
Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

**ANNY KEROLLAYNY GOMES RODRIGUES**

# AGRUPAMENTO FUZZY KERNELIZADO ADAPTADO PARA DADOS FALTANTES

Recife

2019

**ANNY KEROLLAYNY GOMES RODRIGUES**

# AGRUPAMENTO FUZZY KERNELIZADO ADAPTADO PARA DADOS FALTANTES

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

**Área de Concentração**: Estatística Aplicada

Orientador: Raydonal Ospina Martínez
Coorientador: Marcelo Rodrigo Portela Ferreira

Recife

2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

R696a Rodrigues, Anny Kerollayny Gomes
Agrupamento fuzzy kernelizado adaptado para dados faltantes / Anny Kerollayny Gomes Rodrigues. – 2019.
63 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2019.
Inclui referências.

1. Estatística. 2. Dados incompletos. 3. Agrupamento fuzzy. I. Martínez, Raydonal Ospina (orientador). II. Título.

310 CDD (23. ed.) UFPE- MEI 2019-124

**ANNY KEROLLAYNY GOMES RODRIGUES**

# AGRUPAMENTO FUZZY KERNELIZADO ADAPTADO PARA DADOS FALTANTES

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 30 de julho de 2019.

**BANCA EXAMINADORA**

---

Prof.(º) Marcelo Rodrigo Portela Ferreira
UFPB

---

Prof.(º) Getúlio José de Amorim Amaral
UFPE

---

Prof.(º) Francisco de Assis Tenório de Carvalho
UFPE/CIN

Aos meus pais, Aldo e Josileide, dedico
com muito amor e carinho.

## AGRADECIMENTOS

Agradeço primeiramente aos meus pais que sempre incentivaram meus estudos, e sempre me apoiaram nos momentos difícies. As minhas irmãs Line e Ka que sempre me trouxeram alegria nos momentos de estresse.

Ao meu namorado, Reinaldo, pelo carinho e amizade e que sempre esteve ao meu lado pacientemente durante meu período de estudo.

A minha amiga Adenice por seus ensinamentos, palavras de apoio e pelos momentos de alegria vividos durante o período do mestrado.

Ao meu amigo Jodavid pelas pela sua disposição e paciência em me ajudar, pelos incentivos, conselhos e boas conversas.

Aos meus amigos de turma Anabeth, Lucas, Eduardo e Jairo vocês foram imprescindíveis para a construção do saber durante esses dois anos de mestrado, obrigada pelas ótimas risadas e por ter tornado nossos dias na pós graduação mais alegres.

Aos amigos de mestrado e doutorado Pedro, César, João Eudes, Ranah, Larissa, Bruna, Cristine, Luíz Félix, Daniel, Cris, Jonas, Saul, Zé, Jordan, Fernando e Joás. Que estiveram comigo nesta caminhada sempre me ajudando e incentivando.

Aos amigos de João Pessoa, Tainara, Felipe, Clarissa, Diogo, Adriano, Andrée Lukas que estão sempre vibrando a cada conquista minha.

Aos meus professores orientadores Raydonal Ospina e Marcelo Rodrigo por suas orientações, conselhos e paciência e por ter acreditado no meu potencial para a elaboração deste trabalho.

A todos os professores do DE-UPPE, pelos valiosos ensinamentos e incentivo, em especial aos professores Cribari Neto, Getúlio Amorim, Audrey Cysneiros, Gauss Cordeiro e José Havier.

Agradeço à Valéria Bittencourt, secretária da pós-graduação em Estatística, pela competência e atenção com os alunos.

Aos participantes da banca examinadora, pelas sugestões e correções.

A CAPES, pelo apoio financeiro.

## RESUMO

Em muitas áreas da ciência, conjuntos de dados e procedimentos estatísticos são frequentemente afetados por valores ausentes (*missing values*). Na análise de agrupamento, a falta de dados pode prejudicar a formação dos grupos. Muitos métodos de agrupamento para dados incompletos presentes na literatura não levam em consideração os pesos ou a relevância das variáveis na formação dos grupos *clusters*. Este trabalho tem como objetivo propor e avaliar o método de agrupamento de núcleos Fuzzy *C-means* com Kernelização da Métrica via distâncias adaptativas locais (VKFCM-K-LP) sob três tipos de estratégias para dados faltantes. A primeira estratégia denominada como Estratégia de Dados Completos (EDC ou *Whole Data Strategy*), realiza o agrupamento apenas com o conjunto de dados completos, ou seja, nesta estratégia as observações ausentes são excluídas da análise. A EDC pode ser aplicada no agrupamento desde que os valores ausentes não ultrapassem a porcentagem de 25% de todos os valores observados. A segunda abordagem usa a estratégia de distância parcial (EDP ou *Partial Distance Strategy*), onde são calculadas as distâncias parciais entre todos os dados disponíveis e, em seguida, reescalonadas pela recíproca da proporção dos valores observados. A terceira técnica, Estratégia de Conclusão Ótima (ECO ou *Optimal Completion Strategy*), calcula valores ausentes de forma iterativa como variáveis auxiliares na otimização de uma função objetivo. Para a avaliação do método VKFCM-K-LP com as estratégias EDC, EDP e ECO, foram utilizados conjuntos de dados com 5%, 10%, 15% e 20% de valores ofaltantes. Os resultados do agrupamento foram analisados de acordo com as medições CR, FM e OERC. O melhor desempenho do agrupamento foi obtido pelas estratégias EDP e ECO. Nos grupos com a abordagem ECO, novas bases de dados foram derivadas e os valores faltantes foram estimados no processo de otimização. Os resultados do agrupamento com a estratégia ECO apresentaram desempenhos superiores quando comparados aos grupos de resultados obtidos a partir do conjunto de dados em que os valores faltantes foram imputados pela média e mediana dos valores observados.



**Palavras-chave:** Dados Incompletos. Agrupamento Fuzzy. Distância Adaptativa Kernelizada.

## ABSTRACT

In many areas of science. data sets and statistical procedures are often affected by missing values. In clustering analysis, lack of data may impair the formation of groups. Many clustering methods for incomplete data present in the literature do not take into account the weights or the relevance of the variables in the construction of the clusters. This work aims to propose and evaluate the method of Fuzzy C-means kernel clustering with metric kernelization via local adaptive distances (VKFCM-K-LP) under three types of strategies for missing data. The first strategy called Whole Data Strategy (EDC) performs clustering only with the complete data set, is in this strategy the missing patterns are excluded from the analysis. The EDC can be applied in the cluster as long as the missing values do not exceed the percentage of 25% of all observed values. The second approach uses the Partial Distance Strategy (EDP) where is calculated the partial distances between all available resources and then rescaled by the reciprocal of the proportion of observed values. The third technique, Optimal Completion Strategy (ECO), computes missing values iteratively as auxiliary variables in the optimization of an objective function. For the evaluation of the VKFCM-K-LP method with EDC, EDP and ECO strategies, data sets with 5%, 10%, 15% and 20% of missing values were used. The results of the clustering were analyzed according to the CR, FM and OERC measurements. The best performance of the clustering was obtained by the EDP and ECO strategies. In the clusters with the ECO approach, new databases were derived, and the missing values were estimated in the optimization process. The results of clustering with the ECO strategy presented superior performance when compared to the result clusters obtained from the dataset in which the missing values were imputed by the mean and median of the observed values.



**Keywords:** Incomplete Data. Fuzzy clustering. Adaptive Distance Kernel.

## LISTA DE FIGURAS

**Figura 1 – Gráficos dos tipos de padrão de *missings*.** . . . . . 26
**Figura 2 – Gráficos de Dispersão e Boxplots para o conjunto de Dados Iris Plant.** . 39
**Figura 3 – Gráficos dos Padrões e Frequências dos valores faltantes por variável para o banco *Iris Plant*.** . . . . . 40
**Figura 4 – Resultados médios das 100 repetições para a Taxa de Erro no Conjunto *Iris Plant*.** . . . . . 43
**Figura 5 – Gráficos de Dispersão e Boxplots para o conjunto de Dados Thyroid Gland.** . . . . . 45
**Figura 6 – Gráficos de Padrões e Frequências dos valores faltantes por variável para o banco *Thyroid Gland*.** . . . . . 46
**Figura 7 – Gráficos dos Resultados médios nas 100 repetições dos algoritmos para a Taxa de Erro no Conjunto *Thyroid Gland*.** . . . . . 48
**Figura 8 – Gráficos de desempenho dos métodos quando a quantidade de *missings* é variada de 5 a 20%.** . . . . . 51
**Figura 9 – Gráficos da Análise de Componentes Principais.** . . . . . 52
**Figura 10 – Gráficos para o banco *Thyroid Gland* obtido pelos métodos de imputação única.** . . . . . 56

## LISTA DE TABELAS

**Tabela 1 – Matriz de Confusão.** . . . . . 35
**Tabela 2 – Desempenho do algoritmo de agrupamento VKFCM-K-LP sob três tipos de abordagem estudadas EDC, EDP e ECO para o banco *Iris Plant*.** 39
**Tabela 3 – Matrizes de confusão obtidas pelo algoritmo VKFCM-K-LP em conjunto com os métodos EDC, EDP e ECO utilizando 5, 10, 15 e 20% de dados faltantes.** . . . . . 41
**Tabela 4 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP em conjunto com o método EDC sob diferentes porcentagens de dados faltantes.** . . . . . 42
**Tabela 5 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP em conjunto com o método EDP sob diferentes porcentagens de dados faltantes.** . . . . . 42
**Tabela 6 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP em conjunto com o método ECO sob diferentes porcentagens de dados faltantes.** . . . . . 43
**Tabela 7 – Consistência das variáveis para o conjunto de dados *Iris Plant*.** . . . . . 44
**Tabela 8 – Desempenho do algoritmo de agrupamento VKFCM-K-LP sob três tipos de abordagens estudadas para o banco *Thyroid Gland*.** . . . . . 47
**Tabela 9 – Matrizes de confusão obtidas pelo algoritmo VKFCM-K-LP em conjunto com os métodos EDC, EDP e ECO utilizando 5, 10, 15 e 20% de dados faltantes n dadoso conjunto de *Thyroid Gland*.** . . . . . 47
**Tabela 10 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP sob a abordagem EDC para o conjunto *Thyroid Gland*.** . . . . . 48
**Tabela 11 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP sob a abordagem EDP para o conjunto *Thyroid Gland*.** . . . . . 49
**Tabela 12 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP sob a abordagem ECO para o conjunto *Thyroid Gland*.** . . . . . 49
**Tabela 13 – Consistências das variáveis para o conjunto de dados *Thyroid Gland*.** . 50
**Tabela 14 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Média no conjunto de dados *Iris Plant*.** . . . . 53

**Tabela 15 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Mediana no conjunto de dados *Iris Plant*.** . . . 54
**Tabela 16 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Média no conjunto de dados *Thyroid Gland*.** . 54
**Tabela 17 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Mediana no conjunto de dados *Thyroid Gland*.** 55

## LISTA DE ALGORITMOS

**Algoritmo 1 – Método de agrupamento VKFCM-K-LP** . . . . . 24
**Algoritmo 2 – Método de agrupamento VKFCM-K-LP sob a estratégia EDC.** . . 29
**Algoritmo 3 – Método de agrupamento VKFCM-K-LP sob a estratégia EDP.** . . . 31
**Algoritmo 4 – Método de agrupamento VKFCM-K-LP sob a abordagem ECO.** . . 32
**Algoritmo 5 – Geração dos *missings*.** . . . . . 34

## LISTA DE SÍMBOLOS

| | |
|---|---|
| $A$ | Matriz positiva definida |
| $K$ | Número de grupos |
| $p$ | Número de variáveis |
| $n$ | Número de observações |
| $\Omega$ | Conjunto de $n$ observações |
| $\boldsymbol{\lambda}_k$ | Vetor de pesos para o grupo $k$ |
| $\mathbf{v}_k$ | Vetor de protótipos para o grupo $k$ |
| $x_{ij}$ | Descreve o $i$-ésimo dado da $j$-ésima grupo variável |
| $u_{ik}$ | Grau de pertinência para $i$-ésima observação do grupo $k$ |
| $\phi$ | Mapeamento não-linear |
| $\mathscr{F}$ | Espaço de características |
| $\mathbf{M}$ | Matriz indicadora de *missings* |
| $\varphi^2$ | Distância adaptativa local |
| $\varphi^2_{dp}$ | Distância adaptativa local parcial |
| $X$ | Conjunto de dados |
| $\mathscr{P}$ | Partição *a priori* |
| $P$ | Partição rígida |
| $C$ | Número de classes *a priori* |
| $d$ | Medida de consistência das variáveis |
| $\mathbf{U}$ | Matriz de Partição Fuzzy |
| $\mathscr{K}$ | Função Kernel Gaussiana |
| $\mathbf{V}$ | K-upla de protótipos |
| $\mathbf{\Lambda}$ | K-upla de vetores de pesos |
| $T$ | Número máximo de iterações |
| $J$ | Função objetivo |
| $J_{dp}$ | Função objetivo para a estratégia EDP |
| $J_M$ | Função objetivo para a estratégia ECO |

**SUMÁRIO**


| | | |
|---|---|---|
| **1** | **INTRODUÇÃO** | 14 |
| 1.1 | ORGANIZAÇÃO DA DISSERTAÇÃO | 18 |
| **2** | **REFERENCIAL TEÓRICO** | 19 |
| 2.1 | AGRUPAMENTO KERNEL *FUZZY C-MEANS* (KFCM) | 20 |
| 2.2 | AGRUPAMENTO FUZZY BASEADO EM KERNEL COM PONDERAÇÃO AUTOMÁTICA DAS VARIÁVEIS VIA DISTÂNCIA ADAPTATIVA LOCAL | 21 |
| 2.3 | ANÁLISE DE DADOS INCOMPLETOS | 25 |
| **2.3.1** | **Técnicas para lidar com Valores Faltantes** | 27 |
| 2.4 | VKFCM-K-LP SOB A ABORDAGEM DE DADOS FALTANTES | 28 |
| **2.4.1** | **Estratégia de dados Completos (EDC)** | 28 |
| **2.4.2** | **Estratégia da Distância Parcial (EDP)** | 29 |
| **2.4.3** | **Estratégia de Conclusão Ótima (ECO)** | 31 |
| **2.4.4** | **Complexidade computacional dos métodos** | 32 |
| **3** | **DESENHO EXPERIMENTAL** | 33 |
| 3.1 | GERAÇÃO DOS MISSINGS | 33 |
| 3.2 | MEDIDAS DE QUALIDADE | 34 |
| **4** | **RESULTADOS** | 37 |
| 4.1 | CONJUNTO DE DADOS *IRIS PLANT* | 38 |
| 4.2 | CONJUNTO DE DADOS THYORIOD GLAND | 44 |
| 4.3 | COMPARAÇÃO ENTRE OS MÉTODOS DE IMPUTAÇÃO | 51 |
| **5** | **CONSIDERAÇÕES FINAIS** | 57 |
| | **REFERÊNCIAS** | 60 |

# 1 INTRODUÇÃO

O aumento no volume e na variedade dos dados requer avanços em metodologias para entender, processar e resumir os dados automaticamente. A análise de agrupamento é uma das principais técnicas utilizadas para a extração de conhecimento em grandes conjuntos de dados, pois colaboram no processo de compreensão e visualização das estruturas de dados (ESTIVILL-CASTRO, 2002; SHEN *et al.*, 2006).

O objetivo do agrupamento é organizar os dados (observações, itens de dados, vetores de recursos, etc) com base em critérios de similaridade (ou dissimilaridade) de tal forma que as observações pertencentes a um mesmo grupo (*cluster*) apresentem um alto grau de similaridade enquanto observações em grupos distintos apresentem um alto grau de dessimilaridade (JAIN; MURTY; FLYNN, 1999; XU; WUNSCH, 2005). A análise de agrupamento é uma classificação não supervisionada, pois o problema consiste em agrupar um determinado conjuntos de dados não rotulados em grupos significativos (JAIN; MURTY; FLYNN, 1999). Os métodos de agrupamento são aplicados em diversas áreas do conhecimento, tais como: Mineração de Dados, Segmentação de Imagens, Reconhecimento de Padrões, entre outras (FILIPPONE *et al.*, 2008). A depender da aplicação utilizada os grupos obtidos no agrupamento podem apresentar características diferentes.

Dessa forma, diferentes técnicas de agrupamento foram propostas na literatura, sendo as mais populares baseadas em hierarquias e partição. No agrupamento hierárquico são encontradas estruturas que podem ser divididas em camadas onde cada camada apresenta subestruturas que podem ser divididas e assim por diante de forma recursiva. O resultado final desta técnica, é uma estrutura hierárquica de grupos chamada de dendrograma (WARD; JOE, 1963). Nos métodos de agrupamentos Particionais, uma única partição do conjunto de dados é obtida, e geralmente estes métodos são baseados na otimização de uma função objetivo (FILIPPONE *et al.*, 2008). Estes métodos são mais flexíveis que os hierárquicos já que permitem que as observações mudem de grupo em cada passo do algoritmo, se essa mudança apresentar uma melhor solução em termos de variabilidade da partição resultante.

Os métodos de agrupamento particionais se diferenciam em dois tipos, o agrupamento *hard* e o agrupamento *fuzzy*. Nos métodos de agrupamento *hard* os grupos são naturalmente disjuntos, ou seja, o conjunto de dados é particionado em um número predefinido de grupos. Nesta técnica de agrupamento cada dado pode pertencer seomente a um grupo, o que

torna seus resultados menos informativos, pois não existe distinção entre os dados típicos do grupo e os dados que estão no limite de dois ou mais grupos. Em aplicações do mundo real, os limites de grupos geralmente são difíceis de definir, pois é complexo encontrar critérios razoáveis que incluam alguns objetos de dados em um *cluster*, mas excluam outros. Para contornar tal situação, são usadas técnicas que permitem a flexibilização destes critérios, como os métodos de agrupamento *fuzzy* que aceitam o fato de que os grupos geralmente não estão completamente separados. Estes métodos usam ferramentas de lógica *fuzzy* em particular os dos conjuntos *fuzzy* em que são atribuídas aos dados de um grupo, um grau de pertinência que varia entre 0 e 1 (ZADEH, 1965). Logo, um dado pode pertencer a todos os grupos com um certo grau de pertinência.

A teoria da lógica *fuzzy* dos conjuntos *fuzzy* foi inicialmente aplicada em agrupamento no trabalho de Ruspini (1969). A ideia de Zadeh aplicada no contexto de agrupamento é representar a similaridade que uma observação compartilha com cada grupo por meio dos graus de pertinências (BEZDEK; EHRLICH; FULL, 1984). Assim, cada observação da amostra de dados terá um grau de pertinência em cada grupo, as pertinências próximas a 1 significam um alto grau de similaridade entre a observação e o grupo, enquanto pertinências próximas a 0 implicam em pouca similaridade.

Os principais métodos de agrupamento *fuzzy* são descritos nos trabalhos de Evers *et al.* (1999), Jain, Murty e Flynn (1999), Bezdek (1981). O algoritmo de agrupamento *fuzzy* mais popular é o *Fuzzy C-Means* (FCM) (BEZDEK, 1981), a medida de similaridade (ou dissimilaridade) comumente utilizada neste método é a distância euclidiana. Segundo Ferreira e Carvalho (2014) esta distância apresenta bons resultados quando aplicados a conjuntos de dados nos quais os grupos são aproximadamente hiperesféricos e aproximadamente linearmente separáveis, ou seja, com esta limitação grupos que possuem diferentes formas, tamanhos e orientação permitem que os métodos que utilizam a distância euclidiana encontre agrupamentos pouco representativo.

Recentemente, alguns métodos de agrupamento que produzem hipersuperfícies de separação não-linear entre os grupos foram propostos. Dentre os quais estão os métodos de agrupamentos baseados em funções Kernel como: Kernel *C-Means* (BEZDEK, 1981), Kernel *C-Means* Fuzzy (GIROLAMI, 2002), Mapas Auto organizáveis (KOHONEN, 1982; KOHONEN, 2013) e Neural gas (MARTINETZ; BERKOVICH; SCHULTEN, 1993). O uso das funções Kernel permitem um mapeamento não-linear arbitrário $\phi$ do espaço original $p$-dimensional do conjunto de dados $X \subset \mathbb{R}^p$ para um espaço de dimensão mais alta (possivelmente infinita),

chamado espaço de características $\mathscr{F}$. O intuito desta transformação, é que ao passar pra dimensões mais alta pode ser possível obter grupos mais definidos e linearmente separáveis (HAYKIN, 1994).

