UNIVERSIDADE BEIRA INTERIOR

Departamento de Engenharia Informática

Imputação de Valores Omissos em Análise Descritiva de Dados

Luzizila Salambiaku

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutora Maria Paula Prata de Sousa
Co-orientador: Prof. Doutora Maria Eugénia Ferrão

Covilhã, Janeiro de 2021

Dissertação elaborada no Instituto de Telecomunicações - Delegação da Covilhã e no Departamento de Informática da Universidade da Beira Interior (UBI) por Luzizila Salambiaku, Licenciado em Engenharia Informática pela Universidade Kimpa Vita (UNIKIVI/Angola), sob orientação da Doutora Maria Paula Prata de Sousa, Investigadora do Instituto de Telecomunicações e Professora Auxiliar do Departamento de Informática da UBI, e co-orientação da Doutora Maria Eugénia Ferrão, Investigadora do Centro de Matemática Aplicada à Previsão e Decisão Económica e Professora Auxiliar com Agregação do Departamento de Matemática da UBI, e submetida à UBI para discussão em provas públicas.

Este trabalho é financiado pela FCT/MCTES através de fundos nacionais e quando aplicável cofinanciado por fundos comunitários no âmbito do projeto UIDB/50008/2020 e pelo projecto CEMAPRE/REM - UIDB/05069/2020.

Dedicatória

A Deus.
Este trabalho, é uma homenagem póstuma a memória do meu querido pai que em vida se chamou ***Luzizila Paulo***, pela educação que soube dar-me desde a infância até o homem que hoje sou, a minha mãe Maria Jorida Pedro, hoje eu glorifico a Deus...

Agradecimentos

A Deus primeiramente pelo fôlego da vida, pela saúde, força de superar as minhas dificuldades e por tudo que tem feito em minha vida e pela minha família.
Gostava de agradecer a todos que contribuíram para a forma final desta dissertação através de críticas e sugestões. Entre eles destaco:
A minha orientadora, Professora Doutora Maria Paula Prata de Sousa, cuja orientação foi crucial para este trabalho dissertativa, no que tange a colocação de questões pertinentes, na direção da pesquisa, na partilha de experiência científica e pela sua honestidade.
À minha co-orientadora Professora Doutora Maria Eugénia Ferrão, pela dedicação, apoio e ajuda incondicional que tornaram possível o desenvolvimento desta dissertação. Durante uma longa trajetória demonstraram uma paciência, disponibilidade de tempo, os meus sinceros agradecimentos e gratidão.
Aos Prof. MSc. Makiese Mavakala e Mampuya K. Fidele, pelo apoio incontestável nos momentos em que foram mais necessários.
A Prof. Doutora Maria de Fátima, pelas suas ajudas incontáveis, compreensão, etc.
A minha princesa Engrácia Juliana, pela paciência, motivação e sobre tudo pela intercessão.
Aos meus irmãos pelo incentivo e companheirismo.
Aos meus filhos Feliciana Luzizila e Luzizila da Graça, minha benção tê-los.
À família Samba em geral, faltam-me palavras para expressar o que vocês se tornaram para mim e a minha família.
Aos meus padrinhos Pastor Diasilua e Odeth, pelo incetivo e conselhos.
Ao papá Lucas e a mamã Mimosa pelos conselhos.
À mamã Helena pela paciência, conselhos e encorajamento dia e noite.
Gostaria de agradecer também aos meus professores do Curso de Mestrado em Engenharia Informática que permitiram-me atingir até aqui.
Finalmente, gostaria de agradecer á minha família e amigos e companheiros de viagens que ficarão no anonimato, para não correr o risco de esquecer algum, por todas as oportunidades que me proporcionaram e por enriquecerem a minha vida.
O meu muito obrigado a todos.

Resumo

Atualmente lidamos com um grande volume de dados e vários programas que permitem fazer análise destes dados. No entanto, os valores omissos representam um problema frequente no processo de análise destes conjuntos de dados que podem surgir por vários motivos. Por exemplo, podem ser resultados perdidos das análises duma amostra, ou alguns indivíduos não responderem a um determinado questionário. Visto que a maior parte dos programas e algoritmos utilizados para o tratamento de dados requere conjuntos de dados sem valores omissos, isto é, dados completos, a sua existência pode limitar a análise dos dados. Daí, surge a necessidade de recorrer a métodos de imputação de valores omissos. Nesta dissertação foram utilizados e comparados seis métodos distintos de imputação, disponíveis no *software* R e avaliado o seu desempenho em conjuntos de dados relacionados com a área da educação, nomeadamente dados da avaliação nacional do rendimento escolar (Prova Brasil). Foi estudada uma amostra de 20408 estudantes para testar os seis algoritmos em quatro subconjuntos de dados gerados por simulação com diferentes percentagens de valores omissos, considerando 5%, 10%, 15% e 20% nas variáveis de interesse. Foram explorados métodos de imputação simples (Média, Mediana e Moda), métodos baseados em aprendizagem automática (kNN e bPCA) e um método de imputação múltipla (MICE). Foi avaliado o desempenho de cada método adotado neste trabalho calculando os respetivos erros de imputação através as métricas RMSE e MAE. Os resultados obtidos mostram que o método de imputação pela Moda forneceu quase de forma constante menores valores de erro.

Palavras-chave

Imputação, Valores Omissos, Análise Descritiva de Dados, Média, Médiana, Moda, bPCA, kNN, MICE.

Abstract

We currently deal with a large volume of data and several programs that allow analysis of this data. However, missing values represent a frequent problem in the process of analyzing these data sets, which can arise for several reasons. For example, they may be missing results from a sample analysis, or some individuals may not answer a questionnaire. Since most programs and algorithms used for data processing require data sets without missing values, that is, complete data, their existence can limit data analysis. Hence, the need arises to use methods for imputing missing values. In this dissertation, six different imputation methods, available in software R, were used compared. Their performance was evaluated in datasets related to the education area, namely data from the national evaluation of school performance (Prova Brasil). A sample of 20408 students was studied to test the six algorithms in four subsets of data with different percentages of missing values, considering 5%, 10%, 15% and 20% in the variables of interest. Single imputation methods (Mean, Median and Mode), methods based on machine learning (kNN and bPCA) and a multiple imputation method (MICE) were explored. The performance of each method adopted in this work was evaluated by calculating the respective imputation errors using the metrics RMSE and MAE. The results obtained show that the method of imputation by Mode provided almost constantly lower values of error.

Keywords

Imputation, Missing values, Descriptive data analysis, Mean, Median, Mode, bPCA, kNN, MICE.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

API *Application Programming Interface*

bPCA *Bayesian Principal Component Analysis*

BSD *Berkeley Software Distribution*

CD Censo Demográfico

CSV *A Comma Separated Values*

CPU *Central Processing Unit* (Unidade Central de Processamento)

ESS *European Social Survey*

FKM *fuzzy K-means*

UBI Universidade da Beira Interior

KDD *Knowledge Discovery in Databases* (Exploração de Conhecimento em bases de dados)

KNN *K-nearest neighbors* (K-Vizinhos mais Próximos)

LOCF *Last Observation Carried Forward*

MAE Erro Absoluto Médio

MAR *Missing At random* (Falta de dados aleatória)

MCAR *Missing complete At random* (Falta de dados completamente aleatória)

MI *Multiple Imputation* (Imputação Múltipla)

MICE *Multiple Imputations by Chained Equations*

MV Máxima Verossimilhança

NA *Not Available*

NMAR *Not missing At random* (Falta de dados não aleatória)

PCA *Principal component Analysis* (Análise de Componentes Principais)

RMSE *Root Mean Square Error* (Raiz do Erro Médio Quadrático)

SAS *Statistical Analysis System*

SCE *Supervised Classification Error*(Erro de Classificação Supervisionado)

SPSS *Statistical Package for the Social Sciences*

SVD *Singular Value Decomposition* (Decomposição em Valores Singulares)

UCE *Unsupervised Classification Error*(Erro de Classificação não Supervisionado)

VIM *Visualization and imputation of missing values* (Visualização e imputação de valores omissos)

Capítulo 1

Introdução

O presente trabalho foi desenvolvido no Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade da Beira Interior. Neste capítulo é feita uma breve apresentação das diferentes temáticas, dos objetivos que almejamos atingir, da metodologia utilizada para obter os resultados e da estrutura desta dissertação.

Esta dissertação foi planeada e escrita num contexto universitário e académico, com intuito de abordar alguns conceitos de análise de dados, mecanismos de valores omissos bem como a imputação destes valores omissos utilizando o sistema R, podendo ser útil para qualquer profissional ou investigador que necessite deste tipo de trabalho.

As contribuições da dissertação consistem num trabalho de síntese sobre métodos de imputação de valores omissos e na avaliação da aplicação de alguns desses métodos em conjuntos de dados com diferentes percentagens de valores omissos simuladas a partir de um conjunto de dados real.

No entanto, esta dissertação não pretende explicar em detalhe todos os métodos e técnicas de análise de dados, no que diz respeito a imputação de valores omissos nem os seus fundamentos teóricos, que podem ser estudados através da consulta de outros textos especializados. Sendo oferecidos neste trabalho, uma vertente marcadamente prática e simulada de seis métodos de imputação de valores omissos. São ainda fornecidas algumas referências para que o leitor desta dissertação possa aprofundar o seu conhecimento.

1.1 Contextualização

No final da década de 80, foi introduzida pela primeira vez a expressão que atualmente delimita uma área das ciências da computação, a exploração de conhecimento (*KDD - Knowledge Discovery in Databases*) por Gregory Piatetsky-Shapiro, na *International Joint Conference on Artificial Intelligence* [Ps91, p. 68]. Os autores Fayyad, Piatesky-Shapiro e Smyth [FPSS96], definem KDD como um processo não trivial de obtenção de nova informação, válido, compreensível e útil, que se decompõe em diferentes etapas (ver Figura 1.1):

1. Seleção;
2. Pré-processamento;
3. Transformação;
4. Data Mining;
5. Análise/avaliação e Interpretação.

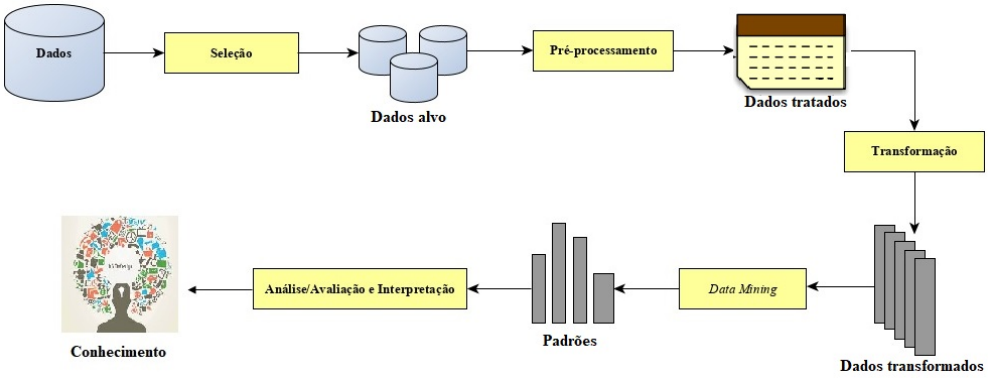


Figura 1.1: Esquema dos processos que constituem o processo de KDD.
Fonte: Adaptado de [FPSS96].

O foco deste trabalho consiste no problema que pode ocorrer na fase de pré-processamento, a existência de valores omissos que por sua vez compromete o desempenho dos algoritmos de análise e exploração de dados e a fiabilidade dos seus resultados.

Não obstante o que se afirmou, ainda no âmbito do tratamento de valores omissos, Rubin [LR02a], afirma que, é comum em pesquisas científicas o problema da ocorrência de dados omissos (*missing data*).

Assim, segundo Little *et al.* [LR87], Zhang [Zha07], a abordagem analítica adequada para conjuntos de dados com observações incompletas é uma questão que pode ser muito delicada quanto aos resultados e conclusões errados que podem ser tomados na utilização de métodos inadequados sobre o estudo dos dados. Nos últimos anos têm-se levado acabo muitas pesquisas sobre o desenvolvimento de métodos estatísticos direcionados a tratar dados omissos.

1.2 Definição do Problema

Segundo Pereira [Per14] a maior parte das técnicas estatísticas foram concebidas ou projetadas para analisar dados completos. Por esta razão sempre se procurou tratar os dados para que se tornem aceitáveis de serem analisados usando técnicas já consolidadas, tornando a inferência mais precisa sobre os dados.

Neste sentido a imputação de valores omissos no pré-processamento de dados é imprescindível. A mesma deve ser feita de forma cuidadosa para que garanta sempre a integridade dos dados, evitando assim a introdução de valores indesejáveis que por sua vez poderá causar agravamento da perda de precisão quando feita a análise final dos mesmos dados. Como não há uma solução única para a maioria dos casos de dados omissos, têm sido investigados e propostos diversos métodos para dar cobro a este problema, onde cada caso apresenta melhores resultados ou desempenho com uma técnica específica. Entre outras, apresentamos alguns exemplos: *K-Nearest Neighbors* (KNN) [DBBA20], *Sequential K-Nearest Neighbors* (SKNN) [LKGTD19], *Singular Value Decomposition*(SVD) [ISS14], *Fuzzy K-Means* (FKM) [SMG15], *Bayesian Principal Component Analysis* (bPCA) [QZH+08,

OST+03], *Multiple Imputations by Chained Equations* (MICE) [NKF10], etc.

1.3 Objetivos

O presente trabalho tem como objetivo principal explorar mecanismos da linguagem R em aplicações de análise de dados, comparando distintas abordagens relacionadas ao tratamento de valores omissos. Pretende-se usar dados reais do sistema de Avaliação Nacional do Rendimento Escolar (Prova Brasil) do ano 2017, disponível em `http://portal.mec.gov.br/prova-brasil`. Simulando, a partir desses dados, vários conjuntos de dados com diferentes percentagems de valores omissos, pretende-se analisar o comportamento de diferentes métodos de imputação, comparando as características do conjunto de dados original com as características do conjunto de dados após imputação.

O erro introduzido por cada método de imputação será avaliado através das métricas erro de raiz quadrático médio (*Root Mean Square Error* - RMSE) e erro absoluto médio (*Mean Absolute Error* - MAE).

1.4 Metodologia

Para efeitos da análise de valores omissos, na elaboração deste trabalho, fez-se uma revisão bibliográfica relacionada a nossa temática, em publicações existentes, livros, artigos, jornais, dissertações (impressos e online) e sites web que nos permitiu compreender os principais conceitos sobre tratamento de valores omissos. As diferentes análises de dados foram realizadas com apoio de *softwares* livres e gratuitos, *RStudio* e R, este último desenvolvido pela *Foundation for Statistical Computing* disponível em `https://www.r-project.org`. Neste site, pode ser encontrada a versão mais recente para (*Microsoft Windows, Linux e MAC OS*. Utilizaram-se diversas bibliotecas apropriadas para análise de dados. A versão mais atual do R, até no memento da escrita desta dissertação, é a 4.0.3, lançada em 10 de Outubro de 2020.

Para a componente prática, obedecemos aos seguintes passos:

1. Fez-se o download e instalação dos softwares (R e RStudio) assim como as bibliotecas necessárias;
2. Seleção da fonte de dados a utilizar;
3. Verificação dos dados a fim de averiguar se havia existência de valores omissos ou não nos ficheiros disponibilizados;
4. Fizemos uma análise preliminar dos dados por intermédio de gráficos estatísticos;
5. Construção de conjuntos de dados com 5%, 10%, 15% e 20% de dados omissos nas variáveis em estudo;
6. Para cada um dos conjuntos de dados, fez-se a imputação de valores omissos pelo conjunto de algoritmos seleccionado;

7. Análise e interpretação dos resultados.

1.5 Estrutura da Dissertação

Como em qualquer projeto onde se pretende utilizar algumas ferramentas de desenvolvimento, é preciso dominar e estar familiarizado com os seus conceitos, conteúdos entre outros. Esta dissertação é composta por seis capítulos organizados da seguinte forma:

- No primeiro Capítulo, procura-se apresentar a contextualização da análise de dados e dos problemas das ocorrências de valores omissos, são apresentados os objetivos do trabalho, incluindo a definição do problema e a metodologia.
- No segundo Capítulo, apresenta-se uma descrição resumida sobre as pesquisas científicas publicadas nas áreas de tratamento de valores omissos, principais conceitos relacionados aos valores omissos a sua caraterização e os seus respetivos mecanismos bem como as métricas de avaliação do desempenho.
- No terceiro Capítulo, abordam-se diferentes métodos e técnicas para o tratamento de valores omissos, desde a simples imputação por eliminação até métodos de imputação múltipla.
- O quarto Capítulo apresenta-se as simulações das análises de dados e comparação de desempenho para as imputações realizadas por diferentes métodos, incluindo métricas para avaliação de erros.
- O quinto Capítulo faz uma análise comparativa dos resultados.
- Por último, no sexto Capítulo são feitas as considerações finais e apresentação de algumas perspetivas para trabalhos futuros nessa linha de pesquisa.

Capítulo 2

Conceitos e Definições

Neste capítulo, apresentamos conceitos base de análise de dados, conceitos e definições sobre valores omissos e métricas de avaliação do erro.

2.1 Análise Descritiva de Dados

A análise descritiva de dados é a fase inicial do processamento de dados. Segundo Reis & Reis [RR02, p. 5], "utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos. As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como percentagens, índices e médias."

No âmbito desta Dissertação usaremos as seguintes estatísticas descritivas : medidas de síntese (média, mediana mínimo e máximo, 1º e 3º quartis e desvio padrão) com a finalidade de descrever e sumariar/sintetizar os conjuntos de dados em estudo através de tabelas e gráficos.

2.2 Tipos de Variáveis e de Dados

As características ou atributos de uma população ou amostra são representadas através de variáveis. Uma variável é qualitativa se representar atributos qualitativos (estado civil, nacionalidade, etc.) e os valores da variável são passíveis de categorização. Quando as caraterísticas ou atributos são quantitativos, tais como peso, altura, número de elementos do agregado familiar, etc, os valores da variável quantitativa são numéricos. Nestes termos, as variáveis podem ser organizadas em diferentes escalas, segundo a possibilidade de mensuração [Ste46]:

Escala ordinal: quando é possível estabelecer uma relação de ordem entre as categorias da variável. Exemplo: classe social. Escala nominal: quando não é possível estabelecer uma relação de ordem entre as categorias da variável, sendo comparadas apenas por igualdade ou diferença. Exemplo: nacionalidade. Escala intervalar: quando é possível quantificar as diferenças entre as medidas, mas não há um ponto zero absoluto. Exemplo: temperatura.

Escala razão/absoluta: quando é possível quantificar a razão entre dois valores da variável e o zero tem significado. Exemplo: peso, altura, etc. As variáveis ou os dados com este tipo de escala podem ainda ser discretas, quando os valores da variável é um conjunto finito ou enumerável, tal como o número de filhos, o número de alunos numa escola etc.;

ou pode ser contínua quando há uma infinidade de valores entre quaisquer dois valores da escala, tal como é o exemplo dos atributos peso ou altura.

Os diferentes conjuntos de dados podem ser classificados num dos seguintes tipos:

- Numérico (dados quantitativos);
- Nominal (dados qualitativos);
- Ordinal (dados qualitativos).

Os dados de tipo numérico são dados quantitativos, que permitem representar valores quantificáveis associados a uma determinada característica ou atributo da unidade estatística observada ou mensurada (e.g. , peso ou altura de uma pessoa, o número de alunos da UBI, o número de páginas de um livro, etc.). As variáveis que representam atributos quantitativos podem ainda ser classificadas como discretas (por exemplo número de elementos do agregado familiar) ou contínuas (por exemplo peso do sujeito) [VB17].

Para o propósito desta Dissertação, serão usadas variáveis qualitativas e quantitativas, com escala ordinal. Ou seja, tal como se descreve detalhadamente no Capítulo 4, são usadas três variáveis da "Prova Brasil 2017" (SSE, TSR e DL). A variável *SSE* é a situação socioeconómica do estudante, *TSR* é a trajetória do estudante sem repetição da série e *DL* representa o desempenho do estudante na leitura.

2.3 Dados Omissos

O enquadramento metodológico para o estudo sobre a imputação de dados omissos realizado nesta Dissertação beneficiou do trabalho prévio das orientadoras, concretamente através da leitura de alguns artigos [FP19], [FPA20], em que são apresentados os conceitos e as definições principais. Nestes termos, segundo [FP19], o conceito de dados omissos é definido como "a diferença entre o conjunto de dados que planeamos recolher e o que realmente é conseguido" (Longford, 2005, p. 13). Os dados omissos ou ausentes são devidos à falta de resposta do item (pergunta/questão), ou seja, participantes de uma pesquisa ou teste que não dão respostas para cada item ou pergunta administrada (item omisso). Em algumas situações, os participantes esperados na *survey* ou teste não aparecem (sujeito omisso). Para Schafer (1999, p. 1), quando os dados omissos compreendem apenas uma pequena fração de todos os casos, a exclusão de casos pode ser uma solução perfeitamente razoável para o problema. Nas configurações multivariadas os dados omissos ocorrem em mais de uma variável; no entanto, os casos incompletos costumam ser uma parte substancial de todo o conjunto de dados. Nesse caso, excluí-los pode ser ineficiente, fazendo com que grandes quantidades de dados sejam descartadas. Além disso, omiti-los da análise tenderá a introduzir viés, na medida em que os casos incompletamente observados diferem sistematicamente dos completamente observados. Contudo, nós não sabemos qual o impacto que realmente tem nos resultados das análises descritivas ou inferenciais, em particular quando se trata de grande volume de dados (big data) [PHW16].

