BOLETIM DE PESQUISA E DESENVOLVIMENTO
24

ISSN 1981-1004
Outubro / 2023

Aproximações em Análise Multivariada para Discriminar Esquemas de Controle de Doenças entre Vinhedos

Embrapa

ISSN 1981-1004
Outubro / 2023

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Uva e Vinho
Ministério da Agricultura e Pecuária

BOLETIM DE PESQUISA E DESENVOLVIMENTO 24

Aproximações em Análise Multivariada para Discriminar Esquemas de Controle de Doenças entre Vinhedos

Fabio Rossi Cavalcanti

Embrapa Uva e Vinho
Bento Gonçalves, RS
2023

Embrapa Uva e Vinho
Rua Livramento, 515 - Caixa Postal 130
95701-008 Bento Gonçalves, RS
Fone: (0xx) 54 3455-8000

www.embrapa.br/uva-e-vinho
www.embrapa.br/fale-conosco/sac

Revisão de texto
Renata Gava

Normalização bibliográfica
Rochelle Martins Alvorcem (CRB-10/1810)

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Renata Gava

Foto da capa
Fábio Rossi Cavalcanti

1ª edição
Publicação digital (2023): PDF

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Uva e Vinho

Aproximações em análise multivariada para discriminar esquemas de controle de doenças entre vinhedos / Fábio Rossi Cavalcanti – Bento Gonçalves: Embrapa Uva e Vinho, 2023.

PDF (26 p.) -- (Boletim de Pesquisa e Desenvolvimento / Embrapa Uva e Vinho, ISSN 1981-1004; 24).

1. Análise de componentes principais. 2. Análise de agrupamento. 3. Análise de correspondência múltipla. 4. Análise fatorial de dados mistos. 5. Ciência de dados. I. Cavalcanti, Fábio Rossi. II. Embrapa Uva e Vinho. III. Série.

CDD (21. ed.) 634.82

Rochelle Martins Alvorcem (CRB-10/1810)

Sumário

Aproximações em Análise Multivariada para discriminar esquemas de controle de doenças entre vinhedos

Fabio Rossi Cavalcanti[1]

Resumo – O Boletim destaca métodos de estatística multivariada (AM) para analisar diferentes esquemas de controle de doenças em vinhedos de várias vinícolas. Esses métodos identificam e categorizam os esquemas considerando variáveis como marca do fungicida, ingrediente ativo, grupo químico e número de aplicações. A construção de variáveis a partir de dados qualitativos é discutida através da Análise de Agrupamento (AA) e Correspondência Múltipla (MCA). A Análise de Componentes Principais (PCA) converte dados qualitativos em numéricos, ampliando as variâncias explicadas nos dois eixos por meio de um método que usa distâncias de Jaccard/Tanimoto. O Boletim apresenta este método. A Análise Exploratória Fatorial (EFA) é empregada para discriminação e interpretação. A PCA prioriza a maximização da variação, enquanto a EFA estabelece uma estrutura com fatores latentes. Ambas têm semelhanças, mas diferem em implementação e interpretação. A EFA forma "*factor loadings*", e a PCA mostra que a marca do fungicida e aplicações são essenciais para a principal componente. O estudo sublinha que, apesar da força das técnicas multivariadas, a coleta humana de dados, sobretudo na avaliação de incidência de doenças, é vital para uma análise integral.

Termos para indexação: análise de componentes principais, análise de agrupamento, análise de correspondência múltipla, análise fatorial de dados mistos, ciência de dados.

[1] Engenheiro-agrônomo, doutor em Fitopatologia, pesquisador da Embrapa Uva e Vinho, Bento Gonçalves, RS

Multivariate Analysis to Discriminate Disease Spraying Schedules among Vineyards

Abstract – The Bulletin highlights multivariate analysis (MVA) methods for analyzing different disease control schemes in vineyards across multiple wineries. These methods identify and categorize schemes considering variables such as fungicide, active ingredient, chemical group and number of applications. The construction of variables from qualitative data is discussed through Cluster Analysis (AA) and Multiple Correspondence (MCA). Principal Component Analysis (PCA) converts qualitative data into numerical data, expanding the variances explained in both axes through a method that uses Jaccard/Tanimoto distances. The Bulletin presents this method. Exploratory Factor Analysis (EFA) is employed for discrimination and interpretation. PCA prioritizes maximizing variation, while EFA establishes a structure with latent factors. Both have similarities, but differ in implementation and interpretation. The EFA forms “factor loadings”, and the PCA shows that the fungicide brand and applications are essential to the main component. The study underscores that, despite the strength of multivariate techniques, human data collection, especially when assessing disease incidence, is vital for a comprehensive analysis.

Index terms: principal component analysis, cluster analysis, multiple correspondence analysis, factor analysis of mixed data, data science.

Introdução

Em Fitopatologia, esquemas de controle químico para a proteção de plantas contra doenças causadas por patógenos adaptados à dispersão são fundamentados em calendários de aplicação de fungicidas. Tais calendários de pulverização possuem um número considerável de variáveis, que vão desde a marca do fungicida, até a dose por hectare, passando por identificação de tipos de vinhedo, grupo químico, detalhes de tecnologia de aplicação (bico, pressão) etc. Nos ciclos produtivos de vinhedos comerciais (manejo convencional da videira), diferentes esquemas de pulverização de fungicidas sintéticos são aplicados, em diferentes vinhedos da mesma vinícola e/ou em vinícolas diferentes em diferentes situações geográficas. Embora a sequência de doenças da videira a serem combatidas ao longo do ciclo produtivo siga basicamente um calendário similar em todo o Estado do Rio Grande do Sul, os esquemas de pulverização frequentemente são montados diferentes, irregulares e heterogêneos entre si, por diferentes técnicos mesmo dentro de uma mesma vinícola.

A análise de um grande número de variáveis é comum em Agronomia, mas muitas vezes apenas algumas são realmente significativas para descrever uma situação ou resultado. A seleção dessas variáveis frequentemente é feita de maneira empírica, baseada na intuição ou experiência do profissional. Além disso, as decisões na agricultura geralmente dependem de vários fatores, tornando a tomada de decisões um processo complexo. A Estatística disponibiliza várias ferramentas para a análise e avaliação de dados. No entanto, muitas dessas ferramentas criadas por estatísticos precisam de aplicação prática. Isso se aplica particularmente à Estatística Multivariada. A recente disponibilidade de computadores e pacotes de software como tidyverse, caret, shiny e dplyr para a plataforma R, e numpy, pandas, scipy e tensorflow para a linguagem Python facilitou a análise de grandes matrizes de covariância, correlação, determinantes, autovalores, inversas e rotações (Kassambara, 2017).

