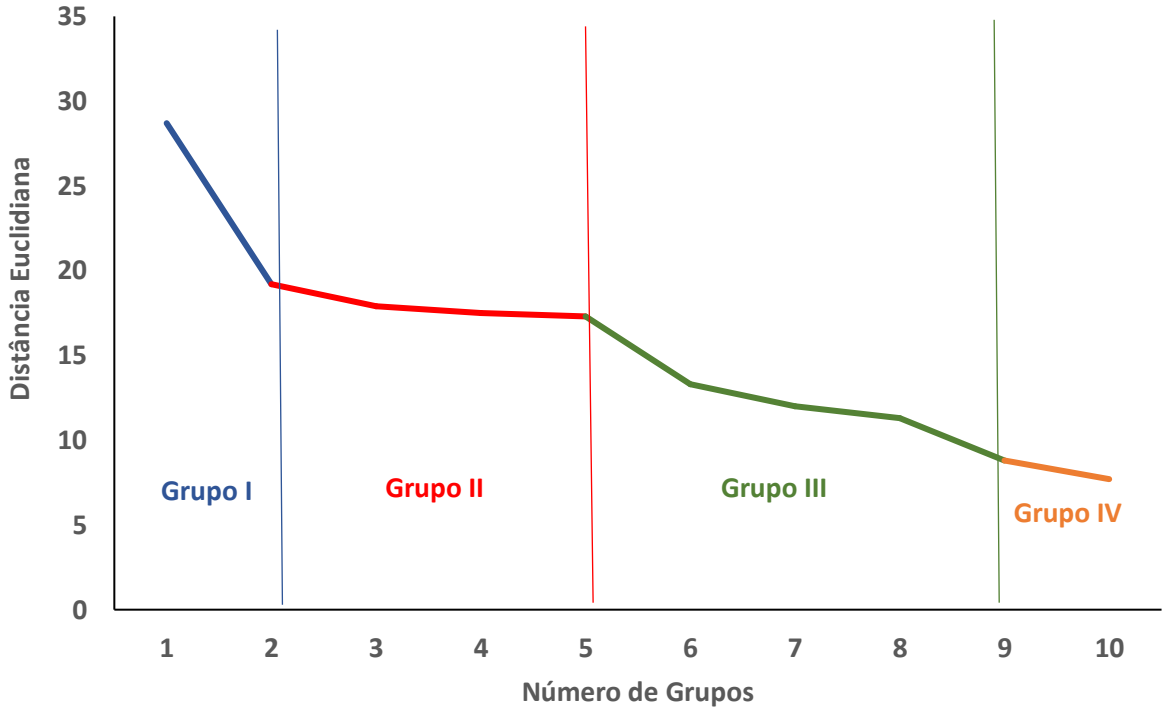



**Figura 21.** Evolução das distâncias euclidianas de aglomeração conforme os números de grupos pelo método da ligação simples

Por outro lado, há vários procedimentos estatísticos que buscam evitar a subjetividade do pesquisador na definição do número de grupos, conforme o método de agrupamento adotado (Dalton et al., 2009; Cruz et al., 2020). Entre os procedimentos, podemos citar o desvio padrão médio e coeficiente de determinação propostos por Khattree e Nair (2002), além do método baseado no tamanho relativo das fusões (distâncias) no dendrograma, por Mojena (1977). No entanto, ressalta-se que não há consenso quanto ao melhor procedimento (Milligan e Cooper, 1985).

## Passo a passo no R: Determinação do Número de Grupos (Linha Fenon)

### 1. Gerando o Dendrograma Base (Ligação Simples)

Primeiro, precisamos calcular a matriz de distâncias a partir da nossa matriz de dados X e, em seguida, realizar o agrupamento hierárquico. Utilizaremos o método de ligação simples como exemplo.

```
# 1. Calcular a matriz de distâncias euclidianas a partir da matriz X
D_euclidiana <- dist(t(X), method = "euclidean")

# 2. Realizar o agrupamento hierárquico pelo método de ligação simples
fit_simples <- hclust(D_euclidiana, method = "single")
```

## 2. Plotando o Dendrograma com as Linhas Fenon

Agora, vamos plotar o dendrograma e adicionar as linhas fenon em diferentes níveis de distância para visualizar os possíveis cortes e a formação dos grupos. As linhas horizontais (h) representam os níveis de corte, enquanto a função rect.hclust desenha retângulos ao redor dos grupos formados por um corte específico.

```
# 1. Plotar o dendrograma (Figura 22)
plot(fit_simples,
     main = "Dendrograma com Linhas Fenon (Ligação Simples)",
     xlab = "Parcelas",
     ylab = "Distância Euclidiana",
     sub = "",
     hang = -1)

# 2. Adicionar linhas de corte fenon em diferentes níveis de distância
abline(h = c(10, 15, 20, 25), col = "red", lty = 2)

# 3. Destacar os grupos formados por um corte específico (ex: d = 15)
rect.hclust(fit_simples, h = 15, border = "blue")
```

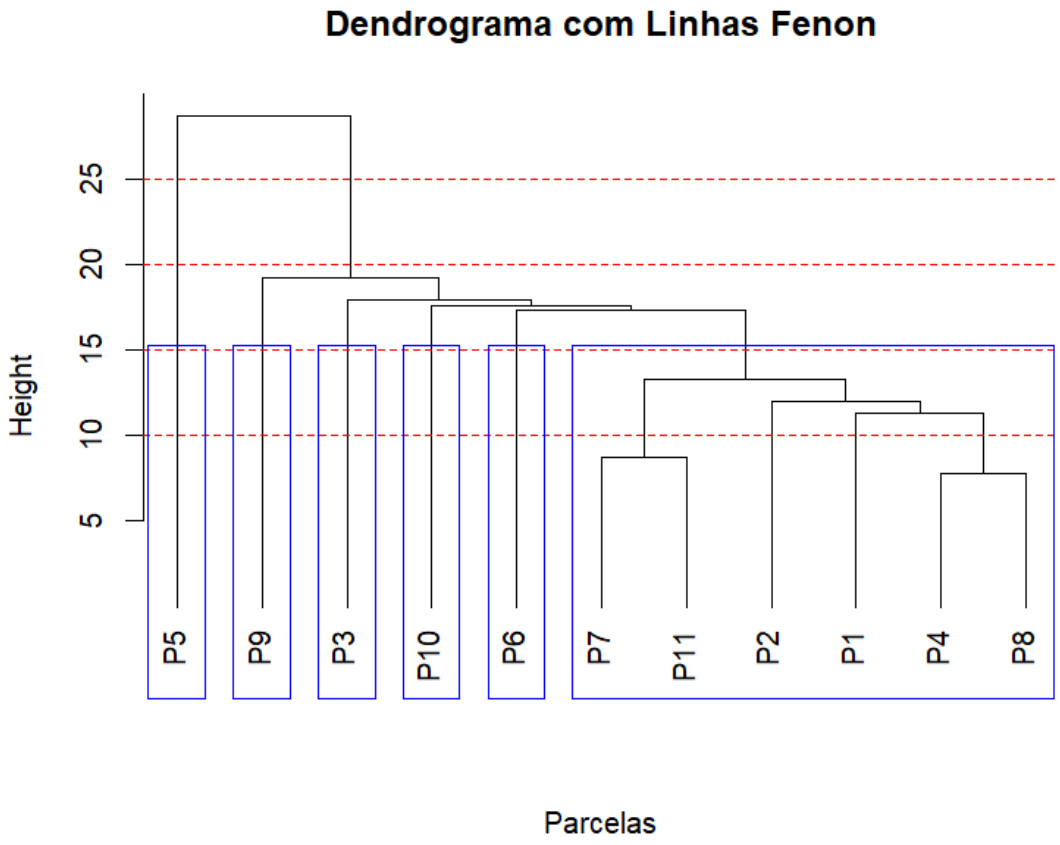



**Figura 22.** Dendrograma representando as sequências das fusões das parcelas, obtido pelo emprego do método de ligação simples, com base na distância euclidiana, considerando linha fenon em d = 10, 15, 20 e 25 e a formação de quatro grupos.

- **Desvio padrão quadrático médio, coeficiente de determinação ($R^2$)** e **$R^2$ semiparcial** (Khattree e Nair, 2002)

Todos os procedimentos hierárquicos continuam até que todos os objetos sejam alocados em um grupo. Para a decisão de quando parar o procedimento de agrupamento, existem algumas medidas estatísticas que podem ser usadas, tais como o desvio padrão quadrático médio (RMSSTD), coeficiente de determinação ($R^2$) e $R^2$ semiparcial (Khattree e Nair, 2002).

O desvio padrão quadrático médio (RMSSTD) é a raiz quadrada da variância do novo grupo formado pela fusão de dois grupos, somada sobre todas as variáveis. Especificamente, se o grupo recém-formado $B_q$ tiver as respectivas variâncias $s_{q1}^2$, $s_{q2}^2$, ..., $s_{qp}^2$, então:

$$RMSSTD_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s_{qj}^2}$$

Obviamente, quanto mais semelhantes forem os dois grupos unidos para formar Bq, mais homogêneo será o grupo Bq, e, consequentemente, menor será o valor de RMSSTDq.

O coeficiente de determinação ($R^2$) para o grupo Bq é definido como:

$$R_q^2 = 1 - \frac{\sum SQW_q}{SQTotal}$$

Em que: **SQTotal** = soma dos quadrados total corrigida de todas as unidades, somada sobre todas as variáveis; e **SQWq** = soma dos quadrados agrupados e corrigidos dentro do cluster, somada sobre todas as variáveis, no **estágio** ou **passo** em que o grupo **Bq** foi formado.

Já o **$R^2$ semiparcial** mede a perda de homogeneidade pela fusão de dois grupos. É definido como a redução no valor de **$R^2$** correspondente quando os dois grupos são mesclados. Os valores de **$R^2$ semiparcial** são derivados do estágio ou passo anterior, tomando as diferenças de valores sucessivos de $\boldsymbol{R_q^2}$ (o primeiro valor é $1 - \boldsymbol{R_q^2}$).

Exemplo: Consideramos os dados brutos de seis pacientes com quatro medidas em cada (Tabela 45) e resultado do agrupamento pelo método de Ward (Tabela 46).

**Tabela 45.** Dados de seis pacientes com as medidas $x_1$, $x_2$, $x_3$ e $x_4$ (Everitt, 1989)

| Paciente | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| **1** | 39,8 | 38,0 | 22,2 | 23,2 |
| **2** | 53,7 | 37,2 | 18,7 | 18,5 |
| **3** | 47,3 | 39,8 | 23,3 | 22,1 |
| **4** | 41,7 | 37,6 | 22,8 | 22,3 |
| **5** | 44,7 | 38,5 | 24,8 | 24,4 |
| **6** | 47,9 | 39,8 | 22,0 | 23,3 |

**Tabela 46.** Resultado da análise de agrupamento pelo método de Ward (Khattree; Nair, 2002)

| Estágio (**q**) | Grupos | N Grupos |
|---|---|---|
| 1 | {1},{2},{4},{5},{3, 6} | 5 |
| 2 | {2},{5},{3, 6},{1, 4} | 4 |
| 3 | {2},{1, 4},{3, 6, 5} | 3 |
| 4 | {2},{1, 4, 3, 6, 5} | 2 |
| 5 | {1, 4, 3, 6, 5, 2} | 1 |

Estágio (q = 1)

| Estágio | Grupos | N Grupos |
|---|---|---|
| 1 | {1},{2},{4},{5},{3, 6} | 5 |

$$\bar{X}_{1(3,6)} = \left(\sum_{i=1}^{n} X_{1i}\right)/n = (47{,}3 + 47{,}9)/2 = 47{,}6$$

$$\bar{X}_{2(3,6)} = \left(\sum_{i=1}^{n} X_{2i}\right)/n = (39{,}8 + 39{,}8)/2 = 39{,}8$$

$$\bar{X}_{3(3,6)} = \left(\sum_{i=1}^{n} X_{3i}\right)/n = (23{,}3 + 22{,}0)/2 = 22{,}65$$

$$\bar{X}_{4(3,6)} = \left(\sum_{i=1}^{n} X_{4i}\right)/n = (22{,}1 + 23{,}3)/2 = 22{,}7$$

## Soma dos Quadrados dentro e entre Grupos

$$SQ_{ij} = \sum_{\substack{i=1 \\ j=1}}^{p,n} \left(X_{ij} - \bar{X}_i\right)^2$$

$$SQ_A = \sum_{i=1}^{n_A} (X_i - \bar{X}_A)^2$$

$$SQ_B = \sum_{i=1}^{n_B} (X_i - \bar{X}_B)^2$$

**Matriz de Soma de Quadrados de AB**

$$S_{AB} = \sum_{\substack{i=1 \\ j=1}}^{p,n_{AB}} \left(X_{ij} - \bar{X}_{i(AB)}\right)^2$$

$$S_{AB}^2 = S_{AB}/(n_{AB} - 1)$$

Em que: $\bar{\boldsymbol{X}}_{\boldsymbol{AB}} = (n_A\bar{X}_A + n_B\bar{X}_B)/(n_A + n_B)$; $\boldsymbol{n_A}$, $\boldsymbol{n_B}$ e $\boldsymbol{n_{AB}} = n_A + n_B$ são os números de pontos da **i**-ésima variável (**i** = 1, ..., **p**) no **j**-ésimo paciente (**j** = 1, ..., **n**) em **A**, **B** e **AB** respectivamente.