Para Rubin (1987), a representação dos dados completos por Z_{com}, pode ser dividida em Z_{obs} (valores observados) e Z_{omi} (valores omissos ou não observados) ou seja:

$$Z_{com} = (Z_{obs}, Z_{omi}) \tag{2.1}$$

Dada uma matriz de dados Q de ordem n $\times$ k, onde n representa o número de objetos e k quantidade (número) de itens (variáveis), em que z_{ij} é o valor da variável$i(i = 1, ..., n)$ e objeto $j(j = 1, ..., k)$. "Onde o autor cria uma variável indicadora Q para fornecer uma distribuição probabilística para estudar o comportamento de valores omissos em que Q_i assume o valor 0 ou 1, designada também como distribuição indicadora.

$$Q \quad = \quad \begin{cases} 1, & \text{se } Z_{ij} \text{ observado} \\ 0, & \text{se } Z_{ij} \text{ não observado} \end{cases} \tag{2.2}$$

Esta distribuição depende da forma como os dados ausentes se distribuem ao longo da matriz de dados, quando o indivíduo não apresentar resultado sobre a variável em estudo ele receberá o valor 0, caso contrário será representado pelo valor 1" [Zha07] [dS12]. Esta distribuição pode ser um instrumento para analisar as causas para a ausência de dados, observando as relações entre os valores omissos.

2.3.1 Mecanismos de Valores Omissos

Antes da realização de uma análise dos resultados, deve-se procurar saber primeiramente o mecanismo da causa de valores omissos no conjunto de dados ou amostra. Esses mecanismos podem ser classificados como: Valores omissos completamente aleatórios (*missing completely at random* - MCAR), valores omissos aleatórios (*missing at random* - MAR) e valores omissos não aleatórios (*missing not at random)* - MNAR.

Para Buuren [vB12, p. 31] "o termo geral do modelo de dados omissos (*missing data model*) é dado pela expressão:"

$$P(Q|Z_{obs}, Z_{omi}, \gamma) \tag{2.3}$$

Onde *P* representa a distribuição de probabilidade, *Q* indicador de valores omissos, Z_{obs} é a parte de valores observados e Z_{omi} a parte de valores omissos, respetivamente. Os parâmetros desconhecidos estão contidos no γ que descreve a relação entre *Q* e os dados. Nesta equação 2.3 diz-se que a probabilidade de *Q* assumir um valor igual a zero ou um pode depender de Z_{obs} e Z_{omi}.

- MCAR

Para Silva [dS12], "os valores omissos completamente aleatórios (MCAR), ocorrem quando a probabilidade de um item ter respostas omissas não depender nem dos valores observados, Z_{obs}, e nem dos valores omissos, Z_{omi}. A distribuição de MCAR é indicada pela

existência de algum parâmetro γ importante para a probabilidade de *Q* assumir um valor 0 ou 1. Entretanto a omissão completa *Q* não está relacionada com os dados (Z_{obs} e Z_{omi}) e desta forma a distribuição pode ser representada da forma seguinte:"

$$P(Q|\gamma) \tag{2.4}$$

Para GRAHAM *et al.*, (1995) citado por Veroneze [Ver11], "o mecanismo é considerado MCAR mesmo se os valores omissos aparecem por algum acontecimento que não seja verdadeiramente aleatório, mas sim provocado por alguma variável, em alguns casos. Acontece quando a causa é uma variável não correlacionada com a que tenha valores omissos."

- MAR

Ao contrário de MCAR, classifica-se como MAR quando a ausência está relacionada com valores observados noutras variáveis, mas a causa da omissão não está relacionada com os valores ausentes em si [ZJM]. O MAR permite as probabilidades do mecanismo de valores omissos dependerem de dados observados e não de valores omissos [SG02], [Rib15]. A probabilidade Q indicador de valores omissos, definição acima. Indicada pela distribuição tem a dependência da proporção de valores observados Z_{obs} por intermédio de algum parâmetro γ relacionando **Z** e **Q**, desta forma que ela pode ser representada como:

$$P(Q|Z_{obs}, \gamma) \tag{2.5}$$

GRAHAM et al., (1995) citado por Veroneze [Ver11, p. 10], "MAR é um mecanismo acessível porque se a causa que levou aos valores omissos pode ser medida e é adicionada na análise podem ser consideradas todas as influências por elas provocadas."

- MNAR

O mecanismo de valores omissos MNAR ou ainda mecanismo inacessível, acontece quando a probabilidade de um registo com dado omisso em uma variável pode depender do valor do item (variável). Para ZHANG [Zha07], quando a distribuição de Q depende dos valores omissos e dos valores observados, verifica-se a seguinte desigualdade:

$$P(Q|Z) \neq P(Q|Z_{obs}) \tag{2.6}$$

Aqui os dados que tendem a ser omitidos geralmente, são os que aparecem nos extremos da distribuição com mais baixos ou mais altos valores em relação ao padrão da amostra. (e.g.: Um indivíduo com nível de renda muito alto ou muito baixo tem menor probabilidade de responder sobre sua renda num inquérito).

A omissão de dados conhecidos como MNAR pode ser designada de não-ignorável, sabendo que se forem ignorados pode causar enviesamentos na análise de dados [Zhu14].

O esquema apresentado na Figura 2.1, traz um resumo dos três mecanismos, evidenciando as suas principais diferenças.

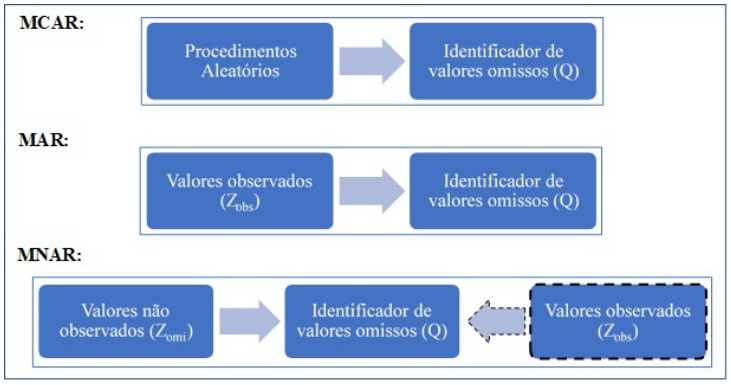


Figura 2.1: Mecanismos de dados omissos.
Fonte: Adaptado de McKnight *et al.* [MMSF07].

2.3.2 Padrões de Valores Omissos

Little & Rubin [LR02b] salientam a importância de detetar o padrão dos valores omissos, que descreve que valores foram observados e os que são omissos numa matriz, para selecionar o procedimento mais adequado no tratamento de valores omissos. Nesta ordem de ideias, identificar o padrão de valores omissos ajuda na descoberta do mecanismo, uma vez que ele preocupa-se na relação que existe entre os valores omissos e os observados [Ver11]. A seguir são descritos os mais frequentes padrões apresentados na Figura 2.2:

- **Padrão univariado:** é o mais frequente em estudos experimentais, quando se tem a presença de valores omissos em apenas uma variável, Figura 2.2 (A).
- **Padrão de item não respondido:** mais frequente em pesquisas por intermédio de questionários, onde os indivíduos respondem alguns itens deixando outros, provocando assim valores omissos com item não respondido para o questionário, Figura 2.2 (B).
- **Padrão monótono:** é comum observar este padrão em estudos longitudinais [1] em decorrência da desistência de participantes da pesquisa ao longo das avaliações por algumas razões, Figura 2.2 (C).
- **Padrão geral (*general pattern*):** neste padrão os valores omissos estão presentes em toda matriz (conjunto de dados), isto é, podem ocorrer em uma ou mais variáveis para qualquer observação [VL18], Figura 2.2 (D). Este último é também chamado de arbitrário.

As áreas sombreadas na Figura 2.2 representam a localização dos valores omissos no conjunto de dados considerando quatro variáveis em estudo.

[1] Método de pesquisa que visa analisar as variações nas características dos mesmo elementos amostrais ao longo de um período de tempo

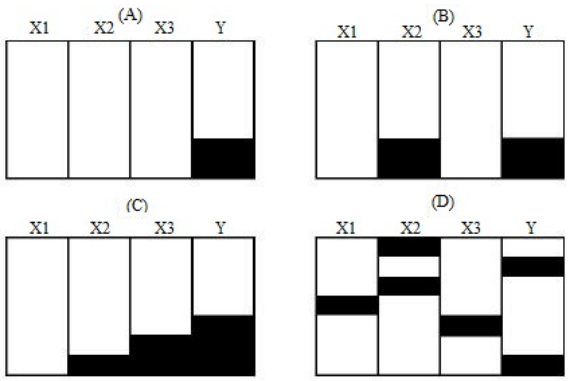


Figura 2.2: Alguns padrões de dados omissos:(A) Padrão univariado, (B) Padrão de item não respondido, (C) Padrão monótono e (D) Padrão geral.
Fonte: Adaptado de Schafer & Granham [SG02]; Enders [End10, p.4]

Como ponto de partida, é importante distinguir entre padrões de valores omissos e mecanismos de valores omissos. Os dois termos têm significados muito diferentes mas são usados algumas vezes de forma alternada por investigadores. Um padrão de valores omissos refere-se à configuração dos valores observados e omissos em um conjunto de dados. Enquanto os mecanismos de valores omissos descrevem as possíveis relações entre variáveis medidas ou calculadas e a probabilidade de valores omissos [End10, p.2].

2.4 Métricas de Avaliação do Erro

Uma métrica serve para medir ou avaliar a qualidade de um modelo segundo os objetivos desejados para uma determinada tarefa. Segundo Mário Filho[2], existem várias funções matemáticas (designados por métricas) que nos permitem avaliar a capacidade de erro e acerto dos nossos modelos. Existem diferentes tipos de métricas que podem ser utilizadas para avaliar um determinado modelo, algumas métricas mais simples, outras mais complexas, sendo que algumas funcionam melhor para conjuntos de dados com determinadas características, ou outras personalizadas de acordo com o objetivo final do modelo.

Para avaliar a eficácia dos diferentes métodos de imputação utilizados nesta dissertação, foram adotadas duas métricas estatísticas. A qualidade dos valores imputados foi avaliada através de raiz quadrada do erro quadrático médio (RMSE [3]) e do erro absoluto médio (MAE [4]).

Para Chai [CD14], apesar de ambas as métricas de erros terem sido usadas nas avaliações de desempenho de modelos há muitos anos, não há um consenso sobre a métrica mais apropriada para erros.

A raiz quadrada do erro quadrático médio (RMSE) tem sido muito usada como uma métrica estatística padrão para avaliar o desempenho do modelo em estudos de meteorologia,

[2] https://www.mariofilho.com/as-metricas-mais-populares-para-avaliar-modelos-de-machine-learning/
[3] Do inglês Root Mean Square Error
[4] Do inglês Mean Absolute Error

qualidade do ar e pesquisas climáticas [CD14, Nad14, SKS07]. A representação matemática do RMSE pode ser:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(O_i - P_i\Big)^2} \tag{2.7}$$

Onde O_i indica o valor real (original) ou observado e P_i, o valor imputado ou preditado e n o número total de observações de O_i. O seu funcionamento, é calcular a diferença entre o valor real do ponto e o valor do mesmo na imputação ou predição; essa diferença, chamada de $residual$, é elevada ao quadrado para depois somar todos e finalmente divide-se a soma dos residuais pelo número total de pontos de dados e calcular a raiz quadrada do quociente, que nos fornecerá a raiz quadrada dos erros quadráticos médios, Equação 2.7. Assim sendo, se o erro for muito maior, como consequência tem-se uma maior influência sobre o erro quadrático total, do que se os erros forem menores (isto é, quando menor for o resultado melhor é o modelo na precisão da imputação realizada).

A segunda métrica usada foi o erro absoluto médio (MAE) que é uma outra métrica importante e amplamente usada em avaliações de modelos. O MAE é representado matematicamente pela seguinte fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\Big|P_i - O_i\Big| \tag{2.8}$$

O erro absoluto médio (MAE) é a diferença média entre pontos de dados imputados (P_i) e observados (O_i) e dada pela fórmula 2.8. Relativamente ao seu cálculo, este envolve a soma dos valores absolutos dos erros, para depois dar o erro total, e a seguir, dividir o erro total pelo tamanho da amostra [WM05].

A principal diferença entre as duas métricas adotadas nesta dissertação consiste em o RMSE atribuir pesos maiores aos erros maiores, pela presença da expressão quadrática, enquanto o MAE mostra a média dos erros absolutos considerando que qualquer tipo de erro tem o mesmo peso.

Capítulo 3

Métodos de Tratamento de Valores Omissos

Schafer [Sch97] e Akande *et al.* [ALR17] definem ”imputação” como um termo genérico para o preenchimento de valores omissos com os dados plausíveis para uma análise dos dados completos posterior. Quando apenas um valor é colocado para cada dado omisso, a imputação diz-se simples, quando há mais de um valor para cada dado omisso, a imputação diz-se múltipla. Ou seja, cada valor omisso é substituído por um conjunto de valores plausíveis que representam a incerteza sobre o valor certo a imputar [Gan16], [Rub87]. Neste capítulo, é apresentada uma breve descrição dos diferentes métodos e das técnicas de imputação utilizadas mais frequentemente em pesquisas relacionadas, a saber: métodos baseados na eliminação, imputação simples, métodos de aprendizagem máquina e imputação múltipla.

3.1 Métodos Baseados na Eliminação

Na maior parte dos estudos em que os dados omissos são identificados, utilizam-se os procedimentos baseados na eliminação de observações [PE04]. Por ser o procedimento padrão dos pacotes estatísticos, acredita-se que vários pesquisadores acabam por utilizar inconscientemente esses métodos [VL18]. Os procedimentos baseados na eliminação de observações são descritos nas seções seguintes.

3.1.1 Eliminação *Listwise*

A análise de caso completo através de eliminação *listwise* (*listwise deletion*) é o método padrão para lidar com dados incompletos em muitos pacotes estatísticos, incluindo SPSS, SAS, Stata, S-PLUS e R. A função *na.omit()* tem o mesmo funcionamento no S-PLUS e R. Este procedimento, elimina todos os casos com um ou mais valores omissos nas variáveis de interesse para análise [vB12],[End10]. ”A conveniência é a grande vantagem do processo de análise completa” [End10, p. 39]. Para Little e Rubin [LR02b], é difícil formular regras de ouro uma vez que as consequências da utilização de exclusões *listwise* dependem de mais do que apenas da taxa de dados omissos.

A Tabela 3.1 mostra um exemplo de três variáveis de dados longitudinais para uma pequena amostra de casos. A abreviatura NA (*Not Available*) indica os valores omissos que foram eliminados, restando apenas as observações completas para cada caso (nas linhas 2 e 5) durante o processo de imputação pela eliminação *listwise*.

Tabela 3.1: Exemplo do conjunto de dados longitudinal imputados pela *Listwise deletion*

	Dados observados			Dados imputados		
Código	X(1)	X(2)	X(3)	X(1)	X(2)	X(3)
1	20	23	NA			
2	27	46	51	27	46	51
3	63	NA	NA			
4	25	NA	26			
5	35	35	37	35	35	37

Fonte: Autor

Encontramos pouca utilização deste método na literatura. Como exemplo, Vinha [VL18] fez o uso deste método tendo como base um conjunto de dados reais de uma avaliação educacional.

3.1.2 Eliminação *Pairwise*

A adoção do procedimento de eliminação (*listwise*) gera perda de uma parcela considerável da informação contida no conjunto de dados, principalmente quando o número de variáveis envolvidas no estudo cresce. O método *pairwise deletion* também conhecido como análise de casos disponíveis, surge como alternativa para reduzir essa perda de informações das exclusões listwise [VL18]. Este método tenta minimizar a perda que ocorre na exclusão *listwise*. Uma maneira fácil de pensar em como a exclusão em pares funciona é pensar na correlação bivariada. A correlação mede a força da associação entre duas variáveis. Para cada par de variáveis para os quais há dados disponíveis, o coeficiente de correlação levará esses dados em consideração. Assim, o método pairwise maximiza a possibilidade de aproveitamento de todos os dados disponíveis. Cada estatística calculada pode ser baseada num conjunto de dados diferente. Embora essa técnica seja geralmente preferida à exclusão listwise, ela também pressupõe que os dados omissos sejam MCAR.

Na Tabela 3.2 reutilizamos os mesmos dados da Tabela 3.1 para mostrar um exemplo de três variáveis de dados longitudinais para uma pequena amostra de casos, considerando apenas uma variável de interesse - **X(3)**, com os valores omissos e aplicar o processo de imputação pela eliminação *Pairwise* (isto é, apenas os valores *NA* de **X(3)** serão removidos.

Tabela 3.2: Exemplo do conjunto de dados longitudinal imputados por *Pairwise deletion*, considerando apenas uma variável de interesse (X3)

	Dados observados			Dados imputados		
Código	X(1)	X(2)	**X(3)**	X(1)	X(2)	**X(3)**
1	20	23	NA			
2	27	46	51	27	46	51
3	63	NA	NA			
4	25	NA	26	25	NA	26
5	35	35	37	35	35	37

Fonte: Autor

3.2 Imputação Simples

A imputação simples que também é chamada de única, serve para preencher cada valor omisso na amostra por um único valor. A imputação única produz conjuntos de dados completos os quais podem ser analisados através de procedimentos analíticos convencionais. Desta forma, existem vários métodos de imputação única e, são apresentados a seguir os principais encontrados na literatura.

3.2.1 Imputação pela Substituição por um Valor de Tendência Central

A substituição de valores omissos por um valor de tendência central constitui um dos métodos mais simples de imputação. A imputação pela média, é um método que foi apresentado pela primeira vez por Wilks em 1932 [Wil32], e possivelmente é um dos mais antigos procedimentos de imputação. A imputação pela média é uma solução rápida e simples para os dados omissos.

A imputação pela média é também considerada como método de imputação por constantes [MMSF07]. Sendo um dos mais comuns, ela consiste em substituir todos os valores omissos de uma variável por um único valor. Fazem parte do grupo de métodos de imputação por constantes, imputação por mediana, imputação de zeros, sendo mais simples a técnica de imputar por zeros, dos exemplos mencionados se no caso for plausível para o conjunto de dados em análise. A imputação pela média poderá ser usada como uma correção rápida somente quando o número de valores omissos é reduzido(pequeno) e deve ser evitada em geral [vB12] [Per14]. É recomendada a utilização do valor da mediana ao invés do valor médio, sempre que existirem valores extremos na amostra [End10].

Trabalhos onde é estudada a imputação pela média são por exemplo: imputação em dados educacionais [VL18]; desenvolvimento de modelos estatísticos de risco para mortalidade cirúrgica após intervenções de cirurgias cardíacas [EPC+01].

3.2.2 Imputação pela Última Observação Realizada (LOCF)

A última observação realizada (LOCF) do inglês *Last Observation Carried Forward*, requer dados longitudinais. A ideia é levar o último valor observado como um substituto para os valores omissos. LOCF é conveniente porque gera um conjunto de dados completo. O método é usado nos ensaios clínicos [vB12]. Esta técnica assume que os resultados não mudam após o último valor observado ou durante o período intermitente em que os valores são omitidos.

A Tabela 3.3 apresenta um exemplo de três variáveis de dados longitudinais para uma pequena amostra de casos. Constata-se que a última observação completa para cada caso ”carrega adiante” para os pontos de valores omissos subsequentes. A abreviatura NA, indica os valores omissos e a cor azul, indica os valores imputados pela última observação realizada (LOCF).

Tabela 3.3: Exemplo do conjunto de dados longitudinal imputados com a última observação transportada

	Dados observados			Dados imputados		
Cód	X(1)	X(2)	X(3)	X(1)	X(2)	X(3)
1	20	23	NA	20	23	23
2	27	46	51	27	46	51
3	63	NA	NA	63	63	63
4	25	NA	26	25	25	26
5	35	35	37	35	35	37

Fonte: Autor

3.2.3 Imputação pela Regressão

A imputação pela regressão[1] incorpora conhecimento de outras variáveis com a ideia de produzir imputações mais inteligentes. A primeira etapa envolve a construção de um modelo a partir dos valores observados. Previsões para os casos incompletos são, então calculadas de acordo com o modelo ajustado, e servem como substitutos para os dados omissos [vB12]. O grau de subestimação depende da variância explicada e da proporção de casos omissos [LR02b]. Imputar valores previstos pode tornar imputações realistas se a previsão é perto da perfeição. Caso isso aconteça, o método reconstrói os dados omissos a partir dos valores disponíveis. É dado improvável surgir este tipo de dados omissos na maioria das aplicações [vB12][Per14].

Encontramos na literatura alguns exemplos de aplicação deste método. Na imputação de valores omissos em [SSQG06]; na imputação de dados laboratoriais omissos e desenvolvimento de um modelo de ajuste de risco para avaliação da qualidade dos serviços cirúrgicos dos hospitais ligados ao *National Veterans affairs* por intermédio das taxas de mortalidade, utilizando a regressão para imputar os dados omissos [KDH+97].

3.2.4 Imputação pelo Método do Indicador

Suponha que se quer usar uma regressão, mas faltam valores em uma das variáveis explicativas. O método do indicador consiste na inclusão de uma variável binária (0,1), que sinaliza o dado omisso. Isto é, adicionando uma nova variável, por indicador, com valor 1 ou 0 indicando se o valor é omisso ou não. É aplicado o procedimento a cada variável incompleta [vB12][Per14]. Este método é frequente em saúde pública e epidemiologia [vB12][Gan16].

3.3 Métodos de Aprendizagem Máquina

3.3.1 K-Nearest Neighbors (KNN)

O K-NN é um algoritmo que se baseia na informação dos k vizinhos mais próximos de um padrão para efetuar a sua classificação e pode ser adaptado para efetuar o preenchimento de valores omissos [JMGL+10].

[1] Em problemas da regressão, o objetivo do modelo é predizer valores numéricos.

Para Ferreira & Rocha [RF17], este método realiza a previsão de novos exemplos sem criar explicitamente um modelo a partir dos dados disponíveis para o treino. Neste sentido, há três pontos necessários para prever um novo exemplo:

1. Comparam-se os valores das variáveis de entrada do novo exemplo com os mesmos valores dos exemplos conhecidos (do conjunto de treino);
2. Escolhem-se os k exemplos mais próximos, usando uma métrica de similaridade aplicada a estes valores (por exemplo, distância Euclidiana, ou outra);
3. O valor do atributo de saída do novo exemplo será igual ao valor do atributo de saída mais comum nos *k* exemplos mais próximos (caso em que a variável de saída é categórica) ou a média dos valores nos k exemplos mais próximos (em regressão).

Neste algoritmo, o valor a preencher pode ser a moda caso a variável seja discreta, e média ou média ponderada caso a variável seja contínua. O uso da média ponderada permite atribuir pesos a cada vizinho de acordo à sua distância do padrão incompleto. Além disso, o algoritmo apresenta dois requisitos, o conhecimento do número de vizinhos a procurar (k) e a escolha da distância entre padrões mais apropriada.

