UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

FCA
FACULDADE DE CIÊNCIAS AGRONÔMICAS
UNESP | BOTUCATU

VALTER CESAR DE SOUZA

APLICAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS NA IMPUTAÇÃO DE VALORES AUSENTES EM DADOS AGROMETEOROLÓGICOS EM ALTA DIMENSÃO

Botucatu

2023

VALTER CESAR DE SOUZA

APLICAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS NA IMPUTAÇÃO DE VALORES AUSENTES EM DADOS AGROMETEOROLÓGICOS EM ALTA DIMENSÃO

Tese apresentada à Faculdade de Ciências Agronômicas da Unesp Campus de Botucatu, para obtenção do título de Doutor em Agronomia (Energia na Agricultura).

Orientador: Sergio Augusto Rodrigues

Botucatu

2023

S729a

Souza, Valter Cesar de

Aplicação de análise de componentes principais na imputação de valores ausentes em dados agrometeorológicos em alta dimensão / Valter Cesar de Souza. -- Botucatu, 2023

133 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agronômicas, Botucatu

Orientadora: Sergio Augusto Rodrigues

1. Agrometeorologia. 2. Dados Meteorológicos. 3. Evapotranspiração. 4. Imputação. 5. Simulação. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agronômicas, Botucatu. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: APLICAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS NA IMPUTAÇÃO DE VALORES AUSENTES EM DADOS AGROMETEOROLÓGICOS EM ALTA DIMENSÃO

AUTOR: VALTER CESAR DE SOUZA
ORIENTADOR: SERGIO AUGUSTO RODRIGUES

Aprovado como parte das exigências para obtenção do Título de Doutor em AGRONOMIA (ENERGIA NA AGRICULTURA), pela Comissão Examinadora:

Prof. Dr. SERGIO AUGUSTO RODRIGUES (Participaçao Presencial)
Bioprocessos e Biotecnologia / Faculdade de Ciencias Agronomicas de Botucatu UNESP

Prof.ª Dr.ª SANDRA CRISTINA DE OLIVEIRA (Participaçao Presencial)
Gestao, Desenvolvimento e Tecnologia / Faculdade de Ciências e Engenharia - FCE - UNESP - Tupã/SP

Prof. Dr. PAULO ANDRÉ DE OLIVEIRA (Participaçao Presencial)
Agronegócio / Faculdade de Tecnologia de Botucatu

Prof. Dr. ALEXANDRE DAL PAI (Participaçao Presencial)
Bioprocessos e Biotecnologia / Faculdade de Ciencias Agronomicas de Botucatu

Prof.ª Dr.ª VALERIA CRISTINA RODRIGUES SARNIGHAUSEN (Participaçao Presencial)
Bioprocessos e Biotecnologia / Faculdade de Ciências Agronômicas de Botucatu - UNESP

Botucatu, 24 de fevereiro de 2023

Para

Minha esposa, Sílvia

Meus filhos, Eduarda e Caio

Minha mãe, Conceição

Em memória,

Meu pai, Antonio

Dedico.

AGRADECIMENTOS

Após uma reflexão sobre os últimos quatro anos, sinto-me extremamente grato por ter contado com a orientação e o apoio de profissionais talentosos, cujo conhecimento e habilidades enriqueceram significativamente esta pesquisa em diversas formas. Gostaria de expressar meu sincero agradecimento às seguintes pessoas, que desempenharam papéis fundamentais em meu caminho até a conclusão desta tese de doutorado.

Ao Prof. Dr. Sérgio Augusto Rodrigues, pela orientação, ensinamentos, paciência e exemplo de professor.

Ao Prof. Dr. Paulo Arbex, pela oportunidade e apoio, sem o qual não seria possível desenvolver esta pesquisa.

Aos membros da banca examinadora, Prof. Dr. Alexandre Dal Pai, Prof. Dr. Paulo André de Oliveira, Profa. Dra. Sandra Cristina de Oliveira, Prof. Dr. Sergio Augusto Rodrigues e Profa. Dra. Valeria Cristina Rodrigues Sarnighausen, por dedicarem seu tempo e expertise para avaliar este trabalho e fornecerem sugestões construtivas que aprimoraram expressivamente a qualidade desta pesquisa.

À minha família pelo amor, encorajamento e apoio incondicionais ao longo de toda a minha jornada acadêmica. Suas palavras de estímulo e confiança foram essenciais para superar desafios e perseverar neste percurso desafiador.

A todos os professores e funcionários da Faculdade de Ciências Agronômicas (FCA).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Uma longa caminhada avança com o segundo passo.

Valter De Souza.

RESUMO

Esta pesquisa aborda o tema *Missing Value Imputation* (MVI) com foco em agrometeorologia, na estimativa e imputação de dados de evapotranspiração de referência. O trabalho foi dividido em três capítulos, entendidos como revisão, metodologia e aplicação. O Capítulo 1 apresenta uma revisão bibliométrica abrangendo cerca de 19.745 publicações entre 1940 e 2022, destacando os principais pesquisadores, artigos, periódicos, instituições e países no assunto MVI. O Capítulo 2 apresenta uma proposta detalhada do planejamento de uma simulação em MVI, abordando o banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho. Essas informações são preparatórias para a aplicação. O Capítulo 3 avalia o desempenho de procedimentos multivariados alternativos de análise de componentes principais na imputação de dados ausentes em base de dados de evapotranspiração. Foram utilizados dados do período de 2012 a 2021 de quarenta e cinco estações meteorológicas automáticas da região de São Paulo, Brasil, simulando diferentes cenários de dados faltantes. Esta pesquisa ressalta a importância da agrometeorologia, a necessidade de tratar dados ausentes por meio de técnicas de MVI e a eficácia de procedimentos alternativos de análise de componentes principais na imputação de valores faltantes.

Palavras-chave: evapotranspiração; planejamento de simulação; bibliometria; conjunto de dados incompleto; qualidade de dados; EM-PCA; NIPALS-PCA.

ABSTRACT

This research addresses the theme of Missing Value Imputation (MVI) with a focus on agrometeorology, in the estimation and imputation of reference evapotranspiration data. The work was divided into three chapters, understood as review, methodology, and application. Chapter 1 presents a bibliometric review covering approximately 19,745 publications between 1940 and 2022, highlighting the main researchers, articles, journals, institutions, and countries in the MVI subject. Chapter 2 presents a detailed proposal for planning a simulation in MVI, addressing the database and type of data, mechanism and fault rate, imputation technique, and performance evaluation method. This information is preparatory for the application. Chapter 3 evaluates the performance of alternative multivariate principal component analysis procedures in imputing missing data into evapotranspiration databases. Data from the period 2012 to 2021 from forty-five automatic meteorological stations in the region of São Paulo, Brazil, were used, simulating different scenarios of missing data. This research underscores the importance of agrometeorology, the need to address missing data through MVI techniques, and the effectiveness of alternative principal component analysis procedures in imputing missing values.

Keywords: evapotranspiration; simulation planning; bibliometrics; incomplete data set; data quality; EM-PCA; NIPALS-PCA.

LISTA DE FIGURAS

Capítulo 1 – Imputação de Valores Ausentes: uma Análise Bibliométrica da Literatura

Capítulo 2 – Planejamento de Simulação para Imputação de Valor Ausente

Capítulo 3 – Comparação de Algoritmos de Análise de Componentes Principais para Imputação em Dados Agrometeorológicos em Alta Dimensão e Tamanho Amostral Reduzido

LISTA DE QUADROS

INTRODUÇÃO GERAL

Capítulo 2 – Planejamento de Simulação para Imputação de Valor Ausente

Capítulo 3 – Comparação de Algoritmos de Análise de Componentes Principais para Imputação em Dados Agrometeorológicos em Alta Dimensão e Tamanho Amostral Reduzido

LISTA DE TABELAS

Capítulo 1 – Imputação de Valores Ausentes: uma Análise Bibliométrica da Literatura

Capítulo 3 – Comparação de Algoritmos de Análise de Componentes Principais para Imputação em Dados Agrometeorológicos em Alta Dimensão e Tamanho Amostral Reduzido

LISTA DE ABREVIATURAS E SIGLAS

BDMEP	Banco de Dados Meteorológicos para Ensino e Pesquisa
BD	Banco de Dados
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CMStatistics	Computational and Methodological Statistics
CPTEC	Centro de Previsão de Tempo e Estudos Climáticos
CSDA	Computational Statistics and Data Analysis
CSV	Comma Separated Values (Valores Separados por Vírgulas)
DNA	DeoxyriboNucleic Acid
DT	Decision Tree
EM	Expectation Maximization
FAO	Food and Agriculture Organization
FCA	Faculdade de Ciências Agronômicas
HDD	High-Dimension Data
HDLSS	High-Dimension and Low-Sample-Size
IASC	International Association for Statistical Computing
IEEE	Institute of Electrical and Electronic Engineers
IM	Imputação pela Média das colunas
INMET	Instituto Nacional de Meteorologia
JASA	Journal of the American Statistical Association
KNN	K-Nearest Neighbor
LR	Linear Regression
LS	Least Squares
LSHTM	London School of Hygiene & Tropical Medicine
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multiple Imputation by Chained Equations
missMDA	Missing Multivariate Data Analysis
MNAR	Missing Not at Random
MVI	Missing Value Imputation
NIPALS	Nonlinear Iterative Partial Least Squares
PCA	Principal Component Analysis (Análise Componentes Principais)
PLOS	Public Library of Science

RF	Random Forest
RHN	Rede Hidrometeorológica Nacional
S	South
SNIRH	Sistema Nacional de Informações sobre Recursos Hídricos
SP	São Paulo
UCL	University College London
UCLA	University of California Los Angeles
UNESP	Universidade Estadual Paulista
UNICAMP	Universidade Estadual de Campinas
USA	United States of America
USP	Universidade de São Paulo
W	West
WOS	Web Of Science

LISTA DE SÍMBOLOS

p	número de variáveis
n	número de amostras
®	marca registrada
ET_o	evapotranspiração de referência
ET_c	evapotranspiração de cultura
R_n	radiação líquida
G	fluxo de calor do solo (considerado zero)
T_{med}	temperatura média
u_2	velocidade do vento (calculada ou medida a 2 m do solo)
e_s	pressão de saturação de vapor d'água do ar
e_a	pressão de vapor d'água do ar
Δ	inclinação da curva de pressão de vapor saturado
γ	coeficiente psicrométrico
T_{max}	temperatura máxima
T_{min}	temperatura mínima
u_h	velocidade do vento (medida a *h* m do solo)
h	altura da medida da velocidade vento em relação ao solo
$e_{s\ max}$	pressão de vapor saturado máxima
$e_{s\ min}$	pressão de vapor saturado mínima
UR_{max}	umidade relativa máxima
UR_{min}	umidade relativa mínima
P	pressão atmosférica (calculada ou medida)
z	altitude
R_{ns}	radiação líquida ondas curtas
R_{nl}	radiação líquida ondas longas
R_s	radiação solar (medida)
α	α = 0,23 (alberdo), cultura de referência de grama verde
σ	constante de Stefan-Boltzmann ($4{,}903*10^{-9}$ MJ K^{-4} m^{-2} dia^{-1})
$T_{max,K}$	temperatura máxima em Kelvin
$T_{min,K}$	temperatura mínima em Kelvin
R_{so}	radiação solar céu-claro (sem nuvens)

R_e	radiação extraterrestre
G_{sc}	constante solar (0,0820 MJ m^{-2} min^{-1})
d_r	inverso da distância relativa terra-sol
φ	latitude
δ	declinação solar
ω_s	ângulo horário do pôr do sol
J	número do dia ano entre 1 (01JAN) e 365 ou 366 (31DEZ)
r	coeficiente de correlação de *Pearson*
x_i	i-ésimo valor observado (i=1, ..., m)
$\bar{x}$	média dos valores observados
y_i	i-ésimo valor imputado (i=1, ..., m)
$\bar{y}$	média dos valores imputados
m	número de missings
ME	mean error
MAE	mean absolute error
pMAE	mean absolute percentual error
MSE	mean squared error
RMSE	root mean squared error
pRMSE	root mean squared percentual error
d	índice de concordância de Willmott
c	coeficiente de confiança

SUMÁRIO

INTRODUÇÃO GERAL

Os fundamentos da agrometeorologia estão enraizados nas ciências físicas e biológicas, essenciais para a definição, aplicação e disponibilização do conhecimento sobre o tempo e o clima (TAKLE, 2015) aos agricultores e tomadores de decisão na produção agrícola (CALANCA, 2014; HOOGENBOOM, 2000; STIGTER, 2007). Nas pesquisas em agrometeorologia, principalmente nos estudos de evapotranspiração, que desempenha um papel fundamental no ciclo hidrológico (KISI et al., 2021), no planejamento (GONSAGA DE CARVALHO et al., 2011; WANG et al., 2022) e na gestão de sistemas de irrigação (ESTÉVEZ; GAVILÁN; GIRÁLDEZ, 2011; MARIN et al., 2019; MARTÍ; ZARZO, 2012), na modelagem da demanda de água (TERINK; IMMERZEEL; DROOGERS, 2013), no monitoramento do estresse hídrico (HART et al., 2009), na estimativa do balanço hídrico (CAI et al., 2009) e em trabalhos hidrológicos e ambientais (MARTÍ; GASQUE, 2010) é essencial contar com fontes de dados meteorológicos de alta qualidade, com acurácia e confiabilidade.

A determinação da evapotranspiração é uma tarefa complexa devido aos custos elevados relacionados às técnicas diretas para instalação, operação e manutenção dos equipamentos de medição (ALLEN et al., 2011; RANA; KATERJI, 2000). Uma abordagem alternativa é o uso de métodos indiretos (ONNABI MILANI et al., 2007) que se baseiam em equações matemáticas adaptáveis às condições climáticas locais, dispensando a necessidade de medições diretas. No entanto, é fundamental dispor de séries históricas de dados meteorológicos.

É comum em séries temporais de dados meteorológicos apresentarem desvios ou falhas (HASAN et al., 2021; JUNNINEN et al., 2004; YOZGATLIGIL et al., 2013), incluindo valores ausentes, conhecidas como *missings* (WHITE; ROYSTON; WOOD, 2011). Diversos fatores contribuem para a ocorrência de *missings*, por exemplo, falha de equipamentos, problemas de coleta de dados, informações incompletas (HASAN et al., 2021). A existência de valores ausentes em um banco de dados resulta em perda de eficiência, introduzindo viés devido às discrepâncias entre os dados faltantes e completos, complicando a análise dos dados (FARHANGFAR; KURGAN; PEDRYCZ, 2007). Eliminar os valores faltantes traz a

desvantagem de reduzir o tamanho amostral, o que resulta em diminuir a precisão da análise estatística. (ROTH, 1994; STRIKE; EMAM; MADHAVJI, 2001).

Uma maneira comum de lidar com conjuntos de dados incompletos é por meio das técnicas de imputação de valor ausente, *missing value imputation* (MVI), que é amplamente utilizada como um método de resolução. Os métodos de MVI podem ser agrupados em técnicas estatísticas ou abordagens baseadas em *machine learning.* Um método de MVI pode ser avaliada por meio de simulações que levam em consideração diferentes aspectos, tais como o escopo e o tipo de dados do banco, o mecanismo de falha, a taxa de dados ausentes, a técnica de imputação empregada e o método utilizado para aferir o desempenho.

Experimentos agrícolas podem apresentar um número relativamente maior de variáveis em relação ao tamanho amostral, cenários de alta dimensão e amostra reduzida, *high-dimension and low-sample-size (HDLSS)*. Na literatura o problema de alta dimensão e tamanho amostral reduzido é abordado por diversos pesquisadores (AHN et al., 2007; CHEN; WIESEL; HERO, 2011; HOYLE, 2008; JUNG; SEN; MARRON, 2012; MULLER et al., 2008; SHEN et al., 2016; SHEN; SHEN; MARRON, 2013; YATA; AOSHIMA, 2010, 2012). As dimensões dos dados indicam o número de características que foram medidas para cada observação, tornando-se uma tarefa desafiadora analisar dados de alta dimensão, normalmente designada pela sigla *HDD (high-dimension data*). Os pesquisadores Johnstone e Paul (2018) discutem e orientam para o caso do número de variáveis ou recursos coletados p superar o número de casos (ou tamanho amostral) n, ou seja, para $p>>n$. Os autores procuram dar orientações para alguns desses fenômenos de alta dimensão, tais como: propagação e viés de autovalor, inconsistência de autovetores. O trabalho de Ayesha, Hanif e Talib (2020) apresenta o estado da arte em técnicas de redução de dimensionalidade e sua adequação para diferentes tipos de dados e áreas de aplicação. Além disso, foram destacadas as questões das técnicas de redução de dimensionalidade que podem afetar a precisão e relevância dos resultados. A maioria das técnicas estatísticas multivariadas são sensíveis ao tamanho amostral (SIDDIQUI, 2013). A literatura apresenta recomendações na busca por estimativas mais precisas das características populacionais (HONG et al., 1999) e influência do tamanho amostral (SHAUKAT; RAO; KHAN, 2016).

Das técnicas estatísticas, para imputação de *missings*, devido a sua capacidade e flexibilidade de aplicação em fenômenos de alta dimensão e amostra reduzida, na presença de *missings*, destaque para o procedimento multivariado *Análise de Componentes Principais* (GARCÍA-DIEGO; ZARZO, 2010; JOSSE; HUSSON, 2012; MARTÍ; ZARZO, 2012), PCA. Introduzida por *Karl Pearson* (1901) e fundamentada por *Hotelling* (1933), é utilizada para análise exploratória e redução da dimensionalidade (MINGOTI, 2005), podendo também ser utilizada como procedimento de imputação de dados ausentes (DE KETELAERE; HUBERT; SCHMITT, 2015). Destaque para a utilização conjunta com métodos iterativos como o algoritmo NIPALS (DE LA FUENTE; GARCÍA-MUÑOZ; BIEGLER, 2010; ESHGHI, 2014; HOWLEY et al., 2006; PATEL; SIVANATHAN; MHASKAR, 2021; VYAS et al., 2021; YANG et al., 2012), *Nonlinear Iterative Partial Least Squares* (WRIGHT, 2017) e o algoritmo EM (BUCIOR-KWACZYŃSKA, 2018; MALAN et al., 2020; NILASHI et al., 2022; XIE et al., 2019), *Expectation-Maximization* (DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977).

Aplicações da análise multivariada no campo da ciências agrárias tiveram um salto com o avanço da tecnologia da informação. Entender as limitações das técnicas multivariadas frente as situações de alta dimensão, amostra reduzida, ocorrência de valores ausentes (*missings)* e aplicar procedimentos de análises consistentes para cada problema agrícola, coloca-se como um desafio na área de análise de dados agronômicos.

Neste contexto, esta pesquisa tem por objetivo apresentar um panorama da literatura científica no tema *Missing Value Imputation*, disponibilizar um roteiro para simulação em MVI e avaliar a performance de métodos multivariados alternativos de análise de componentes principais na imputação de dados ausentes.