**Incremento da soma de quadrados ($\boldsymbol{I_{AB}}$)**

$$I_{AB} = SQ_{AB} - SQ_A - SQ_B$$

Assim:

$$SQ_{36} = \left[\left\{\left(X_{13} - \bar{X}_{1(36)}\right)^2 + \left(X_{16} - \bar{X}_{1(36)}\right)^2\right\} + \left\{\left(X_{23} - \bar{X}_{2(36)}\right)^2 + \left(X_{26} - \bar{X}_{2(36)}\right)^2\right\} + \left\{\left(X_{33} - \bar{X}_{3(36)}\right)^2 + \left(X_{36} - \bar{X}_{3(36)}\right)^2\right\} + \left\{\left(X_{43} - \bar{X}_{4(36)}\right)^2 + \left(X_{46} - \bar{X}_{4(36)}\right)^2\right\}\right]$$

$$SQ_{36} = [\{(47{,}3 - 47{,}6)^2 + (47{,}9 - 47{,}6)^2\} + \{(39{,}8 - 39{,}8)^2 + (39{,}8 - 39{,}8)^2\} + \{(23{,}3 - 22{,}65)^2 + (22{,}0 - 22{,}65)^2\} + \{(22{,}1 - 22{,}7)^2 + (23{,}3 - 22{,}7)^2\}]$$

$$SQ_{36} = 0{,}18 + 0{,}00 + 0{,}845 + 0{,}72 = 1{,}7450$$

$$SQ_{36}^2 = S_{36}/(n_{36} - 1) = 1{,}7450/(2 - 1) = 1{,}7450$$

**Incremento da soma de quadrados ($\boldsymbol{I_{AB}}$)**

$$I_{36} = SQ_{36} - SQ_3 - SQ_6 = 1{,}7450 - 0 - 0 = 1{,}7450$$

**Desvio padrão quadrático médio (RMSSTD)**

$$RMSSTD_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s_{qj}^2} = \sqrt{\frac{1{,}7450}{4}} = 0{,}6605$$

**Coeficiente de determinação ($R^2$) para o grupo Bq**

$$R_{36}^2 = 1 - \frac{I_{36}}{SQTotal} = 1 - \frac{1{,}7450}{170{,}463} = 0{,}9898$$

**$R^2$ semiparcial**

$$R_{36}^2 semiparcial = 1 - R_{36}^2 = 1 - 0{,}9898 = 0{,}0102$$

Estágio (q = 2)

| Estágio | Grupos | N Grupos |
|---|---|---|
| 2 | {2},{5},{3, 6},{1, 4} | 4 |

$$\bar{X}_{1(1,4)} = \left(\sum_{i=1}^{n} X_{1i}\right)/n = (39{,}8 + 41{,}7)/2 = 40{,}75$$

$$\bar{X}_{2(1,4)} = \left(\sum_{i=1}^{n} X_{2i}\right)/n = (38{,}0 + 37{,}6)/2 = 37{,}8$$

$$\bar{X}_{3(1,4)} = \left(\sum_{i=1}^{n} X_{3i}\right)/n = (22{,}2 + 22{,}8)/2 = 22{,}5$$

$$\bar{X}_{4(1,4)} = \left(\sum_{i=1}^{n} X_{4i}\right)/n = (23{,}2 + 22{,}3)/2 = 22{,}75$$

Assim:

$$SQ_{14} = \left[\left\{(X_{11} - \bar{X}_{1(14)})^2 + (X_{14} - \bar{X}_{1(36)})^2\right\} + \left\{(X_{21} - \bar{X}_{2(14)})^2 + (X_{24} - \bar{X}_{2(14)})^2\right\} + \left\{(X_{31} - \bar{X}_{3(14)})^2 + (X_{34} - \bar{X}_{3(14)})^2\right\} + \left\{(X_{41} - \bar{X}_{4(14)})^2 + (X_{44} - \bar{X}_{4(14)})^2\right\}\right]$$

$$SQ_{14} = [\{(39{,}8 - 40{,}75)^2 + (41{,}7 - 40{,}75)^2\} + \{(38{,}0 - 37{,}8)^2 + (37{,}6 - 37{,}8)^2\} + \{(22{,}2 - 22{,}5)^2 + (22{,}8 - 22{,}5)^2\} + \{(23{,}2 - 22{,}75)^2 + (22{,}3 - 22{,}75)^2\}]$$

$$SQ_{14} = 1{,}805 + 0{,}080 + 0{,}180 + 0{,}405 = 2{,}4700$$

$$S_{14}^2 = S_{14}/(n_{12} - 1) = 2{,}47/(2 - 1) = 2{,}47$$

**Incremento da soma de quadrados ($I_{AB}$)**

$$I_{14} = SQ_{14} - SQ_1 - SQ_4 = 2{,}47 - 0 - 0 = 2{,}47$$

**Desvio padrão quadrático médio (RMSSTD)**

$$RMSSTD_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s_{qj}^2} = \sqrt{\frac{2{,}47}{4}} = 0{,}7858$$

**Coeficiente de determinação ($R^2$) para o grupo Bq**

$$R_{14}^2 = 1 - \frac{I_{14}}{SQTotal} = 1 - \frac{2{,}47}{170{,}463} = 0{,}9573$$

**$R^2$ semiparcial**

$$R^2_{14}parcial = R^2_{36} - R^2_{14} = 0{,}9898 - 0{,}9573 = 0{,}0145$$

Estágio (q = 3)

| Estágio | Grupos | N Grupos |
|---|---|---|
| 3 | {2},{1, 4},{3, 6, 5} | 3 |

$$\bar{X}_{1(3,6,5)} = \left(\sum_{i=1}^{n} X_{1i}\right)/n = (47{,}3 + 47{,}9 + 44{,}7)/3 = 46{,}63$$

$$\bar{X}_{2(3,6,5)} = \left(\sum_{i=1}^{n} X_{2i}\right)/n = (39{,}8 + 39{,}8 + 38{,}5)/3 = 39{,}37$$

$$\bar{X}_{3(3,6,5)} = \left(\sum_{i=1}^{n} X_{3i}\right)/n = (23{,}3 + 22{,}0 + 24{,}8)/3 = 23{,}37$$

$$\bar{X}_{4(3,6,5)} = \left(\sum_{i=1}^{n} X_{4i}\right)/n = (22{,}1 + 23{,}3 + 24{,}4)/3 = 23{,}27$$

Assim:

$$SQ_{365} = \left[\left\{(X_{13} - \bar{X}_{1(365)})^2 + (X_{16} - \bar{X}_{1(365)})^2 + (X_{15} - \bar{X}_{1(365)})^2\right\} + \left\{(X_{23} - \bar{X}_{2(365)})^2 + (X_{26} - \bar{X}_{2(365)})^2 + (X_{25} - \bar{X}_{2(365)})^2\right\} + \left\{(X_{33} - \bar{X}_{3(365)})^2 + (X_{36} - \bar{X}_{3(365)})^2 + (X_{35} - \bar{X}_{3(365)})^2\right\} + \left\{(X_{43} - \bar{X}_{4(365)})^2 + (X_{46} - \bar{X}_{4(365)})^2 + (X_{45} - \bar{X}_{4(365)})^2\right\}\right]$$

$$SQ_{365} = [\{(47{,}3 - 46{,}63)^2 + (47{,}9 - 46{,}63)^2 + (44{,}7 - 46{,}63)^2\} + \{(39{,}8 - 39{,}37)^2 + (39{,}8 - 39{,}37)^2 + (38{,}5 - 39{,}37)^2\} + \{(23{,}3 - 23{,}37)^2 + (22{,}0 - 23{,}37)^2 + (24{,}8 - 23{,}37)^2\} + \{(22{,}1 - 23{,}27)^2 + (23{,}3 - 23{,}27)^2 + (24{,}4 - 23{,}27)^2\}]$$

$$SQ_{365} = 5{,}787 + 1{,}127 + 3{,}927 + 2{,}647 = 13{,}4867$$

$$S^2_{365} = S_{365}/(n_{365} - 1) = 13{,}4867/(3 - 1) = 6{,}7433$$

**Incremento da soma de quadrados ($I_{AB}$)**

$$I_{365} = SQ_{365} - SQ_{36} - SQ_5 = 13{,}4867 - 1{,}7450 - 0 = 11{,}7417$$

**Desvio padrão quadrático médio (RMSSTD)**

$$RMSSTD_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s^2_{qj}} = \sqrt{\frac{6{,}7433}{4}} = 1{,}2984$$

**Coeficiente de determinação ($R^2$) para o grupo Bq**

$$R^2_{365} = 1 - \frac{I_{365}}{SQTotal} = 1 - \frac{11{,}7417}{170{,}463} = 0{,}9064$$

**$R^2$ semiparcial**

$$R^2_{365}parcial = R^2_{14} - R^2_{365} = 0{,}9753 - 0{,}9064 = 0{,}0689$$

Estágio (q = 4)

| Estágio | Grupos | N Grupos |
|---|---|---|
| 4 | {2},{1, 4, 3, 6, 5} | 2 |

$$\bar{X}_{1(1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{1i}\right)\Big/ n = (39{,}8 + 41{,}7 + 47{,}3 + 47{,}9 + 44{,}7)/5 = 44{,}43$$

$$\bar{X}_{2(1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{2i}\right)\Big/ n = (38{,}0 + 37{,}6 + 39{,}8 + 39{,}8 + 38{,}5)/5 = 38{,}74$$

$$\bar{X}_{3(1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{3i}\right)\Big/ n = (22{,}2 + 22{,}8 + 23{,}3 + 22{,}0 + 24{,}8)/5 = 23{,}02$$

$$\bar{X}_{4(1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{4i}\right)\Big/ n = (23{,}2 + 22{,}3 + 22{,}1 + 23{,}3 + 24{,}4)/5 = 23{,}06$$

Assim:

$$SQ_{14365} = \left[\left\{\left(X_{11} - \bar{X}_{1(14365)}\right)^2 + \left(X_{14} - \bar{X}_{1(14365)}\right)^2 + \left(X_{13} - \bar{X}_{1(14365)}\right)^2 + \left(X_{16} - \bar{X}_{1(14365)}\right)^2 + \left(X_{15} - \bar{X}_{1(14365)}\right)^2\right\} + \left\{\left(X_{21} - \bar{X}_{2(14365)}\right)^2 + \left(X_{24} - \bar{X}_{2(14365)}\right)^2 + \left(X_{23} - \bar{X}_{2(14365)}\right)^2 + \left(X_{26} - \bar{X}_{2(14365)}\right)^2 + \left(X_{25} - \bar{X}_{2(14365)}\right)^2\right\} + \left\{\left(X_{31} - \bar{X}_{3(14365)}\right)^2 + \left(X_{34} - \bar{X}_{3(14365)}\right)^2 + \left(X_{33} - \bar{X}_{3(14365)}\right)^2 + \left(X_{36} - \bar{X}_{3(14365)}\right)^2 + \left(X_{35} - \bar{X}_{3(14365)}\right)^2\right\} + \left\{\left(X_{41} - \bar{X}_{4(14365)}\right)^2 + \left(X_{44} - \bar{X}_{4(14365)}\right)^2 + \left(X_{43} - \bar{X}_{4(14365)}\right)^2 + \left(X_{46} - \bar{X}_{4(14365)}\right)^2 + \left(X_{45} - \bar{X}_{4(14365)}\right)^2\right\}\right]$$

$$SQ_{14365} = [\{(39{,}8 - 44{,}43)^2 + (41{,}7 - 44{,}43)^2 + (47{,}3 - 44{,}43)^2\} + \{(47{,}9 - 44{,}43)^2 + (44{,}7 - 44{,}43)^2 + (38{,}0 - 38{,}74)^2\} + \{(37{,}6 - 38{,}74)^2 + (39{,}8 - 38{,}74)^2 + (39{,}8 - 38{,}74)^2\} + \{(38{,}5 - 38{,}74)^2 + (22{,}2 - 23{,}02)^2 + (22{,}8 - 23{,}02)^2 + (23{,}3 - 23{,}02)^2 + (22{,}0 - 23{,}02)^2 + (24{,}8 - 23{,}02)^2\} + \{(23{,}2 - 23{,}06)^2 + (22{,}3 - 23{,}06)^2 + (22{,}1 - 23{,}06)^2 + (23{,}3 - 23{,}06)^2 + (24{,}4 - 23{,}06)^2\}]$$

$SQ_{14365} = 49{,}128 + 4{,}152 + 5{,}008 + 3{,}372 = 61{,}600$

$S^2_{14365} = SQ_{14365}/(n_{14365} - 1) = 61{,}600/(5 - 1) = 15{,}415$

**Incremento da soma de quadrados ($I_{AB}$)**

$$I_{14365} = SQ_{14365} - SQ_{1,4} - SQ_{365} = 61{,}6000 - 2{,}4700 - 13{,}4867 = 45{,}7033$$

**Desvio padrão quadrático médio (RMSSTD)**

$$RMSSTD_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s^2_{qj}} = \sqrt{\frac{15{,}415}{4}} = 1{,}9631$$

**Coeficiente de determinação ($R^2$) para o grupo Bq**

$$R^2_{14365} = 1 - \frac{I_{14365}}{SQTotal} = 1 - \frac{45{,}7033}{170{,}463} = 0{,}6383$$

**$R^2$ semiparcial**

$$R^2_{14365}parcial = R^2_{14536} - R^2_{365} = 0{,}9064 - 0{,}6383 = 0{,}2681$$

Estágio (q = 5)

| Estágio | Grupos | N Grupos |
|---|---|---|
| 5 | {1, 4, 3, 6, 5, 2} | 1 |

$$\bar{X}_{1(2,1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{1i}\right)\Big/n = (53{,}7 + 39{,}8 + 41{,}7 + 47{,}3 + 47{,}9 + 44{,}7)/6 = 45{,}85$$

$$\bar{X}_{2(2,1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{2i}\right)\Big/n = (37{,}2 + 38{,}0 + 37{,}6 + 39{,}8 + 39{,}8 + 38{,}5)/6 = 38{,}48$$

$$\bar{X}_{3(2,1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{3i}\right)\Big/n = (18{,}7 + 22{,}2 + 22{,}8 + 23{,}3 + 22{,}0 + 24{,}8)/6 = 22{,}30$$

$$\bar{X}_{4(2,1,4,3,6,5)} = \left(\sum_{i=1}^{p} X_{4i}\right)\Big/n = (18{,}5 + 23{,}2 + 22{,}3 + 22{,}1 + 23{,}3 + 24{,}4)/6 = 22{,}30$$

Assim:

$$SQ_{214365} = \left[\left\{\left(X_{12} - \bar{X}_{1(214365)}\right)^2 + \left(X_{11} - \bar{X}_{1(214365)}\right)^2 + \left(X_{14} - \bar{X}_{1(214365)}\right)^2 + \left(X_{13} - \bar{X}_{1(214365)}\right)^2 + \left(X_{16} - \bar{X}_{1(214365)}\right)^2 + \left(X_{15} - \bar{X}_{1(214365)}\right)^2\right\} + \left\{\left(X_{22} - \bar{X}_{2(214365)}\right)^2 + \left(X_{21} - \bar{X}_{2(214365)}\right)^2 + \left(X_{24} - \bar{X}_{2(214365)}\right)^2 + \left(X_{23} - \bar{X}_{2(214365)}\right)^2 + \left(X_{26} - \bar{X}_{2(214365)}\right)^2 + \left(X_{25} - \bar{X}_{2(214365)}\right)^2\right\} + \left\{\left(X_{32} - \bar{X}_{3(214365)}\right)^2 + \left(X_{31} - \bar{X}_{3(214365)}\right)^2 + \left(X_{34} - \bar{X}_{3(214365)}\right)^2 + \left(X_{33} - \bar{X}_{3(214365)}\right)^2 + \left(X_{36} - \bar{X}_{3(214365)}\right)^2 + \left(X_{35} - \bar{X}_{3(214365)}\right)^2\right\} + \left\{\left(X_{42} - \bar{X}_{4(214365)}\right)^2 + \left(X_{41} - \bar{X}_{4(214365)}\right)^2 + \left(X_{44} - \bar{X}_{4(214365)}\right)^2 + \left(X_{43} - \bar{X}_{4(214365)}\right)^2 + \left(X_{46} - \bar{X}_{4(214365)}\right)^2 + \left(X_{45} - \bar{X}_{4(214365)}\right)^2\right\}\right]$$

$$SQ_{214365} = [\{(53{,}7 - 45{,}85)^2 + (39{,}8 - 45{,}85)^2 + (41{,}7 - 45{,}85)^2 + (47{,}3 - 45{,}85)^2 + (47{,}9 - 45{,}85)^2 + (44{,}7 - 45{,}85)^2\} + \{(37{,}2 - 38{,}48)^2 + (38{,}0 - 38{,}48)^2 + (37{,}6 - 38{,}48)^2\} + \{(39{,}8 - 38{,}48)^2 + (39{,}8 - 38{,}48)^2 + (38{,}5 - 38{,}48)^2\} + \{(18{,}7 - 22{,}3)^2 + (22{,}2 - 22{,}3)^2 + (22{,}8 - 22{,}3)^2 + (23{,}3 - 22{,}3)^2 + (22{,}0 - 22{,}3)^2 + (24{,}8 - 22{,}3)^2\} + \{(18{,}5 - 22{,}3)^2 + (22{,}2 - 22{,}3)^2 + (22{,}3 - 22{,}3)^2 + (22{,}1 - 22{,}3)^2 + (23{,}3 - 22{,}3)^2 + (24{,}4 - 22{,}3)^2\}]$$

$$SQ_{214365} = 123{,}075 + 6{,}128 + 20{,}560 + 20{,}700 = 170{,}463$$

$$S^2_{214365} = SQ_{214365}/(n_{214365} - 1) = 170{,}463/(6 - 1) = 34{,}093$$

**Incremento da soma de quadrados ($I_{AB}$)**

$$I_{214365} = SQ_{214365} - SQ_{2,2} - SQ_{14365} = 170{,}463 - 0{,}000 - 61{,}660 = 108{,}803$$

**Desvio padrão quadrático médio (RMSSTD)**

$$RMSSTD_q = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s^2_{qj}} = \sqrt{\frac{34{,}093}{4}} = 2{,}9194$$

**Coeficiente de determinação ($R^2$) para o grupo Bq**

$$R^2_{214365} = 1 - \frac{SQ_{214365}}{SQTotal} = 1 - \frac{170{,}463}{170{,}463} = 0{,}0000$$

**$R^2$semiparcial**

$$R^2_{214365}parcial = 1 - R^2_{214365} - R^2_{14365} - R^2_{365} - R^2_{14} - -R^2_{36} = 1 - (0{,}0000 + 0{,}2681 + 0{,}0689 + 0{,}0102 + 0{,}0145) = 0{,}6383$$

Os resultados obtidos neste exemplo estão resumidos na Tabela 47.