Outra medida de similaridade frequentemente utilizada é a métrica heterogénea de sobreposição euclidiana (HEOM), introduzida por Wilson e Martinez [WM97].

Encontramos a aplicação deste método em [Oli18] que tem como objetivo de tentar encontrar um método de imputação que permita imputar conjuntos de dados clínicos de uma forma consistente, fazem uma análise comparativa entre três métodos (MICE, kNN e missForest), concluem que o missForest obteve os melhores resultados de imputação.

3.3.2 Bayesian Principal Component Analysis (bPCA)

Análise de componentes principais (PCA) do inglês *Principal Component Analysis*, é um dos mais populares métodos de análise multivariada, cujo objetivo é resumir as informações contidas em dados contínuos (ou seja, quantitativos) multivariados, diminuindo a dimensionalidade dos dados sem perder informações importantes [Kas17].

Para Rocha & Ferreira [RF17] "a PCA consta de um procedimento algébrico que converte as variáveis originais (que são tipicamente correlacionadas) num conjunto de variáveis não correlacionadas (linearmente) que se designam por componentes principais (PC) ou variáveis latentes. Assim sendo, a PCA fornece um mapeamento de um espaço com N dimensões para um espaço com M dimensões, onde N é o número de variáveis originais e M é geralmente menor. É importante notar que a PCA é sensível à escala dos dados, pelo que se recomenda a sua normalização prévia."

O método de estimação de valores omissos baseado em bPCA consiste em três processos essenciais, que são regressão de componente principal (PC), estimativa bayesiana e algoritmo repetitivo semelhante a uma expectativa-maximização (EM). A descrição detalhada de cada processo pode ser encontrada em [GLH15][OST+03].

Encontramos na literatura alguns exemplos de aplicação deste método. Um método de valor omissos baseado em bPCA para dados de volume de fluxo de tráfego em Beijing (Pequim) [QZH+08]. O bPCA é também usado no contexto de saúde em [NLB12].

3.4 Imputação Múltipla

Na década de 70, Rubin introduz uma técnica estatística denominada ”imputação múltipla” (IM) para resolver o problema de não-resposta em pesquisas. A técnica ocupa uma posição de destaque na literatura de valores omissos devido à sua aplicação em uma variedade de contextos [LR87][Sch97]. A ideia da imputação múltipla é que a cada valor omisso são imputados dois ou mais valores (**m**), para representar a incerteza sobre qual valor imputar, ao invés de apenas um valor [LR87].

As **m** imputações atribuídas para cada valor omisso geram **m** conjuntos de dados completos. Os resultados obtidos são agrupados/combinados em uma estimativa pontual final acrescidos do desvio padrão, por regras de Rubin (agrupamento simples) [LR87, LR02b, Rib15]. Na Tabela 3.4 apresentam-se algumas estimativas completas de valores e variações de dados, como exemplo da imputação múltipla de valores omissos; Os $m = 5$ conjuntos de valores imputados para $X3$, são arredondados para números inteiros [Sch97].

Tabela 3.4: Os níveis de colesterol em pacientes com ataque cardíaco observados em dias após o ataque, com m = 5 imputações múltiplas

Dados observados			Dados imputados para X3				
X1	X2	X3	1	2	3	4	5
270	218	156					
236	234	NA	186	259	200	259	227
210	214	242					
142	116	NA	238	50	116	133	197
280	200	NA	187	190	186	222	169
272	276	256					
160	146	142					
220	182	216					
242	288	NA	243	264	295	234	215

Fonte: Adaptado de Schafer [Sch97]

3.4.1 Fases da Imputação Múltipla

O método de imputação múltipla resume-se basicamente em três etapas principais: imputação, análise e combinação/agrupamento. Essas etapas estão detalhados a seguir. A Figura 3.1 ilustra resumidamente como as diferentes etapas da imputação múltipla podem ser representadas.

- **Imputação**

Esta é a fase fundamental da técnica de IM [Zha07]. Para Gandolfi [Gan16], o pesquisador deve definir as variáveis que farão parte do modelo de imputação, e o tipo de modelo que melhor se ajusta à distribuição da variável com valores omissos. Nesta fase da imputação

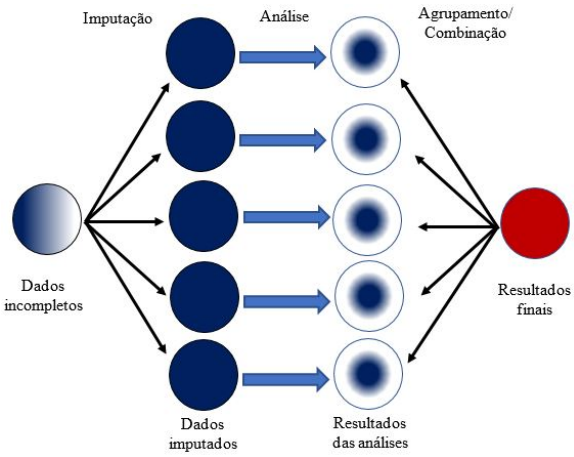


Figura 3.1: Principais fases/passos da imputação múltipla.
Fonte: Adaptado de
`https://www.scielosp.org/media/assets/rbepid/v13n4/05f01.jpg`

são realizadas $m \geq 2$ imputações para cada valor omisso para obter conjuntos de dados completos. A escolha de um valor reduzido de m (conjuntos de dados completos produzidos) pode inflacionar o intervalo de confiança das estimativas e consequentemente diminuir o poder das análises [GOG07]. Molenberghs e Verbeke [MV06] descrevem alta eficácia da imputação múltipla mesmo com pequenos valores de m, (*e.g.* 3 a 5 imputações foram suficientes para terem obtidos bons resultados).

- **Análise**

Na fase da análise são estimados os parâmetros de interesse de cada conjunto de dados imputados. Esta operação é geralmente feita através da aplicação do método de análise que teríamos usado caso os dados fossem completos (sem omissos). Como os dados de entrada são diferentes, os resultados também serão diferentes. Entretanto, é importante notar que as diferenças surgem apenas por causa da incerteza sobre o valor a imputar. Para Baracho [Bar03] os m valores são organizados de tal forma que o primeiro conjunto de dados completos é formado pelo primeiro valor imputado para cada valor omisso, o segundo valor imputado para cada valor omisso forma o segundo conjunto de dados completo, etc. e sendo analisados com o uso de métodos padrão.

- **Combinação/Agrupamento**

A terceira e última fase tem como objetivo combinar todas as análises dos $m \geq 2$ conjuntos de dados completos em apenas um conjunto de resultados produzindo estimativas globais para parâmetros de interesse e para erros padrão. Fórmulas simples para esta fase foram descritas por Rubin [Rub87].

3.4.2 Multiple Imputations by Chained Equations

O método de imputação *Multiple Imputations by Chained Equations* (MICE) pressupõe que os dados omissos em estudo são MAR, isto é dependem apenas dos valores observa-

dos. Neste método são executados vários modelos de regressão sendo a modelação de cada variável com dados omissos condicionada pelas restantes variáveis [ASL07]. Neste método cada variável pode ser modelada de acordo com a sua distribuição e tipo da variável [BO00] [vBGO11].

O processo de imputação tem quatro passos principais [ASL07]:

1. para cada valor omisso é atribuído um valor por um processo de imputação simples com por exemplo a média.
2. para uma das variáveis, X, os omissos voltam a ser colocados.
3. os valores da variável X vão ser estimados por um modelo de regressão em que a variável X é a variável dependente e todas as outras são as variáveis independentes.
4. os valores omissos de X são substituídos pelas imputações preditas calculadas no modelo de regressão. A variável X será depois usada como variável independente nos modelos de regressão para as outras variáveis.

Os passos 2-4 são repetidos para cada variável que tem valores omissos constituindo uma iteração. Finalmente os passos 2-4 são repetidos pelo número de iterações definido pelo utilizador. Após este processo obtém-se um conjunto de dados imputados.

Todo o processo anterior será repetido tantas vezes quantos conjuntos de dados imputados o utilizador quiser construir (geralmente entre 5-10). O processo termina com a análise dos vários conjuntos de dados gerados, incluindo duas fases, primeiro uma análise padrão em cada conjunto de dados e depois combinar as estimativas de cada conjunto de dados.

3.5 Trabalhos Relacionados

A análise da literatura mostra-nos vários trabalhos sobre imputação simples e múltipla, dos quais realçamos:

"*Multiple imputation in big identifiable data for educational research: An example from the Brazilian Education assessment system*" [FPA20], neste artigo os autores fazem uma pequena abordagem sobre o problema da omissão de valores em conjuntos de dados e desafios enfrentados pelos pesquisadores para lidar com casos de valores omissos. O estudo foi realizado utilizando os dados da Prova Brasil 2017, com ajuda do *software* R, concluindo com sugestão do uso da imputação múltipla parace ser um dos mais apropriados métodos para manipular valores omissos.

"*Computing Topics on Multiple Imputation in Big Identifiable data using R: An application to Educational Research*" [FP19], que faz uma abordagem de como lidar com a imputação múltipla em grandes volumes de dados para fins de pesquisa educacional. Neste artigo foram usados dois *packages* do programa estatístico R nomeadamente *BaylorEdPsych* e *mi*. O primeiro *package* foi usado para verificar o mecanismo de dados

omissos existentes no conjunto de dados e o segundo para realização da imputação múltipla. Os resultados obtidos sugerem que a melhoria da qualidade da imputação requer o desenvolvimento de métodos alternativos.

"*Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare*" [DBT18], os autores deste artigo fazem uma comparação de quatro algoritmos de imputação de dados omissos com maior influência na Saúde, nomeadamente Expectativa-Maximização Regularizada (EM), Imputação Múltipla, Imputação kNN e imputação pela média em dois conjuntos de dados reais de saúde, utilizando duas métricas de avaliação (RMSE e tempo de execução). Concluiram que a Expectativa-Maximização Regularizada é melhor algoritmo de imputação de dados omissos por ter apresentado melhor resultado de desempenho do erro RMSE e menos tempo de execução.

"Imputação em datasets médicos – uma comparação entre três métodos" [Oli18], este trabalho faz uma análise comparativa entre três métodos diferentes de imputação disponíveis no R, nomeadamente missForest (método não paramétrico de imputação com base em random forest), KNN (método que imputa o valor omisso com base nos vizinhos mais próximos) e MICE (método de imputação múltipla). Os resultados destacam o método *missForest* foi mais consistente ao imputar os dados omissos, apresentado erro menor de imputação apesar da sua complexidade e o tempo elevado para imputar os dados.

Outra pesquisa intitulada "Dados Ausentes em avaliações educacionais: comparação de métodos de tratamento" [VL18], cujo objetivo principal consistiu na "comparação de desempenho de quatro métodos de tratamentos de valores omissos (imputação pela média, *listwise deletion*, máxima verossimilhança e imputação múltipla), usando modelos de regressão aplicados aos dados da avaliação educacional, com funções do *software* estatístico SAS." Os quatro métodos foram estudados para os três mecanismos de dados omissos existentes MCAR, MAR, NMAR e para variáveis contínuas e categóricas. Para os dados usados os resultados sugerem a utilização dos métodos baseados na máxima verossimilhança e evitar a imputação pela média.

"*A Review on missing value estimation using imputation algorithm*" [AMAS17]. Neste artigo os autores apresentam uma análise de 31 algoritmos de imputação, descrevendo as suas principais caraterísticas e limitações, onde SVDimpute e BPCA mostram melhor desempenho em conjuntos de dados com baixa entropia e, LLSimpute(Local Least Square Imputation) e KNNimpute obtiveram melhor desempenho com os conjuntos de dados de alta entropia.

Em "Uma Análise da Aplicação de Algoritmos de Imputação de Valores Faltantes em Bases de Dados Multirrótulo" [Scr17] foram selecionadas seis bases de dados multirrótulo de diferentes domínios de aplicação. Tendo sido estudados quatro algoritmos de imputação (Moda, Média, Mediana e KNN), foram obtidos os melhores resultados com o KNN.

"*Missing data methods for arbitrary missingness with samples*" [McN17], aborda um estudo de simulação para comparar e avaliar o desempenho das seguintes técnicas de imputação: máxima verossimilhança, eliminação listwise e imputação múltipla. Neste

estudo, o autor conclui que a imputação múltipla teve o melhor desempenho dos métodos utilizados.

"*Missing Value Imputation Using Stratified Supervised Learning for Cardiovascular Data*" [NM16], neste artigo os autores fizeram a imputação de valores omissos no contexto de saúde com aprendizagem estratificada supervisionada para dados cardiovasculares, onde foram usados quatro métodos de imputação (árvore de decisão, rede neural, K-NN e K-Mean Clustering).

"*A Comparison of Six Methods for Missing Data Imputation*" [SMG15], os autores deste artigo realizam um estudo sobre valores omissos e fazem uma comparação de seis diferentes métodos de imputação, nomeadamente: média, KNN, FKM, SVD, bPCA e MICE. O estudos comparativo realizou-se utilizando quatro conjuntos de dados reais de tamanhos diferentes contendo 4 a 65 variáveis, considerando valores omissos MCAR e quatro métricas de avaliação (RMSE, UCE, SCE e tempo de execução). Dos resultados obtidos neste artigo, os autores destacam e sugerem o bPCA e o FKM como dois métodos de imputação que merecem consideração adicional na prática.

"*A Comparison of various imputation methods for missing values in air quality data*" [ZJM], os autores deste artigo apresentam vários métodos de imputação de dados relacionados à qualidade do ar na Malásia, com o objetivo de selecionar o melhor método de imputação. Os métodos estudados por simulação foram média, mediana, EM (*expectation-maximization*), decomposição de valores singular (SVD), KNN e KNN Sequencial (SKNN). Os melhores resultados foram obtidos com os métodos EM, KNN e SKNN.

"*Improving missing values imputation in collaborative filtering with user-preference genre and singular value decomposition* [ISS14], este artigo trata do problema da filtragem colaborativa e a sua sensibilidade na dispersão de dados quando os utilizadores avaliam alguns produtos ou serviço, baseando-se no conjunto de dados *MovieLens* usando 5 vezes a validação cruzada, concluindo que imputar valores omissos com o modelo proposto por eles (SVDUPMedianCF) a partir do algoritmo k-means modificado, de cada usuário juntamente com SVD obtem-se erro menor comparativamente a abordagem tradicional.

"Preenchimento de falhas em dados de correlação de Anomalia da altura geopotencial (500 hPa)" [dMAL14], neste artigo os autores tratam da imputação múltipla de dados omissos usando média preditiva do MICE. Concluem que este processo preencheu os valores omissos com uma qualidade adequada tendo em conta a existência de um forte coeficiente de correlação de Pearson de 0.99, entre o conjunto de dados originais e os dados imputados.

"Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica" [NKF10], Os autores usaram o mesmo conjunto de dados de [NKF09] (apresentado abaixo) com apenas 450 pacientes do estudo original, fazendo uma análise comparativa de dois métodos de imputação (única e múltipla). A imputação múltipla foi realizada pelo método chamado *Bayesian Linear Regression (BLR)* implementado no *package MICE* do programa R e regressões logísticas multivariáveis

para comparar os métodos. As conclusões mostram a importância da imputação múltipla para ter em conta a variabilidade dos dados.

"Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos" [NKF09], os autores deste estudo propõem promover uma divulgação da imputação múltipla para os pesquisadores da área da saúde e também mostrar o ganho considerável que o pesquisador tem em suas análises ao decidir imputar os valores omissos em vez de fazer uma análise restrita aos casos completos. A comparação dos resultados obtidos com os dados imputados e da análise com casos completos foi feita com dados reais, cuja população da pesquisa foi composta por pacientes internados no Hospital de Clínicas de Porto Alegre, com cerca de 651 pacientes, de onde foram usadas apenas 470 pacientes com todas as variáveis de interesse completas. Foram considerados dois métodos de imputação múltipla do *package* MICE ambos baseados em regressão linear.

"*A bPCA based missing value imputing method for traffic flow volume data*" [QZH+08], neste artigo os autores fazem alusão da existência frequente do problema de dados perdidos em vários sistemas de informação de tráfego, e propõe um método de imputação de valores omissos baseado em bPCA para o preenchimento de dados de fluxo de tráfego. O estudo mostra que a imputação baseado em bPCA é capaz de prever valores omissos com uma maior precisão.

"Métodos de Imputação de Dados Aplicados na Área da Saúde" [Nun07], o autor faz uma análise de dados utilizando a imputação múltipla com o pacote MICE do programa R, onde mostra por meio da aplicação de diferentes métodos de imputação de dados que a imputação múltipla tem mais vantagens em relação a imputação única.

Em "*Dealing with missing data in a multi-question depression scale: A comparison of imputation methods*" são estudados seis métodos de imputação: imputação múltipla, imputação única pela regressão, imputação única pela média individual, imputação única pela média geral, resposta precedente do indivíduo e escolha aleatória [SSQG06].

"Uso de Técnicas de Data Mining para Imputação de Dados Uma Aplicação ao Censo Demográfico de 1991" [Tor03], o autor deste trabalho faz um estudo de imputação em pesquisas domiciliares sobre o Censo Demográfico (CD) de 1991, tratando a não resposta caso um determinado quesito estivesse omisso. O processo de preenchimento usado foi o de padronizar as variáveis de fecundidade, utilizando uma comparação de modelos de Regressão logística e Redes Neuronais, em que o modelo de Regressão Logística apresentou melhores resultados de imputação.

"*Prediction of operative mortality after valve replacement surgery*" [EPC+01], os autores deste artigo, utilizaram dados de mais de 90 mil pacientes para o desenvolvimento de modelos estatísticos de risco para casos de mortalidade cirúrgica após cirurgias cardíacas. Para todas as variáveis contínuas, os dados omissos foram preenchidos pela mediana e para as variáveis categóricas o valor imputado foi aquele mais prevalente na população.

3.6 Ferramentas

Nesta seção apresentamos as diferentes ferramentas essenciais para materialização prática deste trabalho na análise de dados.

3.6.1 Instalação e Configuração do ambiente computacional R

O R é um sistema de computação científica e estatística, programável e que permite o tratamento de vários tipos de dados. Na sua versão base, contém um conjunto de ferramentas que permitem armazenar, processar, calcular, analisar e visualizar dados. Além disso, o R possui ainda uma poderosa linguagem de programação, que permite implementar novas funções com o comportamento definido pelo utilizador.

A instalação do R pode ser feita acessando o *site web* do projeto R em `https://www.r-project.org`, onde poderá encontrar a sua versão mais recente disponível, para *Microsoft Windows, Linux* e *MAC OSX*, podendo ser instaladas livremente. A versão mais recente, até à data de desenvolvimento desta dissertação, é a 4.0.3, atualizada em 10 de Outubro de 2020 [Refa].

A pesar de ter um conjunto de menus disponíveis que permitem realizar algumas operações, o R é um ambiente orientado para o uso de uma linha de comandos. Sendo assim, existem os chamados ambientes de integrados de desenvolvimento (IDE), que inclui a linha de comandos e uma gama de ferramentas complementares. Neste âmbito, recorremos à instalação do *RStudio*[2], que permite criar um ambiente de trabalho mais produtivo dadas as inúmeras funcionalidades disponibilizadas.

3.6.2 Instalação das Principais Bibliotecas

As bibliotecas (*packages*), são conjuntos extras de funções que podem ser instaladas além do R. No entanto, existem *packages* que permitem auxiliar as diversas linhas de pesquisas como: estatística, biologia, ciências sociais, econometria, medicina, *machine learning*, etc. Ao instalar um novo package, terá de o carregar para o poder utilizar. No ambiente visual do R, existe o menu *Packages* onde se encontram opções que nos permitem instalar novos *packages* e atualizar os existentes. O RStudio disponibiliza também um ambiente que permite verificar os *packages* instalados e instalar novos, através da opção *Packages* no canto inferior direito do ambiente de trabalho. Para o trabalho desenvolvido nesta pesquisa (dissertação) em particular, adotou-se além das funções incorporadas nas bibliotecas (*packages*), nas nossas análises os *packages* descritos a seguir. Para instalação de um pacote (*packages*), basta executar o comando ilustrado na listagem 3.1.

```
1 install.packages("Nome_do_packages")
```

Listing 3.1: Exemplo de sintaxe da instalação de packages na linha de comando.

Uma vez instalado, o *package* poderá ser carregado na memória para o seu uso no código (com ou sem aspas). Para tal usa-se o comando ilustrado na listagem 3.2.

[2]Disponível em: `https://www.rstudio.com`

```
1 library(Nome_do_packages)
```

Listing 3.2: Exemplo de carregamento de packages na linha de comando.

3.6.2.1 Hmisc

O Hmisc[3] (*Harell Miscelaneous*) escrito por Harrell com contribuições de Dupont *et al.*, contém muitas funções importantes para análise de dados[Har19], e nos permitiu ilustrar os diferentes padrões de valores omissos por meios gráficos entre outras.Este *package* contém ainda uma vasta gama de funções, dentre as quais oferece duas funções para imputação de valores omissos, que são $impute()$ e $aregImpute()$ [Vid16]:

A função $impute()$ faz uma inputação simples atribuindo aos valores omissos utilizando o método estatístico definido pelo utilizador que pode ser (média, máx), por padrão é atribuida a mediana; a função $aregImpute()$ permite a imputação usando regressão aditiva (*additive regression*), autoinicialização (*bootstrap*) e a correspondência média preditiva (*preditive mean matching*).

3.6.2.2 imputeTS

O *package* imputeTS é especializado para imputação de séries temporais. Disponibiliza uma gama de implementações de algoritmos de imputação diferentes, tais como, Média, LOCF, Interpolação, etc. Além dos algoritmos de imputação, o imputeTS também oferece funções que permitem gerar gráficos e visualização das estatísticas de valores omissos [MG19]. Neste trabalho utilizou-se para imputação pela média , mediana (Anexo A.1.2) e moda (Anexo A.1.3, adotando a função $na.mean()$ associada ao método desejado(*"mean", "median"* ou *"mode"* A versão mais recente e suas dependências estão disponíveis em `https://cran.r-project.org/web/packages/imputeTS/index.html`).