ESTRUTURA DA TESE

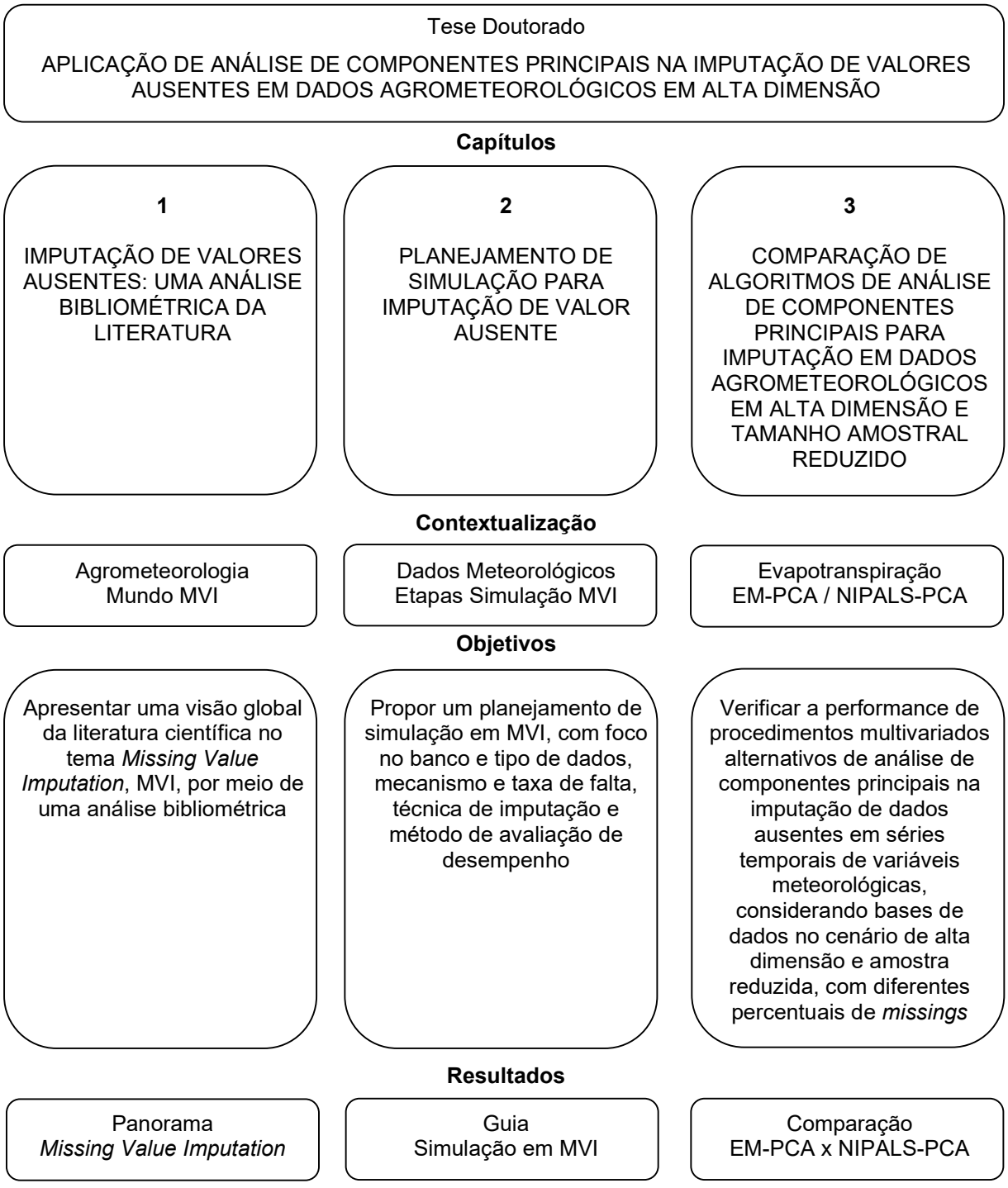
A tese está estruturada em três capítulos, nos quais buscou-se realizar uma análise bibliométrica da temática *Missing Value Imputation*, apresentar as etapas de uma simulação em MVI e comparar procedimentos alternativos para aplicação de técnicas de análise de componentes principais em situações de alta dimensão, amostra reduzida e ausência de dados, por meio de uma avaliação do desempenho frente a dados observados de evapotranspiração de referência considerando diferentes contextos.

O resultado esperado de cada capítulo foi a publicação de um artigo em revistas científicas de impacto na área de ciências agrárias.

Desta maneira, esta tese foi estruturada da seguinte forma: Capítulo 1: Apresenta um panorama das pesquisas mundiais no tema *Missing Value Imputation* por meio de uma análise bibliométrica. Capítulo 2: Propõe um planejamento para estudos de simulação em MVI, com foco no banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho, preparando conceitos para a aplicação, apresentada no capítulo três. Capítulo 3: Compara o desempenho de procedimentos multivariados alternativos de análise de componentes principais na imputação de dados ausentes em séries temporais de variáveis meteorológicas, considerando bases de dados no cenário de alta dimensão e amostra reduzida, com diferentes taxas de dados faltantes.

O Quadro 1 ilustra o desenvolvimento metodológico desta pesquisa por meio de um framework estruturado. Proporciona uma visão panorâmica da organização dos capítulos, enfatizando os resultados obtidos.

Quadro 1 – Framework Metodológico da Pesquisa



Fonte: Autoria Própria.

CAPÍTULO 1

IMPUTAÇÃO DE VALORES AUSENTES: UMA ANÁLISE BIBLIOMÉTRICA DA LITERATURA[1]

MISSING VALUES IMPUTATION: A BIBLIOMETRIC ANALYSIS OF THE LITERATURE

Valter Cesar de Souza[$&*], **Sérgio Augusto Rodrigues**[$&]

[$]São Paulo State University (Unesp), School of Agriculture, Botucatu, São Paulo, Brasil

* Corresponding author
E-mail: valter.souza@unesp.br

[&]These authors contributed equally to this work.

[1] Capítulo redigido de acordo com as normas do periódico PLOS ONE.

RESUMO

A agrometeorologia desempenha um papel crucial ao fornecer informações meteorológicas essenciais para agricultores e tomadores de decisão envolvidos na produção agrícola. Uma área de pesquisa importante nesse campo é a evapotranspiração, que requer o uso de fontes de dados meteorológicos de alta qualidade para estimar o uso de água pelas culturas, especialmente em sistemas de irrigação. No entanto, devido às condições nas quais os dados meteorológicos são coletados, é comum encontrar uma proporção de dados ausentes, conhecidos como *missings*. Independentemente das razões para essas ausências, é fundamental avaliar e tratar adequadamente esses dados faltantes. Neste contexto, o objetivo deste capítulo foi apresentar um panorama das pesquisas mundiais no tema *Missing Value Imputation*, MVI, por meio de uma análise bibliométrica. Para desenvolver a análise bibliométrica utilizou a base de dados *Web Of Science*, com levantamento bibliográfico executado em 2 de janeiro de 2023. Nesta análise bibliométrica o foco foi publicações científicas, no período de 1940 a 2022, relacionados ao assunto MVI. A consulta foi realizada com a *string* de busca: *(missing) AND (value* OR data) AND (imputation OR estimation)*. As métricas bibliométricas utilizadas foram medidas de produtividade e impacto. A construção e visualização de redes bibliométricas foram desenvolvidas pelo software VOSviewer e o pacote Bibliometrix. O levantamento gerou uma amostra com 19.745 publicações, divulgadas por 6.225 fontes científicas, produzidos por cerca de 62.193 autores, gerando 450.673 referências citadas, envolvendo 12.925 instituições, abrangendo cerca de 158 países. Do conjunto das publicações identificadas 70% foram produzidos na última década. Dos autores com maior número de trabalhos, ocorreu uma predominância de pesquisadores Chineses, para as referências identificadas. Os países protagonistas foram os Estados Unidos, seguidos pela China e Inglaterra. O Brasil vem na décima nona posição. As instituições Norte Americanas liberam o ranking das instituições com maior número de publicações em MVI, destaque para Harvard University, University of Michigan, University of Washington, University of North Carolina e University of California. Seguida das universidades do Reino Unido: University College London, University of Oxford, London School of Hygiene & Tropical Medicine. E a mais jovem, a proeminente Chinese Academy of Sciences. Destaque no Brasil para USP, UNICAMP e UNESP. O periódico *Statistics in Medicine*, foi o de maior impacto, seguido pelo *PLOS ONE, Biometrics, Journal of the American Statistical Association, Computational Statistics & Data Analysis, Communications in Statistics-Theory and Methods, Statistical Methods in Medical Research, IEEE Access, BMC Medical Research Methodology e Statistica Sinica.* Dos pesquisadores fundamentais destaque para *Donald B. Rubin, Roderick Joseph Alexander Little, Arthur P. Dempster, Joseph L. Schafer, John W. Graham, Stef van Buuren, Ian R. White, Patrick Royston, James M. Robins e Craig K. Enders.* A distribuição temporal da ocorrência das palavras-chaves, indica fases distintas de pesquisa em MVI, caracterizada pelo início com o *Algoritmo EM*, passando por *Imputação Múltipla* e, na atualidade concentração nas técnicas de *Aprendizado de Máquina*.

Palavras-chave: bibliometria; dados ausentes; imputação; conjunto de dados incompleto; qualidade dos dados.

ABSTRACT

Agrometeorology plays a crucial role in providing essential weather information to farmers and decision-makers involved in agricultural production. An important area of research in this field is evapotranspiration, which requires the use of high-quality meteorological data sources to estimate water use by crops, especially in irrigation systems. However, due to the conditions under which meteorological data is collected, it is common to encounter a proportion of missing data, known as missings. Regardless of the reasons for these absences, it is essential to evaluate and treat these missing data. In this context, the objective of this chapter was to present an overview of global research on Missing Value Imputation (MVI) through bibliometric analysis. The bibliometric analysis was conducted using the Web Of Science database, with a bibliographic survey conducted on January 2, 2023. The focus of this bibliometric analysis was scientific publications related to MVI from the period 1940 to 2022. The search query used was: (missing) AND (value* OR data) AND (imputation OR estimation). The bibliometric metrics used included measures of productivity and impact. The construction and visualization of bibliometric networks were developed using the software VOSviewer and the Bibliometrix package. The survey generated a sample of 19,745 publications, published by 6,225 scientific sources, authored by approximately 62,193 authors, resulting in 450,673 cited references and involving 12,925 institutions across approximately 158 countries. Of the identified publications, 70% were produced in the last decade. Among the authors with the highest number of works, Chinese researchers were predominant in the identified references. The leading countries were the United States, followed by China and England, with Brazil ranking nineteenth. In terms of institutions with the highest number of publications in MVI, North American institutions topped the list, with Harvard University, University of Michigan, University of Washington, University of North Carolina, and University of California taking the lead. The United Kingdom's institutions followed, with University College London, University of Oxford, and London School of Hygiene & Tropical Medicine. The Chinese Academy of Sciences, the youngest institution, was also prominent. Notable Brazilian institutions included USP, UNICAMP, and UNESP. Regarding impact, the journal Statistics in Medicine had the highest impact, followed by PLOS ONE, Biometrics, Journal of the American Statistical Association, Computational Statistics & Data Analysis, Communications in Statistics-Theory and Methods, Statistical Methods in Medical Research, IEEE Access, BMC Medical Research Methodology, and Statistica Sinica. Regarding fundamental researchers in the field, Donald B. Rubin, Roderick Joseph Alexander Little, Arthur P. Dempster, Joseph L. Schafer, John W. Graham, Stef van Buuren, Ian R. White, Patrick Royston, James M. Robins, and Craig K. Enders stood out. The temporal distribution of the occurrence of keywords indicates distinct phases of research in MVI, characterized by the initial phase with the EM Algorithm, followed by Multiple Imputation, and currently concentrating on Machine Learning techniques.

Keywords: bibliometrics; missing data; imputation; incomplete dataset; data quality.

1.1 INTRODUÇÃO

A agrometeorologia, meteorologia agrícola, baseia-se em ciências físicas e biológicas básicas para descobrir, definir e aplicar o conhecimento do tempo e do clima à produção agrícola (TAKLE, 2015). Visa fornecer serviços e informações agrometeorológicas aos agricultores e tomadores de decisão na produção agrícola (CALANCA, 2014; HOOGENBOOM, 2000; STIGTER, 2007), auxiliando na conservação de recursos naturais e proteção do solo, planta e recursos hídricos (TAKLE, 2003). Devido às condições sob as quais os dados meteorológicos são coletados, normalmente contém uma proporção de dados ausentes, *missings*, os quais podem surgir a partir de erros nas medições, falhas na aquisição de dados, entradas impróprias, problemas nos equipamentos, entre outros (HASAN et al., 2021; JUNNINEN et al., 2004; OSBORNE, 2013; YOZGATLIGIL et al., 2013). Quaisquer que sejam as razões, os *missings* devem ser avaliados e tratados de forma adequada para preparação, organização e estruturação dos dados.

Dependendo da taxa de *missings* e área de aplicação, os dados ausentes podem ser removidos da base de dados sem ter um efeito significativo no resultado da análise (STRIKE; EMAM; MADHAVJI, 2001). A desvantagem da exclusão é a redução do tamanho amostral, diminuindo a precisão da análise estatística (ROTH, 1994). Ao contrário da estratégia de exclusão a imputação de valor ausente, *missing value imputation* (MVI), é a solução mais comumente utilizada para tratar o problema de conjunto de dados incompleto. Os métodos de MVI podem ser classificadas em técnicas estatísticas ou *machine learning*. Das técnicas estatísticas destaque para média, *Linear Regression* (LR), *Least Squares* (LS) e *Expectation Maximization* (EM). Das técnicas baseadas em aprendizado de máquina destaque para *Clustering*, *Decision Tree* (DT), *K-nearest neighbor* (KNN) e *Random Forest* (RF) (LIN; TSAI, 2020).

Com intuito de verificar o impacto na literatura científica mundial e os pilares teóricos do tema MVI, foi aplicado uma análise bibliométrica. A bibliometria surgiu no início do século vinte, com o pesquisador Paul Otlet (1934), no tratado *Traité de Documentation*, consolidando-se com o artigo de Pritchard (1969), *Statistical Bibliography or Bibliometrics?*. Sua origem vem da necessidade do estudo e da avaliação das atividades de produção e comunicação científica (ARAÚJO, 2006).

Desenvolve-se através de leis empíricas sobre o comportamento da literatura, com destaque para o método de medição da produtividade de autores dado pela lei de Lotka (1926), a lei de dispersão do conhecimento científico de Bradford (1934) e o modelo de distribuição e frequência de palavras num texto, de Zipf (1949).

A bibliometria compreende uma análise extensa de trabalhos publicados, utilizando ferramentas estatísticas, visando descobrir as tendências, publicações e citações de um determinado tema, por ano, autor, instituição, país, fonte científica e palavra-chave (ZUPIC; ČATER, 2015). Esta abordagem é relevante na verificação de áreas específicas de impacto mundial, como é o caso da MVI, onde a internacionalização e as colaborações podem ser analisadas (VELASCO-MUÑOZ et al., 2018), fato que motivou a utilização da análise bibliométrica.

Considerando os potenciais pesquisadores e usuários interessados nesta análise, pode-se dizer que os resultados apresentados por este capítulo são úteis na avaliação da produção científica no tema MVI. Podem identificar os principais trabalhos e métodos, direcionando oportunidades, investimentos e futuras linhas de pesquisas.

Neste contexto, o objetivo deste capítulo foi apresentar uma visão global da literatura científica no tema *Missing Value Imputation*, MVI, por meio de uma análise bibliométrica. Evidenciando a produtividade e impacto da produção científica em termos de número de publicações por ano acadêmico, autor, instituição e país, número de citações por documentos e fonte, número de palavras-chave e cocitação bibliográfica.

1.2 MATERIAL E MÉTODOS

Para desenvolver a análise bibliométrica utilizou-se o banco de dados *Web Of Science* (CLARIVATE, 2022a). A *Web Of Science* (WOS) é um banco de dados global de citações independente de editores no qual é possível pesquisas temporais em todas as 254 áreas temáticas a partir de quase 1,9 bilhão de referências citadas de mais de 171 milhões de registros, com mais de 9.000 instituições acadêmicas, coorporativas e governamentais, 21.000 revistas acadêmicas de alta qualidade e revisadas por pares, publicadas em todo o mundo (CLARIVATE, 2022b).

Nesta análise bibliométrica o foco foi publicações científicas, no período de 1940 a 2022, relacionados ao tema *Missing Value Imputation*. A consulta foi realizada, no banco de dados *Web Of Science*, no idioma inglês, para todos os campos pesquisáveis (*tópico, título, autor, títulos da publicação, ano de publicação, afiliação, agência financeira, editora, data de publicação, resumo, número de acesso, endereço, identificadores de autor, palavras-chave de autor, conferência, tipo de documento, DOI, editor, número do subsídio, autor grupo, palavra-chave plus®, idioma, ID PubMed, categorias da Web of Science*) nas publicações, com a estrutura de *string* de busca: *(missing) AND (value* OR data) AND (imputation OR estimation)*.

Para cada publicação identificada, foram consideradas na análise bibliométrica as seguintes informações: fonte, título, autores, ano, instituição, país, referências, resumo e palavras-chave.

As métricas bibliométricas utilizadas foram medidas de produtividade e impacto: número de publicações por ano acadêmico, autor, instituição e país, número de citações por documentos e fonte, número de palavras-chave e cocitação (GRÁCIO, 2016).

A análise de cocitação bibliográfica entre duas publicações visa analisar a frequência que são citadas juntas, evidenciando o grau de associação entre pares de documentos, segundo a compreensão da comunidade científica citante (SMALL, 1973).

As técnicas de agrupamento (ou *clustering*) desempenham um papel fundamental na pesquisa bibliométrica, usadas para identificar grupos de publicações, autores ou periódicos relacionados. A bibliometria faz uso de tais técnicas, as quais foram desenvolvidas principalmente em áreas como estatística, ciência da computação e ciência de redes (VAN ECK; WALTMAN, 2017). Os *clusters* são formados por termos relacionados, agrupados pela mesma cor. A proximidade dos termos pode ser interpretada como uma indicação da semelhança do contexto em que ocorrem (WOLSKI et al., 2021). A representação de cada termo pode ser feita por um círculo, com seu tamanho refletindo o número de publicações em que o termo foi encontrado, e a distância entre dois termos oferece uma indicação aproximada da relação dos termos. A similaridade entre dois termos foi

determinada com base em coocorrências, quanto maior o número de publicações em que ambos foram encontrados, mais forte será a relação entre os termos (VAN ECK; WALTMAN, 2017). Para maiores detalhes das técnicas de *clustering* consultar os pesquisadores Waltman e Van Eck (2012, 2013).

A construção e visualização das redes bibliométricas foram realizadas pelos softwares *VOSviewer*[2] (VAN ECK; WALTMAN, 2010; VOSVIEWER, 2022) e o pacote *Bibliometrix*[3] (ARIA; CUCCURULLO, 2017).

O VOSviewer (2023) é um software livre especialmente projetado para gerar a representação gráfica de mapas bibliométricos (rede, sobreposição e densidade), e pode ser usado para construir e visualizar mapas bibliométricos de dados de coautoria (autores, organizações e países), coocorrência de palavras-chave (dos autores), citação (documentos, fontes, autores, organizações e países), cocitação (referências, fontes e autores). Permite realizar a detecção de comunidades científicas usando a técnica de agrupamento (MORESI; PINHO; COSTA, 2021).

O pacote *Bibliometrix*, desenvolvido em linguagem ***R*** pelos pesquisadores Aria e Cuccurullo (2017), oferece recursos para mapas de agrupamentos de documentos e estruturas conceitual, intelectual e social. Os mapas de agrupamentos são figuras que combinam unidades de análise (documentos, autores e fontes), medida de acoplamento (referências, palavras-chave dos autores, títulos e resumos) e medida de impacto (contagens por citações locais ou globais). A estrutura conceitual se baseia na coocorrência de palavras, oferece as opções de rede de coocorrência, mapa e evolução temática de palavras-chave dos autores, título e resumo, com a visualização dos mapas, das tabelas de dados das redes e dos agrupamentos. A estrutura intelectual se baseia nos artigos para obter a rede de cocitação de artigos, de autores e de periódicos e a rede histórica de citação direta. A estrutura social apresenta a rede colaboração de autores, instituições e países (MORESI; PINHO; COSTA, 2021).