A Análise Exploratória de Dados (EDA) possui duas grandes abordagens para analisar variáveis: a estatística univariada e a multivariada. A primeira estuda as variáveis isoladamente, enquanto a segunda engloba todas as variáveis possíveis. Em pesquisas que dependem de várias variáveis, a abordagem univariada pode ocultar informações importantes, por isso a

crescente adoção da Estatística Multivariada é fundamental. Existem vários métodos para tratar dados multivariados. Para variáveis quantitativas, a Análise de Componentes Principais é frequentemente recomendada. Para variáveis qualitativas, existem a Análise de Correspondência Múltipla e para estruturas mistas de dados, a Análise Fatorial de Dados Mistos e a Análise Fatorial Múltipla (Ahmad; Nabi, 2021). A Análise Fatorial Exploratória é um método que identifica relações subjacentes entre as variáveis medidas (Mardia et al., 1979; Kassambara, 2017).

A Análise de *Cluster* e Agrupamento Hierárquico é um método não supervisionado que classifica objetos observando as semelhanças ou diferenças entre eles. Essa análise identifica grupos cujos objetos internos são semelhantes e agrupados em nós, deixando os objetos diferentes em grupos distintos. Além disso, existem algoritmos de segmentação como o K-Means, *Partitioning Around Medoids* (PAM), *Clustering Large Applications* (Clara) e *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), embora não sejam discutidos neste estudo. Em geral, um método multivariado, como o PCA, é um procedimento matemático que reduz um conjunto de variáveis correlacionadas em um pequeno grupo de variáveis independentes através de uma transformação ortogonal. Essas variáveis podem ser chamadas de 'dimensão', 'fator' ou 'componente' (Ferreira et al., 2020; James et al., 2021). Inicialmente, o objetivo da PCA era encontrar planos que se ajustassem a um conjunto de pontos em um espaço multidimensional através de matrizes de correlação e covariância. A componente era determinada pela combinação linear das variáveis que apresentavam a maior variabilidade na matriz de covariância.

Os métodos multivariados também formam a base para técnicas de aprendizado de máquina, ajudando a criar modelos capazes de aprender e fazer previsões com base em várias entradas, simultaneamente. Por exemplo, PCA é comumente usado na redução de dimensionalidade, um passo importante na preparação de dados para algoritmos de aprendizado de máquina. Técnicas de clusterização, como K-Means, são usadas em aprendizado não supervisionado para identificar padrões e segmentar dados seguindo premissas de *clustering*, associação e redução de dimensionalidade (James et al., 2021).

Para tentar sistematizar e discriminar diferentes esquemas de controle químico, oito calendários de pulverização foram utilizados para o

desenvolvimento de métodos de Análise Multivariada (AM) para caracterização da proteção específica de vinhedos contra doenças. O objetivo do presente trabalho foi, a partir dos dados dos calendários de aplicação, discriminar os oito vinhedos mediante métodos de AM em ordem de presença de dados quantitativos: Análise de Agrupamentos (AA), Análise de Correspondência Múltipla (MCA), Análise Fatorial de Dados Mistos (FAMD), Analise Fatorial Múltipla (MFA), Análise Fatorial Exploratória (EFA) e Análise de Componentes Principais (PCA).

A Análise Exploratória e a Estatística Multivariada aparecem como ferramentas promissoras para tentar caracterizar calendários de pulverização usados no controle de doenças. No presente estudo foram utilizados dados de calendários de fungicidas usados por vinícolas de duas regiões do Estado: a Campanha e Serra Gaúcha.

O presente método não pretende caracterizar a natureza de esquemas de pulverização, seus contextos de adoção e aplicações. Apenas, discriminar esses esquemas.

Material e Métodos

Devido à enorme diversidade e variedade de padrões de dados associados aos esquemas de controle químico que são adotados pelas vinícolas gaúchas, foram utilizados calendários de pulverização de oito vinícolas da Campanha e Serra Gaúcha, referentes às safras 2013/2014, 2017/2018 e 2019/2020 para o desenvolvimento de métodos de AM para caracterização da proteção de vinhedos contra doenças.

Os oito vinhedos foram de variedades distintas de *Vitis* sp., com níveis particulares de susceptibilidade às diferentes doenças que ameaçam a safra da videira no Rio Grande do Sul: ‘Cabernet Sauvignon’, ‘Chardonnay’, ‘Malbec’, ‘BRS Magna’, ‘Bordô’, ‘Isabel’, ‘BRS Violeta’ e ‘BRS Carmem’.

Os calendários foram padronizados seguindo os dados fornecidos: intervalo de aplicação, marca comercial do fungicida, ingrediente ativo, grupo químico, dosagem (m/v) em 100 L de calda, produto gasto (g ou mL) e volume de água por hectare gasto. Em alguns calendários mais detalhados foram disponibilizadas informações sobre: quadra, parcela, área, número do

pulverizador (em referência à configuração de aplicação do trator), número do trator, número da prescrição técnica, data de aplicação e até operador. Além disso, a partir das informações disponíveis no calendário, foram gerados canais de dados associados a número de aplicações e intervalo entre aplicações.

Essas variáveis foram usadas incorporadas ou não aos conjuntos de dados (*dataset*) específicos a cada método AM detalhado abaixo.

Análise de *cluster* e Agrupamento Hierárquico (AA)

Por haver calendários com mais aplicações e outros com menos, foram padronizados 20 tratamentos por esquema nas mesmas semanas de cada safra específica, entre si, para a construção do *dataset* de dados. Após a justaposição dos tratamentos (linhas), foram definidas as variáveis para estudo de AA: tratamento, vinhedo, região, variedade (variedd), quantidade de calda utilizada (calda), senha do produto comercial (produto), ingrediente ativo (IA) e grupo químico (grupo). Para a senha do produto comercial foram criados nomes fictícios em substituição às marcas comerciais.

Após o carregamento dos dados (em um *data frame* imediatamente copiado para uma matriz 'x'), uma função particular para calcular o índice de Jaccard/Tanimoto em variáveis categóricas foi definida baseada em *overlappings* (Agresti, 2013; Costa, 2021) das '*strings*' que compuseram os vetores da sequência de 20 tratamentos recebidos por vinhedo, dois a dois (par), entre os oito esquemas de proteção (vinhedos) usados no estudo (Figura 1). Esse método foi rodado para cada variável escolhida para o estudo: produto (senha do mesmo), IA e grupo.

Matrizes de distância ('da' e 'darel') (Figura 1) foram então montadas a partir dos índices calculados para cada par e dendrogramas a partir da matriz de distâncias relativas ('darel') foram gerados pelas classes hclust() e ggdendrogram() dos pacotes stats e ggdendro (R), respectivamente.

Análise de Correspondência Múltipla (MCA)

Para realizar a MCA, um *data frame* ativo foi selecionado do conjunto original de dados com todos os tratamentos recebidos pelos oito vinhedos

(cada qual com quantidades particulares de aplicação), e com cinco variáveis selecionadas para a MCA, todas convertidas a fatores com diferentes níveis: região (regiao, 2 níveis), variedade (variedd, 8 níveis), produto (26), IA (12) e grupo (13).