3.6.2.3 MICE

Este $package$ fornece uma série de métodos de imputação univariados embutidos nele, conforme mostrados na Tabela 3.5. Apesar da especificação de métodos padrões para cada situação, a escolha de um método diferente em casos especiais pode ser melhor [vBGO11]. Foi utilizado neste trabalho com a finalidade de imputar os diferentes conjuntos de dados com valores omissos adotados nesta pesquisa (Anexo A.1.6).

O $package$ mice define três classes distintas de dados:

* $mids$: conjunto de valores resultantes da imputação múltipla;
* $mira$: análises dos valores completos imputados;
* $mipo$: combinação das análises da imputação múltipla.

A Tabela 3.6 descreve as principais funções do *package* $mice$. O algoritmo geral do método pode ser consultado em [Gan16, vBGO11].

[3]Disponível em: `https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf`

Tabela 3.5: Alguns métodos de imputação encontrados no *mice*

Método	Descrição	Tipo
norm	Regressão linear Bayesiana ($Baysianlinearregression$)	Quantitativa
norm.predict	Valores preditos ($Predictedvalues$)	Quantitativa
norm.nob	Regressão estocástica ($Stochasticregression$)	Quantitativa
norm.boot	Imputação normal com *bootstrap* ($Normalimputationwithbootstrap$)	Quantitativa
2L.norm	Modelo normal multinível ($Multilevelnormalmodel$)	Quantitativa
pmm	Média preditiva correspondente ($Predictivemeanmatching$)	Quantitativa
mean	Incondicional imputação média ($Uncondicionalmeanimputation$)	Quantitativa
logreg	Regressão logística ($Logisticregression$)	Binária/Quantitativa
logreg.boot	Regressão logística com *bootstrap* ($Logisticregressionwithbootstrap$)	Binária/Quantitativa
polyreg	Regressão logística multinomial ($Polytomouslogisticregression$)	Quantitativa/Nominal
lda	Análise descriminantes ($Discriminantanalysis$)	Quantitativa/Nominal
sample	Amostragem aleatória a partir dos valores obeservados ($Randomsamplefromtheavailableobservedvalues$)	Qualitativa ou Quantitativa

Fonte: Adaptada de [vBGO11]

Tabela 3.6: As principais funções da biblioteca mice.

Função	Entrada	Saída	Descrição
complete	mids	data.frame	Converte mids em dados completos.
md.pattern	data.frame	matrix	Resumo de padrões dos valores omissos.
mice	data.frame	mids	Criação de conjunto de dados da imputação múltipla.
lm.mids	mids	mira	Regressão linear para dados imputados.
glm.mids	mids	mira	Modelo linear generalizado para dados imputados.
pool	mira	mipo	Combinação das análises repetidas.

3.6.2.4 *Packages* de Visualização Gráfica

Foram utilizados dois diferentes *packages* para visualização gráfica neste trabalho, *Lattice graphics* e VIM[4].

O primeiro *package*, *Lattice graphics* possui um visualizador de dados de forma gráfica para R. O *package* foi usado para visualizar os padrões de valores omissos e para visualizar a distribuição dos valores imputados. Para mais detalhes sobre *lattice* consulte [WKM20]. O segundo *package* visualização e imputação de valores omissos "VIM", permitiu-nos a visualização das atribuições marginais das variáveis, levando em consideração os seus valores omissos, por intermédio da função *marginplot*.

Este último *package*, apresenta novas ferramentas para visualizar valores omissos e/ou imputados, que podem ser utilizadas para explorar dados e a estrutura de valores omissos e/ou imputados. Dependendo da estrutura dos valores omissos, os métodos correspondentes podem ajudar a identificar o mecanismo que gera os valores omissos e permite explorar dados, incluindo os valores omissos. Além disso, a qualidade da imputação pode

[4]Do inglês, *Visualization and imputation of missing values*

ser visualmente explorada usando vários métodos gráficos univariados, bivariados e multivariados. Finalmente, o *package* VIM possui uma interface gráfica (VIMGUI) que permite fácil manuseio dos métodos gráficos implementados. Além disto, o VIM também possui uma função $kNN()$ para imputação de valores omissos [TKAP19].

O ggplot2 é mais um *package* desenvolvido por Hadley Wickham [WG17], a ideia deste *package*, vem de uma obra chamada *The Grammar of Graphics*, que é uma maneira de descrever um gráfico a partir dos seus componentes. Dessa forma, ficaria teoricamente mais fácil entender a construção de gráficos mais complexos [FGM18]. Esse *package* é estruturado de forma que a gramática seja utilizada para um gráfico a partir de múltiplas camadas, que por sua vez são formatadas por dados, transformações estatísticas dos valores, objetos geométricos e ajuste de posicionamento.

3.6.2.5 pcaMethods e bPCA

Os *packages pcaMethods* e *bPCA* foram usados para imputar valores omissos e, assim, obter os conjuntos de valores completos estimados através do método ”bpca” da função ”pca()” (Anexo A.1.5). Além disto, estes *packages* implementam *biplot* (2D e 3D) de dados multivariados com base na análise de componentes principais e ferramentas de diagnóstico da qualidade da redução [JM18], [SRW19, SR15].

Entretanto, existe vária documentação de apoio à utilização dos *packages* apresentados e disponíveis *online*.

3.6.2.6 forecast

O *forecast*, é um dos *packages* R, que foi desenvolvido por Rob J Hyndman, conhecido por exibir e analisar previsões de séries temporais, incluindo suavização exponencial (*exponential smoothing* em inglês) por intermedio de modelos de espaço de estado e modelização automática ARIMA (*autregressive integrated moving average*) através dos métodos e ferramentas nele fornecidos [HAB+20]. Este *packages*, nos foi útil neste trabalho para avaliação do desempenho (cálculo de erros e tempos de execução) dos diferentes métodos de imputação utilizados no trabalho A.1.7.

3.6.2.7 xlsx e tidyverse

O xlsx é um *package* que fornece um controlo programático de ficheiros Excel no R, permite aos utilizadores fazer a leitura, escrita e manipulação de ficheiros (planilhas) xlsx em *data.frame* [DC20]. Já o *tidyverse* é uma coleção de *packages* R projetados para ciência de dados, serviu para ler os ficheiros csv (*comma separated values*), pode ser consultado em [Wic19] e `https://www.tidyverse.org`.

Capítulo 4

Estudo de Simulação

Os valores omissos prejudicam as análises e podem ocorrer devido a diversos mecanismos e fatores. Por exemplo, as pessoas podem deixar itens que lhes pareceram muito difíceis em branco, no meio do teste podem perder o interesse e pular algumas seções ou ainda podem se recusar a responder algum assunto delicado. Quando os inquiridos navegam livremente nos itens e só respondem a um número relativamente pequeno de exercícios de um número elevado que são normalmente fornecidos dentro de ambientes de aprendizagem, também pode surgir uma grande quantidade de valores omissos[WDV10].

Neste capítulo apresentam-se as análises de diferentes métodos de imputação usando vários conjuntos de dados onde se fez variar a percentagem de valores omissos. Apresentam-se, as estatísticas descritivas para o conjunto de dados completo e para os conjuntos dados antes e após a imputação de valores e apresenta-se o erro para cada processo de imputação. Foram utilizadas duas métricas diferentes para calcular os erros, a raiz quadrada da média do quadrado dos erros (RMSE) e a média absoluta dos erros (MAE). As várias etapas do estudo de simulação estão ilustrados na Figura 4.1.

Para instalação e execução deste estudo, utilizamos o computador *HP Pavilion g6 laptop* com instalação do Sistema Operativo Windows 10 Pro 64 bit, processador Intel(R) Core(TM) i5-2450M, *Radeon graphics*, SSD (*Hard disk*) e 8 GB de memória RAM instalada para pré-processamento de dados e os testes necessários.

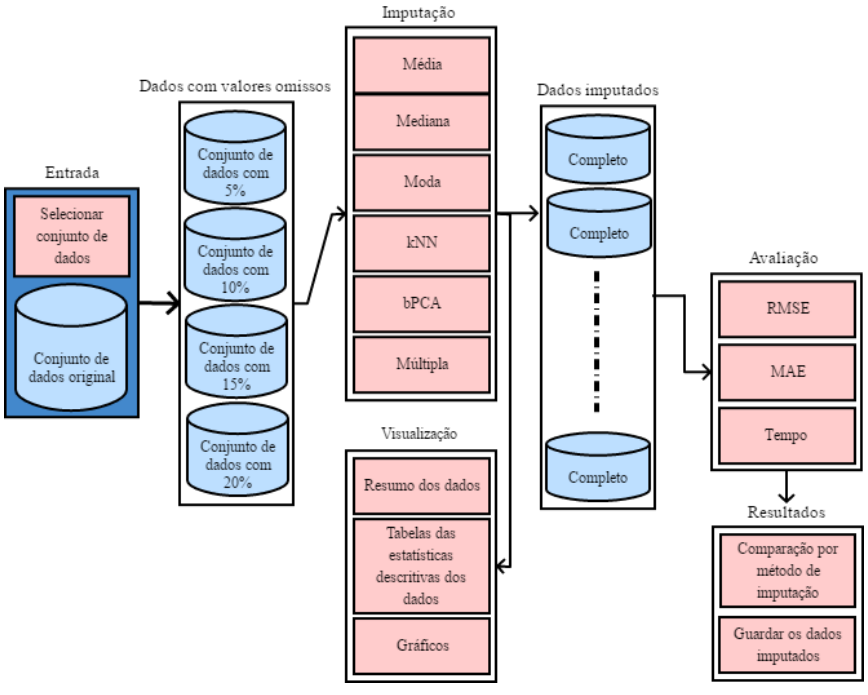


Figura 4.1: Etapas do estudo de simulação para tratamentos de valores omissos.

4.1 Descrição dos dados

Para avaliação dos diferentes métodos de imputação usados neste trabalho, foi considerado um conjunto de dados completo relativo ao estudo de Avaliação Nacional do Rendimento Escolar mais conhecida *Prova Brasil*. É uma avaliação bienal criada em 2005 e realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP/MEC), órgão responsável por desenvolver e aplicar avaliações educacionais bem como pelo censo da educação de todo território brasileiro, sob égide do Sistema Nacional de Avaliação da Educação Básica (SAEB), com o objetivo de avaliar a qualidade do ensino oferecido pelo sistema educacional brasileiro nas escolas públicas por intermédio de testes. É uma aplicação do tipo censos aos alunos de 5º e 9º anos de escolaridade do ensino primário em escolas com 20 ou mais alunos matriculados nestes níveis. Os testes são preparados e em conformidade a Matriz de Referência da Prova do Brasil e do SAEB, que contém a descrição de competências e habilidades dos alunos, padronizados, englobando Língua Portuguesa (leitura) e Matemática (resolução de problemas) além de questionário socioeconómico [Refb].

O conjunto de dados original (Prova Brasil 2017) contém informações relativas a centenas de milhares de estudantes com uma grande ocorrência de valores omissos. No entanto, para a simulação realizada neste trabalho, foi utilizada uma amostra composta com valores completos das variáveis utilizadas. Os mesmos dados utilizados por Ferrão e Prata [FP19], e que foram obtidos por eliminação *listwise*.

Partindo desse conjunto de dados com 20408 registos completos em três variáveis, foram gerados quatro conjuntos de dados com 5% (Miss5), 10% (Miss10), 15% (Miss15) e 20% (Miss20) de valores omissos. As variáveis consideradas foram: Y que adiante designamos por DL (Desempenho do estudante na Leitura), X1 que designamos por SSE (situação socioeconómica do estudante) e X2 que designamos por TSR (trajectória do estudante sem repetição). A geração dos quatro conjuntos de dados foi realizada através de extracções aleatórias nas variáveis DL e SSE, respectivamente no grupo de estudantes com desempenho mais baixo e no grupo de estudantes mais pobres. A variável "TSR" é uma variável dicotómica e sem valores omissos.

De notar que em [FP19] foi mostrado, por aplicação do teste de Little, que estes conjuntos de dados possuem um padrão de dados omissos não ignorável (mecanismo MNAR).

4.2 Estatísticas Descritivas

Na Tabela 4.1, são apresentadas as estatísticas descritivas do conjunto de dados original (completo) à esquerda e as estatísticas descritivas dos quatro subconjuntos de dados com valores omissos.

Tabela 4.1: Estatísticas descritivas dos conjuntos com dados omissos (antes da imputação).

	Originais		**Miss 5%**		**Miss 10%**		**Miss 15%**		**Miss 20%**	
	DL	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.500	4.400	-0.400	4.500	-0.300	4.600	-0.200	4.700
Mediana	0.040	4.960	0.100	5.000	0.200	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.151	5.166	0.208	5.229	0.261	5.289	0.297	5.327
3º Quartil	0.690	5.710	0.700	5.800	0.800	5.800	0.800	5.900	0.900	5.900
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.864**	**1.011**	**0.850**	**0.998**	**0.848**	**1.001**	**0.859**	**1.014**
NA's	0	0	983	969	1970	1979	3083	3115	4078	4036

Fonte: Dados da pesquisa (elaboração do autor)

Observa-se na última linha da tabela a quantidade de valores omissos nas duas variáveis de interesse de acordo com as respetivas percentagens da amostra. Nestes subconjuntos, 1952 estudantes da amostra não tinham as informações da situação socioeconómica (SSE) ou do desempenho na leitura (DL) para o primeiro subconjunto (Miss5); 3949 estudantes da amostra do segundo subconjunto (Miss10), faltaram-lhes informações sobre a situação socioeconómica (SSE) ou do desempenho na leitura (DL); Para o terceiro subconjunto (Miss15), 6198 estudantes não tinham informações sobre a situação socioeconómica (SSE) ou do desempenho na leitura (DL) e por último 8114 estudantes não tinham informações relacionadas com a situação socioeconómica (SSE) ou do desempenho na leitura (DL) na amostra (Miss20).

A Figura 4.2, ilustra o teste de verificação da existência de valores omissos no conjunto de dados original(completo). A cor azul representa que não existem valores omissos no conjunto de dados tendo no total 20408 observações completas para cada variável (TSR, SSE e DL); Os zeros abaixo representam o somatório de valores omissos e à direita representam a quantidade de padrões para cada linha. Este gráfico foi gerado usando a função R *md.pattern()* no *mice* e calcula também as frequências dos padrões de dados omissos.

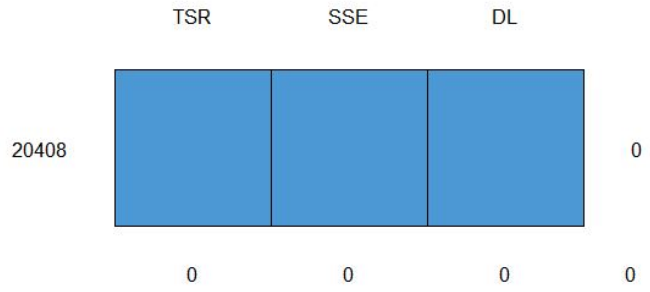
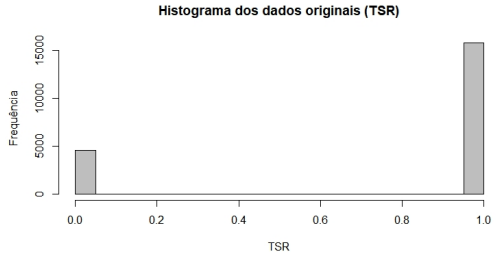
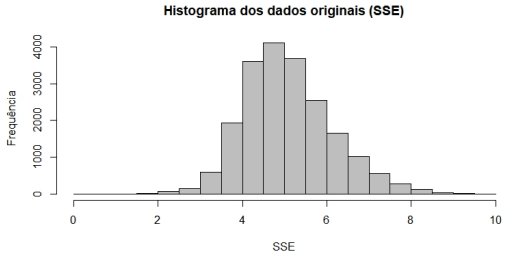


Figura 4.2: Representação gráfica de dados completos.

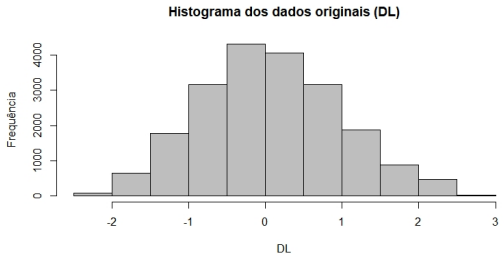
Para um melhor conhecimento do conjunto de dados gerou-se o histograma das três variáveis "TSR" 4.3a, "SSE" 4.3b e "DL" 4.3c de forma separada. Segundo a distribuição de valores verificados na Figura 4.3, nota-se que as variáveis em estudo possuem a maior quantidade de valores concentrados ao centro.



(a) Trajetória do estudante sem repetição (TSR)



(b) Situação Socioeconómica do Estudante (SSE)



(c) Desempenho do Estudante na Leitura (DL)

Figura 4.3: Histogramas das variáveis TSR (a), SSE (b) e DL (c) dos dados originais (sem *missing*).

4.3 Exploração de Valores Omissos

A exploração de valores omissos obedeceu os passos propostos por Harrel [Har19], tendo como base essencialmente ferramentas gráficas. Para o autor um investigador não deve adotar por uma análise de casos completos sem primeiramente fazer a exploração dos seus valores omissos para evitar tomadas de decisões inadequadas nos resultados e nas suas interpretações.

Na Figura 4.4, são mostrados os padrões de valores omissos em cada subconjunto de dados cujas áreas com a cor azul representam os dados observados e a rosa indica a localização dos valores não observados (omissos). No gráfico A (conjunto Miss5) observamos os seguintes padrões de valores omissos: 4,5% de valores omissos apenas na variável DL, 4,4% de valores omissos apenas na variável SSE e 0,3% de valores omissos nas duas variáveis. No gráfico B (conjunto Miss10) observamos 8,3% de omissos em SSE, 8,3% em DL e 1,4% de omissos nas duas variáveis. No gráfico C (Miss15) 12,3% de omissos em SSE, 12,3% em DL e 3,0% em ambas as variáveis. Finalmente no gráfico D (Miss20) 15,3% em DL, 15,0% em SSE e 4,6% em ambas as variáveis.

4.4 Imputação Simples

4.4.1 Imputação de Valores Omissos pela Média

O cálculo da média (ou moda) pode ser realizado através do uso de todas as instâncias observadas ou de todas as instâncias agrupadas por classe, também conhecida como imputação média condicional em que os valores omissos são estimados pela dos casos ob-

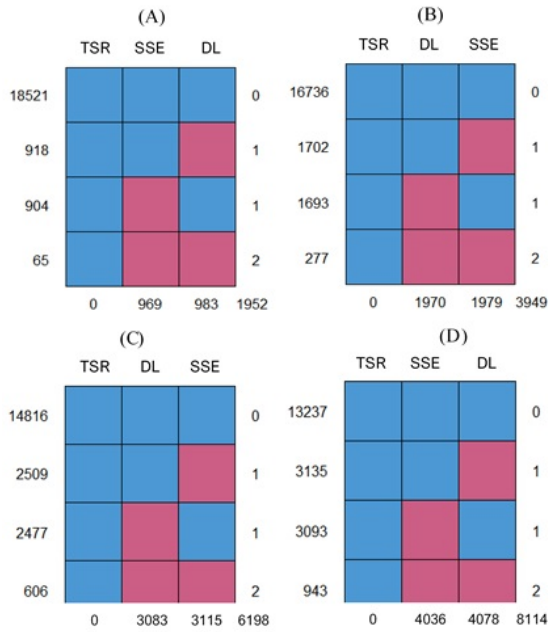


Figura 4.4: Representação gráfica de padrões de valores omissos por variável nos diferentes subconjuntos de dados. Sendo (A) com 5% de valores omissos; (B) com 10% de valores omissos; (C) com 15% de valores omissos e (D) com 20% de valores omissos.

servados (casos completos) que fazem parte da mesma classe como padrão incompleto [GLSGFV10]. A imputação simples pela média foi realizada através da substituição de valores omissos pela valor da média dos mesmos, isto é, referente a cada variável.

Para Schafer [Sch97] e Aljuaid [AS17], nesta abordagem, os valores omissos de um conjunto de dados (vetor) são preenchidos pelos valor médio da variável em todos os casos observados. Levando em consideração que existem valores omissos no *i-ésimo* atributo. Neste caso executa o processo de imputação calculando o estimador médio,

$$\bar{Y} = \frac{1}{N_{obs,i}} \sum_{n=1}^{N} \Big(m_{in}\Big) Y_{in} \tag{4.1}$$

Onde $\bar{Y}$ representa a média, $N_{obs,i}$ indica o número total de valores observados em Y_i, m_{in} valor observado na variável.

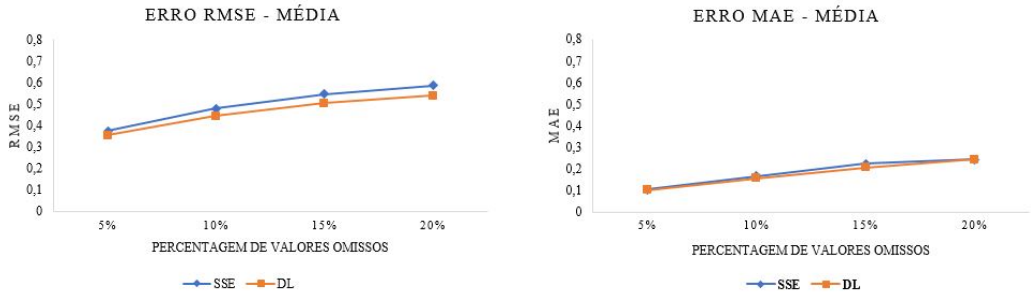
A descrição deste método e suas referências podem ser encontradas na seção 3.2.1 do Capítulo 3 e o respetivo código no Anexo A.1.1 desta dissertação. Foi usada a função ”colMeans()” que permite realizar o cálculo de médias sobre as colunas de uma matriz ou *data frame* (desde que este tenha apenas valores numéricos). Na Tabela 4.2 são apresentados os resultados das estatísticas descritivas para os conjuntos obtido por imputação pela média nas diferentes percentagens. Os valores com a cor azul são os que são diferentes dos apresentados na Tabela 4.1, depois da imputação.

Tabela 4.2: Estatísticas descritivas dos conjuntos com dados imputados pela Média.