---

[2] https://www.vosviewer.com

[3] https://www.bibliometrix.org

1.3 RESULTADOS E DISCUSSÃO

A estrutura de *string* proposta, *(missing) AND (value* OR data) AND (imputation OR estimation),* resultou em um total de 19.745 publicações, as quais foram avaliadas quanto as métricas: número de publicações por ano acadêmico, número de publicações por autor, número de publicações por instituição, número de publicações por país, número de citações por periódico, número de citações por documento, número de palavras-chave e número de cocitações.

1.3.1 NÚMERO DE PUBLICAÇÕES POR ANO

As publicações foram extraídas da base *Web Of Science*, no dia 2 de janeiro de 2023 e suas principais características são apresentadas na Tabela 1. Neste levantamento, os primeiros trabalhos são datados de 1940 (CORNISH, 1940a) e (CORNISH, 1940b), feitos por *E. A. Cornish*, nos quais tratam de dados faltantes em experimentos agrícolas.

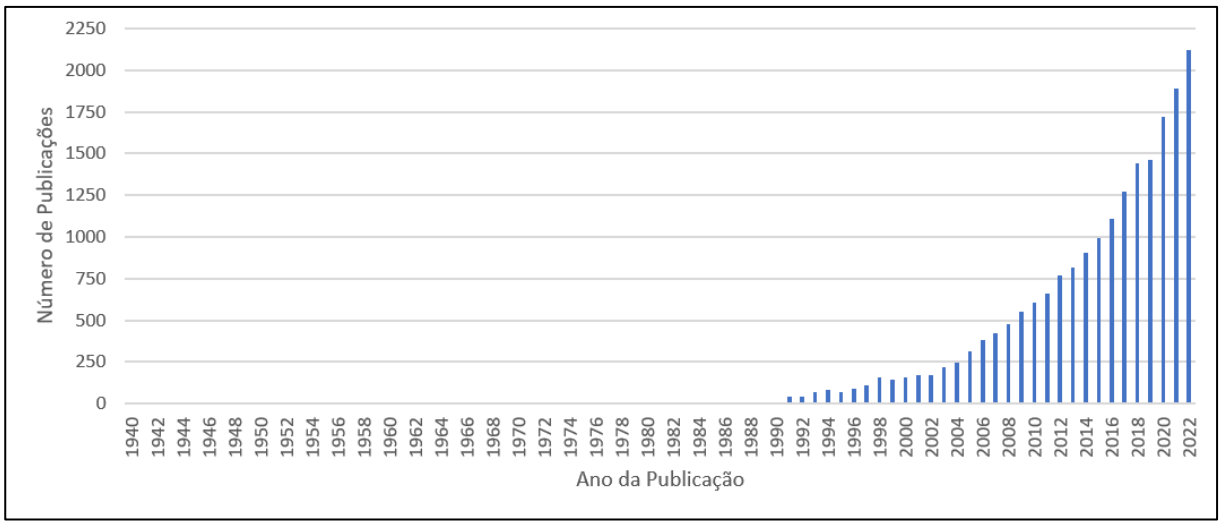
Tabela 1 – Principais características da base de publicações identificadas

Informação	**Resultado**
Período de análise	1940 até 2022
Número de publicações	19.745
Número de autores	62.193
Número de palavras-chave	38.783
Número de referências citadas*	450.673
Número de fontes (periódicos, livros...)	6.225

Fonte: Autoria Própria a partir de dados WOS.
*distintas.

A Figura 1 apresenta a evolução temporal das publicações, no tema *Missing Value Imputation.*

Figura 1 – Evolução temporal das publicações relacionadas com MVI

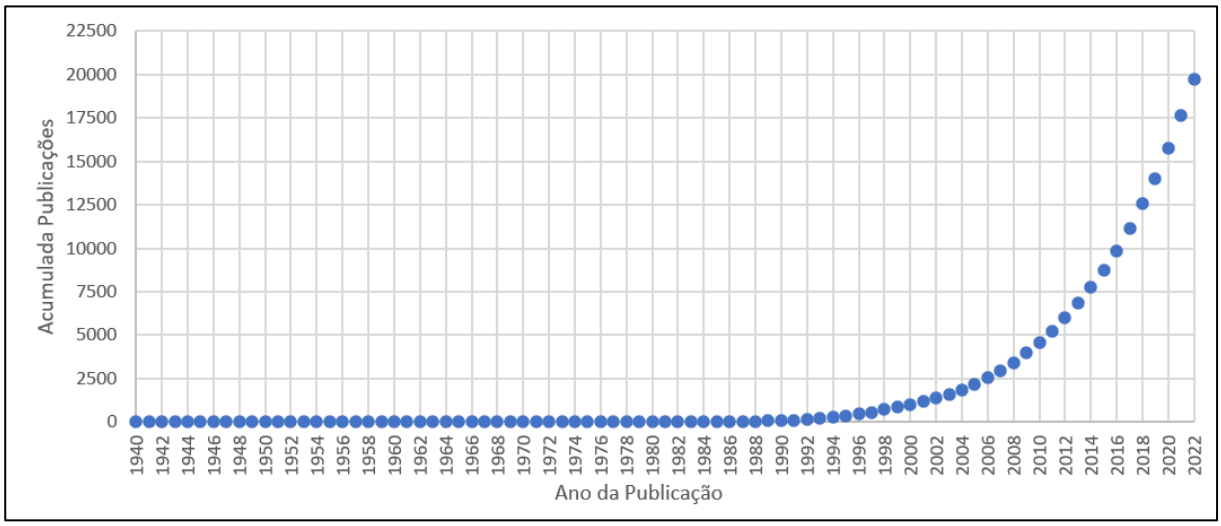


Fonte: Autoria Própria a partir de dados WOS.

Notadamente, a partir de 1990 o número anual de publicações em MVI apresentou um crescimento acentuado, chegando em 2022 a uma produção anual de cerca de 2118 publicações, Figura 1.

Na Figura 2, pode ser observado o crescimento acumulado das publicações, indicando que partir de 1990, houve um padrão exponencial de crescimento.

Figura 2 – Evolução temporal acumulada das publicações em MVI



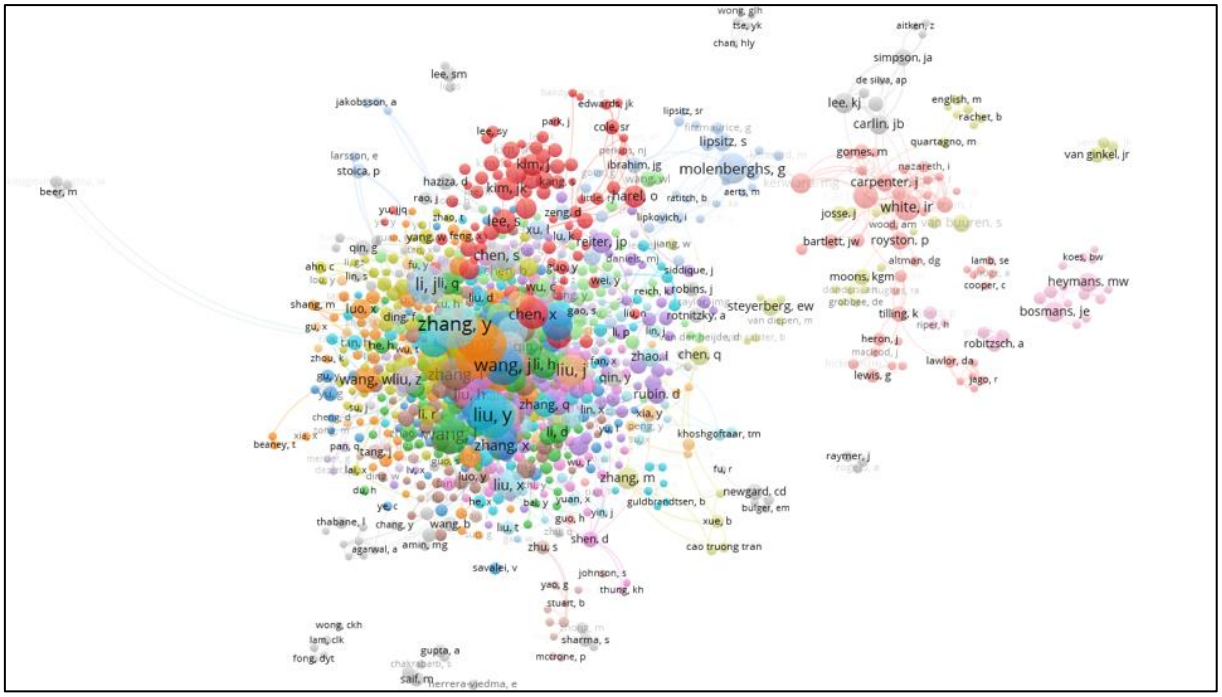
Fonte: Autoria Própria a partir de dados WOS.

Percebe-se que o número de publicações em MVI no período de 1940 a 1990 pode ser considerado muito baixo, correspondendo a 0,3% do total de publicações identificadas, enquanto 99,7% foram publicados entre 1990 e 2022, indicando o crescimento acentuado desta área nos últimos 30 anos, Figura 2. Isto mostra a recente importância do tema *Missing Value Imputation*.

1.3.2 NÚMERO DE PUBLICAÇÕES POR AUTOR

Com o portfólio de análise, de 19.745 publicações identificadas, observou-se um total de 62.193 autores, os quais são identificados os principais na Figura 3, onde pode-se observar o grau de importância relativa dos autores de acordo com os tamanhos dos círculos (em função do número de publicações).

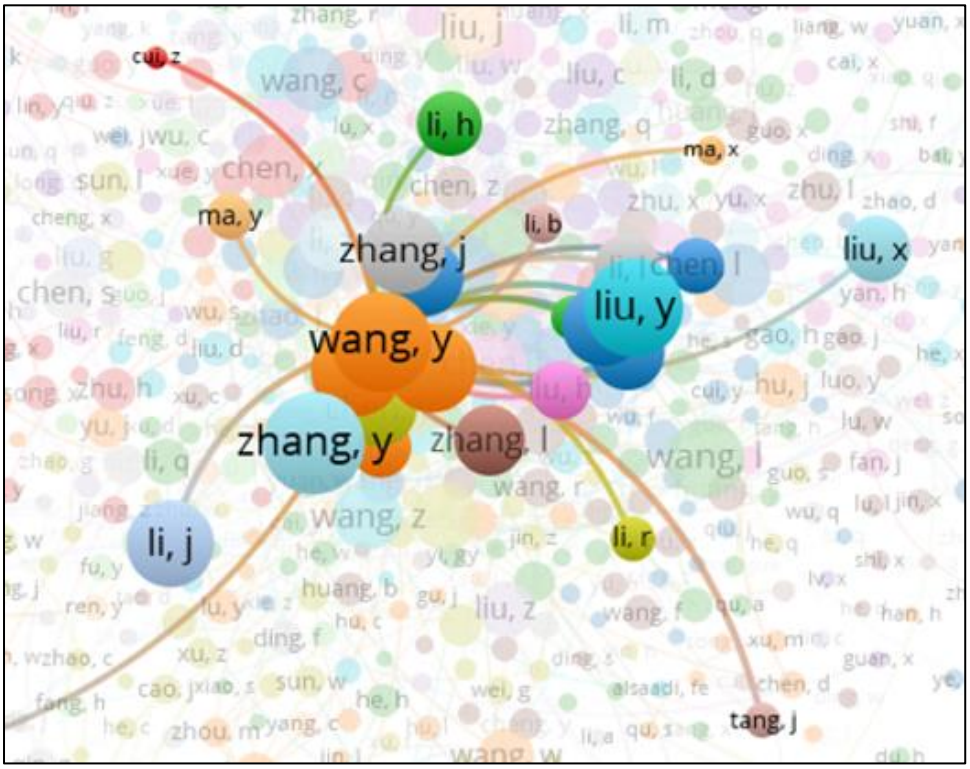
Figura 3 – Rede de colaboração dos autores nas publicações identificas pelo número de publicações



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

A Figura 4 mostra a rede de colaboração com destaque para os autores *Zhang Y*., *Wang Y*., *Liu Y*., autores com maior número de publicações, das referências identificadas.

Figura 4 – Rede de colaboração dos autores com maior número de publicações



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

A Tabela 2 destaca os primeiros dez autores[4], ordenados de forma decrescente (rank) em relação ao número de publicações para as referências identificadas neste estudo, 19.745 publicações. Existe um predomínio de pesquisadores chineses, resultado do forte desenvolvimento econômico focado em inovação e tecnologia (KELLY et al., 2022).

[4] Pela dificuldade de tratativas não foram comentados pesquisadores da Tabela 2.

Tabela 2 – Número de publicações por autor

Rank	AUTOR	PUBLICAÇÕES	%
1	ZHANG Y	174	0,88
2	WANG Y	162	0,82
3	LIU Y	154	0,78
4	LI Y	129	0,65
5	WANG X	129	0,65
6	LI J	128	0,65
7	WANG J	121	0,61
8	ZHANG J	112	0,57
9	WANG L	106	0,54
10	ZHANG Z	104	0,53

Fonte: Autoria Própria a partir de dados WOS.

Considerando as citações da base *Web Of Science*, os primeiros dez autores com maiores números de citações são apresentados na Tabela 3.

Tabela 3 – Número de citações por autor, base *Web Of Science*

Rank	AUTOR	CITAÇÕES
1	ROYSTON, P	14.005
2	SCHAFER, J	13.318
3	WHITE, IR	13.007
4	VAN BUUREN, S	11.719
5	WOOD, A M	7.726
6	RUBIN, D	7.673
7	YANG, J	7.438
8	VISSCHER, P M	6.936
9	GODDARD, M E	6.894
10	ENDERS, C K	6.178

Fonte: Autoria Própria a partir de dados WOS.

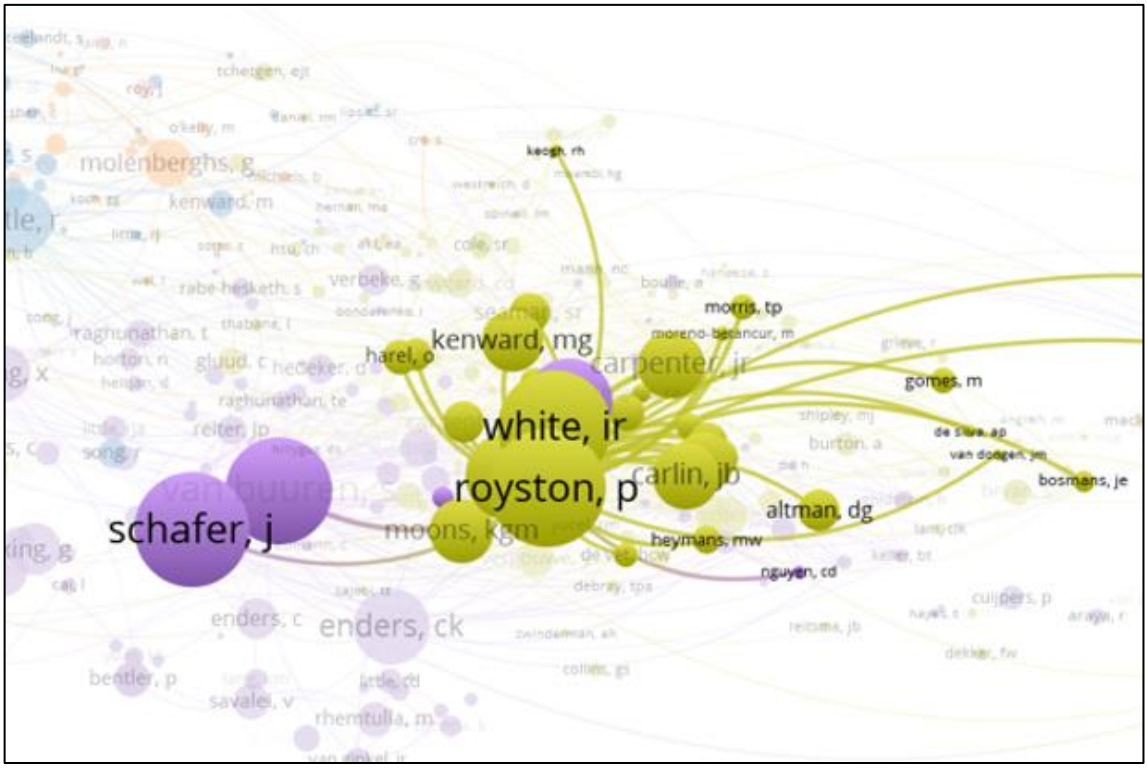
Com base na Tabela 3, dentre os pesquisadores proeminentes destaca-se Patrick Royston, cujas contribuições foram significativas para o desenvolvimento do método de imputação múltipla por meio de equações encadeadas. De acordo com Royston (2004), o método de imputação múltipla é uma abordagem eficaz que permite estimar valores plausíveis para os dados ausentes e capturar a incerteza associada a essas estimativas. No livro de Schafer (1997) aborda a análise de dados multivariados com valores ausentes, explorando as principais técnicas e abordagens e fundamentos teóricos por trás da análise de dados incompletos. Apresenta a abordagem de *Expectation-Maximization* (EM) como uma técnica poderosa para a imputação de valores ausentes. O artigo de White et al. (2009), referência valiosa para pesquisadores e profissionais envolvidos em estudos epidemiológicos e clínicos que enfrentam o desafio dos dados faltantes. Sua visão sobre as vantagens e desafios da imputação de valores ausentes contribuem para uma melhor compreensão e aplicação adequada dessa abordagem em pesquisas na área da saúde. Stef van Buuren é reconhecido como um dos desenvolvedores do pacote estatístico *Multiple Imputation by Chained Equations* (MICE), em (2011), um dos trabalhos mais importantes na área de imputação de valores ausentes. O MICE é uma ferramenta poderosa e flexível para a imputação de valores ausentes em dados multivariados, utiliza modelos específicos para cada variável com valores ausentes e itera entre eles para gerar imputações múltiplas. A pesquisadora Angela M. Wood tem contribuições em várias áreas da pesquisa biomédica, incluindo epidemiologia, estudos clínicos e saúde pública. Donald Rubin (1987) apresentou a teoria da imputação múltipla com fundamentos teóricos sólidos para a imputação de valores faltantes. Os pesquisadores Jian Yang, Peter M. Visscher e Michael E. Goddard são reconhecidos por suas contribuições em genética quantitativa e métodos estatísticos para lidar com dados ausentes em estudos genômicos. Craig K. Enders publicou o livro *Applied Missing Data Analysis* (2010), fornece orientações práticas e metodológicas para lidar com dados faltantes em pesquisas e estudos de diversas áreas, como ciências sociais, saúde, psicologia e educação. Apresenta uma variedade de abordagens para lidar com dados ausentes, incluindo métodos de imputação simples e avançados, como a imputação múltipla e modelagem de equações estruturais com imputação. Discute os princípios e as etapas envolvidas

na aplicação dessas técnicas, destacando suas vantagens, limitações e considerações práticas.

Esses pesquisadores são os mais citados no tema imputação de valores ausentes, cujas contribuições teóricas, metodológicas e práticas têm sido amplamente reconhecidas e aplicadas na comunidade acadêmica e científica.