A aplicação do algoritmo de correspondência (sobre os dados categóricos/níveis dos fatores), obtenção dos escores dos fatores e construção das matrizes de escores, de probabilidade, matriz diagonal e de autovalores (Av) foram computadas em sequência pela classe mca() do FactoMineR, a partir do *data frame* ativo acima descrito. Com a análise, foram construídos o *screeplot* com um eixo de variâncias explicadas e outro com as dimensões disponíveis até um máximo de (5), e mapas de distâncias euclidianas distribuindo os fatores do estudo, sobre os eixos.

As qualidades de representação das categorias sobre as variáveis e indivíduos (tratamentos individualizados) foram estudadas pelos estimadores 'coord', 'cos2', 'contrib' da classe mca(), sendo adotado para apresentação dos dados o 'cos2'.

```
# funcao para Jaccard (1901, 1912) ---
jaccard <- function(a, b) {
  interseccao = length(intersect(a, b))
  uniao = length(a) + length(b) - interseccao
  return (interseccao/uniao)
}
## matrizes
nvinhedos <- 8
entradas <- 18
x <- matrix(0, entradas, nvinhedos)
da <- matrix(0, nvinhedos, nvinhedos)
darel <- matrix(0, nvinhedos, nvinhedos)
# variavel: produto ---
c1 <- 0
for(i in 1:nvinhedos) {
  for(f in 1:entradas) {
    c1 = c1 + 1
    x[f, i] = dados$produto[c1]
  }
}
x
x[,1] # = "a"
jaccard(x[,1], x[,1]) # = jaccard(a,a)

# monta uma matriz de dados com dist. jaccard corrigidas
for(i in 1:nvinhedos) {
  for(f in 1:nvinhedos) {
    da[f, i] = jaccard(x[,f],x[,i])
    darel[f,                i]                =
jaccard(x[,f],x[,i])/jaccard(x[,f],x[,f])
  }
}
da
darel
```

Figura 1. Segmento de *script* em R contendo um método (Jaccard/Tanimoto) para montagem de matrizes de distâncias sobre variáveis categóricas fornecidas por esquemas de proteção de vinhedos contra doença, desdobrado por fatores básicos como: produto comercial (fungicida), ingrediente ativo e grupo químico associado ao produto escolhido pelo técnico em um i-ésimo momento para aplicação.

Análise Fatorial de Dados Mistos (FAMD)

Para realizar a FAMD, um método híbrido entre a MCA e PCA disponível no FactoMineR, um *data frame* ativo foi preparado do conjunto original de dados - analogamente como o da MCA - com todos os tratamentos recebidos pelos oito vinhedos, mas com sete variáveis selecionadas para a FAMD. Duas variáveis qualitativas foram convertidas a fatores com diferentes níveis: região (regiao, 2 níveis) e variedade (variedd, 8 níveis). Cinco variáveis quantitativas foram selecionadas: taxa de aplicação do tratamento em L ha^{-1} (calda), número de aplicações na safra (napl), senha do produto comercial (produto), ingrediente ativo (IA) e grupo químico (grupo).

Para gerar dados numéricos das variáveis 'produto', 'IA' e 'grupo', foi estabelecido o uso do índice de Jaccard/Tanimoto do vinhedo específico calculado para a Análise de Cluster sobre os vinhedos (Figura 1), seguido de simulações em *loop* para a obtenção de dados normais em função do número de tratamentos observados em cada vinhedo. O critério de saída desse *loop* ficou estabelecido como a geração de dados normais sobre os índices que maximizassem a obtenção de fatores e componentes para realização da PCA do FAMD e demais abordagens em PCA no presente trabalho (Figura 2).

```
# noise
for (f in 7:9) {
  dfgeral[,f] <- rnorm(dim(dfgeral)[1], mean = dfgeral[,f], sd = 0.1)
  print(dfgeral[,f])
}
## FAMD
##
ativofamd <- FAMD(dfativo, graph = TRUE)
print(ativofamd)
# obter os eigenvalues
eigenv <- get_eigenvalue(ativofamd)
dfativo <- dfgeral[,-c(1,2)]
head(dfativo, 15)
```

Figura 2. Simulação para gerar valores numéricos associados às variáveis 'produto', 'IA' e 'grupo', para composição de data frame ativo para análises FAMD, EFA e PCA.

Para dados qualitativos, seguindo a MCA, a aplicação do algoritmo de correspondência (iterações para a montagem da CDT, *complete disjuntive table*), obtenção dos escores dos fatores e construção das matrizes de escores, de probabilidade, matriz diagonal e de autovalores (Av) foram

computadas em sequência pela classe famd() do FactoMineR, a partir do data frame ativo acima descrito. Com a análise, foram construídos o *screeplot* com um eixo de variâncias explicadas e outro com as dimensões disponíveis até um máximo de (5), e mapas de distâncias euclidianas distribuindo os fatores do estudo, sobre os eixos.

Também como a MCA, as qualidades de representação das categorias sobre as variáveis e indivíduos (tratamentos individualizados) foram estudadas pelos estimadores 'coord', 'cos2', 'contrib' da classe mca(), sendo adotado para apresentação dos dados o 'cos2'.

Análise Fatorial Exploratória (EFA) e Análise de Componentes Principais (PCA)

Para a matriz de variáveis ativas da EFA e PCA, os calendários de pulverização adotados nos vinhedos estudados foram igualados em 20 tratamentos (n = 20 indivíduos), conforme critério exposto na descrição da AA. A matriz serviu para análise de *clusters* (dendrogramas construídos com hclust(), para cada variável estudada, mas aqui por Distância Euclidiana. Dos dados originais, foram selecionadas cinco variáveis numéricas para compor o *data frame* ativo: quantidade de calda gasta por tratamento (calda), número de aplicações na safra (napl), senha do produto comercial (produto), ingrediente ativo (IA) e grupo químico (grupo). As EFAs e PCAs foram realizadas a partir de autovalores (Av) e matrizes de autovetores extraídos do data frame ativo com os dados não transformados-padronizados (Lobato et al., 2017). Em seguida, foram estudados os critérios de Kaiser (1974) (Av > 0,600), esfericidade de Bartlett ($p < 0,05$), do determinante positivo da matriz convertida do *data frame* ativo, e variância explicada acumulada > 70%. Tais critérios também foram considerados para a saída do *loop* de simulação explicado em FAMD, na tentativa de maximizar fatores e componentes principais para as análises EFA e PCA.

Com as análises, foi construído o *screeplot* com um eixo de autovalores iniciais (*eigenvalues*) e outro com as dimensões disponíveis até um máximo de (5), e calculados os coeficientes de correlação entre as variáveis. Adicionalmente, foi usada uma função fa.parallel() (FactoMineR) que auxilia na definição dos dois componentes principais, além dos critérios acima, das variâncias acumuladas e carregamento de fatores (*factor loadings*).