**Tabela 47.** Resumo das estatísticas de Incremento da soma de quadrados ($I_{AB}$), desvio padrão quadrático médio (RMSSTD), $R^2$parcial (RSQ) e $R^2$semiparcial (SPRSQ)

| Passos | Grupos | Número de Grupos | $I_{AB}$ | RMSSTD | RSQ | SPRSQ |
|---|---|---|---|---|---|---|
| 1 | {1}, {2}, {4}, {5}, {3, 6} | 5 | 1,7450 | 0,6605 | 0,9898 | 0,0102 |
| 2 | {1, 4}, {2}, {5}, {3, 6} | 4 | 2,4700 | 0,7858 | 0,9753 | 0,0145 |
| 3 | {1, 4}, {2}, {3, 6, 5} | 3 | 11,7417 | 1,2984 | 0,9064 | 0,0689 |
| 4 | {1, 4, 3, 6, 5}, {2} | 2 | 45,7033 | 1,9631 | 0,6383 | 0,2681 |
| 5 | {1, 4, 3, 6, 5, 2} | 1 | 108,8033 | 2,9194 | 0,0000 | 0,6383 |

**Passo a Passo: Validação de Khattree e Nair (2002)**

**1. Preparação dos Dados e Agrupamento Hierárquico**

Primeiro, definimos o *dataframe* pacientes e, em seguida, calculamos a matriz de distâncias entre os pacientes. Com base nessa matriz, realizamos o agrupamento hierárquico, que servirá de base para a validação.

```
# 1. Definir o dataframe de exemplo 'pacientes'
pacientes <- data.frame(
  x1 = c(39.8, 53.7, 47.3, 41.7, 44.7, 47.9),
  x2 = c(38.0, 37.2, 39.8, 37.6, 38.5, 39.8),
  x3 = c(22.2, 18.7, 23.3, 22.8, 24.8, 22.0),
  x4 = c(23.2, 18.5, 22.1, 22.3, 24.4, 23.3)
)

# Atribuir nomes às linhas para facilitar a identificação
rownames(pacientes) <- paste0("P",1:6)

# 2. Calcular a matriz de distâncias (ex: euclidiana) entre os pacientes
distancia_pacientes <- dist(pacientes, method = "euclidean")

# 3. Realizar o agrupamento hierárquico utilizando o método Ward.D2
fit_ward <- hclust(distancia_pacientes, method = "ward.D2")
```

**2. Cálculo da Soma Total de Quadrados (SST)**

A Soma Total de Quadrados (SST) representa a variabilidade total dos dados. É um valor fixo que serve como referência para o cálculo do $R^2$.

```
# Calcular a Soma Total de Quadrados (SST)
# A função scale(..., scale = FALSE) centraliza os dados, e a soma dos quadrados é a SST
SST <- sum(scale(pacientes, scale = FALSE)^2)
round(SST, 2)
```

170,46

### 3. Iteração e Cálculo das Métricas para Diferentes Números de Grupos

Agora, vamos iterar sobre um número decrescente de grupos (k) e calcular o SSW (Soma de Quadrados Dentro dos Grupos), o RMSSTD e o $R^2$ para cada k.

```
# Definir o número de grupos a serem testados (de 5 a 1)
k_valores <- 5:1

# Criar um dataframe para armazenar os resultados
resultados_validacao <- data.frame(
  K = k_valores,
  SSW = NA,
  RMSSTD = NA,
  R2 = NA
)

# Loop para calcular as métricas para cada número de grupos (k)
for(i in seq_along(k_valores)){

  # Cortar o dendrograma para obter os grupos para o k atual
  grupos_atuais <- cutree(fit_ward, k = k_valores[i])

  # Inicializar a Soma de Quadrados Dentro dos Grupos (SSW) para o k atual
  SSW_atual <- 0

  # Calcular SSW para cada grupo
  for(g in unique(grupos_atuais)){

    # Selecionar os dados pertencentes ao grupo 'g'
    dados_grupo <- pacientes[grupos_atuais == g, , drop = FALSE]

    # Se o grupo tiver mais de uma observação, calcular a soma dos quadrados
    if(nrow(dados_grupo) > 1){
      SSW_atual <- SSW_atual + sum(scale(dados_grupo, scale = FALSE)^2)
    }
  }

  # Armazenar o SSW calculado
  resultados_validacao$SSW[i] <- SSW_atual

  # Calcular RMSSTD
  # ncol(pacientes) é o número de variáveis (x1, x2, x3, x4)
  # nrow(pacientes) é o número de observações (P1 a P6)
  resultados_validacao$RMSSTD[i] <-
    sqrt(
      SSW_atual /
        (ncol(pacientes) * (nrow(pacientes) - k_valores[i]))
    )

  # Calcular R²
```

```
  resultados_validacao$R2[i] <- 1 - (SSW_atual / SST)
}

# Exibir os resultados parciais
print(round(resultados_validacao, 4))
```

| K | SSW | RMSSTD | R² |
|---|---|---|---|
| 5 | 1.7450 | 0.6605 | 0.9898 |
| 4 | 4.2150 | 0.7259 | 0.9753 |
| 3 | 15.9567 | 1.1531 | 0.9064 |
| 2 | 61.6600 | 1.9631 | 0.6383 |
| 1 | 170.4633 | 2.9194 | 0.0000 |

**4. Cálculo do R² Semiparcial (SPR2)**

O R² semiparcial (SPR2) mede o aumento na proporção da variância explicada ao passar de k+1 grupos para k grupos. Ele ajuda a identificar o ponto onde a adição de um novo grupo não contribui significativamente para a explicação da variância.

```
# Calcular o R² semiparcial (SPR2)
# O primeiro valor de SPR2 é 1 - R² do maior número de grupos
# Os valores subsequentes são a diferença entre os R² consecutivos
resultados_validacao$SPR2 <- c(
  1 - resultados_validacao$R2[1], # Para o maior k (5 grupos)
  -diff(resultados_validacao$R2)   # Diferença entre R² para k e k-1 grupos
)

# Exibir os resultados finais com o SPR2
print(round(resultados_validacao, 4))
```

| K | SSW | RMSSTD | R² | SPR² |
|---|---|---|---|---|
| 5 | 1.7450 | 0.6605 | 0.9898 | 0.0102 |
| 4 | 4.2150 | 0.7259 | 0.9753 | 0.0145 |
| 3 | 15.9567 | 1.1531 | 0.9064 | 0.0689 |
| 2 | 61.6600 | 1.9631 | 0.6383 | 0.2681 |
| 1 | 170.4633 | 2.9194 | 0.0000 | 0.6383 |

- **Método de Mojena (1977)**

Para determinar a *linha fenon* em dendrogramas gerados por meio de técnicas de agrupamento hierárquico e definir o número de grupos, Mojena (1977) sugeriu se basear no tamanho relativo dos níveis de fusão (distância):

a) Selecionar o número de grupos no j-ésimo **passo** que, primeiramente, satisfizer $\boldsymbol{a_j} > \boldsymbol{q_k}$ , em que: $\boldsymbol{a_j}$= distância do nível de fusão correspondente ao **j**-ésimo passo (**j** = 1, 2, ..., g - 1); $\boldsymbol{q_k}$ = valor referencial de corte, dado por: $q_k = \bar{\alpha} + k\hat{\sigma}_\alpha$; $\bar{\boldsymbol{\alpha}}$ = média dos valores de **α**; $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\alpha}}$ = desvio padrão não viesado de **α**; **k** = constante definida como regra de parada na definição do

número de grupos, sugerido por Mojena (1977) entre 2,75, e 3,50 e igual a 1,25 por Milligan e Cooper (1985).

$$\bar{\alpha} = \frac{1}{g-1}\sum_{j=1}^{g-1}\alpha_j$$

$$\hat{\sigma}_\alpha = \sqrt{\frac{\sum_{j=1}^{g-1}\alpha_j^2 - \frac{1}{g-1}\left(\sum_{j=1}^{g-1}\alpha_j\right)^2}{g-2}}$$

Como exemplo, consideremos os resultados obtidos no **método de ligação simples** (Tabela 48).

**Tabela 48.** Níveis de fusão (distância euclidiana) das parcelas com o emprego do **método de ligação simples**, com base na **distância euclidiana**

| Passos | Níveis de Fusão ($a$) | Parcelas | g |
|---|---|---|---|
| 1 | 7,7 | P4, P8 | 2 |
| 2 | 8,8 | P7, P11 | 2 |
| 3 | 11,3 | P1, P4, P8 | 3 |
| 4 | 12,0 | P1, P4, P8, P2 | 4 |
| 5 | 13,3 | P1, P4, P8, P2, P7, P11 | 6 |
| 6 | 17,3 | P1, P4, P8, P2, P7, P11, P6 | 7 |
| 7 | 17,5 | P1, P4, P8, P2, P7, P11, P6, P10 | 8 |
| 8 | 17,9 | P1, P4, P8, P2, P7, P11, P6, P10, P3 | 9 |
| 9 | 19,2 | P1, P4, P8, P2, P7, P11, P6, P10, P3, P9 | 10 |
| 10 | 28,7 | P1, P4, P8, P2, P7, P11, P6, P10, P3, P9, P5 | 11 |

Para o **Passo 3**

$$\bar{\alpha} = \frac{1}{3-1}(7{,}7 + 11{,}3) = \frac{19}{2} = 9{,}5$$

$$\hat{\sigma}_\alpha = \sqrt{\frac{(7{,}7^2 + 11{,}3^2) - \frac{1}{3-1}(7{,}7 + 11{,}3)^2}{3-2}} = \sqrt{\frac{186{,}98 - \frac{1}{2}(19)^2}{1}} = 2{,}55$$

Para **k** = 1,25 (Milligan e Cooper (1985), $\boldsymbol{q_k} = \bar{\alpha} + k\hat{\sigma}_\alpha = 9{,}5 + 1{,}25x2{,}55 = 12{,}68$

O nível de fusão no **passo 3** ($\alpha$ = 11,3) é inferior ao critério $\mathbf{q_k}$ = 12,68, então, no **passo 4**, teremos:

$$\bar{\alpha} = \frac{1}{4-1}(7{,}7 + 11{,}3 + 12) = \frac{31}{3} = 10{,}33$$

$$\hat{\sigma}_\alpha = \sqrt{\frac{(7{,}7^2 + 11{,}3^2 + 12^2) - \frac{1}{4-1}(7{,}7 + 11{,}3 + 12)^2}{4-2}} = \sqrt{\frac{330{,}98 - \frac{1}{3}(31)^2}{2}} = 2{,}31$$

$$q_k = \bar{\alpha} + k\hat{\sigma}_\alpha = 10{,}33 + 1{,}25x2{,}31 = 13{,}22$$

O nível de fusão no **passo 5** ($\alpha = 12$) é inferior ao critério $\mathbf{q_k} = 13{,}22$, então, **no passo 6**, teremos:

$$\bar{\alpha} = \frac{1}{6-1}(7{,}7 + 8{,}8 + 11{,}3 + 12 + 13{,}3) = \frac{53{,}1}{5} = 10{,}62$$

$$\hat{\sigma}_\alpha = \sqrt{\frac{(7{,}7^2+8{,}8^2+11{,}3^2+12^2+13{,}3^2)-\frac{1}{6-1}(7{,}7+8{,}8+11{,}3+12+13{,}3)^2}{6-2}} = \sqrt{\frac{585{,}31-\frac{1}{5}(53{,}1)^2}{4}} = 2{,}31$$

$$q_k = \bar{\alpha} + k\hat{\sigma}_\alpha = 10{,}62 + 1{,}25x2{,}31 = 13{,}51$$

O nível de fusão no **passo 6** ($\alpha = 13{,}3$) é inferior ao critério $q_k = 13{,}51$, então, no **passo 7**, teremos:

$$\bar{\alpha} = \frac{1}{7-1}(7{,}7 + 8{,}8 + 11{,}3 + 12 + 13{,}3 + 17{,}3) = \frac{70{,}4}{6} = 11{,}73$$

$$\hat{\sigma}_\alpha = \sqrt{\frac{(7{,}7^2+8{,}8^2+11{,}3^2+12^2+13{,}3^2+17{,}3^2)-\frac{1}{7-1}(7{,}7+8{,}8+11{,}3+12+13{,}3+17{,}3)^2}{7-2}} = \sqrt{\frac{884{,}6-\frac{1}{6}(70{,}4)^2}{5}} = 3{,}42$$

$$q_k = \bar{\alpha} + k\hat{\sigma}_\alpha = 11{,}73 + 1{,}25x3{,}42 = 16{,}01$$

O nível de fusão no **passo 7** ($\alpha = 17{,}3$) é **superior ao critério $\mathbf{q_k} = 16{,}01$**. Assim, deve-se parar **nesse passo**, ou seja, ficamos com **seis** grupos (Figura 23, saída R): grupo 1 (Parcelas 1, 2, 4, 7, 8 e 11); grupo 2 (Parcela 6); grupo 3 (Parcela 10); grupo 4 (Parcela 3); grupo 5 (Parcela 9); e grupo 6 (Parcela 5).