Na figura 5 verifica-se uma forte colaboração entre *ROYSTON, P.* e *WHITE, I. R.*, tendo como destaque a publicação *Multiple imputation using chained equations: Issues and guidance for practice* (WHITE; ROYSTON; WOOD, 2011), com 4753 citações, a qual descreve a imputação múltipla por meio de equações encadeadas, apresentando os princípios deste procedimento no domínio de pesquisas médicas e epidemiológicas, considerando tanto variáveis categóricas quanto quantitativas.

Figura 5 – Destaque para os autores com maior número de citações



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

1.3.3 NÚMERO DE PUBLICAÇÕES POR INSTITUIÇÃO

As publicações identificadas congregam cerca de 12.925 instituições. Das quais, as universidades Norte Americanas liberam o ranking das instituições com maior número de participações nas publicações em MVI, apresentado na Tabela 4.

Das universidades brasileiras, destaque para as universidades paulistas: USP (73), UNICAMP(28) e UNESP(18).

Tabela 4 – Número de publicações por instituição

Rank	ORGANIZAÇÃO	FREQUÊNCIA	%
1	HARVARD UNIV	319	1,62
2	UNIV MICHIGAN	311	1,58
3	UNIV WASHINGTON	299	1,51
4	CHINESE ACAD SCI	281	1,42
5	UCL	262	1,33
6	UNIV N CAROLINA	253	1,28
7	UNIV CALIF LOS ANGELES	207	1,05
8	UNIV OXFORD	203	1,03
9	LONDON SCH HYG & TROP MED	194	0,98
10	UNIV MELBOURNE	183	0,93
⋮	⋮	⋮	⋮
78	USP	73	0,37
⋮	⋮	⋮	⋮
213	UNICAMP	28	0,14
⋮	⋮	⋮	⋮
494	UNESP	18	0,09

Fonte: Autoria Própria a partir de dados WOS.

Das Universidades Americanas, tem-se: *Harvard University* (1,62%), *University of Michigan* (1,58%), *University of Washington* (1,51%), *University of North Carolina* (1,28%) e *University of California* (1,05%). *Harvard University*, privada localizada em Cambridge, Massachusetts, Estados Unidos. Fundada em 1636, é uma das instituições de ensino mais renomadas e prestigiadas do mundo, conhecida por sua excelência acadêmica em diversas áreas, incluindo ciências, humanidades, negócios, direito, medicina e engenharia (“Harvard University”, 2022). A *University of Michigan*, pública localizada em Ann Arbor, Michigan, Estados Unidos. Fundada em 1817, a instituição oferece uma ampla variedade de programas acadêmicos,

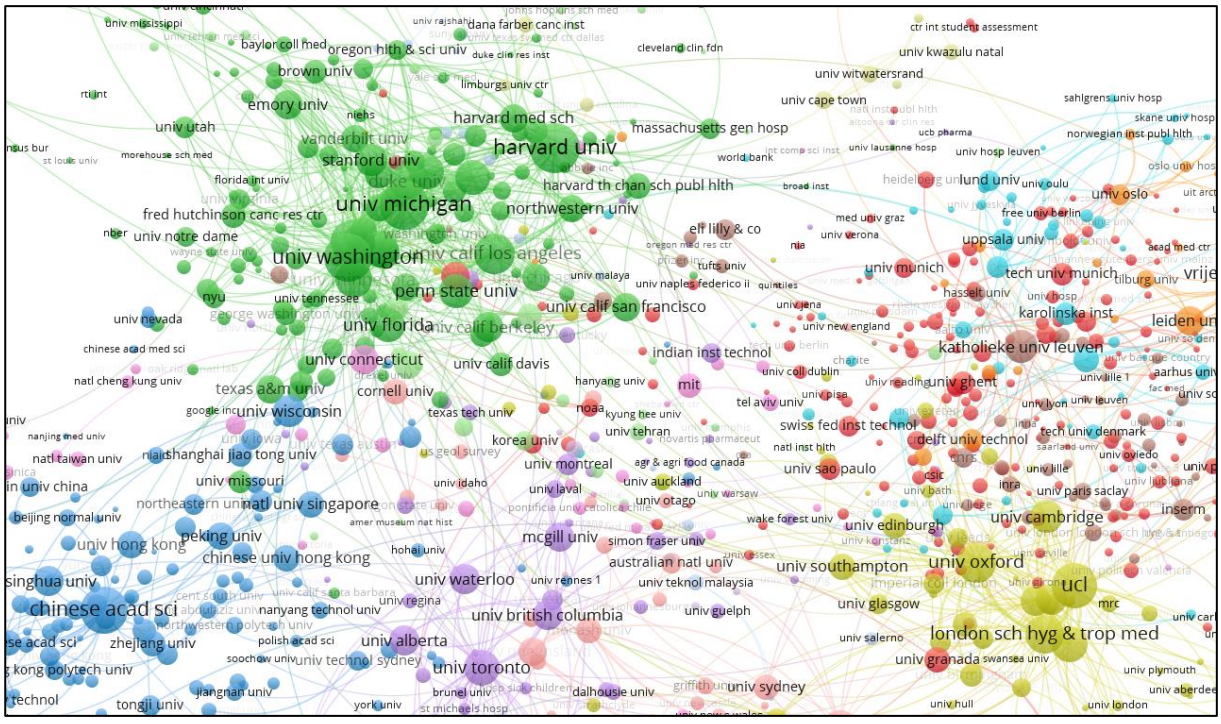
incluindo artes e ciências, engenharia, medicina, direito, negócios, educação, arquitetura. A universidade é conhecida por sua pesquisa de ponta e seu compromisso com a excelência acadêmica (“University of Michigan”, 2022). A *University of Washington* é pública localizada em Seattle, Washington, Estados Unidos. Fundada em 1861, oferece uma ampla variedade de programas acadêmicos em várias disciplinas, incluindo ciências da saúde, ciências sociais, engenharia, ciências naturais e humanidades. A *University of North Carolina*, pública localizada em Chapel Hill, Carolina do Norte, Estados Unidos. Fundada em 1789, uma das instituições mais antigas dos Estados Unidos e oferece uma ampla variedade de programas acadêmicos em várias disciplinas, incluindo ciências sociais, humanidades, ciências da saúde, negócios, direito e ciências naturais (“University of North Carolina”, 2023). A *University of California, Los Angeles* (UCLA), localizada em Los Angeles, Califórnia, Estados Unidos. Fundada em 1919, uma das universidades americanas mais renomadas e conhecida por sua excelência acadêmica em diversas áreas, incluindo artes, ciências sociais, ciências naturais, engenharia, medicina e ciências humanas (“University of California, Los Angeles”, 2023).

Das universidades do Reino Unido, tem-se: *University College London* (1,33%), *University of Oxford* (1,03%), *London School of Hygiene & Tropical Medicine* (0,98%). *University College London*, UCL, pública localizada em Londres, Reino Unido. Fundada em 1826, a instituição é reconhecida por sua excelência em pesquisa e ensino em uma ampla variedade de disciplinas, como ciências sociais, humanidades, ciências da vida, engenharia, medicina e ciências exatas (“University College London”, 2023). A *University of Oxford*, pública localizada em Oxford, Reino Unido. Fundada no século XII, das universidades mais antigas do mundo de língua inglesa e uma das mais prestigiadas do Reino Unido. A universidade oferece uma ampla variedade de programas acadêmicos em várias disciplinas, incluindo ciências, humanidades, medicina, direito e ciências sociais (“University of Oxford”, 2023). A *London School of Hygiene & Tropical Medicine* é uma instituição de ensino e pesquisa localizada em Londres, Reino Unido. Fundada em 1899, especializada em saúde pública, medicina tropical e áreas relacionadas. A instituição desempenha um papel fundamental na pesquisa e na formação de profissionais de saúde, abordando questões globais de saúde e doenças tropicais (LSHTM, 2023).

Entre as dez, ainda tem uma representante da Austrália, *University of Melbourne* (0,93%). Universidade pública localizada em Melbourne, Austrália. É a segunda universidade mais antiga e importante instituição de ensino superior da Austrália. Fundada em 1853, oferece uma ampla gama de programas acadêmicos em diversas áreas, incluindo ciências, artes, engenharia, medicina, negócios e ciências sociais (“University of Melbourne”, 2023). A mais jovem, a proeminente *Chinese Academy of Sciences* (1,42%) instituição de pesquisa de alto nível localizada na China. Fundada em 1949, composta por vários institutos de pesquisa e centros científicos em todo o país, dedicados a uma ampla gama de disciplinas científicas, incluindo ciências naturais, engenharia, ciências da saúde e ciências sociais (“Chinese Academy of Sciences”, 2023).

A rede de colaboração entre as Instituições, cerca de 12.925, apresentada na Figura 6, mostra uma intensa colaboração entre as Universidades Norte Americanas, destacando as *Universidade de Harvard* e *Michigan* (agrupadas na cor verde). Em azul, a rede de colaboração entre as universidades asiáticas, com a *University of Chinese Academy of Sciences* com maior número de publicações, e em amarelo a rede de colaboração entre as universidades do Reino Unido.

Figura 6 – Rede de colaboração das instituições das referências levantadas



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

A Tabela 5 apresenta as dez primeiras instituições mais citadas, de acordo com a base *Web Of Science*. A maioria das instituições são Norte Americanas. Destaque para *University of Pennsylvania*, localizada na cidade da Filadélfia, fundada por Benjamin Franklin em 1740 (“University of Pennsylvania”, 2022).

Tabela 5 – Número de citações por Instituição, base *Web Of Science*

Rank	ORGANIZAÇÃO	FREQUÊNCIA
1	PENN STATE UNIV	26.402
2	HARVARD UNIV	23.944
3	UNIV MELBOURNE	15.864
4	UNIV WASHINGTON	14.711
5	UCL	13.714
6	UNIV OXFORD	12.977
7	UNIV MICHIGAN	10.879
8	UNIV CAMBRIDGE	10.293
9	UNIV CALIF BERKELEY	9.533
10	INST PUBL HLTH	9.160

Fonte: Autoria Própria a partir de dados WOS.

1.3.4 NÚMERO DE PUBLICAÇÕES POR PAÍS

Das 19.745 publicações identificadas nesta pesquisa observou-se o envolvimento de 158 países, destacando-se, na Tabela 6, os dez primeiros países em número de participações nas publicações identificadas, observando como protagonistas os Estados Unidos (7.372), China (2.874) e Inglaterra(2044). O Brasil aparece em décimo nono, com a participação na produção de 290 publicações.

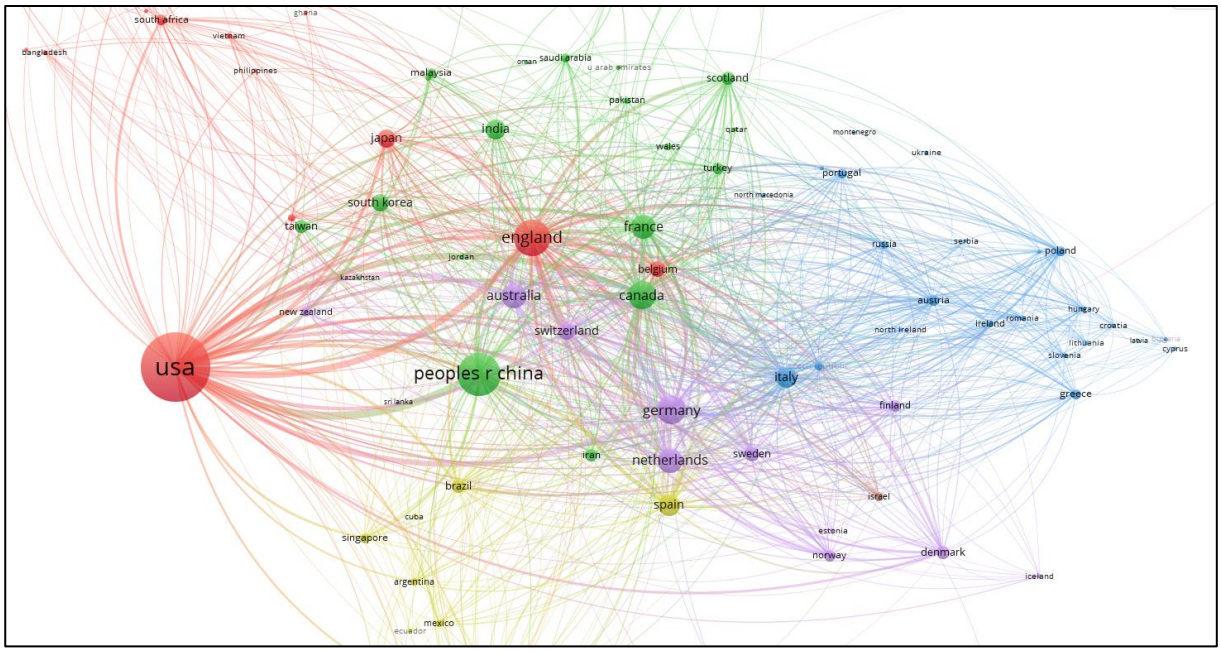
Tabela 6 – Número de publicações por país

Rank	PAÍS	FREQUÊNCIA	%
1	USA	7.372	37,3
2	PEOPLES R CHINA	2.874	14,6
3	ENGLAND	2.044	10,4
4	GERMANY	1.270	6,4
5	CANADA	1.256	6,4
6	NETHERLANDS	934	4,7
7	AUSTRALIA	929	4,7
8	FRANCE	876	4,4
9	SPAIN	707	3,6
10	ITALY	695	3,5
⋮	⋮	⋮	⋮
19	BRAZIL	290	1,5

Fonte: Autoria Própria a partir de dados WOS.

A Figura 7 apresenta por meio de agrupamentos (*clusters*) a rede de colaboração entre os países.

Figura 7 – Rede de colaboração dos países das referências levantadas



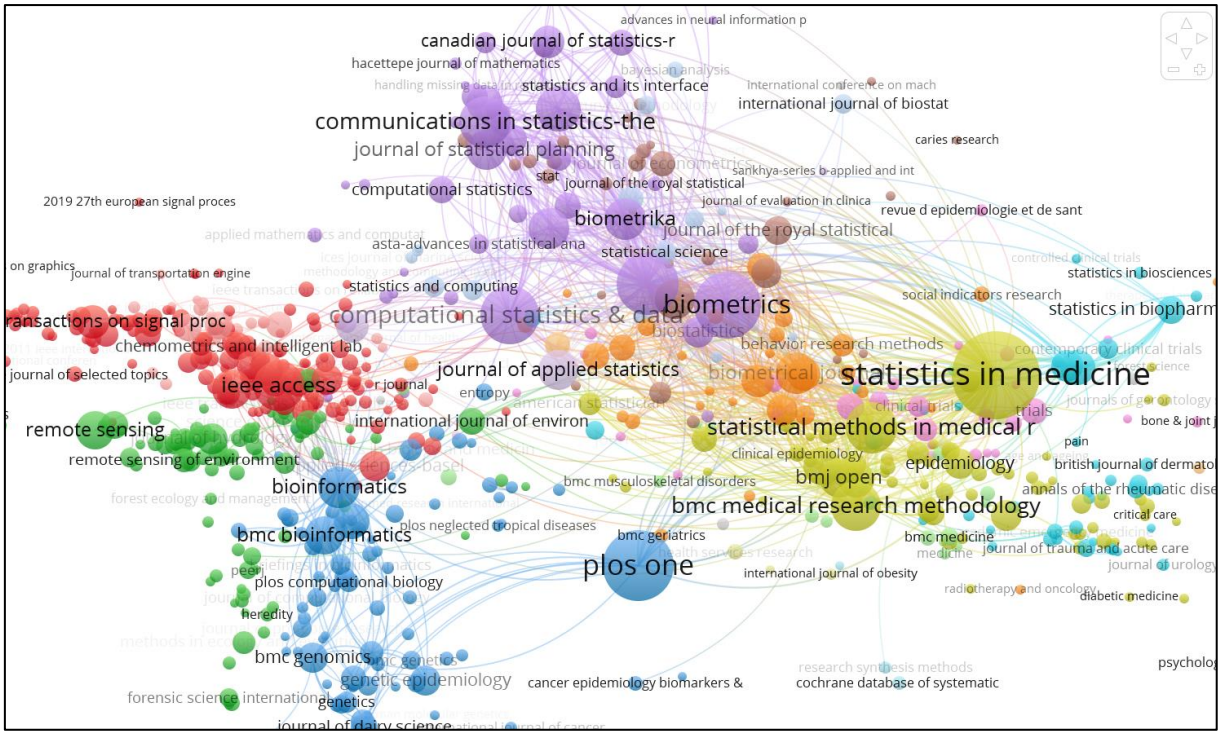
Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

Nesta visualização, Figura 7, os países são agrupados em nove clusters. Destaque para os clusters em *cor vermelha* com os Estados Unidos, Inglaterra e Japão, em *cor verde* China, Canada e França, em *cor roxa* Alemanha, Holanda, Australia, em *azul* Itália e em *amarelo* Espanha e Brasil.

1.3.5 NÚMERO DE PUBLICAÇÕES POR PERIÓDICO

Neste levantamento, 19.745 publicações, foram publicadas em 6.225 fontes científicas. A Figura 8 apresenta a rede de colaboração com base nas citações entre periódicos. O agrupamento roxo observa-se periódicos voltados para área de estatística e estatística computacional. Em azul, periódicos voltados a aplicação da computação na biologia. Em amarelo estão os periódicos voltados a área médica. Em vermelho, são periódicos voltados a quimiometria e computação.

Figura 8 – Rede de colaboração entres os principais periódicos em número de publicações



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

A Tabela 7 lista os dez principais periódicos em número de publicações, de acordo com os documentos identificados.

Tabela 7 – Principais periódicos em número de publicações

Rank	PERIÓDICOS	CITAÇÕES	%
1	STATISTICS IN MEDICINE	384	2,0
2	PLOS ONE	251	1,3
3	BIOMETRICS	224	1,2
4	JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION	193	1,0
5	COMPUTATIONAL STATISTICS & DATA ANALYSIS	173	0,9
6	COMMUNICATIONS IN STATISTICS-THEORY AND METHODS	136	0,7
7	STATISTICAL METHODS IN MEDICAL RESEARCH	134	0,7
8	IEEE ACCESS	131	0,7
9	BMC MEDICAL RESEARCH METHODOLOGY	126	0,6
10	STATISTICA SINICA	103	0,5

Fonte: Autoria Própria a partir de dados WOS.

Estes dez periódicos, apresentados da Tabela 7, são responsáveis por 9,5% das publicações, expressiva representatividade uma vez que se tem 6.225 fontes científicas. Estes periódicos acadêmicos desempenham um papel crucial na disseminação do conhecimento científico e na promoção do avanço do conhecimento.

Voltados para área de medicina e saúde destaque para: *Statistics in Medicine*, *Statistical Methods in Medical Research*, *BMC Medical Research Methodology* e *Biometrics*. O periódico *Statistics in Medicine* é reconhecido como um dos principais veículos para a publicação de pesquisas estatísticas relacionadas à medicina. Foi fundada em 1982, oferece uma plataforma para a divulgação de novas metodologias estatísticas e sua aplicação em problemas médicos (“Stat. Med.”, 2023). O periódico *Statistical Methods in Medical Research* é dedicado especificamente à aplicação de métodos estatísticos na pesquisa médica. Publica estudos que destacam as melhores práticas estatísticas e metodológicas em várias áreas da medicina, como ensaios clínicos, estudos observacionais e meta-análises (“Stat. Methods Med.

