UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Análise de agrupamentos para dados espectrais

**João Pedro Alvarenga Ramos da Silva**

Dissertação de Mestrado do Programa Interinstitucional de Pós-graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

**João Pedro Alvarenga Ramos da Silva**

# Análise de agrupamentos para dados espectrais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Mário de Castro Andrade Filho
Coorientadora: Profa. Dra. Camila Pedroso Estevam de Souza

**USP – São Carlos**
**Maio de 2024**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

P372a

Pedro Alvarenga Ramos da Silva, João
Análise de agrupamentos para dados espectrais / João Pedro Alvarenga Ramos da Silva; orientador Mário de Castro Andrade Filho; coorientadora Camila Pedroso Estevam de Souza. -- São Carlos, 2024.
106 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Estatística. 2. Análise de agrupamento. 3. Dados funcionais. 4. Espectroscopia. 5. Raman. I. de Castro Andrade Filho, Mário, orient. II. Pedroso Estevam de Souza, Camila, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**João Pedro Alvarenga Ramos da Silva**

# Clustering for spectrum data

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Mário de Castro Andrade Filho
Co-advisor: Profa. Dra. Camila Pedroso Estevam de Souza

**USP – São Carlos**
**May 2024**

# AGRADECIMENTOS

Em primeiro lugar, agradeço ao meu orientador Mário de Castro Andrade Filho e coorientadora Camila Pedroso Estevam de Souza pelas sugestões e correções que contribuíram para o meu aprendizado e para o desenvolvimento desta dissertação, o pesquisador Rafael da Silva de Souza que me ajudou na busca de dados para as aplicações práticas e material de estudo.

Por fim, agradeço à minha família e amigos, que sempre estiveram presentes, apoiando minhas escolhas e acreditando em meu potencial.

# RESUMO

SILVA, J. P. A. R. **Análise de agrupamentos para dados espectrais**. 2024. 106 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

A espectroscopia é o estudo que utiliza técnicas para medir o espectro da radiação eletromagnética, incluindo luz visível e radiofrequência, onde buscamos informações como composição química, temperatura, densidade, massa, distância, luminosidade e movimento relativo usando medidas de deslocamento. No caso da espectroscopia Raman, usamos um processo de difusão de luz. Nesta técnica, obtemos informações adicionais sobre as vibrações que permitem a compreensão da estrutura molecular fundamental.

Essas metodologias fornecem uma diversidade de dados, que serão modelados e analisados por meio de técnicas estatísticas e de aprendizado de máquina. Os dados de espectroscopia mostram alta dimensionalidade e forte presença de observações discrepantes (*outliers*) que causam dificuldades no agrupamento devido a falsos positivos na descoberta de novos agrupamentos.

Para o estudo, será feita uma revisão da literatura sobre métodos de agrupamento de dados de espectroscopia que não variam com o tempo. Assim, comparando modelos existentes, aplicando-os a dados reais e a dados simulados.

**Palavras-chave:** Espectroscopia, Dados Funcionais, Raman.

# ABSTRACT

SILVA, J. P. A. R. **Clustering for spectrum data**. 2024. 106 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Spectroscopy is the study that uses techniques to measure the spectrum of electromagnetic radiation, including visible light and radiofrequency, where we search for information such as chemical composition, temperature, density, mass, distance, luminosity, and relative motion using displacement measurements. In the case of Raman spectroscopy, we use a light diffusion process to obtain additional information about vibrations that increase the understanding of the fundamental molecular structure.

These methodologies provide a diversity of data, which will be modeled and analyzed using statistical and machine learning techniques. The spectroscopy data show high dimensionality and a strong presence of outliers that cause difficulties in clustering due to false positives in discovering new clusters.

For the study, a review of the literature on methods for grouping spectroscopy data that do not vary with time will be made. Thus, comparing models with the existing ones, and applying it to real data and simulated data.

**Keywords:** Raman, Spectroscopy, functional data, analysis.

# LISTA DE ILUSTRAÇÕES

# LISTA DE ALGORITMOS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| *outliers* | valores discrepantes |
| ACP | análise de componentes principais |
| ACPF | analise de componentes principais funcional |
| ADF | análise de dados funcionais |
| AEMET | agência estatal de Meteorologia |
| CPM | coeficiente de partição modificado |
| F1 | primeira função simulada |
| F2 | segunda função simulada |
| F3 | terceira função simulada |
| fdp | função de densidade probabilidade |
| fluxo | $f_\lambda = 10^{-17}$ erg/s/$cm^2$/Ang |
| FN | falso negativo |
| FP | falso positivo |
| funHDDC | algoritmo funcional baseado em HDDC |
| HDDC | método de agrupamento de dados de alta dimensão |
| HDDC | método geral de agrupamento para dados de alta dimensão |
| MCMC | Monte Carlo via Cadeias de Markov |
| MLF | modelo de mistura latente funcional |
| MMQ | método dos mínimos quadrados |
| NASA | Administração Nacional da Aeronáutica e Espaço dos Estados Unidos |
| qso | quasares |
| SDSS | Sloan Digital Sky Survey |
| SICV | soma da variação intragrupo |
| vet. a. | vetor aleatório |
| VN | verdadeiro negativo |
| VP | verdadeiro positivo |

# SUMÁRIO

CAPÍTULO

1

# INTRODUÇÃO

A espectroscopia é responsável por medir e interpretar os espectros que resultaram da interação entre radiação e a matéria. Segundo Dumas, Sockalingum e Sule-Suso (2007), a espectroscopia está relaciona com fenômenos físico-químicos, em especial com a radiação eletromagnética, onde átomos ou moléculas realizam a absorção, emissão ou dispersão dessa radiação.

Os fenômenos físico-químicos de acordo com Schulman (2017) abordados no estudo da espectroscopia ocorrem a nível molecular. Neste estudo é observada a ocorrência de alterações no nível energético dos átomos ou algum fenômeno de interação entre moléculas, podendo ser reflexão, difração dentre outras interações possíveis entre os átomos de uma matéria.

Desta forma, o principal interesse ao realizar uma análise espectroscópica é observar a quantidade de radiação que uma superfície de moléculas emite ou absorve quando sujeita a um fenômeno físico-químico, ou é exposta à radiação.

Dependendo do tipo de radiação, podendo ser infravermelho como em Wolkers *et al.* (2004) ou de raios X demonstrado em Vasquez (1991), podemos obter diferentes leituras espectrais, logo, é possível captar informações sobre a matéria ou evento observado.

Nesse sentido, a espectroscopia é amplamente aplicada, resolvendo uma variedade de problemas analíticos, quando desejamos inferir sobre a estrutura de algum elemento ou matéria. O termo espectroscopia é derivado da junção de duas palavras: *spectron* e *skopein*, que significam respectivamente fantasma e a vista no mundo.

Temos que o produto de uma leitura espectroscópica são as curvas observadas do espectro (na Astronomia em escala fluxo e na espectroscopia de Raman é regularmente usado nanômetros), nas quais temos interesse em analisá-las estatisticamente de forma funcional, preservando toda a variabilidade original dos dados amostrados.

Quando falamos de análise funcional podemos ter interesse em diversas características

que são encontras nas ondas ou curvas observadas. Temos em Wang *et al.* (2008) exemplos inferenciais que aplicam algoritmos de aprendizado de máquina e Monte Carlo via Cadeias de Markov (MCMC).

A análise de dados funcionais (ADF) é um ramo da Estatística que analisa dados que fornecem informações sobre curvas, que podem variar em uma determinada escala ou no tempo segundo Morettin (2014) e Ramsey e Silverman (2005). Em sua forma mais geral, sob uma estrutura ADF, cada elemento amostral de dados funcionais é considerado uma função aleatória.

Intrinsecamente, os dados funcionais são de dimensão infinita. A alta dimensionalidade intrínseca desses dados traz desafios para a teoria e para a computação.

Esses desafios variam com a forma como os dados funcionais foram amostrados. A amostragem pode ser encontrada nas áreas de análise espectroscópica (principalmente nos campos da Física, Biologia e Química), acompanhamento meteorológico (temperatura do ar, velocidade e direção do vento, umidade do ar, radiação solar, pressão atmosférica) e econômicas.

Além da alta dimensionalidade (que é uma rica fonte de informação) presente nos dados funcionais, é comumente encontrada a presença de valores discrepantes (*outliers*) (geralmente picos nas curvas) e a presença de ruídos que dificultam as análises. Desta forma, há muitos desafios interessantes para pesquisa em análise funcional.

Frequentemente, os pesquisadores que utilizam a ADF em seus estudos com dados espectrais estão interessados em realizar o agrupamento dos dados, como em Sikirzhytskaya, Sikirzhytski e Lednev (2017) ou Sasdelli *et al.* (2016), pelo desconhecimento dos verdadeiros rótulos dos dados ou curvas amostradas. Para estes trabalhos, a motivação se dá em descobrir se estes rótulos desconhecidos podem ser interpretados como um novo grupo até então desconhecido de corpos celestes, no caso do campo Astronômico, ou até mesmo distinguir conjuntos de substâncias quando nos referimos aos campos da Biologia ou Química.

A abordagem não supervisionada, quando os dados não possuem rótulos ou classificação, segundo Ghahramani (2003), consiste em aplicar métodos que calculam as distâncias entre as observações em um conjunto de dados (por exemplo a distância euclidiana) ou podem trabalhar através da captura de padrões e assim calcular a similaridade entre as observações, e a partir do cálculo destas métricas, as observações são agrupadas de acordo com a proximidade.

Neste trabalho será apresentada uma variedade de métodos de agrupamento funcional presentes na literatura que podem ser aplicados em dados espectrais. Em uma comparação de performance quando aplicados tanto para dados simulados quanto para dados reais de espectroscopia na Astronomia e de espectroscopia de Raman em leituras de material mineral, são propostas alternativas metodológicas que demonstram uma melhor performance frente aos métodos tradicionalmente usados na ADF.

## 1.1 Aplicabilidades da Espectroscopia

Serão explanadas algumas das aplicabilidades da espectroscopia nas áreas de Biologia, Química, Física e Astronomia. Conjuntamente com as aplicações, serão abordadas as metodologias encontradas nos trabalhos e algumas considerações sobre as abordagens propostas na revisão bibliográfica.

### 1.1.1 Espectroscopia de Raman

Vale a pena exemplificar uma das vertentes mais famosas para a obtenção de leituras espectrais, a chamada espectroscopia de Raman. Segundo Rostron, Gaber e Gaber (2016) é uma técnica de espectroscopia molecular que utiliza a interação da luz com a matéria para determinar a composição de um material.

As informações fornecidas por essa variante são resultados de um processo de difusão da luz, que se difere de técnicas espectroscópicas, dados pela absorção da luz. Outro ponto que é um diferencial, é que a partir da espectroscopia de Raman obtemos informações sobre as vibrações moleculares, logo, caso ocorra alguma reação molecular, o método é capaz de detectar sem a necessidade de procedimentos adicionais.

A leitura feita pela espectroscopia Raman resulta em um espectro que caracteriza a molécula observada, esse espectro é conhecido como identidade molecular. A identidade molecular é indispensável quando se deseja identificar uma matéria ou substância.

Em aplicações na área da Biologia, é de interesse frequente dos pesquisadores agrupar os dados a partir das leituras de ondas espectroscópicas com o intuito de saber quais grupos biológicos estão contidos em uma amostra (Virkler e Lednev (2009) e Farber e Kurouski (2018)). Um tipo de análise que pode ser mencionado para esse caso é a forense. A identificação de fluidos corporais é um aspecto muito importante em investigações tais como em Virkler e Lednev (2009), Muro e Lednev (2017) e Sikirzhytskaya, Sikirzhytski e Lednev (2017) .

A maioria dos casos envolvendo evidências biológicas começa com a localização e coleta de fluidos corporais humanos, como sangue, sêmen e saliva. O procedimento usual envolve a realização de um teste destrutivo de presunção (Muro e Lednev (2017) e Sikirzhytskaya, Sikirzhytski e Lednev (2017)), para o caso de amostras sanguíneas e só podem ser realizadas em laboratório, situações que poderiam ser evitadas com as técnicas de Raman.

Para análises sanguíneas da cena de um crime temos o problema de similaridade entre sangue de cães, gatos, roedores e até humanos (Virkler e Lednev (2009) e Sikirzhytskaya, Sikirzhytski e Lednev (2017)). Desse modo, é de suma importância determinar a origem exata do sangue. Nesse sentido, técnicas não supervisionadas são úteis para tal determinação.

Uma análise estatística em dados de espectros metodologicamente apropriada pode ter impacto positivo em diversos estudos. Foi observado na literatura que em várias análises é

aplicada a técnica de análise de componentes principais (ACP) somente ou em conjunto com outras técnicas (Virkler e Lednev (2009), Farber e Kurouski (2018)) com o objetivo de separar os agrupamentos desejados no estudo.

As ferramentas encontradas em aprendizado de máquina são pouco exploradas nessa literatura. É observado que a influência de ruídos, alta dimensionalidade e picos nas leituras prejudicam fortemente as análises e os algoritmos aplicados.

### 1.1.2 Espectroscopia na Astronomia

Uma aplicação usual da espectroscopia é no campo da Astronomia. As leituras das observações astronômicas são coletadas em formato de ondas (ou funções), geralmente interpretadas em dois eixos: comprimento de onda (eixo $x$) e escala $f_\lambda = 10^{-17}$ erg/s/$cm^2$/Ang (fluxo) (eixo $y$). A partir dessas leituras, segundo Bradt (2004), são extraídas informações das componentes químicas, cores, formações metálicas e muitas outras informações de estrelas e galáxias.

Um grupo com grande atuação e relevância nas leituras espectrais astronômicas é a Sloan Digital Sky Survey (SDSS)(Server (2022)). A SDSS utiliza um telescópio com um espelho primário de $2,5$ metros de diâmetro, totalmente dedicado ao projeto de observação espacial. O telescópio está situado no observatório Apache Point Observatory, no Novo México, EUA.

Os pesquisadores desta área estão frequentemente interessados em realizar o agrupamento não supervisionado dos dados, como em Sasdelli *et al.* (2016) e Logan e Fotopoulou (2020). No caso estudado em Sasdelli *et al.* (2016), notamos que muitas vezes as ferramentas de agrupamento são aplicadas sobre leituras com uma incerteza até mesmo da existência de diferentes agrupamentos nos dados estudados. Visto que muitos eventos astronômicos não foram ainda comprovados, dessa forma, não temos a certeza de que estamos observando de fato grupos diferentes ou variações na leitura de um único processo.

Frequentemente, quando são analisados os espectros, os algoritmos de aprendizado de máquina são usados. Nas análises encontradas em McGurk, Kimball e Ivezić (2010), Bailey (2012), Sasdelli *et al.* (2016) e Logan e Fotopoulou (2020), notamos que em todos os estudos são apresentadas técnicas de sumarização de dados, como ACP, redes de *encoder-decoder* e bem como tratativas clássicas de normalização de dados e exclusão de dados faltantes.

## 1.2 Pacotes Usados

Para as análises feitas, foi usada a linguagem de programação R (2023) e alguns pacotes que estão disponíveis na versão 4.2.0, dentre eles:

- Pacote fda (Ramsay, Graves e Hooker (2022)): O pacote *fda* fornece funções que dão suporte à análise de dados funcionais, principalmente quando precisamos realizar as trocas

de bases dos dados e realizar as suavizações das curvas. Esses assuntos serão abordados nos capítulos seguintes;

- Pacote fda.usc (Febrero-Bande e Oviedo de la Fuente (2012)): No pacote fda.usc encontramos formas de análises descritivas, modelos de regressão e classificação supervisionada e não supervisionada. Neste pacote encontramos os modelos de K-Médias para dados funcionais e a ACP funcional;
- Pacote funHDDC (Schmutz e Bouveyron (2021)): Em funHDDC temos alguns algoritmos que nos permitem agrupar dados. Nele encontramos um algoritmo de agrupamento não supervisionado baseado no método de agrupamento de dados de alta dimensão (HDDC);
- Pacote ggplot2: Suporte na análise gráfica;
- Pacote dplyr: Suporte na manipulação de dados.

## 1.3 Organização da Dissertação

Neste trabalho apresentamos no Capítulo 2 uma definição para os dados funcionais que serão abordados neste trabalho. No Capítulo 3 uma descrição metodológica referente aos modelos que serão usados nas aplicações. As notações apresentadas na descrição metodológica referente aos modelos, em alguns momentos, se diferem dos artigos referidos, as modificações foram feitas com o intuito de uniformizar as notações.

Os Capítulos 4 e 5 apresentam exemplos das aplicações tradicionalmente usadas e de algoritmos mais recentes na literatura que trazem como proposta agrupar dados simulados e reais de espectros astronômicos e minerais. Essas abordagens serão de um ponto de vista não supervisionado. No Capítulo 6, temos os principais resultados, conclusões e reflexões sobre todo o trabalho desenvolvido e por fim, no apêndice temos uma aplicação adicional a dados meteorológicos, onde foi abordada a modelagem funcional definida no tempo.

CAPÍTULO

# 2

# REVISÃO SOBRE DADOS FUNCIONAIS

Como em espectroscopia temos curvas observadas, no âmbito técnico, a análise de dados funcionais é comumente aplicada nessa área. Portanto, neste capítulo apresentamos uma revisão para dados funcionais.

Dados funcionais são definidos amplamente como dados em forma de curvas provenientes de funções, que na maioria dos casos, não conhecemos mas sabemos que elas existem e deram origem aos dados que seriam eventualmente analisados. Mas essa definição pode ser muito abrangente, temos que existem algumas características ao fazer uma analise funcional que podem fazer grandes diferenças na hora de escolhermos qual algoritmo será usado.

A seguir apresentamos uma definição mais formal para os dados funcionais de tal maneira que essas informações podem ser utilizadas para obtermos um ajuste mais fino e polido aos dados.

## 2.1 Preliminares

### *2.1.1 Transformada de Fourier*

A teoria da transformada de Fourier (Fourier (1878)) é relevante para as definições e abordagens dos dados funcionais, pois é de interesse definir um espaço básico para formalização das representações espectrais (Morettin (2014)), este será o $L^2$ ($\mathbb{R}$), de funções de quadrado integrável, definidas em $\mathbb{R}$.

Se $f$ for uma função absolutamente integrável (sem perda de generalidade), definimos a transformada de Fourier de $f$ como

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{i\xi x} dx, \quad \xi \in \mathbb{R}, \tag{2.1}$$

sendo aqui $\xi$ a frequência angular (número de ciclos completos em $2\pi$ unidades de tempo) e $i$ é uma unidade imaginaria (demonstrado em Fourier (1878)). O termo $e^{i\xi x}$ vem da fórmula de Euler com o objetivo de relacionar o sistema de exponenciais complexas com o sistema de senos e cossenos escrito como

$$e^{int} = cos(nt) + isen(nt). \tag{2.2}$$

Pode-se também definir uma transformada de Fourier alternativa (Rudin (1959)), nela tem-se uma $\hat{F}(\xi) = F(f(x))$, logo é escrita a transformada de Fourier alternativa como

$$\hat{F}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}dx. \tag{2.3}$$

Por fim, pode-se expressar a $f(x)$ como a transformada inversa de Fourier (Morettin (2014)) de $\hat{F}(\xi)$ (onde se recupera os dados em sua forma original), que é útil para quando deseja-se realizar a interpretação dos dados no formato original das ondas (ou funções) que foram amostradas, expressando a inversa na forma,

$$f(x) = F^{-1}(\hat{F}(\xi)) = \int_{-\infty}^{\infty} \hat{F}(\xi)e^{i\xi x}d\xi. \tag{2.4}$$

### 2.1.2 Estacionariedade

A condição de estacionariedade é importante para a análise e interpretação dos dados funcionais (neste caso espectros) e sua definição abordada. Tecnicamente, há duas formas de estacionariedade: fraca (de segunda ordem ou ampla) e forte (ou estrita).

#### 2.1.2.1 Estacionariedade Forte

Um processo estocástico $(X(t), t \in T)$ em que $t$ representa o momento em que ocorre a observação podendo ser no tempo ou uma marcação de determinada escala dos dados, diz-se estritamente estacionário se todas as distribuições finito dimensionais permanecem as mesmas sob translações, isto é, considere uma função $F(x_1;t_1)$ para um valor de $x_1$ no tempo $t_1$ e adicionamos uma medida $\tau$ em todos os tempos, assim temos a relação,

$$F(x_1,...,x_n;t_1+\tau,...,t_n+\tau) = F(x_1,...,x_n;t_1,...,t_n), \tag{2.5}$$

para quaisquer $t_1,...,t_n,\tau$ de $T$. Por distribuição finito dimensional, deve-se entender que para todo $n \geq 1$, temos conhecimento das distribuições, ou seja, para $n = 1$ conhecemos as distribuições unidimensionais da v.a. $X(t_1), t_1 \in T$, para $n = 2$ conhecemos as distribuições bidimensionais da v.a. $X(t_1), X(t_2), t_1, t_2 \in T$ e assim por diante.

De (2.5) temos que todas as distribuições unidimensionais são invariantes sob translações do tempo, logo a média $\mu(t)$ e a variância $\sigma^2(t)$ são constantes,

$$E[X(t)] = \mu(t) = \mu \tag{2.6}$$

$$\text{e } Var[X(t)] = \sigma^2(t) = \sigma^2, \tag{2.7}$$

para todo $t \in T$. Sem perda de generalidade, podemos supor que $\mu = 0$, caso contrário, considere o processo $X(t) - \mu$, assim podemos escrever a função de autocovariância de um processo estacionário forte como

$$\gamma(\tau) = Cov[X(t), X(t+\tau)] = Cov[X(0), X(\tau)]. \tag{2.8}$$

#### 2.1.2.2 Estacionariedade Fraca

Um processo estocástico $(X(t), t \in T)$, diz-se fracamente estacionário (ou estacionário de segunda ordem) se, e somente se,

- (i) $E[X(t)] = \mu(t) = \mu$, constante, para todo $t \in T$;
- (ii) $E[X^2(t)] < \infty$, para todo $t \in T$;
- (iii) $\gamma(t_1, t_2) = Cov(X(t_1), X(t_2))$ é uma função apenas de $|t_1 - t_2|$.

Esta relação é especialmente interessante pela sua aplicabilidade no estudo de dados funcionais, dessa forma, denominaremos essa classe de processos como simplesmente "processos estacionários".

### 2.1.3 Ruído Branco

Um ruído é uma interferência aleatória causada por situações adversas ao experimento ou amostram realizada. Um ruído é dito branco quando em uma sequência cada valor $\varepsilon$ da série tiver média zero e variância constante, além de não apresentar correlação serial, formalmente definimos que $E[\varepsilon_t] = E[\varepsilon_t] = \cdots = 0$, $Var[\varepsilon_t] = Var[\varepsilon_t] = \cdots = c$ e $E[\varepsilon_i; \varepsilon_j] = 0$ para qualquer $i \neq j$. Para este trabalho, será considerado que o ruído branco $\varepsilon$ segue uma distribuição Normal na forma $\varepsilon \sim N(0;1)$ para a realização das simulações.

### 2.1.4 Observações Discrepantes (Outliers)

Uma observação discrepante ou *outlier* é uma observação que se encontra a uma distância anormal de outros valores em uma amostra aleatória de uma população, um valor que foge da

normalidade. Muitas vezes, cabe ao analista decidir o que será considerado anormal. A causa de um *outlier* pode ser atribuída a qualquer fator interno ou externo ao processo que gere um comportamento anormal na observação. De maneira geral podemos abordar duas, um efeito aleatório (podendo ser um ruído ou erro de medição) ou uma leitura atípica, porém informativa sobre algum evento.

Na espectroscopia astronômica e na espectroscopia de Raman, algumas curvas que serão analisadas nas aplicações, apresentam *outlier* em suas leituras, contudo esses pontos *outlier* são de extrema importância para o experimento, de tal forma que não podemos simplesmente excluí-los do conjunto de dados.

### 2.1.5 Transformação de Variáveis

Serão propostas duas transformações que serão abordadas neste trabalho, a primeira é chamada de padronização ou padronização $Z$, nela temos interesse em deixar todos os atributos com média zero e desvio padrão igual a um.

Se temos um conjunto com $X_1, X_2, ..., X_n$ variáveis, $z_{ij}$ é a padronização $Z$ da $i$-ésima variável na $j$-ésima observação, podemos escrever

$$z_{ij} = \frac{x_{ij} - \overline{X_i}}{\sqrt{\text{Var}(X_i)}}. \tag{2.9}$$

Outra transformação que será usada é a normalização, onde temos o objetivo de deixar todas as $X_1, X_2, ..., X_n$ variáveis do conjunto em um intervalo em comum, neste caso será o intervalo $[0,1]$. Seja $p_{ij}$ normalização da $i$-ésima variável na $j$-ésima observação, escrevemos na forma

$$p_{ij} = \frac{x_{ij} - \text{Min}(X_i)}{\text{Max}(X_i) - \text{Min}(X_i)}. \tag{2.10}$$

## 2.2 Definição de Dados Funcionais

Funções aleatórias podem ser vistas como elementos aleatórios tomando valores em um espaço de Hilbert ou como um processo estocástico. O primeiro é matematicamente conveniente, enquanto o segundo é um pouco mais adequado de uma perspectiva aplicada. Essas duas abordagens coincidem se as funções aleatórias forem contínuas e a condição de média quadrática estiver satisfeita. Nesta pesquisa, será abordada a visão estocástica (Morettin (2014)).

Considerando a definição estocástica de funções aleatórias, seja $X(t), t \in T$ um processo estacionário, em que $T \subset \mathbb{R}$ e com média igual a zero. Considerando todas as trajetórias deste processo estacionário, verifica-se que ao tomar uma trajetória $X^j(t)$ qualquer, $X^j(t)$ não é de quadrado integrável e não é periódica. Sob esses pressupostos, nosso objetivo é aproximar

uma função aleatória $f(t)$ através de uma $f_T(t)$. Sem perda de generalidade, desenvolvemos as representações para o caso mais complexo e utilizado nas aplicações, a situação de tempo contínuo e frequência contínua.

Considerando uma realização particular do processo de $X(t)$, definimos a função

$$Y(t) = \begin{cases} X(t), \text{ se } & -T \leq t \leq T \\ 0, \text{ se } & |t| > T \end{cases} \tag{2.11}$$

assim, pode-se escrever para esta função uma transformada de Fourier alternativa definida como

$$F_Y(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(t)e^{-i\lambda t}dt = \frac{1}{2\pi}\int_{-T}^{T} X(t)e^{-i\lambda t}dt, \tag{2.12}$$

em que $\lambda$ é a frequência angular (que pode ser em unidades de tempo, por exemplo) e $e^{i\lambda x}$ é o coeficiente da transformada de Fourier (fórmula de Euler). Sendo assim, conseguimos escrever que a transformada inversa de Fourier

$$Y(t) = \int_{-\infty}^{\infty} F_Y(\lambda)^{i\lambda t}d\lambda. \tag{2.13}$$

Se a representação $|F_Y(\lambda)|^2 d\lambda$ expressa a contribuição das componentes de $Y(t)$, com frequências em $(\lambda, \lambda + d\lambda)$ temos

$$J^{(T)}(\lambda) = \frac{|F_Y(\lambda)|^2}{2T}, \tag{2.14}$$

que é denominada a função densidade de potência de $Y(t)$. Para caracterizar as propriedades espectrais de $X(t)$ (processo estacionário) será considerada a esperança de $J^{(T)}(\lambda)$ sobre todas as realizações sendo

$$f(\lambda) = \lim_{T\to\infty} E(J^{(T)}(\lambda)). \tag{2.15}$$

Se o limite (2.15) existir, $f(\lambda)$ é denominada de função densidade espectral de $X(t)$ (espectro de $X(t)$). Desta forma, é possível interpretar que dado $J^{(T)}(\lambda)$, $f(\lambda)$ representa a média das componentes de $X(t)$ com frequência em $(\lambda, \lambda + d\lambda)$ (potência total). Podemos escrever que

$$f(\lambda) = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-2T}^{2T} (1 - \frac{|\tau|}{2T})\gamma(\tau)e^{-i\lambda\tau}d\tau, \tag{2.16}$$

em que $\gamma(\tau)$ é a função de autocovariância de $X(t)$ no tempo $\tau$, dada por

$$\gamma(\tau) = COV[X(t), X(t+\tau)]. \tag{2.17}$$

Para o limite (2.16) existir, temos que ter

$$\int_{-\infty}^{\infty} |\gamma(\tau)|d\tau < \infty, \tag{2.18}$$

pode-se dizer que para o caso em que $\gamma(\tau) \to \infty$, os valores do processo estão suficientemente afastados, logo, caso exista correlação entre os valores do processo ela é considerada fraca. Considerando (2.18) temos que

$$f(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \gamma(\tau) e^{-i\lambda\tau} d\tau, \tag{2.19}$$

que é uma leitura do espectro como uma transformada de Fourier da função de autocovariância.

Assim, a função de autocovariância $\gamma$ em função do tempo $\tau$ é expressa na forma

$$\gamma(\tau) = \begin{cases} \int_{-\infty}^{\infty} f(\lambda) d\lambda = Var[X(t)], \text{ se } & \tau = 0 \\ \int_{-\infty}^{\infty} f(\lambda) e^{i\lambda\tau} d\lambda, \text{ se } & \tau > 0 \end{cases} \tag{2.20}$$

logo, seguindo a definição descrita em (2.20) é possível interpretar que a variância do processo é uma relação aditiva das componentes presentes em $X(t)$ que estão definidas no intervalo $(\lambda, \lambda + d\lambda)$, a contribuição das componentes é expressa pela função $f(\lambda)d\lambda$.