Duas abordagens de 'fatoração' foram usadas para calcular a EFA: a da Solução do Fator Principal (PA, *principal factor solution*) e Máxima Verossimilhança (ML, *maximum likelihood*) que se mostrou mais robusta associando os fatores e variáveis e maximizando variâncias. No entanto, a abordagem de 'fatoração' escolhida foi dada pela classe factanal() por adoção de ML e rotação 'varimax' para a matriz do *data frame* ativo. Depois dos procedimentos de 'fatoração', foram confirmados os *factor loadings* dos fatores (dimensões/eixos) principais, e componentes principais.

Para a PCA, na presente oportunidade, diferentemente de Lobato et al., (2017), foi utilizada uma matriz proveniente do *data frame* ativo com os dados transformados-padronizados (média = 0; s = 1) (Figuras 1 e 2). Foi construído o *screeplot* com variâncias acumuladas e outro com os componentes principais até um máximo de (5), e calculados os coeficientes de correlação entre as variáveis. As qualidades de representação das categorias sobre as variáveis e indivíduos (tratamentos individualizados) foram estudadas pelos estimadores 'coord', 'cos2', 'contrib' da classe mca(), sendo adotado para apresentação dos dados o 'cos2' sobre os eixos dos fatores em estudo orientados sobre o plano cartesiano com as componentes principais.

Resultados e Discussão

Uma abordagem para caracterização do controle químico e esquemas de pulverização por AM pôde discriminar e agrupar as diversas abordagens técnicas de controle de doenças em vinhedos distribuídos em vinícolas parceiras, em duas grandes regiões de produção de uva no Estado do Rio Grande do Sul. Os métodos ensaiados neste trabalho conseguiram, em termos gerais, separar os diferentes esquemas de pulverizações em função das variáveis estudadas, mostrando que a maioria dos vinhedos recebeu um calendário relativamente homogêneo, com alguns vinhedos apresentando diferenciações no regime de pulverização.

No caso da AA, variáveis qualitativas convertidas a categóricas e trabalhadas conforme metodologia descrita para obtenção de índices de similaridade a cada par, foram capazes de agrupar diferencialmente vinhedos (Figura 3). Por exemplo, o dendrograma reconstruído sobre os dados de produto comercial (Figura 3A) revelaram um nó distinto para quadras associadas às variedades

viníferas (*Vitis vinifera*). Embora, os dendrogramas reconstruídos com as variáveis IA e grupo químico (Figuras 3B e 3C) não tenham reproduzido a estrutura da primeira árvore, não foi, para a AA, uma questão de confirmação: os mesmos puderam agrupar e classificar por distância diferentes vinhedos tratados de modo específico, em função de cada variável abordada ('produto', 'IA' e 'grupo'), coerentes com as informações fornecidas. Aliás, no escopo do presente trabalho, a AA foi aplicada propositalmente em uma abordagem supervisionada, ou seja, envolvendo a descoberta de padrões para prever o valor de uma variável específica (Provost; Fawcett, 2016).

Seria possível aplicar uma AA a todas as variáveis do presente estudo para reconstrução de um dendrograma 'não supervisionado'. No entanto, talvez fosse perdida informação de interesse específico à consulta de cada variável. Adicionalmente, a AA também foi usada para geração de números associados à variáveis naturalmente qualitativas, por variável, trazidas dos esquemas fornecidos pelas vinícolas, e tais números comporem uma base de dados para as outras abordagens AM presentes neste estudo.

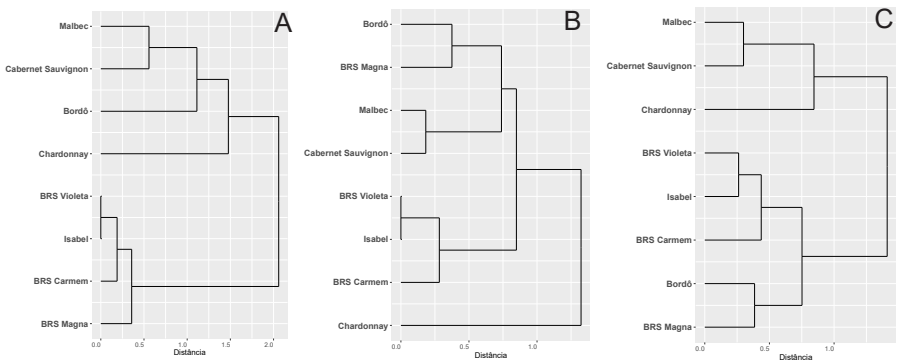

Figura 3. Análise de *cluster* (agrupamento) gerado por método ggdendrogram() para matrizes de distâncias geradas por índices de Jaccard/Tanimoto sobre vetores contendo dados categóricos obtidos de esquemas de proteção de oito vinhedos gaúchos. (A) Marca comercial do produto. (B) Ingrediente ativo do i-ésimo produto. (C) Grupo químico do i-ésimo produto adotado nos esquemas.

A MCA gerada a partir da matriz de correlação de variáveis (categóricas) ativas *versus* fatores (eixos, dimensões) revelou que variáveis descritoras da região e de variedade foram mais relevantes do que informações de marca comercial do fungicida, ingrediente ativo ou grupo químico (Figura 4). Ainda, por esse método de AM (MCA), foram obtidos baixos valores de variâncias e variâncias acumuladas (Figura 4A). Isso normalmente acontece em MCA, porque quando há apenas informações qualitativas disponíveis, métodos de MCA generalizam a correspondência a partir do uso de uma tabela disjuntiva completa (CDT, *complete disjuntive table*) que desdobra as categorias das variáveis (em colunas) e para cada indivíduo (no presente caso, tratamento) marcando zero ou um para cada categoria desdobrada, resultando em uma matriz binária (*dummy*) que possui um espectro informativo limitado considerando variâncias quando comparadas a dados quantitativos (Johnson; Wichern, 2007; Šuligoj, 2018; Corral-de-Witt et al., 2019; Ochoa-Munoz et al., 2019).

Detalhando, o MCA funciona calculando uma 'tabela de Burt', simétrica com as margens de linha e coluna, com a distância qui-quadrado entre cada par de categorias e, em seguida, aplicando uma decomposição de valor singular (SVD) na matriz de resíduos padronizados. Os autovalores são extraídos da diagonal dos valores singulares ao quadrado e os componentes principais são calculados como as coordenadas dos pontos de dados projetados nas novas dimensões (autovetores). É uma técnica de redução de dimensionalidade que fornece uma representação gráfica dos dados, ajudando a revelar padrões ou estruturas subjacentes (Agresti, 2013). De qualquer forma, foi possível pela MCA, fazer uma plotagem em gráfico cartesiano das variáveis estudadas no *data frame* ativo (Figura 4C) e, principalmente, agrupar indivíduos (tratamentos) associados ao controle de doenças adotado em vinícolas da Campanha e em vinícolas da Serra Gaúcha (Figura 4D).