Res.”, 2023). O periódico de acesso aberto *BMC Medical Research Methodology* é especializado em metodologia de pesquisa médica, incluindo aspectos estatísticos e metodológicos. Publica artigos sobre metodologia de pesquisa epidemiológica, ensaios clínicos, metanálise e revisão sistemática (“BMC Medical Research Methodology”, 2023). *Biometrics* é um periódico especializado em promover e ampliar a aplicação de métodos estatísticos e matemáticos nas principais disciplinas das biociências, publicado pela *International Biometric Society*, desde 1945. Publica pesquisas originais que abrangem uma variedade de tópicos, como modelagem estatística de doenças, genômica, epidemiologia e análise de imagem médica (“Biometrics”, 2023).

Periódicos voltados a aplicações de métodos matemáticos e estatísticos destaque para: *PLOS ONE* e *IEEE Access*. *PLOS ONE* é uma comunidade de revistas inclusivas de acesso aberto revisado por pares publicado pela *Public Library of Science* (PLOS) desde 2006, cobre pesquisas primárias de qualquer disciplina dentro da ciência e da medicina (“PLOS One”, 2023). O periódico *IEEE Access* é uma publicação de acesso aberto que abrange várias áreas da engenharia, incluindo a engenharia biomédica. Publicado pelo *Instituto de Engenheiros Elétricos e Eletrônicos* foi criado em 2013 (“IEEE Access”, 2023).

Os periódicos rigorosos em métodos estatísticos e computação destaque para: *Journal of the American Statistical Association*, *Computational Statistics & Data Analysis*, *Communications in Statistics - Theory and Methods* e *Statistica Sinica*. O *Journal of the American Statistical Association* (JASA) é o principal periódico publicado pela *American Statistical Association*, o principal órgão profissional de estatísticos nos Estados Unidos. Os artigos se concentram em aplicações estatísticas, teoria e métodos em ciências econômicas, sociais, físicas, de engenharia e da saúde, fundada em 1922. (“J. Am. Stat. Assoc.”, 2022). O periódico *Computational Statistics and Data Analysi*s (CSDA), uma publicação oficial da rede *Computational and Methodological Statistics* (CMStatistics) e da *International Association for Statistical Computing* (IASC), é uma revista dedicada à divulgação de pesquisas metodológicas e aplicações nas áreas de estatística computacional e análise de dados. (“Computational Statistics & Data Analysis”, 2023). *Communications in Statistics – Theory and Methods* é um periódico dedicado à publicação de avanços teóricos e metodológicos em Probabilidade e Estatística

(“Communications in Statistics - Theory and Methods”, 2023). A *Statistica Sinica* foi criada em 1991, co-patrocinada pelo *Institute of Statistical Science, Academia Sinica* (ISSAS), Taiwan e pela *International Chinese Statistical Association*. Fornece um fórum para publicação de trabalhos inovadores de alta qualidade em todas as áreas de estatística e ciência de dados, incluindo teoria, metodologia e aplicações (“Statistica Sinica”, 2023).

A Tabela 8 apresenta uma classificação para os principais periódicos considerando as informações do Qualis-Periódicos, disponibilizado pela *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES, 2022).

Tabela 8 – Qualis dos principais periódicos em número de publicações

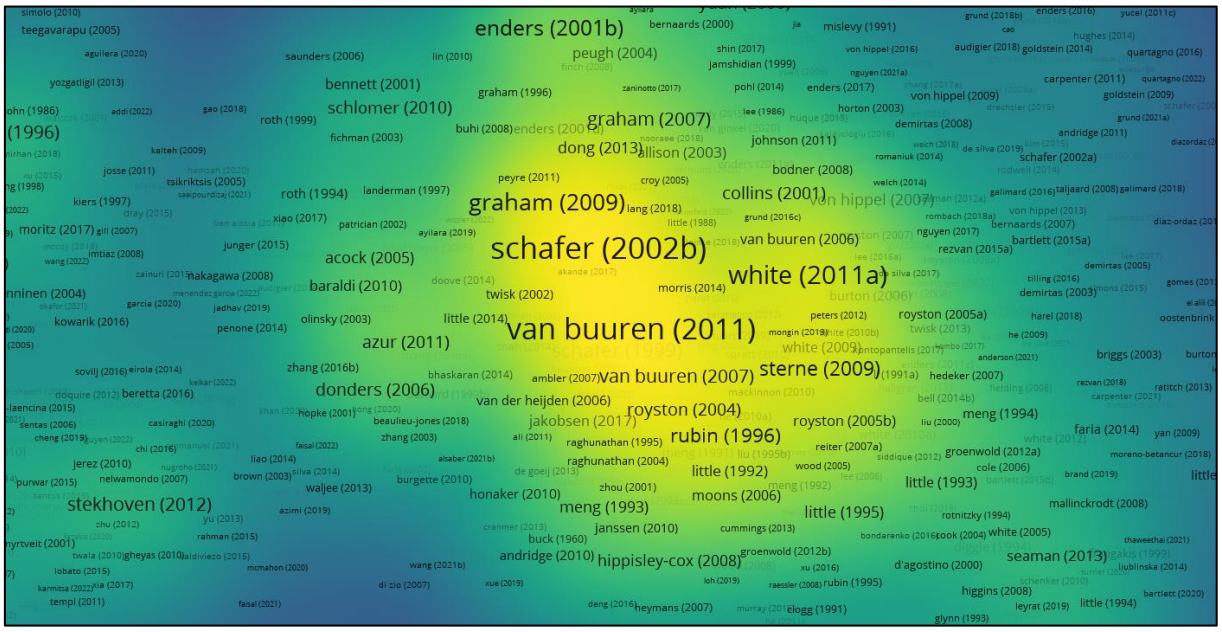
Qualis-Periódicos	**Periódicos**
A1	*Biometrics*
	IEEE Access
	Journal of the American Statistical Association
	PLOS ONE
A2	*BMC Medical Research Methodology*
	Computational Statistics & Data Analysis
	Statistical Methods in Medical Research
	Statistics in Medicine
B2	*Communications in Statistics - Theory and Methods*

Fonte: Autoria Própria a partir de dados da CAPES.

1.3.6 NÚMERO DE CITAÇÕES POR ARTIGO PUBLICADO

O gráfico de densidade, dado pela Figura 9, mostra a importância e intensidade das citações, correspondentes as publicações identificadas. Com a maior intensidade de citações indicada em amarelo e a cor azul menor intensidade. Destaque para os trabalhos Schafer (2002), Van Buuren (2011) e White (2011).

Figura 9 – Gráfico de densidade das publicações mais citados em MVI



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

A Tabela 9 sumariza os primeiros vinte trabalhos com maior número de citações das publicações identificadas e extraídas da base *Web Of Science*, em 2 de janeiro de 2023.

Tabela 9 – Top 20 dos artigos mais citados em MVI

Rank	Título	Autor	Citações
1	*MISSING DATA: OUR VIEW OF THE STATE OF THE ART*	(SCHAFER; GRAHAM, 2002)	7783
2	*MICE: MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS IN R*	(VAN BUUREN; GROOTHUIS-OUDSHOORN, 2011)	7097
3	*MULTIPLE IMPUTATION USING CHAINED EQUATIONS: ISSUES AND GUIDANCE FOR PRACTICE*	(WHITE; ROYSTON; WOOD, 2011)	4753
4	*MISSING DATA ANALYSIS: MAKING IT WORK IN THE REAL WORLD*	(GRAHAM, 2009)	3808
5	*A PRIMER ON MAXIMUM LIKELIHOOD ALGORITHMS AVAILABLE FOR USE WITH MISSING DATA*	(ENDERS, 2001)	3012

(continua)

(continuação)

Rank	Título	Autor	Citações
6	*MULTIPLE IMPUTATION FOR MISSING DATA IN EPIDEMIOLOGICAL AND CLINICAL RESEARCH: POTENTIAL AND PITFALLS*	(STERNE et al., 2009)	2381
7	*MISSING VALUE ESTIMATION METHODS FOR DNA MICROARRAYS*	(TROYANSKAYA et al., 2001)	2333
8	*MULTIPLE IMPUTATION: A PRIMER*	(SCHAFER, 1999)	2332
9	*MULTIPLE IMPUTATION AFTER 18+ YEARS*	(RUBIN, 1996)	2090
10	*MISSFOREST-NON-PARAMETRIC MISSING VALUE IMPUTATION FOR MIXED-TYPE DATA*	(STEKHOVEN; BÜHLMANN, 2012)	1761
11	*MULTIPLE IMPUTATION OF DISCRETE AND CONTINUOUS DATA BY FULLY CONDITIONAL SPECIFICATION*	(VAN BUUREN, 2007)	1686
12	*HOW MANY IMPUTATIONS ARE REALLY NEEDED? - SOME PRACTICAL CLARIFICATIONS OF MULTIPLE IMPUTATION THEORY*	(GRAHAM; OLCHOWSKI; GILREATH, 2007)	1666
13	*THE EXPECTATION-MAXIMIZATION ALGORITHM*	(MOON, 1996)	1651
14	*A COMPARISON OF INCLUSIVE AND RESTRICTIVE STRATEGIES IN MODERN MISSING DATA PROCEDURES*	(COLLINS; SCHAFER; KAM, 2001)	1603
15	*MULTIPLE IMPUTATION OF MISSING BLOOD PRESSURE COVARIATES IN SURVIVAL ANALYSIS*	(VAN BUUREN; BOSHUIZEN; KNOOK, 1999)	1552
16	*MULTIPLE IMPUTATION OF MISSING VALUES*	(ROYSTON; DIVISION, 2004)	1548
17	*ESTIMATION OF REGRESSION COEFFICIENTS WHEN SOME REGRESSORS ARE NOT ALWAYS OBSERVED*	(ROBINS; ROTNITZKY; ZHAO, 1994)	1475
18	*REVIEW: A GENTLE INTRODUCTION TO IMPUTATION OF MISSING VALUES*	(DONDERS et al., 2006)	1467
19	*MULTIPLE IMPUTATION BY CHAINED EQUATIONS: WHAT IS IT AND HOW DOES IT WORK?*	(AZUR et al., 2011)	1363
20	*WHEN CAN CATEGORICAL VARIABLES BE TREATED AS CONTINUOUS? A COMPARISON OF ROBUST CONTINUOUS AND CATEGORICAL SEM ESTIMATION METHODS UNDER SUBOPTIMAL CONDITIONS*	(RHEMTULLA; BROSSEAU-LIARD; SAVALEI, 2012)	1181

Fonte: Autoria Própria a partir de dados WOS.

Destaque para os três primeiros artigos da Tabela 9, o artigo *Missing data: our view of the state of the art*, dos pesquisadores Schafer e Graham (2002). Apresentaram uma revisão do tema MVI, explicando os equívocos e práticas não sólidas aplicadas. Sugerem o uso geral das técnicas: máxima verossimilhança e imputação múltipla. Os pesquisadores Stef van Buuren e Karin Groothuis-Oudshoorn (2011) com o trabalho *MICE: Multivariete Imputation by Chained Equations R* disponibilizaram o pacote MICE, com um conjunto de modelos para tratar *missings*, para dados contínuos, binários, categóricos ordenados e não ordenados. Com vários gráficos de diagnóstico disponíveis para inspecionar a qualidade das imputações. No trabalho *Multiple imputation using chained equations: Issues and Guidance for Practice* dos pesquisadores White I. R., Royston P. e Wood A. M. (2011), descrevem o método de imputação múltipla no tratamento de dados ausentes, para variáveis quantitativas e categóricas. Exemplificam e discutem as limitações do método, por meio de uma análise de um conjunto de dados da área de saúde.

Ao longo dos anos, vários pesquisadores têm contribuído para avanços da área de MVI, desenvolvendo métodos e técnicas que abordam os desafios da imputação de forma eficaz. Com base nos autores apresentados na Tabela 9, em ordem cronológica, um dos primeiros trabalhos relevantes na imputação de dados ausentes foi o estudo dos pesquisadores Robins, Rotnitzky e Zhao (1994). Nesse trabalho, os autores discutem a estimativa de coeficientes de regressão quando alguns regressores estão ausentes. Essa pesquisa foi fundamental para a compreensão dos problemas enfrentados na imputação de dados ausentes em modelos de regressão. Em seguida, Rubin (1996) discute as princípios básicos da imputação múltipla e destaca sua importância na análise de dados incompletos. O pesquisador Todd K. Moon (1996) fez uso do algoritmo de *Expectation-Maximization* (EM) no processamento de sinais na presença de dados faltantes. No trabalho *Multiple Imputation: A Primer* de Schafer (1999) apresenta uma visão abrangente da imputação múltipla, abordando conceitos fundamentais e estratégias práticas para sua aplicação. Os pesquisadores Van Buuren, Boshuizen e Knoo (1999) aplicam a imputação múltipla, no contexto da análise da influência da pressão arterial na sobrevida em idosos, com dados faltantes decorrentes de mortalidade. Troyanskaya *et al.* (2001) abordaram a imputação de valores ausentes em microarranjos de DNA, uma técnica amplamente utilizada em genética e biologia

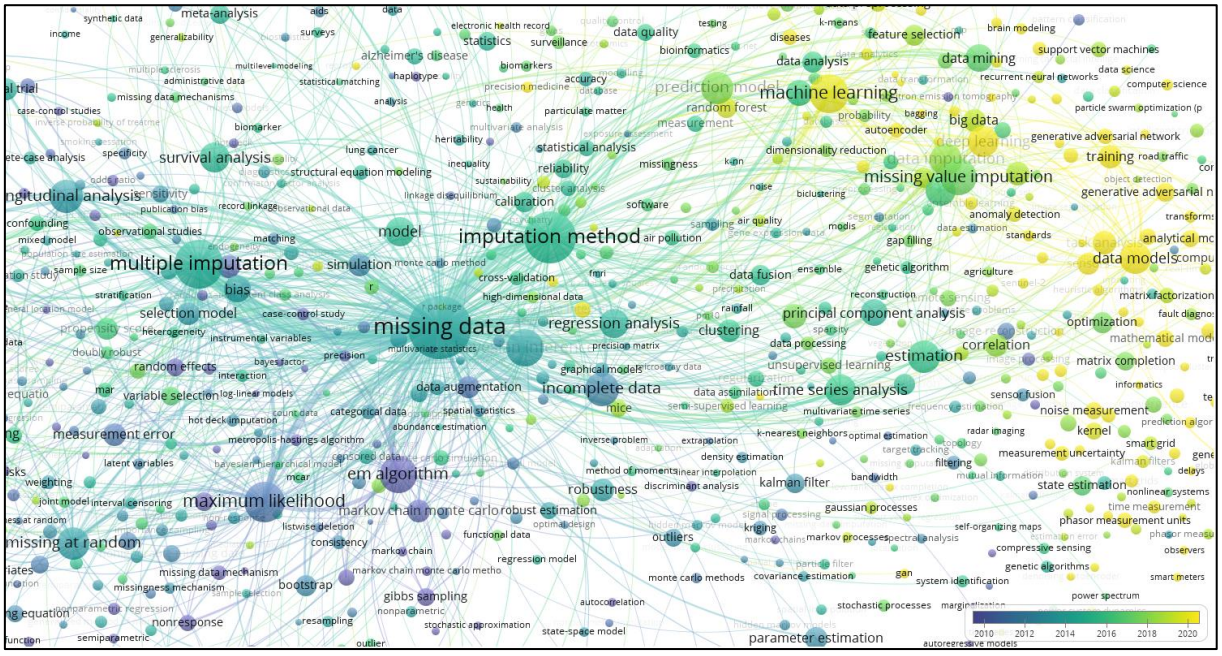
molecular. Collins, Schafer e Kam (2001) realizaram uma comparação entre estratégias inclusivas e restritivas em procedimentos de imputação de dados ausentes. Investigaram diferentes abordagens e suas vantagens e desvantagens em relação à imputação de dados ausentes. Royston e Division (2004) contribuíram para o desenvolvimento da imputação múltipla de valores ausentes, abordando questões-chave e fornecendo diretrizes práticas, discutindo os fundamentos teóricos e as técnicas aplicadas. Donders *et al.* (2006) realizaram uma revisão abrangente e acessível sobre a imputação de valores ausentes. Van Buuren (2007) apresentou um método de imputação múltipla de dados discretos e contínuos baseado em especificações condicionais. Graham, Olchowski e Gilreath (2007) forneceram esclarecimentos sobre a teoria da imputação múltipla, destacando a determinação do número apropriado de imputações necessárias para resultados confiáveis. Graham (2009) apresentou o trabalho *Missing Data Analysis: Making it Work in the Real World*, abordando os desafios práticos enfrentados na análise de dados ausentes, levando em consideração as limitações e as demandas do mundo real. Sterne *et al.* (2009) exploraram o potencial e as armadilhas da imputação múltipla em pesquisas epidemiológicas e clínicas, enfatizando a importância da imputação adequada de dados ausentes e fornecendo orientações sobre as melhores práticas. Azur *et al.* (2011) forneceram uma introdução detalhada à imputação múltipla por equações encadeadas, abrangendo a metodologia e exemplos concretos. Rhemtulla, Brosseau-Liard e Savalei (2012) investigaram as condições em que variáveis categóricas podem ser tratadas como contínuas na estimação de modelos de equações estruturais. E finalmente, Stekhoven e Bühlmann (2012) abordaram a imputação de dados ausentes para conjuntos de dados com variáveis de tipos mistos, combinação de variáveis contínuas e categóricas, apresentando o método *MissForest*.

Esses vinte artigos abrangem uma parte significativa dos temas relacionados à imputação de dados ausentes. Apresentam os fundamentos do campo MVI, fornecendo teorias, métodos e orientações práticas.

1.3.7 DISTRIBUIÇÃO POR PALAVRAS-CHAVE

Das palavras-chave dos autores, cerca de 38.783 palavras, as que ocorrem com maior frequência são distribuídas de acordo com a Figura 10, que mostra e caracterizada o início, das pesquisas em MVI, com o *Algoritmo EM* (DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977; MOON, 1996), passando por *Imputação Múltipla* (RUBIN, 1996; SCHAFER, 1999; STERNE et al., 2009; WHITE; ROYSTON; WOOD, 2011) e atingindo *Aprendizado de Máquina* (BZAI et al., 2022; CAI et al., 2017; HASAN et al., 2021; LIEBAL et al., 2020; NIJMAN et al., 2022; XUE et al., 2021).