Visto que temos as definições para $t \in \mathbb{R}$, vamos apresentar para o caso decorrente de um processo estacionário discreto, ou seja, $t \in \mathbb{Z}$ para $X_t$. No caso discreto, podemos substituir as integrais por somatórios. Se a condição

$$\sum_{k=-\infty}^{\infty} |\gamma_k| < \infty \tag{2.21}$$

estiver satisfeita e assim podemos escrever o espectro de $X_t$ como

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\lambda k}, -\pi < \lambda < \pi. \tag{2.22}$$

## 2.3 Definição de Dados Funcionais nos Reais

Dados funcionais definidos nos reais são chamados por Morettin (2014) de ondaletas. A representação via ondaletas é considerada quando se trabalha funções definidas em espaços como os senos e cossenos, polinômios ortogonais e funções de Haar, dentre outros.

Uma consideração é que nesse momento estamos nos atentando a estender o espaço que a ondaleta está definida para um mais amplo e próximo das aplicações. Do ponto de vista aplicado abordamos o processo estacionário com o método mais usual e simples, que é a representação via Fourier.

Usando a representação de Fourier, toda função periódica, de período $2\pi$ e quadrado integrável, denotada como $L^2(0, 2\pi)$, é gerada por uma superposição de exponenciais, $w_n(x) = e^{inx}$, $n = 0, 1, ...$, obtidas por dilatações da função $w(x) = e^{ix}$ dadas por $w_n(x) = w(nx)$. Dada

essa definição, queremos estender a definição para $L^2(\mathbb{R})$, isto é, gerar esse espaço a partir de uma única função $\psi$. Um caso particular muito útil é dado por

$$\psi_{a,b}(x) = |a|^{-1/2}\psi(\frac{x-b}{a}), \tag{2.23}$$

para $b \in \mathbb{R}$ e $a > 0$. A função $\psi$ é chamada de ondaleta mãe, é usual para esta definição usarmos $a = 2^{-j}$ e $b = k2^{-j}$ para $j,k \in \mathbb{Z}$.

Formalmente, sem perda de generalidade, tem-se que a notação $L^2(\mathbb{R})$ representa o espaço de todas as funções mensuráveis de quadrado integrável sobre $\mathbb{R}$ nesta situação, as funções $f(t)$ devem decair para 0, quando $|t| \to \infty$. Logo, a ideia é considerar dilatações e translações de uma única função $\psi$ de modo a cobrir $\mathbb{R}$. Assim, conseguimos escrever usando $a = 2^{-j}$ e $b = k2^{-j}$ para $j,k \in \mathbb{Z}$

$$\psi_{j,k}(t) = |2|^{-j/2}\psi(2^j t - k). \tag{2.24}$$

A função $\psi_{j,k}(t)$ é obtida de $\psi(t)$ por uma dilatação binária $2^j$ e uma translação diádica $k2^{-j}$.

Uma consideração a se fazer é que as funções $\psi_{j,k}(t), j,k \in \mathbb{Z}$ formam uma base que não precisa ser necessariamente ortogonal. Quando trabalhamos com bases ortogonais, a vantagem é que podemos fazer a reconstrução perfeita do sinal original a partir dos coeficiente da transformada, essa relação é possível aplicando os conceitos da transformação inversa de Fourier.

Visto que temos algumas propriedades desejáveis quando estamos no caso ortogonal, considere $\psi_{j,k}(t), j,k \in \mathbb{Z}$ uma base ortogonal gerada por $\psi$. Para qualquer $f(t)$, com quadrado integrável sobre $\mathbb{R}$, representamos a função na forma

$$f(t) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(t). \tag{2.25}$$

Assim, dizemos que (2.25) é uma série de ondaletas de $f(t)$ (com convergência em média quadrática) e os coeficientes de ondaletas $c_{j,k}$ são dados por

$$c_{j,k} = \int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dt. \tag{2.26}$$

As equações (2.25) e (2.26), são definidas pela transformação inversa de Fourier, alguns dos métodos de agrupamento não supervisionados estudados neste trabalho realizam a extração dos coeficientes das curvas ($c_{j,k}$), para realizar agrupamento, desta forma, é consideradas que as curvas pertencem a um mesmo grupo devido a homogeneidade entre seus coeficientes.

A partir das definições feitas, tanto para a função geral (2.23) de ondaleta mãe, quanto para a aplicação proposta em (2.24) detêm propriedades interessantes, algumas delas são:

- (i) $\int_{-\infty}^{\infty}\psi(t)dt = 0$, chamada admissibilidade;

- (ii) $\int_{-\infty}^{\infty} |\psi(t)|dt < \infty$;
- (iii) $c_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty$, em que $\Psi(\omega)$ é a transformada de Fourier de $\psi(t)$;
- (iv) $\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1$ ou $\int_{-\infty}^{\infty} |\Psi(\omega)|^2 d\omega = 2\pi$;
- (v) Os primeiros $r-1$ momentos de $\psi$ anulam-se, isto é, $\int_{-\infty}^{\infty} t^j \psi(t)dt = 0$ para $j = 0, 1, ..., r-1$, e para algum $r \geq 1$ temos $\int_{-\infty}^{\infty} |t^r \psi(t)|dt < \infty$.

Pode-se entender que quanto maior o valor de $r$ maior será a suavidade da ondaleta mãe, pois o parâmetro $r$ pode ser compreendido como a regularidade de $\psi$. Outra propriedade desejada é em relação ao suporte da ondaleta, algumas ondaletas têm suporte compacto. Sendo que uma definição formal de conjunto compacto pode ser escrita como: Um subconjunto de um espaço topológico é dito compacto quando toda cobertura aberta deste subconjunto admitir uma subcobertura finita.

Em adicional ao conceito de ondaleta mãe, temos diversas maneiras de gerar uma ondaleta. Será apresentada a função escala (também conhecida como ondaleta pai) denotada por $\phi$,

$$\phi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} \varphi_k \psi(2t-k). \tag{2.27}$$

Esta função será uma função ortogonal em $L^2(\mathbb{R})$ (caso de interesse), com (2.27) podemos gerar ondaletas, visto que na prática desconhecemos a forma exata da ondaleta mãe. De maneira análoga feita em (2.24), partindo do caso geral (2.27), será construída uma forma de ondaleta pai em função de $j$ e $k$, assim, aplicando a definição de $\psi$ desenvolvida anteriormente, $\phi_{j,k}$ será denotado como

$$\phi_{j,k}(t) = 2^{j/2} \phi(2^j t - k), \tag{2.28}$$

em que $j, k \in \mathbb{Z}$ sem perda de generalidade, podemos obter então $\psi$ por meio de $\phi$ escrevendo

$$\phi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} h_k \phi(2t-k) \tag{2.29}$$

$$\text{e } h_k = (-1)^k \varphi_{1-k}. \tag{2.30}$$

De maneira prática, os termos $\varphi_k$ e $h_k$ são coeficientes de filtros que usamos para calcular a transformada da ondaleta (nesta abordagem discreta), $\varphi_k$ representa o filtro "passa-baixo"e $h_k$ o filtro "passa-alto", escritos como

$$\varphi_k = \sqrt{2} \int_{-\infty}^{\infty} \phi(t)\phi(2t-k)dt \tag{2.31}$$

$$\text{e } h_k = \sqrt{2}\int_{-\infty}^{\infty} \psi(t)\phi(2t-k)dt. \tag{2.32}$$

Assim, a $f(t) \in L^2(\mathbb{R})$ é escrita como

$$f(t) = \sum_{k=-\infty}^{\infty} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j\geq j_0}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t), \tag{2.33}$$

$$c_{j_0,k} = \int_{-\infty}^{\infty} f(t)\phi_{j_0,k}(t)dt \tag{2.34}$$

$$\text{e } d_{j,k} = \int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dt. \tag{2.35}$$

### 2.3.1 Exemplo

Será apresentado um exemplo clássico com a função de Haar, que é dada na forma

$$\psi^{(H)}(t) = \begin{cases} +1, \text{ se} & 0 \leq t < 1/2, \\ -1, \text{ se} & 1/2 \leq t < 1, \\ 0, \text{ se} & \text{caso contrário.} \end{cases} \tag{2.36}$$

Assim escrevemos a ondaleta pai como $\phi^H(t) = 1$ com $0 \leq t \leq 1$ (suporte compacto) e

$$\psi_{j,k}^{(H)}(t) = \begin{cases} 2^{j/2}, \text{ se} & 2^j k \leq t < 2^{-j}(k+1/2), \\ -2^{j/2}, \text{ se} & 2^{-j}(k+1/2) \leq t < 2^{-j}(k+1), \\ 0, \text{ se} & \text{caso contrário.} \end{cases} \tag{2.37}$$

Reescrevendo (2.27) para o exemplo, temos

$$\phi(t) = \phi(2t) + \phi(2t-1) = \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t) + \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t-1). \tag{2.38}$$

Logo $\varphi_0 = \varphi_1 = \frac{1}{\sqrt{2}}$ e $h_0 = -h_1 = \frac{1}{\sqrt{2}}$.

## 2.4 Considerações

Nosso foco principal é realizar o agrupamento não supervisionado. Contudo, é importante termos a noção das possíveis representações para os dados analisados. Essa visão possibilitou um entendimento metodológico dos algoritmos que serão aplicados.

As definições apresentadas são para processos estacionários. Para o caso não estacionário, as ondaletas sofrem mudanças em média, variância e possuem dependência diretamente do tempo

$(t)$, escrevemos $f(t,\lambda)$. Para construir as funções podemos usar duas abordagens. A primeira seria a partir da função de covariância, escrita na forma $\gamma(t_1 - t_2)$, a segunda forma seria construir um espectro dependente do tempo diretamente a partir do processo observado.

CAPÍTULO

3

# METODOLOGIAS DE AGRUPAMENTO ESTUDADAS

Temos uma vasta gama de métodos que podem ser utilizados quando estamos trabalhando com dados funcionais, seja no ambiente de regressão, classificação ou agrupamento. Cada método tem características específicas e lidam melhor com os dados quando aplicados de maneira correta ao propósito que foram pensados e desenvolvidos.

Sendo assim, apresentamos a seguir metodologias que aplicamos neste trabalho. É de interesse também ponderar sobre uma maneira de avaliar se os ajustes propostos fazem sentido e demonstram uma boa performance diante do desafio de agrupar os dados funcionais não supervisionadamente. Por exemplo, o estudo realizado por Teuling, Pauws e Heuvel (2020) avaliou o desempenho de métodos de agrupamento para dados funcionais, os quais foram avaliados utilizando métricas de concordância calculadas pela distância entre pontos, erro de estimativa e tendência.

Neste trabalho, apesar de não considerarmos nas análises, temos acesso aos rótulos verdadeiros das curvas, e esses serão usados para avaliar a performance do ajuste.

## 3.1 Método Baseado em Agrupamento de Dados de Alta Dimensão

O método geral de agrupamento para dados de alta dimensão (HDDC) foi originalmente proposto para o caso multivariado Bouveyron, Girard e Schmid (2007a). Essa metodologia foi tomada como base para criar o algoritmo funcional baseado em HDDC (funHDDC). Segundo Bouveyron e Jacques (2011), essa estrutura consiste em um modelo de mistura latente funcional que irá inserir os dados funcionais em subespaços. Uma proposta do método funHDDC é trazer boas estimativas unificando os passos de estimação com o de discretização das variáveis, e até

mesmo superando os métodos clássicos de duas etapas (métodos que primeiro sintetizam os dados e depois agrupam).

O método funHDDC, de acordo com as definições encontradas em Jacques e Preda (2014) sobre os diferentes tipos de abordagens funcionais para agrupamento, é um método de filtragem, esse método consiste em primeiro fazer uma releitura das curvas funcionais originalmente amostras, reescrevendo-as em função de uma base de coeficientes, onde cada curva é associada e um vetor de coeficientes (denotado por $\gamma$), e na sequência, será realizado o agrupamento não supervisionado destes vetores de coeficientes.

Para a apresentação de funHDDC, será tomado como base Bouveyron e Jacques (2011) e supor, sem perda de generalidade, que temos $x_1,...,x_n$ que são realizações do processo contínuo $X = [X(t)]$ para $t \in [0,T]$ em $L^2$, ou seja, as observações estão em $L^2[0,T]$. Como na prática fazemos uma amostragem e só temos as realizações da curva funcional em $x_1,...,x_n$, o primeiro passo é reconstruir a função assumindo que esses dados pertencem a um espaço de dimensão finita em uma base (base de funções). Desta forma, escrevemos

$$X(t) = \sum_{j=1}^{p} \gamma_j(X)\psi_j(t), \tag{3.1}$$

em que $\psi$ é a função base especificada, $\gamma = (\gamma_1(X),...,\gamma_p(X))$ é um vetor aleatório em $\mathbb{R}^p$ (esse conjunto definido em $\mathbb{R}^p$ será denotado por $\Gamma$) de realizações em $X$ que é um processo contínuo em $L^2$ (como descrito anteriormente), e $p$ é conhecido e fixado. Para as bases de funções, para o caso mais abrangente onde temos um ruído, a curva em (3.1) pode ser estimada pelo método dos mínimos quadrados (MMQ), apresentada em Moritz (1978).

Desta forma, um conjunto de $n_k$ de curvas, onde cada curva é escrita por um vetor de coeficientes $\gamma_i$, com $i = 1,...,n_k$. Tem-se assim, que os vetores são independentes e pertencem a um único grupo (o $k$-ésimo grupo). Suponha também que o $k$-ésimo agrupamento é funcional pertencente ao intervalo $\mathbb{E}_k \in [0,T]$ escrito como um processo real, $\mathbb{E}_k$ está em $L^2 \in [0,T]$ com dimensões $d_k \leq p$.

Considere que $\mathbb{E}_k$ pode ser extrapolado por $d_k$ por uma base de funções específica do grupo $\varphi_{kj}$ com $j = 1,...,d_k$ e que $\varphi_{kj}$ é obtido de $\psi_j$ por uma transformação linear $\varphi = \sum_{l=1}^{p} q_{k,jl}\psi_l$, que é uma matriz ortogonal $p \times p$ dada por $Q_k = (q_{k,jl}) = [U_k, V_k]$.

Agora, a partir das definições apresentadas, vamos escrever o modelo latente. Se $\lambda_1,...,\lambda_{nk}$ são coeficientes de expansão latentes de $\varphi_{kj}$ em $d_k$, esses coeficientes são provindos de um vetor aleatório (vet. a.) $\Lambda \in \mathbb{R}^{d_k}$, e vemos que as bases $\varphi_{kj}$ e $\psi_j$ têm uma relação, logo é escrita a relação entre $\Lambda$ e $\Gamma$, para o $k$-ésimo grupo como

$$\Gamma = U_k\Lambda + \varepsilon, \tag{3.2}$$

e $\varepsilon$ é um ruído, neste caso segue uma distribuição normal multivariada na forma $\varepsilon \sim N(0,\Xi_k)$.

Para a distribuição de $\Lambda$ temos

$$\Lambda \sim N(m_k, S_k), \tag{3.3}$$

ou seja, $\Lambda$ segue uma distribuição normal multivariada com $S_k = diag(a_{k1}, ..., a_{kd_k})$ sendo a matriz de covariância e $m_k$ as médias. Para $\Gamma$ no $k$-ésimo agrupamento tem-se,

$$\Gamma \sim N(\mu_k, \Sigma_k), \tag{3.4}$$

em que $\mu = U_k m_k$ e $\Sigma_k = U_k S_k U_k^t + \Xi_k$ (sendo $U_k^t$ a matriz transposta de $U_k$ ). A matriz de covariâncias do ruído $\Xi_k$ é escrita na forma $\Delta_k = COV(Q_k^t \Gamma) = Q_k^t \Sigma_k Q_k$, pode-se observar o formato da matriz diagonal em (3.5), no qual o ruído é modelado por $b_k$, além de ser a dimensão do subespaço do $k$-ésimo grupo, que é modelado por $a_k$.

$$\Delta_k = diag(a_{k1}, ..., a_{kd_k}, b_{k(d_k+1)}, ..., b_{k(p)}) \tag{3.5}$$

Sendo assim, fica definido o modelo de mistura latente funcional (MLF).

### 3.1.1 MLF e seus submodelos

Considerando um caso genérico no qual $x_1, ..., x_n$ são curvas $x_i = [x_i(t)]_{t \in [0,T]}$, em que $i = 1, ..., n$. Deseja-se formar $k$ agrupamentos homogêneos. Assumindo que existe uma v.a. $Z = (Z_1, ..., Z_k) \in [0,1]^k$ que é não observada. A variável $Z_k$ é igual a 1 se $X$ pertence ao $k$-ésimo grupo e 0 caso contrario.

Anteriormente $x_i$ era uma amostra de $X$, se $\gamma_i$ é um vetor de coeficientes, como definido anteriormente, temos que agora a distribuição é uma mistura de normais na forma

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma; \mu_k, \Sigma_k), \tag{3.6}$$

em que $\phi$ é a função de densidade probabilidade (fdp) de uma normal, escrita como $\mu_k = U_k m_k$ e $\Sigma_k = Q_k \Delta_k Q_k^t$, temos também uma priori para o $k$-ésimo grupo escrita como $\pi_k = P(Z_k = 1)$. Esse é o modelo $MLF_{[a_{kj}, b_k, Q_k, d_k]}$.

Do ponto de vista interpretativo e aplicado, restringimos os parâmetros em cada grupo e entre grupos do modelo MLF criando submodelos. Por exemplo, se fixarmos o parâmetro $d_k$ no modelo $\text{MLF}_{[a_{kj}, b_k, Q_k, d_k]}$, que é a diagonal de $\Delta_k$ (considerando que é um elemento comum dentro de todos os $k$ grupos), sugerimos que a suposição de que cada matriz $\Delta_k$ contém apenas dois autovalores diferentes definidos anteriormente como $a_k$ e $b_k$, e os outros são hiperparâmetros que serão otimizados. Esse modelo em especifico é escrito como $\text{MLF}_{[a_k, b_k, Q_k, d_k]}$, temos que a Tabela 1 que mostra quais variantes consideramos nesse trabalho para os submodelos MLF,

em Bouveyron, Girard e Schmid (2007b) contêm todos os 28 tipos que podemos montar com $\text{MLF}_{[a_{kj},b_k,Q_k,d_k]}$.

Tabela 1 – Submodelos MLF

| Submodelos |
|---|
| $MLF_{[a_{kj},b_k,Q_k,d_k]}$ |
| $MLF_{[a_{kj},b,Q_k,d_k]}$ |
| $MLF_{[a_k,b_k,Q_k,d_k]}$ |
| $MLF_{[a,b_k,Q_k,d_k]}$ |
| $MLF_{[a_{kj},b,Q_k,d_k]}$ |
| $MLF_{[a,b,Q_k,d_k]}$ |

Fonte – Elaborada pelo autor.

### 3.1.2 Algoritmo FunHDDC via algoritmo EM

Com os coeficientes $\gamma_1,...,\gamma_n$ referentes as curvas $x_1,...,x_n$ observadas, a função log-verossimilhança completa $L(\theta|\gamma;z)$ é escrita na forma

$$l_c(\theta;\gamma_1,...,\gamma_n,z_1,...,z_n) = -\frac{1}{2}\sum_{k=1}^{K}\eta_k\left[\sum_{j=1}^{d_k}\left(\log(\alpha_{kj})+\frac{q_{kl}^t C_k q_{kl}}{\alpha_{kj}}\right)+\sum_{j=1}^{d_k}\left(\log(b_k)+\frac{q_{kl}^t C_k q_{kl}}{b_k}-2\log(\pi_k)\right)\right]+\xi \tag{3.7}$$

em que $\theta$ é o vetor de parâmetros, que depende de qual submodelo será usado no MLF. Será abordado o caso mais geral em que $\theta = (\pi_k, \mu_k, a_{kj}, b_k, q_{kj})$ em que $1 \leq j \leq d_k$ e $1 \leq k \leq K$. Além disso, em 3.7:

- $q_{kj}$ é a $j$-ésima coluna de $Q_k$;
- $C_k = \frac{1}{\eta_k}\sum_{i=1}^{n} z_{ik}(\gamma_i - \mu_k)^t(\gamma_i - \mu_k)$;
- $\eta_k = \sum_{i=1}^{n} z_{ik}$;
- $\xi$ não depende de $\theta$.

Como não se sabe quais são os membros do agrupamento $z_{ik}$ será aplicado o algoritmo EM, proposto por Dempster, Laird e Rubin (1977). Sendo assim, deseja-se estimá-los aplicando o passo E, e depois no passo M será maximizado o valor esperado da função de log-verossimilhança completa (3.7) dado os dados observados e parâmetros da iteração vigente.

No passo E, é calculado $t_{ik}^{(q)} = E[Z_{ik}|\gamma_i, \theta^{(q-1)}]$, em que $\theta^{(q-1)}$ contém os parâmetros obtidos no passo $q-1$ (passo anterior). Para o modelo $\text{MLF}_{a_k,b_k,Q_k,d_k}$ temos que

$$t_{ik}^{(q)} = \frac{1}{\sum_{l=1}^{K} e^{[H_k^{(q-1)}(\gamma_i) - H_l^{(q-1)}(\gamma_i)]}}, \text{ sendo que } H_k^{(q-1)}(\gamma) \text{ é definido em } \gamma \in \mathbb{R}^p, \tag{3.8}$$

$$H_k^{q-1}(\gamma) = \left\| \mu_k^{q-1} - P_k(\gamma) \right\|_{D_k}^2 + \frac{1}{b_k^{(q-1)}} \|\gamma - P_k(\gamma)\|^2 + \sum_{j=1}^{d_k} \log(\alpha_{kj}^{(q-1)}) + (\text{p-d}_k) \log(b_k^{(q-1)}) - 2\log(\pi_k^{(q-1)}), \tag{3.9}$$

em que $||.||_{Dk}^2$ representa uma norma no espaço latente de $\mathbb{E}_k$, que é definido por:

- $||y||_{Dk}^2 = y^t D_{k,y}$;
- $D_k = \hat{Q}\Delta_k^{-1}\hat{Q}^t$;
- $\hat{Q}$ é de dimensão $p \times p$, contendo o vetor $d_k$ de $U_k$, logo tem-se que $\hat{Q}$ é definido em $[U_k, 0_{p-d}]$;
- $P_k(\gamma) = U_k U_k^t(\gamma - \mu_k) + \mu_k$, em suma $P_k$ define o operador de projeção latente no espaço $\mathbb{E}_k$.

De maneira aplicada, estimar $z_{ik}$ através de $t_{ik}$, favorece o caso no qual a projeção de uma nova observação no subespaço está próxima da média de um determinado grupo. Assim, é atribuído preferencialmente uma observação para a classe que está mais próxima, esses dois fatores são ponderados pelos parâmetros $a_k$ e $b_k$.

Por fim, o passo M consiste em maximizar (3.7) substituindo $z_{ik}$ por $t_{ik}^{(q)}$, logo, para os parâmetros do modelo, obtém-se:

$$n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}, \tag{3.10}$$

$$\pi_k^{(q)} = \frac{n_k^{(q)}}{n} = \frac{\sum_{i=1}^{n} t_{ik}^{(q)}}{\sum_{i=1}^{n} t_{ik}}, \tag{3.11}$$

$$\mu_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} \gamma, \tag{3.12}$$

$$C_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} (\gamma_i - \mu_k^{(q)})^t (\gamma_i - \mu_k^{(q)}), \text{ é a matriz de covariâncias da amostra do grupo } k \tag{3.13}$$

e $W = (w_{jk})_{1\leq j,k\leq p} = \int_0^T \psi_j(t)\psi_k(t)$, é a matriz de produtos internos entre as funções de base. (3.14)

Dadas as equações, temos a descrição do Algoritmo 1.

**Algoritmo 1** – Algoritmo do modelo $MLF_{a_k,b_k,Q_k,d_k}$

1: As $d_k$ primeiras colunas de $Q_k$ são atualizadas por $W^{1/2}C_k^{(q)}W^{1/2}$ segundo seus autovalores associados;
2: Para $a_{kj}$ com $j = 1,...,d_k$ são atribuídos os $d_k$ maiores valores dos autovalores de $W^{1/2}C_k^{(q)}W^{1/2}$;
3: Atualizamos $b_k^{(q)}$ pela equação traço$(W^{1/2}C_k^{(q)}W^{1/2}) - \sum_{j=1}^{d_k}(\hat{a_{kj}})^{(q)}$, fazemos esse processo até a convergência (BIERNACKI, 2004).

Para resumir, o algoritmo funHDDC modela e agrupa os dados funcionais por meio de suas projeções em subespaços considerando as principais características de cada curva com o objetivo de criar grupos homogêneos.

#### 3.1.2.1 *Hiperparâmetros*

O algoritmo EM estima a maior parte dos parâmetros, exceto $d_k$ e $K$. Esses dois parâmetros controlam a complexidade do modelo e não podem ser obtidos no passo M do algoritmo. Sendo assim, são estimados baseados nos autovalores da matriz de covariâncias condicional $\Sigma_k$ para o $k$-ésimo agrupamento, método proposto por Bouveyron, Girard e Schmid (2007b). A dimensão selecionada é aquela para a qual as diferenças de autovalores subsequentes são menores do que um limite definido pelo analista.

## 3.2 ACP Funcional

A metodologia de ACP, é aplicada em análises de grandes volumes de dados, sua descrição original pode ser encontrada em Hall, Müller e Wang (2006) e Yao, Müller e Wang (2005).

Em estudos mais atuais, a técnica excedeu as expectativas de ser usada somente como uma tratativa de pré-processamento de dados e passou a ser adotada em alguns algoritmos como parte da estimação de seus parâmetros, como em Krämer, Boulesteix e Tutz (2008), podendo fazer parte de um processo iterativo e até com o intuído inferencial em algumas situações.

ACP é matematicamente definida como uma transformação linear ortogonal, produzindo uma relação entre dois espaços funcionais que vamos denotar $u$ e $v$, assim, são transformados dados para um novo sistema de coordenadas. Por se tratar de uma transformação ortogonal, temos que $T : V \to V$ em um espaço de produto interno real $V$, preserva o produto interno na forma que $< u,v >=< t_u,t_v >$, os comprimentos e os ângulos entre eles.

Por fim, são geradas as componentes, em que a primeira componente detém a maior explicação da variabilidade do conjunto de dados original, a segunda contém a segunda maior explicação da variabilidade do conjunto de dados original e assim por diante. Tem-se que o número de componentes principais é sempre menor ou igual do que o número de variáveis originais.

Para definir a forma funcional desta técnica, sejam $X_1, ..., X_n$ pertencentes a um intervalo compacto dado por $J$, sendo funções aleatórias, independentes e identicamente distribuídas. Sendo assim, escrevemos

- $\int_J E(X^2) < \infty$;
- A média é dada por $\mu = E(X)$;
- A matriz de covariâncias é dada por $\psi(u,v) = COV[X(u), X(v)]$;
- $\psi(u,v) = \sum_{j=1}^{\infty} \theta_j \psi_j(v) \psi_j(u)$ é uma das possíveis decomposições espectrais para $\psi(u,v)$,

em que $\theta_1 \geq \theta_2 \geq ... \geq 0$ representa os autovalores de maneira ordenada de $\psi$ para as autofunções $\psi_j$'s, que formam uma sequência ortonormal em $L_2(J)$.

Considerando $X_i - \mu$ como uma generalização de Fourier, escrevemos $X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_{ij} \psi_j(t)$, sendo $\xi_{ij} = \int_J \psi_j(X_i - \mu)$ referente a pontuação ou efeito aleatório da $j$-ésima componente da analise de componentes principais funcional (ACPF).

Podemos observar que se $\psi_j$ e $\psi_k$ são ortogonais para $j \neq k$, com essa relação escrevemos que para $1 \leq j \leq \infty$ temos que $\xi_{ij}$ são não correlacionados. Esse fato é de extrema importância, principalmente no momento de escolher quais componentes serão utilizadas nos estudos e modelos.

Desta forma, nosso objetivo é realizar a estimação de $\theta_j$ e $\psi_j$. Sendo assim, considerando as definições feitas anteriormente, supondo que para cada $i$ temos os pares $(T_{ij}, Y_{ij})$ observados em que $1 \leq j \leq m_i$. Logo, escrevemos

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \tag{3.15}$$

em que $T_{ij}$ são os pontos de observação, sendo $T_{ij} \in J$ e $\varepsilon_{ij}$ os erros com média 0. De maneira geral, $T_{ij}$ não é necessariamente ser uma v.a. identicamente distribuída. Todavia, sem perda de generalidade, será considerado que para $T_{ij}$ e $\varepsilon_{ij}$, com variância finita, $E(\varepsilon^2) = \sigma^2$. Por fim, temos também que $X_i$, $T_{ij}$ e $\varepsilon_{ij}$ são independentes.

Considerando um conjunto de dados genérico $D = [(T_{ij}, Y_{ij}); 1 \leq j \leq m_i; 1 \leq i \leq n]$ serão estimados $\hat{\theta}_j$ e $\hat{\psi}_j$.