A FAMD é um método de AM que envolve o uso de componentes principais (CP) e se dedica a analisar agrupamentos de dados que contêm variáveis quantitativas e qualitativas, sendo capaz de analisar a similaridade entre indivíduos levando-se em consideração a natureza desse tipo de conjunto de dados. De modo prático, uma análise FAMD é feita por um 'desdobramento' das análises dos dados qualitativos (por MCA) e quantitativos (por PCA), mas é possível explorar a associação entre todas as variáveis, pois as variâncias,

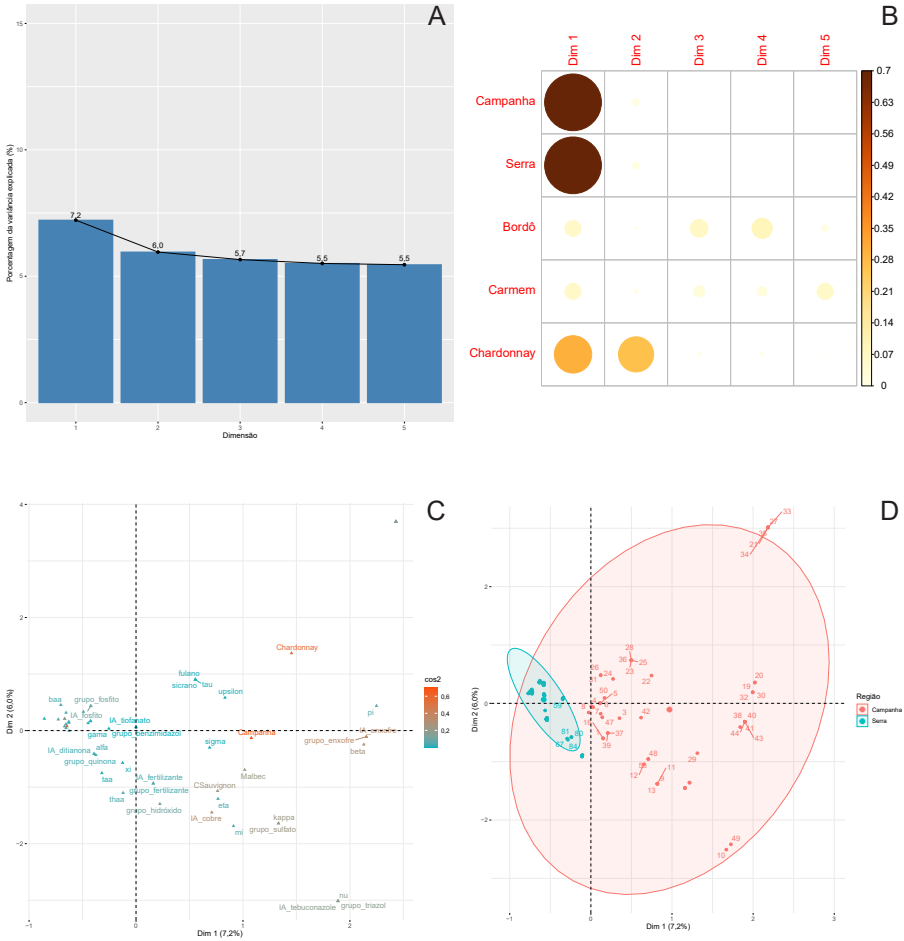

Figura 4. Análise de Correspondência Múltipla (MCA). (A) Proporção da variância explicada por eixos definidos pelo cálculo de autovalores (Av) da análise de correspondência múltipla (*Screeplot* do MCA) do FactoMineR e factoextra, a partir dos dados categóricos fornecidos pelos esquemas de pulverização dos oito vinhedos estudados. (B) Faixas de coeficientes de correlação entre os componentes principais e as variáveis convertidas a fatores. (C) Dispersão das variáveis pelo plano de duas componentes principais. A cor de cada nível do fator (variável) representa a qualidade do fator como representação da variância do mapa de dispersão, pelo estimador cos2 (quadrado da coordenada de posicionamento da variável, no mapa). (D) Gráfico da dispersão dos dados individuais do estudo com estimação de área elíptica de abrangência (elipse de confiança, 95%), desdobrado no fator 'região'.

autovalores e estimadores são levados conjuntamente em consideração (Kassambara, 2017).

Com relação à MCA, a análise em FAMD pôde disponibilizar eixos (dimensões) com variância acumulada explicada marcadamente maior do que na MCA. Um dos motivos para este incremento foi a adoção, para o *data frame* ativo da FAMD, de mais duas variáveis quantitativas, a 'quantidade de calda gasta por tratamento' (calda) e 'número de aplicações na safra' (napl). O conjunto de dados para a FAMD promoveu, então, um aumento no percentual das variâncias explicadas em todos os eixos (dimensões) (Figura 5A), principalmente nas contribuições para as duas CPs (Figura 5B). Pela PCA da FAMD, foi possível estabelecer os eixos associados às variáveis numéricas e as contribuições de cada eixo às duas CPs. As qualidades de contribuições das variáveis estudadas foram analisadas geometricamente, pelas coordenadas geradas a partir da matriz de autovetores *(factor score coefficients*), pela função FactoMineR (var$coord()). Com isso, o estimador 'cos2', que equivale ao quadrado da coordenada aferida, foi usado para expressar a qualidade das contribuições da variável sobre as CPs (Figura 5B) de um modo evidente.

Para o FAMD, foi construído um estudo de indivíduos e a qualidade de suas contribuições, sobre o plano cartesiano das Componentes Principais (Figura 5D). É possível ver claramente dois padrões distintos de pulverização, quando o gráfico de indivíduos e variáveis (Figuras 5C e 5D), um caracterizando o controle de doenças na Serra Gaúcha e outro na Campanha. Esse perfil reproduz parcialmente o desdobramento do fator 'região' na MCA sendo que lá há uma sobreposição de tratamentos na elipse de confiança (95%) sobre os dados das duas regiões, enquanto que na FAMD há uma distinção completa entre as elipses de confiança construídas sobre os dados.