Uma das vantagens dos métodos baseados em funções Kernel, é que os produtos internos no espaço de características podem ser expressos por um kernel de Mercer $\mathscr{K}$ (MERCER, 1909; GIROLAMI, 2002), isto é, $\mathscr{K}(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}')$, em que $\mathbf{x}'\mathbf{x} \in X$, $X \subset \mathbb{R}^p$. Uma das modificações utilizando funções Kernel é a Kernelização da Métrica, onde os protótipos (elementos que pertecem ao conjunto de dados e são usados para caracterizar o grupo) dos grupos são obtidos no espaço original dos dados de entrada, e as distâncias dos dados pra cada protótipo dos grupos são calculadas por meio de funções Kernel (ZHANG; CHEN, 2004). Pesquisas demonstram que os métodos de agrupamento baseado em funções Kernel apresentam melhores desempenhos em relação os métodos tradicionais, por produzirem hipersuperfícies não-lineares de separação entre grupos (ZHANG; CHEN, 2004). Entretanto, na maioria dos domínios, especialmente para conjuntos de dados de alta dimensão algumas variáveis podem ser irrelevantes para a construção dos grupos, e algumas dentre as relevantes, podem ser menos importantes que outras em relação a um grupo específico. Em virtude disso, Ferreira e Carvalho (2014) propuseram uma série de métodos baseados em funções Kernel com ponderação automática das variáveis. Neste contexto foi abordado o agrupamento Kernel Fuzzy *C-Means* com Kernelização da Métrica via distâncias adaptativas locais, utilizando a restrição de que o produto dos pesos das variáveis seja igual a 1, denominado como VKFCM-K-LP. Estes métodos foram amplamente utilizados para descrever as estruturas dos grupos nos conjuntos de dados. No entanto, o método VKFCM-K-LP ainda não foi estudado dentro da abordagem de dados faltantes.

No mundo real, muitas aplicações e procedimentos inferenciais sofrem na presença de *missings* (dados incompletos, faltantes, ausentes). Os dados podem estar ausentes por vários motivos, incluindo procedimentos imperfeitos de entrada manual dos dados, medição incorreta e erros de mensuração de equipamentos, entre outros (FARHANGFAR; KURGAN; PEDRYCZ, 2007). Em muitas áreas, como na Indústria e na área Médica não é incomum encontrar bancos de dados que tenham até 50% ou mais de seus valores ausentes (LAKSHMINARAYAN; HARP; SAMAD, 1999; KURGAN *et al.*, 2005). O problema da falta de dados tem sido comumente estudado, uma razão para isso é o fato de muitas estatísticas terem sido originalmente desenvolvidas para conjuntos de dados sem valores ausentes, e mesmo uma pequena quantidade desses valores no conjunto de dados podem causar problemas nas análises e na tomada de decisão, o que justifica a necessidade de procurar por mecanismos eficientes para lidar com dados incompletos

(HAREL; ZHOU, 2007).

O desenvolvimento de métodos estatísticos para lidar com dados incompletos tem sido alvo de pesquisas nas últimas décadas (RUBIN, 1976; LITTLE; RUBIN, 2014; SCHAFER, 1999). Green *et al.* (2001) identificaram duas alternativas para lidar com valores ausentes; a saber: *Imputação*, em que os valores faltantes são estimados através dos valores observados no banco de dados, as técnicas mais populares são a *Imputação via Média* ou *Mediana*, e *Exclusão*, onde os valores omissos são ignorados do banco de dados. Estas alternativas, embora simples, podem levar a produzir estimativas tendenciosas tanto pela redução do tamanho do conjunto de dados quanto na substituição desses valores por estimativas (LITTLE; RUBIN, 2014). Uma abordagem eficaz é adaptar as análises de dados tradicionais de modo que elas possam se ajustar a conjuntos de dados que possuam valores faltantes.

Diversas abordagens tem sido introduzidas na tentativa de estender as técnicas de agrupamento na presença de valores faltantes. Em reconhecimento de padrões com dados incompletos, Sebestyen (1962), introduz uma abordagem baseada em suposições probabilísticas para lidar com dados faltantes. O algoritmo de Esperança - Maximização (EM) foi usado para lidar com dados incompletos em agrupamento (DEMPSTER; LAIRD; RUBIN, 1977). Vários métodos foram propostos para adaptar o método FCM para dados faltantes (MIYAMOTO; TAKATA; UNAYAHARA, 1998). Wagstaff (2004) descutiu o problema de dados faltantes e propôs o método *C-means* com Restrições Suaves. Poddar e Jacob (2017) apresentaram um método de agrupamento com entrada de dados ausentes usando penalidades de fusão não-convexa.

Hathaway e Bezdek (2001) propuseram estratégias para lidar com valores ausentes em análise de agrupamento utilizando o método FCM. Sob a abordagem de Hathaway e Bezdek (2001), Li, Gu e Zhang (2010) propuseram um método de agrupamento FCM baseado nas observações vizinhas mais próximas. Li, Zhong e Li (2012) estenderam o método FCM e atribuíram uma ponderação por atributo para dados incompletos, no qual o grau de importância de cada atributo é visto como uma variável a ser otimizada durante o agrupamento. Recentemente, Li *et al.* (2017) introduziu o método Kernel para agrupar conjunto de dados com valores faltantes no âmbito de imputação de observações.

Nesta dissertação é adaptado o método de agrupamento VKFCM-K-LP sob três estratégias para lidar com dados faltantes propostas por Hathaway e Bezdek (2001). A primeira, chamada de Estratégia de Dados Completo (EDC), na qual remove todos os dados da amostra que incluem valores faltantes. A segunda é a Estratégia da Distância Parcial (EDP), na qual

calcula a soma das distâncias euclidianas quadradas entre todos os valores disponíveis, e em seguida, pondera pela proporção de valores utilizados no seu cálculo. A última é a Estratégia de Conclusão Ótima (ECO), a qual calcula iterativamente os valores ausentes como variáveis adicionais sobre as quais a função objetivo é minimizada. Na avaliação do método VKFCM-K-LP em conjunto com as estratégias EDC, EDP e ECO, foram utilizados bancos de dados com 5%, 10%, 15% e 20% de *missings*. Os resultados das análises foram quantificados de acordo com as medidas de qualidade: Coeficiente Corrigido de Rand (CR), medida F-*measure* (FM), a Taxa total de Erro de Classificação (OERC) e a medida de consistência das variáveis na estratégia ECO (HUBERT; ARABIE, 1985; BAEZA-YATES; RIBEIRO *et al.*, 2011; BREIMAN *et al.*, 1984; LEE; BERGER; AVICZER, 1996). Além disso, os resultados do agrupamento com estratégia ECO foram comparados com os resultados dos agrupamentos utilizando os métodos de *Imputação via Média* e *Mediana*.

## 1.1 ORGANIZAÇÃO DA DISSERTAÇÃO

Além deste capítulo introdutório, esta dissertação está organizada em mais cinco capítulos, como segue.

No Capítulo 2 estão descritas a teoria das funções Kernel, o método de agrupamento Kernel Fuzzy *C-Means* convencional é mostrado na Seção 2.1. O método de agrupamento VKFCM-K-LP é descrito na Seção 2.2. Na Seção 2.3 são apresentados os mecanismos de dados faltantes com ênfase no Missings Completamente Aleatório (MCA). As abordagens comuns para lidar com dados faltantes e o método VKFCM-K-LP em conjunto com as técnicas EDC, EDP e ECO são mostradas na Subseção 2.3.1 e Seção 2.4, respectivamente.

No Capítulo 3 encontra-se a metodologia, a Seção 3.1 apresenta a geração dos *missings* por simulação e uma descrição mais detalhada das medidas de avaliação CR, FM, OERC e a medida de Consistência e mostrada na Seção 3.2.

No Capítulo 4 estão apresentados os resultados da avaliação do método VKFCM-K-LP com as abordagens EDC, EDP e ECO, em todas as porcentagens de *missings* estudadas, aplicadas a dois conjuntos de dados *Iris Plant* e *Thyroid Gland* amplamente estudados na literatura, nas Seções 4.1 e 4.2 respectivamente. Uma comparação do agrupamento com a técnica ECO e o agrupamento com as técnicas de *Imputação Média* e *Mediana* também é realizada e mostrada na Seção 4.3. Por fim, as conclusões e trabalhos futuros encontram-se no Capítulo 5.

# 2 REFERENCIAL TEÓRICO

Recentemente, os métodos baseados em funções Kernel tem sido alvo de pesquisas no contexto de agrupamento de dados (FILIPPONE *et al.*, 2008; FARHANGFAR; KURGAN; PEDRYCZ, 2007; JAIN, 2010). O objetivo destes métodos é o uso de um mapeamento não-linear arbitrário $\phi$ do espaço original dos dados de entrada para um espaço de mais alta dimensão (possivelmente infinta), chamado espaço de características $\mathscr{F}$.

Seja $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ um conjunto não-vazio, em que $\mathbf{x}_i \in \mathbb{R}^p, \forall_i$. Uma função $\mathscr{K} : X \times X \to \mathbb{R}$ é dita um Kernel de Mercer se, e somente se $\mathscr{K}$ é simétrica, isto é, $\mathscr{K}(\mathbf{x}_k, \mathbf{x}_i) = \mathscr{K}(\mathbf{x}_i, \mathbf{x}_k)$ e a seguinte desingualdade for válida (MERCER, 1909):

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_i c_k \mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \; \forall_n \geq 2; \tag{2.1}$$

em que, $c_r \in \forall_r = 1, \ldots, n$. Cada kernel de Mercer pode ser expresso como:

$$\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_k), \tag{2.2}$$

no qual, $\phi : X \to \mathscr{F}$ executa um mapeamento não-linear do espaço original de $X$ para o espaço de caractéristicas de alta dimensão $\mathscr{F}$. Ao aplicar o mapeamento não-linear em $X$ o espaço original é mapeado por $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_k)$.

Um dos aspectos mais relevantes nas aplicações de métodos baseados em Kernel é a possibilidade de calcular distâncias euclidianas em $\mathscr{F}$ sem que o mapeamento não-linear $\phi$ esteja explicitamente especificado (MÜLLER *et al.*, 2001; SCHÖLKOPF; SMOLA; MÜLLER, 1998). Isto pode ser feito utilizando a *distância Kernel trick* (SCHÖLKOPF; SMOLA; MÜLLER, 1998; SCHOLKOPF; SMOLA, 2001)

$$\begin{aligned} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_k)||^2 &= (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_k))^\top (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_k)) \\ &= \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_k) - 2\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_k) + \phi(\mathbf{x}_i)\phi(\mathbf{x}_k) \\ &= \mathscr{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) + \mathscr{K}(\mathbf{x}_k, \mathbf{x}_k), \end{aligned} \tag{2.3}$$

em que o cálculo das distâncias dos vetores no espaço de características é uma função dos vetores de entrada. As funções Kernel tipicamente utilizadas são (VAPNIK, 2013):

- Linear: $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = \mathbf{x}_i^\top \mathbf{x}_k$,
- Polinomial de grau de $d$: $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = (\gamma \mathbf{x}_i^\top \mathbf{x}_k + \theta)^d$, $\gamma > 0$, $\theta > 0$, $d \in \mathbb{N}$,
- Gaussiana: $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\sigma^2}}$, $\sigma > 0$,
- Laplaciana: $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_k||}$, $\gamma > 0$,
- Sigmóide: $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = tanh(\gamma \mathbf{x}_i^\top \mathbf{x}_k + \theta)$, $\gamma > 0$, $\theta > 0$,

em que, $\gamma$, $\theta$,$\sigma$ e $d$ são parâmetros do Kernel.

Na literatura os métodos de agrupamento baseados em Kernel podem ser divididos em três categorias, Kernelização da Métrica (WU; XIE; YU, 2003; ZHANG; CHEN, 2003), Agrupamento no Espaço de Características (GRAEPEL; OBERMAYER, 1998) e Descrição via Vetores de Suporte (*Support Vector*) (CAMASTRA; VERRI, 2005). No entanto este trabalho restringe-se ao agrupamento baseados em Kernelização da Métrica. Estes métodos de agrupamento buscam por protótipos no espaço original dos dados de entrada e a distância entre um dado $\mathbf{x}_i$ e um protótipo do $k$-ésimo grupo $\mathbf{v}_k$ é determinada por uma função Kernel

$$||\phi(\mathbf{x}_i) - \phi(\mathbf{v}_k)||^2 = \mathscr{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathscr{K}(\mathbf{x}_i, \mathbf{v}_k) + \mathscr{K}(\mathbf{v}_k, \mathbf{v}_k). \tag{2.4}$$

## 2.1 AGRUPAMENTO KERNEL *FUZZY C-MEANS* (KFCM)

Seja $\Omega = \{1, \ldots, n\}$ um conjunto de $n$ observações indexadas por $i$ e descritas por $p$ variáveis. Seja $P = \{P_1, P_2, \ldots, P_k\}$ uma partição de $\Omega$ em $K$ grupos. O objetivo do método de agrupamento Kernel *Fuzzy C-Means* baseado em Kernelização da Métrica é minimizar a seguinte função objetivo

$$\begin{aligned} J &= \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ki})^m ||\phi(\mathbf{x}_i) - \phi(\mathbf{v}_k)||^2 \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ki})^m \{\mathscr{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathscr{K}(\mathbf{x}_i, \mathbf{v}_k) + \mathscr{K}(\mathbf{v}_k, \mathbf{v}_k)\}, \end{aligned} \tag{2.5}$$

sob as restrições,

$$\begin{cases} u_{ki} \in [0,1], & \forall k, i, \\ \sum_{k=1}^{K} u_{ki} = 1, & \forall i, \end{cases} \tag{2.6}$$

em que, $\mathbf{v}_k \in \mathbb{R}^p$ é o protótipo do $k$-ésimo grupo, $u_{ki}$ é o grau de pertinência *fuzzy* para a observação $i$ do $k$-ésimo grupo, $k = 1, \ldots, K$, $i = 1, \ldots, n$ e $m \in \mathbb{R}^+$ é o parâmetro que controla o

grau de imprecisão da pertinência para cada observação $i$. Assim, definimos $\mathbf{U} = [u_{ki}] \in \mathbb{R}^{K\times n}$, como a matriz de partição *fuzzy*. A derivação dos protótipos dos grupos depende da escolha da função Kernel, ao considerarmos o Kernel Gaussiano, o mais utilizado na literatura, temos que $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_i) = 1$, para todo $i = 1, \ldots, n$. Assim, a função objetivo descrita na Eq. (2.5) é reescrita como (GRAVES; PEDRYCZ, 2010)

$$J = 2\sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ki})^m(1 - \mathscr{K}(\mathbf{x}_i, \mathbf{v}_k)), \tag{2.7}$$

desta forma, a equação dos protótipos dos grupos é definida por

$$\mathbf{v}_k^{(t+1)} = \frac{\sum_{i=1}^{n}(u_{ki}^{(t+1)})^m \mathscr{K}(\mathbf{x}_i, \mathbf{v}_k^{(t)})\mathbf{x}_i}{\sum_{i=1}^{n}(u_{ki}^{(t+1)})^m \mathscr{K}(\mathbf{x}_i, \mathbf{v}_k^{(t)})},\ k = 1, \ldots, K. \tag{2.8}$$

Na etapa de atualização da matriz de partição *fuzzy* $\mathbf{U}$, os protótipos dos grupos $\mathbf{v}_k$ são mantidos fixos, assim precisamos encontrar os graus de pertinências *fuzzy* $u_{ki}$ ($k = 1, \ldots, K, i = 1, \ldots, n$). Pelas restrições impostas na Eq. (2.6) e fazendo uso dos multiplicadores de Lagrange para o processo de otimização da função objetivo $J$ (GRAVES; PEDRYCZ, 2010) temos que

$$u_{ki}^{(t+1)} = \left[\sum_{h=1}^{K}\left(\frac{1 - \mathscr{K}(\mathbf{x}_i, \mathbf{v}_k^{(t+1)})}{1 - \mathscr{K}(\mathbf{x}_i, \mathbf{v}_h^{(t+1)})}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{2.9}$$

## 2.2 AGRUPAMENTO FUZZY BASEADO EM KERNEL COM PONDERAÇÃO AUTOMÁTICA DAS VARIÁVEIS VIA DISTÂNCIA ADAPTATIVA LOCAL

Os métodos de agrupamento baseados em Kernel comumente encontrados na literatura, como o Kernel *Fuzzy C-Means* (CHEN, 2002), não levam em consideração os pesos ou a relevância de cada variável na formação dos grupos. Entretanto, para conjuntos de dados de alta dimensão, algumas variáveis podem ter sua relevância diferenciada na construção dos grupos, podendo apresentar pesos pequenos, ou até mesmo irrelevante no processo de agrupamento. Além disso, os grupos podem ser formados por conjuntos de diferentes de variáveis que possuem pesos diferenciados.

A partir da motivação de que possa existir diferenças nos pesos das variáveis e que estas diferenças podem ser mensuradas, de modo que os métodos de agrupamento baseados

em Kernel possam ser melhorado, Ferreira e Carvalho (2014) propuseram vários métodos de agrupamento baseados em Kernel com ponderação automática das variáveis. Dentre os métodos de agrupamento propostos por Ferreira e Carvalho (2014) o VKFCM-K-LP leva em consideração os pesos ou as relevâncias de cada variável para a construção dos *clusters*. Este método de agrupamento é baseado na distância adaptativa local Kernelizada, sob a restrição de que o produto dos pesos das variáveis em cada grupo seja igual a 1.

Nas distâncias adaptativas locais considera-se uma parametrização por um vetor de pesos para cada grupo. Logo, quanto mais próximas as observações estão do protótipo de um dado grupo com relação a uma dada variável, maior será sua importância para este grupo. A restrição imposta sobre o vetor de pesos no método VKFCM-K-LP foi motivada pelos trabalhos de Diday (1977) e Gustafson e Kessel (1979) que foram baseados no agrupamento *hard* via distâncias adaptativas e no *fuzzy* via distâncias quadráticas, ambos definidos por uma matriz $A_k$ positiva definida simétrica, com dimensão $p \times p$ associada ao $k$-ésimo grupo $k = 1, \ldots, K$, sob a restrição que $det(A_k) = 1$.

Se $A_k$ é uma matriz-diagonal, então $j$-ésimo elemento da diagonal representa o peso da $j$-ésima variável no $k$-ésimo grupo. Assim, teremos distâncias adaptativas locais sob a restrição de que o produto dos pesos em cada grupo deve ser igual a 1. A principal ideia desta abordagem é derivada da Proposição 2.1, em que uma função Kernel pode ser reescrita como a soma de funções Kernel aplicadas a cada variável (FERREIRA; CARVALHO, 2014; SCHOLKOPF; SMOLA, 2001).

**Proposição 2.1** *Se* $\mathscr{K}_1 : X_1 \times X_1 \to \mathbb{R}$ *e* $\mathscr{K}_2 : X_2 \times X_2 \to \mathbb{R}$ *são funções Kernel, então a soma,* $\mathscr{K}(\mathbf{x}_1, \mathbf{x}_1') + \mathscr{K}(\mathbf{x}_2, \mathbf{x}_2')$ *é uma função Kernel definida em* $(X_1 \times X_1) \times (X_2 \times X_2)$*, onde* $\mathbf{x}_1, \mathbf{x}_1' \in X_1$ *e* $\mathbf{x}_2, \mathbf{x}_2' \in X_2, X_1, X_2 \subset \mathbb{R}^p$.

*Demonstração:* A demonstração pode ser obtida em Scholkopf e Smola (2001).

Desta forma, se um dado é representado por um vetor com $p$-variáveis podemos particioná-lo em até $p$ partes, e considerar $p$ diferentes funções Kernel, uma para cada parte. Formalmente, temos que $\mathscr{K}(\mathbf{x}_i, \mathbf{x}_k) = \sum_{j=1}^{p} \mathscr{K}_j(x_{ij}, x_{kj})$, onde $\mathscr{K}_j : X_j \times X_j \to \mathbb{R}$ são funções Kernel e $X_j$ é o espaço da $j$-ésima variável com $j = 1, \ldots, p$. Desta forma, a distância baseada na Kernelização da Métrica entre um dado $\mathbf{x}_i$ e o protótipo $\mathbf{v}_k$ com relação à $j$-ésima variável (MÜLLER *et al.*, 2001; SCHÖLKOPF; SMOLA; MÜLLER, 1998) é definida por

$$||\phi_j(x_{ij}) - \phi_j(v_{kj})||^2 = \mathscr{K}_j(x_{ij}, x_{ij}) - 2\mathscr{K}_j(x_{ij}, v_{kj}) + \mathscr{K}_j(v_{kj}, v_{kj}), \tag{2.10}$$

em que $\phi_j = \phi_1, \phi_2, \ldots, \phi_p$ são mapeamentos não lineares de $\mathbf{x}_i \in X$, $X \subset \mathbb{R}^p$ em um espaço de características $\mathscr{F}_j$.