**Passo a passo no R: Método de Mojena**

**1. Gerando o Dendrograma Base e Obtendo os Níveis de Fusão**

Primeiro, precisamos calcular a matriz de distâncias a partir da nossa matriz de dados X (transposta) e, em seguida, realizar o agrupamento hierárquico. A partir do objeto hclust, extraímos os níveis de fusão, que são as alturas em que as uniões ocorrem no dendrograma.

```
# 1. Transpor a matriz X para que as parcelas fiquem nas linhas
X_parcelas <- t(X)

# 2. Calcular a matriz de distâncias (ex: euclidiana) entre as parcelas
D_euclidiana <- dist(X_parcelas, method = "euclidean")
```

```
# 3. Realizar o agrupamento hierárquico pelo método de ligação simples
fit_simples <- hclust(D_euclidiana, method = "single")

# 4. Obter os níveis de fusão (alturas) do dendrograma
niveis_fusao <- fit_simples$height

# Visualizar os níveis de fusão arredondados
print(round(niveis_fusao, 2))
```

```
 7.75  8.77 11.27 12.00 13.27 17.29 17.55 17.92 19.24 28.65
```

## 2. Implementação do Cálculo do Critério de Mojena

Agora, vamos implementar o cálculo do critério de Mojena passo a passo. Iremos iterar sobre os níveis de fusão, calculando a média e o desvio padrão dos níveis de fusão anteriores para determinar o critério $Q_k$. A constante k_mojena é um fator de ajuste.

```
# 1. Definir a constante de Mojena (geralmente entre 1.25 e 1.50)
k_mojena <- 1.25

# 2. Criar um dataframe para armazenar os cálculos
tabela_mojena <- data.frame(
  Passo = 1:length(niveis_fusao),
  Alpha_j = round(niveis_fusao, 2),
  Media_Alpha = NA,
  Desvio_Alpha = NA,
  Qk_Criterio = NA,
  Decisao = NA
)

# 3. Loop para calcular as métricas e aplicar o critério de Mojena
# O loop começa do segundo nível de fusão, pois precisamos de pelo menos um anterior para calcular média/desvio
for(j in 2:(length(niveis_fusao) - 1)) {

  # Níveis de fusão anteriores até o passo j
  alphas_anteriores <- niveis_fusao[1:j]

  # Média e desvio padrão dos níveis de fusão anteriores
  media <- mean(alphas_anteriores)
  desvio <- sd(alphas_anteriores)

  # Cálculo do critério Qk
  qk <- media + (k_mojena * desvio)

  # Armazenar os resultados na tabela
  tabela_mojena$Media_Alpha[j] <- round(media, 4)
  tabela_mojena$Desvio_Alpha[j] <- round(desvio, 4)
  tabela_mojena$Qk_Criterio[j] <- round(qk, 4)

  # Comparar o próximo nível de fusão com o critério Qk do passo atual
```

```
  proximo_alpha <- niveis_fusao[j+1]
  if(proximo_alpha > qk) {
    tabela_mojena$Decisao[j] <- "PARAR AQUI"
  } else {
    tabela_mojena$Decisao[j] <- "CONTINUAR"
  }
}

# 4. Exibir a tabela de cálculos de Mojena
# Excluímos a primeira linha (Passo 1) pois não há cálculo de média/desvio para ela
print(tabela_mojena[2:length(niveis_fusao), c("Passo", "Alpha_j", "Media_Alpha",
"Desvio_Alpha", "Qk_Criterio", "Decisao")])
```

| Passo | Alpha_j | Média Alpha | Desvio Alpha | Qk_Critério | Decisão |
|---|---|---|---|---|---|
| 2 | 8.77 | 8.2281 | 0.7734 | 9.1949 | PARAR AQUI |
| 3 | 11.27 | 9.2418 | 1.8391 | 11.5408 | PARAR AQUI |
| 4 | 11.96 | 9.9209 | 2.0248 | 12.4519 | PARAR AQUI |
| 5 | 13.27 | 10.5901 | 2.3051 | 13.4714 | PARAR AQUI |
| 6 | 17.26 | 11.7022 | 3.4163 | 15.9726 | PARAR AQUI |
| 7 | 17.55 | 12.5376 | 3.8225 | 17.3156 | PARAR AQUI |
| 8 | 17.94 | 13.2134 | 4.0222 | 18.2412 | PARAR AQUI |
| 9 | 19.24 | 13.8825 | 4.2644 | 19.2130 | PARAR AQUI |
| 10 | 28.65 | — | — | — | — |

### 3. Visualização: Dendrograma com Linha de Corte de Mojena

Com base na tabela de Mojena, identificamos o ponto onde a decisão é "PARAR AQUI". O valor de $Q_k$, correspondente a esse ponto, é a altura ideal para traçar a linha de corte no dendrograma, indicando o número de grupos sugerido pelo método.

```
# 1. Plotar o dendrograma (Figura 23)
plot(fit_simples,
     main = "Dendrograma - Corte de Mojena (k=1.25)",
     xlab = "Parcelas",
     ylab = "Distância Euclidiana",
     hang = -1)

# 2. Identificar o valor Qk onde a decisão é "PARAR AQUI" (exemplo: 16.01)
# Este valor deve ser obtido da tabela_mojena gerada no passo anterior
# Para este exemplo, vamos assumir que o primeiro "PARAR AQUI" ocorre com Qk de 16.01

# Encontrar o primeiro Qk onde a decisão é "PARAR AQUI"
indice_parar <- which(tabela_mojena$Decisao == "PARAR AQUI")[1]
if (!is.na(indice_parar)) {
  valor_qk_corte <- tabela_mojena$Qk_Criterio[indice_parar]
} else {
  valor_qk_corte <- max(niveis_fusao) # Se não encontrar, usar o maior nível de fusão como
fallback
  warning("Nenhum ponto de parada encontrado pelo critério de Mojena. Usando o maior nível
de fusão como corte.")
```

```
}

# 3. Desenhar a linha de corte fenon no dendrograma
abline(h = valor_qk_corte, col = "darkgreen", lwd = 2, lty = 2)
text(x = 1.5, y = valor_qk_corte + 1, paste0("Qk = ", round(valor_qk_corte, 2)), col = "darkgreen", font = 2)

# 4. Destacar os grupos resultantes do corte
# O número de grupos será determinado pelo corte em valor_qk_corte
rect.hclust(fit_simples, h = valor_qk_corte, border = "blue")
```

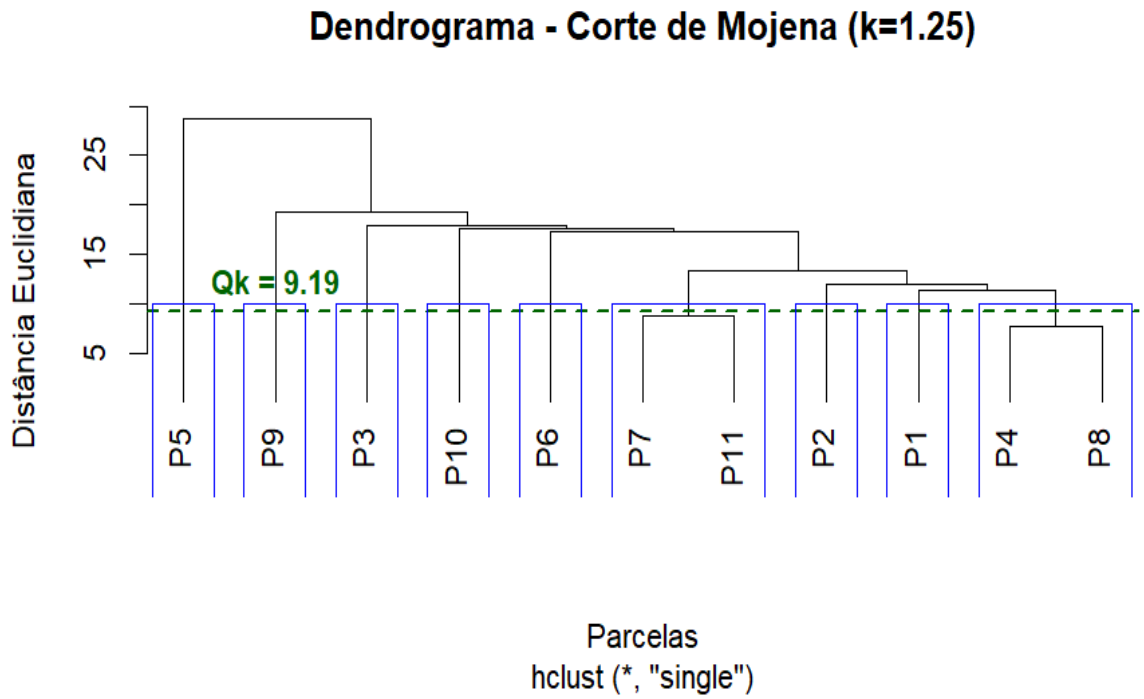



**Figura 23.** Dendrograma obtido pelo emprego do método de ligação simples, com base na distância euclidiana, com destaque do corte determinado pelo método de Mojena (1977).

# Capítulo 5

## Exemplos de uso de análise de agrupamento em Ciência Florestal

Neste capítulo, utilizando-se como palavras-chave agrupamento e *cluster analysis*, foi realizada a pesquisa nos periódicos nacionais diretamente relacionados à área de Recursos Florestais e Engenharia Florestal: Cerne, Ciência Florestal, Revista Árvore e Scientia Forestalis. Pode-se observar, na Tabela 49, considerando o período de 2014 a novembro de 2025, diversos objetivos buscando realizar análise de agrupamento, bem como diversas estratégias para definição de número de agrupamentos, validação de grupos e associação com outras técnicas multivariadas. Por outro lado, observa-se que alguns artigos não explicitam o critério de agrupamento (medida de similaridade ou dissimilaridade) e algoritmo de agrupamento (ligação), o que é uma exigência para realização de inicialização da análise de agrupamento em qualquer estudo.

**Tabela 49.** Alguns detalhes de artigos com utilização de análise de agrupamento publicados no período de 2014 a 2025 nos periódicos Cerne (1), Ciência Florestal (2), Revista Árvore (3) e Scientia Forestalis (4)

| Periódico | Autor(es) | Objetivo(s) | Medida de (Dis)Similaridade | Algoritmo | Definição do Número de Grupos | Validação | Técnica(s) multivariada(s) associadas |
|---|---|---|---|---|---|---|---|
| 1 | Silva et al. (2024) | Caracterizar a diversidade e a estrutura genética de *Anadenanthera colubrina* (Vell.) Brenan em três fragmentos de Floresta Atlântica | Jaccard | UPGMA | Mojena (1977) | Correlação cofenética | |
| | Novak et al. (2019) | Avaliar os atributos físicos e a dinâmica da matéria orgânica do solo sob diferentes coberturas vegetais | Euclidiana | Ligação completa | Linha Fenon | | |
| | Mezzomo et al. (2018) | Selecionar características morfológicas eficientes para separar os isolados de *Fusarium* spp. em grupos de similaridade; | Euclidiana padronizada | UPGMA | Linha Fenon | | |
| | Novak et al. (2017) | Avaliar a qualidade do solo com base em atributos microbiológicos e químicos do solo | Euclidiana | Ligação completa | Linha Fenon | | |
| | Dias Júnior et al. (2017) | Analisar o efeito da adição de madeira de *Eucalyptus grandis* x *E. urophylla* a resíduos madeireiros de origem urbana como estratégia para geração de energia | Manhattan | Ligação média | Linha Fenon | Correlação cofenética | |
| | Corsini et al. (2014) | Analisar a similaridade florística existente entre 26 fragmentos florestais nativos e a equabilidade das fisionomias estudadas e dos grupos de fragmentos formados | Sorensen | UPGMA | Linha Fenon | | |
| | Watzlawick et al. (2014) | Avaliar a influência do grupo ecológico das espécies quanto aos teores de carbono em espécies arbóreas da Floresta Ombrófila Mista Montana | Euclidiana | Ligação completa | Linha Fenon | | |
| 2 | Batista et al. (2025) | Comparar a eficiência e os custos de diferentes processos de amostragem (simples ao acaso, conglomerado sem estratificação e conglomerado com pós-estratificação) em uma floresta de terra firme | Euclidiana | Ward | Linha Fenon | | |
| | Pio et al. (2023) | Assumir que os padrões florístico-estruturais das Florestas Estacionais Sazonalmente Secas podem ser marcantes, devido às condições ambientais impostas nos ambientes e à proximidade com bacias hidrográficas distintas | Sorensen, Bray-Curtis | UPGMA | Permanova | | Componentes principais; Escalonamento Multidimensional Não Métrico; Permanova |
| | Costa Junior et al. (2023) | Classificação da biomassa de povoamentos de acácia negra com base nas variáveis poder calorífico superior, densidade energética, teor de cinzas e estoque de biomassa | Euclidiana, Euclidiana quadrada, Manhattan, Mahalanobis e Minkowski | Ligação simples, Ligação completa, Ligação média, Centróide e Ward | | Correlação cofenética | Discriminante |
| | Fialho et al. (2022) | Selecionar os materiais superiores de *Eucalyptus* spp. a partir de análises estatísticas univariada e multivariada | Euclidiana | Ligação completa | Mojena (1977) | | Componentes principais |
| | Gervazio et al. (2022) | Realizar o levantamento das espécies vegetais em quintais agroflorestais urbanos em Alta Floresta - MT, sul da Amazônia, bem como analisar se contribuem para a conservação da agrobiodiversidade | Bray-Curtis | UPGMA | Linha Fenon | | |
| | Gomes et al. (2022) | Avaliar o efeito da estratificação na prognose da distribuição diamétrica de uma Floresta Ombrófila Mista | Euclidiana | WPGMA | Linha Fenon | Correlação cofenética | |
| | Moraes et al. (2022) | Investigar quintais amazônicos que possuem função de reservatório de agrobiodiversidade | Jaccard | UPGMA | Procedimento pvclust do R | | |

UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento das médias não ponderadas; e WPGMA (*Weighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento por médias ponderadas

*Continua...*

**Tabela 49.** Continuação

| Periódico | Autor(es) | Objetivo(s) | Medida de (Dis)Similaridade | Algoritmo | Definição do Número de Grupos | Validação | Técnica(s) multivariada(s) associadas |
|---|---|---|---|---|---|---|---|
| | Reis et al. (2019) | Agrupar as espécies da Amazônia através das propriedades físico-mecânicas da madeira e realizar a análise discriminante para identificar quais características tecnológicas são mais importantes para o agrupamento | Euclidiana | Ward | Linha Fenon | | Discriminante |
| | Santana et al. (2019) | Avaliar a composição, estrutura e diversidade de espécies regenerantes de uma floresta urbana, oriunda de um projeto paisagístico | Morisita-Horn | UPGMA | Linha Fenon | Correlação cofenética | Correspondência Distendida |
| | Silva et al. (2019) | Classificação dos fragmentos florestais identificados na Bacia do Ribeirão Anhumas, Campinas-SP, utilizando as seguintes métricas de paisagem como indicadores: tamanho, área nuclear e índice de circularidade | Euclidiana | Ward | Linha Fenon | | |
| | Tavares et al. (2019) | Classificar as espécies florestais quanto às exigências em fertilidade do solo, bem como identificar grupos com características semelhantes, utilizando técnicas da estatística multivariada | Euclidiana | Ligação Completa | Linha Fenon | | Componentes principais; Discriminante |
| | Costa et al. (2018) | Delimitar e caracterizar os estratos verticais, observando a regularidade da distribuição de abundância de espécies no gradiente hipsométrico de árvores, em uma comunidade arbórea na Floresta Estacional Subtropical | Bray-Curtis | Ligação Completa | Linha Fenon | | Redundância; Correspondência destendenciada |
| | Ricken et al. (2018) | Analisar relações existentes entre o incremento médio anual em diâmetro e variáveis biométricas, morfométricas e ambientais em uma floresta nativa de *Araucaria angustifolia* | R-Pearson | Ward | Linha Fenon | | Componentes principais; Análise fatorial |
| **2** | Cardoso et al. (2017) | Avaliar a influência da intensidade do tráfego urbano na disponibilidade de elementos e sólidos em suspensão, bem como identificar a potencialidade no acúmulo e retenção das substâncias em diferentes espécies arbóreas, utilizando suas folhas como biomonitores passivos e de acumulação. | Euclidiana | Ward | Linha Fenon | | Componentes principais |
| | Cordeiro et al. (2017) | Caracterizar a dinâmica espaço-temporal do Índice de Vegetação por Diferença Normalizada (NDVI) dos grandes grupos vegetais do Rio Grande do Sul, agrupar regiões homogêneas com base na variabilidade temporal do NDVI e analisar o padrão de variabilidade anual do NDVI dos diferentes grupos | Euclidiana | K-médias | Diferentes modos de inicialização de centroides (por partição aleatória, por pontos semente aleatória e pelo eixo diagonal) | | |
| | Encinas Dardengo et al. (2017) | Caracterizar a anatomia das folhas de *Theobroma speciosum*, registrando as diferenças estruturais observadas entre folhas de sol e folhas de sombra | Euclidiana média padronizada | UPGMA | Linha Fenon | | |
| | Giustina et al. (2017) | Avaliar 50 genótipos de teca com base em marcadores moleculares ISSR | Jaccard transformado para distância | UPGMA; Tocher | Linha Fenon – UPGMA | | |
| | Zavala et al. (2017) | Realizar o inventário florístico de um fragmento florestal sobre área de tensão ecológica no Planalto da Bodoquena, MS, Brasil, e avaliar suas relações fitogeográficas com outras florestas estacionais e cerrados do Centro-Oeste e Sudeste do Brasil. | Sorensen | UPGMA | | | |
| | Callegaro et al. (2015) | Diagnosticar variações estruturais entre os grupos florísticos de uma Floresta Ombrófila Mista Montana, em Nova Prata – RS | Euclidiana | Ward | | | |

UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento das médias não ponderadas; e WPGMA (*Weighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento por médias ponderadas

*Continua...*

**Tabela 49.** Continuação

| Periódico | Autor(es) | Objetivo(s) | Medida de (Dis)Similaridade | Algoritmo | Definição do Número de Grupos | Validação | Técnica(s) multivariada(s) associadas |
|---|---|---|---|---|---|---|---|
| 2 | Lingner et al. (2015) | Caracterizar os remanescentes da Floresta Ombrófila Densa no estado de Santa Catarina | Euclidiana | Ward | Linha Fenon | Correlação cofenética | Correspondência Retificada |
| | Mallmann e Schmitt (2014) | Analisar a riqueza e a composição da comunidade de samambaias em fragmentos de mata ciliar do rio Cadeia, sob diferentes níveis de antropização | Jaccard | UPGMA | Linha Fenon | | Escalonamento multidimensional |
| 3 | Sousa Júnior et al. (2025) | Estabelecer grupos homogêneos de materiais genéticos de *Eucalyptus* spp. e *Corimbia* spp para produção de polpa celulósica, destinada à produção de papel, com base em características morfológicas dos principais elementos anatômicos da madeira, rendimento depurado da polpa marrom e índices de qualidade das fibras (Índices de Runkel, enfeltramento, Mulsteph e coeficiente de flexibilidade) | Euclidiana | Ligação Completa | Linha Fenon | | |
| | Jambers et al. (2025) | Avaliar o potencial energético de diferentes clones de eucalipto, visando sua utilização em agroindústrias | Euclidiana | Não informado | Dois grupos definidos pelos autores | Correlação cofenética | |
| | Wang et al. (2025) | Utilizando marcadores de DNA polimórfico amplificado aleatório (RAPD), analisar a diversidade e a distância genética de *Thuja. koraiensis* | Distância genética com base no marcador molecular RAPDs | UPGMA | | | |
| | Nogueira et al. (2023) | Identificar e propor parâmetros socioambientais que possam ser utilizados na gestão do uso público em ecossistemas de montanha | Euclidiana | Ligação completa | Linha Fenon | | |
| | Souza et al. (2022) | Avaliar as mudanças ocorridas no estrato de regeneração natural em um fragmento de Mata Atlântica com e sem a ocorrência de fogo | Jaccard | UPGMA | Linha Fenon | | |
| | Souza et al. (2021) | Verificar a diversidade genética entre 68 genótipos de *Amburana cearensis* provenientes de diferentes localidades | Euclidiana | Gower, Ligação Média, UPGMA, Ward, Ligação completa, WPGMA, Tocher | Linha Fenon (métodos aglomerativos) | Correlação cofenética | |
| | Silva et al. (2020) | Avaliar o efeito da sazonalidade na ocorrência de gêneros de fungos associados a espécies florestais em uma área de transição Cerrado-Caatinga no estado do Piauí, Brasil | Jaccard | Ward | Linha Fenon | | |
| | Reategui-Betancourt et al. (2021) | Identificar descritores morfológicos em mudas de teca para auxiliar no processo de proteção dos clones da espécie | Jaccard | UPGMA | Linha Fenon | | |

UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento das médias não ponderadas; e WPGMA (*Weighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento por médias ponderadas

*Continua...*

**Tabela 49.** Continuação

| Periódico | Autor(es) | Objetivo(s) | Medida de (Dis)Similaridade | Algoritmo | Definição do Número de Grupos | Validação | Técnica(s) multivariada(s) associadas |
|---|---|---|---|---|---|---|---|
| | Hersen et al. (2019) | Desenvolver e analisar uma taxonomia que particione as unidades federativas do Brasil em grupos com características similares, distribuídas sobre os indicadores das quatro dimensões do desenvolvimento sustentável | Euclidiana quadrada | Ward | Linha Fenon e Método de duas etapas (Soma de quadrados entre e dentre grupos) | | |
| | Palermo e Souza (2019) | Investigar a variabilidade genética populacional de *Annona crassifl ora*, por meio de dados morfométricos de frutos e sementes, em quatro populações | Euclidiana | Ward | Linha Fenon | | |
| | Luz et al. (2018) | Identificar as características de maior importância para a seleção de matrizes de *Corymbia citriodora* para a produção de madeira e óleo essencial | Euclidiana | Tocher | | | |
| | Santos et al. (2018) | Caracterizar a diversidade genética de duas populações de pequi a partir das características físicas dos frutos, mediante análises univariadas e multivariadas, bem como avalias suas implicações para a domesticação e o melhoramento da espécie | Euclidiana | Ward | Linha Fenon | Correlação cofenética | Componentes principais |
| | Baldin et al. (2016) | Descrever anatomicamente os troncos de *Calycophyllum candidissimum*, *C. multiflorum*, *C. spruceanum* e *C. spruceanum f. brasiliensis*, e verificar a similaridade/dissimilaridade deles, com base em caracteres anatômicos | Jaccard | UPGMA | Linha Fenon | Correlação cofenética | |
| **3** | Souza et al. (2016) | Obter informações sobre as características de *Poincianella pyramidalis* (Tul.) através de caracteres morfológicos | Euclidiana Média | UPGMA | Pacote "NbClust" Software R | Correlação cofenética | |
| | Vasconcelos Neto et al. (2016) | Efetuar o agrupamento ecológico e funcional de espécies florestais na Amazonia Sul Ocidental | Euclidiana padronizada | Ward | Linha Fenon | | Discriminante |
| | Ferreira et al. (2015) | Analisar como a distribuição e a riqueza de espécies raras, em fragmentos de Floresta Ombrófila Mista, ocorrem ao longo de um gradiente altitudinal | Jaccard | UPGMA | Linha Fenon | | |
| | Gois et al. (2014) | Caracterizar geneticamente, por meio de marcadores RAPD, indivíduos de *Spondias lutea* L. (cajá), com a finalidade de elaborar estratégias de produção de sementes para a recuperação de mata ciliar | Jaccard | UPGMA, Tocher | Linha Fenon (UPGMA) | | |
| | Guidini et al. (2014) | Avaliar a invasão por espécies arbóreas exóticas em dois fragmentos de Floresta Ombrófila Mista | R-Spearman | Ward | Linha Fenon | Correlação cofenética | |
| | Souza et al. (2014) | Agrupar as unidades de trabalho da unidade de produção anual em classes homogêneas de estoque volumétrico e analisar a composição florística, a diversidade e as estruturas horizontal e diamétrica, por classe de estoque | Euclidiana | Ward | Linha Fenon | | Discriminante |
| 4 | Costa et al. (2022) | Investigar possíveis variações na estrutura e na composição das comunidades arbustivo-arbóreas em três setores de um fragmento de Mata Atlântica | Sorensen, Kulczynski e Canberra | Ward | Linha Fenon | | |

UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento das médias não ponderadas; e WPGMA (*Weighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento por médias ponderadas

*Continua...*

**Tabela 49.** Continuação

| Periódico | Autor(es) | Objetivo(s) | Medida de (Dis)Similaridade | Algoritmo | Definição do Número de Grupos | Validação | Técnica(s) multivariada(s) associadas |
|---|---|---|---|---|---|---|---|
| | Vieira et al. (2022) | Avaliar e comparar o uso do mapa auto-organizável de Kohonen e da análise de agrupamento para a classificação da capacidade produtiva de florestas inequiâneas | Euclidiana | Ward | Linha Fenon | | Discriminante |
| | Delvaux et al. (2021) | Determinar a influência da sazonalidade e da interrupção do fluxo de fotoassimilados da parte aérea para a rizosfera, pelo anelamento de árvores em duas fases de crescimento pós-plantio, sobre a estrutura da comunidade microbiana do solo em florestas de eucalipto | Dice | UPGMA | | | |
| | Silva et al. (2021) | Identificar classes de capacidade produtiva da floresta e valorar os produtos florestais comerciais | Euclidiana | Não explicitado | Linha Fenon | | |
| | Vieira et al. (2021ª) | Avaliar a estrutura diamétrica e o padrão espacial de quatro espécies madeireiras de importância econômica | Euclidiana | Ward | Linha Fenon | | Discriminante |
| | Vieira et al. (2021b) | Avaliar a estrutura diamétrica e a distribuição espacial de *Dipteryx odorata* (Aubl.) Willd. (cumaru) | Euclidiana | Ward | Linha Fenon | | Discriminante |
| | Fritzsons et al. (2020) | Determinar grupos climáticos em que a erva-mate (*Ilex paraguariensis* S.T. Hill) ocorre naturalmente | Não explicitada(o) | Ward | | | Componentes principais |
| | Zanatto et al. (2020) | Avaliar a divergência genética entre progênies de polinização aberta de *Eucalyptus tereticornis* com base em caracteres de crescimento e de qualidade da madeira | Euclidiana | Ward; K-médias | Mojena (1977) | | Componentes principais |
| | Ferreira et al. (2020 | Investigar o potencial de resíduos do processamento de *Schizolobium parayba* var. *amazonicum* (Huber ex Ducke) Barneby para a produção de briquetes | Não explicitada(o) | Não explicitado | Linha Fenon | | |
| 4 | Barbosa et al. (2017) | Analisar procedimentos de pós-estratificação em inventário florestal da vegetação arbóreo-arbustiva | Euclidiana | Ward | Linha Fenon | Análise discriminante | Análise discriminante |
| | Fritzsons et al. (2017) | Investigar a influência da precipitação e da temperatura sobre a presença natural da araucária e identificar diferentes locais onde a espécie tem sua distribuição natural | Não explicitada(o) | Não explicitado | Linha Fenon - Gráfico "Distância de Aglomeração" | | |
| | Kratz et al. (2017) | Avaliar a viabilidade de uma série de materiais para formulação de substratos, com foco em resíduos agroindustriais, e a influência de suas propriedades físico-químicas na produção de mudas de *Eucalyptus benthamii*, e verificar a semelhança entre os componentes testados. | Euclidiana | Ligação simples | Linha Fenon | | |
| | Protásio et al. (2017) | Verificar a viabilidade de uso da relação siringil/guaiacil da lignina e da produtividade de massa seca na classificação de clones de *Eucalyptus*, visando à geração de calor e a produção de carvão vegetal, empregando-se técnicas univariadas e multivariadas | Mahalanobis | UPGMA | Mojena (1977) | Correlação cofenética | |
| | Lobão et al. (2016) | Determinar a idade e a periodicidade do crescimento do tronco do lenho das árvores de *Cedrela* sp. no estado do Acre, em diferentes condições de crescimento (sítios e microsítios), para propiciar os subsídios ao manejo florestal sustentável. | Euclidiana | Não explicitado | Linha Fenon | | Componentes principais |
| | Silveira et al. (2016) | Identificar marcadores moleculares que possam ser usados para caracterizar genótipos responsivos às auxinas usadas para induzir o enraizamento in vitro de explantes de guanandi (*Calophyllum brasiliense)* | Distância genética de Nei e Li | UPGMA; Agrupamento de Vizinhos (*Neighbor-Joining*) | | | |

UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento das médias não ponderadas; e WPGMA (*Weighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento por médias ponderadas

*Continua...*

**Tabela 49.** Continuação

| Periódico | Autor(es) | Objetivo(s) | Medida de (Dis)Similaridade | Algoritmo | Definição do Número de Grupos | Validação | Técnica(s) multivariada(s) associadas |
|---|---|---|---|---|---|---|---|
| 4 | Almeida et al. (2015) | Avaliar a diversidade genética inter e intrapopulacional de *Cenostigma tocantinum*, usada em projetos de arborização urbana, por meio de marcadores AFLP | Distância genética corrigida de Nei | UPGMA | | Correlação cofenética | Escalonamento multidimensional não métrico |
| | Conto et al. (2015) | Caracterizar o perfil vertical do dossel de um trecho de Mata Atlântica através de escaneamento laser aerotransportado | Euclidiana | Ligação simples, Ligação média, Ligação completa, K-médias | Linha Fenon (métodos hierárquicos) | Correlação cofenética | Componentes principais |
| | Teixeira et al. (2015) | Estimar os coeficientes de correlação e avaliar a divergência fenotípica entre 21 genótipos de castanha-do-brasil cultivados em um sistema agroflorestal | Euclidiana média | UPGMA, Tocher | Mojena (1977) | Correlação cofenética | Componentes principais |
| | Pereira et al. (2015) | Classificar sítios florestais com base em propriedades físicas e químicas do solo e, posteriomente, avaliar a influência das propriedades agrupadas em cada sítio sobre o incremento radial do cedro | Não explicitada(o) | UPGMA | Estatística de Mantel | Silhueta | |
| | Protásio et al. (2015) | Selecionar clones jovens de *Eucalyptus* para produção de carvão vegetal de uso siderúrgico | Euclidiana quadrada | UPGMA | Linha Fenon | Correlação cofenética | |
| | Fritzsons et al. (2014) | Buscar indicativos da relação entre as condições edafoclimáticas do sítio e o crescimento diamétrico de procedências de grevílea, em três áreas distintas | Euclidiana Média | Ward | Linha Fenon - Gráfico "Distância de Aglomeração" | | |
| | Lima et al. (2014) | Determinar a equação volume com casca e a classificação da capacidade produtiva de madeira para *Mora paraensis* (Ducke) | Euclidiana | Ward | Linha Fenon | | Discriminante |

UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento das médias não ponderadas; e WPGMA (*Weighted Pair-Group Method using Arithmetic Averages*) = Método de agrupamento por médias ponderadas

# Considerações Finais

Na aplicação das técnicas de agrupamento, várias questões têm surgido, tais como: escolha das variáveis e da medida de distância ou de similaridade; determinação do número de grupos; e escolha do método de agrupamento. A questão relativa à escolha das variáveis a serem utilizadas e do seu número é de fundamental importância na análise de agrupamento, pois a maior ou menor extensão de similaridade entre entidades depende não somente dos atributos (variáveis) medidos, mas também do seu número. Não existem bases teóricas para determinação do número de variáveis a serem utilizadas e, em muitos dos casos, adotam-se procedimentos que possam auxiliar nessa escolha. Quanto à medida de distância ou de similaridade a ser empregada, não há resposta absoluta. A escolha dependerá do conjunto de dados disponíveis e, principalmente, do significado da semelhança que se deseja medir. Com relação ao número de grupos, existem alguns métodos que podem ser adotados para solucionar o problema. No entanto, o que se faz comumente é utilizar vários números de grupos e, por algum critério de otimização, selecionar o mais conveniente ou adotar a análise discriminante para verificação da adequação da partição obtida. A escolha do método de agrupamento também não é um problema simples de ser resolvido, uma vez que existem muitos métodos, e cada um apresenta vantagens e desvantagens, ficando, portanto, a escolha por conta do usuário.