Similar à Análise de Componentes Independentes (*Independent Component Analysis*, ICA) e à PCA, a EFA (Análise Exploratória Fatorial) busca encontrar CPs nos dados. Como uma generalização da PCA, a EFA requer que o número de componentes seja menor do que o número original de variáveis numéricas ofertada nos dados. A otimização da EFA depende de iterações em um subespaço vetorial por estimadores, que podem ser, por exemplo, Máxima Verossimilhança, Solução do Fator Principal, dentre outros, onde toda observação nos dados representa um ponto de amostra nesse

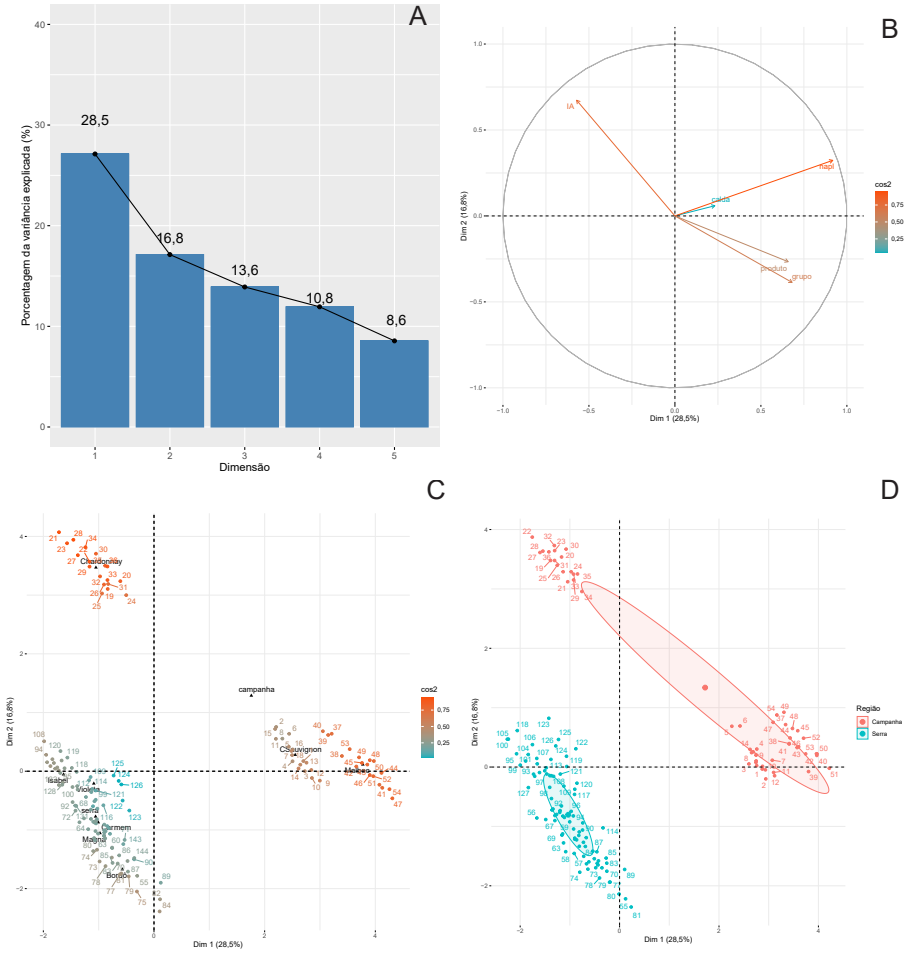

Figura 5. Análise Fatorial de Dados Mistos (FAMD). (A) Proporção da variância explicada por componentes principais definidas pelo cálculo de autovalores (Av) por Análise Fatorial de Dados Mistos (*Screeplot* do FAMD), a partir dos dados mistos (categóricos e numéricos) fornecidos pelos esquemas de pulverização dos oito vinhedos estudados. (B) Componentes principais e as variáveis numéricas que compõem as variâncias explicadas de cada eixo. A cor de cada nível do fator (variável) representa a qualidade do fator como representação da variância total explicada pelo estimador cos2. (C) Dispersão dos dados no plano das duas componentes ortogonais. (D) Gráfico da dispersão dos dados individuais (indivíduos) após Análise Fatorial Múltipla (MFA) com estimação de área elíptica de abrangência (elipse de confiança, 99%), desdobrado no fator 'região'.

subespaço (Dinov, 2018). Enquanto que a PCA assume que os estimadores sejam esféricos (área geométrica), a EFA constrói uma matriz de covariância diagonal para definir um subespaço vetorial que contenha a matriz de covariância (Dinov, 2018).

Sendo assim, embora EFA e PCA usem um arcabouço matemático muito próximo, permitindo que ambos os métodos compartilhem a mesma visão e estratégia de implementação e, não raro, sejam usadas de modo complementar, apontando resultados similares em EDA, há diferenças substanciais entre os métodos (Hardle; Simar, 2015). Diferentemente da PCA, a EFA impõe uma estrutura específica com um número fixo de fatores comuns (latentes) enquanto que a PCA determina p-fatores em decréscimo de ordem de importância. O fator mais importante da PCA é aquele que maximiza a variância projetada. O fator mais importante na EFA é aquele que, após a rotação (e existem vários métodos para rotação), entrega o máximo de interpretação das variáveis sobre os eixos das CPs. Assim, pode haver discrepâncias entre os métodos no que tange as direções das projeções das variáveis sobre as CPs, principalmente a primeira CP. Do ponto de vista de implementação (uso de ferramenta estatística multivariada, por exemplo, o próprio FactoMineR), a PCA se baseia em um algoritmo simples, claro e bem definido. Por outro lado, a modelagem EFA envolve uma variedade e complexidade maior de procedimentos numéricos, o que o torna mais subjetivo e, talvez, por isso venha sempre complementado por PCAs para efeito comparativo (Hardle; Simar, 2015; Ruppert; Matteson, 2015).

No presente estudo, a EFA foi utilizada basicamente para construção de cargas dos fatores *(factor loadings*) como um critério para reforçar a robustez das estimativas de qualidade das variáveis/fatores para a construção das CPs, em comparação com os var$coord e var$cos2 da PCA. De modo complementar, a EFA entregou valores indicativos de fatores e CPs para a montagem da EDA deste estudo. Os resultados da EFA sobre os esquemas de proteção de vinhedos fornecidos pelas vinícolas para oito vinhedos estão resumidos na Figura 6, com as correlações entre as variáveis da EFA (Figura 6A) e o *screeplot* de autovalores iniciais da EFA sobre os dados (Figura 6B). Também foram calculadas as construções de correspondência entre as variáveis de estudo e as CPs definidas, tanto por ‘Solução do Fator Principal’ (PA) como por ‘Máxima Verossimilhança’ (MLE) com rotação ‘varimax’.