A partir da Eq. (2.10) é possível introduzir pesos representando a relevância de cada variável. Seja $\varphi^2(\mathbf{x}_i, \mathbf{v}_k)$ uma medida de distância baseada em Kernelização da Métrica entre uma observação $\mathbf{x}_i$ e um protótipo $\mathbf{v}_k$ do $k$-ésimo grupo. Então, a distância adaptativa local com a restrição de que o produto dos pesos das variáveis em cada grupo seja igual a 1, é dada por (FERREIRA; CARVALHO; SIMÕES, 2016)

$$\varphi^2(\mathbf{x}_i, \mathbf{v}_k) = \varphi^2_{\lambda_k}(\mathbf{x}_i, \mathbf{v}_k) = \sum_{j=1}^{p} \lambda_{kj} ||\phi_j(x_{ij}) - \phi_j(v_{kj})||^2, \tag{2.11}$$

no qual $\boldsymbol{\lambda}_k = \left(\lambda_{k1}, \ldots, \lambda_{kp}\right)$ sujeito à

$$\begin{cases} \lambda_{kj} > 0, & \forall i, j, \\ \prod_{j=1}^{p} \lambda_{kj} = 1, & \forall k, \end{cases}$$

é o vetor de pesos para o $k$-ésimo grupo. A partir das Equações (2.10) e (2.11) podemos definir uma função objetivo $J$ que mede o ajuste entre os grupos e seus protótipos, dada por

$$\begin{aligned} J &= \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ki})^m \varphi^2(\mathbf{x}_i, \mathbf{v}_k) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ki})^m \sum_{j=1}^{p} \lambda_{kj} ||\phi_j(x_{ij}) - \phi_j(v_{kj})||^2, \end{aligned} \tag{2.12}$$

sob as restrições definidas na Eq. (2.6), em que $u_{ki}$ é o grau de pertinência *fuzzy* para a observação $i$ no $k$-ésimo grupo $k = 1, \ldots, K$, $i = 1, \ldots, n$ e $\mathbf{v}_k \in \mathbb{R}^p$ é prótotipo do $k$-ésimo grupo.
Ao considerarmos o Kernel Gaussiano a função objetivo descrita na Eq. (2.12) é reescrita como

$$J = 2 \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ki})^m \sum_{j=1}^{p} \lambda_{kj} (1 - \mathscr{K}(x_{ij}, v_{kj})). \tag{2.13}$$

Na derivação dos protótipos dos grupos, os graus de pertinências *fuzzy* e os pesos das variáveis são mantidos fixos. Então, o protótipo do $k$-ésimo grupo $\mathbf{v}_k = (v_{k1}, \ldots, v_{kp})$ $(k = 1, \ldots, K)$ que minimiza o critério $J$ na Eq. (2.13), tem seus componentes $v_{kj}$ $(j = 1, \ldots, p)$ definidos pela seguinte equação

$$v_{kj}^{(t+1)} = \frac{\sum_{i=1}^{n} (u_{ki}^{(t+1)})^m \mathscr{K}_j(x_{ij}, v_{kj}^{(t)}) x_{ij}}{\sum_{i=1}^{n} (u_{ki}^{(t+1)})^m \mathscr{K}_j(x_{ij}, v_{kj}^{(t)})}. \tag{2.14}$$

em que, $t = 1, \ldots, T$ onde $T$ é o número máximo de iterações. Para a determinação dos pesos, os graus de pertinências *fuzzy* $u_{ki}$ e os protótipos dos grupos $\mathbf{v}_k$ são mantidos fixos. Então, o vetor de pesos $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kp})$ que minimiza o critério $J$, sob $\lambda_{kj} > 0$, $\forall_{kj}$ e $\prod_{j=1}^{p} \lambda_{kj} = 1$, $\forall_k$, tem seus componentes $\lambda_{kj}$ $(j = 1, \ldots, p)$ e $(k = 1, \ldots, K)$ determinados por

$$\lambda_{kj}^{(t+1)} = \frac{\prod_{l=1}^{p} \{\sum_{i=1}^{n} (u_{ki}^{(t+1)})^m ||\phi(x_{il}) - \phi(v_{kl}^{(t+1)})||^2\}^{\frac{1}{p}}}{\sum_{i=1}^{n} (u_{ki}^{(t+1)})^m ||\phi(x_{ij}) - \phi(v_{kj}^{(t+1)})||^2}. \tag{2.15}$$

No cálculo dos graus de pertinências, os protótipos dos grupos $\mathbf{v}_k$ e os pesos das variáveis são mantidos fixos. Desta forma, os graus de pertinências que minimizam a função objetivo $J$ dada na Eq. (2.6), são atualizados por

$$u_{ki}^{(t+1)} = \left[ \sum_{h=1}^{K} \left( \frac{\varphi^2(\mathbf{x}_i, \mathbf{v}_k^{(t+1)})}{\varphi^2(\mathbf{x}_i, \mathbf{v}_h^{(t+1)})} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{2.16}$$

o termo $\varphi^2(\mathbf{x}_i, \mathbf{v}_k)$ é definido na Eq. (2.11). No Algoritmo 1 é possível visualizar os passos para realizar o agrupamento utilizando o método VKFCM-K-LP. Vale ressaltar que as propriedades de convergência do método foram demonstradas no trabalho de Ferreira e Carvalho (2014).

**Algoritmo 1:** Método de agrupamento VKFCM-K-LP

1: Inicialização

   Fixe $K$ (número de grupos), $2 \leq K < n$; fixe $m$, $1 < m < \infty$; fixe $T$ (número máximo de iterações); e fixe $\varepsilon$, $0 < \varepsilon < 1$. Inicialize aleatoriamente os graus de pertinências *fuzzy* $u_{ki}$ sob as restrições dadas na Eq. (2.6); Inicialize todos os pesos uniformemente com $1/p$. Faça $t = 1$.

2: Atualize o vetor de protótipos de $\mathbf{v}_k$, de acordo com a Eq. (2.14).

3: Atualize o vetor de pesos $\boldsymbol{\lambda}_k$ de acordo com a Eq. (2.15).

4: Atualize os graus de pertinências $u_{ki}$ dada na Eq.( 2.16).

5: SE $|J^{t+1} - J^t| \leq \varepsilon$ ou $t > T$

   PARE

   SE NÃO faça $t = t + 1$ e volte ao passo (2).

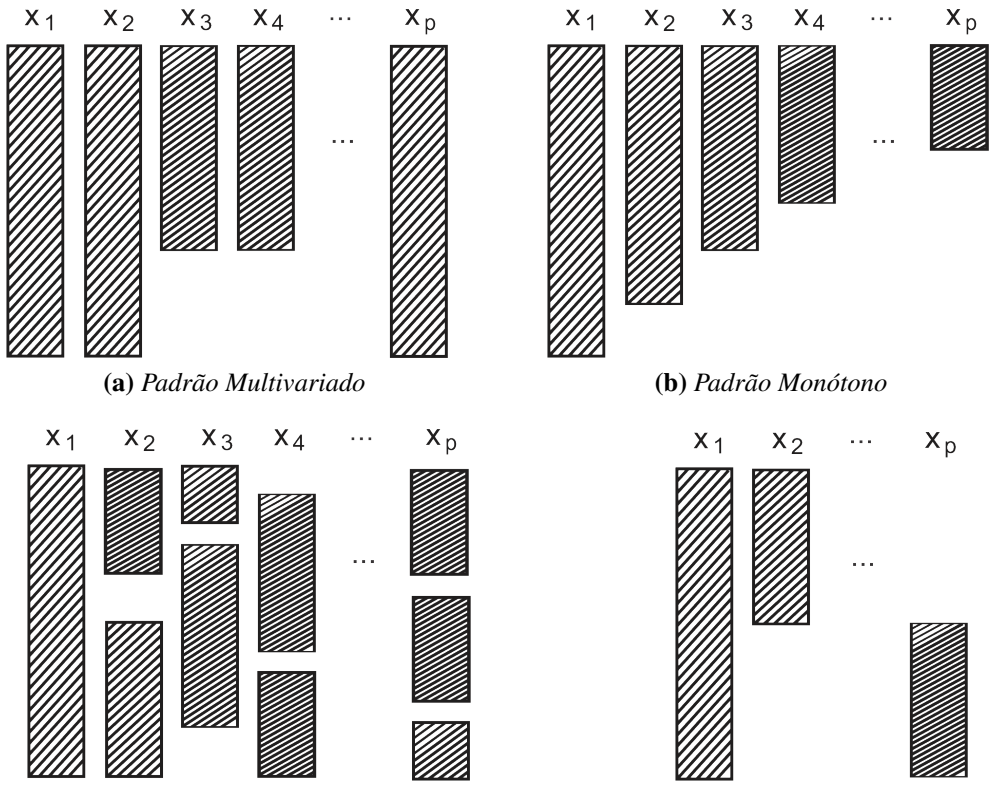
## 2.3 ANÁLISE DE DADOS INCOMPLETOS

A qualidade dos dados é um dos fatores mais importantes que podem afetar os resultados das análises estatísticas. Problemas durante a coleta de dados ou no pré-processamento, podem gerar valores incertos, incorretos ou até mesmo ausentes. A análises de dados com valores faltantes é um problema comumente discutido em todas as áreas da ciências pois, estas análises foram originalmente projetadas para conjuntos de dados sem valores ausentes. Embora as razões para a falta de dados sejam diversas, na literatura existem poucos padrões de dados ausentes resultantes dos valores ausentes nos conjuntos de dados. O padrão de dados ausentes descreve quais valores são observados e quais valores estão em falta no conjunto de dados (LITTLE; RUBIN, 2014).

De modo geral, os padrões de dados ausentes mais comuns são, o padrão multivariado, monótono, geral e o *file-matching* (LITTLE; RUBIN, 2014). No padrão multivariado (Figura 1a), os valores ausentes ocorrem em um grupo de atributos que são completamente observados ou ausentes. O padrão monótono (Figura 1b) geralmente ocorre nos resultados de estudos longitudinais e têm uma disposição dos valores em forma de escada quando organizados em uma matriz de dados. O padrão *file-matching* (Figura 1d) acontece quando os dados são obtidos a partir de várias fontes diferentes, consequentemente, o conjunto de dados combinados terá atributos completamente observados e atributos que nunca são observados juntos. O padrão geral (Figura 1c) de dados ausentes se caracteriza pela forma arbitrária no conjunto de dados e pode ser observado na prática por exemplo, na omissão de respostas em um questionário ou em perdas de dados no pré-processamento.

Enquanto os padrões de dados ausentes descrevem quais valores estão em falta no conjunto de dados, os mecanismos de geração dados faltantes fornecem informações sobre a ocorrência. Os mecanismos de geração de dados ausentes referem-se à relação entre o *missing* e os valores de atributo das variáveis no conjunto de dados. Enquanto os padrões de dados ausentes indicam quais valores de dados podem ser usados para as análise estatísticas, os mecanismos fornecem uma indicação de como os valores disponíveis devem ser tratados durante a análise de dados para obter os melhores resultados.

Os primeiros trabalhos que abordaram os mecanismos de geração de dados faltantes, foram propostos por Rubin (1976). Nestes trabalhos os autores criaram um sistema de classificação de dados ausentes utilizados até hoje, denominados como: Missing Completamente




**(a)** *Padrão Multivariado* **(b)** *Padrão Monótono*

**(c)** *Padrão Geral* **(d)** *Padrão* file-matching

**Figura 1 – Gráficos dos tipos de padrão de *missings*.**

Aleatório (MCA), Missing não Aleatório (MNA) e Missing Aleatório (MA). Estes mecanismos descrevem a relação entre as variáveis analisadas e a porcentagem de valores faltantes (BARALDI; ENDERS, 2010; RUBIN, 2004) na matriz de dados. Neste trabalho, nos concentramos em estratégias para lidar com dados faltantes do tipo MCA.

Little e Rubin (2014) definem o mecanismo de dados faltantes atráves da probabilidade de que um valor esteja disponível ou em falta no conjunto de dados. Seja $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, uma matriz de dados, definimos o vetor $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ $p$-dimensional, para $1 \leq i \leq n$ e $1 \leq j \leq p$, onde $x_{ij}$ é o $i$-ésimo valor da observação $i$ na $j$-ésima variável. Podemos reescrever $X$ como $X = X_{obs} \cup X_M$, onde $X_{obs} = \{x_{ij}\}$, se este valor for observado em $X$, e $X_M = \{x_{ij} = NA\}$ se este valor for ausente em $X$. Neste contexto, definimos uma matriz indicadora de *missings* $\mathbf{M} = [m_{ij}]$ que indica se o valor da observação $x_{ij}$ é faltante $m_{ij} = 1$ ou se $x_{ij}$ é observado, $m_{ij} = 0$. O mecanismo de geração de dados faltantes é definido como a probabilidade condicional de $\mathbf{M}$ dado $X$, $Pr(\mathbf{M}|X, \theta)$, onde $\theta$ denota os parâmetros desconhecidos de uma determinada distribuição de probabilidade. Dessa forma, os valores ausentes são definidos como MCA, se o Missing não depende do conjunto de dados, independentemente desses valores estarem ausentes

ou serem observados (RUBIN, 1976). Formalmente, este mecanismo é definido como:

$$Pr(\mathbf{M}|X,\theta) = Pr(\mathbf{M}|\theta), \text{ para todo } x_{ij} \in X,\theta. \tag{2.17}$$

De uma perspectiva prática, os mecanismos de dados faltantes operam como suposições que ditam quais técnicas devem ser utilizadas para lidar com estes valores (BARALDI; ENDERS, 2010).

### 2.3.1 Técnicas para lidar com Valores Faltantes

Tradicionalmente, os pesquisadores usam uma grande variedade de técnicas para lidar com valores faltantes. Entretanto, o melhor método seria evitar esses valores nos dados, através de um melhor mapeamento do estudo ou a repetição da coleta dos dados. No entanto, a investigação e a repetição das etapas quando os valores ausentes ocorrem se torna inviável ou impossível. Logo, existe a necessidade de se manipular técnicas que lidam com valores faltantes na matriz de dados. Na literatura existe três abordagens comuns para manipulação de valores faltantes (LITTLE; RUBIN, 2014):

- **Exclusão**: Esta técnica é melhor empregada quando a porcentagem de *missings* no banco de dados é relativamente pequena. A abordagem consiste em ignorar os itens de dados ausentes ou os padrões que contém esses valores. Assim, a análise dos dados é realizada sobre o conjunto de dados disponíveis, denominada como Análise de Caso Completo (ACC). A principal vantagem da exclusão é que ela produz um conjunto de dados completos, que por sua vez permite o uso de técnicas de análise de dados padrão (BARALDI; ENDERS, 2010). Como desvantagem temos que o tamanho da amostra é reduzido drasticamente, especialmente para conjuntos de dados que incluem uma grande proporção de dados ausentes.
- **Imputação**: Esta abordagem é bastante comum, e consiste em substituir os valores ausentes por valores estimados que geralmente são derivados dos dados disponíveis, definida como Imputação de Valores Ausentes (IVA). As técnicas de IVA variam de simples como substituir os *missings* pela Média ou Mediana ou mais sofisticadas que utilizam Regressão, Máxima Verossimilhança, etc (RUBIN, 2004). A desvantagem dessa abordagem é que a qualidade dos resultados da análise de dados é afetada pelo método de imputação utilizado, já que os valores imputados são tratados como valores observados.

Como vantagem temos que as técnicas de análise padrão podem ser empregadas já que os valores ausentes foram preenchidos.

- **Adaptação de métodos padrão para dados incompletos**: Uma abordagem eficaz é adaptar os métodos de análise de dados de forma que eles possam manipular conjuntos de dados que possuem valores faltantes. Esses métodos incluem a estimação dos valores ausentes durante a análise de dados e a distinção entre os valores observados e imputados. A principal vantagem da abordagem de adaptação é que todos os dados observados podem ser usados para a análise de dados, evitando as desvantagens da imputação do valor ausente.

## 2.4 VKFCM-K-LP SOB A ABORDAGEM DE DADOS FALTANTES

O método de agrupamento VKFCM-K-LP proposto por Ferreira e Carvalho (2014) não pode ser aplicado diretamente em um conjuntos de dados com valores ausentes. Assim como os métodos clássicos de agrupamento, este método também exige que todos os valores na matriz de dados estejam presentes para o cálculo dos protótipos e das medidas de distâncias. Com o objetivo de lidar com dados incompletos vários métodos foram propostos na literatura. Hathaway e Bezdek (2001) definiram estratégias para agrupar dados incompletos através do algoritmo Fuzzy *C-Means* (FCM). Nesta seção, descrevemos três estratégias propostas por Hathaway e Bezdek (2001) para adaptar o algoritmo de agrupamento VKFCM-K-LP a dados incompletos.

### 2.4.1 Estratégia de dados Completos (EDC)

Esta estratégia consiste em omitir os itens de dados incompletos e aplicar o algoritmo VKFCM-K-LP na matriz de dados completa (HATHAWAY; BEZDEK, 2001). Este método pode ser considerado como uma ACC, já que os valores ausentes não estão incluídos no cálculo dos protótipos dos grupos. Esta estratégia pode ser aplicada ao agrupamento de dados incompletos desde que a porcentagem desses valores seja relativamente pequena. Hathaway e Bezdek (2001) sugerem uma porcentagem menor que 25% de todos os valores completos no conjunto de dados.

Como os protótipos são calculados a partir do conjunto de dados completos, esta estratégia é recomendada se a porcentagem de valores observados forem representativos para todo o conjunto de dados. No entanto, as observações incompletas não são ignoradas totalmente da análise, ao final do agrupamento com conjunto de dados completos os dados incompletos são particionados utilizando o esquema do protótipo mais próximo com base nas distâncias parciais

cada dado incompleto para cada um dos protótipos dos grupos computados. O Algoritmo (2) descreve os passos para a estratégia EDC.

**Algoritmo 2:** Método de agrupamento VKFCM-K-LP sob a estratégia EDC.

1: Inicialização

Fixe $K$ (número de grupos), $2 \leq K < n$; fixe $m$, $1 < m < \infty$; fixe $T$ (número máximo de iterações); e fixe $\varepsilon$, $0 < \varepsilon < 1$. Inicialize aleatoriamente os graus de pertinências *fuzzy* $u_{ki}$; Inicialize todos os pesos uniformemente com $1/p$. Faça $t = 1$.

2: Atualize o vetor de protótipos de $\mathbf{v}_k$, de acordo com a Eq. (2.14).

3: Atualize o vetor de pesos $\boldsymbol{\lambda}_k$ de acordo com a Eq. (2.15).

4: Atualize os graus de pertinências $u_{ki}$ dada na Eq.( 2.16).

5: SE $|J^{t+1} - J^t| \leq \varepsilon$ ou $t > T$

**Particione $X_M$ de acordo com a Eq.(2.18)**

PARE

SE NÃO faça $t = t + 1$ e volte ao passo (2).

### 2.4.2 Estratégia da Distância Parcial (EDP)

Esta estratégia é recomendada por Dixon (1979) quando $X_M$ é suficientemente grande, de modo que a EDC não possa ser utilizada. A proposta consiste em estimar a distância entre duas observações utilizando a função da Distância Parcial (DP). A função da DP calcula a soma das distâncias euclidianas quadradas entre todos os valores de recursos disponíveis e, em seguida, os pondera pela a proporção de valores utilizados no seu cálculo. Ao adaptar o algoritmo VKFCM-K-LP, no qual se utiliza a distância Kernel adaptativa local, temos que sua versão parcial é dada por

$$\varphi_{dp}^2(\mathbf{x}_i, \mathbf{v}_k) = \frac{p}{I_i} \sum_{j=1}^{p} \lambda_{kj} ||\phi(x_{ij}) - \phi(v_{kj})||^2 I_{ij}, \tag{2.18}$$

no qual, $I_i = \sum_{j=1}^{p} I_{ij}$ para $1 \leq i \leq n$ e $1 \leq j \leq p$. A função indicadora $I_{ij}$ é definida por

$$I_{ij} = \begin{cases} 1, & \text{se } x_{ij} \in X_{obs}, \\ 0, & \text{se } x_{ij} \in X_M. \end{cases} \tag{2.19}$$

em que, $X_{obs}$ e $X_M$ são definidos na Sessão 2.3. Assim, a função objetivo para esta estratégia é definida por

$$J_{dp}(\mathbf{V},\mathbf{U},\mathbf{\Lambda}) = \sum_{k=1}^{K}\sum_{i=1}^{n}(u_{ki})^m \varphi_{dp}^2(\mathbf{x}_i,\mathbf{v}_k), \tag{2.20}$$

em que, $\mathbf{V} = \{\mathbf{v_1},\ldots,\mathbf{v}_K\} \in \mathbb{R}^{K\times p}$, $\mathbf{\Lambda} = \{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_K\} \in \mathbb{R}_+^{K\times p}$, e $\varphi_{dp}^2(\mathbf{x}_i,\mathbf{v}_k)$ é definida na Eq. (2.18) e nomeada como Distância Kernel Parcial Adaptativa local na qual se utiliza a restrição de que o produto dos pesos das variáveis em cada grupo seja igual a 1. Na primeira iteração do algoritmo, os graus de pertinências são atualizados de acordo com a Eq. (2.18). Na segunda interação do algoritmo VKFCM-K-LP os protótipos e os pesos são atualizados apenas pelos valores em $X_{obs}$, definidos por

$$v_{kj}^{(t+1)} = \frac{\sum_{k=1}^{n}(u_{ik}^{(t+1)})^m \mathscr{K}(x_{ij},v_{kj}^{(t)})x_{ij}I_{ij}}{\sum_{i=1}^{n}(u_{ki}^{(t+1)})^m \mathscr{K}(x_{ij},v_{kj}^{(t)})I_{ij}}, \tag{2.21}$$

em que, o termo $\mathscr{K}(.)$ é o Kernel Gaussiano. Os pesos das variáveis são obtidos de acordo com a Eq.(2.18).

$$\lambda_{kj}^{(t+1)} = \frac{\prod_{l=1}^{p}\{\sum_{i=1}^{n}(u_{ki}^{(t+1)})^m||\phi(x_{il})-\phi(v_{kl}^{(t+1)})||^2 I_{il}\}^{\frac{1}{p}}}{\sum_{i=1}^{n}(u_{ki}^{(t+1)})^m||\phi(x_{ij})-\phi(v_{kj}^{(t+1)})||^2 I_{ij}}, \tag{2.22}$$

para $1 \leq k \leq K$ e $1 \leq l \leq p$.