Figura 10 – Distribuição temporal das palavras-chave dos autores

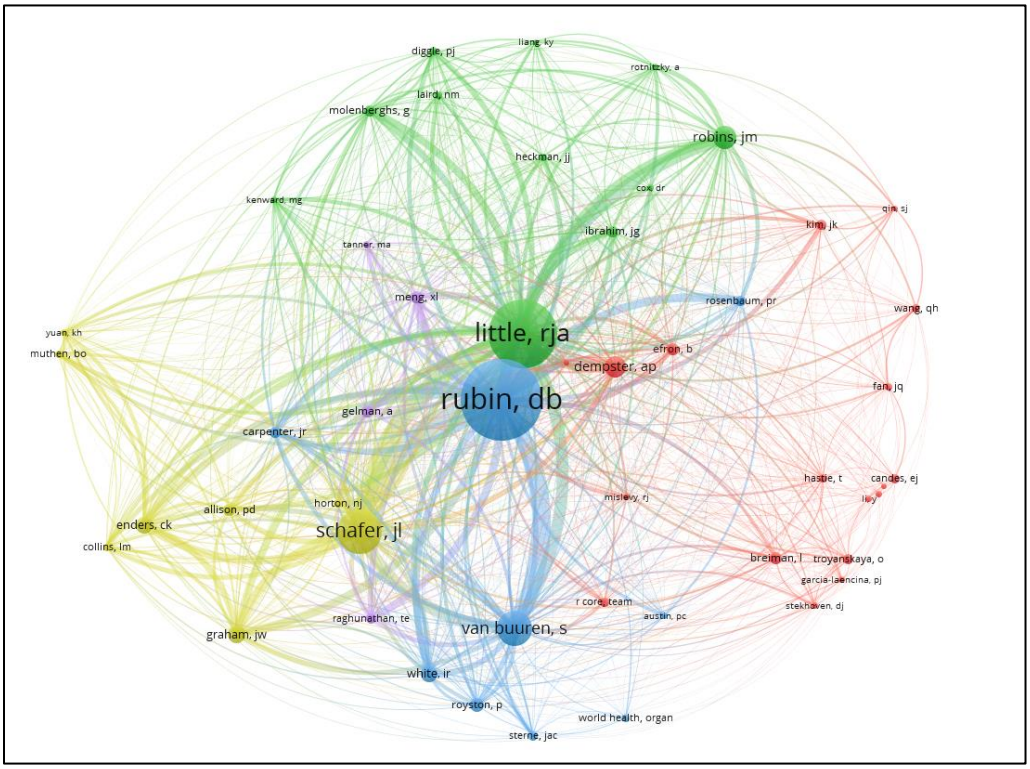


Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

1.3.8 COCITAÇÃO DE AUTORES E REFERÊNCIAS

Em torno de 450.673 referências, produzidos por 239.656 autores, foram utilizadas pelas 19.745 publicações identificadas, o que resulta em uma média de 12 autores e 22 referências por publicação. Os autores fundamentais, podem ser determinados pela análise de cocitação, a Figura 11 mostra a rede de colaboração com os primeiros cinquenta autores de maior impacto.

Figura 11 – Rede de colaboração entre os principais autores cocitados



Fonte: Autoria Própria a partir de dados WOS (VOSviewer).

Os primeiros dez autores com maiores números de cocitações são apresentados na Tabela 10.

Tabela 10 – Número de cocitações por autor

Rank	Autor	Número de Cocitações
1	RUBIN, DB	7798
2	LITTLE, RJA	6394
3	SCHAFER, J	3869
4	VAN BUUREN, S	2925
5	ROBINS, JM	1812
6	DEMPSTER, AP	1609
7	GRAHAM, JW	1270
8	WHITE, IR	1255
9	ENDERS, CK	1206
10	ROYSTON, P	1014

Fonte: Autoria Própria a partir de dados WOS.

Todos estes pesquisadores têm formação em estatística fizeram contribuições significativas ao tema MVI, sete são professores de instituições Norte Americanas, o que ratifica o fato de serem mais produtivas.

Professores na *Harvard University* os pesquisadores *Donald B. Rubin* (“Donald B. Rubin”, 2023), *James M. Robins* (“James M. Robins”, 2023). e *Arthur P. Dempster* (“Arthur P. Dempster”, 2023). Professor Robins conhecido por suas contribuições para a modelagem causal e a análise de dados longitudinais em presença de dados faltantes. Professor Dempster desenvolveu o método *Expectation-Maximization* (EM), amplamente utilizado para estimar parâmetros em presença de dados faltantes (DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977).

Pesquisador *Roderick Joseph Alexander Little* fez importantes contribuições para a modelagem estatística de dados faltantes e é coautor do livro *Statistical Analysis with Missing Data* (LITTLE; RUBIN, 1987), junto com Donald B. Rubin. Professor de bioestatística da *University of Michigan* (“Roderick J. A. Little”, 2023).

O professor *Joseph L. Schafer* (“Joseph L. Schafer”, 2023) da *Pennsylvania University* e o professor *John W. Graham* (“John W. Graham”, 2023) da *Pennsylvania State University* tem papel ativo em pesquisa e ensino da estatística, especialmente no campo da análise de dados faltantes. Publicaram o artigo *Missing data: our view of the State of the art* (SCHAFER; GRAHAM, 2002), *Missing data analysis: Making it work in the real world* (GRAHAM, 2009), o livro *Analysis of Incomplete Multivariate Data* (SCHAFER, 1997), referências fundamentais na área de MVI.

Da *University of California, Los Angele*s, UCLA, professor *Craig K. Enders* (“Craig Enders”, 2023) destaque por suas pesquisas em MVI. Publicou o livro *Applied Missing Data Analysis* (2010), apresenta uma abordagem prática e acessível para a análise de dados faltantes.

Os pesquisadores *Ian R. White* (“Ian R. White”, 2023) e *Patrick Royston* (“Patrick Royston”, 2023) professores na *University College London*. Publicaram o artigo *Multiple Imputation Using Chained Equations: Issues and Guidance for Practice* (WHITE; ROYSTON; WOOD, 2011), referência nas pesquisas de imputação múltipla por equações encadeadas com dados faltantes.

E finalmente o professor *Stef van Buuren* (“Stef van Buuren”, 2023) da *University of Utrecht*, Holanda. Desenvolveu o pacote *MICE* em *R* (2011), que implementa a técnica de imputação múltipla por equações encadeadas para lidar com dados faltantes.

A Tabela 11 apresenta as principais referências, fundamentação teórica da área de MVI. Destaque para *JM Rubin*, *RJA Little* e *JL Schafer* pesquisadores de maior impacto e cocitados das publicações identificadas e analisadas.

Tabela 11 – Principais referências cocitadas

Rank	Título	Autor	Cocitações
1	*INFERENCE AND MISSING DATA*	(RUBIN, 1976)	2000
2	*MULTIPLE IMPUTATION FOR NONRESPONSE IN SURVEYS*	(RUBIN, 1987)	1819
3	*STATISTICAL ANALYSIS WITH MISSING DATA*	(LITTLE; RUBIN, 2019)	1624
4	*MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA THE EM ALGORITHM*	(DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977)	1493
5	*ANALYSIS OF INCOMPLETE MULTIVARIATE DATA*	(SCHAFER, 1997)	1286
6	*STATISTICAL ANALYSIS WITH MISSING DATA*	(LITTLE; RUBIN, 1987)	1179
7	*MISSING DATA: OUR VIEW OF THE STATE OF THE ART*	(SCHAFER; GRAHAM, 2002)	1039
8	*MICE: MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS IN R*	(VAN BUUREN; GROOTHUIS-OUDSHOORN, 2011)	1018
9	*MULTIPLE IMPUTATION USING CHAINED EQUATIONS: ISSUES AND GUIDANCE FOR PRACTICE*	(WHITE; ROYSTON; WOOD, 2011)	707
10	*MISSING VALUE ESTIMATION METHODS FOR DNA MICROARRAYS*	(TROYANSKAYA et al., 2001)	704
11	*MULTIPLE IMPUTATION AFTER 18+ YEARS*	(RUBIN, 1996)	671
12	*ESTIMATION OF REGRESSION COEFFICIENTS WHEN SOME REGRESSORS ARE NOT ALWAYS OBSERVED*	(ROBINS; ROTNITZKY; ZHAO, 1994)	516
13	*A COMPARISON OF INCLUSIVE AND RESTRICTIVE STRATEGIES IN MODERN MISSING DATA PROCEDURES*	(COLLINS; SCHAFER; KAM, 2001)	467
14	*MULTIPLE IMPUTATION: A PRIMER*	(SCHAFER, 1999)	449

(continua)

(continuação)

Rank	Título	Autor	Cocitações
15	*APPLIED MISSING DATA ANALYSIS*	(ENDERS, 2010)	412
16	*A MULTIVARIATE TECHNIQUE FOR MULTIPLY IMPUTING MISSING VALUES USING A SEQUENCE OF REGRESSION MODELS*	(RAGHUNATHAN et al., 2001)	408
17	*MULTIPLE IMPUTATION OF MISSING BLOOD PRESSURE COVARIATES IN SURVIVAL ANALYSIS*	(VAN BUUREN; BOSHUIZEN; KNOOK, 1999)	403
18	*MULTIPLE IMPUTATION OF DISCRETE AND CONTINUOUS DATA BY FULLY CONDITIONAL SPECIFICATION*	(VAN BUUREN, 2007)	391
19	*MISSING DATA ANALYSIS: MAKING IT WORK IN THE REAL WORLD*	(GRAHAM, 2009)	389
20	*MISSFOREST—NON-PARAMETRIC MISSING VALUE IMPUTATION FOR MIXED-TYPE DATA*	(STEKHOVEN; BÜHLMANN, 2012)	380

Fonte: Autoria Própria a partir de dados WOS.

Considerando a análise de cocitações feita nesta pesquisa, chega-se aos dez principais autores que formam os pilares da área de MVI, designados pelos pesquisadores: *Donald B. Rubin, Roderick Joseph Alexander Little, Arthur P. Dempster, Joseph L. Schafer, John W. Graham, Stef van Buuren, Ian R. White, Patrick Royston, James M. Robins e Craig K. Enders.*

1.4 CONCLUSÕES

Neste capítulo foi apresentada uma revisão das pesquisas mundiais no tema *Missing Value Imputation*, MVI. Para alcançar este objetivo foi realizado uma análise bibliométrica, utilizando a base dados *Web Of Science*, considerando o período de 1940 a 2022, foram identificadas em 2 de janeiro de 2023, uma amostra de 19.745 trabalhos.

Neste levantamento com 19.745 artigos, observou-se que 70% foram produzidos na última década. Estas publicações envolveram 158 países, com protagonismo dos Estados Unidos (7.372), seguidos pela China (2.874) e Inglaterra (2044). Dos autores com maior número de trabalhos, ocorreu uma predominância de pesquisadores Chineses, para as referências consideradas. Na décima nona posição vem o Brasil com 290 publicações. As publicações estão distribuídas em

6.225 fontes científicas, destaque para os periódicos: *Statistics in Medicine, PLOS ONE, Biometrics, Journal of the American Statistical Association, Computational Statistics & Data Analysis, Communications in Statistics-Theory and Methods, Statistical Methods in Medical Research, IEEE Access, BMC Medical Research Methodology e Statistica Sinica.* Estiveram envolvidos nestas publicações cerca de 62.193 autores, gerando 450.673 referências citadas, destaque para os pesquisadores fundamentais: *Donald B. Rubin, Roderick Joseph Alexander Little, Arthur P. Dempster, Joseph L. Schafer, John W. Graham, Stef van Buuren, Ian R. White, Patrick Royston, James M. Robins e Craig K. Enders.*

As instituições Norte Americanas liberam o ranking das instituições com maior número de publicações em MVI, destaque para *Harvard University* (319), *University of Michigan* (311), *University of Washington* (299), *University of North Carolina* (253) e *University of California* (207). Seguida das universidades do Reino Unido: *University College London* (262), *University of Oxford* (203), *London School of Hygiene & Tropical Medicine* (194). Entre as dez, ainda tem uma representante da Austrália, *University of Melbourne* (183). E a mais jovem, a proeminente *Chinese Academy of Sciences* (281). No Brasil, destaque para as Universidades Paulistas: *USP* (73), *UNICAMP* (28) e *UNESP* (18).

A distribuição temporal da ocorrência das palavras-chave, caracterizada pelo início com o *Algoritmo EM*, passando por *Imputação Múltipla* e atingindo *Aprendizado de Máquina*, técnicas de Inteligência Artificial.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

ARAÚJO, C. A. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**, v. 12, n. 1, p. 11–32, 10 dez. 2006.

ARIA, M.; CUCCURULLO, C. bibliometrix: An R-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959–975, 1 nov. 2017.

Arthur P. Dempster. Disponível em: <https://statistics.fas.harvard.edu/people/arthur-p-dempster>. Acesso em: 8 jun. 2023.

AZUR, M. J. et al. Multiple imputation by chained equations: what is it and how does it work? **International Journal of Methods in Psychiatric Research**, v. 20, n. 1, p. 40–49, mar. 2011.

Biometrics. Disponível em: <https://onlinelibrary.wiley.com/page/journal/15410420/homepage/productinformation.html>. Acesso em: 8 jun. 2023.

BMC Medical Research Methodology. Disponível em: <https://bmcmedresmethodol.biomedcentral.com/about>. Acesso em: 8 jun. 2023.

BRADFORD, S. C. Sources of Information on Scientific Subjects. **Engineering: An Illustrated Weekly Journal**, n. 137, p. 85–86, 1934.

BZAI, J. et al. Machine Learning-Enabled Internet of Things (IoT): Data, Applications, and Industry Perspective. **Electronics 2022, Vol. 11, Page 2676**, v. 11, n. 17, p. 2676, 26 ago. 2022.

CAI, G. et al. A Review on Micromixers. **Micromachines 2017, Vol. 8, Page 274**, v. 8, n. 9, p. 274, 11 set. 2017.

CALANCA, P. Weather Forecasting Applications in Agriculture. **Encyclopedia of Agriculture and Food Systems**, p. 437–449, 1 jan. 2014.

CAPES. **Plataforma Sucupira**. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.xhtml>. Acesso em: 4 jan. 2023.

Chinese Academy of Sciences. Disponível em: <https://pt.wikipedia.org/wiki/Academia_Chinesa_de_Ciências>. Acesso em: 29 maio. 2023.

CLARIVATE. **Plataforma Web of Science**. Disponível em: <https://clarivate.com/webofsciencegroup/solutions/webofscience-platform/>. Acesso em: 14 nov. 2022a.

CLARIVATE. **Web of Science™ base de dados de citação global independente mais confiável do mundo - Web of Science Group**. Disponível em: <https://clarivate.com/webofsciencegroup/campaigns/web-of-science-base-de-dados-de-citacao-global-independente-mais-confiavel-do-mundo/>. Acesso em: 14 nov. 2022b.

COLLINS, L. M.; SCHAFER, J. L.; KAM, C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. **Psychological Methods**, v. 6, n. 4, p. 330–351, 2001.

Communications in Statistics - Theory and Methods. Disponível em: <https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=lsta20>. Acesso em: 8 jun. 2023.

Computational Statistics & Data Analysis. Disponível em: <https://www.sciencedirect.com/journal/computational-statistics-and-data-analysis/about/aims-and-scope>. Acesso em: 8 jun. 2023.

CORNISH, E. A. The estimation of missing values in incomplete randomized block experiments. **Annals of Eugenics**, v. 10, p. 112–118, 1940a.

CORNISH, E. A. The estimation of missing values in quasi-factorial designs. **Annals of Eugenics**, v. 10, p. 137–143, 1940b.

Craig Enders. Disponível em: <https://www.psych.ucla.edu/faculty-page/cenders/>. Acesso em: 9 jun. 2023.

DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal ofthe Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977.

Donald B. Rubin. Disponível em: <https://statistics.fas.harvard.edu/people/donald-b-rubin>. Acesso em: 8 jun. 2023.

DONDERS, A. R. T. et al. Review: A gentle introduction to imputation of missing values. **Journal of Clinical Epidemiology**, v. 59, n. 10, p. 1087–1091, out. 2006.

ENDERS, C. K. A Primer on maximum likelihood algorithms available for use with missing data. **Structural Equation Modeling**, v. 8, n. 1, p. 128–141, 2001.

ENDERS, C. K. **Applied missing data analysis**. New York: THE GUILFORD PRESS, 2010.

GRÁCIO, M. C. C. Acoplamento bibliográfico e análise de cocitação: revisão teórico-conceitual. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 21, n. 47, p. 82, 12 set. 2016.

GRAHAM, J. W. Missing data analysis: making it work in the real world. **Annual review of psychology**, v. 60, p. 549–576, jan. 2009.

GRAHAM, J. W.; OLCHOWSKI, A. E.; GILREATH, T. D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. **Prevention Science**, v. 8, n. 3, p. 206–213, set. 2007.

Harvard University. Disponível em: <https://pt.wikipedia.org/wiki/Universidade_Harvard#cite_ref-10>. Acesso em: 24 dez. 2022.

HASAN, M. K. et al. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). **Informatics in Medicine Unlocked**, v. 27, p. 1–23, 1 jan. 2021.

HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology**, v. 103, n. 1–2, p. 137–157, 1 jun. 2000.

Ian R. White. Disponível em: <https://www.mrcctu.ucl.ac.uk/about-us/our-staff/ian-white/>. Acesso em: 8 jun. 2023.

IEEE Access. Disponível em: <https://ieeeaccess.ieee.org/about-ieee-access/learn-more-about-ieee-access/>. Acesso em: 8 jun. 2023.

James M. Robins. Disponível em: <https://www.hsph.harvard.edu/profile/james-m-robins/>. Acesso em: 8 jun. 2023.

John W. Graham. Disponível em: <https://scholar.google.com/citations?user=Oiw7kDwAAAAJ&hl=en>. Acesso em: 8 jun. 2023.

Joseph L. Schafer. Disponível em: <https://academictree.org/math/publications.php?pid=300624>. Acesso em: 9 jun. 2023.

Journal of the American Statistical Association. Disponível em: <https://en.wikipedia.org/wiki/Journal_of_the_American_Statistical_Association>. Acesso em: 24 dez. 2022.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, 1 jun. 2004.

KELLY, D. et al. **Guide to China**. Disponível em: <https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/market-insights/guide-to-china.pdf>. Acesso em: 24 dez. 2022.

LIEBAL, U. W. et al. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. **Metabolites**, v. 10, n. 6, p. 1–23, 1 jun. 2020.

LIN, W.-C.; TSAI, C.-F. Missing value imputation: a review and analysis of the literature (2006–2017). **Artificial Intelligence Review**, v. 53, p. 1487–1509, 2020.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. 1st Editio ed. New York: [s.n.].

LITTLE, R. J. A.; RUBIN, D. B. **Statistical analysis with missing data**. [s.l: s.n.].

LOTKA, A. J. The frequency distribution of scientific productivity. **ournal of the Washington Academy of Sciences**, v. 16, n. 12, p. 317–323, 1926.

LSHTM. **Escola de Higiene de Londres e Medicina Tropical**. Disponível em: <https://www.lshtm.ac.uk/>. Acesso em: 29 maio. 2023.

MOON, T. K. The expectatio maximization algorithm. **IEEE SIGNAL PROCESSING MAGAZINE**, p. 47–60, 1996.

MORESI, E. A. D.; PINHO, I.; COSTA, A. P. **BIBLIOMETRIC ANALYSIS: A QUANTITATIVE AND QUALITATIVE APPROACH**. 18th CONTECSI – INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT. **Anais**...São Paulo: out. 2021

NIJMAN, S. W. J. et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. **Journal of clinical epidemiology**, v. 142, p. 218–229, 1 fev. 2022.

OSBORNE, J. W. **Best Practices in Data Cleaning: a complete guide to everything you need to do before and after collecting your data**. Los Angeles: SAGE Publications Inc., 2013.

OTLET, P. **Traité de Documentation - Le Livre sur le Livre - Théorie et Pratique**. Bruxelles: Editiones Mundaneum, 1934.

Patrick Royston. Disponível em: <https://scholar.google.com/citations?user=cRJ-qjUAAAAJ&hl=en>. Acesso em: 8 jun. 2023.

PLOS One. Disponível em: <https://journals.plos.org/plosone/s/journal-information>. Acesso em: 8 jun. 2023.

PRITCHARD, A. Statistical Bibliography or Bibliometrics. **Journal of Documentation**, n. 25, p. 348–349, 1969.