O primeiro passo é calcular $\hat{\mu}$ por $E(X)$. Assim, a matriz de covariâncias $\hat{\psi}(u,v)$, onde $(u,v)$ são dois espaços funcionais quaisquer e $\psi_j$'s são as autofunções destes espaços, é escrito na forma

$$\hat{\psi}(u,v) = \sum_{j=1}^{\infty} \hat{\theta}_j \hat{\psi}_j(v)\hat{\psi}_j(u). \tag{3.16}$$

Considerando o ajuste por mínimos quadrados, temos que $v$ e $u \in J$. Logo, se as larguras das bandas utilizadas são descritas por $h_\mu$ e $h_\phi$, escolhemos $(\hat{a},\hat{b}) = (a,b)$ de maneira a minimizar a expressão

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}[Y_{ij} - a - b(u - T_{ij})]^2 K(\frac{T_{ij}-u}{h_\mu}). \tag{3.17}$$

Assumimos que $\hat{\mu}(u) = \hat{a}$, dessa forma, escolhemos $(\hat{a_0},\hat{b_1},\hat{b_2}) = (a_0,b_1,b_2)$ que minimizem a (3.17) e reescrevemos como

$$\sum_{i=1}^{n}\sum_{j,k:1\leq j\neq k\leq m_i}\left[Y_{ij}Y_{ik} - a_0 - b_1(\mu - T_{ij}) - b_2(v - T_{ik})\right]^2 K\left(\frac{T_{ij}-\mu}{h_\phi}\right)K\left(\frac{T_{ik}-v}{h_\phi}\right). \tag{3.18}$$

Assim, $\phi(u,v) = E[X(u)X(v)]$, sendo $\hat{\phi}(u,v)$ escrita em relação a matriz de covariâncias da maneira $\hat{\psi}(u,v) = \hat{\phi}(u,v) - \hat{\phi}(u)\hat{\mu}(v)$.

Por fim, devemos considerar que $\theta$ são os autovalores de $\psi$. Logo, pela relação $\psi(u,v) = \sum_{j=1}^{\infty}\theta_j\psi_j(v)\psi_j(u)$ obtemos $\theta$ por $\psi$. Todavia, $\psi$ não é positiva definida, mas é simétrica, o que valida a relação. Definimos

- $U_{ij} = u - T_{ij}$;
- $V_{ik} = v - T_{ik}$;
- $Z_{ijk} = Y_{ij}Y_{ik}$;
- $W_{ij} = K(\frac{T_ij-u}{h_\mu})$;
- $W_{ijk} = K(\frac{T_ij-u}{h_\phi})K(\frac{T_ik-v}{h_\phi})$.

Estimamos $\hat{\mu}(u)$ e $\hat{\phi}(u,v)$ seguindo as expressões

$$\hat{\mu}(u) = \frac{S_2R_0 - S_1R_1}{S_0S_2 - S_1^2}, \tag{3.19}$$

$$\hat{\psi}(u,v) = (A_1R_00 - A_2R_10 - A_3R_01)B^{-1}, \tag{3.20}$$

$$
\begin{aligned}
&A_1 = S_{20}S_{02} - S_{11}^2,\ A_2 = S_{10}S_{02} - S_{01}S_{11},\ A_3 = S_{01}S_{20} - S_{10}S_{11},\\
&S_r = \sum_{i=1}^{n}\sum_{j=1}^{m_i} U_{ij}^r W_{ij},\\
&R_r = \sum_{i=1}^{n}\sum_{j=1}^{m_i} U_{ij}^r Y_{ij} W_{ij},\\
&S_{rs} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} U_{ij}^r V_{ik}^s W_{ijk},\\
&S_{rs} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} U_{ij}^r V_{ik}^s Z_{ijk} W_{ijk},\\
&B = A_1 S_{00} - A_2 S_{10} - A_3 S_{01} =\\
&[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} (U_{ij} - \overline{U})^2 W_{ijk}]\\
&[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} (V_{ij} - \overline{V})^2 W_{ijk}] -\\
&[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} (V_{ij} - \overline{V})(U_{ij} - \overline{U}) W_{ijk}]^2 \geq 0,\\
&\overline{Q} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} Q_{ij} W_{ijk}}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k:j<k}^{m_i} W_{ijk}}.
\end{aligned}
\tag{3.21}
$$

## 3.3 K-Médias Funcional

O agrupamento via algoritmo de K-Médias (García, García-Ródenas e Gómez (2015)) é um método amplamente usado quando estamos fazendo um agrupamento não supervisionado. Originalmente foi pensado no processamento de sinais. Esse algoritmo que visa dividir $n$ observações em $k$ grupos, em que cada observação pertence a um grupo com a média mais próxima a um centroide, servindo como um ponto focal do grupo que está se formando.

Isso resulta em uma divisão do espaço de dados em células Voronoi, que são as resultantes da divisão de um plano, regiões próximas de determinado grupo de objetos no plano de duas dimensões (caso mais simplificado) cada coordenada nos eixos $x$ e $y$ é pertencente a um subgrupo, de tal forma que se juntamos todos os subgrupos formamos o plano original.

O K-Médias, é um método de distâncias (Jacques e Preda (2014)), esse método usa métricas para calcular distâncias entre as observações funcionais, e assim realiza o agrupamento das observações, focando em minimizar a distância entre membros do mesmo grupo e maximizar a distância entre os grupos.

O K-Médias busca minimizar as variâncias dentro de cada agrupamento através da redução das distâncias. A mais famosa é a distância euclidiana quadrada. No contexto de dados funcionais, temos uma diversidade de versões desse algoritmo. Em García, García-Ródenas e

Gómez (2015) e Zambom, Collazos e Dias (2019), temos algumas versões sendo que uma delas traz uma proposta usando testes de hipóteses na construção do modelo. O algoritmo completo com explicações detalhadas pode ser encontrado em Hartigan e Wong (1979) e Abraham *et al.* (2003).

Como o objetivo é dividir as $n$ curvas em $k$ grupos (pré-determinados), assim temos que as componentes de $\beta^i$ são os coeficientes das curvas escritas por uma $f(x)$ e $\hat{\beta}^i \in \mathbb{R}^{K+d+1}$. Ao aplicar o K-Médias temos interesse em, a partir do conjunto $[\hat{\beta}^1,...,\hat{\beta}^n]$, selecionar um conjunto $z = [c^1,...,c^k]$ onde cada $c^1$ está associado e pertence a $\mathbb{R}^{K+d+1}$, assim deseja-se realizar a minimização da expressão

$$\frac{1}{n}\sum_{i=1}^{n} \min_{c\in z} ||\hat{\beta}^j - c||^2, \tag{3.22}$$

sendo que ao escrever a notação $||.||$ representamos uma norma, usualmente trazemos a euclidiana. Aplicando a norma euclidiana em (3.22), essa proposta é equivalente a escrever

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{\hat{\beta}^i\in C^j}^{n} ||\hat{\beta}^j - c^j||^2, \tag{3.23}$$

em que estamos procurando uma partição $[C^1,...,C^k]$ de $[\hat{\beta}^1,...,\hat{\beta}^n]$ em $k$ classes ($\hat{\beta}^j$ representa um vetor aleatório qualquer). Assim, (3.23) atinge seu mínimo quando $c^i$ é o centro de $C^i$.

Como temos uma sequência de vetores aleatórios de coeficientes identicamente distribuídos $\beta^n = (\beta^1,...,\beta^n)$, podemos reescrever (3.23) associando cada $\beta^n$ a um centroide $z = [c^1,...,c^2] \subset \mathbb{R}^{K+d+1}$,

$$u_n(\beta^n, z) := \frac{1}{n}\sum_{i=1}^{n} \min_{c\in z}||\beta^i - c||^2, \tag{3.24}$$

sendo assim, associamos as coordenadas das funções a um centro, que desejamos organizar de tal forma a minimizar (3.24). O K-Médias é afirmado como um procedimento consistente, e pode ser verificado a sua prova e definições em Lemaire (1983).

Podemos escrever o algoritmo que estima esse modelo seguindo o Algoritmo 2.

**Algoritmo 2** – Algoritmo K-Médias

1: No primeiro passo vamos dar palpites iniciais para os centros dos agrupamentos propostos (podem atribuídas observando os quantis amostrais);
2: Vamos classificar $\hat{\beta}^i$ cada a partir dos centros, até que todos os $\hat{\beta}^i$ estejam pertencendo a um agrupamento;
3: Partindo dos resultados obtidos na etapa anterior, devemos agora recalcular os $c^j$'s centroides de cada agrupamento a partir das médias dos $\hat{\beta}^i$ que o $j-ésimo$ grupo contém;
4: Após realizarmos os cálculos se os centros dos agrupamentos são os mesmos, então o algoritmo para. Caso contrário, repita os passos anteriores.

### 3.3.1 Agrupamento de Ruídos

A (3.22) que temos como objetivo minimizar, pode ser influenciada por ruídos presentes nos dados. Como a soma dos membros de um grupo deve ser igual a unidade, temos que uma observação deve ser atribuída a somente uma das classes.

Assim, mesmo um *outlier* deve pertencer a uma classe. Isso pode criar o problema no momento da minimização entre as distâncias (da observação e os pontos de referência dos grupos criados). Mesmo no caso de associações difusas, um *outlier* ainda deve ser atribuído a uma ou mais classes.

Uma solução ideal seria aquela em que os pontos de ruído fossem identificados e removidos dos dados. Como na prática essa solução é impossível, é proposta uma forma de tratar os ruídos para que sua influência seja mínima ou anulada nos agrupamentos.

Segundo Dave (1991), podemos propor sempre um número de grupos maior do que o número de classes de objetos. Por exemplo, em um caso de duas classes para realizar um agrupamento, serão propostos três grupos, sendo que esperamos que de maneira assertiva, as observações sem ruído sejam separadas corretamente em duas classes, e as observações ruidosas sejam agrupadas na terceira classe proposta. Sendo assim, retiramos a influência dos *outliers*, no agrupamento das demais curvas.

Se um protótipo de ruído é uma entidade universal tal que está sempre a mesma distância de todos os elementos do conjunto de dados e $v_c$ é um protótipo de ruído, $x_k$ é uma observação de uma variável, com ambas sendo definidas em $\mathbb{R}^p$. Logo, o protótipo de ruído é dado, tal que a distância entre $v_c$ e $x_k$ é escrita como $d_{ck}$, na forma

$$d_{ck} = \delta, \tag{3.25}$$

onde todos os elementos estão a uma distância $\delta$ de $v_c$. Logo, todos têm igual probabilidade, em um primeiro momento, de pertencer ao agrupamento de ruídos e a cada passo essa probabilidade é otimizada pelo algoritmo.

Podemos considerar que $c-1$ grupos são os verdadeiros agrupamentos presentes nos dados e o $c-$ésimo grupo é o agrupamento dos ruídos, desta forma, escrevemos

$$J_N(U,v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(d_{ik})^2, \tag{3.26}$$

onde $c$ são os grupos propostos, $c \in \mathbb{Z}$, $k = 1,...,n$ sendo $n$ o número de observações, $v$ são os protótipos ($c-1$ são os agrupamentos verdadeiros e o $c-$ésimo é o agrupamento de ruídos), $m$ é um expoente definido $1 < m < \infty$, $u_{ik}$ é o membro do $k$-ésimo ponto pertencente a classe $i$,

restritivamente temos que

$$\sum_{i=1}^{c}(u_{ik}) = 1, \tag{3.27}$$

onde em um caso normal é dado valor um para $u_{ik}$ se ele pertence à classe e zero caso contrario. As distâncias $d_{ik}$ são definidas como

$$(d_{ik})^2 = < x_k - v_i >^T A_i < x_k - v_i > \text{ para } i = 1, ..., c-1, \tag{3.28}$$

$$(d_{ik})^2 = \delta^2 \text{ para } i = c, \tag{3.29}$$

sendo $A_i$ é a matriz positiva definida, que no caso Euclidiano é a matriz identidade.

Sendo assim, queremos definir a distância $\delta$. Uma definição errada pode significar que muitos elementos além dos ruídos serão atribuídos ao agrupamento de ruídos, e uma definição muito restritiva pode não conseguir capturar todos os indivíduos desejados, uma possível parametrização para $\delta$ pode ser escrita como

$$\delta^2 = \lambda \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(d_{ik})^2}{n(c-1)}, \tag{3.30}$$

sendo $\lambda$ um múltiplo usado para obter $\delta$ da média das distâncias.

### 3.3.2 K-Medoides Difuso Funcional

O método de K-Medoides (Kaufman e Rousseeuw (2009)) é semelhante ao K-Médias, em que o objetivo é minimizar a distância entre os pontos observados e os pontos designados como o centro dos agrupamentos. A diferença é que o K-Medoides usa métricas mais interessantes para calcular os centroides e assim a dissimilaridade média para todos os objetos do grupo é minimizada. O algoritmo de K-Medoides também possui características desejadas como ser robusto na presença de ruído ou *outliers*.

Em adicional, é proposta também uma visão difusa no agrupamento não supervisionado, isto é, a restrição escrita na (3.27) é valida, só que para cada $u_{ik}$ podemos obter um valor qualquer entre $[0,1]$. Com essa informação, podemos determinar uma regra mais complexa para o agrupamento da curva observada. Uma regra utilizada no agrupamento difuso é a atribuição do grupo de acordo com maior valor observado em $u_{ik}$.

Uma abordagem do K-Medoides é a minimização da soma da variação intragrupo (SICV), detalhado em Park e Jun (2009) e Chiang, Russell e Braatz (2001).

Seja um conjunto de dados $X = (x_1, ..., x_n)$ com $n$ variáveis os grupos do modelo K-Medoides são gerados selecionando um conjunto de $k$ membros de $X$ como medoide, e atribuindo

cada não membro selecionado de $X$ para seu medoide mais próximo. Desta forma definimos o SICV, como

$$SICV = \sum_{i=1}^{n} \sum_{j=1, x_i \in C_j}^{k} d(x_i, m_j), \tag{3.31}$$

sendo $d(x_i, m_j)$ a distância entre a $i-$ésima variável $x_i$ e o $j-$ ésimo medoide $m_j$ ($d(x_i, m_j)$ também pode ser chamada de dissimilaridade).

Considerando um agrupamento válido no conjunto $X$, escrito como $C = [C_1, ..., C_k]$ temos as propriedades:

- $C_i \neq \emptyset$, $1 \leq i \leq k$;
- $\cup_{i=1}^{k} C_i = X$;
- $C_i \cap C_j = \emptyset$, $i \neq j$, $1 \leq i, j \leq k$.

Como desejamos o melhor desempenho no agrupamento difuso proposto, é interessante validar com outras métricas os agrupamentos construídos pelo modelo. Nesta abordagem usaremos o coeficiente de partição modificado (CPM) em que desejamos quantificar a qualidade de separação conjuntamente com a quantidade de membros que pertencem ao grupo de acordo com a partição difusa. Esse é escrito na forma

$$CPM = 1 - \frac{c}{c-1}(1 - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{C} (u_{ik})^2). \tag{3.32}$$

Esta métrica é definida no intervalo $[0, 1]$, em que $CPM = 0$ corresponde à máxima difusão e $CPM = 1$ a uma partição robusta.

## 3.4 Rede Gás Neural

A rede gás neural (Martinetz, Berkovich e Schulten (1993)) é um algoritmo de rede neural que apresenta um bom desempenho quando trabalhamos com a abordagem não supervisionada. A regra empregada pelo algoritmo é uma adaptação *soft-max* ao procedimento de K-Médias. Nesta melhoria é realizada uma penalização que ajuda tanto na qualidade do ajuste quanto na velocidade de convergência.

A função *soft-max* é uma função que transforma um vetor de valores reais em um vetor de valores reais que somados resultam em 1 que será denotado $K$. A ideia desta função é transformar qualquer entrada (sendo positiva ou negativa) em valores que possam ser interpretados como probabilidades.

Definimos, assim, o espaço total de vizinhanças como $V \subset \mathbb{R}$ e $v \in V$, dessa forma, para cada vetor $v$ determinamos um ordenamento da vizinhança denotado como $w_{i_0}, w_{i_1}, ..., w_{i_{N-1}}$. Sendo $w_i \in \mathbb{R}^{\mathbb{D}}$ (em que $D$ representa o número de dimensões do banco de dados) para $i = 1, ..., N$, em que a referência é $w_{i_0}$ (ponto inicial da ordenação) pois é o mais próximo de $v$, com $k = 0, ..., N-1$.

Assim, uma maneira de dividir o espaço total das vizinhanças $V$ é descrito na metodologia de Voronoi (Reem (2009)), em que para cada $w_j$ temos $||v - w_j|| < ||v - w_{i_k}||$ para $j = 1, ..., N$ e $j \neq i$. Se denotarmos o número $k$ associado aos $w_i$ como $k_i(v,w)$ dependente de $v$ e de todo o conjunto $w_1, ..., w_N$ escrevemos o ajuste de $w_i's$ como

$$\Delta w_i = \varepsilon \cdot h_\lambda(k_i(v,w)) \cdot (v - w_i) \text{ para } i = 1, ..., N, \tag{3.33}$$

sendo $\varepsilon \in [0,1]$, $h_\lambda(k_i(v,w))$ é igual a 1 para $k_i = 0$ e tende a 0 conforme incrementamos $k_i$ e $\lambda$ é uma constante de decaimento. Podemos usar $h_\lambda(k_i(v,w)) = e^{\frac{-k_i(v,w)}{\lambda}}$ e se $\lambda \to 0$ neste caso (3.33) se torna equivalente ao K-Médias.

A função custo que desejamos minimizar será denotada por $E_{ng}(w, \lambda)$ e é definida por

$$E_{ng}(w, \lambda) = \frac{1}{2C(\lambda)} \sum_{i=1}^{N} \int d^D v P(v) h_\lambda(k_i(v,w))(v - w_i)^2 \tag{3.34}$$

$$\text{em que } C(\lambda) = \sum_{k=0}^{N-1} h_\lambda(k), \tag{3.35}$$

sendo $\lambda$ é um fator de normalização, $d$ é o erro (chamado de erro de distorção). A função $E_{ng}$ é baseada no agrupamento difuso, associando os dados em $v$ a um vetor de referência $w_{i(v)}$, em que $p_i(v)$ é chamado de membro de difusão de $v$ no agrupamento $i$. Logo, escrevemos $p_i(v) = \frac{h_\lambda(k_i(v,w))}{C(\lambda)}$, em que $p_i(v)$ indica a probabilidade do $w_{i(v)}$ referenciado pertencer a vizinhança $v$.

### 3.4.1 Arquitetura da Rede Neural

O algoritmo de rede gás neural conta com dois fatores para determinar a sua arquitetura de rede, a minimização da função custo 3.34 e o mapa Kohonen (Kohonen (1982)). O mapa Kohonen é uma forma de estruturar a topologia da rede usando uma técnica auto-organizável.

A técnica de Kohonen é amplamente usada no aprendizado de máquina não supervisionado para produzir uma representação com redução de dimensionalidade, preservando a estrutura topológica dos dados (no espaço de entrada). Logo, a camada de entrada dos dados tem dimensões equivalente ao número de cováriaveis, a camada intermediaria faz a auto-organização essencialmente com as características dos dados de entrada, e agrupando de acordo com a semelhança.

Logo, todos os nós nesta topologia estão conectados diretamente ao vetor de entrada, mas não uns aos outros, o que significa que os nós não sabem os valores de seus vizinhos, e

apenas atualizam os pesos de suas conexões em função das entradas dadas. Ao aplicar os dados de treinamento na rede o mapa se organiza a cada iteração.

É comum que os mapas auto-organizáveis usem a aprendizagem competitiva, desta forma, é feito uma seleção do nó vencedor. O nó vencedor é aquele que mais se assemelha aos dados de entrada. Para selecionar o nó vencedor, um a um os nós são ativados em cada iteração. Assim, cada vez que um vetor de dados $v$ é apresentado, não apenas o nó vencedor é ajustado, mas seus vizinhos próximos e os vetores de referência $w_i$, está etapa de adaptação pode ser escrita como

$$\Delta w_i = \varepsilon \cdot h_\sigma(i, i(v)) \cdot (v - w_i) \text{ para } i = 1, ..., N, \tag{3.36}$$

tal que, $h_\sigma(i, j)$ (sendo $(i, j)$ dois pontos quaisquer) é uma função unimodal que diminui monotonicamente a medida que $||i - j||$ aumenta, com uma constante de decaimento $\sigma$. Sendo que $i(v)$ e $i$ representa o nó vencedor e seus adjacentes.

## 3.5 Métricas de Performance

Neste trabalho usaremos algumas métricas para avaliar se os modelos aplicados, tanto a dados simulados quanto a dados reais, obtiveram sucesso em realizar o agrupamento não supervisionado minimamente coerente.

### 3.5.1 Matriz de Confusão e Acurácia

Iremos analisar os agrupamentos realizados pelos modelos em duas óticas de acurácia, a primeira é mais geral, considerando-se todos os grupos ajustados pelo modelo obteve êxito em atingir uma porcentagem mínima estipulada de 80%. Desta forma, a equação que define a acurácia total do modelo é dado por

$$AC_{total} = \frac{\text{Número de Casos Agrupados Corretamente}}{\text{Número de Casos Totais}}. \tag{3.37}$$

Será considerada também a acurácia de cada agrupamento individualmente. O intuito de levar em consideração esta métrica é que esperamos que os modelos utilizados façam uma boa separação espacial das curvas. Sendo assim, cada grupo deve ficar bem definido, de tal forma que os agrupamentos sejam consistentes e evitado o caso em que um agrupamento único "englobe" todas as observações. Definimos então a acurácia de $i-$ésimo grupo como

$$AC_{\text{grupo i}} = \frac{\text{Número de Curvas do Grupo } i \text{ Agrupados Corretamente}}{\text{Número Total de Curvas do Grupo } i}. \tag{3.38}$$

Um outro cuidado que será tomado é que para os modelos terem a liberdade de criar um nova classe, será exigido que o modelo atribua no mínimo 30% das curvas do total de observações disponíveis na base, para a classe nova criada, para que ela seja elegível.

Essas informações podem ser extraídas da chamada matriz de confusão, que é definida pela Tabela 2.

Tabela 2 – Definição da Matriz de Confusão.

| Matriz de Confusão | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | A | B |
| Agrupamento do Modelo | A | VP | FP |
| | B | FN | VN |

Fonte – Elaborada pelo autor.

Onde temos que verdadeiro positivo (VP) são os casos em que foi classificado corretamente a classe de interesse, falso positivo (FP) que indica a classificação errônea da classe de interesse, verdadeiro negativo (VN) aponta que a classe que não é de interesse foi prevista de maneira correta e por fim, falso negativo (FN) que informa quando a classe que não é de interesse foi classificada de forma errada.

A matriz de confusão pode ser interpretada da seguinte forma: a diagonal principal da matriz mostra os grupos que foram agrupados de forma correta e a soma da diagonal principal indica o número total de casos agrupados corretamente, as demais caselas apontam os grupos que foram agrupados erroneamente.

### 3.5.2 Medida V

Para este trabalho será considerada uma métrica para avaliar a homogeneidade dos agrupamentos e sua completude.

Ao analisar a homogeneidade de um grupo, é esperado que todos os objetos daquele grupo sejam de tal forma que todos detenham um único rótulo. Considerando que temos um conjunto de $N$ observações e que $k$ número de grupos definidos pelo modelo avaliado e $c$ as verdadeiras classes.

A qualidade do ajuste realizado pelo modelo é denotado por $h$, logo essa propriedade é definida matematicamente como

$$h = 1 - \frac{H(C|K)}{H(C)}, \tag{3.39}$$

$$\text{em que } H(C|K) = -\sum_{c,k} \frac{n_{ck}}{N} \log(\frac{n_{ck}}{n_k}), \tag{3.40}$$

sendo $n_{ck}$ é o número de observações rotuladas originalmente $c$ no agrupamento $k$, $n_k$ é o número total de observações no grupo $k$ e $N$ é o número total de observações.

Um agrupamento completo é aquele em que todos os pontos de dados pertencentes à mesma classe são agrupados no mesmo grupo, podemos assim definir que $H(C|K)$ é completo quando $H(K|C) = 0$. A completude de um agrupamento é dada letra $\zeta$, expressa na forma

$$\zeta = 1 - \frac{H(K|C)}{H(K)}, \tag{3.41}$$

$$\text{em que } H(K|C) = -\sum_{c,k} \frac{n_{ck}}{N} log(\frac{n_{ck}}{n_c}), \tag{3.42}$$

$n_{ck}$ é o número observações rotuladas originalmente $c$ no agrupamento $k$, $n_c$ é o número total de observações originalmente rotuladas como $c$ e $N$ é o número total de observações, pode-se dizer que $H(K|C)$ é denominado entropia condicional. A entropia condicional quantifica a quantidade de informação necessária para descrever o resultado de uma variável aleatória dado que o valor de uma variável conhecida.

Sendo assim, segundo Rosenberg e Hirschberg (2007), podemos definir a medida V ($V_\beta$) como

$$V_\beta = (1+\beta)\frac{h\zeta}{\beta h + \zeta}, \tag{3.43}$$

sendo $\beta$ um parâmetro de penalização com o valor padrão de $\beta$ sendo 1, $h$ o parâmetro de homogeneidade e $\zeta$ o parâmetro de completude.

Podemos interpretar a medida V, como uma avaliação da qualidade do agrupamento realizado pelo modelo proposto. Nela temos dois pontos, o primeiro é a homogeneidade do agrupamento, se o grupo apresenta coerência nas observações contidas e um segundo ponto levado em consideração é a completude, se o grupo abrange um número razoável de observações. Esta métrica é definida no intervalo $[0,1]$, sendo que quanto mais perto de 1 melhor a qualidade do agrupamento feito pelo modelo.

### 3.5.3 Validação do Número de Grupos

Neste trabalho será feito um estudo, que visa compreender qual seria o número ideal de grupos, para a situação de dados reais e simulados, caso o número real de grupos não fosse conhecido ou dispusesse uma grande incerteza sobre tal valor.

Desta forma, para avaliar essa situação será usados duas métricas para avaliação o Índice Calinski-Harabasz (calinhara) e a Pontuação de Silhouette (silhouette) (Caliński e Harabasz (1974) e Rousseeuw (1987a), respectivamente)

#### 3.5.3.1 Índice Calinski-Harabasz (calinhara)

O Índice Calinski-Harabasz também conhecido como calinhara ou Critério da Razão de Variância (Caliński e Harabasz (1974)), é uma métrica estimada com base em distâncias, onde é avaliado se as observações em cada agrupamento feita pelos modelos propostos estão próximas (ou seja, o grupo tem alta densidade), enquanto as distancias entre os agrupamentos estão mais distantes uns dos outros (logo, estão bem separados). Sendo que quanto maior o valor do índice, mais provável é que o número de agrupamentos associado ao índice de maior valor seja o número verdadeiro de grupos do conjunto de dados

Nesse índice é calculado a dispersão inter-agrupamentos ou soma de quadrados entre grupos ($SQEG$) é calculada como

$$SQEG = \sum_{k=1}^{K} n_k ||C_k - C||^2, \tag{3.44}$$

sendo $n_k$ o número de observações e $C_k$ o centroide no agrupamento $k$, $C$ o centroide do conjunto de dados e $K$ o número total de agrupamentos propostos.

Então, é calculado também dispersão intra-agrupamentos ou a soma de quadrados dentro do grupo (SQDG)

$$SQDG = \sum_{k=1}^{K} SQDG_k = \sum_{i=1}^{n_k} ||X_{ik} - C_k||^2, \tag{3.45}$$

onde, $n_k$ o número de observações, $C_k$ o centroide e $X_{ik}$ a $i-$ésima observação no agrupamento $k$ e $K$ o número total de agrupamentos propostos.

Por fim escrevemos o calinhara como

$$CH = \frac{SQEG}{SQDG} \frac{N-K}{K-1}, \tag{3.46}$$

com o valor de $N$ sendo o número total de observações.

#### 3.5.3.2 Pontuação de Silhouette (silhouette)

O método de Silhouette (Rousseeuw (1987a)), é uma metodologia baseada em distancias, dela é possivel extrair a pontuação de silhouette que é uma métrica usada para calcular a qualidade de um ajuste de agrupamento realizado.

Seu valor varia de $-1$ a 1, sendo que valores próximos a 1 indicam que os agrupamentos estão bem separados uns dos outros, valores próximos a 0 pode-se dizer que a distância entre os agrupamentos não é significativa e coeficientes próximos a $-1$ indicam que os agrupamentos são atribuídos de forma errada.

A equação que define o pontuação de silhouette, pode ser escrita na forma

$$S(i) = \frac{b(i) - a(i)}{max[a(i), b(i)]}, \tag{3.47}$$

onde $S(i)$ é a pontuação da $i-$ésima observação, $a(i)$ é um valor médio de distancia entre a $i-$ésima e as demais observações inter-agrupamento e $b(i)$ é um valor médio de distancia entre a $i-$ésima e as observações intra-agrupamento. Desta forma para encontrar a pontuação de silhouette de cada proposta (de número real de agrupamento, quando desconhecido), é calculado a média dos $S(i)$, para cada ajuste proposto.

CAPÍTULO

# 4

# ESTUDO DE SIMULAÇÃO

Com o objetivo de avaliar o desempenho e a assertividade das metodologias propostas em relação aos métodos mais usuais que foram encontrados na literatura, foram realizados dois estudos de simulações, com o objetivo de evidenciar algumas situações recorrentes nas análises de dados funcionais. Todas as análises e simulações foram feitas usando a linguagem R.

No primeiro estudo ou simulação 1 foram abordadas duas funções, nomeadas de primeira função simulada (F1) e segunda função simulada (F2). Neste caso, as simulações apresentam uma diferença nas curvas que pode ser evidenciada até visualmente (apresentado na Figura 1).

Já no segundo estudo ou simulação 2 foram adotadas as mesmas funções F1 e F2 da simulação 1, e também, foi construída uma terceira função nomeada de terceira função simulada (F3), que tem por objetivo trazer mais um comportamento desafiando as metodologias a realizar um bom agrupamento não supervisionado e confundir-se com as simulações de F1 e F2 por apresentar um comportamento semelhante (Figura 7).