	Originais		**Miss 5%**		**Miss 10%**		**Miss 15%**		**Miss 20%**	
	DL	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.400	4.500	-0.300	4.600	-0.200	4.700	-0.100	4.800
Mediana	0.040	4.960	0.151	5.100	0.208	5.200	0.261	5.289	0.297	5.327
Média	0.073	5.082	0.151	5.166	0.208	5.227	0.261	5.289	0.297	5.327
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.843**	**0.987**	**0.808**	**0.948**	**0.781**	**0.922**	**0.769**	**0.908**

Fonte: Dados da pesquisa (elaboração do autor)

A Figura 4.5 mostra a evolução dos erros, para cada uma das variáveis em estudo, quando se aumenta a percentagem de valores omissos. Apresentam-se os resultados para as duas métricas já referidas, RMSE 4.5a à esquerda e MAE 4.5b à direita comparando os valores observados e os valores imputados pela média.



(a) RMSE (b) MAE

Figura 4.5: Evolução de erros RMSE (a) e MAE (b) nas variações percentuais de valores omissos no conjunto de dados imputados pela Média.

Observando o comportamento dos gráficos, conclui-se que quanto mais cresce a percentagem de valores omissos maior é o erro e MAE $\leq$ RMSE.

4.4.2 Imputação de Valores Omissos pela Mediana

O método de imputação de valores omissos pela mediana é considerado não condicional, visto que preenche o valor omisso pelo valor da mediana dos valores observados da variável em análise na amostra. Na Tabela 4.3 apresentam-se os resultados das estatísticas descritivas utilizando os valores imputados pelo método de imputação pela mediana para DL e SSE, para as diferentes percentagens de valores omissos. O código utilizado encontra-se no Anexo A.1.2, cuja a vantagem é de permitir uma simples implementação em preencher todos valores omissos por uma constante, a mediana da mesma variável ou atributo; adaptando o método ”median” da função ”na.mean” disponível no *package* ”imputeTS”. Mais detalhes e referências sobre *imputeTS* podem ser consultados na subsubseção 3.6.2.2.

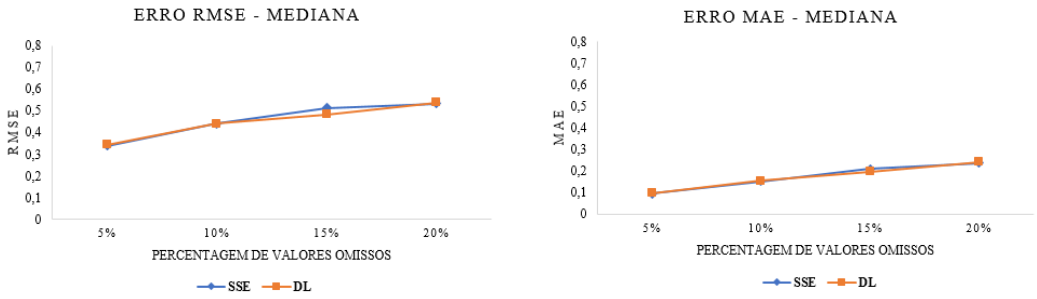
Tabela 4.3: Estatísticas descritivas dos conjuntos com dados imputados pela Mediana

	Originais		Miss 5%		Miss 10%		Miss 15%		Miss 20%	
	DL	SSE	DL	SSE	DL	SSE	DL	SSE	DL	SSE
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.400	4.500	-0.300	4.600	-0.200	4.700	-0.100	4.800
Mediana	0.040	4.960	0.100	5.000	0.200	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.148	5.158	0.207	5.217	0.252	5.275	0.298	5.302
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.843**	**0.988**	**0.808**	**0.949**	**0.782**	**0.922**	**0.769**	**0.910**

Fonte: Dados da pesquisa (elaboração do autor)

Tal como no método anterior, foram adotadas duas métricas para o cálculo de erros para avaliação dos resultados das imputação em diferentes percentagens de valores omissos, sendo a raiz quadrada da média do quadrado dos erros (RMSE) e a média absoluta dos erros (MAE) e podem ser obtidos pelas fórmulas (2.7 e 2.8) apresentadas e descritas na Subseção 2.4. O código R utilizado é apresentado no Anexo A.1.7.

A Figura 4.6 mostra a evolução dos erros das simulações de acordo com cada variável de interesse nos quatro subconjuntos com valores omissos, usando RMSE 4.6a à esquerda e MAE 4.6b à direita para os valores observados e valores imputados pela mediana.



(a) RMSE (b) MAE

Figura 4.6: Evolução de erros RMSE (a) e MAE (b) nas variações percentuais de valores omissos no conjunto de dados imputados pela Mediana.

Observando o comportamento dos gráficos para os tipos de erro, conclui-se que quanto mais cresce a percentagem de valores omissos maior é o erro. Também nota-se que o RMSE $\geq$ MAE.

4.4.3 Imputação de Valores Omissos pela Moda

Sabendo que a moda de uma série de valores é o valor que aparece com maior frequência; a imputação pela moda, foi utilizada para preencher os valores omissos pelo valor que mais aparece na variável em estudo, tendo em conta todas observações da amostra.

A Tabela 4.4 ilustra os resultados das estatísticas descritivas na aplicação da imputação de valores omissos pela moda para as variáveis DL e SSE, nas diferentes percentagens de valores omissos, adaptando o método ”mode” da função ”na.mean” implementado no

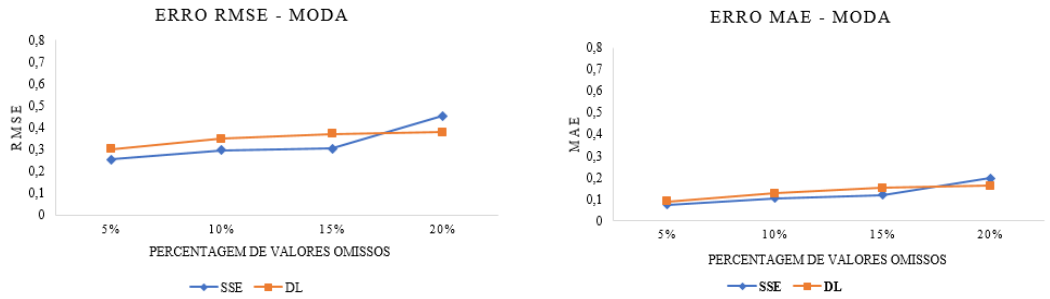
package "imputeTS". Mais detalhes e referências sobre *imputeTS* podem ser consultados na subsubseção 3.6.2.2. O código utilizado pode ser encontrado no Anexo A.1.3.

Tabela 4.4: Estatísticas descritivas dos conjuntos com dados imputados pela Moda

	Originais		**Miss 5%**		**Miss 10%**		**Miss 15%**		**Miss 20%**	
	DL	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.400	4.500	-0.300	4.600	-0.200	4.600	-0.100	4.800
Mediana	0.040	4.960	0.000	5.000	0.000	5.000	0.000	5.000	0.000	5.000
Média	0.073	5.082	0.139	5.139	0.178	5.168	0.206	5.184	0.218	5.263
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.845**	**0.994**	**0.813**	**0.966**	**0.792**	**0.955**	**0.785**	**0.917**

Fonte: Dados da pesquisa (elaboração do autor)

A Figura 4.7 mostra a evolução dos erros das simulações de acordo com cada variável de interesse nos quatro subconjuntos com valores omissos, usando RMSE 4.7a à esquerda e MAE 4.7b à direita para os valores observados e valores imputados pela moda.



(a) RMSE (b) MAE

Figura 4.7: Evolução de erros RMSE (a) e MAE (b) nas variações percentuais de valores omissos no conjunto de dados imputados pela Moda.

Ao observar a Figura 4.7 nota-se que o comportamento do erro cresce ligeiramente na medida que cresce a percentagem de valores omissos. Também nota-se que o RMSE $\geq$ MAE.

4.5 Imputação de Dados com Métodos Baseados em Aprendizagem Automática

Neste trabalho, foram utilizados dois métodos e *packages* diferentes para imputar valores omissos por *K-nearest neighbors* (kNN) e *Bayesian Principal Component Analysis* (bPCA) disponíveis no *software R*. Além destes existem outras técnicas e *packages* de imputação baseadas em Aprendizagem Automática disponibilizadas em *softwares* estatísticos convencionais, por exemplo *sklearn, tensorflow* do Python, etc.

4.5.1 Imputação de Valores Omissos com kNN

Como foi dito anteriormente no Capítulo 3, no método kNN, os k vizinhos são escolhidos em função de alguma medida de distância e usa a sua média como estimativa de imputação. A descrição detalhada e referências do método k vizinhos mais próximos (kNN) utilizado nesta dissertação, podem ser encontradas na subseção 3.3.1 do Capítulo anterior. O código está disponível no Anexo A.1.4, com o parâmetro *k = 6* vizinhos na função kNN disponível na biblioteca.

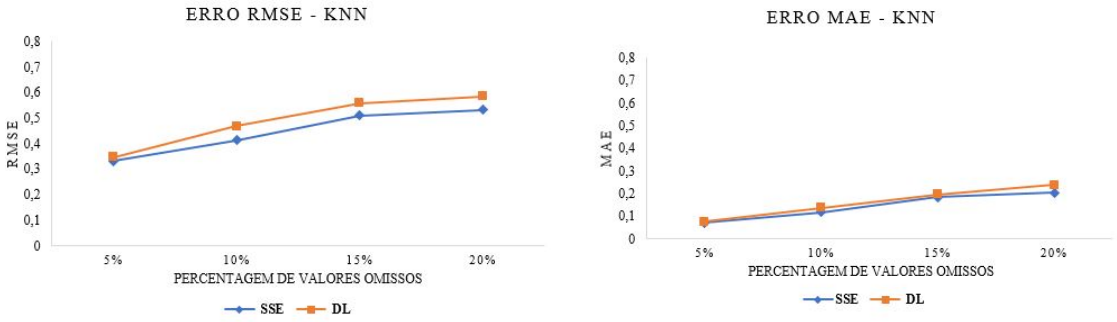
A Tabela 4.5 mostra-nos as estatísticas descritivas dos valores imputados usando a imputação com k-vizinhos mais próximos (kNN) para cada variável observada, na simulação dos quatro diferentes subconjuntos de dados em estudo.

Tabela 4.5: Estatísticas descritivas dos conjuntos com dados imputados com kNN.

	Originais		**Miss 5%**		**Miss 10%**		**Miss 15%**		**Miss 20%**	
	DL	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.500	4.450	-0.400	4.600	-0.300	4.700	-0.100	4.750
Mediana	0.040	4.960	0.100	5.000	0.150	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.144	5.152	0.205	5.199	0.244	5.265	0.306	5.281
3º Quartil	0.690	5.710	0.700	5.700	0.800	5.700	0.800	5.800	0.800	5.800
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.853**	**0.994**	**0.828**	**0.966**	**0.849**	**0.937**	**0.800**	**0.939**

Fonte: Dados da pesquisa (elaboração do autor)

A Figura 4.8 mostra a evolução dos erros das simulações de acordo com cada variável em estudo e para os quatro subconjuntos com valores omissos, utilizando duas métricas diferentes, RMSE 4.8a à esquerda e MAE 4.8b à direita para os valores observados e valores imputados pelo método de k-Vizinhos mais próximos (kNN).



(a) RMSE (b) MAE

Figura 4.8: Evolução de erros RMSE (a) e MAE (b) nas variações percentuais de valores omissos no conjunto de dados imputados com KNN

Conclui-se que quanto mais cresce a percentagem de valores omissos maior é o erro, com e o MAE $\leq RMSE$.

4.5.2 Imputação de Valores Omissos com bPCA

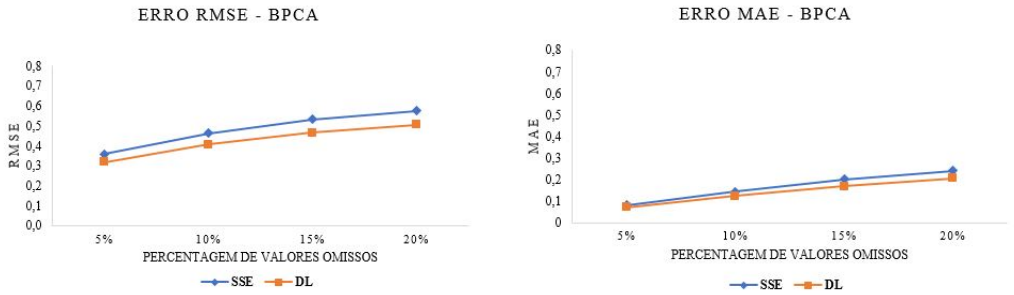
As estatísticas descritivas relacionadas às imputações realizadas com análise de componentes principais bayesiano (bPCA), são mostradas na Tabela 4.6 para cada percentagem de valores omissos; utilizando o método ”bpca” da função ”pca()” com número de componentes (nPcs = 2) disponível nos *bpca* que é similar ao imputePCA do *pcaMethods*. O código utilizado está disponível no Anexo A.1.5.

Tabela 4.6: Estatísticas descritivas dos conjuntos com dados imputados com bPCA.

	Originais		**Miss 5%**		**Miss 10%**		**Miss 15%**		**Miss 20%**	
	DL	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.489	4.500	-0.400	4.600	-0.293	4.700	-0.200	4.800
Mediana	0.040	4.960	0.100	5.000	0.200	5.100	0.200	5.200	0.300	5.292
Média	0.073	5.082	0.141	5.163	0.193	5.225	0.240	5.283	0.274	5.320
3º Quartil	0.690	5.710	0.700	5.700	0.700	5.700	0.700	5.700	0.700	5.700
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.848**	**0.988**	**0.816**	**0.949**	**0.794**	**0.924**	**0.786**	**0.911**

Fonte: Dados da pesquisa (elaboração do autor)

A Figura 4.9 mostra os erros RMSE (4.9a à esquerda) e MAE (4.9b à direita) da imputação utilizando o método bPCA com diferentes percentagens de valores omissos. Calculados através das fórmulas 2.7 para RMSE e 2.8 para MAE, cujo código R utilizado pode ser encontrado no Anexo A.1.7.



(a) RMSE (b) MAE

Figura 4.9: Evolução de erros RMSE (a) e MAE (b) nas variações percentuais de valores omissos no conjunto de dados imputados com bPCA

4.6 Imputação Múltipla

A imputação múltipla neste trabalho, foi realizada através *Multivariate Imputation by Chained Equations* (MICE) disponível no *software* R.

Com intuito de tentar construir um método que reflita a incerteza sobre as previsões de valores omissos, Rubin [LR87] descreve o método de imputação múltipla, o qual substitui cada valor omisso por um conjunto de valores plausíveis que representa esta incerteza

sobre o valor a ser imputado. Mais detalhes e referências deste método encontram-se nas subseções 3.4, 3.4.2 e 3.6.2.3.

A seguir, na Tabela 4.7, são apresentadas as estatísticas descritivas ou medidas resumos dos quatro subconjuntos de valores imputados com MICE, em função da percentagem de valores omissos de cada subconjunto e considerando as duas variáveis em estudo.

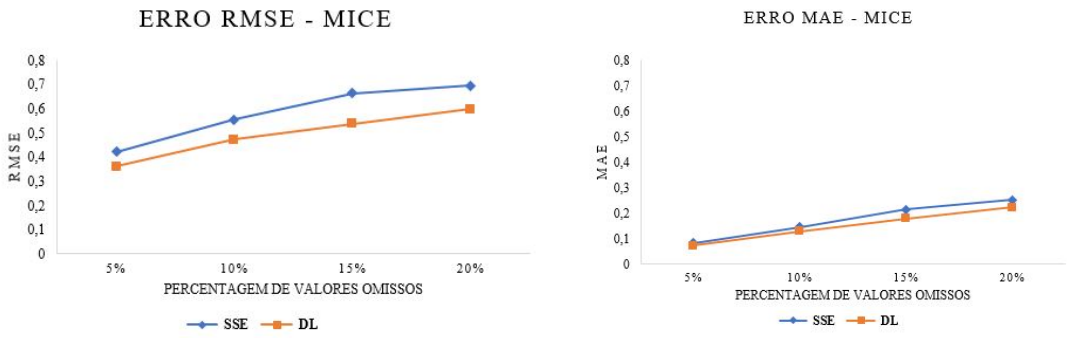
Utilizou-se a função *mice()* e o método "PMM" (*Predictive Mean Matching*) disponível no *package* "mice", com m=5 número de conjuntos de dados imputados, 50 iterações realizadas para imputar os valores omissos, o valor da semente de gerador aleatório de 500, *md.pattern* para visualização do resumo de padrões dos valores omissos. O código usado está disponível no Anexo A.1.6.

Tabela 4.7: Estatísticas descritivas dos conjuntos com dados imputados com MICE.

	Originais		**Miss 5%**		**Miss 10%**		**Miss 15%**		**Miss 20%**	
	DL	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**	**DL**	**SSE**
Mínimo	-2.320	0.340	-2.300	0.300	-2.300	0.300	-2.300	1.200	-2.300	1.500
1º Quartil	-0.570	4.350	-0.500	4.400	-0.400	4.500	-0.300	4.600	-0.200	4.700
Mediana	0.040	4.960	0.100	5.000	0.100	5.100	0.200	5.200	0.300	5.200
Média	0.073	5.082	0.141	5.162	0.192	5.222	0.231	5.282	0.263	5.309
3º Quartil	0.690	5.710	0.700	5.700	0.800	5.800	0.800	5.900	0.800	5.900
Máximo	2.510	9.710	2.500	9.700	2.500	9.700	2.500	9.700	2.500	9.700
Desvio Pad.	**0.907**	**1.051**	**0.865**	**1.013**	**0.851**	**1.001**	**0.852**	**1.005**	**0.866**	**1.009**

Fonte: Dados da pesquisa (elaboração do autor)

A Figura 4.10 mostra os erros RMSE (4.10a à esquerda) e MAE (4.10b à direita), da previsão em função da percentagem de valores omissos em cada subconjunto de dados, seja para a variável SSE como para DL, o erro após a imputação com MICE parece crescer mais do que outros à medida que a percentagem de valores omissos aumenta. O código e funcionamento utilizado para o cálculo de erros encontra-se no Anexo A.1.7.

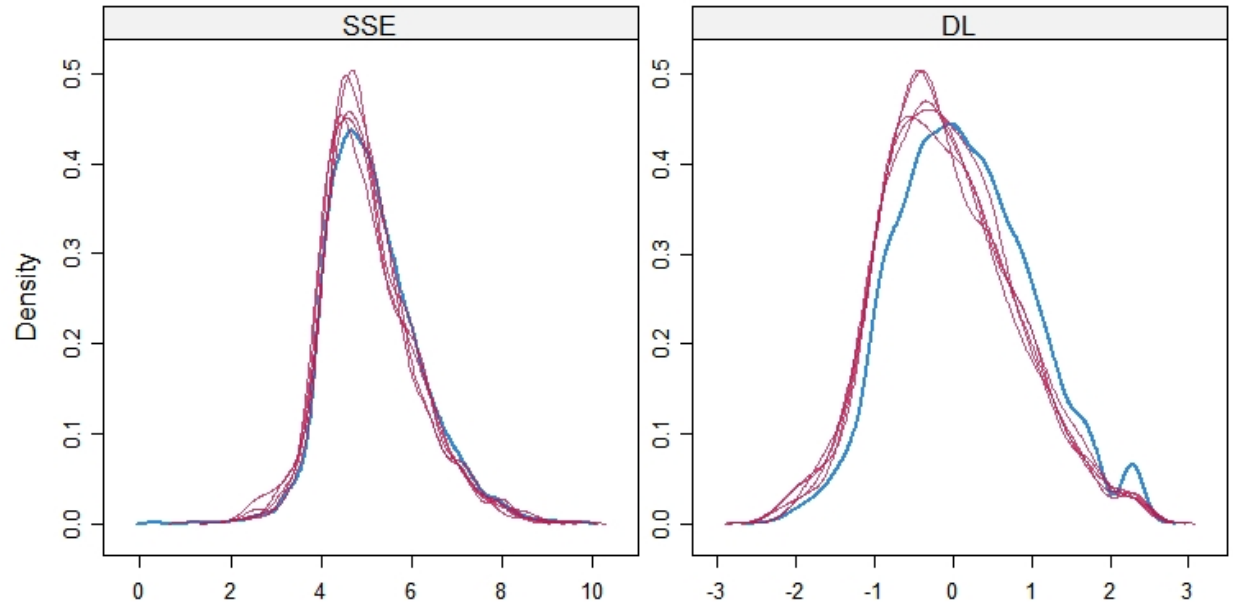


(a) RMSE (b) MAE

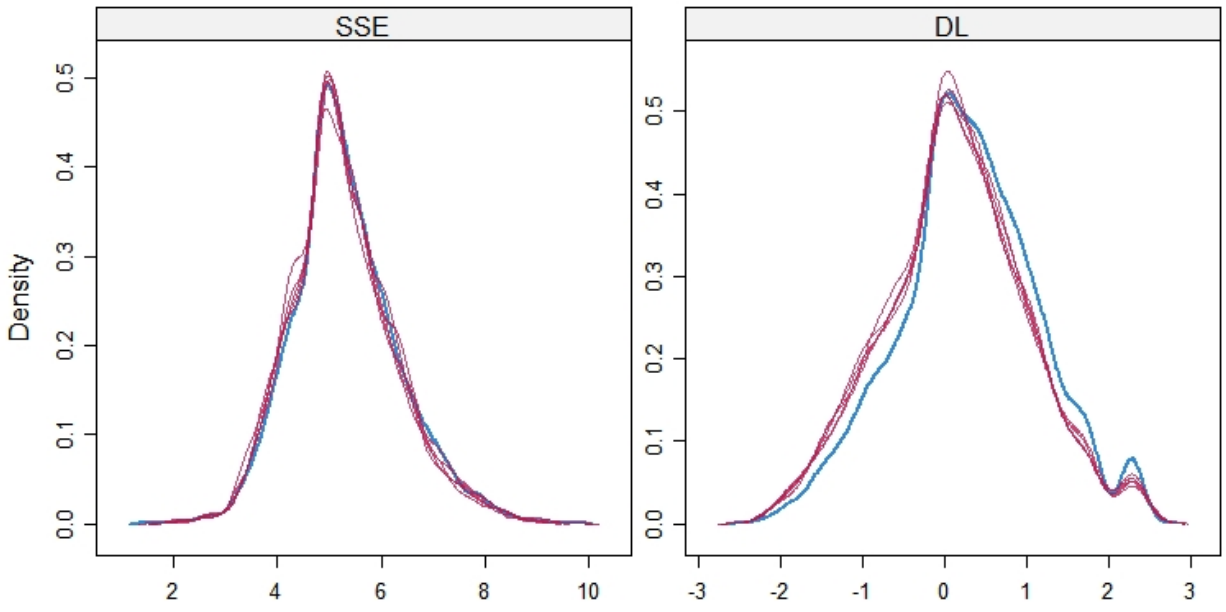
Figura 4.10: Evolução de erros RMSE (a) e MAE (b) nas variações percentuais de valores omissos no conjunto de dados imputados com MICE

Na Figura 4.11, a distribuição dos valores imputados para os conjuntos de dados *Miss5* e *Miss20* apresentada em magenta, sendo a distribuição dos valores observados ilustrados em cor azul. A Figura 4.12 mostra as distribuições de cada variável como ponto individual,

para os mesmos conjuntos de dados.