RAGHUNATHAN, T. E. et al. A Multivariate technique for multiply imputing missing values using a sequence of regression models. **Survey Methodology**, v. 27, n. 1, p. 85–95, 2001.

RHEMTULLA, M.; BROSSEAU-LIARD, P. É.; SAVALEI, V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. **Psychological Methods**, v. 17, n. 3, p. 354, set. 2012.

ROBINS, J. M.; ROTNITZKY, A.; ZHAO, L. P. Estimation of regression coefficients when some regressors are not always observed. **Journal of the American Statistical Association**, v. 89, n. 427, p. 846–866, 1994.

Roderick J. A. Little. Disponível em: <https://en.wikipedia.org/wiki/Roderick_J._A._Little>. Acesso em: 8 jun. 2023.

ROTH, P. L. MISSING DATA: A CONCEPTUAL REVIEW FOR APPLIED PSYCHOLOGISTS. **Personnel Psychology**, v. 47, n. 3, p. 537–560, 1 set. 1994.

ROYSTON, P.; DIVISION, C. Multiple Imputation of Missing Values. **The Stata Journal**, v. 4, n. 3, p. 227–241, 1 ago. 2004.

RUBIN, D. B. Inference and Missing Data. **Biometrika**, v. 63, n. 3, p. 581–592, 1976.

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1987.

RUBIN, D. B. **Multiple Imputation after 18+ YearsJournal of the American Statistical Association**, 1996.

SCHAFER, J. L. **Analysis of Incomplete Multivariate Data**. New York: Chapman & Hall/CRC, 1997.

SCHAFER, J. L. Multiple imputation: a primer. **Statistical Methods in Medical Research**, v. 8, n. 1, p. 3–15, 2 fev. 1999.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: Our view of the state of the art. **Psychological Methods**, v. 7, n. 2, p. 147–177, 2002.

SMALL, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. **Journal of the American Society for Information Science**, v. 24, n. 4, p. 265–269, 1 jul. 1973.

Statistica Sinica. Disponível em: <https://www3.stat.sinica.edu.tw/statistica/>. Acesso em: 8 jun. 2023.

Statistical Methods in Medical Research. Disponível em: <https://journals.sagepub.com/description/SMM>. Acesso em: 8 jun. 2023.

Statistics in Medicine. Disponível em: <https://onlinelibrary.wiley.com/page/journal/10970258/homepage/productinformation.html>. Acesso em: 8 jun. 2023.

Stef van Buuren. Disponível em: <https://stefvanbuuren.name/>. Acesso em: 8 jun. 2023.

STEKHOVEN, D. J.; BÜHLMANN, P. MissForest—non-parametric missing value imputation for mixed-type data. **Bioinformatics**, v. 28, n. 1, p. 112–118, 1 jan. 2012.

STERNE, J. A. C. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. **BMJ**, v. 338, n. 7713, p. 157–160, 29 jun. 2009.

STIGTER, C. J. From basic agrometeorological science to agrometeorological services and information for agricultural decision makers: A simple conceptual and diagnostic framework. **Agricultural and Forest Meteorology**, v. 142, n. 2–4, p. 91–95, 12 fev. 2007.

STRIKE, K.; EMAM, K. EL; MADHAVJI, N. Software cost estimation with incomplete data. **IEEE Transactions on Software Engineering**, v. 27, n. 10, p. 890–908, 2001.

TAKLE, E. S. AGRICULTURAL METEOROLOGY AND CLIMATOLOGY. **Encyclopedia of Atmospheric Sciences**, p. 54–60, 1 jan. 2003.

TAKLE, E. S. Agricultural Meteorology and Climatology. **Encyclopedia of Atmospheric Sciences: Second Edition**, p. 92–97, 1 jan. 2015.

TROYANSKAYA, O. et al. Missing value estimation methods for DNA microarrays. **Bioinformatics**, v. 17, n. 6, p. 520–525, 1 jun. 2001.

University College London. Disponível em: <https://pt.wikipedia.org/wiki/University_College_London>. Acesso em: 29 maio. 2023.

University of California, Los Angeles. Disponível em: <https://www.britannica.com/topic/University-of-California-Los-Angeles>. Acesso em: 29 maio. 2023.

University of Melbourne. Disponível em: <https://www.britannica.com/topic/University-of-Melbourne>. Acesso em: 29 maio. 2023.

University of Michigan. Disponível em: <https://pt.wikipedia.org/wiki/Universidade_de_Michigan>. Acesso em: 24 dez. 2022.

University of North Carolina. Disponível em: <https://pt.wikipedia.org/wiki/Universidade_da_Carolina_do_Norte>. Acesso em: 29 maio. 2023.

University of Oxford. Disponível em: <https://www.britannica.com/topic/University-of-Oxford>. Acesso em: 29 maio. 2023.

University of Pennsylvania. Disponível em: <https://pt.wikipedia.org/wiki/Universidade_da_Pensilvânia>. Acesso em: 24 dez. 2022.

VAN BUUREN, S. Multiple imputation of discrete and continuous data by fully conditional specification. **Statistical Methods in Medical Research**, v. 16, p. 219–242, dez. 2007.

VAN BUUREN, S.; BOSHUIZEN, H. C.; KNOOK, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. **Statistics in Medicine**, v. 18, n. 6, p. 681–694, 1999.

VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. mice: Multivariate imputation by chained equations in R. **Journal of Statistical Software**, v. 45, n. 3, p. 1–67, 2011.

VAN ECK, N. J.; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, v. 84, n. 2, p. 523–538, 2010.

VAN ECK, N. J.; WALTMAN, L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. **Scientometrics**, v. 111, n. 2, p. 1053–1070, 1 maio 2017.

VELASCO-MUÑOZ, J. F. et al. Advances in water use efficiency in agriculture: A bibliometric analysis. **Water (Switzerland)**, v. 10, n. 4, 2018.

VOSVIEWER. **Visualizzing scientific landscapes**. Disponível em: <https://www.vosviewer.com/>. Acesso em: 14 nov. 2022.

VOSVIEWER. **VOSviewer :: Download**. Disponível em: <https://www.vosviewer.com/download>. Acesso em: 16 jan. 2023.

WALTMAN, L.; VAN ECK, N. J. A new methodology for constructing a publication-level classification system of science. **Journal of the American Society for Information Science and Technology**, v. 63, n. 12, p. 2378–2392, 1 dez. 2012.

WALTMAN, L.; VAN ECK, N. J. A smart local moving algorithm for large-scale modularity-based community detection. **The European Physical Journal B 2013 86:11**, v. 86, n. 11, p. 1–14, 13 nov. 2013.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. **Statistics in Medicine**, v. 30, n. 4, p. 377–399, 2011.

WOLSKI, L. Z. et al. MINERAÇÃO DE TEXTO E CLUSTERIZAÇÃO EM ESTUDOS BIBLIOMÉTRICOS:O MAPEAMENTO CIENTÍFICO DE TESES E DISSERTAÇÕES DE UM PROGRAMA DE PÓS-GRADUAÇÃO. **Anais do Congresso Internacional de Conhecimento e Inovação – ciki**, v. 1, n. 1, 16 mar. 2021.

XUE, Y. et al. Multi-objective Feature Selection with Missing Data in Classification. **IEEE Transactions on Emerging Topics in Computational Intelligence**, v. 6, n. 2, p. 355–364, 18 abr. 2021.

YOZGATLIGIL, C. et al. Comparison of missing value imputation methods in time series: The case of Turkish meteorological data. **Theoretical and Applied Climatology**, v. 112, n. 1–2, p. 143–167, 2013.

ZIPF, G. K. **Human behavior and the principle of least effort**. Cambridge, Massachusetts: Addison-Wesley, 1949.

ZUPIC, I.; ČATER, T. Bibliometric Methods in Management and Organization. **Organizational Research Methods**, v. 18, n. 3, p. 429–472, 2015.

CAPÍTULO 2

PLANEJAMENTO DE SIMULAÇÃO PARA IMPUTAÇÃO DE VALOR AUSENTE[1]

SIMULATION PLANNING FOR MISSING VALUE IMPUTATION

Valter Cesar de Souza[$&*], **Sérgio Augusto Rodrigues**[$&]

[$]São Paulo State University (Unesp), School of Agriculture, Botucatu, São Paulo, Brasil

* Corresponding author
E-mail: valter.souza@unesp.br

[&]These authors contributed equally to this work.

[1] Capítulo redigido de acordo com as normas do periódico PLOS ONE.

RESUMO

A agrometeorologia estuda as relações entre o clima e a atividade agrícola, exigindo informações meteorológicas precisas, confiáveis e de alta qualidade. Em particular, estudos sobre evapotranspiração são importantes para o planejamento e gestão de sistemas de irrigação, bem como para trabalhos hidrológicos e ambientais. No entanto, é comum encontrar uma proporção de dados faltantes em grandes conjuntos de dados coletados em estações meteorológicas. Os dados podem estar faltando devido a diversos motivos, como falha de equipamentos, entrada inadequada de dados, indisponibilidade de informações ou problemas durante a coleta de dados. Se os valores ausentes não forem tratados de forma adequada, podem resultar em perda de eficiência, viés devido a diferenças entre dados faltantes e completos, e complicações na análise e interpretação dos dados. Uma solução comum para lidar com dados incompletos é a imputação de valores ausentes, conhecida como *Missing Value Imputation* (MVI), em vez da remoção dos dados faltantes. Compreender como os procedimentos de MVI funcionam em aplicações e locais específicos é um desafio importante a ser superado. Neste contexto, o objetivo deste capítulo foi propor um planejamento de simulação em MVI, com foco no banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho, preparando conceitos para a aplicação apresentada no capítulo três: Comparação de algoritmos de análise de componentes principais para imputação em dados agrometeorológicos em alta dimensão e tamanho amostral reduzido.

Palavras-chave: planejamento de simulação; agrometeorologia; evapotranspiração; qualidade de dados.

ABSTRACT

Agrometeorology studies the relationships between climate and agricultural activity, requiring accurate, reliable, and high-quality meteorological information. In particular, studies on evapotranspiration are important for the planning and management of irrigation systems, as well as for hydrological and environmental work. However, it is common to find a proportion of missing data in large datasets collected from weather stations. Data may be missing due to a variety of reasons, such as equipment failure, improper data entry, unavailability of information, or problems during data collection. If missing values are not handled properly, they can result in loss of efficiency, bias due to differences between missing and complete data, and complications in data analysis and interpretation. A common solution to dealing with incomplete data is to impute missing values, known as Missing Value Imputation (MVI), instead of removing the missing data. Understanding how MVI procedures work in specific applications and locations is an important challenge to overcome. In this context, the objective of this chapter was to propose a simulation planning in MVI, focusing on the database and type of data, mechanism and failure rate, imputation technique and performance evaluation method, preparing concepts for the application presented in chapter three: Comparison of principal component analysis algorithms for imputation in agrometeorological data in high dimension and reduced sample size.

Keywords: simulation planning; agrometeorology; evapotranspiration; data quality.

2.1 INTRODUÇÃO

Dados meteorológicos com acurácia, qualidade e confiabilidade, são exigidos nas pesquisas de agrometeorologia, especialmente nos estudos de evapotranspiração, a qual revela-se importante componente no ciclo hidrológico (KISI et al., 2021; MARTÍ; GASQUE, 2010), no planejamento (GONSAGA DE CARVALHO et al., 2011; WANG et al., 2022) e gestão de sistemas de irrigação (MARIN et al., 2019; MARTÍ; ZARZO, 2012), na modelagem da demanda de água (TERINK; IMMERZEEL; DROOGERS, 2013), no monitoramento do estresse hídrico (HART et al., 2009), na estimativa do balanço hídrico (CAI et al., 2009). É comum em bases de dados obtidas em estações meteorológicas observar uma proporção de dados ausentes (WHITE; ROYSTON; WOOD, 2011), *missings*, que devem ser tratados.

A distribuição da ocorrência dos *missings* em uma base dados pode estar relacionada a um dos mecanismos descritos por Little e Rubin (2002), em: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Missing Not at Random* (MNAR). Os valores ausentes em bases de dados, se não tratados adequadamente, podem levar a problemas de perda de eficiência, viés resultante de diferenças entre dados faltantes e completos, complicações no manuseio e nos resultados das análises dos dados. (FARHANGFAR; KURGAN; PEDRYCZ, 2007).

A imputação de valor ausente, *missing value imputation* (MVI), é uma abordagem utilizada para tratar o problema de conjunto de dados incompletos, ao invés da remoção dos dados ausentes do conjunto. Entre os métodos de MVI, observa-se aqueles que utilizam procedimentos estatísticos e os de *machine learning*. Dos procedimentos estatísticos, destaque para o procedimento multivariado Análise de Componentes Principais (GARCÍA-DIEGO; ZARZO, 2010; JOSSE; HUSSON, 2012a; MARTÍ; ZARZO, 2012) em conjunto com os algoritmos *Nonlinear Iterative Partial Least Squares* (WRIGHT, 2017) e *Expectation Maximization* (DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977).

No entanto, entender o desempenho dos procedimentos de MVI em aplicações e locais específicos apresenta-se como um importante desafio. A performance de um método MVI, pode ser verificada por meio de um processo de simulações, que levam em consideração o domínio do banco de dados, o tipo de

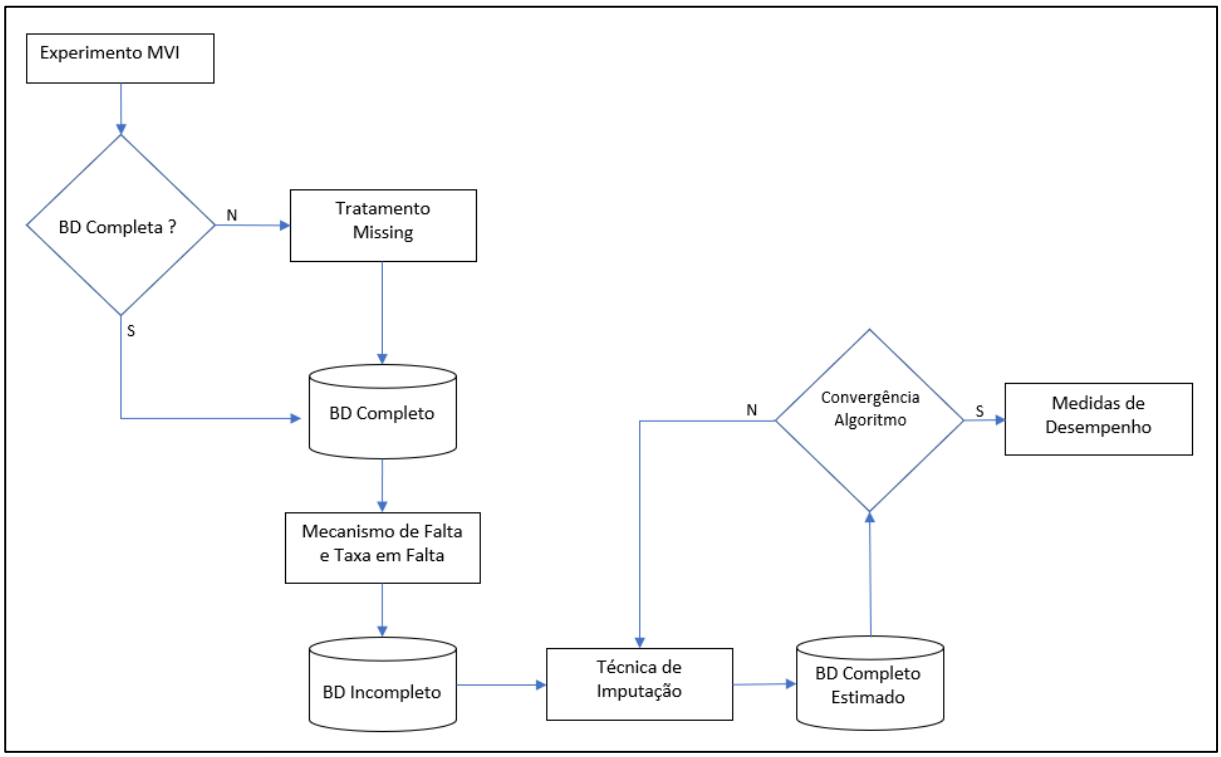
dados, o mecanismo e a taxa de falta, a técnica de imputação e o método de avaliação de desempenho.

Neste contexto, o objetivo deste capítulo foi propor um planejamento de simulação em MVI, com foco no banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho, preparando conceitos para a aplicação apresentada no capítulo três: *Comparação de algoritmos de análise de componentes principais para imputação em dados agrometeorológicos em alta dimensão e tamanho amostral reduzido*.

2.2 PLANEJAMENTO DE SIMULAÇÃO

Existem algumas definições e etapas importantes ao se planejar um estudo de simulação em MVI, que são: banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho. Normalmente os estudos de simulação visam verificar a performance dos métodos de imputação, como esquematizado na Figura 1, considerando as interações entre tipo de aplicação, tipos de dados, mecanismos e taxa de falta.

Figura 1 – Etapas de um experimento MVI



Fonte: Autoria Própria.

2.3 BANCO DE DADOS

Bancos de dados de diversas áreas estão disponíveis para estudos em domínios públicos na *internet*, entre eles os domínios de dados médicos (CIOS; WILLIAM MOORE, 2002) (HARPER, 2005), de imagem, medição e projeto de software (KHOSHGOFTAAR; VAN HULSE, 2008), dados financeiros, dados baseados em questionários, dados aeroespaciais (TIAN et al., 2014), dados industriais (LAKSHMINARAYAN; HARP; SAMAD, 1999). A Universidade da Califórnia em Irvine apresenta um considerável repositório de bancos de dados, centenas de conjuntos de dados nos mais variados domínios (DUA, D. E GRAFF, 2019). Em uma aplicação, considerar vários domínios de dados, tem a vantagem de mostrar a escalabilidade de domínio de um método MVI. Os dados podem ser do tipo: categóricos (SCHAFER, 1997) (RAHMAN; ISLAM, 2013), discretos, contínuos (SCHAFER, 1997) (TROYANSKAYA et al., 2001) (STÄDLER; STEKHOVEN; BÜHLMANN, 2014) ou mistos (STEKHOVEN; BÜHLMANN, 2012) (LIAO et al., 2014) (ZHU; HE; LIATSIS, 2012).

Na área de agrometeorologia, existem domínios de fontes, por exemplo, os portais *HidroWeb*, *INMET*, *CPTEC*, *BDMEP*. O portal HidroWeb[2] é uma ferramenta integrante do Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH) e oferece o acesso ao banco de dados que contém todas as informações coletadas pela Rede Hidrometeorológica Nacional (RHN), reunindo dados de níveis fluviais, vazões, chuvas, climatologia, qualidade da água e sedimentos (SNIRH, 2022). O Instituto Nacional de Meteorologia[3] disponibiliza dados de precipitação, temperatura e outros parâmetros meteorológicos de estações automáticas, analógicas e de radiossonda (INMET, 2022). O Centro de Previsão de Tempo e Estudos Climáticos[4], do Instituto Nacional de Pesquisas Espaciais, disponibiliza em diversos formatos de dados meteorológicos regionais (CPTEC, 2022). O Banco de Dados Meteorológicos do INMET[5] (BDMEP) abriga dados meteorológicos diários em forma digital, de séries

[2] https://www.snirh.gov.br/hidroweb/

[3] https://portal.inmet.gov.br/

[4] https://www.cptec.inpe.br/

[5] https://portal.inmet.gov.br/servicos/bdmep-dados-historicos/

históricas das várias estações meteorológicas da rede de estações do INMET com milhões de informações, referentes às medições diárias, de acordo com as normas técnicas internacionais da Organização Meteorológica Mundial (BDMEP, 2022).