Para avaliar o desempenho dos algoritmos, analisaremos sua performance em três pontos: considerando o agrupamento total realizado (se no geral, tivemos uma boa separação dos grupos), a assertividade de cada uma das classes (se não houve uma mistura de grupos que são evidentemente originários de funções distintas) e a homogeneidade e completude dos agrupamentos através da medida V.

## 4.1 Simulação 1: Duas Classes de Funções

Para o caso da simulação 1, foram adotadas duas funções (F1 e F2) que apresentam comportamentos bem diferentes, podendo até mesmo serem identificadas visualmente em um gráfico de duas dimensões. O objetivo principal deste estudo é mostrar como as técnicas em diferentes graus de complexidade se comportam em uma situação de extrema simplicidade quando queremos realizar o agrupamento não supervisionado. A situação abordada na simulação

1 é simples, mas pode acontecer no contexto aplicado, um exemplo é o estudo em Bogetoft e Otto (2010).

Foram realizadas 30 curvas sintéticas das funções 1 e 2, cada uma contendo 200 pontos. Com o intuito de que os dados sintéticos produzidos sejam parecidos com dados espectrais reais, as funções F1 e F2 são dadas por

$$F1(x_t) = (0.8 \cdot F1(x_{t-1})) + \varepsilon \tag{4.1}$$

$$\text{e } F2(x_t) = (0.2 \cdot F2(x_{t-1})) + \varepsilon^2, \tag{4.2}$$

para $t = 1, 2, ..., 200$ e $\varepsilon$ é um ruído branco aleatório, podemos observar duas curvas simuladas provindas de (4.1) e (4.2) na Figura 1. Para realizar as simulações foi considerado o algoritmo (3).

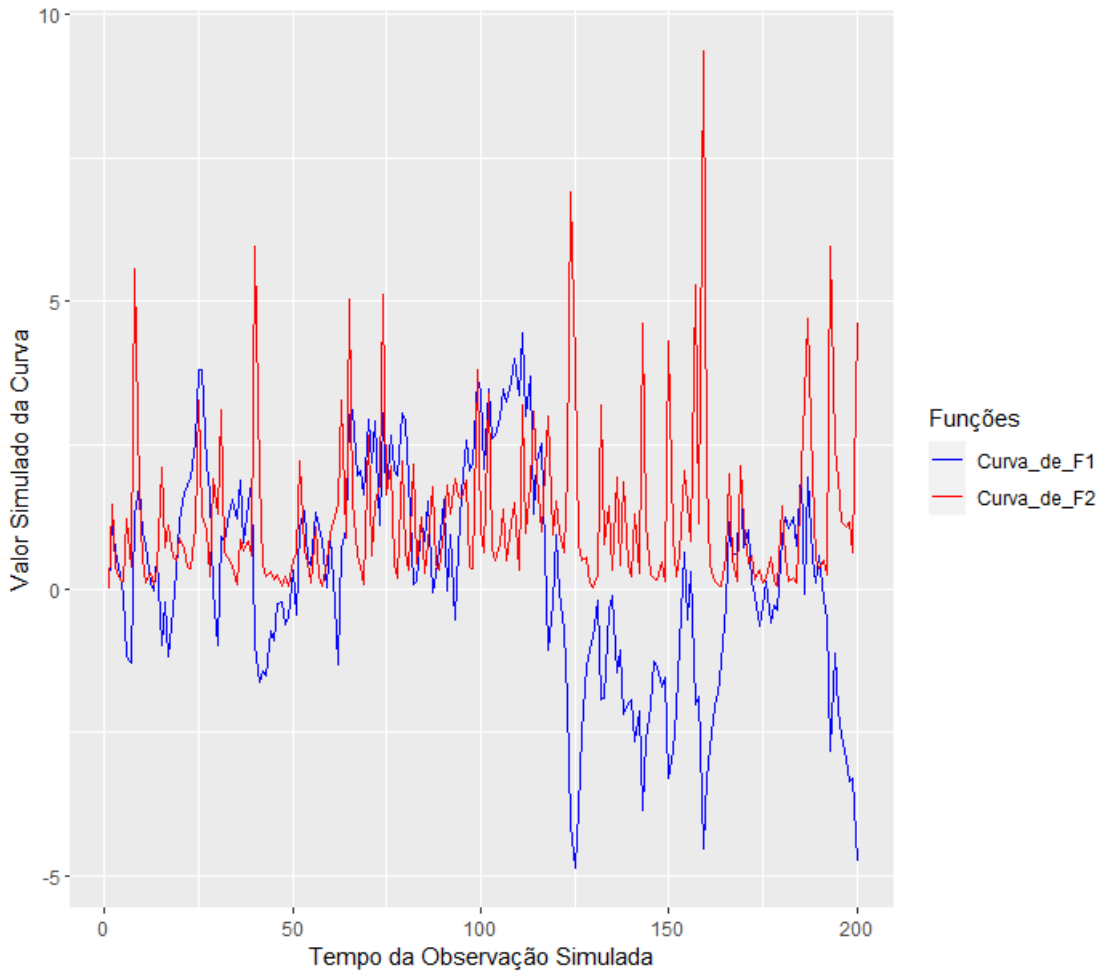


Figura 1 – Exemplificação das curvas simuladas pelas funções F1 e F2.

Fonte – Elaborado pelo autor.

**Algoritmo 3** – Pseudocódigo para a Simulação.

1: Inicio: código
2: Definir a semente "1234";
3: $x_t$ = vetor e declarar um chute inicial (recomendado valor "0");
4: Inicializar $x_t mat$ como uma matriz com 30 linhas e 200 colunas
5: Inicio: looping em $j$ de 1 á 30;
6: Inicio: looping $i$ variando de 2 á 200;
7: $prob$ = uma uniforme (U[0;1])";
8: $ruidobranco$ = uma normal (N[0;1]) usando a $prob$;
9: $x_t[i] = f(x)_{\text{função proposta da simulação}} + ruidobranco$ (esse é o valor simulado das funções no ponto $i$ da curva $j$);
10: Fim: looping $i$;
11: $x_t mat[j,] = x_t[i]$;
12: Fim: looping $j$;
13: Fim: código.

Após as simulações realizadas, a primeira abordagem feita foi considerando somente o ACPF. Como esta abordagem em especifico tem uma saída diferente das demais metodologias que foram aplicadas, iremos avaliar somente de forma visual através de gráficos. Podemos observar na Figura 2 o agrupamento das curvas de F1 e F2.

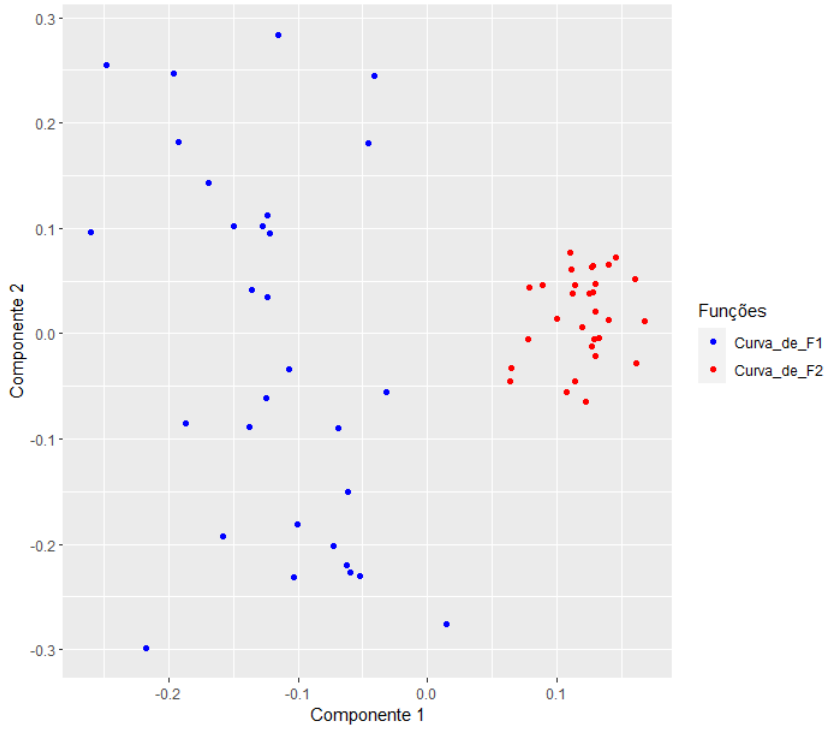


Figura 2 – Agrupamento das curvas pelo método ACPF.

Fonte – Elaborado pelo autor.

Podemos observar na Figura 2 que a assertividade do algoritmo foi de 100% no agrupamento total e não ocorreu superposição de grupos.

Em alguns casos, é comumente aplicado um certo número componentes da ACPF em outros modelos substituindo as variáveis de entrada original, com o intuito de que essas metodologias conjuntas tragam resultados melhores do que cada uma delas individualmente.

Desta forma, foram aplicadas de 2 a 30 componentes ao K-Médias. A quantidade de componentes que produziu o melhor desempenho foi a de duas componentes, ou seja, essa foi a seleção que produziu o melhor resultado quando foi aplicado ao K-Médias. Sua matriz de confusão pode ser observada na Tabela 3. Temos também que a Figura 3 ilustra o agrupamento de cada curva assim como a curva média de cada um dos agrupamentos.

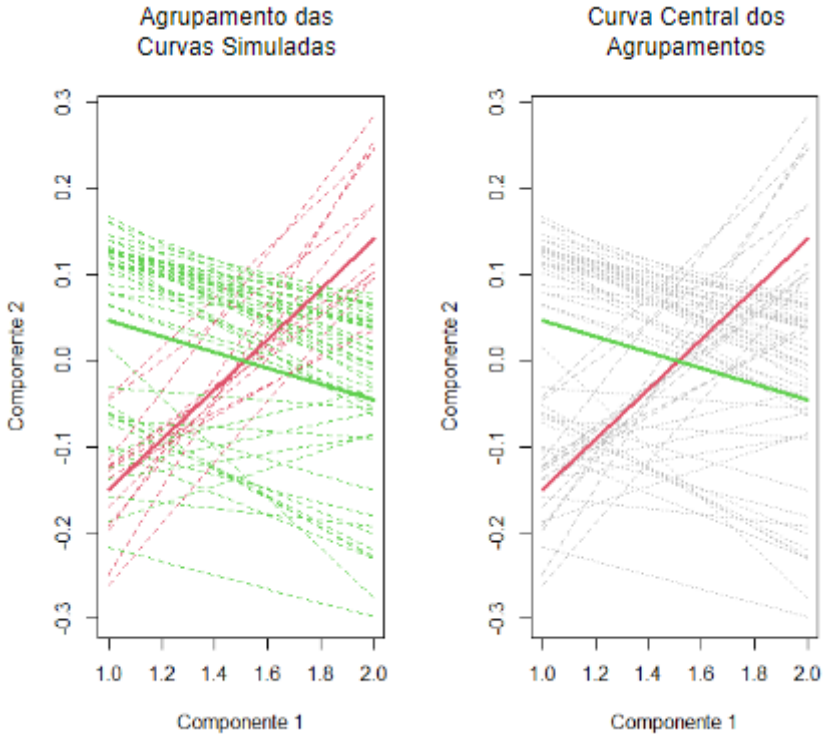


Figura 3 – Visualização do agrupamento do K-Médias, aplicado conjuntamente com as duas primeiras componentes principais resultantes do ACPF. A cor verde representa F1 e a vermelha F2.

Fonte – Elaborado pelo autor.

Tabela 3 – Matriz de confusão resultante da metodologia aplicando ACPF e o K-Médias na sequencia, nos dados simulados de duas funções.

| Matriz de Confusão K-Médias com ACPF | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | Função 1 | Função 2 |
| Agrupamento do Modelo | Função 1 | 14 | 0 |
| | Função 2 | 16 | 30 |

Fonte – Elaborada pelo autor.

Na Tabela 3 vemos que a acurácia total do agrupamento foi de 73%. Todavia para a F1 a assertividade foi de 46% e para F2 foi de 100%. Logo, observa-se que mesmo obtendo uma

acurácia total razoavelmente boa de 73%, quando olhamos o agrupamento individual de cada curva, para a F1 tivemos menos de 50% de acurácia, sendo assim, podemos concluir que neste problema a abordagem conjunta de ACPF e K-Médias não foi benéfica.

Visto que conjuntamente não tivemos bons resultados, será observado como o K-Médias se comporta sozinho. Na Figura 4 temos visualmente o agrupamento individual de cada curva e na Tabela 4 o resultado na matriz de confusão.

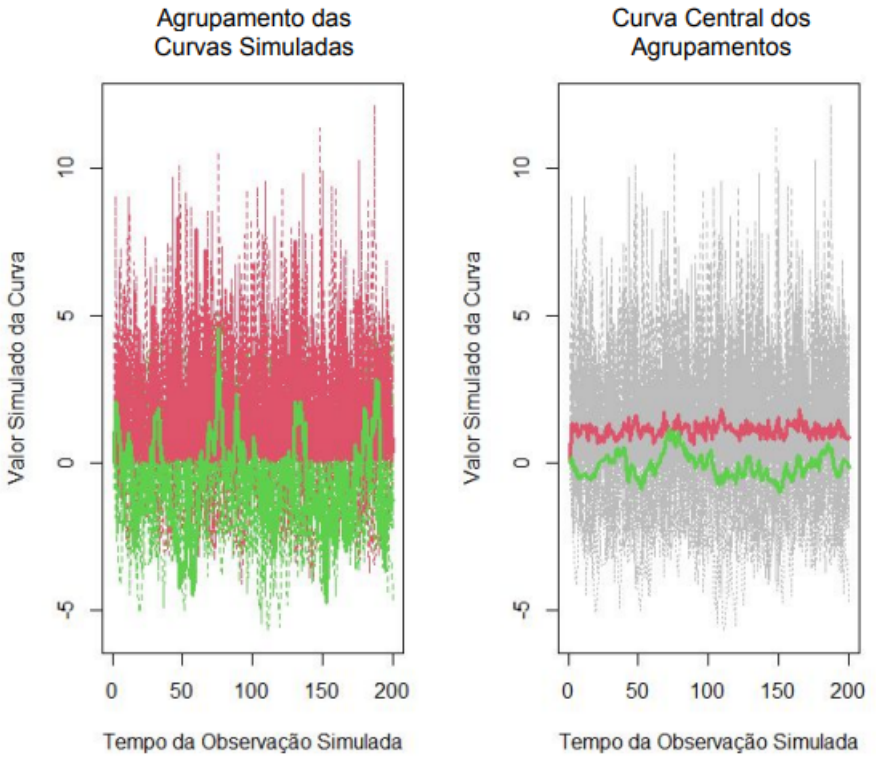


Figura 4 – Visualização do agrupamento do K-Médias, aplicado sozinho. A cor verde representa F1 e a vermelha F2.

Fonte – Elaborado pelo autor.

Tabela 4 – Matriz de confusão resultante do K-Médias, nos dados simulados de duas funções.

| Matriz de Confusão K-Médias | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | Função 1 | Função 2 |
| Agrupamento do Modelo | Função 1 | 23 | 0 |
| | Função 2 | 7 | 30 |

Fonte – Elaborada pelo autor.

Na Tabela 4 vemos que a assertividade total do K-Médias foi de 88%, e para F1 e F2 foi respectivamente, 76% e 100%. Para este caso, temos que somente o ACPF individualmente teve uma performance superior ao K-Médias com ACPF e K-Médias somente.

Observamos que em (4.1) e (4.2) as funções apresentam uma dependência do valor anterior. Nesse sentido, uma outra analise se mostrou interessante, considerando o funHDDC, já que é adequado para quando temos uma dependência. Os resultados estão na Figura 5 e na Tabela 5.

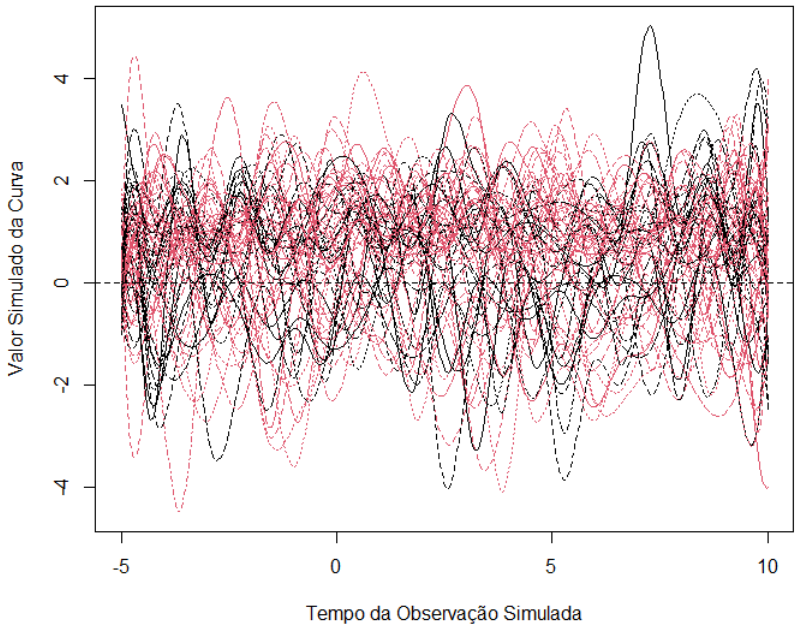


Figura 5 – Visualização do agrupamento do funHDDC, aplicado conjuntamente com as alterações de base e suavizações. A cor preta representa F1 e a vermelha F2.

Fonte – Elaborado pelo autor.

Tabela 5 – Matriz de confusão resultante do funHDDC, nos dados simulados de duas funções.

| Matriz de Confusão funHDDC | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | Função 1 | Função 2 |
| Agrupamento do Modelo | Função 1 | 12 | 9 |
| | Função 2 | 18 | 21 |

Fonte – Elaborada pelo autor.

Nesta situação simplificada, o funHDDC não obteve um desempenho tão bom, sendo possivelmente o pior entre o ACPF e o K-Médias.

### 4.1.1 Simulações com Duas Classes de Funções e Repetições

Foram feitas 200 repetições das simulações mencionadas na Seção 4.1, com o objetivo de entender se em uma gama maior de conjuntos de dados os modelos analisados são consistentes nos agrupamentos propostos.

Serão considerados os modelos funHDDC e K-Médias, iremos considerar também uma abordagem difusa com os modelos de rede gás neural e K-Médias difuso, sendo que para o K-Médias o parâmetro de difusão é modelado por uma forma polinomial.

Para analisar a qualidade dos agrupamentos não supervisionados será usada a medida V. O resultado pode ser visto na Figura 6.

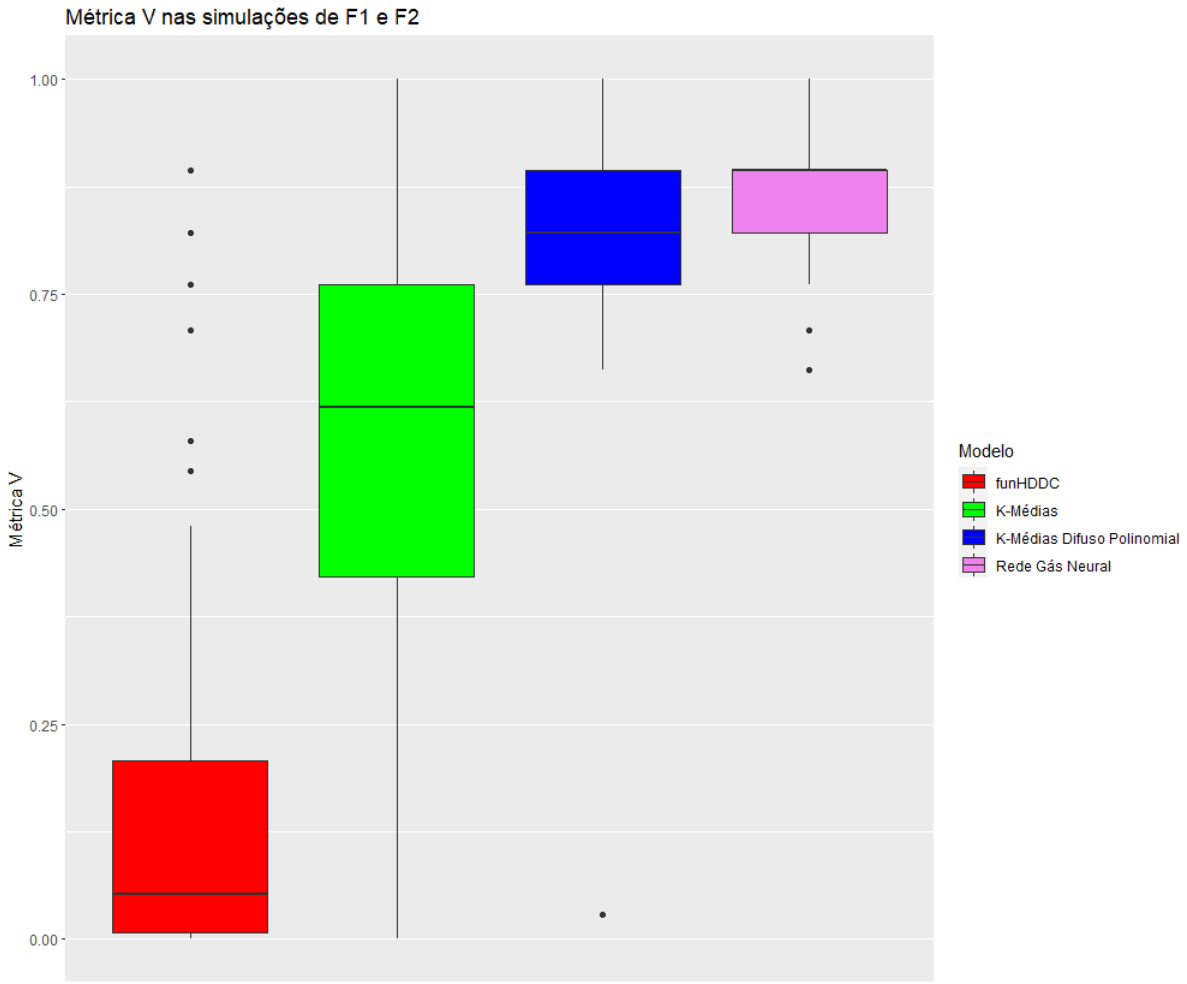


Figura 6 – Métrica V para as 200 repetições das simulações das funções F1 e F2.

Fonte – Elaborado pelo autor.

Observamos que a abordagem difusa trouxe bons resultados no agrupamento, os dois melhores modelos foram K-Médias difuso e a rede neural. Na analise da Seção 4.1 foi observado que a abordagem K-Médias que obteve bons resultados quando aplicada a uma amostra mais ampla não se mostrou consistente. Entretanto, modelando com metodologia difusa, obtivemos uma melhora significativa nos resultados.

### 4.1.2 Considerações sobre Simulações com Duas Classes de Funções

Na simulação 1, a partir do estudo realizado podemos observar que o melhor desempenho foi obtido usando somente o ACPF. Uma possível justificativa para este resultado é que devido à simplicidade e à fácil identificação até mesmo visual das funções F1 e F2.

Por mais que as outras metodologias não tenham tido uma boa precisão nos agrupamentos, não significa que são ruins, só que a situação tem peculiaridades nas quais os modelos não se sobressaíram. A visão difusa em particular, foi benéfica aos modelos e à análise, observando a métrica medida V, temos que a melhora foi significativa.

A critério de assegurar os resultados visualizados, será feito mais simulações, para compreender em um cenário mais amplo se os resultados observados se mantêm.

## 4.2 Simulação 2 : Três Classes de Funções

Temos que por intuito, em um contexto mais complexo, os algoritmos tendem a piorar suas acurácias (ou mantê-las), em relação a um contexto mais simplificado.

Para a simulação 2 adotamos novamente as funções (4.1) e (4.2), só que agora iremos incluir uma terceira função F3. A função F3 vem com o intuito de confundir os modelos e complicar o agrupamento. Assim, entender se em uma situação mais desafiadora, as metodologias aplicadas na simulação 1 mantêm o mesmo nível de performance, pioram ou melhoram. A F3 é dada por

$$F3(x_t) = (0.5 \cdot F3(x_{t-1})) \cdot (0.1 \cdot F3(x_{t-2})^2) + \varepsilon, \tag{4.3}$$

$t = 1, 2, ..., 200$ e $\varepsilon$ é um ruido branco aleatório.

Para obter as curvas de F3 foi usado novamente o Algoritmo 3. A Figura 7 apresenta as três curvas para ilustrar qual é a situação que iremos usar como base do estudo.

De maneira análoga à simulação 1, será aplicado o ACPF considerando as duas primeiras componentes e verificar a partir da Figura 8 como foi a assertividade da metodologia.

Na Figura 8 notamos um comportamento semelhante à simulação 1. Para as curvas F1 e F2, só que para F3 o método agrupou todos os pontos bem no centro dos pontos dispersos de F2. Sendo assim, o ACPF por mais que tenha tido um bom desempenho para F1 e F2, não conseguiu encontrar uma boa divisão para as curvas provenientes de F3. Vemos que a assertividade no agrupamento não supervisionado para cada uma das classes foi de 100% para F2, mas para F1 e F3 não ocorreu uma distinção que pudesse ser minimamente razoável.

Como na simulação 1 foram consideradas as duas primeiras componentes, para o segundo estudo (que foi também o número de componentes principais que produziu a melhor acurácia total), foram aplicadas novamente as duas primeiras componentes do ACPF no K-Médias para

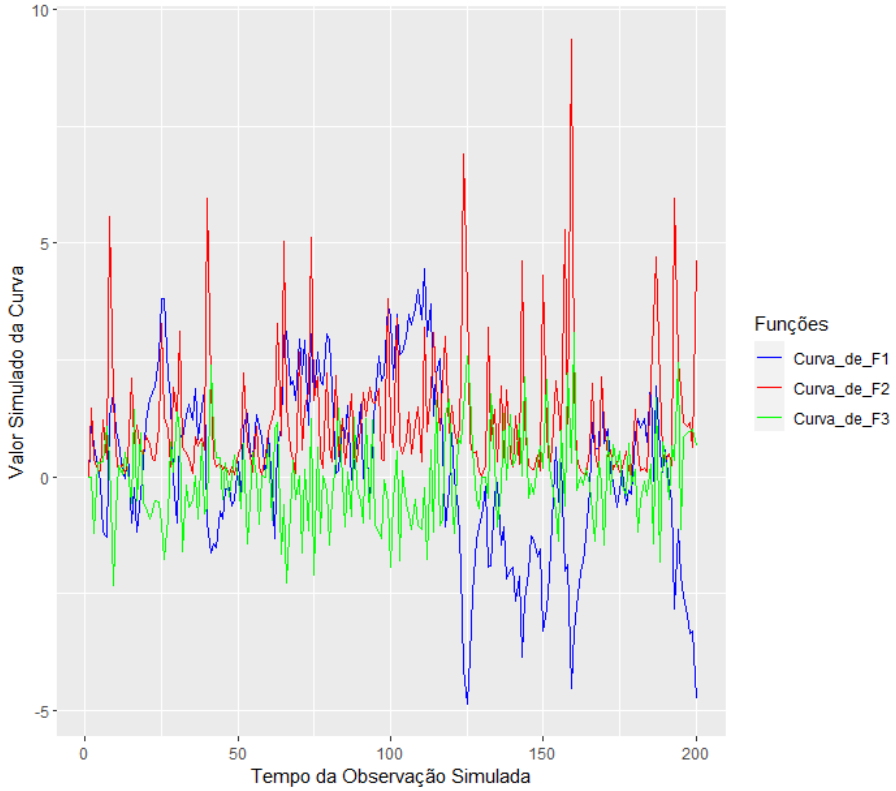


Figura 7 – Exemplificação das curvas simuladas pelas funções F1, F2 e F3.

Fonte – Elaborado pelo autor.

observar qual seria o desempenho das metodologias conjuntas. Os resultados estão na Figura 9 e Tabela 6.

Tabela 6 – Matriz de confusão resultante da metodologia aplicando ACPF e o K-Médias na sequencia, nos dados simulados de três funções.

| Matriz de Confusão | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Função 1 | Função 2 | Função 3 |
| Agrupamento do Modelo | Função 1 | 13 | 0 | 2 |
| | Função 2 | 4 | 30 | 28 |
| | Função 3 | 13 | 0 | 0 |

Fonte – Elaborada pelo autor.

A assertividade total foi de 47% e por grupo tivemos para F1, F2 e F3, 43%, 100% e 0% respectivamente. Pela Tabela 6, temos a situação em que o K-Médias considerou como um único grupo as curvas provindas das funções F2 e F3 (a menos de 2 repetições), além de fazer uma divisão para as curvas de F1, atribuindo parte dela como um agrupamento e a outra parte um agrupamento distinto.

Será analisado agora como ficaria o agrupamento das funções considerando somente o K-Médias. A Figura 10 apresenta como cada curva foi agrupada, juntamente com a curva média de cada grupo, e a Tabela 7 mostra os resultados do modelo.