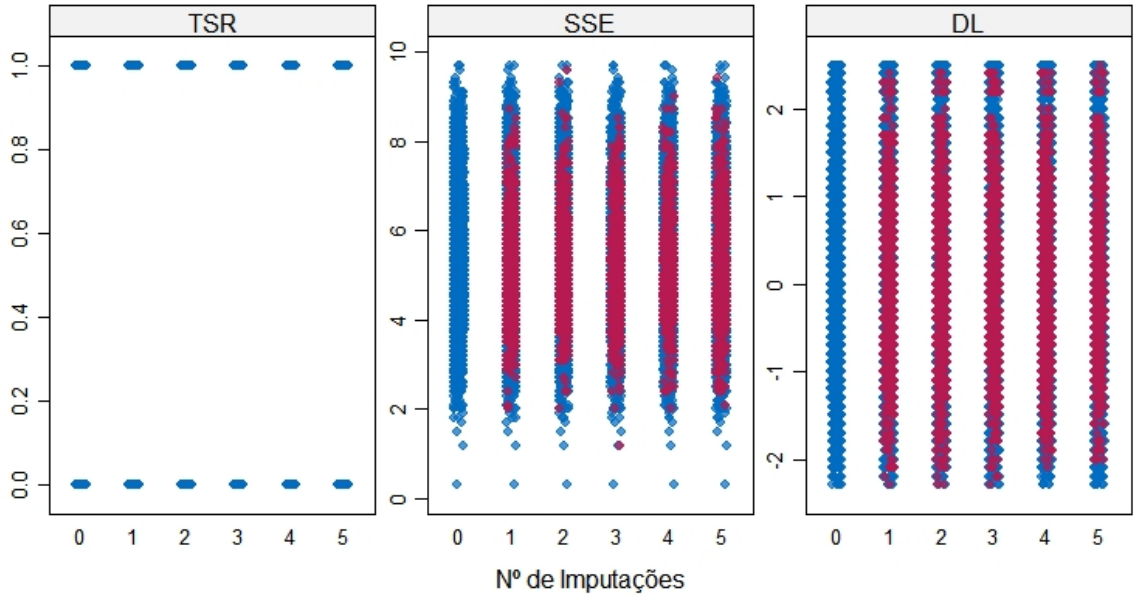


(a) Valores imputados do subconjunto com 5% de valores omissos

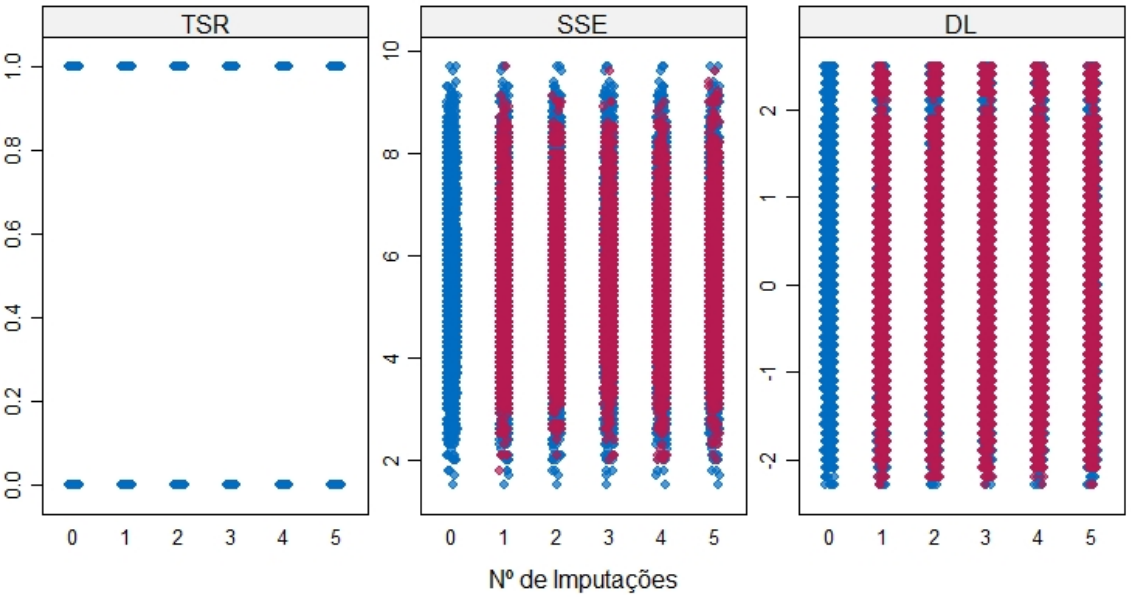


(b) Valores imputados do subconjunto com 20% de valores omissos

Figura 4.11: Função densidade dos valores observados e dos valores imputados em Miss5 e Miss20 com MICE.



(a) Valores imputados do subconjunto com 5% de valores omissos



(b) Valores imputados do subconjunto com 20% de valores omissos

Figura 4.12: Distribuição dos valores observados e dos valores imputados em Miss5 e Miss20 com MICE.

Capítulo 5

Análise dos Resultados

Neste capítulo é apresentada uma análise comparativa dos resultados obtidos no capítulo anterior. Comparam-se os valores do erro RMSE e MAE obtidos para cada método, por variável e por percentagem de valores omissos e adicionalmente apresentam-se os tempos de execução de cada um dos métodos. Finalmente, apresentamos os gráficos com a função densidade de probabilidade para o conjunto de dados observados e para os dados imputados pelo método que apresentou melhores resultados (Moda) e para o método de imputação simples que apresentou piores resultados (Média) para ambas as variáveis e para todas as percentagens de valores omissos estudados.

A Tabela 5.1 apresenta o erro quadrático médio (RMSE) para a variável SSE, para todas as percentagens de valores omissos estudadas e para todos os métodos de imputação. Como se pode observar a imputação pela Moda foi o método que obteve os menores valores de erro. A imputação múltipla, por MICE, obteve os piores resultados. Podemos também observar que, como seria de esperar, quanto maior a percentagem de valores omissos maior o erro obtido, e este padrão ocorre em todos os casos estudados. A Tabela 5.2 apresenta o erro RMSE para a variável DL, para todas as percentagens de valores omissos estudadas e para todos os métodos de imputação. Também para esta variável o menor erro RMSE ocorre com a imputação pela Moda. Os piores casos ocorrem com o MICE, excepto para o conjunto Miss15 em que o pior método foi o kNN.

As tabelas 5.3 e 5.4 apresentam o erro absoluto médio (MAE) para todos os métodos de imputação para todas as percentagens de valores omissos, respectivamente para a variável SSE e DL. Aqui, a imputação pela Moda continua a ser o melhor para os conjuntos com mais valores omissos mas já não é sempre a que tem menor erro. Para a variável SSE no conjunto Miss5, é o método kNN que apresenta melhores valores. Para a variável DL, é o método bPCA que apresenta os melhores resultados para os conjuntos Miss5 e Miss10. O método que apresenta o erro MAE mais elevado é a Média, com uma única exceção para Miss20 na vaiável SSE em que MICE é pior.

Tabela 5.1: Comparação do erro (RMSE) de métodos das imputações para SSE nas diferentes percentagens de valores omissos.

Erro RMSE para SSE				
	Miss5	**Miss10**	**Miss15**	**Miss20**
Método de imputação	**SSE**	**SSE**	**SSE**	**SSE**
Média	0,374	0,479	0,547	0,586
Mediana	0,338	0,440	0,514	0,533
Moda	**0,255**	**0,296**	**0,304**	**0,453**
KNN	0,332	0,414	0,511	0,532
bPCA	0,360	0,464	0,532	0,574
MICE	0,421	0,556	0,663	0,696

Fonte: Dados de pesquisa (elaboração do autor).

Tabela 5.2: Comparação do erro (RMSE) de métodos das imputações para DL nas diferentes percentagens de valores omissos.

Erro RMSE para DL				
	Miss5	**Miss10**	**Miss15**	**Miss20**
Método de imputação	**DL**	**DL**	**DL**	**DL**
Média	0,355	0,443	0,504	0,538
Mediana	0,344	0,440	0,482	0,538
Moda	**0,301**	**0,350**	**0,373**	**0,378**
KNN	0,346	0,469	0,559	0,585
bPCA	0,319	0,408	0,465	0,507
MICE	0,362	0,472	0,538	0,599

Fonte: Dados de pesquisa (elaboração do autor).

Tabela 5.3: Comparação do erro (MAE) de métodos das imputações para SSE nas diferentes percentagens de valores omissos.

Erro MAE para SSE				
	Miss5	**Miss10**	**Miss15**	**Miss20**
Método de imputação	**SSE**	**SSE**	**SSE**	**SSE**
Média	0,103	0,166	0,224	0,243
Mediana	0,095	0,153	0,211	0,237
Moda	0,076	**0,105**	**0,119**	**0,197**
KNN	**0,071**	0,118	0,184	0,204
bPCA	0,081	0,143	0,200	0,240
MICE	0,082	0,146	0,214	0,252

Fonte: Dados de pesquisa (elaboração do autor).

Tabela 5.4: Comparação do erro (MAE) de métodos das imputações para DL nas diferentes percentagens de valores omissos.

Erro MAE para DL				
	Miss5	**Miss10**	**Miss15**	**Miss20**
Método de imputação	**DL**	**DL**	**DL**	**DL**
Média	0,101	0,156	0,207	0,243
Mediana	0,098	0,155	0,198	0,243
Moda	0,089	0,126	**0,153**	**0,163**
KNN	0,075	0,136	0,195	0,237
bPCA	**0,072**	**0,123**	0,169	0,206
MICE	0,073	0,129	0,178	0,223

Fonte: Dados de pesquisa (elaboração do autor).

Nas Figuras 5.1 e 5.2, apresentam-se graficamente os resultados das tabelas anteriores, respectivamente para o erro RMSE e para o erro MAE. Aqui podemos observar mais facilmente que o maior errro RMSE foi introduzido pelo MICE no caso da variável SSE, e para a variável DL o MICE e o kNN partilham os piores resultados. Para o erro MAE, as diferenças são bastante diluídas. A Moda é o método que na generalidade dos casos apresenta menor erro, mas a diferença para os outros métodos só é clara para os conjuntos com mais valores omissos.

Enquanto com o RMSE o MICE foi o pior método com valores de erro mais acentuados a seguir da média, a moda foi consistentemente melhor método entre as diferentes percentagens de valores omissos e métricas de avaliação de desempenho, superando todos os outros métodos quando aplicado ao conjunto de dados estudados com base nos critérios de RMSE e MAE.

Na Tabela 5.5, são apresentados os valores dos tempos médios de execução para cada método de imputação usado, e a Figura 5.3 apresenta graficamente os mesmos resultados. Para medir os tempos foram utilizadas duas funções $tic()$ e $toc()$ disponíveis no *package* $tictoc$.

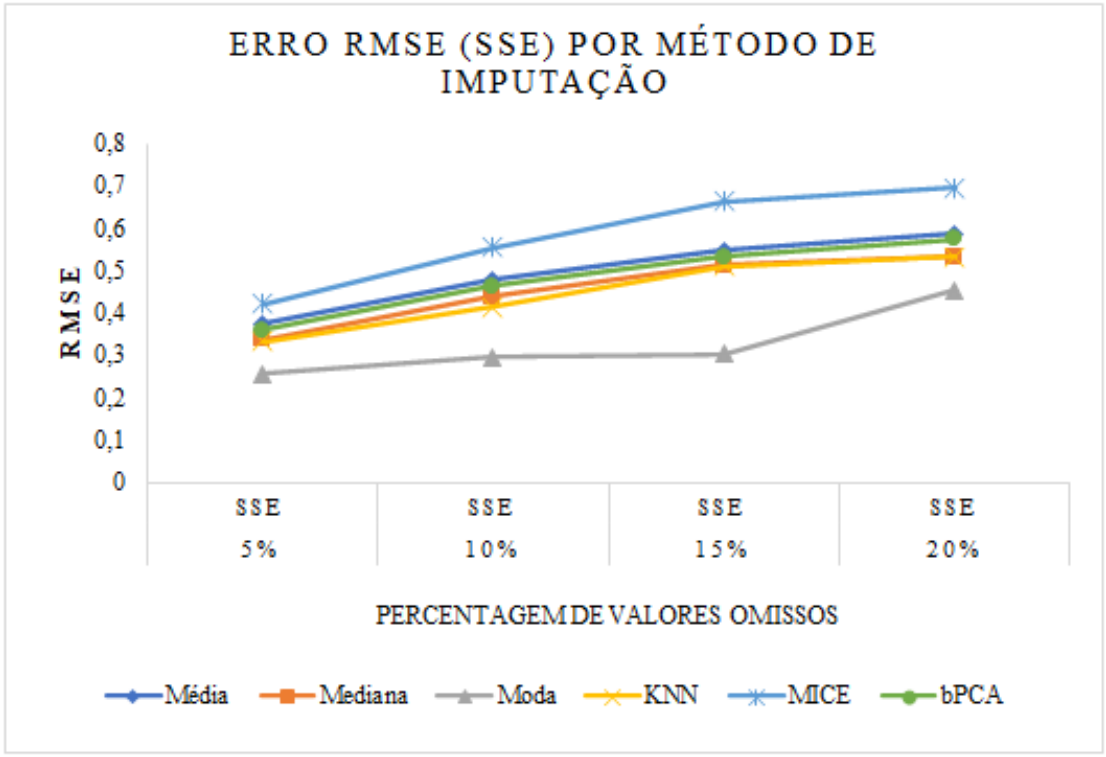
Podemos observar que os métodos de imputação pela Média, Mediana e Moda foram todos muito mais rápidos, com uma duração de algumas décimas de segundos.

Os métodos kNN e bPCA, para estes conjuntos de dados, tem tempos de execução na ordem das dezenas de segundos, enquanto o MICE tem um tempo de execução na ordem das várias centenas de segundos. De notar que no MICE o número de iterações depende da convergência da imputação que neste estudo não foi avaliada.

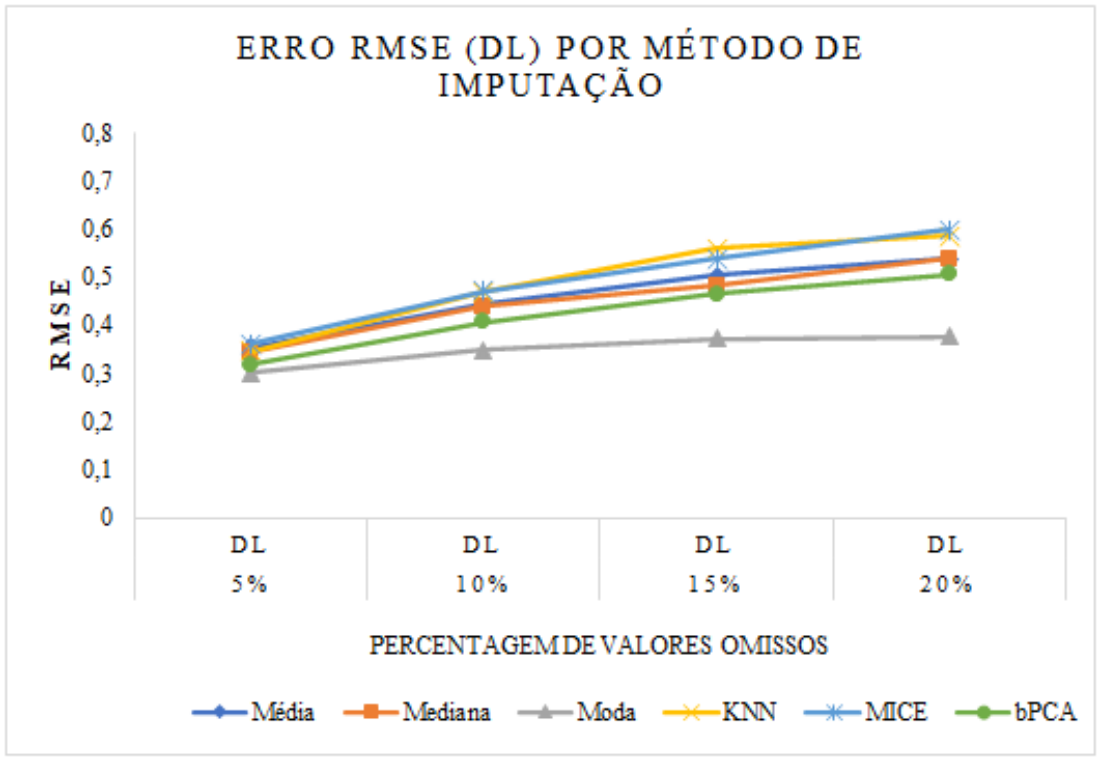
Tabela 5.5: Tempos médios de execução em segundos por método de imputação

	Média	**Mediana**	**Moda**	**KNN**	**bPCA**	**MICE**
Miss5	0,08	**0,07**	0,17	34,49	26,39	236,71
Miss10	0,1	**0,09**	0,17	49,39	48,22	504,08
Miss15	**0,08**	**0,08**	0,17	81,19	75,44	739,25
Miss20	**0,09**	**0,09**	0,16	104,46	99,52	887,09

Fonte: Dados da pesquisa (elaboração do autor)



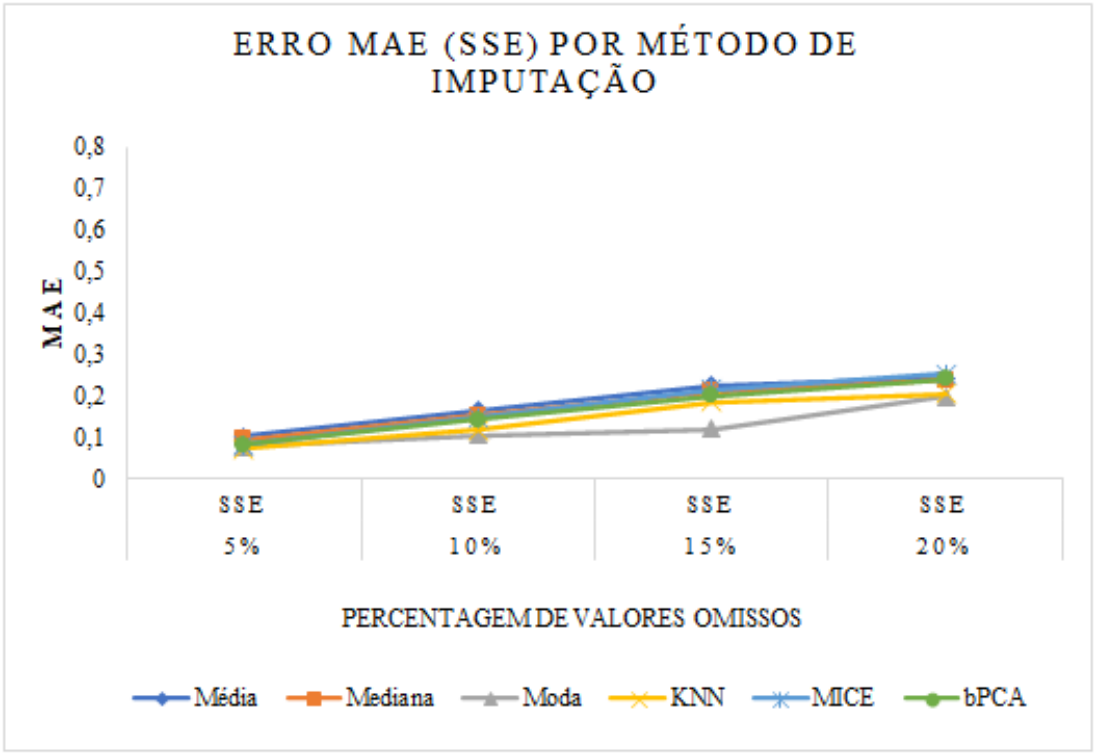
(a) Situação Socioeconómica do Estudante - SSE



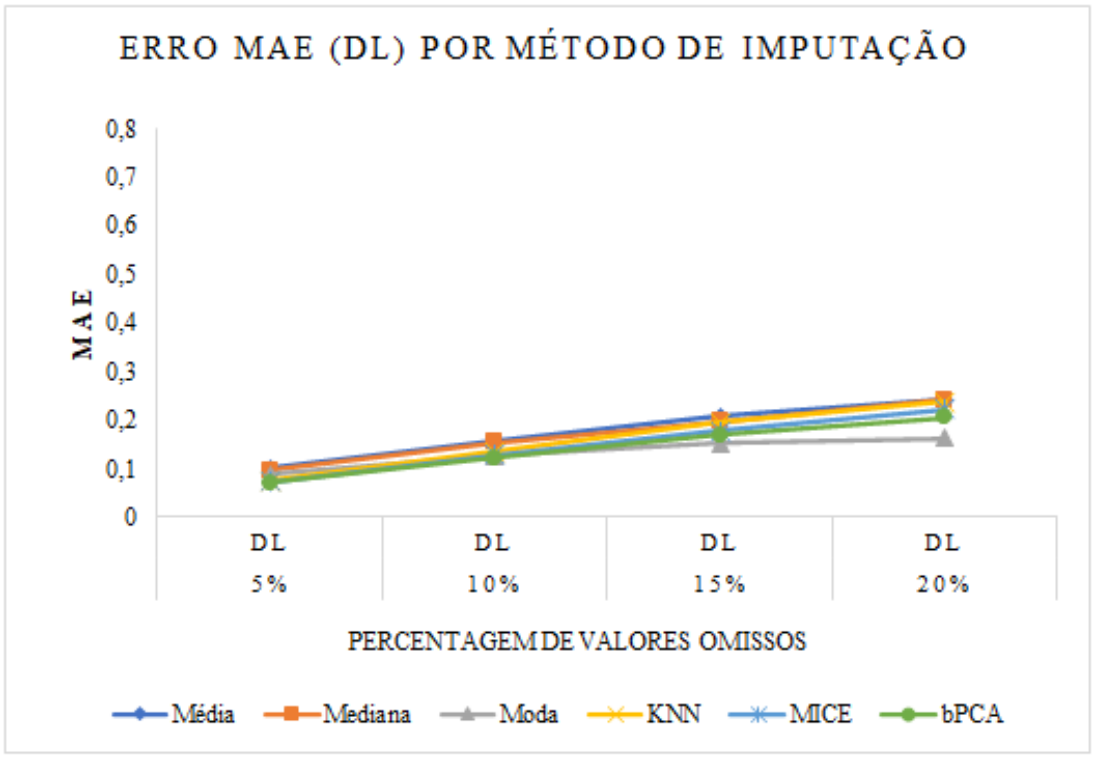
(b) Desempenho do Estudante na Leitura - DL

Figura 5.1: Comportamento de RMSE para os vários métodos de imputação nos diferentes conjuntos de dados imputados.