A técnica de análise de agrupamento relaciona-se com outras técnicas multivariadas. A análise discriminante é comumente utilizada para verificar a adequação da partição obtida. A análise fatorial, ou a análise de componentes principais, também é aplicada para redução do número de variáveis a serem utilizadas na análise de agrupamento.

De forma geral, atualmente as técnicas de análise multivariada, em particular a análise de agrupamento, vêm despertando maior interesse de pesquisadores e estudiosos na área florestal, principalmente em razão da a implementação de recursos computacionais e de *softwares* desenvolvidos nesta linha, tornando mais acessível a utilização da referida técnica. Entre os ***softwares*** utilizados, pode-se citar, entre outros, GENES, MATLAB, SAEG, SAS, SPSS, STATISTICA e R.

# Referências Bibliográficas


ADAMS, M.W., WIERSMA, J.V. **An adaptation of principal components analysis to an assessment of genetic distance**. East Lansing: Michigan State University; Agricultural Experiment Station East Lansing, 1978. 8p. (Farm Science. Research Report, 347). Disponível em: https://babel.hathitrust.org/cgi/pt?id=uiug.30112019632667&seq=5. Acesso em: 07 nov. 2025.

ALMEIDA, F. V.; LOPES, M. T. G.; VALENTE, M. S. F.; BENTES, J. L. S. Diversidade genética entre e dentro de populações de *Cenostigma tocantinum* Ducke. **Scientia Forestalis**, v. 43, n. 108, p.753-762, 2015. https://doi.org/10.18671/scifor.v43n108.1.
ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H.-P.; SANDER, J. OPTICS: ordering points to identify the clustering structure. **ACM SIGMOD Record**, v. 28, n. 2, p.49-60, 1999. https://doi.org/10.1145/304181.304187.

BAILEY, K.D. **Typologies and taxonomies**: an introduction to classification techniques. Thousand Oaks: Sage Publications, 1994. 89p. (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-102).

BALDIN, T.; SIEGLOCH, A. M.; MARCHIORI, J. N. C. Compared anatomy of species of *Calycophyllum* DC. (Rubiaceae). **Revista Árvore**, v.40, n.4, p.759-768, 2016. https://doi.org/10.1590/0100-67622016000400020.

BARBOSA, G. P.; NOGUEIRA; G. S.; OLIVEIRA; M. L. R.; MACHADO, E. L. M.; CASTRO, R. V. O.; DUTRA, G. C. Pós-estratificação em inventário florestal da vegetação arbórea-arbustiva. **Scientia Forestalis**, v. 45, n. 115, p.445-453, 2017. https://doi.org/10.18671/scifor.v45n115.03.

BATISTA, F. J.; CASTRO, E. L.; LEITE, F. K. R. G.; FRANCEZ, L. M. B.; SOUZA, D. V.; SCHUH, M. S.; SANTOS, T. R.; FERREIRA, J. E. R. Efficiency and costs of different sampling methods in an inventory of a Terra Firme Forest, Pará state, Brazil. **Ciência Florestal**, v. 35, e88427, 2025. https://doi.org/10.5902/1980509888427.

BORCARD, D., GILLET, F., LEGENDRE, P. Cluster analysis. In: Borcard, D.; Gillet, F.; Legendre, P. **Numerical ecology with R**. New York: Springer, 2011. p.53-114. https://doi.org/10.1007/978-1-4419-7976-6_4.

BORCARD, D.; GILLET, F.; LEGENDRE, P. **Numerical Ecology with R**. 2.ed. Cham: Springer International Publishing, 2018. 435p. https://doi.org/10.1007/978-3-319-71404-2.

BOUROCHE, J.M., SAPORTA, G. **Análise de dados**. Rio de Janeiro: Zahar, 1982. 117p.

BOX, G.E.P.; COX, D.R. An analysis of transformations. **Journal of the Royal Society**, v.26, n.2, p.211-252, 1964. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

BUSSAB, W.O., MIAZAKI, E.S., Andrade, D.F. **Introdução à análise de agrupamentos**. São Paulo: IME-USP, 1990. 105p.

CALLEGARO, R.M.; LONGHI, S.J.; ANDRZEJEWSKI, C. Variações estruturais entre grupos florísticos de um remanescente de floresta ombrófila mista montana em Nova Prata – RS. **Ciência Florestal**, v. 25, n. 2, p.337-349, 2015. https://doi.org/10.5902/1980509818452.

CARDOSO, K.M.; PAULA, A.; SANTOS, J.S.; SANTOS, M.L.P. Uso de espécies da arborização urbana no biomonitoramento de poluição ambiental. **Ciência Florestal**, v. 27, n. 2, p.535-547, 2017. https://doi.org/10.5902/1980509827734.

CATTELL, R. B. A note on correlation clusters and cluster search methods. **Psychometrika**, v. 9, p.169-184, 1944. https://doi.org/10.1007/BF02288721.

CATTELL, R. B.; COULTER, M. A. Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. **The British Journal of Mathematical and Statistical Psychology**, v. 19, n. 2, p.237-269, 1966. https://doi.org/10.1111/j.2044-8317.1966.tb00370.x.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. NbClust: an R package for determining the relevant number of clusters in a data set. **Journal of Statistical Software**, v. 61, n. 6, p.1-36, 2014. https://doi.org/10.18637/jss.v061.i06.

COMARCK, R.M. A review of classifications. **Journal of the Royal Statistical Society. Series A (General)**, v.134, n. 3, p.321-353, 1971. https://doi.org/10.2307/2344237.

CONTO, T.; GÖRGENS, E. B.; SILVA, A. G. P.; LARANJA, D. C. F.; ESTRAVIZ RODRIGUEZ, L. C. Caracterização do perfil vertical do dossel de um trecho de Mata Atlântica através de escaneamento laser aerotransportado. **Scientia Forestalis**, v. 43, n. 108, p.873-884, dez. 2015. https://doi.org/10.18671/scifor.v43n108.12.

CORDEIRO, A.P.A.C.; BERLATO, M.A.; FONTANA, D.C.; MELO, R.W.; SHIMABUKURO, Y.E.; FIOR, C.S. Regiões homogêneas de vegetação utilizando a variabilidade do NDVI. **Ciência Florestal**, v. 27, n. 3, p.883-896, 2017. https://doi.org/10.5902/1980509828638.

CORSINI, C. R.; SCOLFORO, J. R. S.; OLIVEIRA, A. D.; MELLO, J. M.; MACHADO, E. L. M. Diversidade e similaridade de fragmentos florestais nativos situados na Região Nordeste de Minas Gerais. **Cerne**, v. 20, n. 1, p.1-10, 2014. https://doi.org/10.1590/S0104-77602014000100001.

COSTA JUNIOR, S.; SILVA, D. A.; BEHLING, A.; KOEHLER, H. S.; TRAUTENMÜLLER, J. W.; COSTA, A. Classificação da qualidade da biomassa de árvores de acácia-negra para fins energéticos. **Ciência Florestal**, v. 33, n. 2, e71436, 2023. https://doi.org/10.5902/1980509871436.

COSTA, M. P.; LONGHI, S. J.; FÁVERO, A. A. Arquitetura e estrutura vertical da comunidade arbórea de uma floresta estacional subtropical. **Ciência Florestal**, v. 28, n. 4, p.1443-1454, 2018. https://doi.org/10.5902/1980509835052.

COSTA, V. P. P.; HOLANDA, A. C.; COSTA, M. P.; LOPES-NUNES, A. L. S. Estrutura da vegetação como indicador de distúrbio e resiliência em Unidade de Conservação na Mata Atlântica. **Scientia Forestalis**, v. 50, e3815, 2022. https://doi.org/10.18671/scifor.v50.14.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento das plantas**. 1990. Tese (Doutorado) – Universidade de São Paulo, Piracicaba, 1990. https://doi.org/10.11606/T.11.1990.tde-20231122-093446.

CRUZ, C.D., REGAZZI, A.J.; Carneiro, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v.1. Viçosa: Editora UFV, 2012. 514p.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** v.2. Viçosa: Editora UFV, 2013. 585p.

CRUZ, C.D; FERREIRA, F.M; PESSONI, L.A. **Biometria aplicada ao estudo de diversidade genética**. 2.ed. Viçosa: UFV, 2020. 620p.

DALTON, L.; BALLARIN, V.; BRUN, M. Clustering algorithms: on learning, validation, performance, and applications to genomics. **Current Genomics**, v.10, n.6, p.430-445, 2009. https://doi.org/10.2174/138920209789177601.

DE MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D. L. The Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems**, v. 50, n. 1, p.1-18, 2000. https://doi.org/10.1016/S0169-7439(99)00047-7.

DELVAUX, J. C.; MIGUEL, P. S. B.; OLIVEIRA, M. N. V.; FRANCO, M. H. R.; CAMARGO, R.; BORGES, A. C. Efeito da sazonalidade e disponibilidade de recursos sobre comunidades microbianas em solo de floresta de eucalipto. **Scientia Forestalis**, v. 49, n. 130, e3593, 2021. https://doi.org/10.18671/scifor.v49n130.18.

DIAS JÚNIOR, A. F.; ANUTO, R. B.; ANDRADE, C. R.; SOUZA, N. D.; TAKESHITA, S.; BRITO, J. O.; NOLASCO, A. M. Influence of *Eucalyptus* wood addition to urban wood waste during combustion. **Cerne**, v.23, n.4, p.455-464, 2017. https://doi.org/10.1590/01047760201723042337.

EDWARDS, A.W.F., CAVALLI-SFORZA, L.L. A method for cluster analysis. **Biometrics,** v.21, n. 2, p.362-365, 1965. https://doi.org/10.2307/2528096.

ENCINAS DARDENGO, J. F.; ROSSI, A. A. B.; SILVA, I. V.; PESSOA, M. J. G.; SILVA, C. J. Análise da influência luminosa nos aspectos anatômicos de folhas de *Theobroma speciosum* Willd Ex Spreng. (Malvaceae). **Ciência Florestal**, v. 27, n. 3, p.843-851, 2017. https://doi.org/10.5902/1980509828634.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining (KDD-96), 2., 1996, Portland. **Proceedings**… Portland: AAAI Press, 1996. p.226–231. Disponível em: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf. Acesso em: 03 abr. 2025.

EVERITT, B. S. **Statistical methods for medical investigations**. London: Edward Arnold, 1989. 208p.

EVERITT, B. S.; DUNN, G. **Applied multivariate data analysis**. 2.ed. London: Hodder Education, 2001. 342p.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHI, D. **Cluster analysis**. 5.ed. London: John Wiley, 2011. 352p.

FACELI, K.; CARVALHO, A. C. P. L. F.; SOUTO, M. C. P. **Validação de algoritmos de agrupamento**. São Carlos: USP; ICMC, 2005. 38p. (USP; ICMC. Relatórios

Técnicos, 254). Disponível em: https://repositorio.usp.br/bitstreams/868d51f1-7b35-4223-87e7-1e2f750f7c10 . Acesso em: 27 mar. 2025.

FARRIS, J. S. On the cophenetic correlation coeficient. **Systematic Zoology**, v. 18, n. 3, p.279-285, 1969. https://doi.org/10.2307/2412324.

FERREIRA, T. S.; HIGUCHI, P.; SILVA, A. C.; MANTOVANI, A.; MARCON, A. K.; SALAMI, B.; BUZZI JUNIOR, F.; ANSOLIN, R. D.; BENTO, M. A.; ROSA, A. D. Distribuição e riqueza de espécies arbóreas raras em fragmentos de floresta ombrófila mista ao longo de um gradiente altitudinal, em Santa Catarina. **Revista Árvore**, v. 39, n. 3, p.447-455, 2015. https://doi.org/10.1590/0100-67622015000300005.

FERREIRA, V. R. S.; CADEMARTORI, P. H. G.; LIMA, E. A.; FERRAZ, F. A.; AGUIAR, O. J. R.; SILVA, D. A. Produção e avaliação de briquetes de *Schizolobium parahyba* var. *amazonicum* (Huber ex Ducke) Barneby. **Scientia Forestalis**, v. 48, n. 128, e3284, 2020. https://doi.org/10.18671/scifor.v48n128.12.

FIALHO, L.F.; CARNEIRO, A.C.O.; FIGUEIRÓ, C.G.; PERES, L.C.; CARNEIRO, A.P.S.; SURDI, P.G. Application of univariate and multivariate statistical analyzes in clonal selection of *Eucalyptus* spp. for charcoal production. **Ciência Florestal**, v. 32, n. 3, p.1659-1683, 2022. https://doi.org/10.5902/1980509840443.

FREI, F. **Introdução à análise de agrupamentos**: teoria e prática. São Paulo: Editora Unesp, 2006. 111p.

FRITZSONS, E.; MATTOS, P. P.; AGUIAR, A. V.; BRAZ, E. M.; GRABIAS, J.; FERRAZ, M. Crescimento da *Grevillea robusta* em diferentes sítios edafoclimáticos no Estado do Paraná. **Scientia Forestalis**, v. 42, n. 103, p.383-392, 2014. Disponível em: https://www.ipef.br/publicacoes/scientia/nr103/cap08.pdf. Acesso em: 29 abr. 2025.

FRITZSONS, E.; WREGE, M. S.; MANTOVANI, L. E. Fatores climáticos limitantes para a distribuição da araucária no estado de São Paulo. **Scientia Forestalis**, v. 45, n. 116, p.663-672, 2017. https://doi.org/10.18671/scifor.v45n116.07.

FRITZSONS, E.; WREGE, M. S.; SOARES, M. T. S.; AGUIAR, A. V.; SOUSA, V. A.; SCARANTE, A. G.; AMBRÓSIO, M. F. M. Proposta metodológica para subsidiar conservação e melhoramento genético da erva mate no sul do Brasil, baseada em grupos climáticos. **Scientia Forestalis**, v. 48, n. 128, e3267, 2020. https://doi.org/10.18671/scifor.v48n128.22.

GAMA, M. de P. **Bases da análise de grupamentos (*cluster analysis*)**. São Paulo: Dialética, 2022. 200p.

GERVAZIO, W.; Y., O. M.; ROBOREDO, D.; BERGAMASCO, S. M. P. P.; FELITO, R. A. **Quintais agroflorestais urbanos no sul da Amazônia: os guardiões da agrobiodiversidade? Ciência Florestal**, Santa Maria-RS, v. 32, n. 1, p.163-186, 2022. https://doi.org/10.5902/1980509843611.

GIUSTINA, L.D.; ROSSI, A.A.B.; VIEIRA, F.S.; TARDIN, F.D.; NEVES, L.G.; PEREIRA, T.N.S. Variabilidade genética em genótipos de teca (*Tectona grandis* Linn. F.) baseada em marcadores moleculares ISSR e caracteres morfológicos. **Ciência Florestal**, v. 27, n. 4, p.1311-1324, 2017. https://doi.org/10.5902/1980509829894.

GOIS, I. B.; FERREIRA, R. A.; SILVA-MANN, R.; BLANK, M. F. A.; SANTOS NETO, E. M. Diversidade genética entre indivíduos de *Spondias lutea* L. procedentes do baixo São Francisco Sergipano, por meio de marcadores RAPD. **Revista Árvore**, v. 38, n. 2, p.261-269, 2014. https://doi.org/10.1590/S0100-67622014000200006.