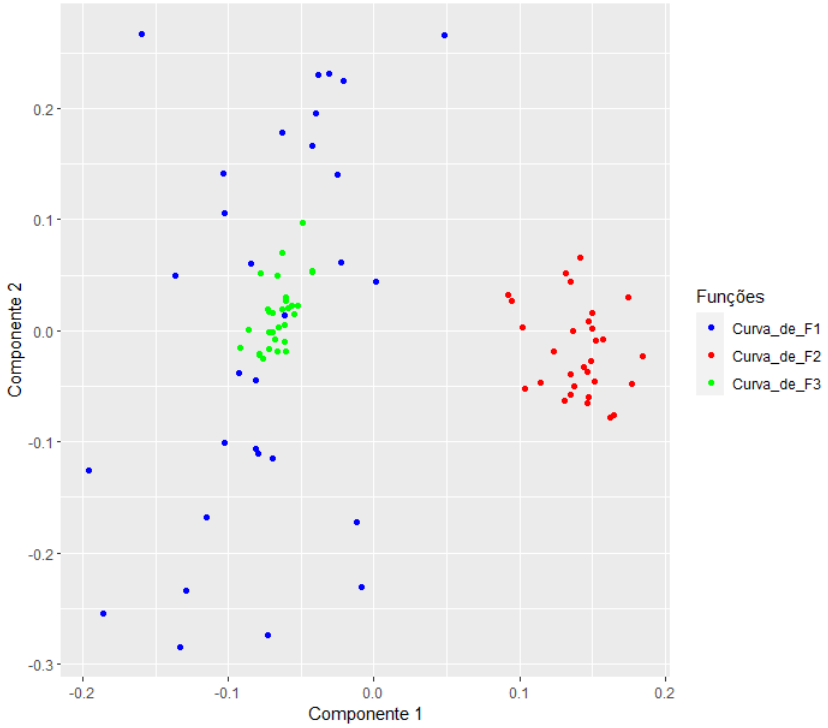


Figura 8 – Agrupamento das curvas pelo ACPF, onde visualizamos somente as duas primeiras componentes no plano de duas dimensões.

Fonte – Elaborado pelo autor.

Tabela 7 – Matriz de confusão resultante do K-Médias, nos dados simulados de três funções.

| Matriz de Confusão | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Função 1 | Função 2 | Função 3 |
| Agrupamento do Modelo | Função 1 | 13 | 0 | 2 |
| | Função 2 | 0 | 30 | 0 |
| | Função 3 | 17 | 0 | 28 |

Fonte – Elaborada pelo autor.

Pela Tabela 7, notamos uma melhora em relação à aplicação conjunta do ACPF e o K-Médias. A assertividade total foi de 78% superior aos 43% anteriores e por segmentos tivemos 43% para F1, 100% para F2 e 93% para F3. Para F2 e F3 tivemos uma significativa melhora em relação ao caso que aplicamos o ACPF, porém a divisão das curvas de F1 em dois grupos distintos apareceu novamente.

Aplicando o funHDDC aos dados simulados, temos a Figura 11 e a matriz de confusão resultante pode ser observada na Tabela 8. Pela Tabela 8, este modelo obteve uma acurácia total de 66% e 40%, 70% e 90% para as curvas da F1, F2, e F3, respectivamente.

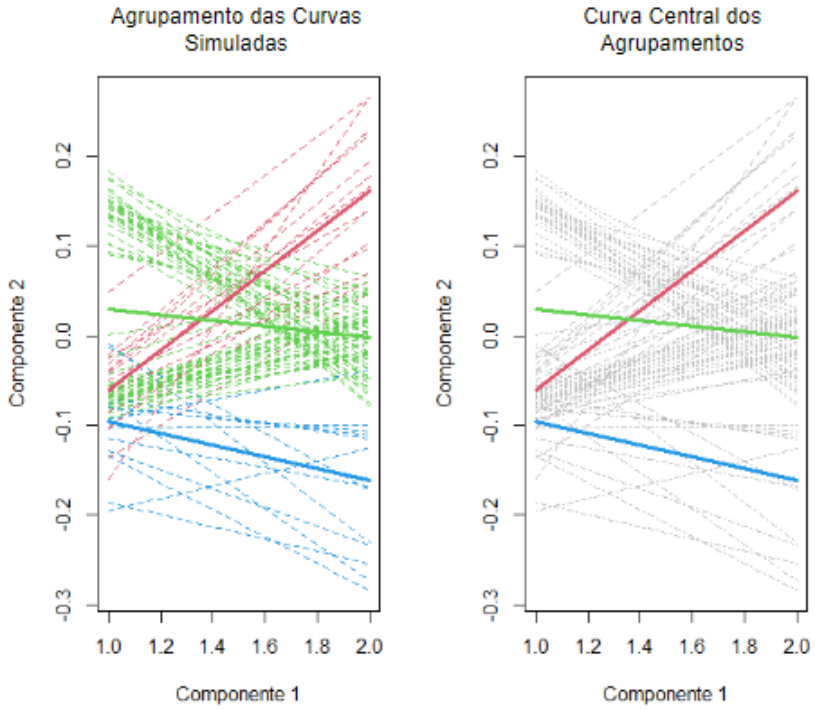


Figura 9 – Visualização do agrupamento do algoritmo K-Médias, aplicado conjuntamente com as duas primeiras componentes principais resultantes do ACPF. A cor verde representa F3, a vermelha F2 e a azul F1.

Fonte – Elaborado pelo autor.

Tabela 8 – Matriz de confusão resultante do algoritmo funHDDC, nos dados simulados de três funções.

| Matriz de Confusão | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Função 1 | Função 2 | Função 3 |
| Agrupamento do Modelo | Função 1 | 12 | 9 | 3 |
| | Função 2 | 0 | 21 | 0 |
| | Função 3 | 18 | 0 | 27 |

Fonte – Elaborada pelo autor.

### 4.2.1 Simulações com Três Classes de Funções e Repetições

Foram simulados 200 conjuntos de dados com 90 curvas, sendo 30 curvas de cada função (F1, F2 e F3). Desta forma, os modelos funHDDC, K-Médias, rede gás neural difusa e K-Médias com difusão polinomial. Podemos observar os resultados na Figura 12.

Temos que para o caso com três grupos diferentes, não ocorreu uma grande diferença de assertividade e homogeneidade dos grupos para os modelos K-Médias, K-Médias com difusão polinomial e o funHDDC. Contudo, a rede neural teve um ajuste ligeiramente melhor do que os demais modelos.

Considerando as 200 repetições que totalizando 6000 observações de cada uma das funções, foi realizado um estudo, onde foram adotados as métricas de Pontuação de Silhouette (silhouette) e Índice Calinski-Harabasz (calinhara), que visa em um cenário de incerteza sobre o

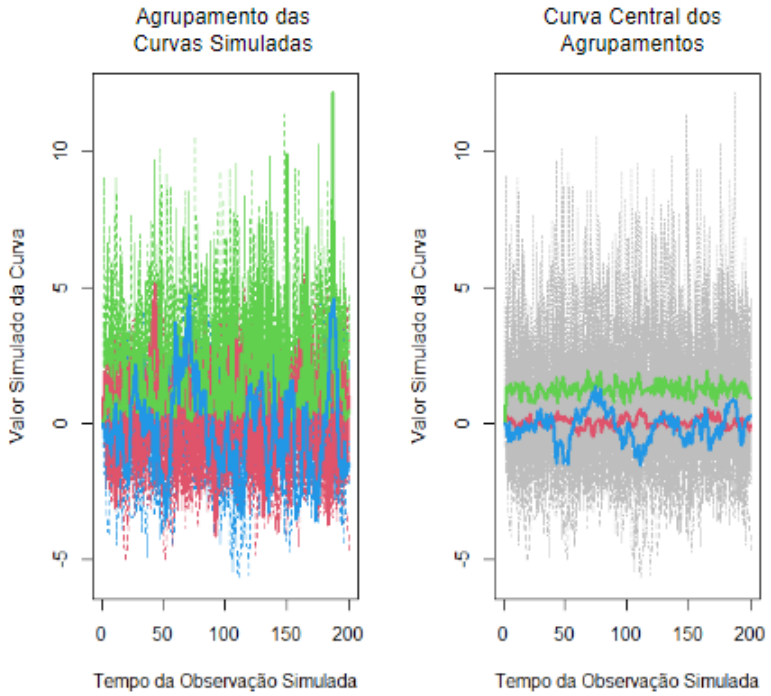


Figura 10 – Visualização do agrupamento do algoritmo K-Médias, aplicado sozinho. A cor verde representa F2, a vermelha F3 e a azul F1.

Fonte – Elaborado pelo autor.

número real de agrupamentos, realizar uma sugestão com base nos ajustes feitos pelos modelos de qual seria o número real de agrupamentos do conjunto de dados, os resultados podem ser visualizados na Figura 13.

Nota-se que na Figura 13, temos o indicativo de que devido a proximidade das curvas simuladas, as métricas apontam que o número ideal de agrupamentos seria 2, na métrica de silhouette, os valores estão próximos de 0 o que mostra uma incerteza da métrica, indicando que os agrupamentos realizados não tenham ficado bem definidos, e em calinhara, para a Rede Gás Neural e o K-Médias Difuso, que são os modelos que alcançaram as melhores pontuações na Medida V, apontam que 2 agrupamentos seria o valor ideal.

### 4.2.2 Considerações sobre Simulações com Três Classes de Funções

Para a simulação 2, o melhor desempenho foi encontrado aplicando o K-Médias sozinho. Diferentemente da simulação 1, o ACPF não conseguiu realizar uma separação razoável das curvas. Um ponto que devemos levar em conta é que o funHDDC teve uma performance muito superior neste caso mais complexo do que na simulação 1.

Esse estudo reafirma a suposição de que alguns modelos são melhores do que outros em determinadas situações, e que por mais que em uma situação de maior simplicidade um algoritmo teve uma baixa performance, ele pode ter um desempenho melhor em um contexto semelhante, só que mais complexo. Uma estratégia interessante a se pensar quando abordamos

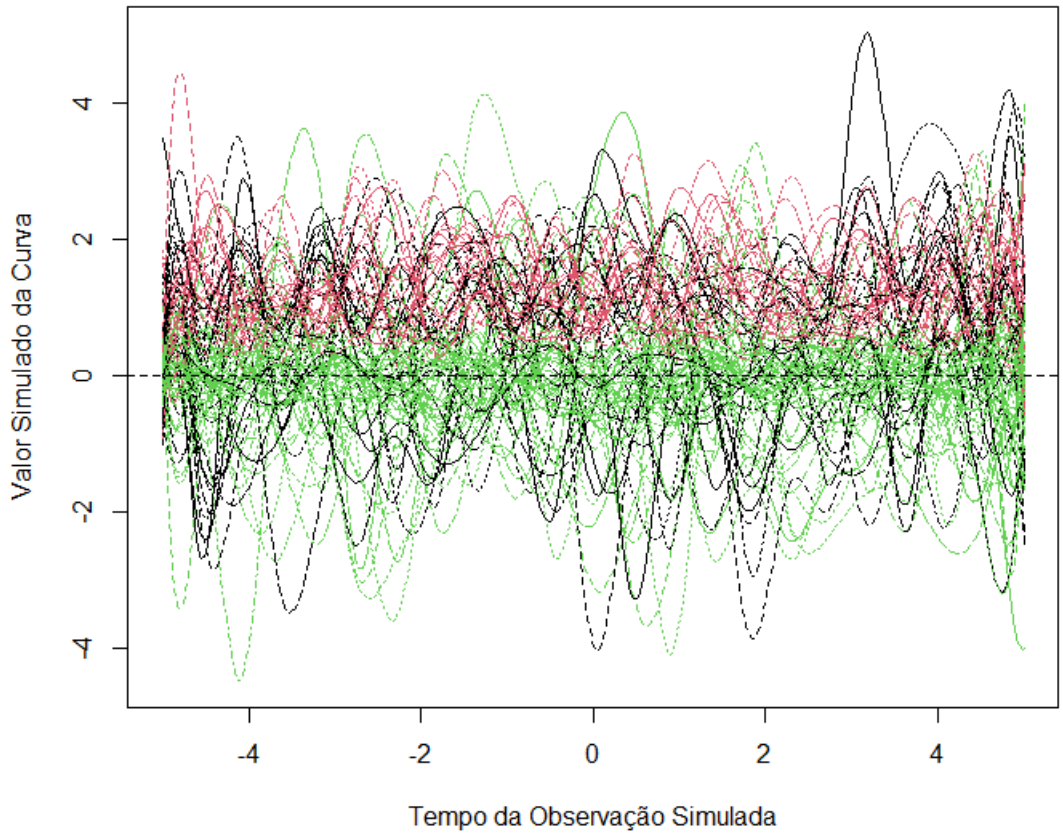


Figura 11 – Visualização do agrupamento do algoritmo funHDDC, aplicado conjuntamente com as alterações de base e suavizações. A cor preta representa F1, a vermelha F2 e verde F3.

Fonte – Elaborado pelo autor.

a modelagem de dados é entender a proposição das metodologias e aplicá-las de acordo com o contexto dos dados. Assim, podemos usufruir dos principais pontos fortes de cada um dos modelos propostos.

Diferente da simulação 1, na qual a visão difusa fez grande diferença independentemente do modelo, no caso da simulação 2 a visão difusa no K-Médias não contribuiu significativamente para a performance do modelo. Entretanto, a rede neural difusa se destacou positivamente no agrupamento.

Para este estudo, será considerado mais simulações, a fim de compreender como os algoritmos se comportam considerando uma vasta gama de variabilidade simulada.

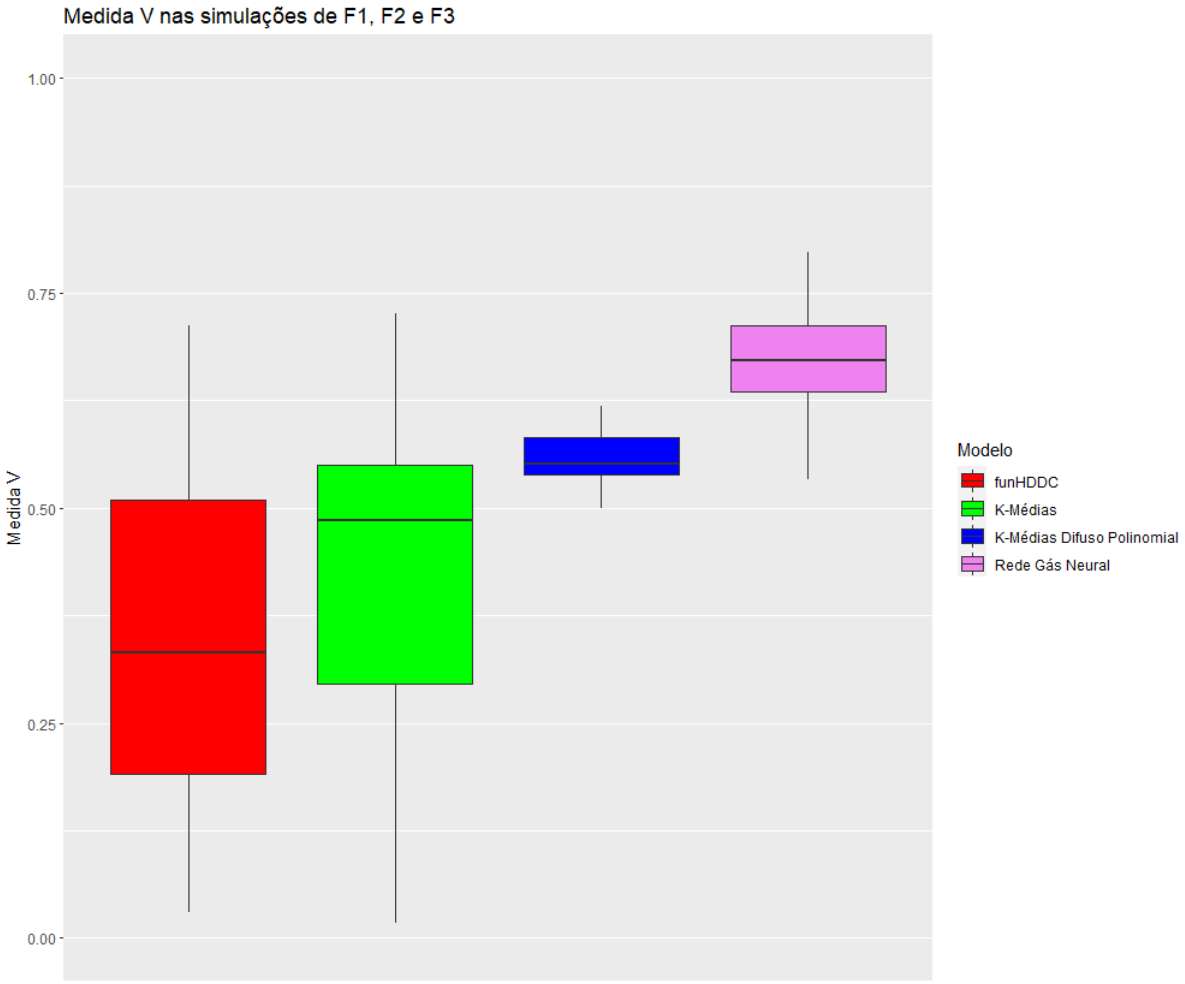


Figura 12 – Medida V para as 200 repetições das simulações das funções F1, F2 e F3.

Fonte – Elaborado pelo autor.

## 4.3 Simulação 3 : Duas Classes de Funções com Independência entre os Pontos da Curva

Para a simulação 3 foram adotadas algumas suposições como independência entre os pontos da curva e um ruído que segue uma distribuição Normal para cada curva, alterando a variância da Normal entre funções. Será denotada a função $F4_A$ e $F4_B$ como

$$F4_A(x_t) = 2 + \varepsilon_a, \text{ onde } \varepsilon_a \sim \text{ N(0,2)}, \tag{4.4}$$

$$F4_B(x_t) = cos(r) + sen(r) + \varepsilon_b, \text{ onde } \varepsilon_b \sim \text{ N(0,1)}, \tag{4.5}$$

$t = 1,2,...,200$ e $r$ um valor aleatório de $[0;2\pi]$. Para está simulação, serão amostradas 350 curva de $F4_A(x_t)$ e $F4_B(x_t)$, totalizando 700 observações com cada curva contendo 200 pontos. O

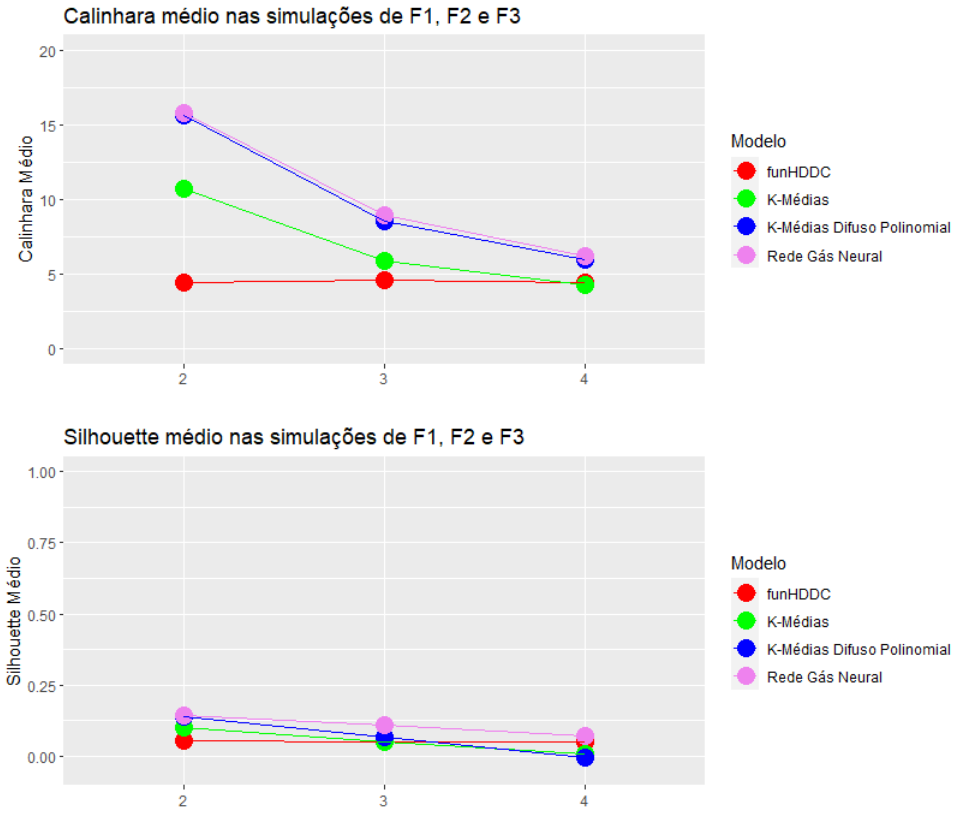


Figura 13 – Silhouette e Calinhara para as 200 repetições das simulações das funções F1, F2 e F3.

Fonte – Elaborado pelo autor.

intuito desta simulação é compreender como os modelos que supõem normalidade se comportam, diante aos demais modelos.

Para obter as curvas de F4 foi usado novamente o Algoritmo 3. A Figura 14 apresenta as duas curvas para ilustrar qual é a situação que iremos usar como base do estudo.

Pela Figura 14, nota-se a semelhança entre as curvas $F4_A(x_t)$ e $F4_B(x_t)$, no qual o destaque é a dispersão superior de $F4_A(x_t)$.

Ao aplicar o modelo de ACPF, demonstrado na Figura 15, temos uma boa separação das curvas $F4_A(x_t)$ e $F4_B(x_t)$ observando somente as duas primeiras componentes. Desta forma, será ajustado o K-Médias usando as duas primeiras componentes do ACPF (que foi também o número de componentes principais que produziu a melhor acurácia total), os resultados estão na Figura 16 e Tabela 9.

Nota-se pela Figura 16 e Tabela 9 que as metodologias conjuntas geram um viés, visto que o K-Médias considerou quase todas as curvas pertencentes a um único grupo. A assertividade total foi de 51%, porém, para $F4_A$ e $F4_B$ foi alcançado uma acurácia de 11% e 92%, respectivamente, desta forma é possível inferir que o modelo não conseguiu discernir uma curva da outra, na aplicação conjunta. Aplicando somente o K-Médias, obteve-se a Figura 17 e Tabela 10.

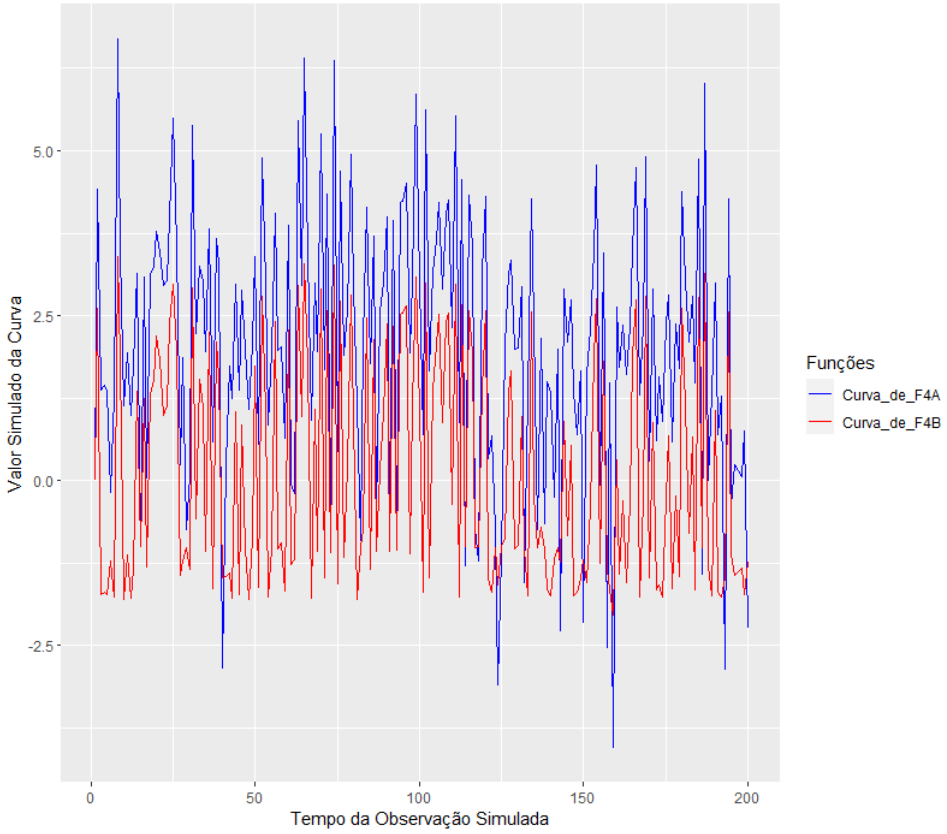


Figura 14 – Exemplificação das curvas simuladas pelas funções $F4_A$ e $F4_B$.

Fonte – Elaborado pelo autor.

Tabela 9 – Matriz de confusão resultante da metodologia aplicando ACPF e o K-Médias na sequencia, nos dados simulados de duas funções.

| Matriz de Confusão K-Médias com ACPF | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | Função 4A | Função 4B |
| Agrupamento do Modelo | Função 4A | 40 | 27 |
| | Função 4B | 310 | 323 |

Fonte – Elaborada pelo autor.

Observando Figura 17 e Tabela 10 temos que a curva média dos grupos ficou bem definida e separada, exceto pelo inicio onde ocorreu uma distorção, logo, temos uma evidencia de que ocorreu uma boa separação das curvas pelo modelo. Ao analisar a matriz de confusão, a assertividade total foi de 98% e para $F4_A$ e $F4_B$ foi alcançado uma acurácia de 97% e 99%, respectivamente, tendo assim uma assertividade total e por curva alta. Por fim, foi ajustado o funHDDC que gerou os resultados da Figura 18 e Tabela 11 .

Analisando a Tabela 11, temos que o funHDDC obteve uma assertividade total de 99% e para as curvas individualmente foi de 100% e 98%, para $F4_A$ e $F4_B$ respectivamente, são assertividades altas o suficiente para serem consideradas boas. O funHDDC obteve a melhor coerência para a classificação das curvas mostrando que o modelo conseguiu realizar o discernimento nas

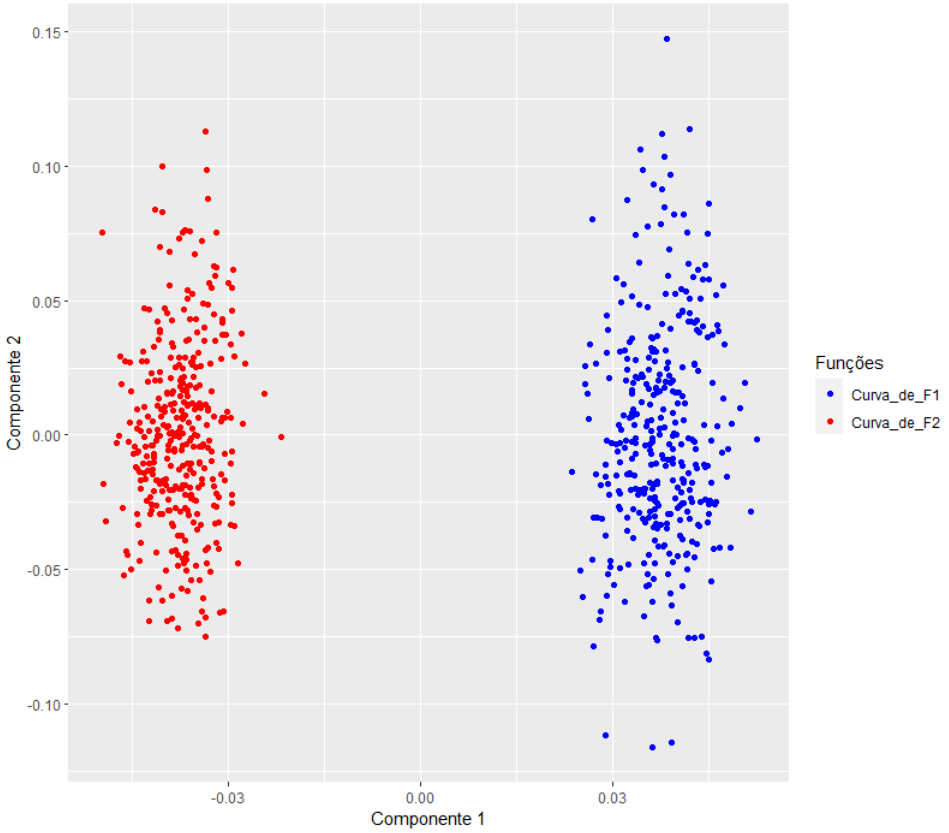


Figura 15 – Agrupamento das curvas pelo ACPF, onde visualizamos somente as duas primeiras componentes no plano de duas dimensões.

Fonte – Elaborado pelo autor.

Tabela 10 – Matriz de confusão resultante do K-Médias, nos dados simulados de duas funções.

| Matriz de Confusão K-Médias | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | Função 4A | Função 4B |
| Agrupamento do Modelo | Função 4A | 341 | 1 |
| | Função 4B | 9 | 349 |

Fonte – Elaborada pelo autor.

curvas.

Uma possível justificativa para essa situação é que o funHDDC tem certas suposições que foram atendidas nesta simulação, como por exemplo a normalidade nos ruídos das curvas, e como $F4_A$ e $F4_B$ são curvas visualmente próximas, com ruídos diferentes (porém de uma distribuição Normal), apresenta um desafio que o modelo de funHDDC tem uma ótima performance.