Segundo Hathaway e Bezdek (2001) o fator de escala $p/I_i$ na Eq. (2.18) não produz efeito sobre o cálculo dos protótipos na Eq. (2.21) e consequentemente no cálculo dos pesos na Eq.(2.22). Este fator de escala também não produz efeito sobre $u_{ki}$, pois aparece tanto na parte superior como na inferior da equação podendo ser omitido da Eq. (2.18).

$$u_{ki}^{(t+1)} = \left[\sum_{h=1}^{K}\left(\frac{\varphi_{dp}^2(\mathbf{x}_i,\mathbf{v}_k^{(t)})}{\varphi_{dp}^2(\mathbf{x}_i,\mathbf{v}_h^{(t)})}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{2.23}$$

O algoritmo para a estratégia EDP é obtido por meio de três modificações no VKFCM-K-LP (1) básico. Estas modificações estão descritas no Algoritmo 3.

**Algoritmo 3:** Método de agrupamento VKFCM-K-LP sob a estratégia EDP.

1: Inicialização

Fixe $K$ (número de grupos), $2 \leq K < n$; fixe $m$, $1 < m < \infty$; fixe $T$ (número máximo de iterações); e fixe $\varepsilon$, $0 < \varepsilon < 1$. Inicialize aleatoriamente os graus de pertinências *fuzzy* $u_{ki}$; Inicialize todos os pesos uniformemente com $1/p$. Faça $t = 1$.

2: Atualize o vetor de protótipos de $\mathbf{v}_k$, de acordo com a **Eq. (2.21)**.

3: Atualize o vetor de pesos $\boldsymbol{\lambda}_k$ de acordo com a **Eq. (2.22)**.

4: Atualize os graus de pertinências $u_{ki}$ dada na **Eq.( 2.23)**.

5: SE $|J^{t+1} - J^t| \leq \varepsilon$ ou $t > T$

PARE

SE NÃO faça $t = t + 1$ e volte ao passo (2).

### 2.4.3 Estratégia de Conclusão Ótima (ECO)

A ideia principal desta estratégia é calcular iterativamente os valores ausentes em $X_M$, como variáveis auxiliares no processo de otimização da função objetivo $J_M$ (HATHAWAY; BEZDEK, 2001) definida na Eq. (2.24).

$$J_M(\mathbf{V}, \mathbf{U}, \boldsymbol{\Lambda}, X_M) = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ki})^m \varphi^2_{\boldsymbol{\lambda}_k}(\mathbf{x}_i, \mathbf{v}_k), \tag{2.24}$$

em que,

$$\varphi^2(\mathbf{x}_i, \mathbf{v}_k) = \sum_{j=1}^{p} \lambda_{kj} ||\phi(x_{ij}) - \phi(v_{kj})||^2 = 2 \sum_{j=1}^{p} \lambda_{kj} (1 - \mathscr{K}(x_{ij}, v_{kj})). \tag{2.25}$$

os protótipos $v_{kj}$ e os pesos $\lambda_{kj}$ são definidos conforme as Eqs. (2.14) (2.15). Dessa forma, os valores ausentes são atualizados por (HATHAWAY; BEZDEK, 2001)

$$X_M^{(t+1)} = \underset{X_M}{arg\ min} \{J_M(\mathbf{U}^{(t+1)}, \mathbf{V}^{(t+1)}, \boldsymbol{\Lambda}^{(t+1)}, X_M^{(t)})\}, \tag{2.26}$$

para todo $x_{ij} \in X_M$ obtido pela minimização da Eq. (2.26)

$$x_{ij}^{(t+1)} = \frac{\sum_{k=1}^{K} (u_{ki}^{(t+1)})^m v_{kj}^{(t+1)}}{\sum_{k=1}^{K} (u_{ki}^{(t+1)})^m}, \tag{2.27}$$

em que, as pertinências $u_{ki}$ são definadas na Eq. (2.16), em que, $1 \leq i \leq n$ e $1 \leq j \leq p$. Na estratégia ECO os valores ausentes são imputados pelas médias ponderadas de todos os protótipos dos grupos em cada etapa da iteração. Além disso, os valores ausentes $X_M$ são inicializados por valores aleatórios. A fórmula na Eq. (2.27) é obtida atráves das derivadas parciais da função objetivo dada na Eq. (2.24) mantendo os prótotipos, pesos e pertinências fixas. O algoritmo 4 descreve os passos do método VKFCM-K-LP sob a abordagem ECO. A vantagem desta abordagem é que os valores ausentes são imputados durante o processo de agrupamento.

**Algoritmo 4:** Método de agrupamento VKFCM-K-LP sob a abordagem ECO.

1: Inicialização

Fixe $K$ (número de grupos), $2 \leq K < n$; fixe $m$, $1 < m < \infty$; fixe $T$ (número máximo de iterações); e fixe $\varepsilon$, $0 < \varepsilon < 1$. Inicialize aleatoriamente $X_M$; Inicialize os graus de pertinências *fuzzy* $u_{ki}$ sob as restrições dadas em (2.6); Inicialize todos os pesos uniformemente com $1/p$; Faça $t = 1$.

2: Atualize o vetor de protótipos de $\mathbf{v}_k$, de acordo com a Eq. (2.14).

3: Atualize o vetor de pesos $\boldsymbol{\lambda}_k$ de acordo com a Eq. (2.15).

4: Atualize os graus de pertinências $u_{ki}$ de acordo com a Eq. (2.16).

5: **Para todo $x_{ij} \in X_M$ atualize de acordo com a Eq. (2.27)**

6: SE $|J^{t+1} - J^t| \leq \varepsilon$ ou $t > T$

PARE

SE NÃO faça $t = t + 1$ e volte ao passo (2).

### 2.4.4 Complexidade computacional dos métodos

A complexidade computacional do VKFCM-K-LP padrão (FERREIRA; CARVALHO, 2014) é $O(K^2np)$, em que, $n$ é o número de observações, $K$ é o número de grupos e $p$ é o número de variáveis. Para a estratégia EDC, em qual se excluí os valores faltantes, e em seguida realiza o agrupamento do VKFCM-K-LP padrão no conjunto dos valores completos em $X_{obs}$, temos que a complexidade computacional é dada por $O(K^2|X_{obs}|p)$, em que, $|X_{obs}|$ é a cardinalidade de $X_{obs}$. Na estratégia EDP que usa a soma das distâncias parciais euclidianas quadradadas na Eq. (2.11) para o cálculo dos prótipos, pesos e da matriz de partição *fuzzy*, a complexidade computacional não é aumentada, logo esta estratégia tem complexidade definada por $O(K^2np)$. Na estratégia ECO, o cálculo da matriz de partição *fuzzy* requer $K^2np$ operações, e os cálculos de protótipos dos grupos e valores faltantes requerem operações $Knp$. Assim, a complexidade computacional da estratégia ECO é dada por $O(K^2np)$.

# 3 DESENHO EXPERIMENTAL

Ferreira e Carvalho (2014) propuseram vários métodos de agrupamento baseados em Kernel com podenração automática das variáveis. Dentre estes métodos encontra-se o VKFCM-K-LP, no qual utiliza a distância adaptativa local kernelizada. Este método apresentou um bom desempenho em detectar classes *a priori*. No entanto, seu desempenho ainda não foi avaliado no contexto de dados incompletos. Assim, este trabalho se propõe a utilizar o método VKFCM-K-LP sob três abordagens de dados incompletos definidas por Hathaway e Bezdek (2001). Para a avaliação do método, houve a necessidade de implementar um gerador de valores faltantes, a fim de criar conjuntos de dados com valores faltantes sobre o qual os métodos presentes neste trabalho serão avaliados. A implementação do mecanismo de geração de dados faltantes nas bases de dados de teste e as representações gráficas foram realizadas com o auxílio dos pacotes presentes no *software* R (TEAM *et al.*, 2019). Os principais pacotes do R utilizados foram o `ggplot2, VIM` e `naniar`. Após a geração dos conjuntos de dados com *misings*, os métodos de agrupamento foram implementados por meio de programas em linguagem C. As avaliações foram realizadas no computador com um processador *Intel Core (TM) I3-3217U CPU1.80GHz* e memória *RAM 4GB* usando o sistema operacional Linux.

## 3.1 GERAÇÃO DOS MISSINGS

O gerador de valores faltantes (*misings*) utilizado nesse estudo, remove os valores do conjunto de dados completos com uma determinada probabilidade, de acordo com o mecanismo de dados faltantes MCA. Na geração dos valores ausentes do tipo MCA, assumimos independência entre a distribuição conjunta de $(\mathbf{x}_i, \mathbf{M})$, portanto, a probabilidade de que um valor $x_{ij}$ seja observado, independe dos valores em $X$ ou $\mathbf{M}$ (LITTLE; RUBIN, 2014). Ao considerar a distribuição de Bernoulli com parâmetro $\theta$, $0 \leq \theta \leq 1$, para a variável indicadora $\mathbf{M}_i$, com probabilidade $Pr(\mathbf{M}_i = 1|x_i, \theta)$ dado que $x_i$ é um valor faltante. Temos que se os valores faltantes são independentes de $X$, então $Pr(\mathbf{M}_i = 1|x_i, \theta) = \theta$, como a constante independe dos valores em $X$, temos a geração do mecanismo do tipo MCA.

Em termos computacionais um conjunto de dados completos $X$ foi selecionado, e posteriormente modificado para se obter um conjunto de dados incompletos, selecionando aleatoriamente uma porcentagem especificada de seus componentes $\{x_{ij}\}$, designados como *missings*. Os valores $\{x_{ij}\}$ foram tidos como *missings*, quando o elemento $m_{ij}$ da amostra gerada

para variável indicadora **M**, foi igual $m_{ij} = 1$. Assim, o valor do elemento $x$ da posição $ij$ foi excluído do conjunto de dados completo e designado como ausente. Na geração das amostras Bernoulli, para a variável **M**, como descrito no Algoritmo 5. Em nosso estudo foi considerado o parâmetro $\theta = \{0.05, 0.10, 0.15, 0.20\}$.

**Algoritmo 5:** Geração dos *missings*.

1: Inicialização

Seja $X$ uma matriz de dados completa (sem *missings*), e seja $Z = vec(X)$ de tamanho $N$, em que $vec(.)$ é um operador de vetorização.

2: Gere $M = \{M_1, \ldots, M_N\}$, em que

$M_i \sim Bernoulli(\theta)$

3: Para $Pr(M = 1|\theta)$, ou seja $m_i = 1$

Substitua os elementos da posição {i} de $Z$ por *missings*.

## 3.2 MEDIDAS DE QUALIDADE

Na comparação dos métodos de agrupamento propostos neste trabalho, foram utilizadas as seguintes medidas de qualidade, a medida F- *measure* (FM) (BAEZA-YATES; RIBEIRO *et al.*, 2011), o Coeficiente de Rand Corrigido (CR) (HUBERT; ARABIE, 1985), e a Taxa Total de Erro de Classificação (OERC) (BREIMAN *et al.*, 1984). Essas medidas são obtidas a partir de uma Matriz de Confusão definida na Tabela 1. Outra medida abordada nesse trabalho, foi a Consistência das variáveis definida por (ANTAL; SZABÓ, 2017; LEE; BERGER; AVICZER, 1996). Sejam $\mathscr{P} = \{\mathscr{P}_1, \ldots, \mathscr{P}_c, \ldots, \mathscr{P}_C\}$ a partição *a priori* de $\Omega = \{1, 2, \ldots, n\}$, em $C$ classes e seja $P = \{P_1, \ldots, P_k, \ldots, P_K\}$ a partição rígida de $\Omega = \{1, 2, \ldots, n\}$ fornecida pelo método de agrupamento, em $K$ grupos, em que $\cup_{k=1}^{k} P_k = \Omega$ e $\cap_{k=1}^{k} P_k = \emptyset$. No algoritmo VKFCM-K-LP, a partição rígida foi obtida a partir da matriz de partição *fuzzy*, alocando cada observação $\mathbf{x}_i$ a um grupo rígido, fornecido pelo método de agrupamento, $P_k$ se

$$k = \underset{1 \leq h \leq K}{\arg\max}\, u_{hi}. \tag{3.1}$$

As quantidades $n_{ck}$, na Tabela 1 representam as observações que estão na classe $\mathscr{P}_c$ e no grupo $P_k$ para $c = 1, 2, \ldots, C$ e $k = 1, 2, \ldots, K$.

**Tabela 1 – Matriz de Confusão.**

| | Grupos | | | | |
|---|---|---|---|---|---|
| Classes | $P_1$ | $\dots$ | $P_k$ | $\dots$ | $P_K$ |
| $\mathscr{P}_1$ | $n_{11}$ | $\dots$ | $n_{1k}$ | $\dots$ | $n_{1K}$ |
| $\vdots$ | $\vdots$ | $\dots$ | $\vdots$ | $\dots$ | $\vdots$ |
| $\mathscr{P}_c$ | $n_{c1}$ | $\dots$ | $n_{ck}$ | $\dots$ | $n_{cK}$ |
| $\vdots$ | $\vdots$ | $\dots$ | $\vdots$ | $\dots$ | $\vdots$ |
| $\mathscr{P}_C$ | $n_{C1}$ | $\dots$ | $n_{Ck}$ | $\dots$ | $n_{CK}$ |

O CR mede o grau de concordância entre uma partição $\mathscr{P}_c$ *a priori* e uma partição $P_k$ fornecida pelo método de agrupamento. Esta medida não é sensível ao número de partições ou a distribuição dos padrões no grupo (HUBERT; ARABIE, 1985). A medida CR é definida no intervalo de $[-1,1]$, o valor 1 indica uma concordância perfeita entre as partições, enquanto os valores próximos a $-1$ representam uma má concordância entre as partições. Formalmente, a medida CR é definida como

$$CR = \frac{\sum_{c=1}^{C}\sum_{k=1}^{K}\binom{n_{ck}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{C}\binom{n_{c\bullet}}{2}\sum_{k=1}^{K}\binom{n_{\bullet k}}{2}}{\frac{1}{2}\left[\sum_{c=1}^{C}\binom{n_{c\bullet}}{2} + \sum_{k=1}^{K}\binom{n_{\bullet k}}{2}\right] - \binom{n}{2}^{-1}\sum_{c=1}^{C}\binom{n_{c\bullet}}{2}\sum_{k=1}^{K}\binom{n_{\bullet k}}{2}}, \tag{3.2}$$

em que, $n_{ck}$ representa o número de observações que estão na classe $\mathscr{P}_c$ e no grupo $P_k$, $n_{c\bullet} = \sum_{k=1}^{K} n_{ck}$, representa o número de observações da classe $\mathscr{P}_c$, $n_{\bullet k} = \sum_{c=1}^{C} n_{ck}$, representa o número de observaçõess no grupo $P_k$ e $n$ é o número total de observações.

A precisão entre uma classe $\mathscr{P}_c$ e um grupo $P_k$, é definida como a razão entre o número de observações na classe $\mathscr{P}_c$ e $P_k$ e o número de observações no grupo $P_k$, dada por

$$Precisão(\mathscr{P}_c, P_k) = \frac{n_{ck}}{n_{\bullet k}}. \tag{3.3}$$

A Revocação entre uma classe $\mathscr{P}_c$ e um grupo $P_k$ é definida como a razão entre o número de observaçõess que estão na classe $\mathscr{P}_c$ e no grupo $P_k$ e o número de observações da classe $\mathscr{P}_c$, dada por

$$Revocação(\mathscr{P}_c, P_k) = \frac{n_{ck}}{n_{c\bullet}}. \tag{3.4}$$

A medida $FM$ entre uma classe $\mathscr{P}_c$ e um grupo $P_k$ é a média harmônica entre a Precisão e a Revocação. Esta medida assume valores no intervalo $[0,1]$ no qual 1 indica concordância perfeita

entre as partições. E pode ser definida como

$$FM(\mathscr{P}_c, P_k) = 2\frac{Precisão(\mathscr{P}_c, P_k) \times Revocação(\mathscr{P}_c, P_k)}{Precisão(\mathscr{P}_c, P_k) + Revocação(\mathscr{P}_c, P_k)}. \tag{3.5}$$

Por fim, a medida $F$ entre a partição *a priori* $\mathscr{P}$ e partição rígida $P$ é dada por

$$FM(\mathscr{P}, P) = \frac{1}{n}\sum_{c=1}^{C} n_{c\bullet} \max_{1\leq k\leq K} FM(\mathscr{P}_c, P_k), \tag{3.6}$$

A medida OERC tem como objetivo mensurar a capacidade de um de um algoritmo de agrupamento detectar as classes *a priori* em um conjunto de dados. Esta medida de qualidade é definida no intervalo $[0,1]$ no qual valores próximos de zero indicam maior habilidade de um algoritmo de agrupamento encontrar classes *a priori*. Sua expressão é definida por

$$OERC \quad = \quad \sum_{k=1}^{K} \frac{n_{\bullet k}}{n}\left[1 - \arg\max_{1\leq c\leq C}\left(\frac{n_{ck}}{n_{\bullet k}}\right)\right]. \tag{3.7}$$

Ao final do agrupamento com o método VKFCM-K-LP sob a abordagem ECO, obtivemos um conjunto de dados completo, sobre o qual resultou os melhores valores das medidas CR, OERC e FM. Assim, a fim de verificar se os valores imputados de acordo com a ECO se assemelha ao padrão de cada variável, foram calculadas as medidas de Consistência (LEE; BERGER; AVICZER, 1996) definida por

$$d_k(j) = \frac{|\mu_{p0}(j) - \mu_{p1}(j)|}{\sqrt{\sigma^2_{p0}(j) + \sigma^2_{p1}(j)}}, \tag{3.8}$$

em que, $k$ denota o k-ésimo grupo, $1 \leq j \leq p$ e $p$ representa as variáveis a serem analisadas, $\mu_{p0}$ e $\sigma^2_{p0}$ são a média e a variância do conjunto de dados com *missings*, respectivamente, e por fim, $\mu_{p1}$ e $\sigma^2_{p1}$ referem-se à média e variância do conjunto de dados com valores imputados. Então, espera-se que quanto melhor o agrupamento com a ECO, as consistência estejam próximas a zero, o que indicará que os valores imputados não foram discrepantes em relação as escalas originais das variáveis, no banco com *missings*.

# 4 RESULTADOS

Este Capítulo apresenta uma avaliação do método de agrupamento *fuzzy* baseado em Kernel, com ponderação automática das variáveis, utilizando distâncias adaptativas local VKFCM-K-LP. O VKFCM-K-LP foi estudado sob três tipos de abordagens para dados faltantes, EDC, EDP e ECO. Nestas avaliações foram utilizados bancos com 5%, 10%, 15% e 20% de *missings*, que foram gerados através da metodologia descrita na Seção 3.1.