A Figura 5.3, ilustra a evolução dos tempos médios da execução nos diferentes conjuntos de dados por método de imputação.



(a) Situação Socioeconómica do Estudante - SSE



(b) Desempenho do Estudante na Leitura - DL

Figura 5.2: Comportamento de MAE para os vários métodos de imputação nos diferentes conjuntos de dados imputados.

A Figura 5.4, representa os gráficos da função densidade de probabilidade antes e depois da imputação pela Média, método de imputação simples com piores resultados, para todos

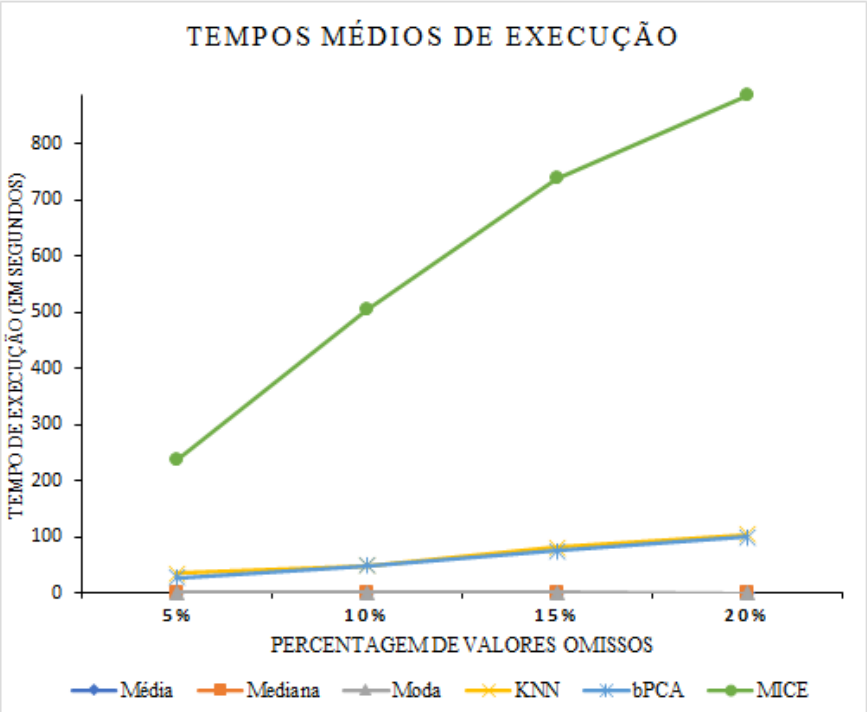
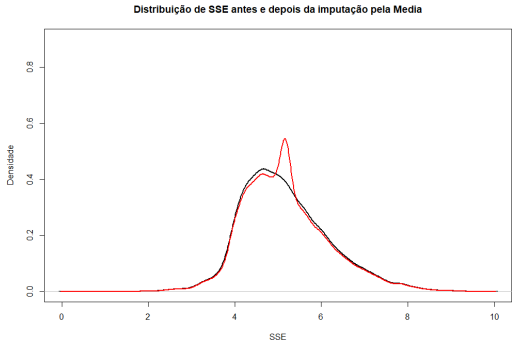
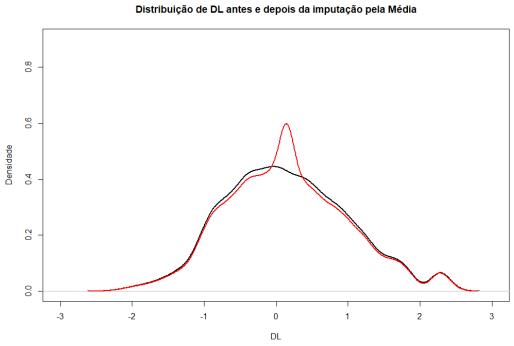


Figura 5.3: Evolução dos tempos médios da execução nas diferentes conjuntos de dados por método de imputação.

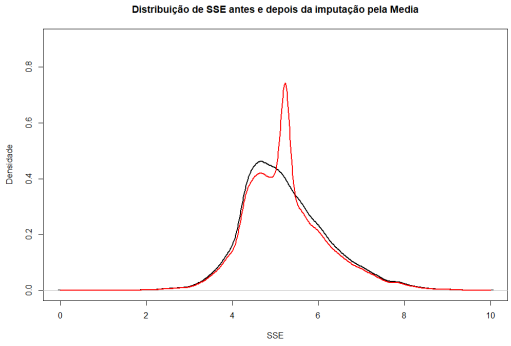
subconjuntos e para as duas variáveis consideradas. A preto estão os dados observados e a vermelho os dados após imputação. A Figura 5.5 representa os mesmos gráficos para a imputação pela Moda, método com os valores de erro mais baixos.



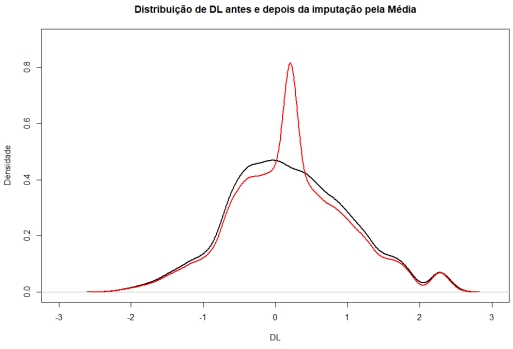
(a) Situação Socioeconómica do Estudante - SSE (5%)



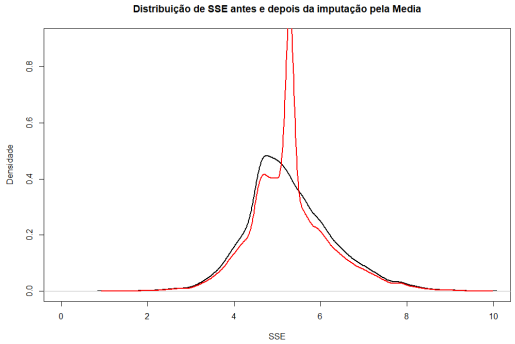
(b) Desempenho do Estudante na Leitura - DL(5%)



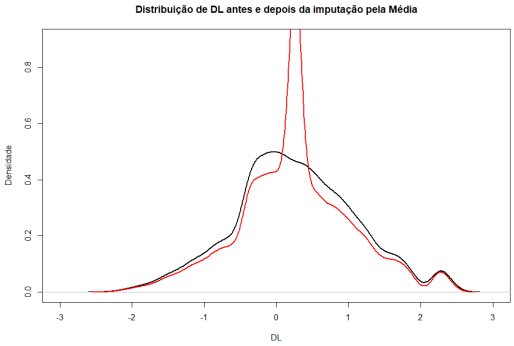
(c) Situação Socioeconómica do Estudante - SSE(10%)



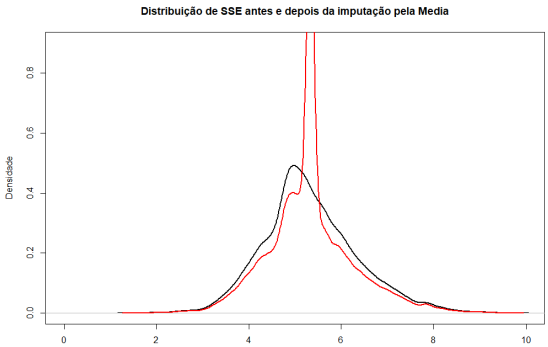
(d) Desempenho do Estudante na Leitura - DL(10%)



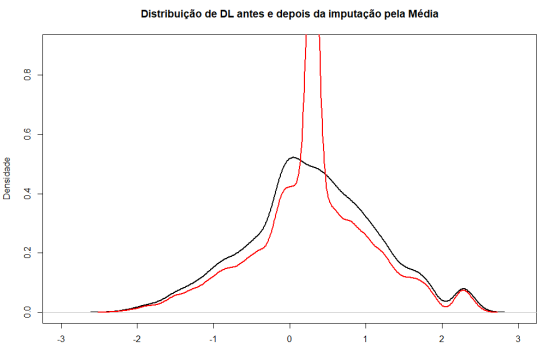
(e) Situação Socioeconómica do Estudante - SSE(15%)



(f) Desempenho do Estudante na Leitura - DL(15%)

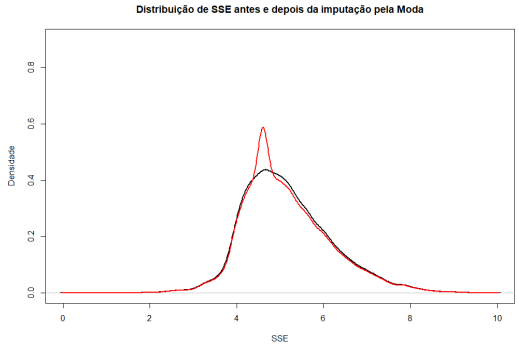


(g) Situação Socioeconómica do Estudante - SSE(20%)

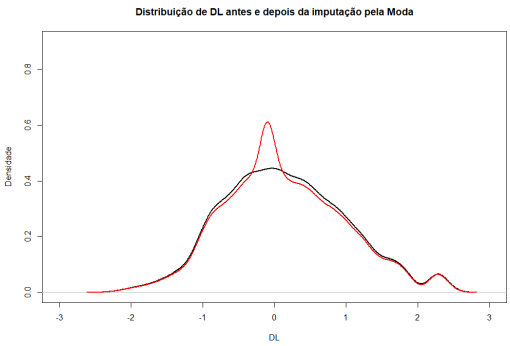


(h) Desempenho do Estudante na Leitura - DL(20%)

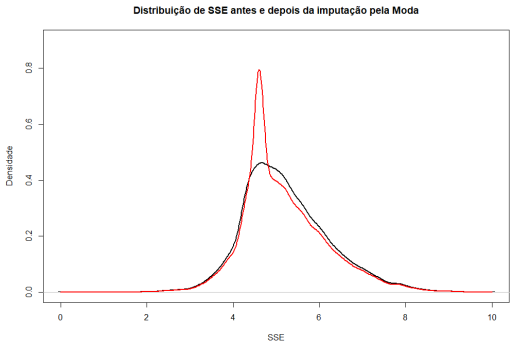
Figura 5.4: Comportamento da distribuição de SSE e DL antes e depois das imputações pela Média com 5, 10, 15 e 20% de valores omissos.



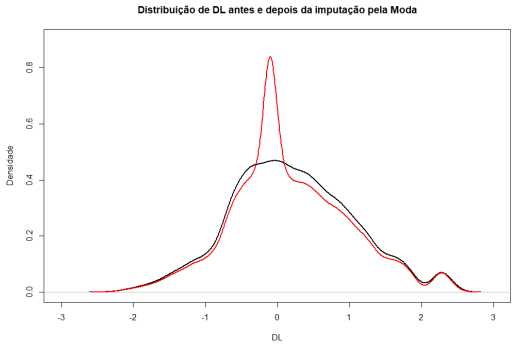
(a) Situação Socioeconómica do Estudante - SSE (5%)



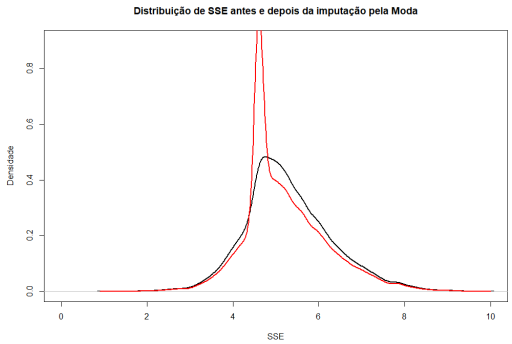
(b) Desempenho do Estudante na Leitura - DL(5%)



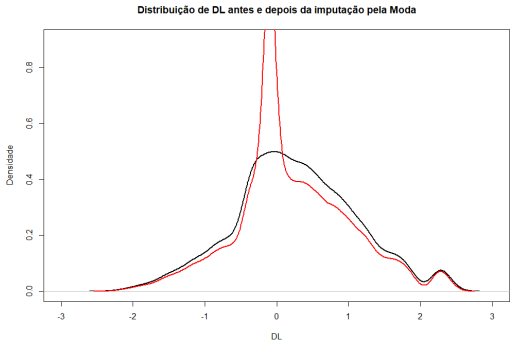
(c) Situação Socioeconómica do Estudante - SSE(10%)



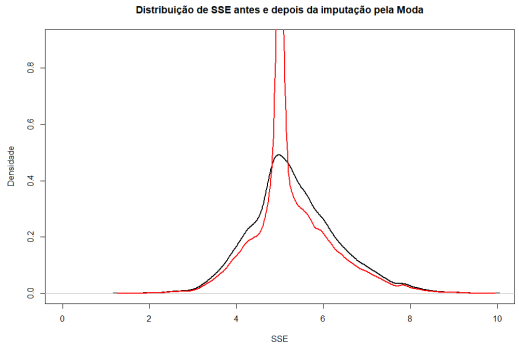
(d) Desempenho do Estudante na Leitura - DL(10%)



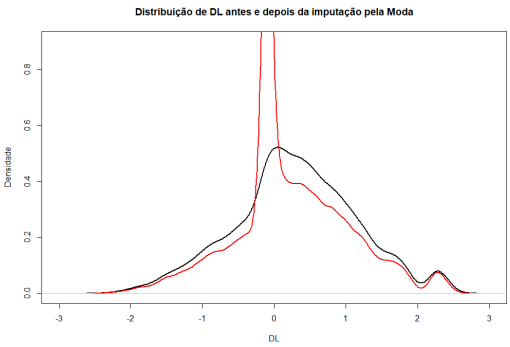
(e) Situação Socioeconómica do Estudante - SSE(15%)



(f) Desempenho do Estudante na Leitura - DL(15%)



(g) Situação Socioeconómica do Estudante - SSE(20%)



(h) Desempenho do Estudante na Leitura - DL(20%)

Figura 5.5: Comportamento da distribuição de SSE e DL antes e depois das imputações pela Moda com 5, 10, 15 e 20% de valores omissos.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Principais Conclusões

Um problema frequente na análise de dados é a existência de valores omissos em conjuntos de dados, pelo que encontrar as melhores maneiras de lidar com estes valores é uma área de estudo importante. Na presente dissertação foram realizadas algumas análises da aplicação de métodos (algoritmos) de imputação de valores omissos no conjunto de dados do estudo de Avaliação Nacional do Rendimento Escolar (Prova Brasil). Foram testados seis algoritmos de imputação de valores omissos sendo três de imputação por valore constantes (Média, Mediana e Moda), dois baseados em aprendizagem automática (kNN e bPCA) e um baseado na imputação múltipla com MICE, cujo objetivo é de comparar e determinar melhores métodos e técnicas de alcançar valores dos resultados mais próximos aos originais.

Os seis algoritmos de imputação foram estudados em quatros subconjuntos de dados dos estudantes com distintas percentagens de valores omissos, tais como 5%, 10%, 15% e 20% de valores omissos gerados aleatoriamente. A seguir cada subconjunto de valores imputados passou por avaliação de resultados com duas métricas de erro RMSE e MAE. Calcularam-se estatísticas descritivas simples para o conjunto de dados original (completo) e para cada conjunto de valores imputados com os diferentes métodos. Perante os resultados obtidos, pode-se concluir o seguinte:

- Em relação aos valores originais, observa-se em todas as percentagens de valores omissos (5%, 10%, 15% e 20%) imputados pela Moda, valores menores na maioria das posições.
- Em 15 e 20% de valores omissos, o método de imputação pela Moda forneceu menores valores das estimativas de erros nas duas variáveis de interesse. Isto é valores mais próximos dos originais.
- Na medida que cresce a percentagem de valores omissos, o valor da média também aumenta e o desvio padrão entre as imputações decresce na maior parte dos casos.
- O método que apresentou um erro mais levado foi na generalidade dos casos a imputação por MICE. No entanto, neste estudo não se avaliou a convergência dos valores imputados, tendo sido realizadas 50 iterações em todos os casos.
- Dos métodos de imputação simples, a imputação pela média foi a que apresentou maiores valores de erro.
- Por se tratar de resultados obtidos em situações particulares, utilizando conjuntos de dados educacionais, não podem ser generalizados.

Muitas pesquisas tem sido publicados sobre diversos métodos de imputação principalmente na área da saúde entre outras, não obstante, na literatura encontramos poucos trabalhos publicados sobre o assunto na área da educação.

6.2 Trabalhos Futuros

Tendo em conta importância do assunto, de certa maneira existem ainda muitos aspectos a serem explorados e barreira por ultrapassar. Como trabalhos futuros sobre esta temática sugerimos:

- Dar sequência e aprofundar o estudo da análise de valores omissos nas diversas áreas de aplicação, testando mais métodos de imputação.
- Realização das análises feitas nesta dissertação utilizando outros conjuntos de dados.
- Definir outras métricas de avaliação do erro, diferentes de RMSE e MAE que foram utilizadas neste trabalho.

Bibliografia

[ALR17] Olanrewaju Akande, Fan Li, and Jerome Reiter. An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *American Statistician*, 71(2):162–170, 2017. 12

[AMAS17] Roslan Armina, Azlan Mohd Zain, Nor Azizah Ali, and Roselina Sallehuddin. A Review on Missing Value Estimation Using Imputation Algorithm. *Journal of Physics: Conference Series*, 892(1), sep 2017. 20

[AS17] Tahani Aljuaid and Sreela Sasi. Proper imputation techniques for missing values in data sets. *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016*, 2017. 31

[ASL07] Melissa J. Azur, Elizabeth A. Stuart, and Constantine Frangakis & Philip J. Leaf. A dimensional approach to developmental psychopathology. *International Journal of Methods in Psychiatric Research*, 16(S1):S16–S23, 2007. Available from: `http://www.scopus.com/inward/record.url?eid=2-s2.0-79951999327{&}partnerID=tZOtx3y1`. 19

[Bar03] Stella Maris Lemos Nunes Baracho. Tratamento de Dados Ausentes e Estudos Longitudinais. 2003. 18

[BO00] S. Van Buuren and C. G. M. Oudshoorn. *Multivariate Imputation by Chained Equations Date. MICE V1.0 User's manual*. TNO Prevention and Health, 2000. Available from: `http://www.multiple-imputation.comhttps://stefvanbuuren.name/publications/MICEV1.0ManualTNO000382000.pdf` [cited 15/06/2019]. 19

[CD14] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014. 10, 11

[DBBA20] Kaouthar Driss, Wadii Boulila, Amreen Batool, and Jawad Ahmad. A Novel approach for classifying diabetes' patients based on imputation and machine learning. *2020 International Conference on UK-China Emerging Technologies, UCET 2020*, pages 14–17, 2020. 2

[DBT18] Tan Duy Le, Razvan Beuran, and Yasuo Tan. Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018*, pages 247–251, 2018. 20

[DC20] Adrian Dragulescu and Cole Arendt. Package ' xlsx '. 2020. 26

[dMAL14] Rildo Gonçalves de Moura, José Antônio Aravéquia, and Alexandre Boleira Lopo. Preenchimento de falhas em dados de correlação de Anomalia da altura geopotencial (500 hPa). *Ciência e Natura*, 36(2):503–509, 2014. 21

[dS12] Maria Joseane Cruz da Silva. Imputação múltipla : comparação e eficiência em experimentos multiambientais Maria Joseane Cruz da Silva Piracicaba. Master's thesis, Universidade de São Paulo, Piracicaba, 2012. 7

[End10] Craig K Enders. *Applied missing data analysis*. The Guilford Press, New York London, 2010. 10, 12, 14

[EPC+01] Fred H. Edwards, Eric D. Peterson, Laura P. Coombs, Elizabeth R. DeLong, W. R.Eric Jamieson, A. Laurie W. Shroyer, and Frederick L. Grover. Prediction of operative mortality after valve replacement surgery. *Journal of the American College of Cardiology*, 37(3):885–892, 2001. Available from: `http://dx.doi.org/10.1016/S0735-1097(00)01202-X`. 14, 22

[FGM18] Paulo Felipe De Oliveira, Saulo Guerra, and Robert Mcdonnell. *Ciência de Dados com R - Introdução*. Editora IBPAD, 2018. 26

[FP19] Maria Eugénia Ferrão and Paula Prata. *Computing Topics on Multiple Imputation in Big Identifiable Data Using R: An Application to Educational Research*, pages 12–24. 06 2019. 6, 19, 28

[FPA20] Maria Eugénia Ferrão, Paula Prata, and Maria Teresa Gonzaga Alves. Multiple imputation in big identifiable data for educational research: An example from the Brazilian Education assessment system. *Ensaio*, 28(108):599–621, 2020. 6, 19

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–53, 1996. 1, 2

[Gan16] Marina Gandolfi. Imputação Múltipla via Algorítmo Mice E Método Imld. Dissertação apresentado ao programa de pós-graduação em bioestatística, Universidade de Maringá, Brasil, 2016. 12, 15, 17, 24

[GLH15] Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*, volume 72. Springer, 2015. 16

[GLSGFV10] Pedro J. García-Laencina, José Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282, 2010. 31