GOMES, M. S.; DIAS, A. N.; FIGUEIREDO FILHO, A.; RETSLAFF, F. A. S.; LANSSANOVA, L. R. Estratégias de projeção da estrutura diamétrica em Floresta Ombrófila Mista. **Ciência Florestal**, v. 32, n. 2, p.902-922, 2022. https://doi.org/10.5902/1980509863311.

GORE JR., P. A. Cluster analysis. In: Tinsley, H.E.A.; Brown, S.D. (Eds.). **Handbook of applied multivariate statistics and mathematical modeling**. London: Academic Press, 2000. Chap. 11, p.297-321. https://doi.org/10.1016/B978-012691360-6/50012-4.

GOTELLI, N. J.; ELLISON, A. M. **A primer of ecological statistics**. Sunderland: Sinauer Associates, 2004. 614p.

GOWER, J.C. A comparison of some methods of cluster analysis. **Biometrics,** v. 23, n. 4, p.623-637, 1967. https://doi.org/10.2307/2528417.

GOWER, J.C. A general coefficient of similarity and some of its properties. **Biometrics,** v. 27, n. 4, p.857-871, 1971. https://doi.org/10.2307/2528823.

GUIDINI, A.L.; SILVA, A.C.; HIGUCHI, P.; DALLA ROSA, A.; SPIAZZI, F. R.; NEGRINI, M.; FERREIRA, T. S.; SALAMI, B.; MARCON, A. K.; JUNIO, F. B. Invasão por espécies arbóreas exóticas em remanescentes florestais no Planalto Sul Catarinense.

**Revista Árvore**, v. 38, n. 3, p.469-478, 2014. https://doi.org/10.1590/S0100-67622014000300009.

GÜNGÖR, E.; ÖZMEN, A. Distance and density based clustering algorithm using Gaussian kernel. **Expert Systems with Applications**, v. 69, p.10-20, 2017. https://doi.org/10.1016/j.eswa.2016.10.022.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. 6.ed. Porto Alegre: Bookman, 2009. 687p.

HARTIGAN, J.A. **Clustering algoritms**. New York: John Wiley and Sons, 1975. 366p.

HERSEN, A.; TIMOFEICZYK JUNIOR, R.; SILVA, D. A.; SILVA, J. C. G. L.; LIMA, J. F. Sustainable development in Brazil: a conglomerated analysis for federative units. **Revista Árvore**, v. 43, n. 6, e430604, 2019. https://doi.org/10.1590/1806-90882019000600004.

HINNEBURG, A.; GABRIEL, H.-H. DENCLUE 2.0: Fast clustering based on kernel density estimation. In: Berthold, M. R.; Shawe-Taylor, J.; Lavrač, N. (Eds.). **Advances in intelligent data analysis VII (IDA 2007)**. Heidelberg: Springer-Verlag, 2007. p.70-80. (Lecture Notes in Computer Science - LNISA, 4723). https://doi.org/10.1007/978-3-540-74825-0_7.

HINNEBURG, E.; KEIM, D.A. An efficient approach to clustering in large multimedia databases with noise. In: International Conference on Knowledge Discovery and Data Mining (KDD 1998), 4., 1998, New York. **Proceedings**… New York: AAAI Press, 1998. p.58-65. Disponível em: https://cdn.aaai.org/KDD/1998/KDD98-009.pdf. Acesso em: 03 abr. 2025.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p.651–666, 2010. https://doi.org/10.1016/j.patrec.2009.09.011.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Englewood Cliffs: Prentice Hall, 1988. 320p.

JAMBERS, V. T.; AMORIM, C. N.; CAMPOS, C. M.; ROCHA, L. L.; CARNEIRO, A. C. O.; CARVALHO, A. M. M. L.; OLIVEIRA, A. C.; PEREIRA, B. L. C. Firewood quality of *Eucalyptus* spp. clones for agroindustry supply in Mato Grosso, Brazil. **Revista Árvore,** v. 49, n. 1, e4921, 2025. https://doi.org/10.53661/1806-9088202549263901.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6.ed. Upper Saddle River: Pearson Prentice Hall, 2007. 808p.

JOLLIFFE, I. T. **Principal component analysis**. 2.ed. New York: Springer, 2002. 487p.

KASSAMBARA, A. **Practical guide to cluster analysis in R** - unsupervised machine learning. 1.ed. Sydney: STHDA; Alboukadel Kassambara, 2017. 187p.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data**. An introduction to cluster analysis. Hoboken: Wiley-Interscience, 2005. 342p.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS software**. Cary: SAS Institute, 2000. 338p.

KRATZ, D.; NOGUEIRA, A. C.; WENDLING, I.; MELLEK, J. E. Physic-chemical properties and substrate formulation for *Eucalyptus* seedlings production. **Scientia Forestalis**, v. 45, n. 113, p.63-76, 2017. https://doi.org/10.18671/scifor.v45n113.06.

LANCE, G.N., WILLIAMS, W.T. A general theory of classificatory sorting strategies. **Computer Journal**, v. 9, n. 4, p.373-380, 1967. https://doi.org/10.1093/comjnl/9.4.373.

LEDOIT, O.; WOLF, M. A well-conditioned estimator for large-dimensional covariance matrices. **Journal of Multivariate Analysis**, v. 88, n. 2, p. 365-411, 2004. https://doi.org/10.1016/S0047-259X(03)00096-4.

LEDOIT, O.; WOLF, M. Honey, I shrunk the sample covariance matrix. **The Journal of Portfolio Management**, v. 30, n. 4, p. 110-119, 2003. https://doi.org/10.3905/jpm.2004.110.

LEGENDRE, P.; LEGENDRE, L. **Numerical Ecology**. 3.ed. Amsterdam: Elsevier, 2012. 1006p. (Developments in Environmental Modelling, 24).

LIMA, R. B.; APARÍCIO, P. S.; FERREIRA, R. L. C.; SILVA, W. C.; GUEDES, M. C.; OLIVEIRA, C. P.; SILVA, D. A. S.; BATISTA, A. P. B. Volumetria e classificação da capacidade produtiva para *Mora paraensis* (Ducke) no estuário amapaense. **Scientia Forestalis**, v. 42, n. 101, p.141-154, 2014. Disponível em: https://www.ipef.br/publicacoes/scientia/nr101/cap13.pdf. Acesso em: 26 abr. 2025.

LINGNER, D. V.; SCHORN, L. A.; SEVEGNANI, L.; GASPER, A. L.; MEYER, L.; VIBRANS, A. C. Floresta Ombrófila Densa de Santa Catarina - Brasil: agrupamento e

ordenação baseados em amostragem sistemática. **Ciência Florestal**, v. 25, n. 4, p.933-946, 2015. https://doi.org/10.5902/1980509820595.

LLOYD, S. P. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, v. IT-28, n. 2, p.129-137, 1982. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1056489. Acesso em: 13 nov. 2025.

LOBÃO, M. S.; ORTEGA, G. P.; AMARAL, E.; AMORIM, P. G. R.; AMARO, M. A.; ROIG, F. A.; TOMAZELLO FILHO, M. Análise de similaridade das árvores de *Cedrela* sp. sob diferentes condições de crescimento no leste do estado do Acre, Brasil. **Scientia Forestalis**, v. 44, n. 109, p.231-239, 2016. https://doi.org/10.18671/scifor.v44n109.22.

LUZ, O. S. L.; HONÓRIO, A. B. M.; FIDELIS, R. R.; NASCIMENTO, I. R.; MORAES, C. B.; LEAL, T. C. A. B. Characteristics for the selection of parents of *Corymbia citriodora* aiming to the production of wood and essential oil. **Revista Árvore**, v. 42, n. 1, e420118, 2018. https://doi.org/10.1590/1806-90882018000100018.

MAHALANOBIS, P. C. On the generalised distance in statistics. **Proceedings of the National Institute of Sciences of India**, v. 2, n. 1, p. 49-55, 1936. Disponível em: https://insa.nic.in/writereaddata/UpLoadedFiles/PINSA/Vol02_1936_1_Art05.pdf. Acesso em: 13 mai. 2026.

MALHOTRA, N. K. **Pesquisa de marketing**: uma orientação aplicada. 7.ed. Porto Alegre: Bookman Editora, 2019. 800p.

MALLMANN, I. T.; SCHMITT, J. L. Riqueza e composição florística da comunidade de samambaias na mata ciliar do Rio Cadeia, Rio Grande do Sul, Brasil. **Ciência Florestal**, v. 24, n. 1, p.97-109, 2014. https://doi.org/10.5902/1980509813327.

MANLY, B.F.J.; NAVARRO ALBERTO, J.A. **Métodos estatísticos multivariados**: uma introdução. 4.ed. Porto Alegre: Bookman, 2019. 254p.

MARDIA, K.V., KENT, J.T., TAYLOR, C. **Multivariate analysis**. 2.ed. London: Jonh Wiley, 2023. 592p.

MATTE, M. K.; NICOLETTI, M. C. Revisão de estratégias para a aceleração do algoritmo k-means. **Anais do WCF**, v. 6, p.1-6, 2019. Disponível em: https://www.cc.faccamp.br/anaisdowcf/edicoes_anteriores/wcf2019/01/paper_01.pdf. Acesso em: 17 nov. 2023.

MATTEUCCI, S.D., COLMA, A**. Metodología para el estudio de la vegetación**. Washington: Secretaria General de la Organización de los Estados Americanos, 1982. 167p. Disponível em: https://www.researchgate.net/publication/44553298. Acesso em: 22 jun. 2025.

McCUNE, B.; GRACE, J. B. **Analysis of ecological communities**. Gleneden Beach: MjM Software Design, 2002. 300p.

MEZZOMO, R.; ROLIM, J. M.; POLETTO, T.; ROSENTHAL, V. C.; SAVIAN, L. G.; REINIGER, L. R. S.; MUNIZ, M. F. B. Morphological and molecular characterization of *Fusarium* spp. pathogenic to *Ilex paraguariensis*. **Cerne**, v. 24, n. 3, p.209-218, 2018. https://doi.org/10.1590/01047760201824032535.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in data set. **Psychometrica**, v. 50, n.2, p.159-179, 1985. https://doi.org/10.1007/BF02294245.

MOJENA, R. Hierarchical grouping method and stopping rules: an evaluation. **Computer Journal**, v. 20, n. 4, p.359-363, 1977. https://doi.org/10.1093/comjnl/20.4.359.

MORAES, H.C.S.; KATO, O.R.; SABLAYROLLES, M.G.P.; AZEVEDO, C.M.B.C.; OLIVEIRA, J.S.R. Inovação nos quintais agrobiodiversos da Cooperativa D'Irituia, Pará. **Ciência Florestal**, v. 32, n. 1, p.309-332, 2022. https://doi.org/10.5902/1980509854864.

MURTAGH, F.; LEGENDRE, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? **Journal of Classification**, v. 31, p.274-295, 2014. https://doi.org/10.1007/s00357-014-9161-z.

NOGUEIRA, B. G. S.; SAVI, M.; SANTOS, J. F. L.; TETTO, A. F.; PAULA, E. V.; STEINER NETO, P. J. Social and environmental parameters for public use management in mountain ecosystems. **Revista Árvore**, v. 47, e4723, 2023. https://doi.org/10.1590/1806-908820230000023.

NOVAK, E.; CARVALHO, L. A.; SANTIAGO, E. F.; TOMAZI, M. Changes in the soil structure and organic matter dynamics under different plant covers. **Cerne**, v. 25, n. 2, p.230-239, 2019. https://doi.org/10.1590/01047760201925022618.

NOVAK, ELAINE CARVALHO, SANTIAGO, L. A.; PORTILHO, E. F.; ROSA, I. I. Chemical and microbiological attributes under different soil cover. **Cerne**, v. 23, n. 1, p.19-30, 2017. https://doi.org/10.1590/01047760201723012228.

ORLÓCI, L. **Multivariate analysis in vegetational research**. 2.ed. The Hague: Dr. W. Junk B.V., Publishers, 1978. 451p.

PALACIO, F. X.; APODACA, M. A.; CRISCI, J. V. **Análisis multivariado para datos biológicos**: teoría y su aplicación utilizando el lenguaje R. Ciudad Autónoma de Buenos Aires : Fundación de Historia Natural Félix de Azara, 2020. 268p.

PALERMO, A. C.; SOUZA, A. M. Morphometric analysis of fruits and seeds of *Annona crassiflora* Mart. (Annonaceae) from Central Brazil. **Revista Árvore**, v. 43, n. 3, e430304, 2019. https://doi.org/10.1590/1806-90882019000300004.

PEREIRA, J. C. R. **Análise de dados qualitativos**: estratégias metodológicas para as ciências da saúde, humanas e sociais. 2.ed. São Paulo: Editora da Universidade de São Paulo, 1999. 156p.

PEREIRA, L. D.; FLEIG, F. D.; REICHERT, J. M.; RODRIGUES, M. F.; SCHRÖDER, T.; MEYER, E. A. Classificação de sítio com base em propriedades físicas e químicas do solo para incremento radial de *Cedrela fissilis*. **Scientia Forestalis**, v. 43, n. 108, p.791-799, 2015. Disponível em: https://www.ipef.br/publicacoes/scientia/nr108/cap05.pdf. Acesso em: 26 abr. 2025.

PIMENTEL, M.C.P. **Uma técnica para agrupar tratamentos na análise de variância**. Porto Alegre: UFRGS, 1987. Trabalho de conclusão de graduação (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul. Disponível em: http://hdl.handle.net/10183/198135. Acesso em: 14 abr. 2025.

PIO, A. D.; OLIVEIRA, L. R.; SPINOLA, C. M.; COSTA, J. P.; SANTOS, L. C. S.; VALE, V. S. Padrões florístico-estruturais, riqueza e diversidade de Florestas Estacionais Semideciduais no Cerrado. **Ciência Florestal**, v. 33, n. 3, e69612, 2023. https://doi.org/10.5902/1980509869612.

POHLMANN, M. C. Análise de conglomerado. In: Corrar, L. J.; Paulo, E.; Dias Filho, J. M. (Coord.). **Análise multiviariada**: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007. Cap. 6, p.324-388.

PROTÁSIO, T. P.; COUTO, A. M.; TRUGILHO, P. F.; GUIMARÃES JUNIOR, J. B.; LIMA JUNIOR, P. H.; SILVA, M. M. O. **Scientia Forestalis**, v. 43, n. 108, p.801-816, 2015. https://doi.org/10.18671/scifor.v43n108.6.

PROTÁSIO, T. P.; TRUGILHO, P. F.; ARAÚJO, A. C. C.; BASTOS, T. A.; ROSADO, S. C. S.; PINTO, J. F. N. Classificação de clones de *Eucalyptus* por meio da relação siringil/guaiacil e das características de crescimento para uso energético. **Scientia Forestalis**, v. 45, n. 114, p.327-341, 2017. https://doi.org/10.18671/scifor.v45n114.09.

QUEIROZ, W. T. **Análise multivariada em inventário florestal contínuo**. 1.ed. Belo Horizonte: Poisson, 2021. 271p. Disponível em: https://livros.poisson.com.br/individuais/Multivariada_Florestal/Multivariada_Florestal.pdf. Acesso em: 10 Mai. 2026.

RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. 3.ed. Lavras: UFLA, 2012. 305p.

RAO, R.C. **Advanced statistical method in biometric research**. New York: John Wiley and Sons, 1952. 390 p.