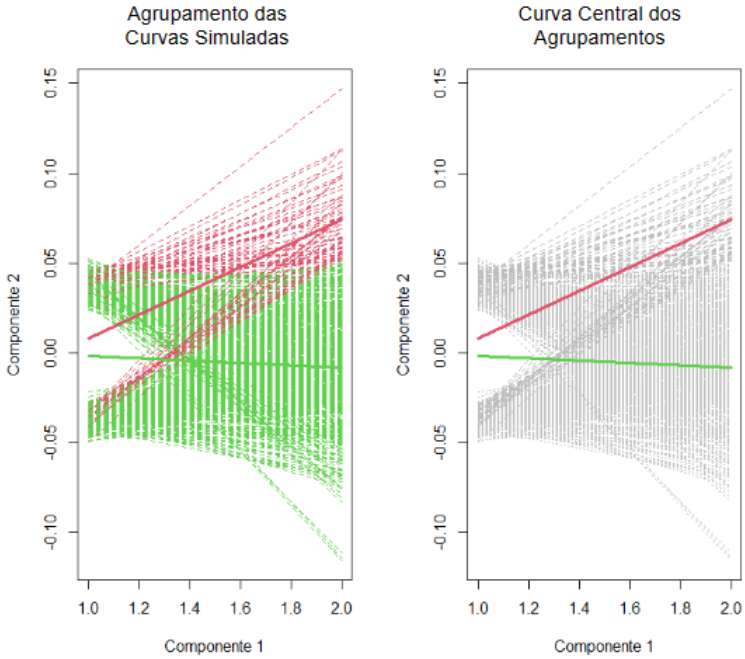


Figura 16 – Visualização do agrupamento do algoritmo K-Médias, aplicado conjuntamente com as 2 primeiras componentes principais resultantes do ACPF. A vermelha $F4_A(x_t)$ e a verde $F4_B(x_t)$

Fonte – Elaborado pelo autor.

Tabela 11 – Matriz de confusão resultante do funHDDC, nos dados simulados de duas funções.

| Matriz de Confusão funHDDC | | | |
|---|---|---|---|
| | | Agrupamento Real | |
| | | Função 4A | Função 4B |
| Agrupamento do Modelo | Função 4A | 350 | 4 |
| | Função 4B | 0 | 346 |

Fonte – Elaborada pelo autor.

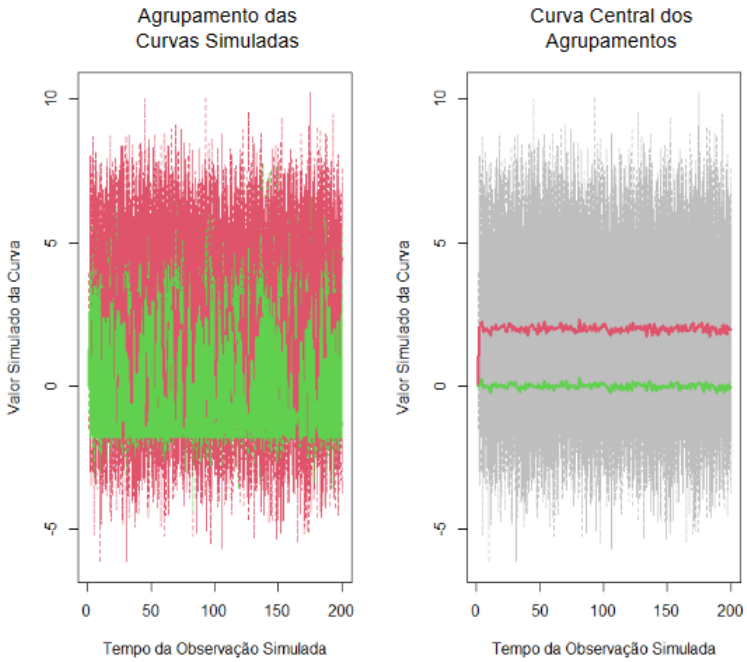


Figura 17 – Visualização do agrupamento do algoritmo K-Médias, aplicado sozinho. A vermelha $F4_A(x_t)$ e a verde $F4_B(x_t)$.

Fonte – Elaborado pelo autor.

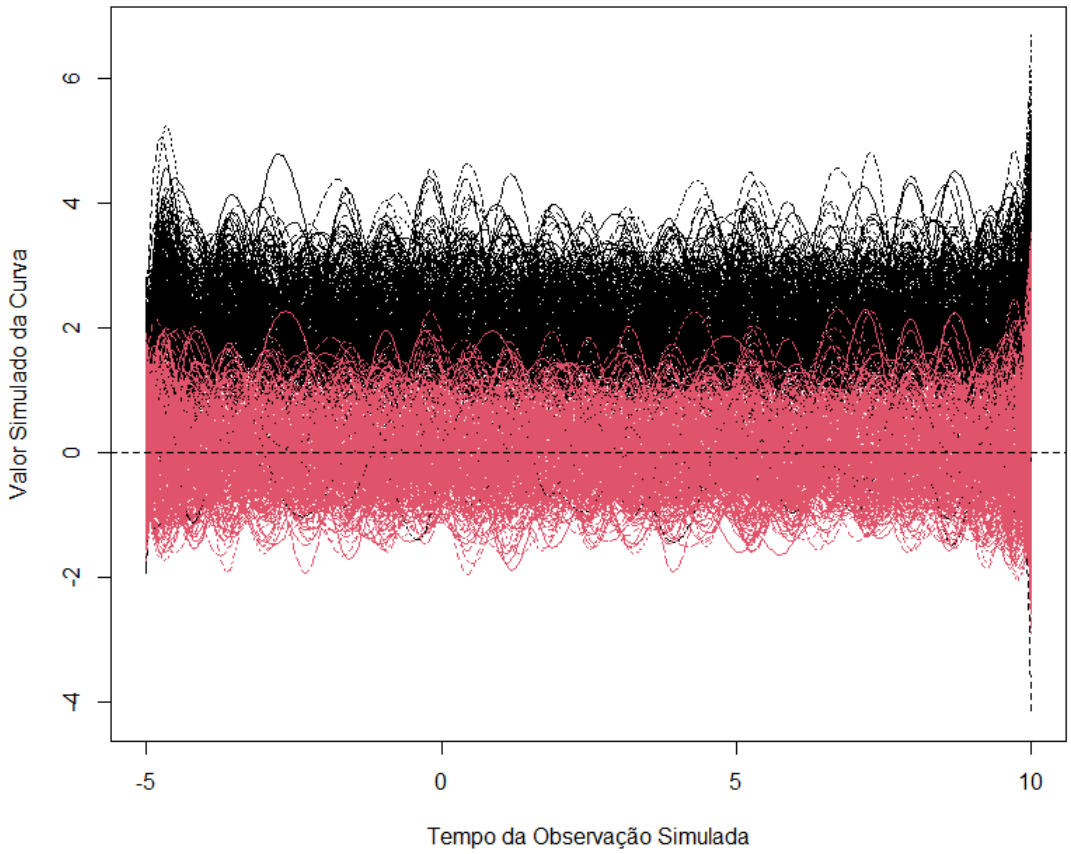


Figura 18 – Visualização do agrupamento do funHDDC. A preta representa as curvas da função $F4_A(x_t)$ e a vermelha da função $F4_B(x_t)$.

Fonte – Elaborado pelo autor.

CAPÍTULO

# 5

# APLICAÇÃO

Após avaliar como os métodos K-Médias, ACPF, funHDDC e rede gás neural se comportam em um ambiente controlado e bem definido através das simulações, é de interesse deste trabalho entender o desempenho desses métodos quando aplicados em dados reais.

Quando tratamos de dados funcionais, temos que eles podem estar definidos tanto no tempo quanto em escala. Na aplicação será tratada a abordagem em escala, que é onde os espectros são definidos, com o intuito de verificar se o agrupamento não supervisionado é capaz de separar de maneira minimamente razoável os grupos distintos presentes em cada um dos conjuntos de dados.

Temos as análises feitas com dados astronômicos, que são apresentados em escalas de ondas. Esse conjunto de dados apresenta leituras espaciais provindas do observatório de Apache, nele encontramos leituras funcionais de galáxias, quasares e estrelas.

As galáxias são definidas como objetos astronômicos, que pode conter corpos celestes, como estrelas e quasares. Devida a está condição, ocorre uma semelhança teórica entre galáxias e os corpos celestes, pois temos conjuntos de agrupamentos podendo estar contidos em outro conjunto. Para simplificar as definições e resultados, neste trabalho os três objetos observados (galáxias, quasares e estrelas) serão denominados como corpos celestes.

Será também analisado os dados de minérios, em que temos os espectros derivados da espectroscopia de Raman. Temos um conjunto de dados em que foram observados minérios de diamante, tremolite e quartzo.

Uma consideração que deve ser feita é a respeito da composição química dos minérios. O diamante é formato basicamente por carbonos, entre tanto, tremolite e quartzo contém tetraedro de silício. O quartzo é composto por tetraedro de silício e oxigênio, já a tremolite contém além do tetraedro de silício outras componentes químicas em sua estrutura.

Nessas condições, o diamante é afastado quimicamente da tremolite e quartzo. E a

tremolite e quartzo tem uma componente química em comum.

Usaremos novamente as métricas de acurácia entre grupos e a total, além da medida V para avaliar se o agrupamento proposto é satisfatório em comparação com as verdadeiras classes, conjuntamente com o auxílio gráfico e da matriz de confusão.

## 5.1 Dados Astronômicos

Os dados astronômicos que iremos estudar são provindos da SSDS, que é um grupo fundado pela Universidade de Chicago em 2000. O observatório está localizado em Apache no Novo México.

Em Astronomia, ao observar um objeto no espaço, é normal não ter 100% de clareza do que está sendo observado. Em Sasdelli *et al.* (2016), com dados de supernovas, os pesquisadores afirmam que não existe uma definição certa das diferentes classes (caso elas existam) de supernovas, mas é de interesse dentro deste contexto, a partir de modelos não supervisionados, identificar e realizar possíveis agrupamentos dos dados para até mesmo provar as teorias que são escritas mas ainda não comprovadas.

Uma características das curvas astronômicas é que elas são definidas na escala fluxo, ou seja, matérias diferentes emitem luzes diferentes, que são visíveis quando fazemos a observação do objeto astronômico em uma determinada faixa do fluxo (Bradt (2004)). Sendo assim, os pesquisadores conseguem inferir a situação do corpo celeste a partir do quão intensas são as leituras feitas em uma determinada escala de onda. Essa condição é ilustrada na Figura 19.

No estudo que será apresentado, temos uma base de dados contendo curvas espectroscópicas de estrelas, galáxias e quasares (qso). Um quasar pode ser definido com um buraco negro que emite uma forte luz muito semelhante a estrelas, mas em sua existência tende a consumir toda à matéria a sua volta, produzindo uma forte luz e uma grande emissão de sinais de rádio, condição que permitiu a sua descoberta no espaço.

Usaremos 10 observações de cada corpo celeste observados em 2021 e 2022, totalizando 30 observações. Tentaremos realizar o agrupamento não supervisionado de cada leitura. Os corpos amostrados se encontram na Tabela 18, que contém seus códigos de identificação e podem ser consultados no Server (2022).

Para exemplificar o que as curvas na Figura 19 representam, a Figura 21 mostra como é a imagem observada de uma estrela, galáxia e quasar. Podemos notar a grande semelhança entre a imagem de uma estrela e um quasar, além da semelhança teórica entre galaxia e estrelas.

Na Figura 20, notamos que todos os grupos contêm leituras em todos os comprimentos de onda, mas na pratica isso não ocorre, pois os corpos celestes estão definidos em intervalos de comprimentos de ondas distintos. Para uniformizar a base de dados, nos comprimentos em que o corpo celeste não estava definido, foi atribuído o valor 0.

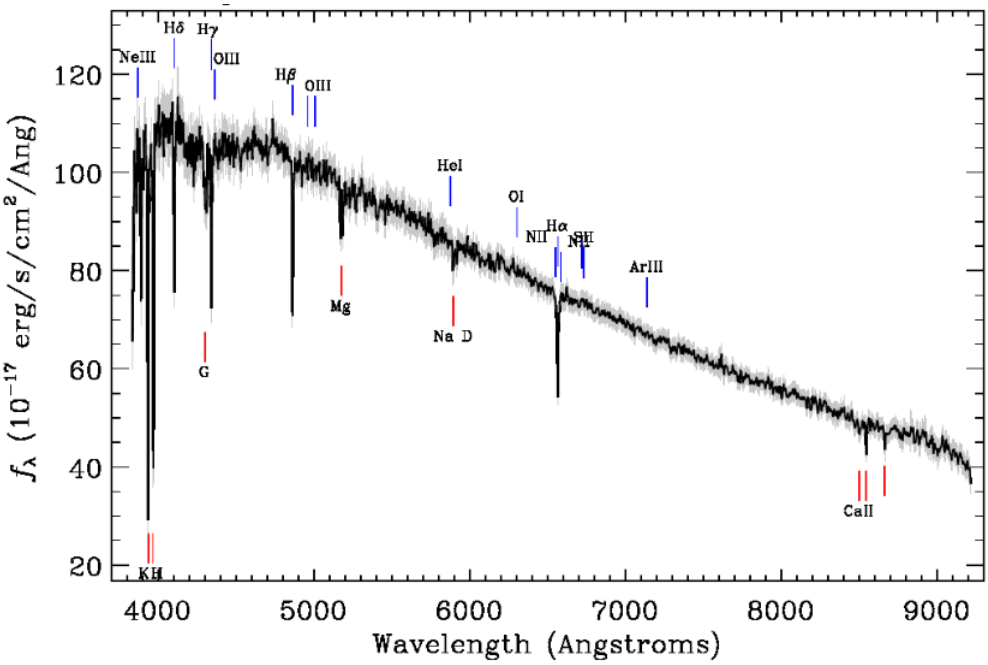


Figura 19 – Curva obtida da estrela 2276569950720124928 observada em 2022.

Fonte – Essa figura foi elaborado por Server (2022) e está disponível publicamente.

Para essa aplicação, que apresenta uma alta complexidade por apresentar curvas oscilatórias, pontos *outliers*, curvas definidas em diferentes escalas e diferentes corpos celestes apresentam leituras muito próximas, foram feitos alguns pré-processamentos nos dados. O primeiro foi definir todas as leituras em uma única faixa de escala. Sendo assim, foi observado qual leitura apresentava a maior faixa na escala de comprimento de onda. Desta forma, esta leitura com a maior faixa foi definida para ser a base de todas as leituras. Para as curvas que não apresentavam leituras em um determinado ponto da escala foi atribuído valor 0 naquele ponto, de modo que, ao final todas as curvas tivessem informações em todos os comprimentos de ondas. Devido à alta oscilação das curvas e presença de *outliers*, as curvas foram transformadas usando a definição de padronização dos dados.

Os algoritmos de K-Médias e funHDDC não convergiram para um resultado e o ACPF não foi capaz de realizar uma separação satisfatória, como podemos observar na Figura 22.

Como os métodos estudados anteriormente não convergiram quando aplicados os dados astronômicos, será ajustado somente a rede neural baseada em uma regra de adaptação *soft-max* e o K-Medoides juntamente com a abordagem de agrupamento difuso. O estudo para qual seria o número ideal de grupos segundo os modelos propostos, em Sasdelli *et al.* (2016) é feita de acordo com uma análise das curvas médias presentes em cada agrupamento. Na literatura são encontradas métricas como o índice de Davies-Bouldin (Davies e Bouldin (1979)) ou o coeficiente Silhouette (Rousseeuw (1987b)) para avaliar os resultados de modelos que realizam agrupamentos não supervisionados, para os quais é analisada a qualidade dos agrupamentos feitos pela similaridade das observações contidas dentro de cada agrupamento e a heterogeneidade

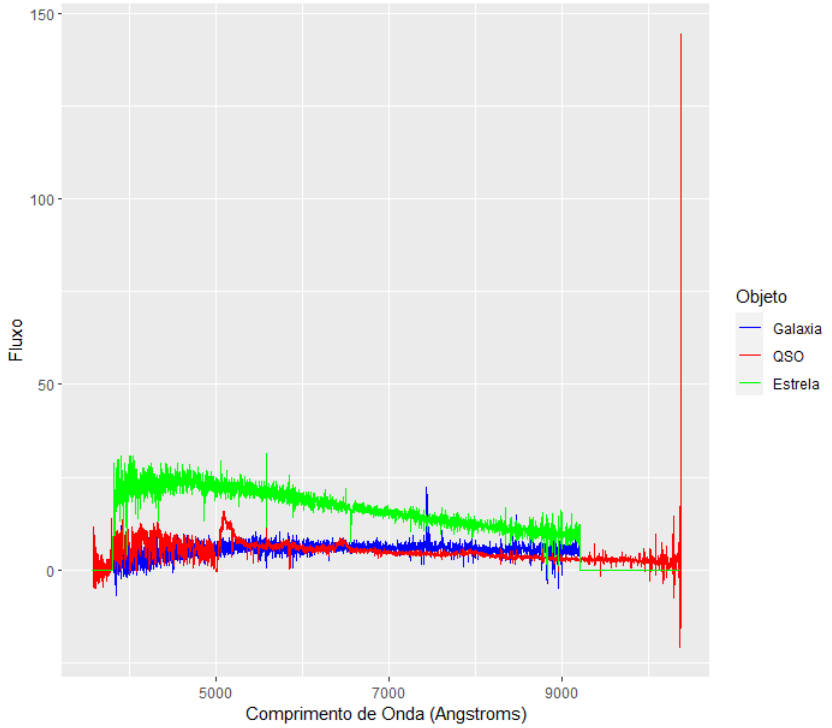


Figura 20 – Curvas dos três grupos presentes no conjunto de dados.

Fonte – Elaborada pelo autor.

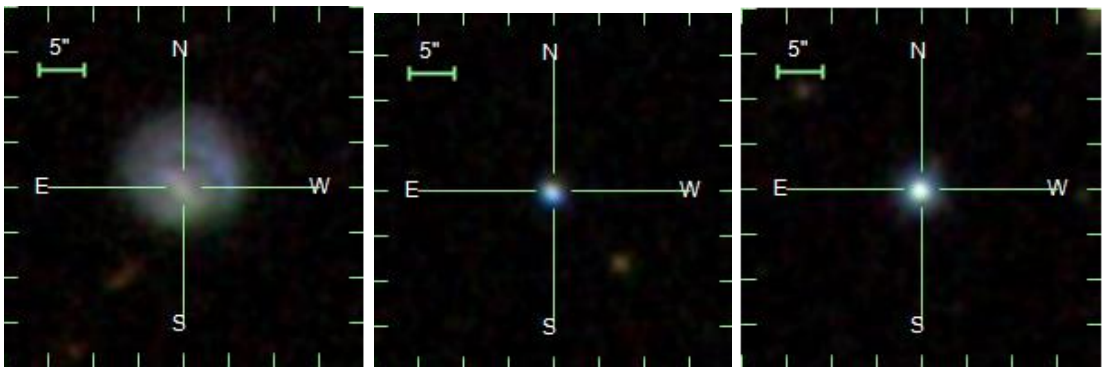


Figura 21 – Imagens de uma galaxia, qso e estrela, respectivamente.

Fonte – Essa figura foi elaborado por Server (2022) e está disponível publicamente.

entre grupos distintos. Em Hu e Xu (2004) e Zhao, Hautamaki e Fränti (2008) são comparadas e exemplificadas formas de analisar o número ideal de agrupamentos via critério de informação de Akaike e critério de informação Bayesiana.

Serão considerados três grupos (número correto de grupos distintos) para analisar como os modelos agrupariam as curvas, o resultado pode ser observado na Tabela 12, nota-se que o K-Medoides apresentou um desempenho melhor. Os resultados da rede neural podem ser observados na Tabela 13.

Em uma visão geral, para o modelo de redes neurais tivemos uma acurácia total de

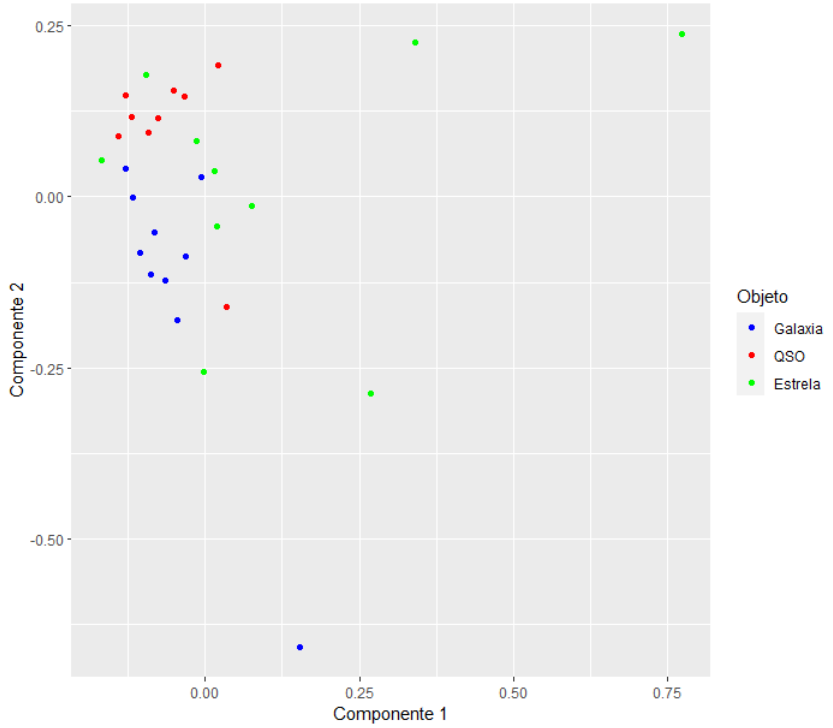


Figura 22 – Resultado do modelo ACPF.

Fonte – Elaborada pelo autor.

Tabela 12 – Resultados da medida V.

| Tabela Medida V | | | |
|---|---|---|---|
| | | Modelo | |
| | | Rede Gás Neural | K-Medoides |
| Número de Grupos | 3 | 0,567 | 0,659 |

Fonte – Elaborada pelo autor.

60%, o que não é um resultado considerado bom. O modelo encontrou dificuldades na distinção entre galáxias e estrelas. Todavia, ocorreu uma assertividade de 80% para os quasares que frenquentemente são confundidos com estrelas.

Analisando os resultados obtidos com a rede neural, entende-se que estamos em um caso em que temos além da semelhança entre algumas curvas dos corpos celestes, uma forte presença de ruídos, e esta união de características é um caso em que modelos difusos com tratamento de ruídos têm uma boa performance.

Foi construído um K-Medoides difuso com agrupamento de ruídos. Foi escolhido via otimização que o melhor valor para o parâmetro de difusão seria 3, o número inicial usado para otimizar este hiperparâmetro foi 5, o coeficiente de ponderação para o índice de silhueta difusa foi fixado em 1. O resultado pode ser observado na Tabela 14 abordagem de difusão com tratamento de ruídos obteve uma acurácia total de 80%, e foi obtida uma assertividade de 90%,

Tabela 13 – Matriz de confusão resultante da rede neural.

| Matriz de Confusão Rede Gás Neural | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Galáxia | QSO | Estrela |
| Agrupamento do Modelo | Galáxia | 2 | 1 | 2 |
| | QSO | 2 | 8 | 0 |
| | Estrela | 6 | 1 | 8 |

Fonte – Elaborada pelo autor.

Tabela 14 – Matriz de confusão resultante da metodologia K-Medoides difuso com detecção de ruídos em agrupamentos.

| Matriz de Confusão: K-Medoides | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Galáxia | QSO | Estrela |
| Agrupamento do Modelo | Galáxia | 9 | 1 | 0 |
| | QSO | 0 | 7 | 2 |
| | Estrela | 1 | 2 | 8 |

Fonte – Elaborada pelo autor.

70% e 80%, para o grupo de galáxia, qso e estrelas, respectivamente.

Apesar de uma precisão de 70% no grupo de quasares, as leituras agrupadas erradas apresentavam visualmente características diferentes do resto dos quasares, o que justifica as três leituras agrupadas erradas. O K-Medoides difuso com detecção de ruídos teve um resultado satisfatório, principalmente pelo fato de reconhecer a presença de três agrupamentos distintos, diferentemente da rede neural aplicada que só realizou uma boa distinção dos quasares e das demais metodologias, que não obtiveram convergência.

Como os dados astronômicos apresentaram um grande desafio e incerteza de muitas leituras ou reconhecimento da existência de alguns corpos celestes, uma assertividade total de 80% e superior a 70% nos agrupamentos individuais, deixa evidente a presença de pelo menos três grupos com leituras funcionais distintas, podendo até mesmo ocorrer a presença de um quarto grupo derivado de uma subclasse de quasar na amostra estudada.

Para os dados astronômicos, os pesquisadores da área sempre tem um certo grau de incerteza quando ao valor verdadeiro de agrupamentos presentes em uma amostra de espectros observada, sendo assim, para os dados coletados neste estudo, foi feito um estudo usando as métricas de Pontuação de Silhouette (silhouette) e Índice Calinski-Harabasz (calinhara), para os modelos de K-Medoides e Rede Gás Neural obteve-se a Figura 23.

Observando a Figura 23, temos que segundo as métricas o número ideal de agrupamentos seria 2, ocorrendo uma concordância entre as duas métricas na visão média. Tem-se que as curvas de galáxias e estrelas tem certa proximidade nas leituras e o modelo de Rede Neural teve

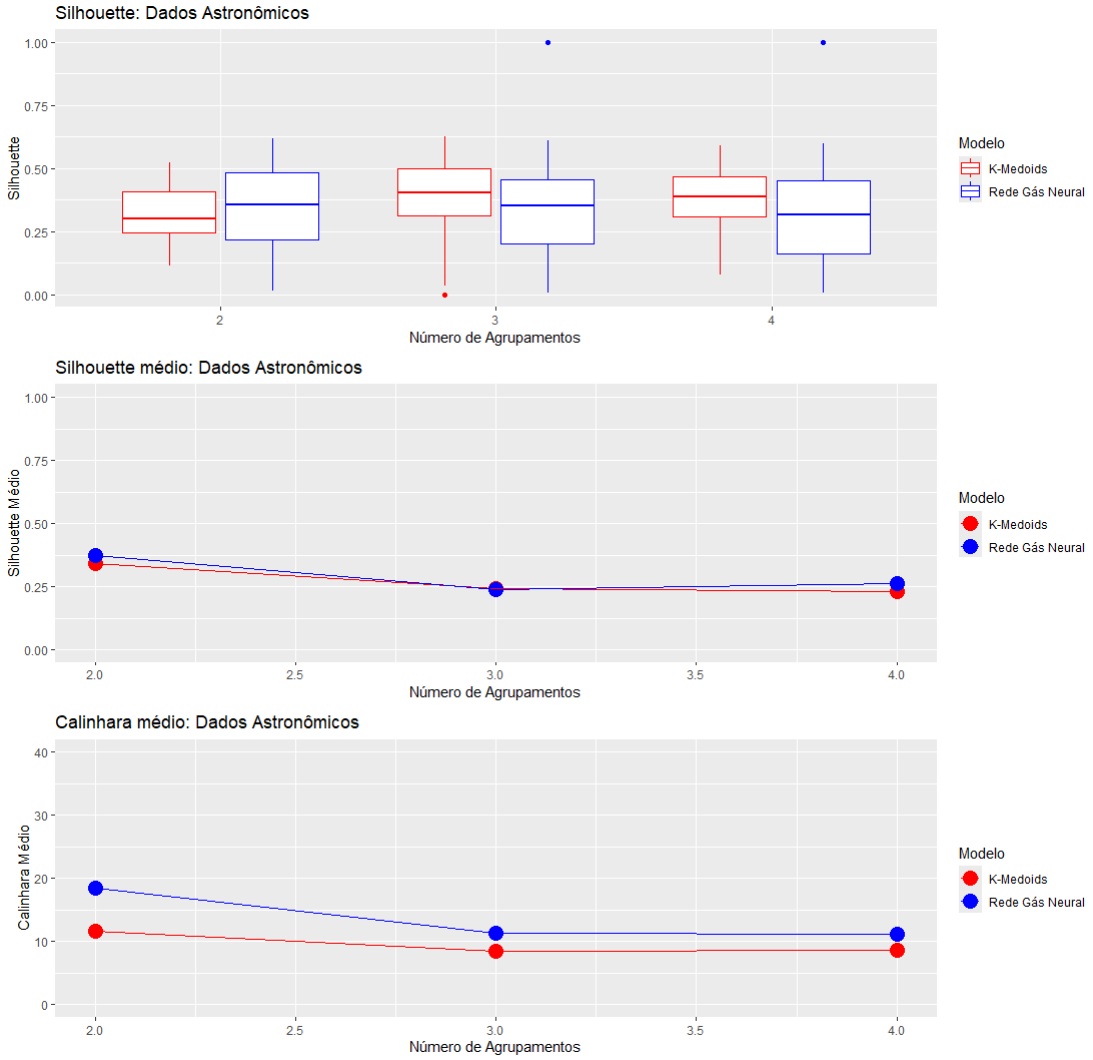


Figura 23 – Silhouette e Calinhara para dados astronômicos.

Fonte – Elaborado pelo autor.

uma dificuldade de separar esses grupos e o K-Medoides confundiu algumas curvas de estrelas com QSO e galáxia, todavia para o K-Medoides não ocorreu uma grande disparidade entre as pontuações, demostrando uma certa incerteza quando ao número ideal de grupos.

Parte desta incerteza pode ser justificada pelo fato das galáxias conterem majoritariamente estrelas, e o QSO apresentar tanto em teoria quanto em leitura funcional, uma diferença dos outros dois.