[GOG07] John W. Graham, Allison E. Olchowski, and Tamika D. Gilreath. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213, 2007. 18

[HAB+20] Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *Package 'forecast'*. 2020. 26

[Har19] Frank E . Jr. Harrell. Package "Hmisc": R packlage version version 4.2-0. 2019. Available from: `https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf` [cited 20/05/2019]. 24, 30

[ISS14] Wanapol Insuwan, Ureerat Suksawatchon, and Jakkarin Suksawatchon. Improving missing values imputation in collaborative filtering with user-preference genre and singular value decomposition. *Proceedings of the 2014 6th International Conference on Knowledge and Smart Technology, KST 2014*, pages 87–92, 2014. 2, 21

[JM18] Clarice Garcia Borges Demétrio José Cláudio Faria and Ivan Bezerra Allaman Maintainer. Package 'bpca': Biplot of Multivariate Data Based on Principal Components Analysis. pages 1–45, 2018. Available from: `https://cran.r-project.org/web/packages/bpca/bpca.pdf`. 26

[JMGL+10] José M. Jerez, Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2):105–115, 2010. 15

[Kas17] Alboukadel Kassambara. *Practical Guide to Principal Component Methods in R*. Multivariate Analysis. STHDA, 1 edition, 2017. 16

[KDH+97] Shukri F. Khuri, Jennifer Daley, William Henderson, Kwan Hur, James O. Gibbs, Galen Barbour, John Demakis, George Irvin, John F. Stremple, Frederick Grover, Gerald McDonald, Edward Passaro, Peter J. Fabri, Jeannette Spencer, Karl Hammermeister, and Bradley J. Aust. Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: Results of the National Veterans Affairs surgical risk study. *Journal of the American College of Surgeons*, 185(4):325–338, 1997. 15

[LKGTD19] Wai Yan Lai, Kuok King Kuok, Shirley Gato-Trinidad, and Kuo Xiong Ling Derrick. A study on sequential K-nearest neighbor (SKNN) imputation for treating missing rainfall data. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(3):363–368, 2019. 2

[LR87] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, USA, 1987. 2, 17, 36

[LR02a] Roderick J. A. LITTLE and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, second edition, 2002. 2

[LR02b] Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Hoboken : John Wiley & Sons, New York, NY, second edition edition, 2002. 9, 12, 15, 17

[McN17] Daniel McNeish. Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, 44(1):24–39, 2017. Available from: `https://doi.org/10.1080/02664763.2016.1158246`. 20

[MG19] Steffen Moritz and Sebastian Gatscha. Package ? imputeTS ': Time Series Missing Value Imputation. page 29, 2019. Available from: `https://github.com/SteffenMoritz/imputeTS`. 24

[MMSF07] Patrick E. McKnight, Katherine M. McKnight, Souraya Sidani, and Aurelio José Figueredo. *Missing Data: A Gentle Introduction*, volume 62. The Guilford Press, New York London, 2007. 9, 14

[MV06] Geert Molenberghs and Geert Verbeke. Models for discrete longitudinal data. *http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=epidoc1.in]/?t2000=023788/(100)*, 01 2006. 18

[Nad14] Tamil Nadu. for Software Fault Prediction in Class Level and Package Level Metrics. (978):1–5, 2014. 11

[NKF09] LN Nunes, MM Klück, and JM Fachel. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos Multiple imputations for missing data: a simulation with epidemiological data. *Cadernos de Saúde Pública*, 25(2):268–278, 2009. 21, 22

[NKF10] Luciana Neves Nunes, Mariza Machado Klück, and Jandyra Maria Guimarães Fachel. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. *Revista Brasileira de Epidemiologia*, 13(4):596–606, 2010. 3, 21

[NLB12] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine*, 55(1):37–50, 2012. 17

[NM16] Davis ND and Rahman MM. Missing Value Imputation Using Stratified Supervised Learning for Cardiovascular Data. *Global Journal of Technology and Optimization*, 01(S1):1–11, 2016. 21

[Nun07] Luciana Neves Nunes. *Métodos de Imputação de Dados Aplicados na Área da Saúde*. Tese de doutoramento, Universidade Federal do Rio Grande do Sul, 2007. 22

[Oli18] João Carlos Fidalgo Pinho Oliveira. Imputação em datasets médicos – uma comparação entre três métodos. *Journal of Chemical Information and Modeling*, 53(9), 2018. 16, 20

[OST+03] Shigeyuki Oba, Masa Aki Sato, Ichiro Takemasa, Morito Monden, Ken Ichi Matsubara, and Shin Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003. 3, 16

[PE04] James L. Peugh and Craig K. Enders. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, 74(4):525–556, 12 2004. Available from: `http://journals.sagepub.com/doi/10.3102/00346543074004525`. 12

[Per14] Edna Alessandra Pereira. Algumas propostas para imputação de dados faltantes em teoria de resposta ao item. Master's thesis, Universidade de Brasília, Dissertação (Mestrado em Estatística), Brasília, 2014. 2, 14, 15

[PHW16] Maria Pampaka, Graeme Hutcheson, and Julian Williams. Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research and Method in Education*, 39(1):19–37, 2016. 6

[Ps91] Gregory Piatetsky-shapiro. Knowledge Discovery in Real Databases : A Report on the IJCAI-89 Workshop. *AI Magazine*, 11(5):3, 1991. 1

[QZH+08] Li Qu, Yi Zhang, Jianming Hu, Liyan Jia, and Li Li. A BPCA based missing value imputing method for traffic flow volume data. *IEEE Intelligent Vehicles Symposium, Proceedings*, D(10):985–990, 2008. 3, 17, 22

[Refa] [online]Available from: `https://www.r-project.org/` [cited 29 Jul. 2019]. 23

[Refb] [online]Available from: `http://portal.mec.gov.br/prova-brasil/`. 28

[RF17] Miguel Rocha and Pedro G. Ferreira. *Análise e Exploração de Dados com R*. Lisboa - Portugal, 1 edition, 2017. 16, 66

[Rib15] Elisalvo Alves Ribeiro. Imputação de Dados Faltantes via Algoritmo EM e Rede Neural MLP com o Método de Estimativa de Máxima Verossimilhança para Aumentar a Acurácia das Estimativas. Dissertação de mestrado, Universidade Federal de Sergipe, São Cristóvão, 2015. 8, 17

[RR02] Edna Afonso Reis and Ilka Afonso Reis. Análise Descritiva de Dados. 2002. Available from: `http://www.est.ufmg.br`. 5

[Rub87] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987. 12, 18

[Sch97] J L Schafer. *Analysis of Incomplete Multivariate Data*, volume Analytical. Chapman & Hall/CRC, USA, 1997. 12, 17, 31

[Scr17] Adriana Scrobote. Uma análise da aplicação de algoritmos de imputação de valores faltantes em bases de dados multirrótulo. 2017. 20

[SG02] Josepn L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002. 8, 10

[SKS07] Parvinder Singh Sandhu, Sunil Kumar, and Hardeep Singh. Intelligence System for Software Maintenance Severity Prediction. *Journal of Computer Science*, 3(5):281–288, 2007. 11

[SMG15] Peter Schmitt, Jonas Mandel, and Mickael Guedj. A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*, 06(01):1–6, 2015. 2, 21

[SR15] Wolfram Stacklies and Henning Redestig. Handling of data containing outliers. pages 1–3, 2015. 26

[SRW19] Wolfram Stacklies, Henning Redestig, and Kevin Wright. Package 'pcaMethods' : A collection of PCA methods. 2019. 26

[SSQG06] Fiona M. Shrive, Heather Stuart, Hude Quan, and William A. Ghali. Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology*, 6:1–10, 2006. Available from: `https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-6-57`. 15, 22

[Ste46] S.S. Stevens. On the Theory of Scales of Measurement. *Science*, 103(2684):677–680, 1946. 5

[TKAP19] Matthias Templ, Alexander Kowarik, Andreas Alfons, and Bernd Prantner. VIM: Visualization and Imputation of Missing Values, Version 4.8.0. pages 1–69, 2019. Available from: `https://github.com/statistikat/VIM`. 26

[Tor03] Ana Cristina Pessanha Torres. *Uso de Técnicas de Data Mining para Imputação de Dados Uma Aplicação ao Censo Demográfico de 1991*. Dissertação de mestrado em estudos populacionais e pesquisas sociais, área de concentração em estudos populacionais e demografia., Instituto Brasileiro de Geografia e Estatística, 2003. 22

[vB12] Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/-CRC Press, Boca Raton London New York, 2012. 7, 12, 14, 15

[VB17] José Braga de Vasconcelos and Alexandre Barão. *Ciência dos Dados nas Organizações: Aplicações em Python*. FCA - Editora de Informática, Lda, Lisboa, 1ª edition, 2017. 6

[vBGO11] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. 19, 24, 25

[Ver11] Rosana Veroneze. Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla. Dissertação de mestrado apresentada à faculdade de engenharia elétrica e de computação, Universidade Estadual de Campinas, SP-Brasil, 2011. 8, 9

[Vid16] Vidhya. Tutorial on 5 powerful R Packages used for imputing missing values, 2016. Available from: `https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/`. 24

[VL18] Luis Gustavo do Amaral Vinha and Jacob Arie Laros. Dados ausentes em avaliações educacionais: comparação de métodos de tratamento. *Estudos em Avaliação Educacional*, (x):1, 2018. 9, 12, 13, 14, 20

[WDV10] K. Wauters, P. Desmet, and W. Van Den Noortgate. Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6):549–562, 2010. 27

[WG17] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, volume 48. 2017. 26

[Wic19] Hadley Wickham. Package ‘ tidyverse ’. pages 1–5, 2019. 26

[Wil32] S. S. Wilks. Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples. *The Annals of Mathematical Statistics*, 3(3):163–195, 1932. 14

[WKM20] Kevin Wright, Neil Klepeis, and Paul Murrell. Package ‘ lattice ’. 2020. 25

[WM97] D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6(February):1–34, 1997. 16

[WM05] Cort Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82, 2005. Available from: `www.int-res.com`. 11

[Zha07] Paul Zhang. Multiple Imputation: Theory and Method. *International Statistical Review*, 71(3):581–592, 2007. 2, 7, 8, 17

[Zhu14] Xiaoping Zhu. Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study. *Open Journal of Statistics*, 04(11):933–944, 2014. 8

[ZJM] Nuryazmin Ahmat Zainuri, Abdul Aziz Jemain, and Nora Muda. A Comparison of Various Imputation Methods for Missing Values in Air Quality Data (Perbandingan Pelbagai Kaedah Imputasi bagi Data Lenyap untuk Data Kualiti Udara). *Sains Malaysiana*, (3):449–456. 8, 21

Apêndice A

Anexos

A.1 Scripts utilizados para o estudo da simulação dos modelos em R

A.1.1 Imputação pela Média

```
            ##########################################
            #          IMPUTACAO PELA MEDIA          #
            ##########################################
#Importação das bibliotecas
library(tidyverse) # jácarrega o readr para leitura de dados
library(Hmisc)
library(VIM)
set.seed(2019)
#Leitura do ficheiro
dataMiss <- read.csv(file.choose(), header = T, sep=";")
Y<-matrix(dataMiss, ncol = ncol(dataMiss),nrow = nrow(dataMiss),
   byrow=T)
NDImputados<-dataMiss
Medias_col<- round(colMeans(NDImputados,na.rm=T), digits = 2)
for(i in 1:nrow(NDImputados))
{
  for(j in 1:ncol(NDImputados))
    {
    if(is.na(NDImputados[i,j]))
    {
      NDImputados[i,j]=Medias_col[j]
      cat("o elemento",i,j,"foi substituido por:", Medias_col[j
   ], "\n")
    }
  }
}
#Guardar os resultados num ficheiro csv
write.table(NDImputados, "D:/FICHEIROS_RESULTANTES_R/Dados_compl
   _Imput10.csv", sep=";")
Estat_ImpMedNDImputados <- summary(NDImputados) #Resumo das
   estatisticas
```

Listing A.1: Exemplo do codigo utilizado na Imputacao pela Media.

A.1.2 Imputação pela Mediana

```
1  ##########################################
2  #          IMPUTACAO PELA MEDIANA        #
3  ##########################################
4
5  ## Importação de bibliotecas
6  library(tidyverse)
7  library(xlsx)
8  library(imputeTS)
9  set.seed(2019)
10
11 ## Ler o ficheiro
12 dataMiss <- read.csv(file.choose(), header = T, sep=";")
13 names(dataMiss) = c("TSR", "SSE", "DL")
14
15 ## Imputacao pela Mediana
16 dataImputMediana <- round(na.mean(dataMiss, option = "median"),
       2)
17
18 ## Guardar os resultados num ficheiro
19 write.xlsx(dataImputMediana, "D:/FICHEIROS_RESULTANTES_R/IMPUT_
       MEDIANA/Dados_compl_ImputMediana20.xlsx")
20
21 ## Calculo da media e desvio padrao dos valores imputados
22 media.desvio<- function(x, perdido) {
23 if (is.numeric(x)) {
24   c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm =
       perdido))
25   }
26 }
27 m_sd_ImpMediana<-round(apply(X = dataImputMediana, MARGIN = 2,
       FUN = media.desvio, perdido = T), digits = 3)
28 ## Resumo das estatisticas
29 summary(dataImputMediana)
```

Listing A.2: Exemplo do codigo utilizado na Imputacao pela Mediana.

A.1.3 Imputação pela Moda

```
1  ##########################################
2  #          IMPUTACAO PELA MODA           #
3  ##########################################
4
5  ## Importação de bibliotecas
6  library(tidyverse)
7  library(xlsx)
8  library(imputeTS)
9  set.seed(2019)
10
11 ## Ler o ficheiro
12 dataMiss <- read.csv(file.choose(), header = T, sep=";")
13 names(dataMiss) = c("TSR", "SSE", "DL")
14
15 ## Imputacao pela Moda
16 dataImputModa <- round(na.mean(dataMiss, option = "mode"), 2)
17
18 ## Guardar os resultados num ficheiro
19 write.xlsx(dataImputModa, "D:/FICHEIROS_RESULTANTES_R/IMPUT_MODA
     /Dados_compl_ImputModaX.xlsx")
20
21 ## Calculo da media e desvio padrao dos valores imputados
22 media.desvio<- function(x, perdido) {
23 if (is.numeric(x)) {
24   c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm =
      perdido))
25   }
26 }
27 m_sd_ImpModa<-round(apply(X = dataImputModa, MARGIN = 2, FUN =
     media.desvio, perdido = T), digits = 3)
28
29 ## Resumo das estatisticas
30 summary(dataImputModa)
```

Listing A.3: Exemplo do codigo utilizado na Imputacao pela Moda.

A.1.4 Imputação com KNN

Algorithm 1 Pseudocódigo para o KNNImpute

```
% M - Matriz com valores omissos
% k - Quantidade desejada de vizinhos
% Mimp - Matriz com os valores imputados
Início
Leitura do conjunto de dados csv:
Mimp ⟵ M
vs ⟵ calculo_da_similaridade (M)
for cada valor omisso (i, j) de Mimp do
    v ⟵ knn(M, vs, (i, j), k)
    (i, j) ⟵ p(v)
until Imputados com sucesso
```

Em que:

* calculo_da_similaridade é a função que permite calcular a similaridade entre todas as linhas da matriz M.
* knn é a função que permite achar os k vizinhos mais próximos para a imputação dos valores omissos (i, j).
* p é uma função que calcula o valor a ser imputado.

```
            ##########################################
            #            IMPUTACAO COM KNN            #
            ##########################################

#Importação de bibliotecas
library(tidyverse)
# library(DMwR)
library(naniar)
library(VIM) #KNN imputation
set.seed(2019)

#Leitura do ficheiro
dataMiss <- read.csv(file.choose(), header = T, sep=";")

#Imputação com a função kNN() do VIM
data_imput_knn <- kNN(dataMiss, variable = c("SSE", "DL"), k=6)

#Imputação com a função knnImputation() do DMwR
#KNNcompleto <- knnImputation(dataESS) #Utilizando library(DMwR)

write.xlsx(data_imput_knn, "D:/FICHEIROS_RESULTANTES_R/Dados_
   ImputKNN_M5.xlsx")
Estat_ImputKNN <- summary(data_imput_knn)
#write.xlsx(KNNcompleto, "D:/FICHEIROS_RESULTANTES_R/Dados_
   ImputKNN_M5.xlsx")
#Estat_ImputKNN <- summary(KNNcompleto)

## Calculo da media e desvio padrao dos valores imputados
media.desvio<- function(x, perdido) {
  if (is.numeric(x)) {
    c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm =
   perdido))
  }
}
m_sd_ImpKNN<-round(apply(X = data_imput_knn, MARGIN = 2, FUN =
   media.desvio, perdido = T), digits = 3)
```

Listing A.4: Exemplo do codigo utilizado na Imputacao pela metodo KNN.

A.1.5 Imputação de Valores Omissos com bPCA

```
1  ##########################################
2  #          IMPUTACAO COM bPCA            #
3  ##########################################
4
5  #Importação de bibliotecas
6  library(bpca)
7  library(pcaMethods)
8  #Leitura do ficheiro
9  dataMiss <- read.csv(file.choose(), header = T, sep=";")
10
11 set.seed(2019)
12 databPCA <- pca((dataMiss), method="bpca", nPcs=2)
13
14 ## Estimativas de eixo principal (loadings)
15 loadings <- loadings(databPCA)
16
17 ## Estimativas de resultados (scores)
18 Resultados <- scores(databPCA)
19
20 ## Estimativas de observacoes completas
21 ObservCompletos <- completeObs(databPCA)
22
23 ## Calculo da media e desvio padrao dos valores imputados
24 media.desvio<- function(x, perdido) {
25 if (is.numeric(x)) {
26   c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm =
      perdido))
27   }
28 }
29 m_sd_ImpbPCA<-round(apply(X = ObservCompletos, MARGIN = 2, FUN =
       media.desvio, perdido = T), digits = 3)
30
31 ## Visualizacao de resultados
32 slplot(databPCA)
```

Listing A.5: Exemplo do codigo utilizado na imputacao de valores omissos com bPCA.

A.1.6 Imputação com mice

```
            ##########################################
            #          IMPUTACAO COM MICE            #
            ##########################################

#Importação de bibliotecas
library(VIM)
library(mice)
set.seed(2019)

#Leitura do ficheiro
dataMiss <- read.csv(file.choose(), header = T, sep=";")
imput_Dados_mice <- mice(dataMiss, m=5, maxit = 50, method = '
   pmm', seed = 500)
summary(imput_Dados_mice)
md.pattern(imput_Dados_mice) #Resumo de padrões dos valores
   omissos

## Obter dados completos
Dadoscompletos <- mice::complete(imput_Dados_mice)
print(Dadoscompletos)

write.xlsx(Dadoscompletos, "D:/FICHEIROS_RESULTANTES_R/Dados_
   ImputMICE2_M5.xlsx")
Estat_ImputMICE <- summary(Dadoscompletos)
## Calculo da media e desvio padrao dos valores imputados
media.desvio<- function(x, perdido) {
if (is.numeric(x)) {
  c(media = mean(x, na.rm = perdido), desvio = sd(x, na.rm =
   perdido))
  }
}
m_sd_ImpMICE<-round(apply(X = Dadoscompletos, MARGIN = 2, FUN =
   media.desvio, perdido = T), digits = 3)
```

Listing A.6: Exemplo do codigo utilizado na imputacao de dados com mice.

A.1.7 Cálculos de Erros

```
            ##########################################
            #           METRICAS DE ERROS            #
            ##########################################
## Leitura de dados
dataOriginal <- read.csv(file.choose(), header = T, sep=";")
dataImputado <- read.csv(file.choose(), header = T, sep=";")

library(forecast) # Para as métricas

## Funcao que rotorna Root Mean Squared Error
rmse <- function(error)
{
  sqrt(mean((error)^2))
}
## Funcao que retorna Mean Absolute Error
mae <- function(error)
{
  mean(abs(error))
}
## Selecao das variaveis
atual_SSE <- c(dataOriginal[,2])
predicted_SSE <- c(dataImputado[,2])

atual_DL <- c(dataOriginal[,3])
preditado_DL <- c(dataImputado[,3])

error <- preditado_SSE - atual_SSE
error <- preditado_DL - atual_DL

## Chamada das funcoes
rmse_SSE <- round(rmse(error),3)
mae_SSE <- round(mae(error), 3)
rmse_SSE
mae_SSE

rmse_DL <- round(rmse(error),3)
mae_DL <- round(mae(error), 3)
rmse_DL
mae_DL
```

Listing A.7: Codigo para calculos de erros (RMSE e MAE).

Glossário

Tabela A.1: Alguns Termos Técnicos

PORTUGUÊS EUROPEU	PORTUGÊS DO BRASIL	INGLÊS
âmbito (do projeto)	escopo (do projeto)	*scope*
base de dados	banco de dados	*database*
equipa	time/equipe	*team*
ficheiros	arquivos	*files*
omissos	ausentes/faltantes	*missing*
software livre	programas livres	*open source*
package	pacote	*package*
planeamento	planejamento	*planning*
sistema operativo	sistema operacional	*operating system*
partes interessadas	*stakeholders*	*stakeholders*
treino	treinamento	*training*
utilizador	usuário	*user*
acurácia	viés	*accuracy*

Tabela A.2: Algumas funções Estatísticas e Matemáticas mais utilizadas

Função	Descrição
abs()	Devolve o valor absoluto
IQR()	Dá o chamado intervalo interquartil (a diferênça entre os valores do quartil 75% e os do quartil 25%)
log()	Calcula o logaritmo natural
mad()	Calcula o desvio absoluto médio
mean()	Calcula a média aritmética de um conjunto de valores
median()	Calcula a mediana de conjunto de valores
sum()	Calcula o valor da soma de um conjunto de valores
max()	Determina o valor máximo de um conjunto de valores
min()	Determina o valor mínimo de um conjunto de valores
mode()	Valor modal (moda) de dados discretos de um conjunto de valores
sd()	Calcula o desvio-padrão de um conjunto de valores
summary()	Resumo Estatístico de um conjunto de valores
sqrt()	Calcula a raiz quadrada
quantile()	Calcula os quartis do conjunto de valores
var()	Calcula a variância de um conjunto de valores

Fonte: Adaptada de Rocha e Ferreira [RF17, pp. 88-89]