Os algoritmos de agrupamento aplicados aos conjuntos de dados foram executados 100 vezes atráves de um experimento de Monte Carlo, com incializações aleatórias. Em cada réplica, foi observado o ajuste entre grupos e protótipos até a sua convergência no valor de $\varepsilon = 10^{-10}$ ou $t > T$ com $T = 300$ e o melhor resultado de acordo com a função objetivo $J$ é selecionado. Em termos de comparação foram calculadas as medidas de avaliação CR, FM, OERC para as melhores soluções dos métodos de agrupamento. Também foram calculadas as médias e os desvios padrão para estas medidas nas 100 repetições de cada algoritmo. O número de grupos $K$ foi definido como sendo igual ao número de classes *a priori* $C$. O parâmetro de imprecisão $m$ foi fixado em 2.0 de acordo com o estudo de Ferreira e Carvalho (2014). Em relação aos termos $2\sigma_j^2$ das funções Kernel Gaussianas, para $\{j = 1,\ldots,p\}$, foram estimados como a média entre os quantis $0,1$ e $0,9$ de $||x_{ij} - x_{kj}||^2$ para $i \neq k$ $i,k = 1,\ldots,n$ (FERREIRA; CARVALHO, 2014; FERREIRA; CARVALHO; SIMÕES, 2016). Posteriormente, com a avaliação do método VKFCM-K-LP sob a abordagem ECO foram calculadas as consistências das variáveis nos bancos de dados completos que foram derivados deste método. Uma comparação do agrupamento com o método ECO e os agrupamentos utilizando a imputação dos *missings* via Média e Mediana, também foi realizada. Para mostrar a eficácia dos método de agrupamento VKFCM-K-LP sob as abordagens EDC, EDP e ECO foram utilizados dois conjuntos de dados *Iris Plant* (ANDERSON, 1935) e *Thyroid Gland* (QUINLAN, 1986) ambos obtidos do Repositório de Aprendizagem de Máquina da Universidade da Califórnia, Irvine, Estados Unidos (*UCI Machine Learning Repository*) (BACHE; LICHMAN, 2013). A esolha destes conjuntos de dados deve-se ao fato dos grupos apresentarem estruturas diferentes, em particular o conjunto de dados *Thyroid Gland* apresenta uma maior sobreposição dos grupos em relação ao conjunto *Iris Plant*. O desempenho dos métodos nestes conjuntos de dados estão descritos neste capítulo.

## 4.1 CONJUNTO DE DADOS *IRIS PLANT*

O conjunto de dados *Iris Plant* (ANDERSON, 1935) é amplamente estudado na área de reconhecimento de padrões. Este conjunto possui três classes *a priori* ($K = 3$) com 50 observações em cada, referentes as Espécies setosa (Classe 1), virginica (Classe 2) e versicolor (Classe 3), de um género de plantas com flor chamada íris, ou seja, temos um total de $n = 150$ observações. Para cada espécie foram observadas quatro variáveis ($p = 4$), Comprimento da Sépala (SL), Largura da Sépala (SW), Comprimento da Pétala (PL) e Largura da Pétala (PW).

Os gráficos das Figuras 2a-2b mostram a dispersão dos valores das variáveis para este conjunto de dados, e os Boxplots para cada espécie. É possível observar uma relação aparentemente linear entre as variáveis PL e SL e as variáveis PW e SW nas classes versicolor e virginica. Nota-se também que as variáveis são diretamente proporcionais se considerarmos as espécies versicolor e virginica, ou seja, nestas espécies o aumento do tamanho de SL implica no aumento no tamanho de PL, o mesmo é observado paras as medidas de largura SW e PW.

Além disso, temos que as três espécies se diferenciam em relação as variáveis, sendo a diferença mais acentuada observada na espécie setosa, quando comparada com as demais, ou seja, esta espécie é linearmente separável das outras duas. Os gráficos de Boxplots na Figura 2a evidenciam uma maior variabilidade nos dados da espécie virginica, nas variáveis SL e PL. Para a Figura 2b, os boxplots mostram uma menor variabilidade nas espécies quando consideramos a variável SW.

Os gráficos das Figuras 3a-3d, apresentam os padrões dos *missings* para o Banco *Iris Plant*, distribuídos nas quatro variáveis. No gráfico, o eixo do $x$ estão as variáveis e no eixo do $y$ encontram-se as observações, a cor preta indica um valor faltante no banco. As figuras também mostram os números de *missings* por variável, para cada porcentagem analisada, sendo a variável PL a que possui o maior número desses valores faltantes, considerando todas as porcentagens de *missings* estudadas. Para os bancos com 5%, 10% e 20% de *missings*, a variável SW possui a menor quantidade destes valores ausentes. As observações da Classe 1 estão no intervalo de $1|-50$, o intervalo de $51|-100$ constitue as observações do Classe 2 e $101|-150$ as observações para o Classe 3, visualizadas nas Figuras 3a - 3d. Além disso, ao consideramos as padrões dos *missings* no conjunto de dados *Iris Plant* observa-se o padrão Geral.

A Tabela 2 apresenta os melhores resultados dos índices CR, FM e OERC, dentre as 100 réplicas de Monte Carlo do algoritmo de agrupamento VKFCM-K-LP, em conjunto

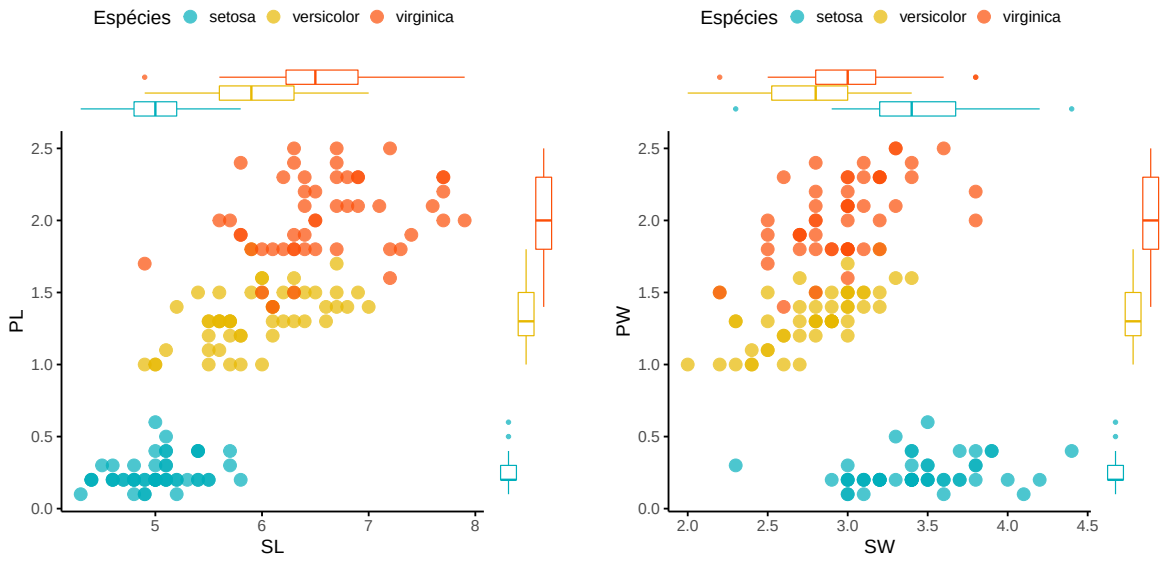


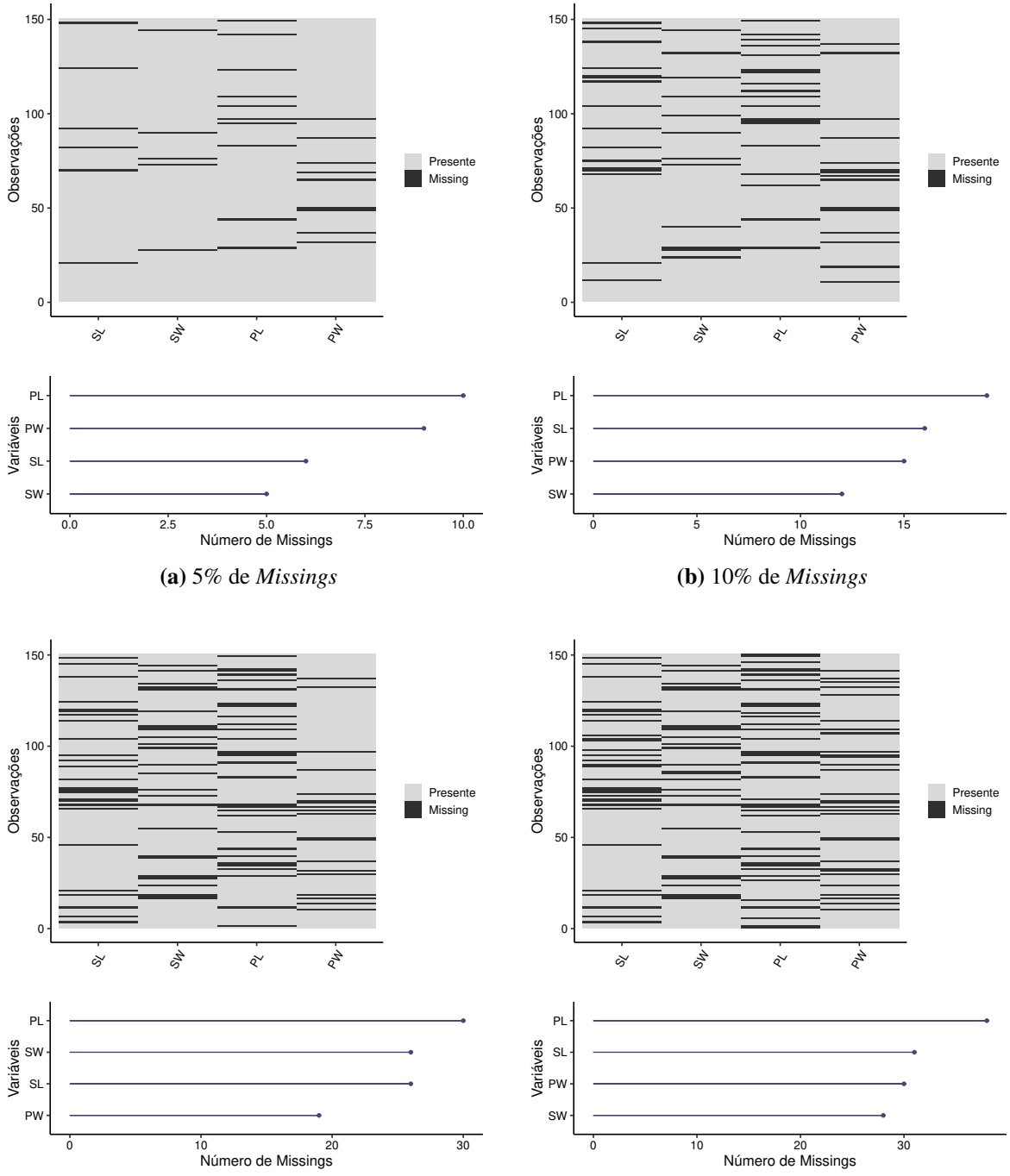

**Figura 2 – Gráficos de Dispersão e Boxplots para o conjunto de Dados Iris Plant.**

com os métodos EDC, EDP e ECO. É possível observar que para todas as porcentagens de *missings* estudadas, os índices CR e FM encontram-se próximos ao valor 1, o que indica uma boa concordância entre as classes *a priori* e os grupos fornecidos pelos métodos de agrupamento estudados. Para 5% de *missings*, no conjunto *Iris Plant*, o melhor desempenho foi obtido para o método EDP.

Entretanto, ao analisar os dados com as porcentagens de 10%, 15% e 20% de *missings*, este método possui um desempenho inferior aos demais, apresentando maiores valores do índice OERC observado, ou seja, nessa estratégia o algoritmo apresentou uma maior dificuldade de detectar as classes *a priori*. De modo geral, o aumento na porcentagem de *missings* no banco de dados piora o desempenho dos algoritmos. Este comportamento também é verificado nas porcentagens de 5% a 10% para a abordagem EDP e 15% a 20% para as abordagens EDC e ECO.

**Tabela 2 – Desempenho do algoritmo de agrupamento VKFCM-K-LP sob três tipos de abordagem estudadas EDC, EDP e ECO para o banco *Iris Plant*.**

| % | CR | | | FM | | | OERC | | |
|---|---|---|---|---|---|---|---|---|---|
| Missing | EDC | EDP | ECO | EDC | EDP | ECO | EDC | EDP | ECO |
| **5** | 0.7429 | 0.8018 | 0.7861 | 0.8991 | 0.9261 | 0.9198 | 0.1000 | 0.0733 | 0.0800 |
| **10** | 0.8016 | 0.7561 | 0.8015 | 0.9266 | 0.9065 | 0.9266 | 0.0733 | 0.0933 | 0.0733 |
| **15** | 0.8176 | 0.7561 | 0.8175 | 0.9333 | 0.9065 | 0.9333 | 0.0666 | 0.0933 | 0.0666 |
| **20** | 0.8018 | 0.7561 | 0.7859 | 0.9261 | 0.9065 | 0.9199 | 0.0733 | 0.0933 | 0.0800 |




**(a)** 5% de *Missings*

**(b)** 10% de *Missings*

**(c)** 15% de *Missings*

**(d)** 20% de *Missings*

**Figura 3 – Gráficos dos Padrões e Frequências dos valores faltantes por variável para o banco *Iris Plant*.**

Como o intuito de investigar o poder preditivo do algoritmo VKFCM-K-LP sob as três abordagens para dados faltantes, a Tabela 3 apresenta as matrizes de confusão obtidas para cada método, e em cada porcentagem de *missings* estudada. Nas colunas encontram-se as classes *a priori*, e nas linhas estão os grupos fornecidos pelos métodos de agrupamento. Os grupos fornecidos pelos métodos de agrupamento foram identificados como, Grupo 1 (setosa), Grupo 2

(virginica) e Grupo 3 (versicolor). Ao analisar as matrizes de confusão na Tabela 3, percebe-se pelos métodos de agrupamento e para todas as porcentagens de *missings* analisadas, que todas as observações pertencentes a espécie setosa no banco *Iris Plant* foram adequadamente agrupadas no Grupo 1 fornecido pelo algoritmo. Este fato é justificado, por esta espécie apresentar uma melhor separabilidade quando comparada com as demais, como mostrado nas Figuras 2a-2b. Quando observamos o Grupo 3 (Tabela 3), verificamos um menor desempenho do algoritmo, em detectar as observações que pertencem a Classe 2 da espécie virginica. Nota-se também que os Grupos 2 e 3, foram os que apresentaram um maior número de observações agrupadas incorretamente, isto acontece por estas grupos não serem linearmente separáveis como observado no Grupo 1.

**Tabela 3 – Matrizes de confusão obtidas pelo algoritmo VKFCM-K-LP em conjunto com os métodos EDC, EDP e ECO utilizando 5, 10, 15 e 20% de dados faltantes.**

| Métodos | Grupos | 5% | | | 10% | | | 15% | | | 20% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| EDC | 1 | 50 | 0 | 0 | 50 | 0 | 0 | 50 | 0 | 0 | 50 | 0 | 0 |
| | 2 | 0 | 47 | 12 | 0 | 46 | 7 | 0 | 44 | 4 | 0 | 42 | 3 |
| | 3 | 0 | 3 | 38 | 0 | 4 | 43 | 0 | 6 | 46 | 0 | 8 | 47 |
| EDP | 1 | 50 | 0 | 0 | 50 | 0 | 0 | 50 | 0 | 0 | 50 | 0 | 0 |
| | 2 | 0 | 47 | 8 | 0 | 45 | 9 | 0 | 45 | 9 | 0 | 45 | 9 |
| | 3 | 0 | 3 | 42 | 0 | 5 | 41 | 0 | 5 | 41 | 0 | 5 | 41 |
| ECO | 1 | 50 | 0 | 0 | 50 | 0 | 0 | 50 | 0 | 0 | 50 | 0 | 0 |
| | 2 | 0 | 46 | 8 | 0 | 45 | 6 | 0 | 45 | 5 | 0 | 45 | 7 |
| | 3 | 0 | 4 | 42 | 0 | 5 | 44 | 0 | 5 | 45 | 0 | 5 | 43 |

As Tabelas 4, 5 e 6 fornecem os pesos das variáveis em cada grupo. De modo geral, observa-se que nos três métodos abordados e para todas as porcentagens de *missings* avaliadas, as variáveis PL e PW foram as mais relevantes para a construção dos grupos. A variável PL obteve a maior relevância em todos os grupos, mesmo apresentando o maior número de valores faltantes, como mostrado nas Figuras 3a-3d. Em contrapartida, observa-se um decréscimo nos pesos da variável PL com o aumento da porcentagem de *missings* no Grupo 2 para os métodos EDP e ECO, este comportamento também é verificado para os pesos da variável PW no Grupo 1 no método EDP. No método EDC verifica-se que o aumento da porcentagens de *missings* a variável PW torna-se mais relevantes para a formação dos grupos.

A Figura 4 apresenta os resultados do desempenho dos algoritmos ECO, EDP e EDC, nas 100 repetições de Monte Carlo. No método EDC observa-se os maiores desvios em relação as taxas médias de erros quando comparados com os demais. Para a abordagem EDP, observou-se taxas de erros médias crescentes e decrescente, ao longo das porcentagens

**Tabela 4 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP em conjunto com o método EDC sob diferentes porcentagens de dados faltantes.**

| % Missings | Grupos | Pesos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| **5** | 1 | 0.5037 | 0.1256 | **4.9758** | **3.1759** |
| | 2 | 0.6373 | 0.4769 | **2.2666** | **1.4512** |
| | 3 | 0.5558 | 0.5889 | **2.3945** | **1.2758** |
| **10** | 1 | 0.4921 | 0.1092 | **5.3030** | **3.5064** |
| | 2 | 0.5829 | 0.4588 | **2.3282** | **1.6057** |
| | 3 | 0.6278 | 0.6350 | **2.2436** | **1.1177** |
| **15** | 1 | 0.5193 | 0.1112 | **4.9059** | **3.5269** |
| | 2 | 0.6167 | 0.4545 | **2.0929** | **1.7041** |
| | 3 | 0.5142 | 0.6845 | **2.0667** | **1.3744** |
| **20** | 1 | 0.4840 | 0.0961 | **4.7588** | **4.5156** |
| | 2 | 0.5645 | 0.4023 | **2.3154** | **1.9013** |
| | 3 | 0.5618 | 0.6328 | **2.4836** | **1.1322** |

**Tabela 5 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP em conjunto com o método EDP sob diferentes porcentagens de dados faltantes.**

| % Missings | Grupos | Pesos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| **5** | 1 | 0.4825 | 0.1349 | **5.1574** | **2.9772** |
| | 2 | 0.5606 | 0.5797 | **2.4196** | **1.2713** |
| | 3 | 0.6293 | 0.4658 | **2.2921** | **1.4879** |
| **10** | 1 | 0.4753 | 0.1317 | **5.3963** | **2.9595** |
| | 2 | 0.6459 | 0.4525 | **2.2600** | **1.5135** |
| | 3 | 0.5799 | 0.6772 | **2.2523** | **1.1304** |
| **15** | 1 | 0.5011 | 0.1345 | **5.1671** | **2.8709** |
| | 2 | 0.7530 | 0.4172 | **2.2037** | **1.4443** |
| | 3 | 0.5390 | 0.8078 | **2.1712** | **1.0575** |
| **20** | 1 | 0.5011 | 0.1345 | **5.1671** | **2.8709** |
| | 2 | 0.7530 | 0.4172 | **2.2037** | **1.4443** |
| | 3 | 0.5390 | 0.8078 | **2.1712** | **1.0575** |

analisadas. Na estratégia ECO nota-se uma taxa de erro crescente, a partir de 10% de *missings*, sendo o método com o comportamento mais acentuado, pois com o incremento desses valores faltantes foi observado um acréscimo na taxa de erro. O método ECO apresentou também os menores desvios em relação a taxa de erro média quando comparado com os métodos EDC e EDP.