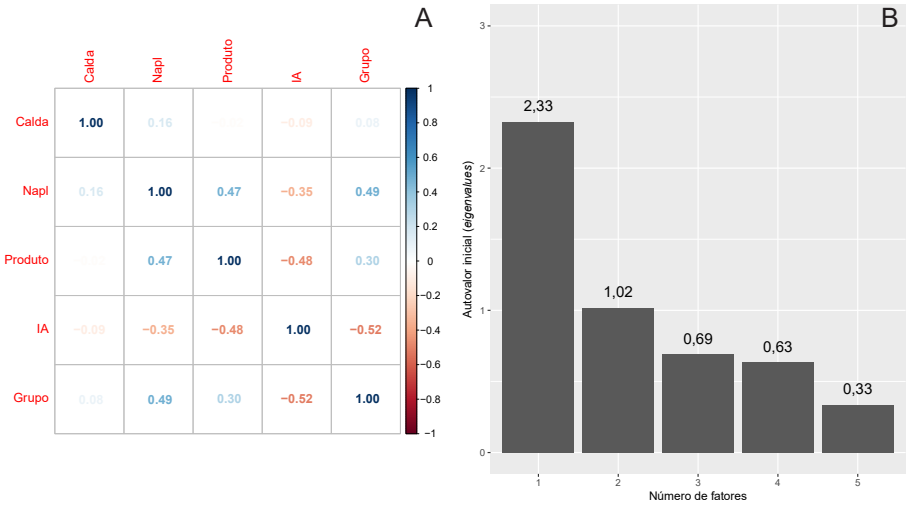

Figura 6. Análise Fatorial Exploratória (EFA) sobre dados quantitativos de esquemas de proteção de oito vinhedos contra doenças, no Rio Grande do Sul. (A) Faixas de coeficientes de correlação entre os fatores. (B) Autovalores (Av) iniciais calculados a partir de matriz de covariância sobre um conjunto de dados em cinco variáveis.

A Análise de Componentes Principais (PCA) gerada a partir da matriz de correlação de variáveis ativas (*data frame* ativo) *versus* CPs revelou que as variáveis descritoras do controle químico, ou seja, a marca comercial do fungicida (produto), o ingrediente ativo (IA), o grupo químico (G) e o número de aplicações associados a cada vinhedo (napl), explicaram a CP-1 (Figura 7), às quais atingiram um valor percentual de contribuição (var$contrib()) acima de 40%. Oportuno salientar que, sobre os dados fornecidos para o presente estudo, a CP-1 explica quase 50% da variância explicada acumulada (Figura 7), na abordagem de PCA com o método de Jaccard/Tanimoto proposto (Figura 1). Cabe explicar que, a cada simulação, os valores e as estimativas se alteram um pouco por se tratar de números aleatórios em torno dos índices de Jaccard/Tanimoto obtidas na AA e também usados nas PCAs.

A variável 'calda' (quantidade de calda gasta por tratamento em L ha^{-1}) explicou sozinha a CP-2, com valor de contribuição sempre acima de 90% (Figura 7). Por sobre o Plano Fatorial-2D construído, foi possível, então,

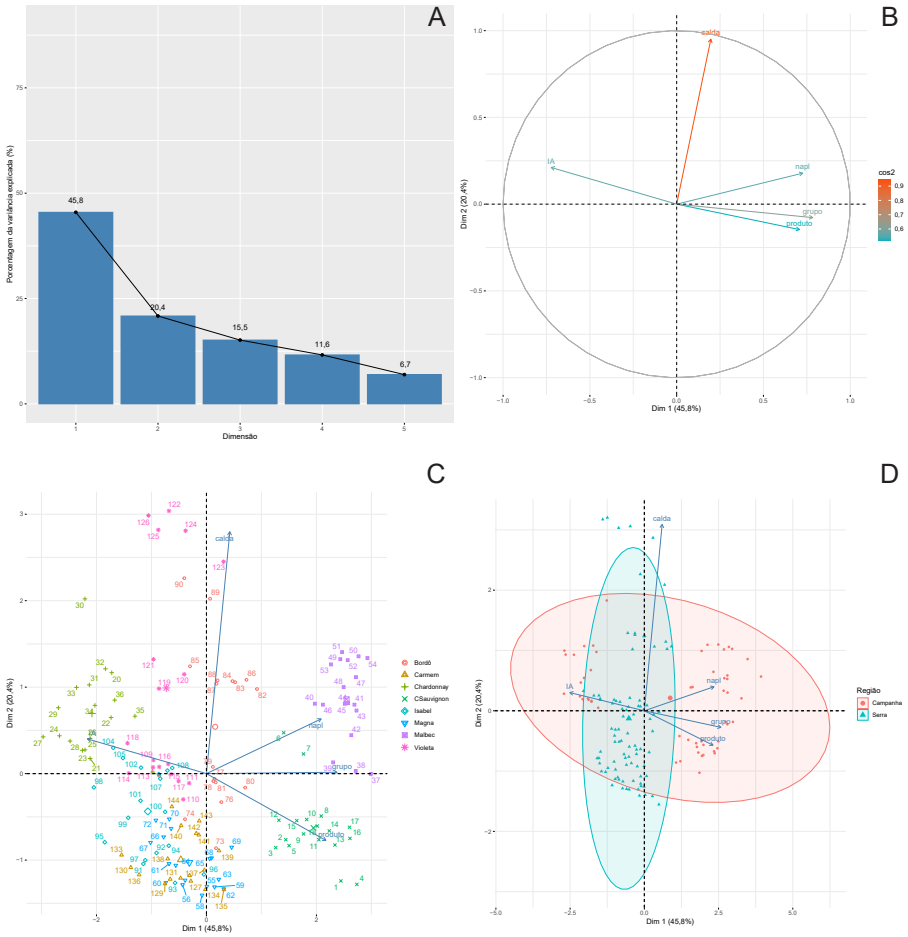

Figura 7. Análise de Componentes Principais (PCA). (A) Proporção da variância explicada por componentes principais definidas pelo cálculo de autovalores (Av) por Análise de Componentes Principais (*Screeplot* do PCA), a partir dos dados mistos (categóricos e numéricos) fornecidos pelos esquemas de pulverização dos oito vinhedos estudados. (B) Componentes principais e as variáveis numéricas que compõem as variâncias explicadas de cada eixo. A cor de cada nível do fator (variável) representa a qualidade do fator como representação da variância total explicada pelo estimador cos2. (C) Dispersão dos dados no plano das duas componentes ortogonais, por indivíduos. (D) Gráfico da dispersão dos dados individuais (indivíduos) após Análise de Componentes Principais (PCA) com estimação de área elíptica de abrangência (elipse de confiança, 95%), desdobrado no fator 'região'.

discriminar os vinhedos do presente estudo desdobrados por variedade (variedd) e região (região) (Figura 7D), tanto no estudo de indivíduos (Figura 7C) quanto no *biplot* resultante (Figura 8).

Nas diferentes safras dos vinhedos das regiões da Campanha e Serra Gaúcha, todos os vinhedos protegidos, mesmo os discrepantes, apresentaram eficiência no controle de doenças, em termos gerais. O fator safra foi uma variável que foi propositalmente deixada de fora das análises, para facilitação do foco sobre as variáveis típicas de um esquema de proteção de vinhedos contra doenças.