## 5.2 Dados de Minérios

Os dados de minérios que iremos analisar foram amostrados usando a espectroscopia de Raman aplicada à identificação mineral e é uma iniciativa do projeto RRUFF, proposto pelo Dr. Robert Downs, enquanto estava trabalhando em um espectrômetro para ser acoplado ao veiculo

rover Mars, a pedido da Administração Nacional da Aeronáutica e Espaço dos Estados Unidos (NASA).

O RRUFF é sediado pela Universidade do Arizona. O principal intuito desta iniciativa é a criação de um conjunto de dados espectrais de alta qualidade a partir de minerais bem caracterizados. É desejado o compartilhamento de informações com todo o mundo.

Podemos observar a curva de quartzo na Figura 24. Notamos nela algumas características, dentre elas, a alta dimensionalidade de uma única curva (cada curva contém entre 1100 e 1400 leituras) e a presença de alguns *outliers*. Vale destacar que todas as curvas de minérios são dados funcionais definidos na escala Raman Shift ($cm^{-1}$).

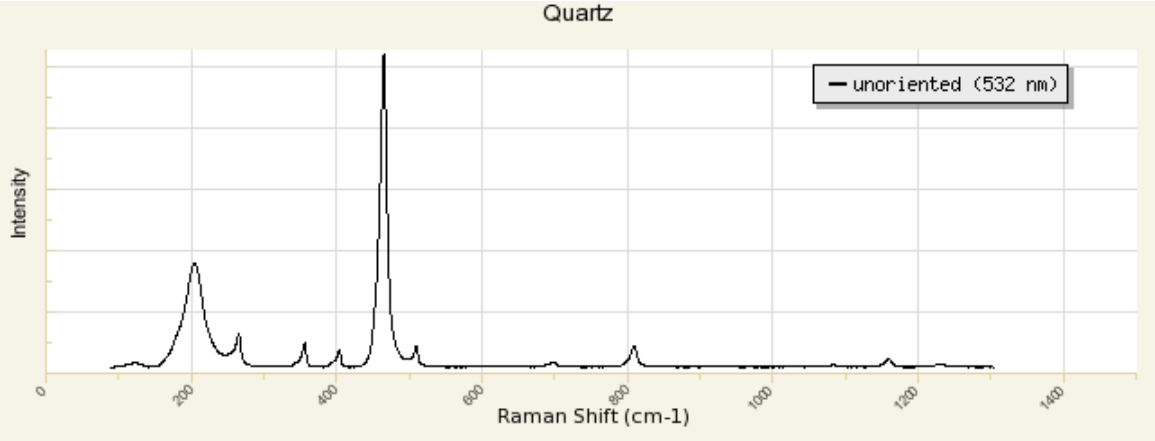


Figura 24 – Imagens da curva de uma minério de quartzo.

Fonte – Essa figura foi elaborado por Mineral (2022) e está disponível publicamente.

Os dados de minérios são observados definindo um marcador fiducial ou fiduciário. Essa definição é a resolução que o objeto é observado, para todas as observações foi definido o valor de 532 nm de marcador. Assim, teremos resolução igualitária para todas as leituras. Outro ponto é que temos valores de escalas diferentes entre os minérios de diamante, quartzo e tremolite. A Figura 25 ilustra cada um dos minérios mencionados.


Figura 25 – Imagens do minério de diamante, tremolite e quartzo, respectivamente.

Fonte – Essa figura foi elaborado por Mineral (2022) e está disponível publicamente.

Foi aplicada a mesma faixa de escala para todas as observações. Caso não tenha leitura em um certo ponto da escala, foi fixado o valor 0 para esta ocasião. Pode-se visualizar as leituras e uma comparação das curvas de cada tipo de minério na Figura 26.

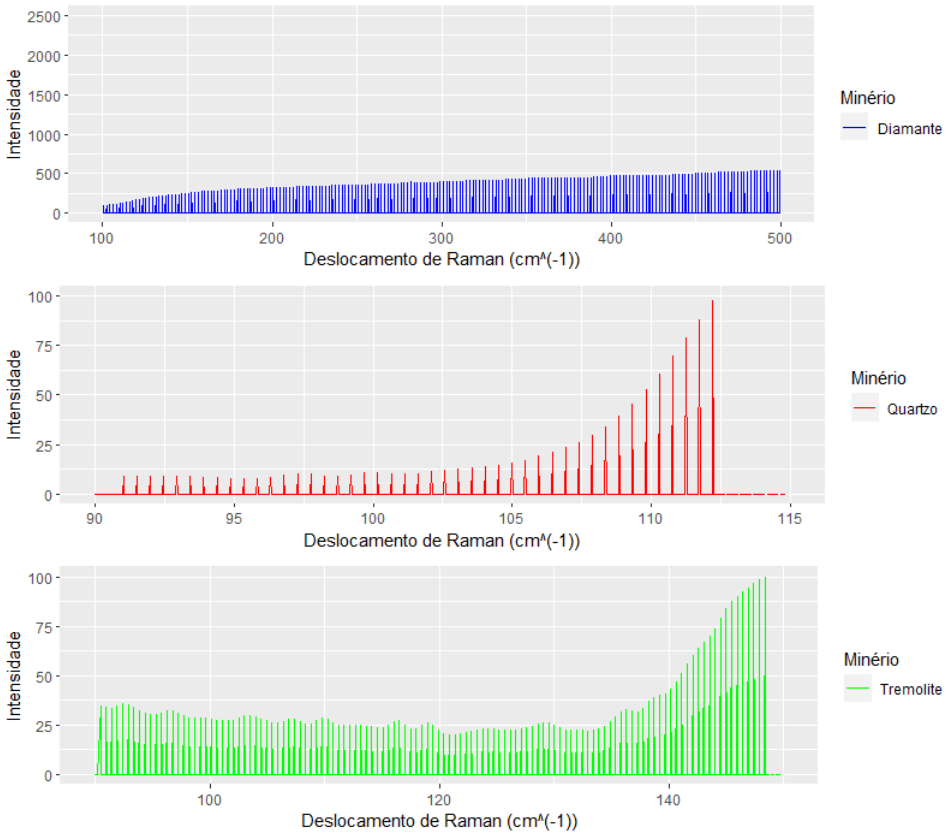


Figura 26 – Curvas dos minérios de diamante, quartzo e tremolite.

Fonte – Elaborada pelo autor.

No total temos oito espectros de diamante e sete espectros de quartzo e de tremolite, que podem ser consultados em Mineral (2022) e seus códigos de identificação estão presentes na Tabela 19.

Para a análise foi estudado se uma transformação de variáveis seria benéfica às análises. Sendo assim, os dados foram padronizados de forma que todos os espectros estudados apresentassem após a transformação, média igual a 0 e desvio padrão igual a 1.

Os únicos modelos que foram capazes de convergir com este conjunto de dados foram redes gás neural, K-Medoides difuso com detecção de ruídos em agrupamentos e o ACPF. Contudo, o resultado do ACPF neste caso não foi satisfatório. Podemos observar o resultado do modelo pela Figura 27. Nota-se que os pontos que representam os minérios de tremolite e quartzo ficaram sobrepostos quase totalmente, além dos pontos do grupo diamante estarem próximos dos demais. Concluímos assim que o ACPF não fez um bom ajuste ao conjunto de dados.

Mesmo conhecendo os grupos e as classes de cada observação, será analisado como seria a seleção do número de agrupamentos ideal segundo os métodos de redes e K-Medoides. Quando se analisa o número ideal de agrupamentos, pode-se usar métricas como o índice de Davies-Bouldin (Davies e Bouldin (1979)) ou o coeficiente Silhouette (Rousseeuw (1987b)), nos

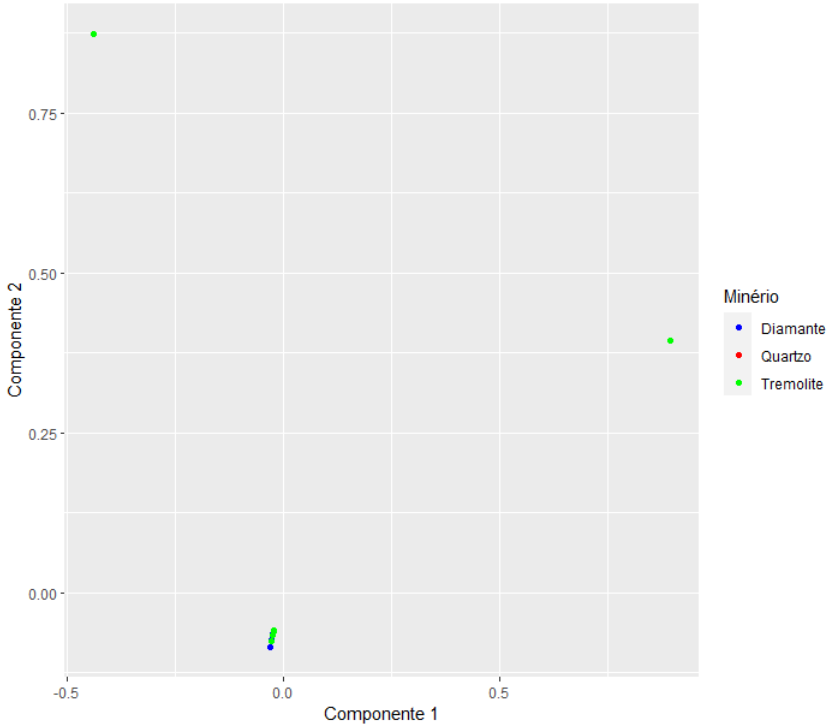


Figura 27 – Resultado do modelo ACPF no conjunto de minérios.

Fonte – Elaborada pelo autor.

Tabela 15 – Resultados da medida V.

| Tabela Medida V | | | |
|---|---|---|---|
| | | Modelo | |
| | | Rede Gás Neural | K-Medoides |
| Número de Grupos | 3 | 0,863 | 0,554 |

Fonte – Elaborada pelo autor.

trabalhos desenvolvidos por Zhao, Hautamaki e Fränti (2008) e Hu e Xu (2004) é analisados o número ideal de grupos via Critério de informação de Akaike e Critério de Informação Bayesiana. Contudo, tem-se conhecimento dos rótulos originais dos grupos, assim usaremos a medida V ajustada para 3 grupos (número correto de grupos) avaliando qualidade dos ajustes realizados pelos métodos, os resultados estão na Tabela 15.

Notamos que segundo a medida V, a rede gás neural fornece a maior precisão e homogeneidade dos grupos quando fixamos o número ideal de agrupamentos como 3, com um coeficiente de medida V de $0,309$ maior que o K-Medoides. Será analisado o desempenho individual dos modelos observando a Tabela 16 e Tabela 17.

A acurácia total foi de 81% e 50%, para os modelos de rede neural e K-Medoides, respectivamente. Desta forma, nota-se que o K-Medoides encontrou dificuldades para agrupar as curvas de tremolite, que foram confundidas com as curvas de quartzo.

Tabela 16 – Matriz de confusão resultante da metodologia K-Medoides difuso com detecção de ruídos em agrupamentos.

| Matriz de Confusão: K-Medoides | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Quartzo | Diamante | Tremolite |
| Agrupamento do Modelo | Quartzo | 5 | 3 | 5 |
| | Diamante | 0 | 5 | 0 |
| | Tremolite | 2 | 0 | 2 |

Fonte – Elaborada pelo autor.

Tabela 17 – Matriz de confusão resultante da metodologia rede gás neural.

| Matriz de Confusão: Rede Gás Neural | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Quartzo | Tremolite | Diamante |
| Agrupamento do Modelo | Quartzo | 6 | 3 | 0 |
| | Tremolite | 1 | 4 | 0 |
| | Diamante | 0 | 0 | 8 |

Fonte – Elaborada pelo autor.

Todavia, o modelo de rede gás neural teve uma precisão perfeita para agrupar as curvas de diamante, e analogamente encontrou dificuldades para agrupar corretamente as tremolites, sua assertividade foi de 57%. Contudo, segundo a medida V na Tabela 15, teve um resultado de $0,86$, podendo ser considerado um ajuste razoável aos dados.

Para os dados de minérios também foi proposto um estudo do número ideal de agrupamentos, considerando desconhecer o valor real de agrupamento, podendo ser visualizado na Figura 28

Segundo a Figura 28, temos que a métrica de Calinhara para o modelo de K-Medoides acusou o valor correto de agrupamentos, porem, para a Rede Neural, devido a confusão na classificação de Tremolite e Quartzo (Tabela 17) o modelo apontou sendo 2 o valor ideal de agrupamentos. Para a métrica de Silhouette, obteve-se valores muito próximos para as três suposições, e todos próximos a 0 demonstrando uma incerteza quando ao valor ideal.

Pode-se justificar que neste caso que devido a tremolite e quartzo apresentarem um elemento em comum, apesar da rede neural apresentar maior distinção em seus grupos, apresenta uma incerteza quando ao número real de agrupamentos.

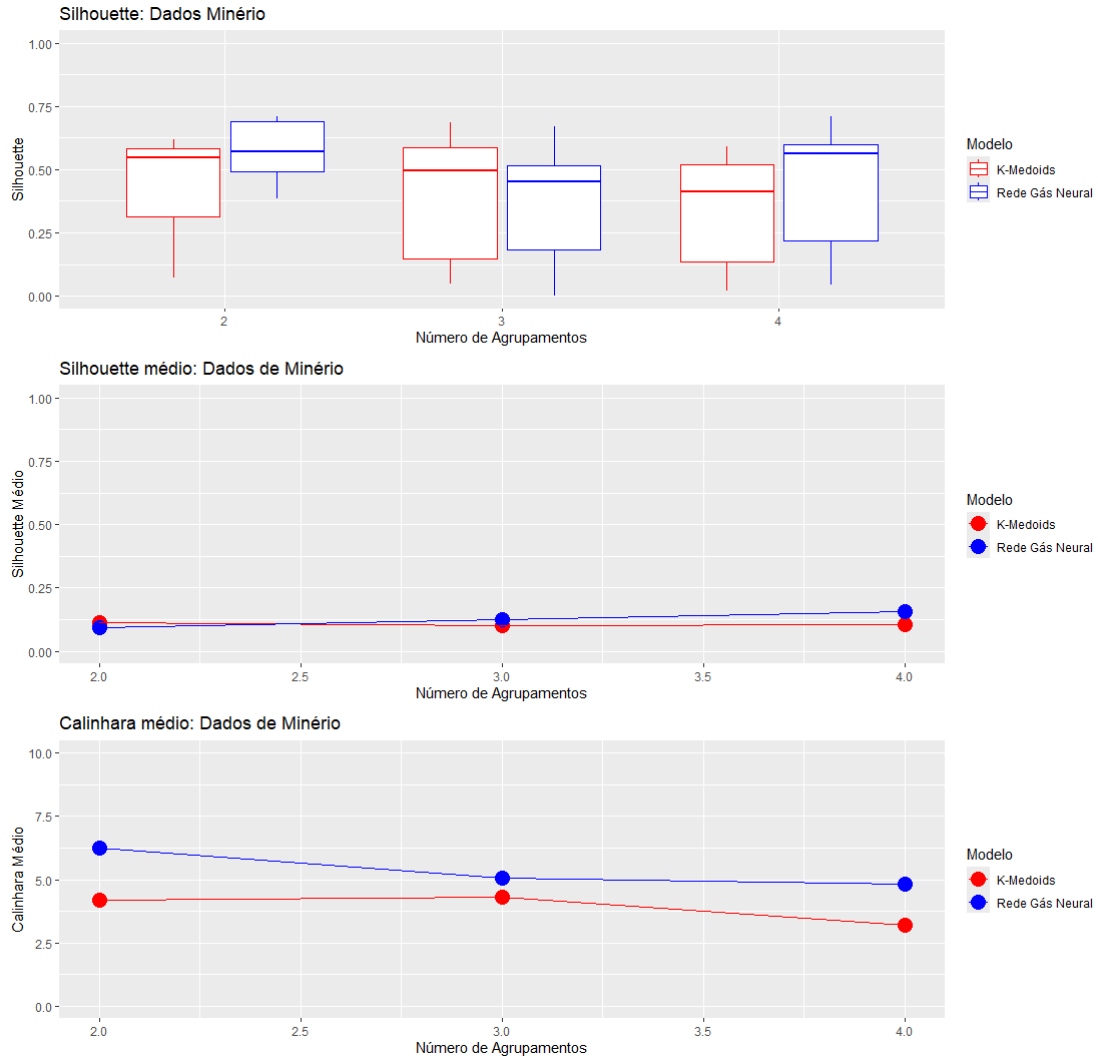


Figura 28 – Silhouette e Calinhara para os dados de minérios.

Fonte – Elaborado pelo autor.

Tabela 18 – Códigos dos corpos celeste amostrados no estudo, encontrados em Server (2022).

| Corpos Celestes Amostrados |
|---|
| Galáxia: 2275444044907702272 |
| Galáxia: 2275444319785609216 |
| Galáxia: 2275444594663516160 |
| Galáxia: 2275444869541423104 |
| Galáxia: 2275445144419330048 |
| Galáxia: 2276570225598031872 |
| Galáxia: 2276570500475938816 |
| Galáxia: 2276571325109659648 |
| Galáxia: 2276571599987566592 |
| Galáxia: 2276572424621287424 |
| QSO: 2275446243930957824 |
| QSO: 2275446518808864768 |
| QSO: 2275447068564678656 |
| QSO: 2275451191733282816 |
| QSO: 2275454490268166144 |
| QSO: 2276571874865473536 |
| QSO: 2276573249255008256 |
| QSO: 2276574898522449920 |
| QSO: 2276577647301519360 |
| QSO: 2276579021691054080 |
| Estrela: 2275447893198399488 |
| Estrela: 2275452566122817536 |
| Estrela: 2275458338558863360 |
| Estrela: 2275459712948398080 |
| Estrela: 2275460262704211968 |
| Estrela: 2276569950720124928 |
| Estrela: 2276570775353845760 |
| Estrela: 2276575173400356864 |
| Estrela: 2276577097545705472 |
| Estrela: 2276578197057333248 |

Fonte – Elaborada pelo autor.

Tabela 19 – Códigos dos minérios amostrados no estudo, encontrados em Mineral (2022).

| Minérios Amostrados |
|---|
| Diamante: R150086 |
| Diamante: R150087 |
| Diamante: R150088 |
| Diamante: R150089 |
| Diamante: R150105 |
| Diamante: R150106 |
| Diamante: R150107 |
| Diamante: R150108 |
| Quartzo: R100134 |
| Quartzo: R110104 |
| Quartzo: R110108 |
| Quartzo: R150074 |
| Quartzo: R150091 |
| Quartzo: X080015 |
| Quartzo: X080016 |
| Tremolite: R040109 |
| Tremolite: R050210 |
| Tremolite: R050498 |
| Tremolite: R060311 |
| Tremolite: R061087 |
| Tremolite: R070422 |
| Tremolite: R150094 |

Fonte – Elaborada pelo autor.

CAPÍTULO

6

# CONCLUSÕES E TRABALHOS FUTUROS

Após os estudos de simulação e aplicações a dados reais, é possível compreender que a modelagem para dados funcionais não é simples e direta, é preciso se aprofundar no entendimento dos dados e compreender as características das curvas estudadas.

Na aplicação em Astronomia, podemos usar a modelagem funcional para compreender a existência de grupos e subgrupos de fenômenos espaciais e corpos celestes, bem como em Sasdelli *et al.* (2016). Desejamos encontrar evidências de fenômenos que atualmente são descritos teoricamente mas ainda não foram comprovados na prática.

Já na aplicação para o conjunto de dados dos minérios, foram agrupados espectros provindos da espectroscopia de Raman. Essa aplicação pode ser usada para realizar o agrupamento de diferentes minérios, de tal forma que permite compreender através do agrupamento não supervisionado a existência de diferentes tipos de grupos rochosos por meio dos dados amostrados pelas analises espectroscópicas.

Pensando que em muitas vezes a análise desses dados é feita por equipes de profissionais multidisciplinares, e em varias situações são feitas análises simplificadas ou os dados são primeiramente tabelados e depois analisados com o intuito de aplicar nas leituras funcionais métodos convencionais. Foi elaborado um passo a passo simplificado descrito na Figura 29 com o objetivo de orientar os profissionais na análise de dados funcionais.

Este trabalho explorou diversas técnicas de agrupamento não supervisionado para dados funcionais a fim de entender qual seria a melhor utilização de cada algoritmo. Nota-se que para os casos em que a natureza apresenta uma separação dos objetos estudados, a Rede Gás Neural tem uma performance melhor. Já quando os objetos apresentam uma proximidade, onde os conjuntos de objetos podem estar contidos uns nos outros, o K-Medoides pode apresentar uma performance melhor.

Com esse estudo foi possível construir um diagrama para auxiliar na modelagem das

curvas, contemplando os casos mais encontrados na análise de dados funcionais. Como trabalho futuro, é sugerido um estudo de simulação para avaliar a situação em que pode ocorrer a sobreposição entre agrupamentos diferentes.

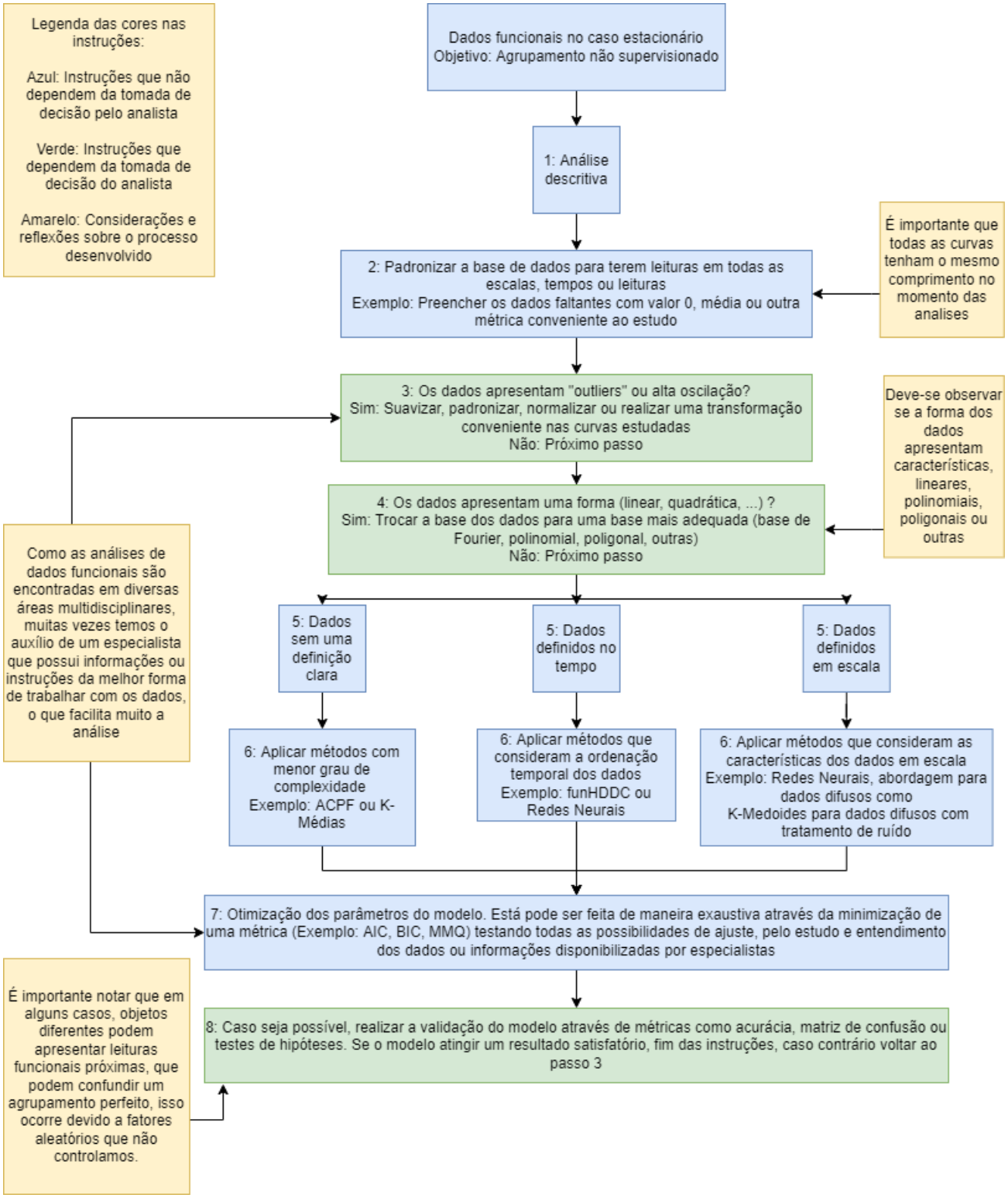


Figura 29 – Instruções para modelagem de dados funcionais.

Fonte – Elaborada pelo autor.

CAPÍTULO

7

# APÊNDICE

## 7.1 Aplicação Adicional

Como a análise espectral em Astronomia e na espectroscopia de Raman são definidas em escala, será exemplificado um caso em particular em que as curvas funcionais estão definidos no tempo.

### *7.1.1 Dados Meteorológicos*

A estação meteorológica é uma ferramenta fundamental para monitorar as condições meteorológicas em diversas áreas, por exemplo na agronomia de modo a auxiliar agricultores a tomarem decisões.

A observação meteorológica de superfície, realizada nas estações meteorológicas, consiste da coleta diária (podendo ser em intervalos de horas, minutos ou dias) de dados referentes às diversas variáveis atmosféricas que caracterizam o estado da atmosfera. Muitas vezes as leituras são interpretadas como curvas ou funções a fim de entender as mudanças climáticas.

Para uma coleta de dados de precisão é necessário seguir algumas normas com relação à localização, tipo e instalação dos equipamentos, e padronização dos horários de observação e dos procedimentos operacionais, como calibração e aferição dos instrumentos de medição.

Temos muitos estudos que fazem análises de dados funcionais tais como Lim, Oh e Cheung (2019), em que o trabalho é voltado ao agrupamento não supervisionado de leituras climáticas (temperatura) que apresentam uma particularidade de serem multi-escala, ou seja, cada curva estudada é definida em uma escala especifica que não necessariamente é igual entre as curvas do conjunto de dados. Os autores usam algoritmos de particionamento recursivo e particionamento combinando escalas nas análises.

A agência estatal de Meteorologia (AEMET) do governo espanhol faz o monitoramento

de todos os aspectos meteorológicos da Espanha, em casos de situações de risco a agência emite alertas avisando os cidadãos espanhóis de possíveis riscos à segurança humana. Desta forma, a AEMET é uma grande fornecedora de dados públicos na área de Meteorologia.

Os dados que serão analisados são uma série de resumos diários de 73 estações meteorológicas espanholas selecionadas no período 1980 a 2009. O conjunto de dados contém informações geográficas de cada estação e a média para o período 1980-2009 de temperatura diária (Celsius), precipitação diária (milímetros) e velocidade diária do vento (metros/segundo).

Usaremos as leituras de temperatura das estações, que podem ser vistas na Figura 30 e todas as curvas de cada uma das estações podem ser observadas na Figura 31.

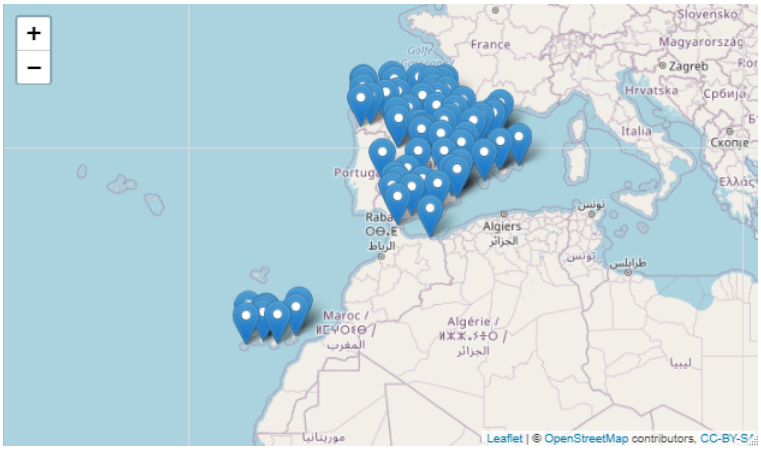


Figura 30 – Localizações geográficas de cada ponto observado no conjunto AEMET.

Fonte – Elaborado pelo autor.

Como o objetivo deste trabalho é entender como o agrupamento não supervisionado para dados funcionais se comporta em diferentes situações, neste caso as curvas estão definidas no tempo, e esperamos diversas peculiaridades destas leituras, como dependência temporal entre pontos da mesma curva e sazonalidade.

Para a realização dos estudos, foi separada uma amostra de supostos três agrupamentos (baseando-se na localização geográfica da estação meteorológica) que poderiam ocorrer neste conjunto de dados. O grupo nomeado de Norte apresenta as observações de temperatura do norte da Espanha com oito curvas. O grupo Centro apresenta as leituras de temperatura do sul da Espanha com 12 curvas e por fim, o grupo Sul com nove curvas, que apresenta a temperatura de uma ilhota próxima ao Marrocos. Cada um dos pontos amostrados representados por grupo estão dispostos na Figura 32. A ideia com esta segregação é compreender se em uma abordagem de agrupamento não supervisionado temos uma diferença entre esses agrupamentos propostos ou se na prática as leituras são todas iguais e não temos diferenças geográficas de temperaturas.

Como uma abordagem inicial e mais simplificada, foi usado o ACPF e o K-Médias, tanto individualmente quanto em uma abordagem conjunta. A visualização gráfica da separação feita pelo ACPF está na Figura 33.