**Tabela 6 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP em conjunto com o método ECO sob diferentes porcentagens de dados faltantes.**

| % Missings | Grupos | Pesos | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| **5** | 1 | 0.4876 | 0.1363 | **5.1732** | **2.9078** |
| | 2 | 0.6364 | 0.4663 | **2.2082** | **1.5256** |
| | 3 | 0.5590 | 0.5811 | **2.4149** | **1.2746** |
| **10** | 1 | 0.4744 | 0.1317 | **5.3949** | **2.9642** |
| | 2 | 0.6447 | 0.4468 | **2.1660** | **1.6024** |
| | 3 | 0.5545 | 0.6754 | **2.2918** | **1.1648** |
| **15** | 1 | 0.5062 | 0.1338 | **5.1193** | **2.8821** |
| | 2 | 0.7345 | 0.4044 | **2.1258** | **1.5835** |
| | 3 | 0.5219 | 0.7826 | **2.2277** | **1.0989** |
| **20** | 1 | 0.4576 | 0.1210 | **5.4762** | **3.2964** |
| | 2 | 0.7185 | 0.4041 | **2.0228** | **1.7024** |
| | 3 | 0.5122 | 0.7171 | **2.4697** | **1.1020** |

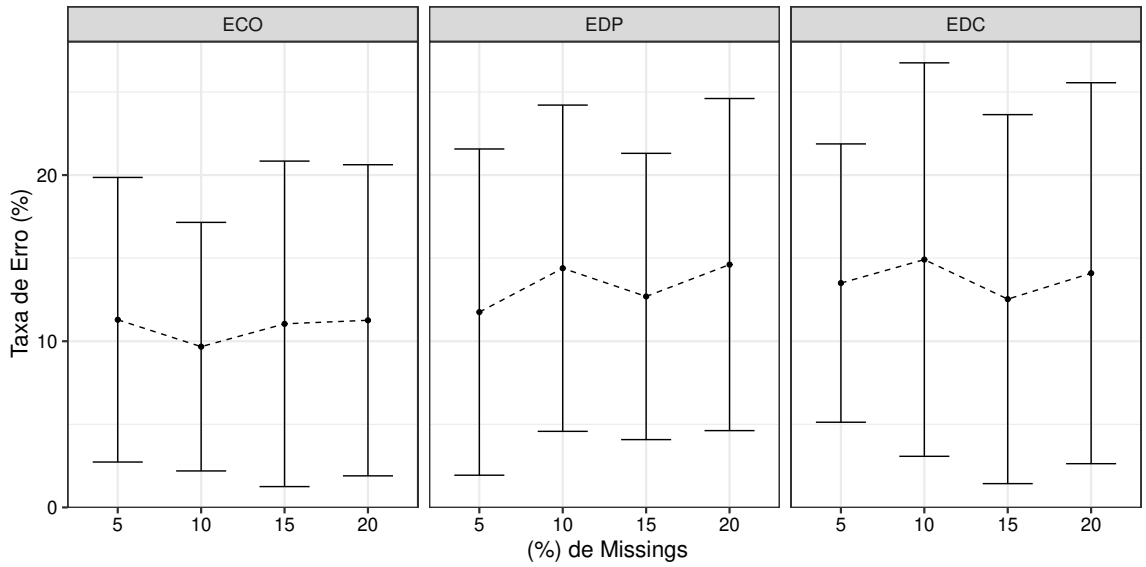



**Figura 4 – Resultados médios das 100 repetições para a Taxa de Erro no Conjunto *Iris Plant*.**

Ao analisar as medidas de consistência das variáveis na Tabela 7, para o banco de dados obtido após o agrupamento com o algoritmo VKFCM-K-LP, em conjunto o método ECO, temos que essas medidas estão bem próximas à zero. Este contexto evidencia uma boa qualidade no agrupamento, ou seja, os valores imputados para os dados faltantes através do método ECO, não foram discrepantes em relação a escala original das variáveis do conjunto de dados *Iris Plant*.

**Tabela 7 – Consistência das variáveis para o conjunto de dados *Iris Plant*.**

| % Missings | Grupos | Variáveis | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| **5** | 1 | 0.00025 | 0.00115 | 0.00534 | 0.00067 |
| | 2 | 0.00355 | 0.00035 | 0.00845 | 0.06201 |
| | 3 | 0.00583 | 0.00271 | 0.01433 | 0.00000 |
| **10** | 1 | 0.00040 | 0.00485 | 0.00563 | 0.00001 |
| | 2 | 0.00785 | 0.00154 | 0.01366 | 0.06181 |
| | 3 | 0.00952 | 0.01112 | 0.03800 | 0.00295 |
| **15** | 1 | 0.00012 | 0.01161 | 0.01162 | 0.00029 |
| | 2 | 0.00537 | 0.00181 | 0.06160 | 0.06618 |
| | 3 | 0.01623 | 0.02967 | 0.03063 | 0.00266 |
| **20** | 1 | 0.00673 | 0.02212 | 0.02356 | 0.00263 |
| | 2 | 0.01598 | 0.00622 | 0.10373 | 0.12329 |
| | 3 | 0.02470 | 0.03045 | 0.01900 | 0.11264 |

## 4.2 CONJUNTO DE DADOS THYORIOD GLAND

Nesta seção os métodos de agrupamento sob a abordagem de dados faltantes são testados no conjunto de dados *Thyoriod Gland* (QUINLAN, 1986). Este conjunto é um dos 6 conjuntos que formam o banco *Thyroid Disease*. O *Thyoriod Gland* possui três classes ($K = 3$), normal (Classe 1) com 150 observações, hyper (Classe 2), com 35 observações e hypo (Classe 3) com 30 observações, que foram construídas a fim de determinar se um paciente encaminhado para o hospital tem hipotireoidismo. Este conjunto de dados possui um total de $n = 250$ observações e cinco variáveis ($p = 5$), Captação de Resina T3 (T3), Total de tiroxina sérica (TTS), Triiodotironina sérica total (TST), Hormônio estimulante da tiróide basal (TSH), Diferença absoluta máxima do valor do TSH após a injeção de 200 microgramas de hormônio liberador de tireotropina (DTSH).

As Figuras 5a e 5b apresentam a Dispersão e os gráficos de Boxplots para as variáveis T3, TST e TST, TTS, observa-se que a Classe 2 se encontra mais dispersa que as demais, este comportamento é evidenciado nos gráficos de Boxplots para as variáves analisadas. Na Figura 5b é possível notar um relação linear entre as variáveis TST e TTS, nas Classes 1 e 2. Além disso, estas classes apresentam uma menor variabilidade nos dados quando consideramos a variável TST.

As Figuras 6a- 6d mostram os *missings* no banco de dados *Thyoriod Gland* distribuídos nas cinco variáveis. A variável T3 apresenta uma maior quantidade de valores faltantes para as porcentagens 15% e 20%. Para todos os bancos com *missings* analisados, a variável DTSH

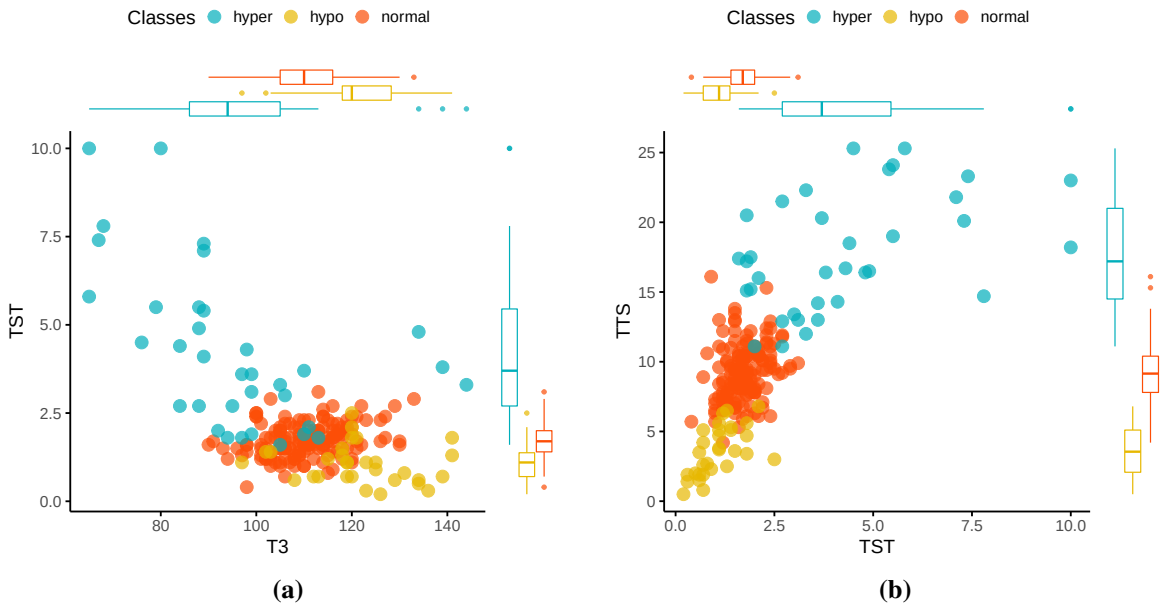
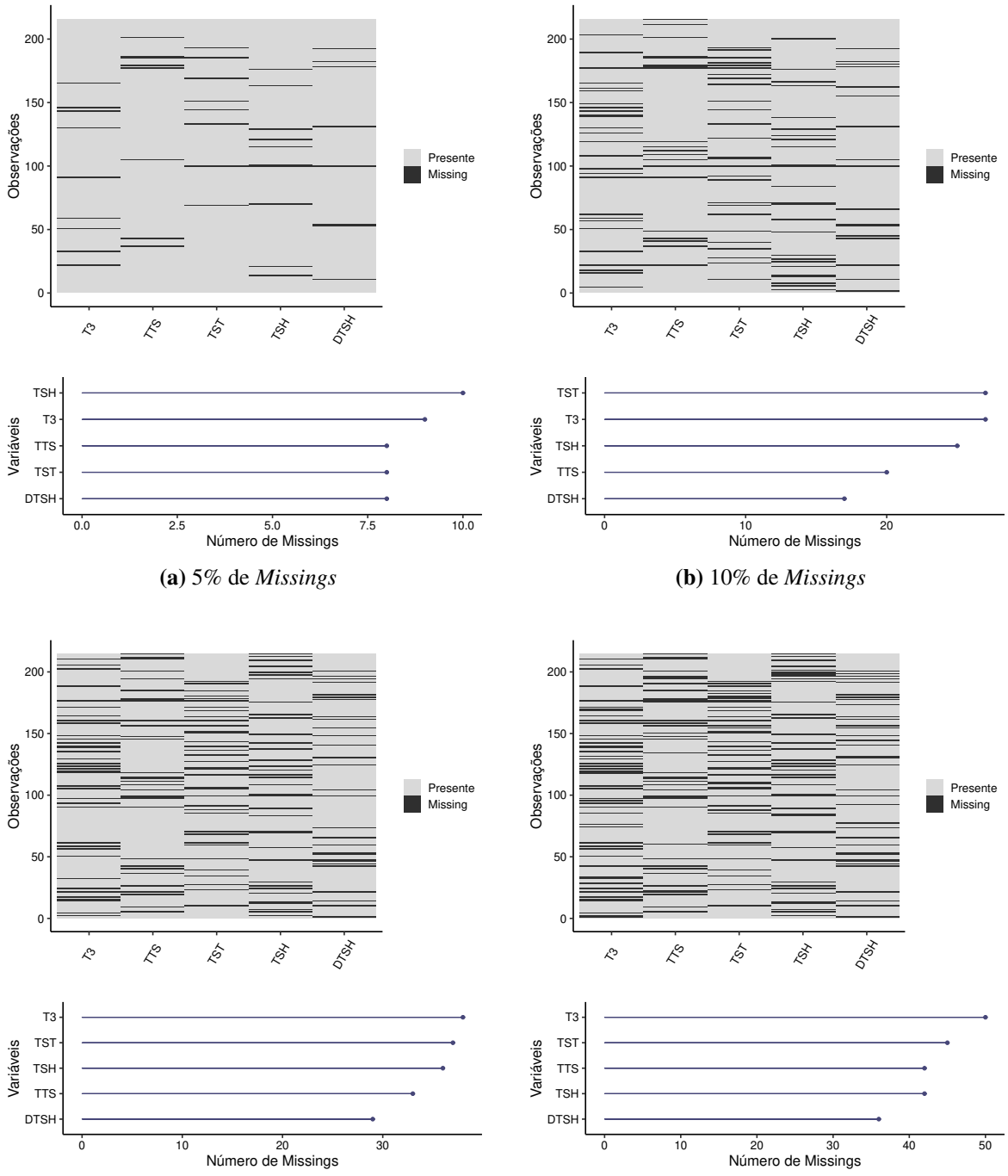



**Figura 5 – Gráficos de Dispersão e Boxplots para o conjunto de Dados Thyroid Gland.**

possui a menor quantidade desses valores. Nota-se também que estes valores encontram-se bem distribuídos entre as variáveis. As observações que estão no intervalo de $1|-150$, constituem Classe 1 - normal, o intervalo de $151|-175$ caracteriza a Classe 2-hyper e o intevalo de $175|-215$ estão as observações para a Classe 3-hypo.

A Tabela 8 apresenta os melhores resultados das medidas de qualidade do agrupamento, dentre as 100 repetições do algoritmo VKFCM-K-LP sob os três tipos de estratégias para dados faltantes. Para uma porcentagem de 5% de *missings*, nota-se que os maiores índices, foram para o método EDC, obtendo um CR igual a 0.818 e um FM no valor de 0.943. Como estes valores encontram-se próximos a 1, pode-se concluir que houve uma boa concordância entre as classes *a priori* e os grupos fornecidos pelo método de agrupamento. Neste contexto, a medida OERC encontrada foi de 5.5%, o que aponta uma dificuldade de 0.055 do algoritmo em fornecer as classes *a priori*. Nos métodos EDP e ECO o aumento do número de *missings* no banco *Thyroid Gland* influencia a qualidade do agrupamento, pois houve um decréscimo nos valores dos índices estudados. Destes, a estratégia EDP apresentou os melhores desempenhos para as medidas de qualidade analisadas em todas as porcentagens de *missings*.

Na interpretação das matrizes de confusão da Tabela 9, os grupos fornecidos pelos métodos de agrupamento foram identificados como Grupo 1 (normal), Grupo 2 (hyper) e Grupo 3 (hypo). As matrizes de confusão na Tabela 9, evidenciam que houve uma maior dificuldade nos algoritmos de agrupamento em identificar os Grupos 1 e 2 em todos os métodos analisados. Estes grupos correspondem as classes normal e hyper, que nas Figuras 5a e 5b mostram-se mais




**(a)** 5% de *Missings*

**(b)** 10% de *Missings*

**(c)** 15% de *Missings*

**(d)** 20% de *Missings*

**Figura 6 – Gráficos de Padrões e Frequências dos valores faltantes por variável para o banco *Thyroid Gland*.**

sobrepostas quando comparadas com a Classe 3, o que dificulta o desempenho dos métodos de agrupamento.

A Figura 7 apresenta as Taxas de Erros médias para as 100 repetições do algoritmo VKFCM-K-LP, sob as abordagens EDC, EDP e ECO no banco *Thyroid Gland*. As taxas de erros médias para os métodos de agrupamento EDP e ECO obtiveram um comportamento crescente,

**Tabela 8 – Desempenho do algoritmo de agrupamento VKFCM-K-LP sob três tipos de abordagens estudadas para o banco *Thyroid Gland*.**

| % | **CR** | | | **FM** | | | **OERC** | | |
|---|---|---|---|---|---|---|---|---|---|
| Missing | EDC | EDP | ECO | EDC | EDP | ECO | EDC | EDP | ECO |
| **5** | 0.818 | 0.803 | 0.775 | 0.943 | 0.939 | 0.930 | 0.055 | 0.060 | 0.069 |
| **10** | 0.509 | 0.734 | 0.656 | 0.838 | 0.918 | 0.892 | 0.176 | 0.083 | 0.111 |
| **15** | 0.787 | 0.633 | 0.586 | 0.935 | 0.885 | 0.868 | 0.065 | 0.120 | 0.139 |
| **20** | 0.753 | 0.441 | 0.434 | 0.923 | 0.809 | 0.807 | 0.074 | 0.204 | 0.200 |

**Tabela 9 – Matrizes de confusão obtidas pelo algoritmo VKFCM-K-LP em conjunto com os métodos EDC, EDP e ECO utilizando 5, 10, 15 e 20% de dados faltantes n dadoso conjunto de *Thyroid Gland*.**

| Métodos | Grupos | 5% | | | 10% | | | 15% | | | 20% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| EDC | 1 | 144 | 0 | 6 | 118 | 1 | 4 | 143 | 2 | 5 | 147 | 5 | 8 |
| | 2 | 6 | 35 | 0 | 32 | 34 | 1 | 7 | 33 | 0 | 3 | 30 | 0 |
| | 3 | 0 | 0 | 24 | 0 | 0 | 25 | 0 | 0 | 25 | 0 | 0 | 22 |
| EDP | 1 | 143 | 0 | 6 | 138 | 1 | 5 | 130 | 1 | 5 | 117 | 2 | 8 |
| | 2 | 7 | 35 | 0 | 12 | 34 | 0 | 20 | 34 | 0 | 33 | 33 | 1 |
| | 3 | 0 | 0 | 24 | 0 | 0 | 25 | 0 | 0 | 25 | 0 | 0 | 21 |
| ECO | 1 | 141 | 0 | 6 | 133 | 1 | 6 | 127 | 1 | 6 | 115 | 2 | 7 |
| | 2 | 9 | 35 | 0 | 17 | 34 | 0 | 23 | 34 | 0 | 35 | 33 | 1 |
| | 3 | 0 | 0 | 24 | 0 | 0 | 24 | 0 | 0 | 24 | 0 | 0 | 22 |

ao longo das porcentagens de *missings* avaliadas. Para 20% de *missings* a Taxa média de Erro Total de classificação para estes métodos foram aproximadamente 0.20. Embora, o método EDP apresente taxas menores entre 5% a 10% de *missings*, quando comparado com a estratégia ECO. As maiores variações são observadas no método EDC para as porcentagens 5% e 15% de *missings*. Este método obteve uma Taxa crescente entre 5% e 10% e decrescente a partir de 10%.

Ao analisar os pesos das variáveis em cada grupo, sob as abordagens EDC, EDP e ECO, nas Tabelas 10, 11 e 12 nota-se que as variáveis TST e TSH foram as mais relevantes para compor o Grupo 1. Na formação do Grupo 2 as variáveis mais importantes foram TSH e a DTSH, para o Grupo 3 as mais relevantes foram as variáveis TTS e TST. Além disso, na estatégia EDC, nota-se que as variáveis TTS e TSH tornaram-se mais relevantes para a formação dos Grupo 3 e 2, respectivamente, com o aumento do número de *missings*. Este comportamento também é observado para a variável TST no Grupo 1 nas estratégia EDP e ECO. Em contrapartida, com o aumento de número de *missings* na variável DTSH, houve um descrésimo da sua importância para a construção do Grupo 2 na estratégia ECO.