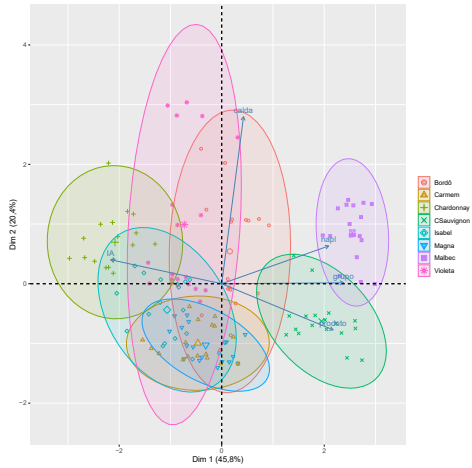

Figura 8. Gráfico da dispersão dos dados individuais (indivíduos) após Análise de Componentes Principais (PCA) com estimação de área elíptica de abrangência (elipse de confiança, 95%), desdobrado no fator 'vinhedos'.

O Plano Fatorial-2D com as variáveis associadas aos esquemas de controle químico adotados nas duas regiões poderia incorporar facilmente mais dados (variáveis suplementares), produzindo outros estudos em conjunto com a PCA, dirigidos. Esses estudos, por sua vez, produziriam novas inferências e caracterizações. Um exemplo de bloco de dados a ser naturalmente acrescentado ao método seria uma coluna de variáveis

associadas à incidência das doenças (Cavalcanti, 2021), o que poderia trazer uma boa abordagem racional para comparar eficiência de controle de doenças.

No entanto, embora a PCA tenha aptidão para trabalhar com grande massa de informações (Ciência de Dados), a aquisição de certos tipos de dados pode ser difícil e trabalhosa por prospecção humana, como, por exemplo, justamente o levantamento de incidência de doenças envolvendo avaliadores humanos. Por isso, soluções envolvendo amostragem territorial ainda se fazem necessárias, mesmo que isso acarrete perda de acurácia da AM. Paralelamente, com o advento de novas tecnologias de imagens envolvendo sensores hiperespectrais em satélites e veículos aéreos não tripulados (drones), levantamentos mais extensivos de incidência de doença podem ser realizados, aumentando o poder de entrada para o método de AM proposto.

Conclusões

a) As abordagens de AM aplicadas no presente trabalho são capazes de classificar e discriminar diferentes esquemas de controle de doenças;

b) É possível gerar números associados à similaridade/semelhança comparativa, a partir de dados qualitativos e categóricos para produção de inferências matemáticas capazes de discriminar dados provenientes de esquemas de controle de doença. As análises AA, MCA e FAMD concorrem para discriminação de esquemas de aplicação de fungicida para proteção de plantas;

c) O esforço em se produzir números que possam compor variáveis para análises EFA e PCA a partir de simulações sobre índices de similaridade/semelhança obtidos (neste trabalho por distância de Jaccard/Tanimoto) pode gerar análises confiáveis com esses métodos, com aumentos de percentuais de variâncias acumuladas explicadas.

Referências

AGRESTI, A. Categorical data analysis. 3rd ed. Hoboken: J. Wiley, 2013. 742 p. (Wiley series in probality and statistics, 792).

AHMAD, L.; NABI, F. **Agriculture 5.0:** Artificial Intelligence, IoT and Machine Learning. CRC Press, 2021. 243 p.

CAVALCANTI, F. R. Manejo fitossanitário: manejo de doenças da videira na Campanha Gaúcha. In: SILVEIRA, S.V.; PROTAS, J. F. S. (ed.). **Vinhos finos da região da Campanha Gaúcha:** tecnologias para a vitivinicultura e para estruturação de Indicação Geográfica. Bento Gonçalves, RS: Embrapa Uva e Vinho, 2021. p. 117-134. (Embrapa Uva e Vinho. Documentos, 130). Disponível em: http://www.alice.cnptia.embrapa.br/alice/handle/doc/1142105. Acesso em: 21 dez. 2022.

CORRAL-DE-WITT, D.; CARRERA, E. V.; MUNOZ-ROMERO, S.; TEPE, K.; ROJO-ALVAREZ, J. L. Multiple correspondence analysis of emergencies attended by integrated security services. **Applied Sciences**, v. 9., n. 7, p. 1-24, 2019. DOI 10.3390/app9071396.

COSTA, L. F. Further generalizations of the Jaccard Index. **HAL Science Ouverte**, hal-03384438, v. 4, oct. 2021. Disponível em: https://hal.science/hal-03384438v4. Acesso em: 21 dez. 2022.

DINOV, I. D. **Data science and predictive analytics:** biomedical and health applications using R. Springer Nature, 2018. 851p.

FERREIRA, R. R. M.; PAIM, F. A. de P.; RODRIGUES, V. G. S.; CASTRO, G. S. A. **Análise de cluster não supervisionado em R:** agrupamento hierárquico. Campinas: Embrapa Territorial, 2020. 43 p. (Embrapa Territorial. Documentos, 133).

HÄRDLE, W. K.; SIMAR, L. **Applied multivariate statistical analysis**. 4th ed. New York: Springer-Verlag, 2015. 581 p.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**: with applications in R. 2nd ed. Springer, 2021. 607 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th ed. Upper Saddle River, NJ: Pearson Prentice-Hall International, 2007. 773 p.

KASSAMBARA, A. **Practical guide to principal component methods in R**: PCA, (M)CA, FAMD, MFA, HCPC, factoextra. STHDA Editors, 2017. 264 p.

LOBATO, A. C. B.; GALARZA, B.; CAVALCANTI, F. R. Desenvolvimento de um método de caracterização de esquemas de controle químico em vinhedos por Análise Explanatória Multivariada. In: ENCONTRO DE INICIAÇÃO CIENTÍFICA, 15., 2017; ENCONTRO DE PÓS-GRADUANDOS DA EMBRAPA UVA E VINHO, 11., 2017, Bento Gonçalves. **Resumos...** Bento Gonçalves: Embrapa Uva e Vinho, 2017, p. 55. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/162551/1/Anais-15IC-2017-55.pdf. Acesso em: 22 dez. 2022.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Academic, 1979. 518 p.

OCHOA-MUNOZ, A. F.; GONZÁLES-ROJAS, V. M.; PARDO, C. E. Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm. **DYNA**, v. 86, n. 211, p. 249-257, 2019.

PROVOST, F.; FAWCETT, T. Agrupamento hierárquico. In: PROVOST, F.; FAWCETT, T. (ed.). **Data science para negócios**: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro, Alta Books, 2016. p. 165-181.

RUPPERT, D.; MATTESON, D. S. **Statistics and data analysis for financial engineering**: with R examples. 2nd ed. New York: Springer, 2015. 719 p. DOI: https://doi.org/10.1007/978-1-4939-2614-5.

ŠULIGOJ, F. **Spatial patient registration in robotic neurosurgery**. 2018. 86 p. Thesis (Doctoral) - University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture, Zagreb.

Embrapa

Uva e Vinho

MINISTÉRIO DA AGRICULTURA E PECUÁRIA

GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

CGPE 018276