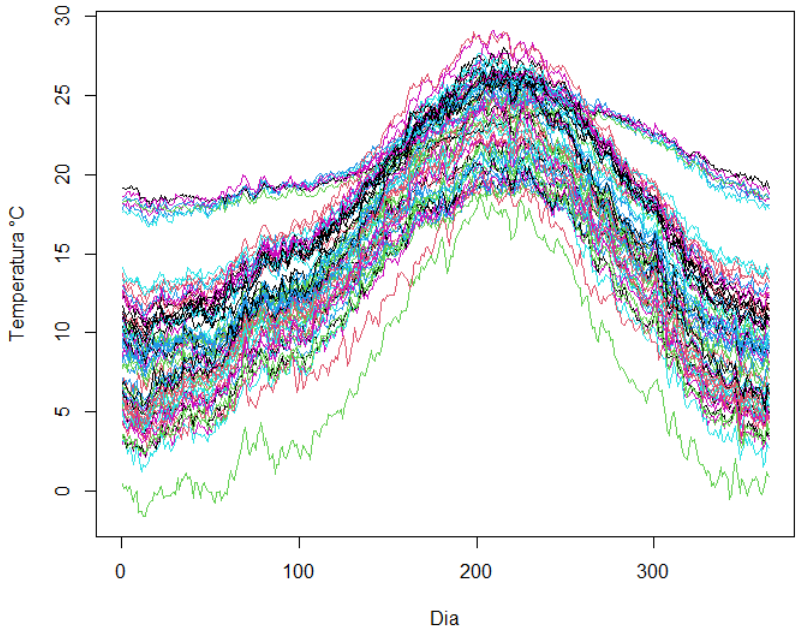


Figura 31 – Visualização das temperaturas no período de 1980 até 2009 de todas as curvas presentes no conjunto AEMET de temperaturas diárias médias pelas estações meteorológicas espanholas.

Fonte – Elaborado pelo autor.

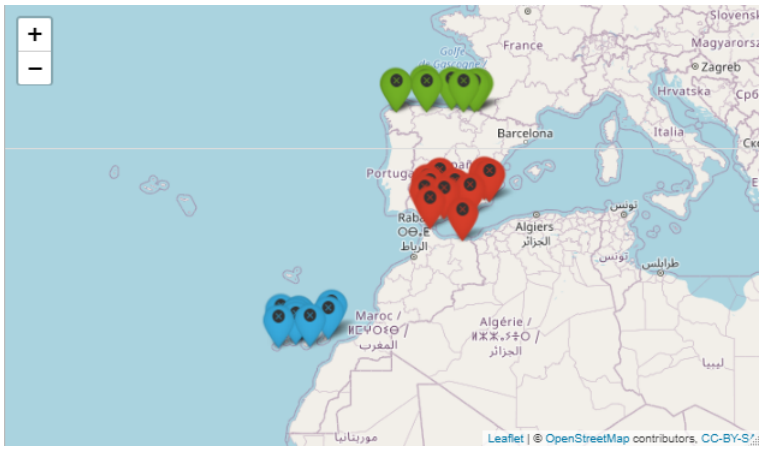


Figura 32 – Visualização das localizações geográficas das medições que foram usadas no trabalho com o intuito de agrupamento não supervisionado. Em azul temos o grupo sul, em vermelho o grupo centro e em verde o grupo norte.

Fonte – Elaborado pelo autor.

Observamos que visualmente o ACPF até realizou uma separação razoável, mas se observarmos a Figura 33 vemos que tivemos dois pontos do grupo Sul e um ponto do grupo Centro que ficaram mais perto da concentração dos pontos do grupo Norte. No geral, a acurácia total calculada foi de 89%, com 100% para o grupo Norte, 91% e 77% para os grupos Centro e Sul, respectivamente.

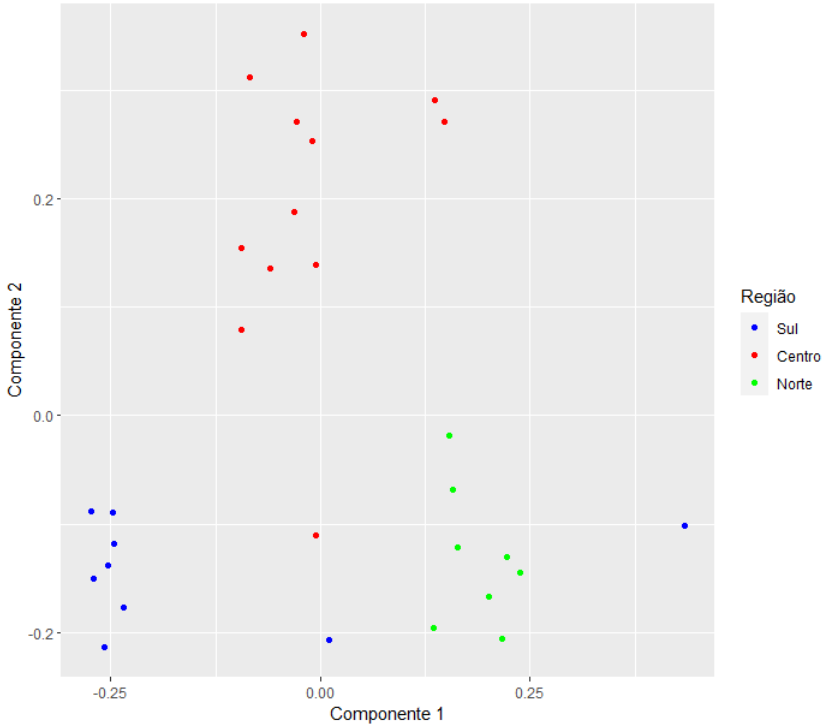


Figura 33 – Agrupamento feito usando o ACPF, selecionando as duas primeiras componentes principais.

Fonte – Elaborado pelo autor.

Visto o bom resultado do ACPF, foi verificado qual seria o melhor número de componentes do ACPF para ser aplicado ao K-Médias. Neste caso foram selecionadas as duas primeiras componentes, e o resultado desta junção pode ser visto na Figura 34 e Tabela 20.

Tabela 20 – Matriz de confusão resultante da metodologia aplicando ACPF e o K-Médias na sequencia, no conjunto AEMET.

| Matriz de Confusão | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Sul | Norte | Centro |
| Agrupamento do Modelo | Sul | 8 | 1 | 0 |
| | Norte | 1 | 7 | 1 |
| | Centro | 0 | 0 | 11 |

Fonte – Elaborada pelo autor.

Observando a Tabela 20, é notável que a aplicação em sequência do ACPF e K-Médias produziu uma assertividade total com 90%, semelhante ao ACPF sozinho. Sendo assim, para comparação das metodologias, os agrupamentos obtidos aplicando somente o K-Médias estão na Figura 35 e Tabela 21.

Na Tabela 21, é possível afirmar que para este estudo o K-Médias não foi capaz de produzir um resultado satisfatório para o conjunto de dados das curvas de temperatura.

Tendo em vista a complexidade e o desafio de fazer um ajuste otimizado ao conjunto

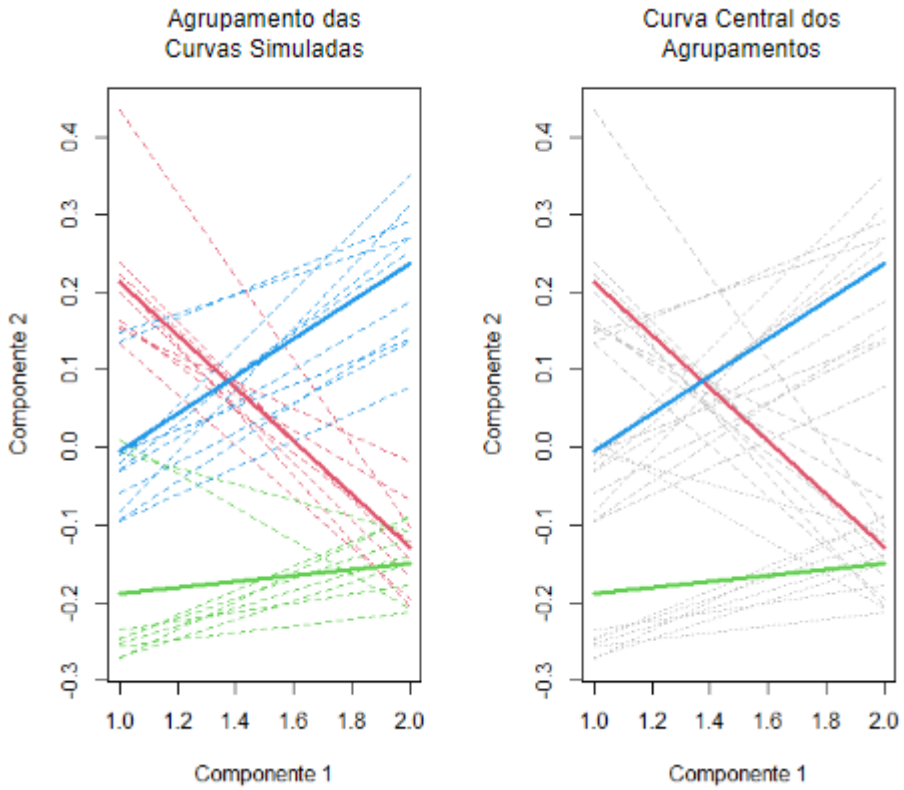


Figura 34 – Agrupamento feito usando o ACPF selecionando as duas primeiras componentes principais e o K-Médias, em verde temos o grupo sul, em azul o grupo centro e em vermelho o grupo norte.

Fonte – Elaborado pelo autor.

Tabela 21 – Matriz de confusão resultante do K-Médias, no conjunto AEMET.

| Matriz de Confusão | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Sul | Norte | Centro |
| Agrupamento do Modelo | Sul | 5 | 1 | 0 |
| | Norte | 1 | 6 | 0 |
| | Centro | 3 | 1 | 12 |

Fonte – Elaborada pelo autor.

de dados, foi feito um estudo mais aprofundado baseado em trocas de bases e suavizações das curvas.

Como as curvas amostradas não têm um comportamento linear simples (observando a análise descritiva na Figura 31), foram testadas algumas trocas de bases e as mais promissoras foram base polinomial, base poligonal e base da transformada de Fourier. Dentre as três melhores, a base poligonal foi a que apresentou o melhor desempenho para o problema.

Também na Figura 31 se observa uma forte oscilação nas curvas, que pode causar confusão e induzir a um agrupamento errado por parte do modelo. Sendo assim, foi aplicada

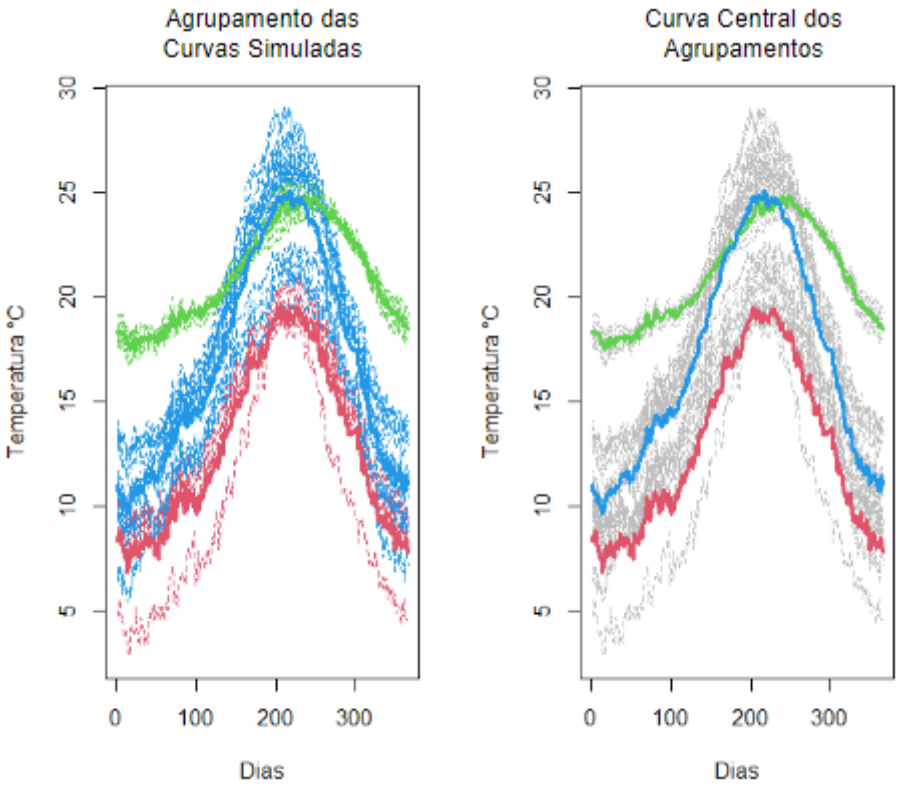


Figura 35 – Agrupamento feito usando o K-Médias, em verde temos o grupo sul, em vermelho o grupo norte e em azul o grupo centro.

Fonte – Elaborado pelo autor.

uma suavização nas curvas baseada na penalidade de *roughness*.

Após todos esses pré-processamentos, o conjunto de dados foi submetida ao funHDDC. Neste algoritmo, foram testados os modelos FLM apresentados na Tabela 1. O que obteve o melhor resultado foi o "AkBQkDk".

O algoritmo funHDDC precisa de um passo de inicialização, foi usado o K-Médias, que mesmo produzindo resultados considerados ruins, forneceu uma ajuda conveniente na convergência do funHDDC.

Por fim no ajuste funHDDC 20%, dos dados foram usados como *threshold*, sendo utilizado para selecionar as dimensões intrínsecas específicas do grupo. Os resultados do ajuste estão na Figura 36.

Analisando a Tabela 22, notamos que a acurácia total do modelo foi de 93%, sendo que só obtivemos erro no grupo sul com 77%, mas vale observar que as curvas foram distribuídas uma para o grupo norte e outra para o centro, ou seja, o erro não foi concentrado em apenas um único grupo, evidenciando que essa atribuição foi devido a algum fator aleatório.

Para os dados meteorológicos de curvas de temperatura, ao realizarmos o agrupamento das curvas de maneira não supervisionada, notamos que o K-Médias não conseguiu ter um bom

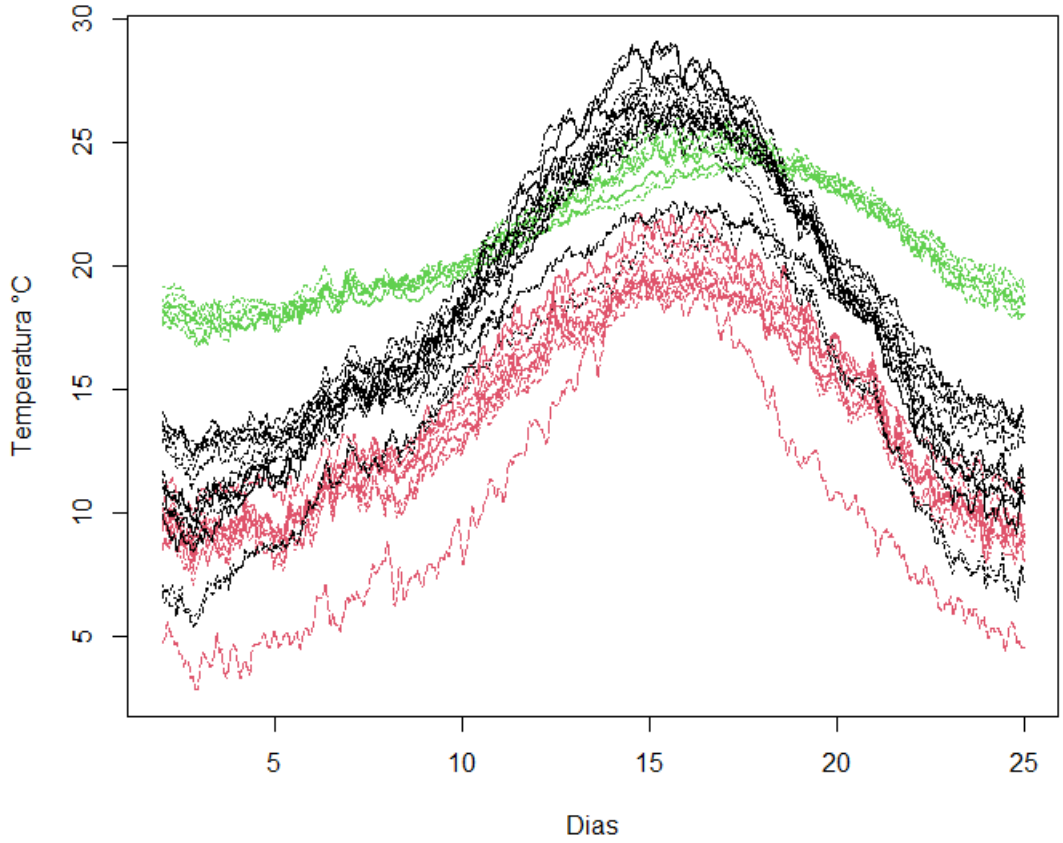


Figura 36 – Agrupamento feito usando o funHDDC, em verde temos o grupo sul, em preto o grupo centro e em vermelho o grupo norte.

Fonte – Elaborado pelo autor.

Tabela 22 – Matriz de confusão resultante do funHDDC, no conjunto AEMET.

| Matriz de Confusão | | | | |
|---|---|---|---|---|
| | | Agrupamento Real | | |
| | | Centro | Norte | Sul |
| Agrupamento do Modelo | Centro | 12 | 0 | 1 |
| | Norte | 0 | 8 | 1 |
| | Sul | 0 | 0 | 7 |

Fonte – Elaborada pelo autor.

desempenho. Em contrapartida o ACPF conseguiu (pelo menos para a amostra selecionada) realizar um bom agrupamento. Caso seja de interesse realizar um ajuste otimizado para os dados, a melhor opção seria usar o funHDDC conjuntamente com o tratamento de mudança de base e suavização de curva. Essa abordagem se mostrou mais promissora do que as demais analisadas.

# REFERÊNCIAS

ABRAHAM, C.; CORNILLON, P.-A.; MATZNER-LØBER, E.; MOLINARI, N. Unsupervised curve clustering using b-splines. **Scandinavian journal of statistics**, Wiley Online Library, v. 30, n. 3, p. 581–595, 2003. Citado na página 46.

BAILEY, S. Principal component analysis with noisy and/or missing data. **Publications of the Astronomical Society of the Pacific**, IOP Publishing, v. 124, n. 919, p. 1015, 2012. Citado na página 24.

BIERNACKI, C. Initializing em using the properties of its trajectories in gaussian mixtures. **Statistics and Computing**, Springer, v. 14, n. 3, p. 267–279, 2004. Citado na página 42.

BOGETOFT, P.; OTTO, L. **Benchmarking with dea, sfa, and r**. [S.l.]: Springer Science & Business Media, 2010. v. 157. Citado na página 58.

BOUVEYRON, C.; GIRARD, S.; SCHMID, C. High-dimensional data clustering. **Computational statistics & data analysis**, Elsevier, v. 52, n. 1, p. 502–519, 2007. Citado na página 37.

____. High-dimensional data clustering. **Computational statistics & data analysis**, Elsevier, v. 52, n. 1, p. 502–519, 2007. Citado nas páginas 40 e 42.

BOUVEYRON, C.; JACQUES, J. Model-based clustering of time series in group-specific functional subspaces. **Advances in Data Analysis and Classification**, Springer, v. 5, n. 4, p. 281–300, 2011. Citado nas páginas 37 e 38.

BRADT, H. **Astronomy methods: A physical approach to astronomical observations**. [S.l.]: Cambridge University Press, 2004. Citado nas páginas 24 e 78.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado nas páginas 53 e 54.

CHIANG, L. H.; RUSSELL, E. L.; BRAATZ, R. D. **Pattern Classification**. [S.l.]: Springer, 2001. 27–31 p. Citado na página 48.

DAVE, R. N. Characterization and detection of noise in clustering. **Pattern Recognition Letters**, Elsevier, v. 12, n. 11, p. 657–664, 1991. Citado na página 47.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, n. 2, p. 224–227, 1979. Citado nas páginas 79 e 85.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citado na página 40.

DUMAS, P.; SOCKALINGUM, G. D.; SULE-SUSO, J. Adding synchrotron radiation to infrared microspectroscopy: what's new in biomedical applications? **Trends in biotechnology**, Elsevier, v. 25, n. 1, p. 40–44, 2007. Citado na página 21.

FARBER, C.; KUROUSKI, D. Detection and identification of plant pathogens on maize kernels with a hand-held raman spectrometer. **Analytical chemistry**, ACS Publications, v. 90, n. 5, p. 3009–3012, 2018. Citado nas páginas 23 e 24.

FEBRERO-BANDE, M.; Oviedo de la Fuente, M. Statistical computing in functional data analysis: The R package fda.usc. **Journal of Statistical Software**, v. 51, n. 4, p. 1–28, 2012. Disponível em: <http://www.jstatsoft.org/v51/i04/>. Citado na página 25.

FOURIER, J. B. J. B. **The analytical theory of heat**. [S.l.]: The University Press, 1878. Citado nas páginas 27 e 28.

GARCÍA, M. L. L.; GARCÍA-RÓDENAS, R.; GÓMEZ, A. G. K-means algorithms for functional data. **Neurocomputing**, Elsevier, v. 151, p. 231–245, 2015. Citado nas páginas 45 e 46.

GHAHRAMANI, Z. **Unsupervised learning**. [S.l.], 2003. 72-112 p. Citado na página 22.

HALL, P.; MÜLLER, H.-G.; WANG, J.-L. Properties of principal component methods for functional and longitudinal data analysis. **The annals of statistics**, Institute of Mathematical Statistics, v. 34, n. 3, p. 1493–1517, 2006. Citado na página 42.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the royal statistical society. series c (applied statistics)**, JSTOR, v. 28, n. 1, p. 100–108, 1979. Citado na página 46.

HU, X.; XU, L. Investigation on several model selection criteria for determining the number of cluster. **Neural Information Processing-Letters and Reviews**, v. 4, n. 1, p. 1–10, 2004. Citado nas páginas 80 e 86.

JACQUES, J.; PREDA, C. Functional data clustering: a survey. **Advances in Data Analysis and Classification**, Springer, v. 8, n. 3, p. 231–255, 2014. Citado nas páginas 38 e 45.

KAUFMAN, L.; ROUSSEEUW, P. J. [S.l.]: John Wiley & Sons, 2009. Citado na página 48.

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological cybernetics**, Springer, v. 43, n. 1, p. 59–69, 1982. Citado na página 50.

KRÄMER, N.; BOULESTEIX, A.-L.; TUTZ, G. Penalized partial least squares with applications to b-spline transformations and functional data. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 94, n. 1, p. 60–69, 2008. Citado na página 42.

LEMAIRE, J. Propriétés asymptotiques en classification (consistance des solutions de problèmes approchés). **Statistique et analyse des données**, v. 8, n. 1, p. 41–58, 1983. Citado na página 46.

LIM, Y.; OH, H.-S.; CHEUNG, Y. K. Multiscale clustering for functional data. **Journal of Classification**, Springer, v. 36, n. 2, p. 368–391, 2019. Citado na página 95.

LOGAN, C.; FOTOPOULOU, S. Unsupervised star, galaxy, qso classification-application of hdbscan. **Astronomy & Astrophysics**, EDP Sciences, v. 633, p. A154, 2020. Citado na página 24.

MARTINETZ, T. M.; BERKOVICH, S. G.; SCHULTEN, K. J. 'neural-gas' network for vector quantization and its application to time-series prediction. **IEEE transactions on neural networks**, IEEE, v. 4, n. 4, p. 558–569, 1993. Citado na página 49.

MCGURK, R. C.; KIMBALL, A. E.; IVEZIĆ, Ž. Principal component analysis of sloan digital sky survey stellar spectra. **The Astronomical Journal**, IOP Publishing, v. 139, n. 3, p. 1261, 2010. Citado na página 24.

MINERAL, E. R. para I. **Dados de Minério RRUFF, acessado em 24/07/2022 ás 21h.** [S.l.], 2022. Disponível em: <https://rruff.info/>. Citado nas páginas 15, 84, 85 e 90.

MORETTIN, P. A. **Ondas e Ondaletas: da análise de Fourier à análise de ondaletas de séries temporais**. [S.l.]: São Paulo: EDUSP, 2014. Citado nas páginas 22, 27, 28, 30 e 32.

MORITZ, H. Least-squares collocation. **Reviews of geophysics**, Wiley Online Library, v. 16, n. 3, p. 421–430, 1978. Citado na página 38.

MURO, C. K.; LEDNEV, I. K. Race differentiation based on raman spectroscopy of semen traces for forensic purposes. **Analytical chemistry**, ACS Publications, v. 89, n. 8, p. 4344–4348, 2017. Citado na página 23.

PARK, H.-S.; JUN, C.-H. A simple and fast algorithm for k-medoids clustering. **Expert systems with applications**, Elsevier, v. 36, n. 2, p. 3336–3341, 2009. Citado na página 48.

R, E. **R para Computação Estatística, acessado em 15/10/2023 ás 19h.** [S.l.], 2023. Disponível em: <https://www.R-project.org>. Citado na página 24.

RAMSAY, J. O.; GRAVES, S.; HOOKER, G. **fda: Functional Data Analysis**. [S.l.], 2022. R package version 6.0.3. Disponível em: <https://CRAN.R-project.org/package=fda>. Citado na página 24.

RAMSEY, J.; SILVERMAN, B. **Functional Data Analysis (2nd ed.)**. [S.l.], 2005. Citado na página 22.

REEM, D. **An Algorithm for Computing Voronoi Diagrams of General Generators in General Normed Spaces**. [S.l.]: IEEE, 2009. Citado na página 50.

ROSENBERG, A.; HIRSCHBERG, J. **V-measure: A conditional entropy-based external cluster evaluation measure**. [S.l.: s.n.], 2007. 410–420 p. Citado na página 53.

ROSTRON, P.; GABER, S.; GABER, D. Raman spectroscopy, review. **laser**, v. 21, p. 24, 2016. Citado na página 23.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. Citado nas páginas 53 e 54.

____. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. Citado nas páginas 79 e 85.

RUDIN, W. [S.l.]: American Mathematical Society, 1959. Citado na página 28.

SASDELLI, M.; ISHIDA, E.; VILALTA, R.; AGUENA, M.; BUSTI, V.; CAMACHO, H.; TRINDADE, A.; GIESEKE, F.; SOUZA, R. de; FANTAYE, Y. *et al.* Exploring the spectroscopic diversity of type ia supernovae with dracula: a machine learning approach. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 461, n. 2, p. 2044–2059, 2016. Citado nas páginas 22, 24, 78, 79 e 91.

SCHMUTZ, A.; BOUVEYRON, J. J. . C. **funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces**. [S.l.], 2021. R package version 2.3.1. Disponível em: <https://CRAN.R-project.org/package=funHDDC>. Citado na página 25.

SCHULMAN, S. G. **Fluorescence and phosphorescence spectroscopy: physicochemical principles and practice**. [S.l.]: Elsevier, 2017. Citado na página 21.

SERVER, G. S. S. A. **Dados Astronômicos SDSS, acessado em 04/07/2022 ás 19h.** [S.l.], 2022. Disponível em: <https://dr14.sdss.org/optical/spectrum/search?id=316543>. Citado nas páginas 15, 24, 78, 79, 80 e 89.

SIKIRZHYTSKAYA, A.; SIKIRZHYTSKI, V.; LEDNEV, I. K. Determining gender by raman spectroscopy of a bloodstain. **Analytical chemistry**, ACS Publications, v. 89, n. 3, p. 1486–1492, 2017. Citado nas páginas 22 e 23.

TEULING, N. D.; PAUWS, S.; HEUVEL, E. van den. A comparison of methods for clustering longitudinal data with slowly changing trends. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–28, 2020. Citado na página 37.

VASQUEZ, R. X-ray photoelectron spectroscopy study of sr and ba compounds. **Journal of Electron Spectroscopy and Related Phenomena**, Elsevier, v. 56, n. 3, p. 217–240, 1991. Citado na página 21.

VIRKLER, K.; LEDNEV, I. K. Blood species identification for forensic purposes using raman spectroscopy combined with advanced statistical analysis. **Analytical chemistry**, ACS Publications, v. 81, n. 18, p. 7773–7777, 2009. Citado nas páginas 23 e 24.

WANG, Y.; ZHOU, X.; WANG, H.; LI, K.; YAO, L.; WONG, S. T. Reversible jump mcmc approach for peak identification for stroke seldi mass spectrometry using mixture model. **Bioinformatics**, Oxford University Press, v. 24, n. 13, p. i407–i413, 2008. Citado na página 22.

WOLKERS, W. F.; OLIVER, A. E.; TABLIN, F.; CROWE, J. H. A fourier-transform infrared spectroscopy study of sugar glasses. **Carbohydrate research**, Elsevier, v. 339, n. 6, p. 1077–1085, 2004. Citado na página 21.

YAO, F.; MÜLLER, H.-G.; WANG, J.-L. Functional data analysis for sparse longitudinal data. **Journal of the American statistical association**, Taylor & Francis, v. 100, n. 470, p. 577–590, 2005. Citado na página 42.

ZAMBOM, A. Z.; COLLAZOS, J. A.; DIAS, R. Functional data clustering via hypothesis testing k-means. **Computational Statistics**, Springer, v. 34, n. 2, p. 527–549, 2019. Citado na página 46.

ZHAO, Q.; HAUTAMAKI, V.; FRÄNTI, P. Knee point detection in bic for detecting the number of clusters. Springer, p. 664–673, 2008. Citado nas páginas 80 e 86.


ICMC
USP
SÃO CARLOS