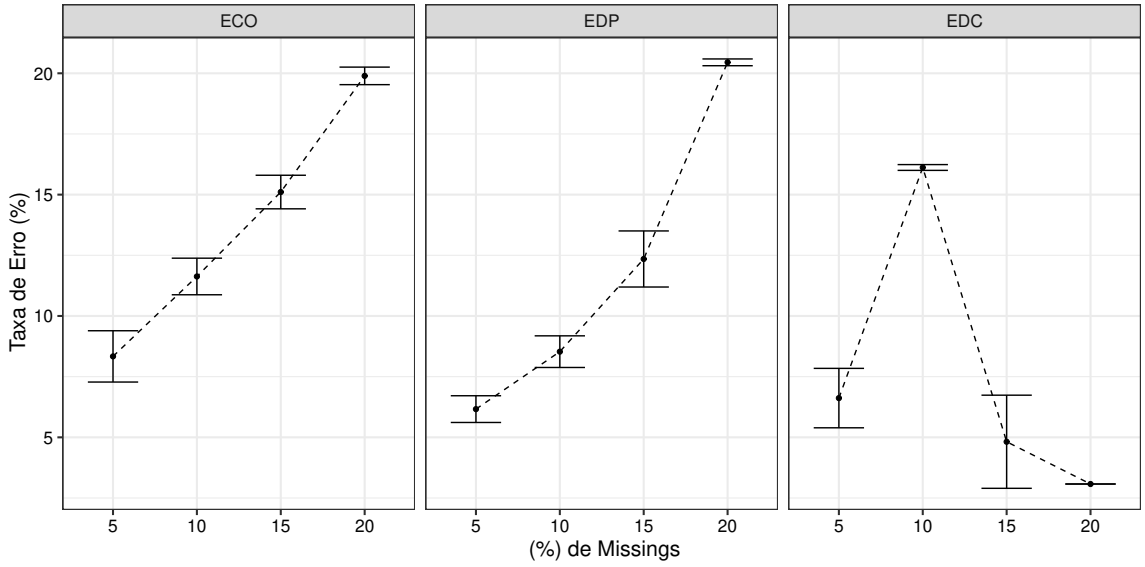



**Figura 7 – Gráficos dos Resultados médios nas 100 repetições dos algoritmos para a Taxa de Erro no Conjunto *Thyroid Gland*.**

**Tabela 10 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP sob a abordagem EDC para o conjunto *Thyroid Gland*.**

| % Missings | Grupos | Pesos | | | | |
|---|---|---|---|---|---|---|
| | | T3 | TTS | TST | TSH | DTSH |
| **5** | 1 | 0.4469 | 0.6842 | **0.9406** | **5.3455** | 0.6502 |
| | 2 | 0.2463 | 0.2615 | 0.1708 | **15.3753** | **5.9078** |
| | 3 | 1.3989 | **3.2601** | **3.1756** | 0.2943 | 0.2345 |
| **10** | 1 | 0.5495 | 0.7670 | **1.1412** | **2.9349** | 0.7082 |
| | 2 | 0.2504 | 0.2753 | 0.1993 | **17.6822** | **4.1143** |
| | 3 | 1.8367 | **3.3095** | **3.0556** | 0.2484 | 0.2166 |
| **15** | 1 | 0.4591 | 0.7436 | **1.1079** | **3.9960** | 0.6614 |
| | 2 | 0.2034 | 0.2707 | 0.1294 | **19.3835** | **7.2343** |
| | 3 | 1.3438 | **3.9067** | **3.1396** | 0.2588 | 0.2343 |
| **20** | 1 | 0.4863 | 0.7701 | 1.2057 | 3.9571 | 0.5595 |
| | 2 | 0.1909 | 0.2899 | 0.1010 | **21.5217** | **8.3067** |
| | 3 | 1.1509 | **4.3158** | **3.4958** | 0.2505 | 0.2298 |

**Tabela 11 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP sob a abordagem EDP para o conjunto *Thyroid Gland*.**

| % Missings | Grupos | **Pesos** | | | | |
|---|---|---|---|---|---|---|
| | | T3 | TTS | TST | TSH | DTSH |
| **5** | 1 | 0.4379 | 0.6721 | **0.9378** | **5.5693** | 0.6503 |
| | 2 | 0.2197 | 0.2613 | 0.1653 | **16.3745** | **6.4314** |
| | 3 | 1.3780 | **3.4215** | **3.2403** | 0.2928 | 0.2234 |
| **10** | 1 | 0.4666 | 0.6584 | **0.9903** | **5.1460** | 0.6385 |
| | 2 | 0.2303 | 0.2467 | 0.1711 | **15.662** | **6.5622** |
| | 3 | 1.4038 | **3.3756** | **3.5690** | 0.2763 | 0.2139 |
| **15** | 1 | 0.4792 | 0.6781 | **1.0123** | **5.2929** | 0.5742 |
| | 2 | 0.2548 | 0.2261 | 0.1784 | **16.4694** | **5.9023** |
| | 3 | 1.2320 | **3.6878** | **3.5524** | 0.2785 | 0.2224 |
| **20** | 1 | 0.4973 | 0.6409 | **1.0404** | **6.0578** | 0.4976 |
| | 2 | 0.3227 | 0.2245 | 0.2006 | **15.8473** | **4.3391** |
| | 3 | 1.1033 | **4.8855** | **4.0218** | 0.2199 | 0.2097 |

**Tabela 12 – Pesos das variáveis em cada grupo ajustados pelo algoritmo VKFCM-K-LP sob a abordagem ECO para o conjunto *Thyroid Gland*.**

| % Missings | Grupos | Pesos | | | | |
|---|---|---|---|---|---|---|
| | | T3 | TTS | TST | TSH | DTSH |
| **5** | 1 | 0.4403 | 0.6634 | **0.9334** | **5.5916** | 0.6557 |
| | 2 | 0.2187 | 0.2711 | 0.1691 | **16.4355** | **6.0643** |
| | 3 | 1.3729 | **3.4176** | **3.2644** | 0.2926 | 0.2230 |
| **10** | 1 | 0.4888 | 0.6737 | **0.9699** | **5.0058** | 0.6253 |
| | 2 | 0.2361 | 0.2653 | 0.1863 | **16.7371** | **5.1163** |
| | 3 | 1.3996 | **3.3082** | **3.6077** | 0.2827 | 0.2117 |
| **15** | 1 | 0.4979 | 0.7017 | **1.0009** | **4.6589** | 0.6137 |
| | 2 | 0.2656 | 0.2554 | 0.2031 | **17.3437** | **4.1813** |
| | 3 | 1.2947 | **3.4309** | **3.2439** | 0.3200 | 0.2168 |
| **20** | 1 | 0.4921 | 0.6576 | **1.0223** | **5.1672** | 0.5848 |
| | 2 | 0.3335 | 0.2552 | 0.2292 | **16.5093** | **3.1038** |
| | 3 | 1.1178 | **4.3838** | **3.4333** | 0.2940 | 0.2021 |

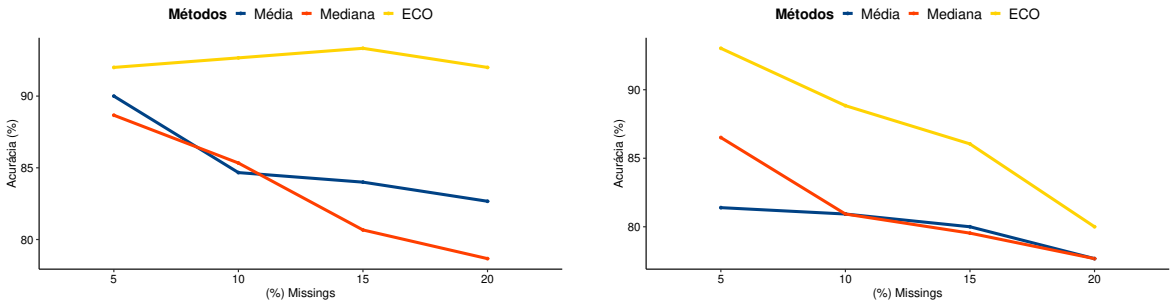
Para avaliar as consistências das variáveis em cada grupo na Tabela 13, foram considerados os bancos antes e depois do agrupamento, com os valores faltantes imputados através do método ECO. Neste contexto, observa-se que as consistências obtidas para as variáveis nos grupos foram próximas a zero, o que indica um bom desempenho do método em imputar os valores faltantes. Nota-se também que as maiores consistências encontram-se nos Grupos 2 e 3 para todas as porcentagens de *missings* avaliadas.

**Tabela 13 – Consistências das variáveis para o conjunto de dados *Thyroid Gland*.**

| % Missings | Grupos | Variáveis | | | | |
|---|---|---|---|---|---|---|
| | | T3 | TTS | TST | TSH | DTSH |
| **5** | 1 | 0.01258 | 0.00092 | 0.00030 | 0.00238 | 0.00529 |
| | 2 | 0.00407 | 0.04769 | 0.04562 | 0.00813 | 0.06072 |
| | 3 | 0.00000 | 0.00691 | 0.00414 | 0.00000 | 0.01321 |
| **10** | 1 | 0.01514 | 0.03953 | 0.02637 | 0.00196 | 0.00063 |
| | 2 | 0.02250 | 0.09938 | 0.14733 | 0.01706 | 0.22179 |
| | 3 | 0.01783 | 0.12019 | 0.00961 | 0.00378 | 0.01297 |
| **15** | 1 | 0.02515 | 0.02872 | 0.01589 | 0.00372 | 0.01526 |
| | 2 | 0.02845 | 0.16012 | 0.22214 | 0.01981 | 0.27982 |
| | 3 | 0.03377 | 0.13482 | 0.00927 | 0.06684 | 0.05777 |
| **20** | 1 | 0.04188 | 0.03056 | 0.02345 | 0.00493 | 0.00116 |
| | 2 | 0.04946 | 0.28002 | 0.34943 | 0.02413 | 0.40891 |
| | 3 | 0.03002 | 0.20914 | 0.02891 | 0.11276 | 0.08566 |

## 4.3 COMPARAÇÃO ENTRE OS MÉTODOS DE IMPUTAÇÃO

Nesta subseção são apresentados os resultados oriundos das comparações dos agrupamentos com VKFCM-K-LP utilizando os métodos ECO, *Imputação via Média e Mediana*. As Figuras 8a-8b mostram as acurácias dos agrupamentos utilizandos o método ECO, Imputação via Média e Mediana nos conjuntos de dados *Iris Plant* e *Thyroid Gland*, quando a quantidade de *missings* é variada de 5 a 20%. Para o agrupamento por *Imputação via Média e Mediana* foram utilizadas as estimativas médias e medianas dos valores observados no banco de dados para cada variável.

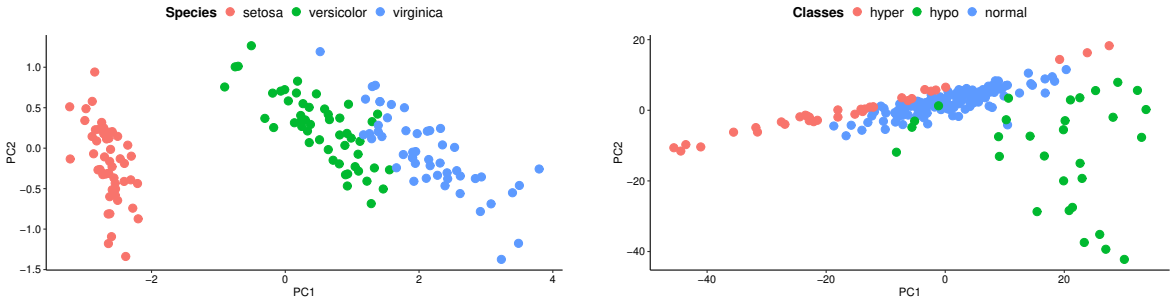



**(a)** *Iris Plant* **(b)** *Thyroid Gland*

**Figura 8 – Gráficos de desempenho dos métodos quando a quantidade de *missings* é variada de 5 a 20%.**

No banco de dados *Iris Plant* com 5% de *missings*, as acurácias foram próximas a 0.90, isto indica que os valores imputados não prejudicaram a qualidade do agrupamento nos três tipos de métodos analisados. No entanto, para o conjunto de dados *Thyroid Gland* nesta mesma porcentagem de *missings*, observa-se uma diferença nas acurácias como mostrado na Figura 8a. A diferença na performance dos métodos com 5% de *missings* nos dois conjuntos de dados analisados, é justificado pelo fato das classes no conjunto de dados *Thyroid Gland* serem mais sobrepostos que as classes do conjunto *Iris Plant*. Com o intuito de visualizar e compreender a sobreposição dos dados, foi aplicada uma Análise de Componentes Principais (ACP) e gerados gráficos com os dois primeiros componentes (Figuras 9a-9b). No ACP as características são ortogonais e ordenadas de acordo com a informação, assim é possível identificar padrões e extrair características (DING; HE, 2004). Temos que, mesmo com a aplicação da ACP os dados das classes no conjunto *Thyroid Gland* encontram-se mais sobrepostos que os dados das classes do conjunto *Iris Plant*. Assim, existe uma maior dificuldade dos algoritmos estudados em agrupar

as classes do conjunto de dados *Thyroid Gland* que as classes no conjunto de dados *Iris Plant*, e esta dificuldade é acentuada com o aumento do número de *missings* no banco de dados como mostrado na Figura 8b. Vale ressaltar também, que como as classes não apresentam sobreposição no conjunto de dados *Iris Plant* isto favorece o desempenho dos métodos de agrupamento mesmo com o aumento na porcetangem de *missings*, o que justifica a acurácia crescente ao longo das porcentagens de *missings* para o método ECO na Figura 8b.




**(a)** *Iris Plant* **(b)** *Thyroid Gland*

**Figura 9 – Gráficos da Análise de Componentes Principais.**

Analisando as porcentagens seguintes, temos que com 10%, 15% e 20% de *missings* imputados, as acurácias do agrupamento com a imputação da Média e Mediana no banco *Thyroid Gland* são bem próximas. Este comportamento não é verificado no conjunto de dados *Iris Plant* na Figura 8a. Além disso, as Tabelas 14 15 16 17 apresentam as consistências das variáveis com a Imputação dos *missings* via Média e Mediana nos dois conjuntos de dados analisados. Ao observar estas tabelas temos que as consistências obtidas foram maiores que as consistências para a estratégia ECO nas Tabelas 7 13, o que implica que os valores imputados via Média e Mediana apresentaram uma maior alteração na escala original das variáveis nos dois conjuntos de dados, quando comparadas com as consistência na estratégia ECO.

Com intuito de mostrar a dispersão das observações imputados em cada método de imputação única, foram selecionadas as variáveis T3 e TST do conjunto de dados *Thyroid Gland* com 5% e 15% de *missings*. Esta escolha deve-se ao fato deste conjunto de dados apresentar variáveis com diferentes escalas de medidas e por apresentar uma maior dificuldade no processo de agrupamento que o conjunto de dados *Iris Plant*. As variáveis T3 e TST foram as que obtiveram um maior número de *missings* na geração, (ver Figuras 6b e 6c). Logo, é importante visualizar graficamente a relação desses valores imputados com os demais presentes no banco, como mostrado nas Figuras 10a-10f. Além disso, também é apresentado os Boxplots para os valores imputados e os valores completos (cor azul). A cor vermelha representa os valores

imputados para a variável T3 e a cor amarela para a varíavel TST. Se os valores forem imputados em ambas variáveis eles recebem a cor preta.

Com 5% de *missings* nas Figuras 10a, 10b e 10c, observa-se que grande parte dos valores imputados encontram-se próximos a distribuição dos valores completos. Os Boxplots dos valores imputados para a variável TST apresentam uma melhor semelhança com os Boxplots dos valores completos, ou seja a dispersão dos dados antes e depois da imputação não apresentaram muitas discrepâncias. Para a variável T3, os valores imputados apresentaram uma menor variabilidade que as observações completas (sem *missings*). Ao avaliar os bancos imputados com 15% de *missings*, nas Figuras 10d- 10f, nota-se que os valores imputados com o método ECO apresentaram uma melhor configuração nos dados em relação aos valores imputados pela Média e Mediana, quando comparados com o gráfico de Dispersão na Figura 5a no conjunto de dados original.

Na imputação via Média e Mediana, nota-se que os valores concentram-se em um mesmo valor formando uma linha reta de inclinação zero. Isto implica que o conjunto de valores imputados através destes métodos apresentam uma correlação igual a zero entre as variáveis T3 e TST. Tabachnick, Fidell e Ullman (2007) argumenta que a imputação dos *missings* com valores de medidas de tendência central como a média, afeta a correlação entre as variáveis e a variância é subestimada.

**Tabela 14 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Média no conjunto de dados *Iris Plant*.**

| % Missings | Grupos | Variáveis | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| **5** | 1 | 0.01731 | 0.01634 | 0.17422 | 0.25311 |
| | 2 | 0.00890 | 0.05284 | 0.03431 | 0.05032 |
| | 3 | 0.02456 | 0.00624 | 0.21387 | 0.00000 |
| **10** | 1 | 0.05468 | 0.05008 | 0.17244 | 0.32902 |
| | 2 | 0.01209 | 0.07793 | 0.08767 | 0.06302 |
| | 3 | 0.12788 | 0.00598 | 0.38896 | 0.09438 |
| **15** | 1 | 0.18425 | 0.12677 | 0.41515 | 0.43382 |
| | 2 | 0.07385 | 0.10341 | 0.15307 | 0.03371 |
| | 3 | 0.12372 | 0.02019 | 0.41156 | 0.09245 |
| **20** | 1 | 0.17918 | 0.12613 | 0.53173 | 0.49124 |
| | 2 | 0.10224 | 0.11974 | 0.18223 | 0.03556 |
| | 3 | 0.19865 | 0.01895 | 0.50715 | 0.23976 |

**Tabela 15 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Mediana no conjunto de dados *Iris Plant*.**

| % Missings | Grupos | Variáveis | | | |
|---|---|---|---|---|---|
| | | SL | SW | PL | PW |
| **5** | 1 | 0.01575 | 0.01845 | 0.18338 | 0.25776 |
| | 2 | 0.00551 | 0.04521 | 0.02264 | 0.02167 |
| | 3 | 0.02632 | 0.00879 | 0.16321 | 0.00000 |
| **10** | 1 | 0.04675 | 0.05812 | 0.18284 | 0.33365 |
| | 2 | 0.03029 | 0.06797 | 0.03022 | 0.02191 |
| | 3 | 0.14485 | 0.00464 | 0.30564 | 0.08778 |
| **15** | 1 | 0.17198 | 0.13849 | 0.42178 | 0.44085 |
| | 2 | 0.09633 | 0.09032 | 0.04010 | 0.05905 |
| | 3 | 0.13630 | 0.04332 | 0.32237 | 0.07927 |
| **20** | 1 | 0.16297 | 0.13849 | 0.53969 | 0.49929 |
| | 2 | 0.13870 | 0.10174 | 0.02639 | 0.10506 |
| | 3 | 0.21722 | 0.04332 | 0.40975 | 0.19766 |

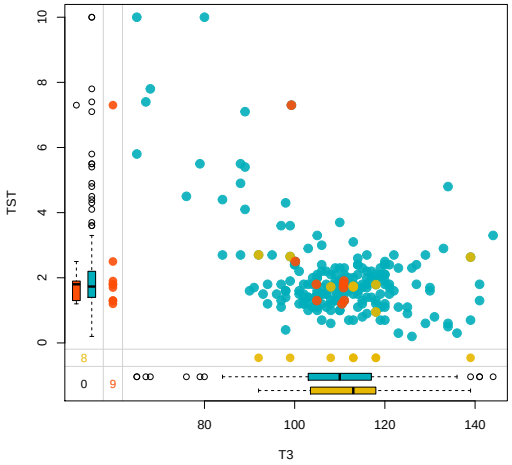
**Tabela 16 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Média no conjunto de dados *Thyroid Gland*.**

| % Missings | Grupos | Variáveis | | | | |
|---|---|---|---|---|---|---|
| | | T3 | TTS | TST | TSH | DTSH |
| **5** | 1 | 0.00891 | 0.00020 | 0.01779 | 0.09343 | 0.03028 |
| | 2 | 0.02228 | 0.14048 | 0.02442 | 0.17675 | 0.23741 |
| | 3 | 0.00000 | 0.17499 | 0.02278 | 0.00000 | 0.00214 |
| **10** | 1 | 0.05014 | 0.00643 | 0.04407 | 0.25214 | 0.05116 |
| | 2 | 0.13232 | 0.16339 | 0.08427 | 0.24967 | 0.39328 |
| | 3 | 0.02532 | 0.23810 | 0.03123 | 0.01299 | 0.00230 |
| **15** | 1 | 0.03076 | 0.01280 | 0.08501 | 0.25876 | 0.08239 |
| | 2 | 0.16093 | 0.20753 | 0.12637 | 0.23315 | 0.43141 |
| | 3 | 0.07188 | 0.36764 | 0.03087 | 0.16200 | 0.03564 |
| **20** | 1 | 0.03848 | 0.01942 | 0.08920 | 0.25748 | 0.10013 |
| | 2 | 0.22618 | 0.32794 | 0.17708 | 0.22526 | 0.51254 |
| | 3 | 0.06937 | 0.44036 | 0.00793 | 0.21541 | 0.11537 |

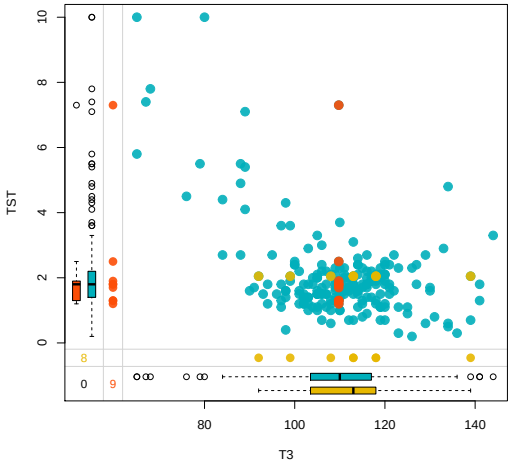
**Tabela 17 – Consistência das variáveis no agrupamento VKFCM-K-LP com a imputação dos *missings* via Mediana no conjunto de dados *Thyroid Gland*.**

| % Missings | Grupos | Variáveis | | | | |
|---|---|---|---|---|---|---|
| | | T3 | TTS | TST | TSH | DTSH |
| **5** | 1 | 0.01004 | 0.00392 | 0.00398 | 0.02499 | 0.00407 |
| | 2 | 0.02254 | 0.14674 | 0.03345 | 0.05759 | 0.21036 |
| | 3 | 0.00000 | 0.16661 | 0.00844 | 0.00000 | 0.00121 |
| **10** | 1 | 0.05661 | 0.02413 | 0.01215 | 0.02623 | 0.00971 |
| | 2 | 0.13441 | 0.17277 | 0.10566 | 0.08908 | 0.36343 |
| | 3 | 0.02321 | 0.22150 | 0.00423 | 0.01625 | 0.00121 |
| **15** | 1 | 0.04232 | 0.01905 | 0.00615 | 0.03655 | 0.02563 |
| | 2 | 0.16424 | 0.22067 | 0.15439 | 0.08908 | 0.39172 |
| | 3 | 0.06671 | 0.34919 | 0.00423 | 0.18079 | 0.04952 |
| **20** | 1 | 0.04668 | 0.01045 | 0.00728 | 0.03869 | 0.01671 |
| | 2 | 0.22855 | 0.34195 | 0.21072 | 0.08908 | 0.47284 |
| | 3 | 0.06671 | 0.41992 | 0.03112 | 0.24134 | 0.13076 |

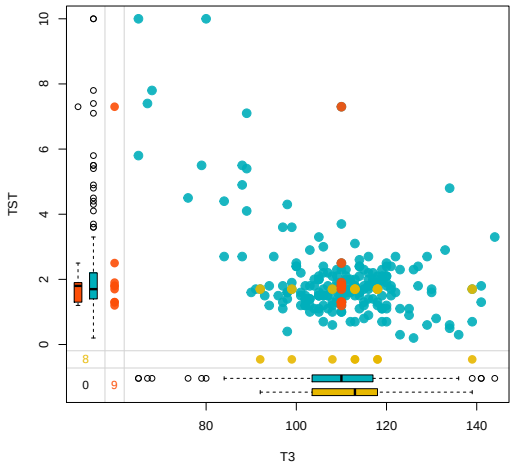
De fato, ao analisarmos as correlações das variáveis T3 e TST, obtivermos $\rho = -0.528$ e $\rho = -0.529$, na imputação da média e mediana, respectivamente. Enquanto, a correlação para o conjunto original (sem *missings*) é de $\rho = -0.536$. A variabilidade dos dados também é prejudicada, ao observarmos os desvios padrão (*sd*) para as variáveis T3 e TST, no conjunto de dados completos temos $sd = 13.145$ e $sd = 1.419$, respectivamente. Enquanto, para o conjunto com 15% dos valores imputados, os desvios padrão são $sd = 11.87$ e $sd = 1.35$ para as variáveis T3 e TST respectivamente, o que indica uma subestimação da variância. Logo, embora as técnicas de imputação via Média e Mediana sejam facéis de implementar, os resultados dos agrupamentos realizados não são satisfatórios, já que a estrutura da correlação das variáveis é modificada e consequentemente estes novos valores podem não apresentar relação com seu grupo de origem, como mostrado nas Figuras 10b 10e 10f 10f. Portanto, o método VKFCM-K-LP sob a abordagem ECO apresentou um melhor desempenho em identificar classes *a priori* 10d 10a de acordo com as acurácias observadas nas Figuras 8a e 8b, logo o conjunto dos valores imputados através deste método são mais próximos ao conjunto das observações do banco de dados original na Figura 5a.