REATEGUI-BETANCOURT, J.L.; ARRIEL, D.A.A.; CALDEIA, S.F.; HIGA, A.R.; MARQUES, R.M.; GONÇALVES, I.S.; MARTINEZ, D.T. Morphological descriptors for the characterisation of Teak clones cuttings (*Tectona grandis* L.F.). **Revista Árvore**, v. 45, e4522, 2021. https://doi.org/10.1590/1806-908820210000022.

REIS, E. **Estatística multivariada aplicada**. 2.ed. Lisboa: Edições Silabo, 2001. 343p.

REIS, P.C.M.R.; REIS, L.P.; SOUZA, A.L.; CARVALHO, A.M.M.L.; MAZEEI, L.; REIS, A.R.S.; TORRES, C.M.M.E. Agrupamento de espécies madeireiras da Amazônia com base em propriedades físicas e mecânicas. **Ciência Florestal**, v. 29, n. 1, p.336-346, 2019. https://doi.org/10.5902/1980509828114.

RENCHER, A. C. **Methods of multivariate analysis**. 2.ed. New York: John Wiley & Sons, 2002. 708p.

RICKEN, P.; HESS, A. F.; BORSOI, G. A. Relações biométricas e ambientais no incremento diamétrico de *Araucaria angustifolia* no Planalto Serrano Catarinense. **Ciência Florestal**, v. 28, n. 4, p.1592–1603, 2018. https://doi.org/10.5902/1980509835107.

ROHLF, F.J. Adaptative hierarchical clustering schemes. **Systematic Zoology**, v.19, n.1, p.58-82, 1970. https://doi.org/10.1093/sysbio/19.1.58.

ROMESBURG, H.C. **Cluster analysis for researchers**. Morrisville: Lulu.com, 2004. 334p.

RUSSEL, P. F., RAO, T. R. On habitat and association of species of anopheline larvae in south-eastern Madras. **Journal of the Malaria Institute of India**, v. 3, n. 1, p.154-178, 1940.

SANTANA, L. D.; FONSECA, C. R.; CARVALHO, F. A. Aspectos ecológicos das espécies regenerantes de uma floresta urbana com 150 anos de sucessão florestal: o risco das espécies exóticas. **Ciência Florestal**, v. 29, n. 1, p.1-13, 2019. https://doi.org/10.5902/1980509830870.

SANTOS, P. H. R.; GIORDANI, S. C. O.; SOARES, B. C.; SILVA, F. H. L.; ESTEVES, E. A.; FERNANDES, J. S. C. Genetic divergence in populations of *Caryocar brasiliense* Camb. from the physical characteristics of the fruits. **Revista Árvore**, v. 42, n. 1, e420116, 2018. https://doi.org/10.1590/1806-90882018000100016.

SCHÄFER, J.; OPGEN-RHEIN, R.; ZUBER, V.; AHDESMÄKI, M.; DUARTE SILVA, A. P.; STRIMMER, K. **corpcor**: efficient estimation of covariance and (partial) correlation. R package version 1.6.10, 2021. Disponível em: https://strimmerlab.github.io/software/corpcor/. Acesso em: 10 Mai. 2026.

SCHÄFER, J.; STRIMMER, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. **Statistical Applications in Genetics and Molecular Biology**, v. 4, n. 1, e32, 2005. https://doi.org/10.2202/1544-6115.1175.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, v. 30, n. 3, p.507-512, 1974. https://doi.org/10.2307/2529204.

Silva, A.L.; Longo, R.M.; Bressane, A.; Carvalho, M.F.H. Classificação de fragmentos florestais urbanos com base em métricas da paisagem. **Ciência Florestal**, v. 29, n. 3, p.1254-1269, 2019. https://doi.org/10.5902/1980509830201.

SILVA, H. F.; SANTOS, A. M. G.; SANTOS, M. V. O.; BEZERRA, J. L.; LUZ, E. D. M. N. Seasonal variation in the occurrence of fungi associated with forest species in a

Cerrado-Caatinga transition area. **Revista Árvore**, v. 44, e4409, 2020. https://doi.org/10.1590/1806-908820200000009.

SILVA, K. D. A.; SILVA JÚNIOR, A. L.; SOUZA, M. C.; SOUZA, L. C.; MIRANDA, F. D.; CALDEIRA, M. V. W.; AZEVEDO, C. DOS S.; SOARES, T. C. B. Effects of forest fragmentation on natural populations of *Anadenanthera colubrina* (Vell.) Brenan: Insights for conservation and sustainable management. **Cerne**, v. 20, e103316, 2024. https://doi.org/10.1590/01047760202430013316.

SILVA, M. K. V. S.; GAMA, J. R. V.; OLIVEIRA, F. A.; RIBEIRO, R. B. S.; MELO, L. O.; ANDRADE, D. F. C.; SILVA, A. A.; CRUZ, G. S. Capacidade produtiva em floresta de planalto no baixo Tapajós, Amazônia Oriental. **Scientia Forestalis**, v. 49, n. 130, e3460, 2021. https://doi.org/10.18671/scifor.v49n130.02.

SILVEIRA, S. S.; SCHÜHLI, G. S.; DEGENHARDT-GOLDBACH, J.; QUOIRIN, M. G. G. Análise da capacidade de enraizamento *in vitro* de *Calophyllum brasiliense* utilizando marcadores RAPD. **Scientia Forestalis**, v. 44, n. 110, p.527-535, 2016. https://doi.org/10.18671/scifor.v44n110.25.

SNEATH, P.H.A., SOKAL, R.R. **Numerical taxonomy.** The principles and practice of numerical classification. San Francisco: W.H. Freeman, 1973. 573 p.

SOARES, C.P.B.; LEITE, H.G.; CAMPOS, J.C.C. Um novo modelo de crescimento e produção. **Revista Árvore**, v.25, n.2, p.265-270, 2001. Disponível em: https://books.google.com.br/books?id=t3iaAAAAIAAJ&printsec=frontcover&hl=pt-BR&rview=1&lr=#v=onepage&q&f=false. Acesso em: 31 out. 2023.

SOKAL, R. R.; ROHLF, F. J. **Biometry**: the principles and practice of statistics in biological research. 3.ed. New York: W. H. Freeman and Company, 1995. 899p.

SOKAL, R.R.; MICHENER, C.D. A statistical method for evaluating systematics relationships. **Science Bulletin**, v.38, part 2, p.1409-1438, 1958. Disponível em: https://ia600703.us.archive.org/5/items/cbarchive_33927_astatisticalmethodforevaluatin1902/astatisticalmethodforevaluatin1902.pdf. Acesso em: 11 abr. 2025.

SOKAL, R.R.; ROHLF, F.J. The comparison of dendrograms by objective methods. **Taxon**, v.11, n.2, p.33-40, 1962. https://doi.org/10.2307/1217208.

SOKAL, R.R.; SNEATH, P.H. **Principles of numerical taxonomy**. San Francisco: WH Freeman Company, 1963. 375p.

SOUSA JÚNIOR, W. P.; CARVALHO, A. M. M. L.; CARNEIRO, A. C. O.; DEMUNER, I. F.; OLIVEIRA, R. S.; RECIS, L. A. C.; LOPES, L. A.; MORETTI, S. D. A. Fiber quality indices of *Corymbia* spp. and *Eucalyptus* spp. wood for the selection of genetic materials for pulp production. **Revista Árvore,** v. 49, n. 1, e4904, 2025. https://doi.org/10.53661/1806-9088202549263840.

SOUZA, A. L.; MEDEIROS, R. M.; MATOS, L. M. S.; SILVA, K. R.; CORRÊA, P. A.; FARIA, F. N. Estratificação volumétrica por classes de estoque em uma Floresta Ombrófila Densa, no Município de Almeirim, Estado do Pará, Brasil. **Revista Árvore**, v. 38, n. 3, p.533-541, 2014. https://doi.org/10.1590/S0100-67622014000300016.

SOUZA, A. V. V.; CARVALHO, J. R. S.; COSTA, E. S. S.; SOUZA, C. S.; SILVA, H. F. J.; LUZ, J. M. Q.; MACIEL, G. M.; SIQUIEROLI, A. C. S. Genetic diversity in Amburana (*Amburana cearensis*) accessions: hierarchical and optimization methods. **Revista Árvore**, v. 45, e4508, 2021. https://doi.org/10.1590/1806-908820210000008.

SOUZA, F. M. S.; RODRIGUES, V. B.; TORRES, F. T. P. Fire effects on natural regeneration in seasonal semideciduous forest. **Revista Árvore**, v. 46, e4614, 2022. https://doi.org/10.1590/1806-908820220000014.

SOUZA, L. S.; MOREIRA, R. F. C.; FREITAS, T. A. S.; MENDONÇA, A. V. R.; AFONSO, S. D. J.; LEDO, C. A. S. Grouping of Catingueira genotypes based on morphological characteristics. **Revista Árvore**, v. 40, n. 3, p.427-434, 2016. https://doi.org/10.1590/0100-67622016000300006.

STEVENS, S. S. On the theory of scales of measurement. **Science**, v.103, n.2684, p.677-680, 1946. https://doi.org/10.1126/science.103.2684.677.

SUGAR, C. A.; JAMES, G. M. Finding the number of clusters in a data: an information-theoretic approach. **Journal of the American Statistical Association**, v. 98, n. 463, p.750-763, 2003. https://doi.org/10.1198/016214503000000666.

TAVARES, L.S.; VALADÃO, F.C.A.; WEBER, O.C.L.; MARTÍNEZ ESPINOSA, M. Análise multivariada de espécies florestais nativas em relação aos atributos químicos e texturais do solo na região de Cotriguaçu- MT. **Ciência Florestal**, v. 29, n. 1, p. 281-291, 2019. https://doi.org/10.5902/1980509823577.

TEIXEIRA, R. A.; PEDROZO, C. Â.; COSTA, E. K. L.; BATISTA, K. D.; TONINI, H.; PESSONI, L. A. Correlações e divergência fenotípica entre genótipos cultivados de

castanha-do-Brasil. **Scientia Forestalis**, v. 43, n. 107, p.523-531, 2015. Disponível em: https://www.ipef.br/publicacoes/scientia/nr107/cap03.pdf. Acesso em: 26 abr. 2025.

THOMPSON, J.P., REIS, R.G. Pattern analysis in epidemiological evaluation of cultivar resistance. **Phytopathology**, v. 69, n. 6, p.545-549, 1979. https://doi.org/10.1094/Phyto-69-545.

VASCONCELOS NETO, E. L.; AZEVEDO, C.; RIBAS, L.; D'OLIVEIRA, M. N. Ecological and functional clustering of forest species in the South Western Amazonia. **Revista Árvore**, v. 40, n. 6, p.1083-1090, 2016. https://doi.org/10.1590/0100-67622016000600014.

VIEIRA, D. S.; GOMES, K. M. A.; SANTOS, L. E.; OLIVEIRA, M. L. R.; GAMA, J. R. V.; MENDONÇA, E. L. M.; LAFETÁ, B. O.; MOURA, C. C.; FIGUEIREDO, A. E. S. Estrutura diamétrica e espacial de espécies madeireiras de importância econômica na Amazônia. **Scientia Forestalis**, v. 49, n. 129, e3438, 2021a. https://doi.org/10.18671/scifor.v49n129.21.

VASCONCELOS, E. S.; CRUZ, C. D.; BHERING, L. L.; RESENDE JÚNIOR, M. F. R. Método alternativo para análise de agrupamento. **Pesquisa Agropecuária Brasileira**, v. 42, n. 10, p.1421-1428, 2007. https://doi.org/10.1590/S0100-204X2007001000008.

VIEIRA, D. S.; OLIVEIRA, M. L. R.; GAMA, J. R. V.; FIGUEIREDO, A. E. S.; LAFETÁ, B. O. Estrutura diamétrica e distribuição espacial de *Dipteryx odorata* (Aubl.) Willd. no oeste do estado do Pará, Brasil. **Scientia Forestalis**, v. 49, n. 131, e3626, 2021b. https://doi.org/10.18671/scifor.v49n131.11.

VIEIRA, D. S.; OLIVEIRA, M. L. R.; GAMA, J. R. V.; LAFETÁ, B. O. Classificação da capacidade produtiva de florestas inequiâneas por meio de mapas auto-organizáveis. **Scientia Forestalis**, v. 50, e3794, 2022. https://doi.org/10.18671/scifor.v50.28.

WANG, J.; ZHANG, Y.; WANG, R.; LAN, X.; YUAN, J.; YU, Z., DU, F. RAPD analysis of genetic diversity in *Thuja koraiensis* Nakai population. **Revista Árvore,** v. 49, n. 1, e4929, 2025. https://doi.org/10.53661/1806-9088202549263914.

WARD, J.H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v.58, n. 301, p.236-244, 1963. https://doi.org/10.1080/01621459.1963.10500845.

WATZLAWICK, L. F.; MARTINS, P. J.; RODRIGUES, A. L.; EBLING, Â. A.; BALBINOT, R.; LUSTOSA, S. B. C. Teores de carbono em espécies da floresta ombrófila mista e efeito do grupo ecológico. **Cerne**, v.20, n.4, p. 613-620, 2014. https://doi.org/10.1590/01047760201420041492.

ZANATTO, B.; PAULA, R. C.; PAULA, N. F.; FREITAS, M. L. M.; ARAÚJO, M. J. Divergência genética entre progênies de polinização aberta de *Eucalyptus tereticornis* Smith para caracteres de crescimento e de qualidade da madeira. **Scientia Forestalis**, v. 48, n. 128, e3327, 2020. https://doi.org/10.18671/scifor.v48n128.15.

ZAVALA, C.B.R; FERNANDES, S.S.L.; PEREIRA, Z.V.; SILVA, S.M. Análise fitogeográfica da flora arbustivo-arbórea em ecótono no Planalto da Bodoquena, MS, Brasil. **Ciência Florestal**, v. 27, n. 3, p.907-921, 2017. https://doi.org/10.5902/1980509828640.

ZHU, Y.; TING, K. M.; JIN, Y.; ANGELOVA, M. Hierarchical clustering that takes advantage of both density-peak and density-connectivity. **Information Systems**, v. 103, e101871, 2022. https://doi.org/10.1016/j.is.2021.101871.

ZUUR, A. F.; IENO, E. N.; SMITH, G. M. **Analyzing ecological data**. New York: Springer, 2007. 672p.

As técnicas estatísticas multivariadas têm ampla aplicação na Ciência Florestal. Entre elas, destaca-se a Análise de Agrupamento (*Cluster Analysis*), com aplicações em conservação da natureza, manejo florestal, técnicas e operações florestais, silvicultura, tecnologia e utilização de produtos florestais.

No livro "Análise de Agrupamento em Ciência Florestal com Aplicações em R", são abordados tópicos sobre conceitos, princípios e objetivos, medidas de distância e similaridade, classificação dos processos de agrupamento, validação do método de agrupamento, determinação do número de grupos e exemplos de uso de análise de agrupamento em Ciência Florestal, buscando-se detalhar suas funcionalidades e cálculos conforme o método utilizado, além de mostrar procedimentos do Software R.

O livro foi elaborado com o objetivo de disponibilizar material didático para estudantes e profissionais da Engenharia Florestal e áreas afins. Trata-se, portanto, de uma obra que poderá contribuir substancialmente para o ensino de graduação e de pós-graduação, atuando como suporte para treinamento, capacitação de recursos humanos, bem como para o atendimento ao setor florestal e segmentos correlacionados.

UFRPE
Editora Universitária da UFRPE
UFV Universidade Federal de Viçosa
CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
ISBN: 978-65-02-25134-8