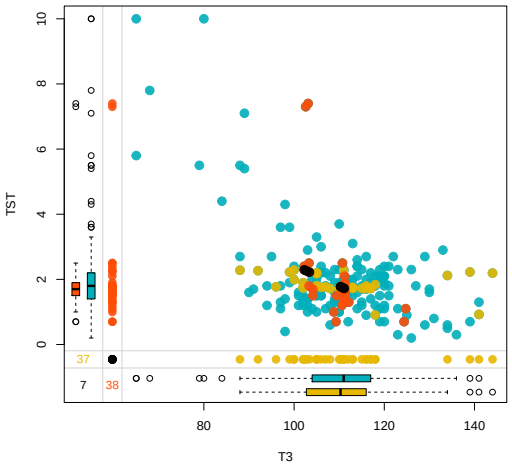

**(a)** Imputação via ECO com 5% de *Missings*.



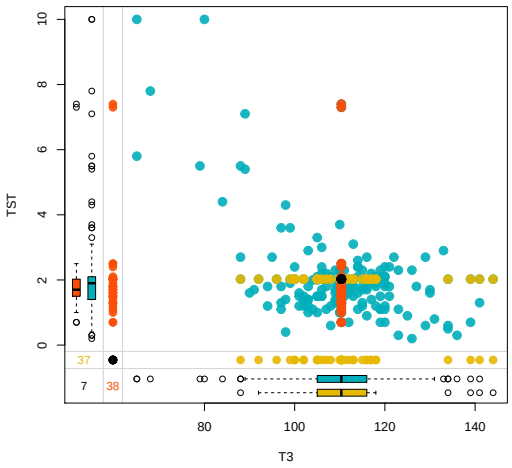

**(b)** Imputação via Média com 5% de *Missings*.



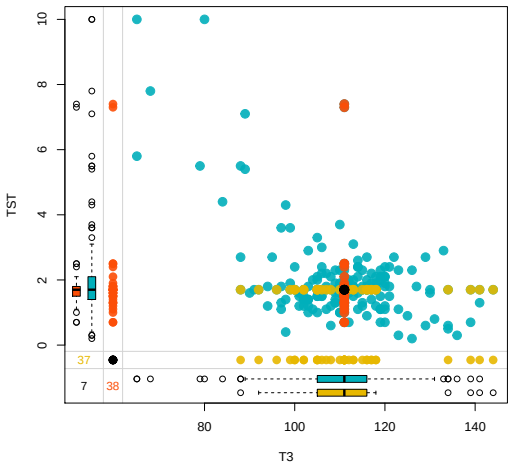

**(c)** Imputação via Mediana com 5% de *Missings*.




**(d)** Imputação via ECO com 15% de *Missings*.




**(e)** Imputação via Média com 15% de *Missings*




**(f)** Imputação via Mediana com 15% de *Missings*.

**Figura 10 – Gráficos para o banco *Thyroid Gland* obtido pelos métodos de imputação única.**

# 5 CONSIDERAÇÕES FINAIS

O problema de *missings* em dados é comumente discutido em várias áreas da ciência, pois as técnicas estatísticas utilizadas para as análises de dados foram originalmente propostas para conjuntos de dados sem valores ausentes, como a análise de agrupamento. Uma alternativa a esta abordagem foi adaptar os métodos de agrupamento para que eles possam manipular conjuntos de dados incompletos. Nesta dissertação foi estudado o método de agrupamento VKFCM-K-LP sob três tipos de estratégias para lidar com dados faltantes, EDC, ECO e EDP. Para a avaliação dos métodos de agrupamento no contexto de dados faltantes foram utilizados dois conjuntos de dados: *Iris Plant* e *Thyroid Gland*. A partir destes conjuntos de dados foram gerados bancos com 5%, 10%, 15%, 20% de *missings*. Os resultados dos algoritmos de agrupamento foram avaliados de acordo com os índices de avaliação CR, FM e OERC.

Os resultados dos agrupamentos com o conjunto de dados *Iris Plant* mostraram-se satisfatórios, com as medidas CR e FM próximas a 1 e a medida OERC próxima a zero, em todos os métodos analisados e para todas as porcentagens de *missings*, o que evidenciou um bom desempenho do método VKFCM-K-LP sob as abordagens EDC, ECO e EDP em identificar classes *a priori*.

Com 5% de *missings* a melhor performance do algoritmo de agrupamento VKFCM-K-LP foi observada na estratégia EDP. Entretanto, ao analisar o gráfico de desempenho para as 100 repetições do algoritmo notou-se que nas porcentagens de 10%, 15% e 20% de *missings*, este método obteve um desempenho inferior aos demais. Na análise das matrizes de confusão para todos os métodos estudados, observou-se que as observações pertencentes a Classe 1 setosa no banco *Iris Plant* foram adequadamente agrupados no Grupo 1.

Em relação aos pesos das variáveis em cada grupo, a variável PL mostrou-se mais relevante, mesmo apresentando o maior número de valores faltantes nos bancos com 5% 10%, 15% e 20% de *missings*. As medidas de consistência das variáveis para os bancos obtidos do agrupamento com o algoritmo VKFCM-K-LP, em conjunto com o método ECO foram próximas a zero, o que mostrou uma boa qualidade do método de agrupamento, ou seja, os valores imputados nos dados faltantes através do método ECO, não foram discrepantes em relação a escala original da variável.

Na geração de *missings* para o conjunto de dados *Thyroid Gland* a variável T3 apresentou uma maior quantidade destes valores para as porcentagens 15% e 20%. As melhores

medidas de qualidade para este conjunto de dados foram observadas nos métodos EDP e ECO. Além disso, estes métodos apresentaram uma taxa de erro média crescente quando analisado o gráfico de desempenho nas 100 repetições do algoritmo.

As matrizes de confusão para o banco *Thyroid Gland* evidenciaram a sobreposição entre as Classes 1 e 2 em todos os métodos analisados, o que justificou o maior número de observações agrupadas incorretamente quando comparadas com a Classe 3. As variáveis TSH e DTSH obtiveram os maiores pesos na construção do Grupo2 em todos os casos analisados. Em contrapartida, a variável T3 apresentou pouca influência na formação dos grupos. As consistências das variáveis obtidas para o método ECO, no banco *Thyroid Gland* foram próximas a zero, o que significa um bom desempenho do método em imputar os valores faltantes.

Na comparação dos resultados oriundos dos agrupamentos com o VKFCM-K-LP, utilizando os métodos de imputação ECO, Média e Mediana, temos que as melhores acurácias foram observadas para o método ECO em todas as porcentagens de *missings* estudadas nos dois conjuntos de dados analisados. Os resultados dos agrupamentos do algoritmo VKFCM-K-LP utilizando os métodos de imputação com a Média e a Mediana não apresentaram resultados satisfatórios, pois foram observados dois problemas: o conjunto de valores imputados afetou a correlação geral das variáveis no banco e houve uma distorção na variabilidade dos dados, o que prejudicou a qualidade do agrupamento. Além disso, as consistências obtidas com a Imputação dos *missings* via Média e Mediana apresentaram os maiores valores quando comparadas as da estratégia ECO.

De modo geral, o algoritmo de agrupamento VKFCM-K-LP em conjunto com as estatratégias para dados faltantes EDC, EDP e ECO apresentaram resultados satisfatórios nos conjuntos de dados com 5% 10%, 15% e 20% de *missings*. Os melhores desempenhos obtidos pelo método de agrupamento foram observados nas estratégias EDP e ECO. Nos grupos com a abordagem ECO, novas bases de dados foram derivadas e os valores faltantes foram estimados no processo de otimização. Os resultados do agrupamento com a estratégia ECO apresentaram desempenhos superiores quando comparados aos grupos de resultados obtidos a partir do conjunto de dados em que os valores faltantes foram imputados pela média e mediana dos valores observados.

Como trabalhos futuros temos, a extensão dos métodos de agrupamento com ponderação automáticas das variáveis propostos por Ferreira e Carvalho (2014), utilizando distâncias adaptativas globais no espaço de características sob a abordagem de dados faltantes. Estudar o desempenho do algoritmo VKFCM-K-LP com as abordagens EDC, EDP e ECO sobre a presença

de *missings* do tipo MNA e MA (HIMMELSPACH; CONRAD, 2010). Aplicar os métodos apresentados nesta dissertação em outras bases de dados. Comparar o método de agrupamento VKFCM-K-LP sob a bordagem ECO com os métodos de imputação múltiplas.

# REFERÊNCIAS


ANDERSON, E. The irises of the gaspe peninsula. **Bulletin of the American Iris society**, v. 59, p. 2–5, 1935.

ANTAL, M.; SZABÓ, L. Z. Some remarks on a set of information theory features used for on-line signature verification. In: IEEE. **2017 5th International Symposium on Digital Forensic and Security (ISDFS)**. [S.l.], 2017. p. 1–5.

BACHE, K.; LICHMAN, M. **UCI machine learning repository**. 2013.

BAEZA-YATES, R.; RIBEIRO, B. d. A. N. *et al.* **Modern information retrieval**. [S.l.]: New York: ACM Press; Harlow, England: Addison-Wesley,, 2011.

BARALDI, A. N.; ENDERS, C. K. An introduction to modern missing data analyses. **Journal of school psychology**, Elsevier, v. 48, n. 1, p. 5–37, 2010.

BEZDEK, J. C. Pattern recognition with fuzzy objective function algorithms. Plenum, New York, 1981.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. Fcm: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, Elsevier, v. 10, n. 2-3, p. 191–203, 1984.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. Wadsworth, 1984.

CAMASTRA, F.; VERRI, A. A novel kernel method for clustering. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 27, n. 5, p. 801–805, 2005.

CHEN, D. Z. S. Fuzzy clustering using kernel method. **IEEE, Nanjing, China**, 2002.

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.

DIDAY, E. Classification automatique avec distances adaptatives. **RAIRO Informatique Computer Science**, v. 11, n. 4, p. 329–349, 1977.

DING, C.; HE, X. K-means clustering via principal component analysis. In: ACM. **Proceedings of the twenty-first international conference on Machine learning**. [S.l.], 2004. p. 29.

DIXON, J. K. Pattern recognition with partly missing data. **IEEE Transactions on Systems, Man, and Cybernetics**, IEEE, v. 9, n. 10, p. 617–621, 1979.

ESTIVILL-CASTRO, V. Why so many clustering algorithms: a position paper. **SIGKDD explorations**, v. 4, n. 1, p. 65–75, 2002.

EVERS, F. T.; HÖPPNER, F.; KLAWONN, F.; KRUSE, R.; RUNKLER, T. **Fuzzy cluster analysis: methods for classification, data analysis and image recognition**. [S.l.]: John Wiley & Sons, 1999.

FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. A novel framework for imputation of missing values in databases. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, IEEE, v. 37, n. 5, p. 692–709, 2007.

FERREIRA, M. R.; CARVALHO, F. D. A. D. Kernel fuzzy c-means with automatic variable weighting. **Fuzzy Sets and Systems**, Elsevier, v. 237, p. 1–46, 2014.

FERREIRA, M. R.; CARVALHO, F. d. A. de; SIMÕES, E. C. Kernel-based hard clustering methods with kernelization of the metric and automatic weighting of the variables. **Pattern Recognition**, Elsevier, v. 51, p. 310–321, 2016.

FILIPPONE, M.; CAMASTRA, F.; MASULLI, F.; ROVETTA, S. A survey of kernel and spectral methods for clustering. **Pattern recognition**, Elsevier, v. 41, n. 1, p. 176–190, 2008.

GIROLAMI, M. Mercer kernel-based clustering in feature space. **IEEE Transactions on Neural Networks**, Citeseer, v. 13, n. 3, p. 780–784, 2002.

GRAEPEL, T.; OBERMAYER, K. Fuzzy topographic kernel clustering. In: **Proceedings of the 5th GI Workshop Fuzzy Neuro Systems**. [S.l.: s.n.], 1998. v. 98, p. 90–97.

GRAVES, D.; PEDRYCZ, W. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. **Fuzzy sets and systems**, Elsevier, v. 161, n. 4, p. 522–543, 2010.

GREEN, P. D.; BARKER, J.; COOKE, M.; JOSIFOVSKI, L. Handling missing and unreliable information in speech recognition. In: **AISTATS**. [S.l.: s.n.], 2001.

GUSTAFSON, D. E.; KESSEL, W. C. Fuzzy clustering with a fuzzy covariance matrix. In: IEEE. **1978 IEEE conference on decision and control including the 17th symposium on adaptive processes**. [S.l.], 1979. p. 761–766.

HAREL, O.; ZHOU, X.-H. Multiple imputation: review of theory, implementation and software. **Statistics in medicine**, Wiley Online Library, v. 26, n. 16, p. 3057–3077, 2007.

HATHAWAY, R. J.; BEZDEK, J. C. Fuzzy c-means clustering of incomplete data. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 31, n. 5, p. 735–744, 2001.

HAYKIN, S. **Neural networks: a comprehensive foundation**. [S.l.]: Prentice Hall PTR, 1994.

HIMMELSPACH, L.; CONRAD, S. Clustering approaches for data with missing values: Comparison and evaluation. In: IEEE. **2010 Fifth International Conference on Digital Information Management (ICDIM)**. [S.l.], 2010. p. 19–28.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, Springer, v. 2, n. 1, p. 193–218, 1985.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern recognition letters**, Elsevier, v. 31, n. 8, p. 651–666, 2010.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys (CSUR)**, Acm, v. 31, n. 3, p. 264–323, 1999.

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

KOHONEN, T. Essentials of the self-organizing map. **Neural networks**, Elsevier, v. 37, p. 52–65, 2013.

KURGAN, L.; CIOS, K. J.; SONTAG, M.; ACCURSO, F. J. Mining the cystic fibrosis data. **Next generation of data-mining applications**, New York: IEEE Press, p. 415–444, 2005.

LAKSHMINARAYAN, K.; HARP, S. A.; SAMAD, T. Imputation of missing data in industrial databases. **Applied intelligence**, Springer, v. 11, n. 3, p. 259–275, 1999.

LEE, L. L.; BERGER, T.; AVICZER, E. Reliable on-line human signature verification systems. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, IEEE, n. 6, p. 643–647, 1996.

LI, D.; GU, H.; ZHANG, L. A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. **Expert Systems with Applications**, Elsevier, v. 37, n. 10, p. 6942–6947, 2010.

LI, D.; ZHONG, C.; LI, J. An attribute weighted fuzzy c-means algorithm for incomplete data sets. In: IEEE. **2012 International Conference on System Science and Engineering (ICSSE)**. [S.l.], 2012. p. 449–453.

LI, T.; ZHANG, L.; LU, W.; HOU, H.; LIU, X.; PEDRYCZ, W.; ZHONG, C. Interval kernel fuzzy c-means clustering of incomplete data. **Neurocomputing**, Elsevier, v. 237, p. 316–331, 2017.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2014. v. 333.

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.

MERCER, J. Xvi. functions of positive and negative type, and their connection the theory of integral equations. **Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character**, The Royal Society London, v. 209, n. 441-458, p. 415–446, 1909.

MIYAMOTO, S.; TAKATA, O.; UNAYAHARA, K. Handling missing values in fuzzy c-means. In: KOREAN INSTITUTE OF INTELLIGENT SYSTEMS. **Proceedings of the Korean Institute of Intelligent Systems Conference**. [S.l.], 1998. p. 139–142.

MÜLLER, K.-R.; MIKA, S.; RÄTSCH, G.; TSUDA, K.; SCHÖLKOPF, B. An introduction to kernel-based learning algorithms. **IEEE transactions on neural networks**, v. 12, n. 2, 2001.

PODDAR, S.; JACOB, M. Clustering of data with missing entries using non-convex fusion penalties. **arXiv preprint arXiv:1709.01870**, 2017.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

RUBIN, D. B. Inference and missing data. **Biometrika**, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.

RUBIN, D. B. **Multiple imputation for nonresponse in surveys**. [S.l.]: John Wiley & Sons, 2004. v. 81.

RUSPINI, E. H. A new approach to clustering. **Information and control**, Elsevier, v. 15, n. 1, p. 22–32, 1969.

SCHAFER, J. L. Multiple imputation: a primer. **Statistical methods in medical research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 8, n. 1, p. 3–15, 1999.

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. **Neural computation**, MIT Press, v. 10, n. 5, p. 1299–1319, 1998.

SCHOLKOPF, B.; SMOLA, A. J. **Learning with kernels: support vector machines, regularization, optimization, and beyond**. [S.l.]: MIT press, 2001.

SEBESTYEN, G. S. Decision-making processes in pattern recognition. Macmillan, 1962.

SHEN, H.; YANG, J.; WANG, S.; LIU, X. Attribute weighted mercer kernel based fuzzy clustering algorithm for general non-spherical datasets. **Soft Computing**, Springer, v. 10, n. 11, p. 1061–1073, 2006.

TABACHNICK, B. G.; FIDELL, L. S.; ULLMAN, J. B. **Using multivariate statistics**. [S.l.]: Pearson Boston, MA, 2007. v. 5.

TEAM, R. C. *et al.* R: A language and environment for statistical computing. Vienna, Austria, 2019.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 2013.

WAGSTAFF, K. Clustering with missing values: No imputation required. In: **Classification, clustering, and data mining applications**. [S.l.]: Springer, 2004. p. 649–658.

WARD, J.; JOE, H. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, Taylor & Francis Group, v. 58, n. 301, p. 236–244, 1963.

WU, Z.-d.; XIE, W.-x.; YU, J.-p. Fuzzy c-means clustering algorithm based on kernel method. In: IEEE. **Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003**. [S.l.], 2003. p. 49–54.

XU, R.; WUNSCH, D. C. Survey of clustering algorithms. Institute of Electrical and Electronics Engineers (IEEE), 2005.

ZADEH, L. A. Fuzzy sets. **Information and control**, Elsevier, v. 8, n. 3, p. 338–353, 1965.

ZHANG, D. Q.; CHEN, S. C. Kernel-based fuzzy and possibilistic c-means clustering. In: **Proceedings of the International Conference Artificial Neural Network**. [S.l.: s.n.], 2003. v. 122, p. 122–125.

ZHANG, D.-Q.; CHEN, S.-C. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. **Artificial intelligence in medicine**, Elsevier, v. 32, n. 1, p. 37–50, 2004.