2.3.1 EXTRAÇÃO DADOS METEOROLÓGICOS DO INMET

Para baixar dados meteorológicos em séries históricas do INMET, várias etapas são necessárias:

1. Entrar no site do INMET: *https://bdmep.inmet.gov.br/*
2. Escolher a opção pacote de dados anuais de todas as estações automáticas separadas por ano, sendo remetido a página para dados históricos anuais;
3. Escolher os anos de interesse, entre os quais estão disponíveis dados a partir do ano 2000. Para cada ano selecionado, um arquivo no formato *.csv*[6] estará disponível para cada estação;
4. Escolher as estações de interesse da pesquisa em particular. Para escolher as estações de interesse, visualizar a distribuição geográfica das estações no mapa de estações no link: *https://mapas.inmet.gov.br/* ;
5. Renomear todos os arquivos (*.csv*), esta ação pode ser feita de maneira manual ou automática. Preferível a forma automática, devido ao número de arquivos a serem manipuladas por uma rotina de tratamento dos dados, exemplificando, para uma escolha de 45 estações para um período de 10 anos, tem-se 450 arquivos. Para junção e um tratamento automático dos dados contidos nestes arquivos, ocorre a necessidade de padronizar os nomes. Passando de um nome completo, por exemplo, *INMET_SE_SP_**A725**_AVARE_01-01-2011_A_31-12-**2011*** para um nome reduzido ***A725_2011***;
6. Para facilitar a rotina de leitura destes arquivos, criar uma pasta para cada estação com os arquivos de planilhas referentes aos anos de interesse;
7. Criar rotina por meio de um script[7] no ambiente ***R*** que faça a leitura automática dos arquivos de dados obtidos, considerando os seguintes passos:

---

[6] *Comma Separated Values*

[7] O *script* disponível por meio de contato via e-mail: valter.souza@unesp.br

- Fazer a leitura dos arquivos com os dados de cada estação de todos os anos da pesquisa;
- Excluir as 9 primeiras linhas, pois são informações relativas as estações meteorológicas da fonte de dados da pesquisa (INMET);
- Colocar nomes de colunas comuns para todas as bases de dados lidas no ambiente ***R***;
- Substituir todos os valores “-9999” por “NA”;
- Converter a data-hora (fuso horário de Greenwich) para o horário local, com um ajuste específico para São Paulo, subtraindo três horas.
- Criar um banco de dados agregando todos os arquivos;
- Recalcular as variáveis de interesse para uma base diária.

2.3.2 EVAPOTRANSPIRAÇÃO DE REFERÊNCIA

A evapotranspiração ocorre quando os processos de evaporação e transpiração da água acontecem simultaneamente (PEREIRA et al., 1999). A evaporação é um processo natural que ocorre quando a água líquida é convertida em vapor de água e removida da superfície de evaporação, como o solo ou a vegetação úmida (MCILROY, 1984). Por sua vez, a transpiração ocorre pela vaporização da água líquida contida nos tecidos vegetais e na remoção de vapor para a atmosfera (KOOL et al., 2014). Ambos os processos ocorrem por meio da mudança de estado da água, passando de líquida para vapor, o que consume energia, fornecida pela radiação solar direta e temperatura ambiente do ar, pelo gradiente de pressão de vapor e do vento (ALLEN et al., 1998).

Fatores ou parâmetros meteorológicos importantes a serem considerados ao avaliar o processo de evapotranspiração são a radiação solar, a temperatura do ar, a umidade do ar e a velocidade do vento. O potencial de evaporação da atmosfera é expresso pela evapotranspiração de referência (ET_o). O tipo de cultura, a variedade, o estágio de desenvolvimento, a cobertura do solo, a densidade das plantas, o teor de água do solo e manejo são fatores considerados ao avaliar a evapotranspiração de cultura. O potencial de evapotranspiração de cultura em condições padrão (ET_c) refere-se à demanda de evaporação de cultura que é cultivada em condições ideais de disponibilidade de água, solo e condições meteorológicas. Quando as condições

de campo diferem das condições padrão, ou seja, condições reais de campo, fatores de correção são necessários para ajustar o ET_c, maiores detalhes nos trabalhos de DOORENBOS e PRUITT (1977) e ALLEN *et al.* (1998).

A Figura 2 mostra os fatores considerados na evapotranspiração que são: fatores *meteorológicos*, fatores da *cultura* e fatores do *solo e manejo*.

Figura 2 – Evapotranspiração e seus fatores de influência

Fonte: Autoria Própria.

2.3.3. MÉTODO PENMAN-MONTEITH

A taxa de evapotranspiração de uma superfície de referência é chamada de evapotranspiração de cultura de referência ou evapotranspiração de referência e é denotada como ET_o. A superfície de referência é uma hipotética cultura de referência de gramíneas com características específicas. Este conceito foi introduzido para estudar a demanda evaporativa da atmosfera independentemente do tipo de cultura, desenvolvimento da cultura, tipo de solo e práticas de manejo. Os únicos fatores que afetam a evapotranspiração de referência são os parâmetros meteorológicos. Para o cálculo da ET_o utiliza-se o método de Penman-Monteith da FAO 56 (ALLEN et al., 1998), faz uso de dados meteorológicos medidos ou derivados de dados comumente medidos. O procedimento de cálculo da evapotranspiração de referência (ET_o) a partir de dados meteorológicos são apresentados nesta seção.

2.3.3.1 EQUAÇÃO DE PENMAN-MONTEITH

O modelo de Penman-Monteith (ALLEN et al., 1998) para o cálculo da evapotranspiração de referência, ET_o, dada pela equação:

$$ET_o = \frac{0{,}408\,\Delta\,(R_n - G) + \gamma\,\frac{900}{T_{med}+273}\,u_2\,(e_s - e_a)}{\Delta + \gamma\,(1 + 0{,}34\,u_2)} \qquad (1)$$

sendo:

ET_o evapotranspiração de referência ..... [mm dia^{-1}]

Rn radiação líquida ..... [MJ m^{-2} dia^{-1}]

G fluxo de calor do solo (considerado zero) ..... [MJ m^{-2} dia^{-1}]

T_{med} temperatura média do ar ..... [°C]

u_2 velocidade do vento (medida ou calculada 2 m do solo) .. [m s^{-1}]

e_s pressão de saturação de vapor d'água do ar ..... [kPa]

e_a pressão de vapor d'água do ar ..... [kPa]

Δ inclinação da curva de pressão de vapor saturado ..... [kPa ºC^{-1}]

γ coeficiente psicrométrico ..... [kPa ºC^{-1}]

Nota: considera uma superfície de um cultivo hipotético de alfafa ou grama em estado de desenvolvimento ativo, cobrindo completamente o solo, sem deficiência hídrica e livre de pragas e doenças. A grama (*Paspalum notatum Flüggé*), superfície hipotética de referência, assume uma altura de aproximadamente 0,12 m, com resistência da superfície igual a 70 s m^{-1} e um albedo de 0,23. A temperatura e velocidade são medidos a 2 m do solo.

2.3.3.2 TEMPERATURA MÉDIA DO AR

Temperatura média, calculada por:

$$T_{med} = \frac{T_{max} + T_{min}}{2} \qquad (2)$$

sendo:

T_{med} temperatura média do ar ..... [°C]

T_{max} temperatura máxima..... [°C]

T_{min} temperatura mínima .....[°C]

2.3.3.3 VELOCIDADE MÉDIA

A velocidade média medida na altura *h*, pode ser corrigida a altura de 2 m, pela equação descrita abaixo:

$$u_2 = u_h \frac{4{,}87}{ln(\,67{,}8\,h - 5{,}42\,)} \quad (3)$$

sendo:

u_2 velocidade do vento (2 m do solo) .....[m s^{-1}]

u_h velocidade do vento (medida a *h* m do solo).....[m s^{-1}]

h altura medida velocidade do vento em relação ao solo ....[m]

2.3.3.4 PRESSÃO DE VAPOR SATURADO

A pressão de vapor saturado, calculada por:

$$e_s = \frac{e_{s\,max} + e_{s\,min}}{2} \quad (4)$$

com:

$$e_{s\,max} = 0{,}6108\ exp\left[\frac{17{,}27\,T_{max}}{T_{max}+237{,}3}\right] \quad (5)$$

e

$$e_{s\,min} = 0{,}6108\ exp\left[\frac{17{,}27\,T_{min}}{T_{min}+237{,}3}\right] \quad (6)$$

sendo:

e_s pressão de vapor saturado .....[kPa]

$e_{s\,max}$ pressão de vapor saturado máxima.....[kPa]

$e_{s\,min}$ pressão de vapor saturado mínima.....[kPa]

T_{max} temperatura máxima .....[°C]

T_{min} temperatura mínimo.....[°C]

2.3.3.5 PRESSÃO DE VAPOR

Utilizando os dados de umidade relativa, a pressão de vapor pode ser calculada por:

$$e_a = \frac{e_{s\,min}\frac{UR_{max}}{100} + e_{s\,max}\frac{UR_{min}}{100}}{2} \quad (7)$$

sendo:

e_a pressão vapor ..... [kPa]

$e_{s\,max}$ pressão de vapor saturado máxima..... [kPa]

$e_{s\,min}$ pressão de vapor saturado mínima ..... [kPa]

UR_{max} unidade relativa máxima..... [%]

UR_{min} unidade relativa mínimo..... [%]

2.3.3.6 INCLINAÇÃO DA CURVA DE PRESSÃO DE VAPOR SATURADO

Inclinação da curva de pressão de vapor saturado, calculada por:

$$\Delta = \frac{4098\left[0{,}6108\ exp\left(\frac{17{,}27\,T_{med}}{T_{med}+237{,}3}\right)\right]}{(T_{med}+237{,}3)^2} \quad (8)$$

sendo:

Δ inclinação da curva de pressão de vapor saturado ..... [kPa $^{o}C^{-1}$]

T_{med} temperatura média do ar ..... [°C]

2.3.3.7 COEFICIENTE PSICROMÉTRICO

O coeficiente psicrométrico, calculado por:

$$\gamma = 0{,}665 * 10^{-3} * P \quad (9)$$

com:

$$P = 101{,}3\left(\frac{293-0{,}0065z}{293}\right)^{5{,}26} \quad (10)$$

sendo:

γ coeficiente psicrométrico.....[kPa $^{o}C^{-1}$]

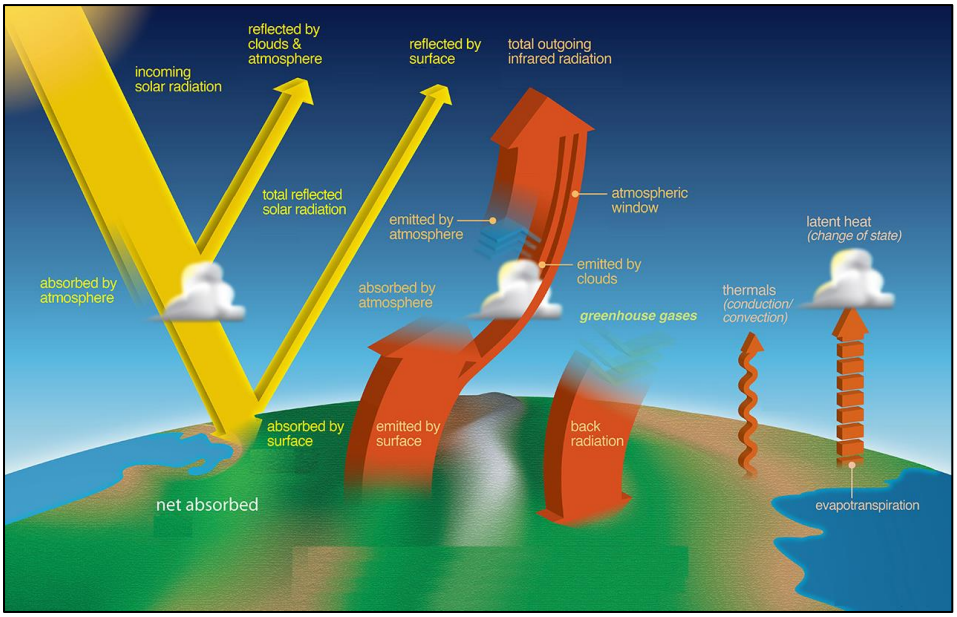
P pressão atmosférica (calculada ou medida).....[kPa]

z altitude .....[m]

2.3.3.8 RADIAÇÃO LÍQUIDA

Para o cálculo da radiação líquida é feito com base na medida da radiação solar medida (FRITSCHEN, 1967) por meio de um balanço de energia (PEREIRA et al., 2014), Figura 3, dado pela diferença da radiação de ondas curtas e ondas longas.

Figura 3 – Balanço de radiação solar



Fonte: Adaptado de (GILLARD, 2017).

Radiação líquida, calculada por:

$$R_n = R_{ns} - R_{nl} \quad (11)$$

sendo:

R_n radiação liquida.....[MJ m^{-2} dia^{-1}]

R_{ns} radiação líquida ondas curtas .....[MJ m^{-2} dia^{-1}]

R_{nl} radiação líquida ondas longas.....[MJ m^{-2} dia^{-1}]

2.3.3.9 RADIAÇÃO LÍQUIDA ONDAS CURTAS

Radiação líquida ondas curtas, é a diferença entre a radiação solar e a radiação solar refletida, calculada por:

$$R_{ns} = (1 - \alpha)R_s \quad (12)$$

sendo:

R_{ns} radiação líquida solar..... [MJ m^{-2} dia^{-1}]

R_s radiação solar (medida)..... [MJ m^{-2} dia^{-1}]

α α = 0,23 (albedo), cultura de referência de grama verde.. [ ]

2.3.3.10 RADIAÇÃO LÍQUIDA ONDAS LONGAS

Radiação líquida de ondas longas, calculada por:

$$R_{nl} = \sigma \left[\frac{T_{max,K}^4 + T_{min,K}^4}{2}\right] \left(0{,}34 - 0{,}14\sqrt{e_a}\right) \left(1{,}35\frac{R_s}{R_{so}} - 0{,}35\right) \quad (13)$$

sendo:

R_{nl} radiação líquida de ondas longas..... [MJ m^{-2} dia^{-1}]

σ constante de Stefan-Boltzmann (σ = 4,903*10^{-9})..... [MJ K^{-4} m^{-2} dia^{-1}]

$T_{max,K}$ Temperatura máxima..... [K]

$T_{min,K}$ Temperatura mínima ..... [K]

e_a pressão vapor..... [kPa]

R_s radiação solar (medida)..... [MJ m^{-2} dia^{-1}]

R_{so} radiação solar céu-claro (sem nuvens)..... [MJ m^{-2} dia^{-1}]

2.3.3.11 RADIAÇÃO SOLAR CÉU-CLARO

Radiação solar céu-claro, sem nuvens, dado por:

$$R_{so} = (0{,}75 + 2 * 10^{-5}z)R_e \quad (14)$$

sendo:

R_{so} radiação solar céu-claro (sem nuvens)..... [MJ m^{-2} dia^{-1}]

R_e radiação extraterrestre .....[MJ m^{-2} dia^{-1}]

z altitude .....[m]

2.3.3.12 RADIAÇÃO EXTRATERRESTRE

A radiação extraterrestre, R_e, para cada dia do ano e para diferentes latitudes pode ser estimado a partir da constante solar, da declinação solar e da época do ano por:

$$R_e = \frac{24*60}{\pi} G_{sc} * d_r[\omega_s \, sin(\varphi) \, sin(\delta) + \cos(\varphi) \, cos(\delta) \, sin(\omega_s)] \qquad (15)$$

sendo:

R_e radiação extraterrestre .....[MJ m^{-2} dia^{-1}]

G_{sc} constante solar (G_{sc} = 0,0820) .....[MJ m^{-2} min^{-1}]

d_r inverso da distância relativa terra-sol.....[ ]

φ latitude .....[rad]

δ declinação solar .....[rad]

ω_s ângulo horário do pôr do sol .....[rad]

2.3.3.13 INVERSO DA DISTÂNCIA RELATIVA TERRA-SOL

O inverso da distância relativa entre a terra e sol, calculado por:

$$d_r = 1 + 0{,}033 \, cos\left(\frac{2\pi}{365} J\right) \qquad (16)$$

sendo:

d_r inverso da distância relativa terra-sol.....[ ]

J número dia ano entre 1 (01JAN) e 365 ou 366 (31DEZ)...[ ]

2.3.3.14 TRANSFORMAÇÃO LATITUDE

A latitude, φ, expressa em radianos é positiva para o hemisfério norte e negativa para o hemisfério sul. A conversão de graus decimais para radianos é dada por:

$$\varphi[radianos] = \frac{\pi}{180}\varphi[graus\ decimais] \quad (17)$$

sendo:

$\varphi[radianos]$ ..... [rad]

$\varphi[graus\ decimais]$ ..... [º]

2.3.3.15 DECLINAÇÃO SOLAR

A declinação solar, calculada por:

$$\delta = 0{,}409\ sin\left(\frac{2\pi}{365}J - 1{,}39\right) \quad (18)$$

sendo :

δ declinação solar ..... [rad]

J número dia ano entre 1 (01Jan) e 365 ou 366 (31Dez) .... [ ]

2.3.3.16 ÂNGULO HORÁRIO DO PÔR DO SOL

O ângulo horário do *pôr do sol*, calculado por:

$$\omega_s = arc\ cos[-tg(\varphi)]tg(\delta) \quad (19)$$

sendo:

ω_s ângulo horário do pôr do sol ..... [rad]

φ latitude ..... [rad]

δ declinação solar ..... [rad]

2.3.3.17 VARIÁVEIS E CONSTANTES

A descrição de todas as variáveis e constantes utilizadas para o cálculo de ET_o são apresentadas nos Quadros 1, 2 e 3.

Quadro 1 – Constantes para o cálculo da evapotranspiração de referência

Constante	Descrição	Unidade
α	α = 0,23 (alberdo),	[ ]
	cultura de referência de grama verde.	
G	fluxo de calor do solo (considerado zero)	[MJ m^{-2} dia^{-1}]
G_{sc}	constante solar (G_{sc} = 0,0820)	[MJ m^{-2} min^{-1}]
σ	constante de Stefan-Boltzmann (σ = 4,903*10^{-9})	[MJ K^{-4} m^{-2} dia^{-1}]

Fonte: Autoria Própria.

Quadro 2 – Dados para o cálculo da evapotranspiração de referência

Variável	Descrição	Unidade
P	pressão atmosférica	[kPa]
	nota: a medida é feita milibar (mb),	
	necessita converter para kPa (1kPa = 10mb)	
T_{max}	temperatura máxima	[°C]
T_{min}	temperatura mínima	[°C]
R_s	radiação solar diária	[MJ m^{-2} dia^{-1}]
	nota: obtida pelo acúmulo da radiação horária	
	dada em [J m^{-2} h^{-1}]	
UR_{max}	unidade relativa máxima	[%]
UR_{min}	unidade relativa mínimo	[%]
u_2	velocidade do vento	[m s^{-1}]
	nota: medida ou calculada a 2m do solo	
z	altitude	[m]
φ	latitude	[rad]

Fonte: Autoria Própria.

Quadro 3 – Equações para o cálculo da evapotranspiração de referência

Variável	Descrição	Unidade	Equação
T_{med}	temperatura média do ar	[°C]	[02]
u_2	velocidade do vento (medida a 2 m do solo)	[m s^{-1}]	[03]
e_s	pressão de saturação de vapor d'água do ar	[kPa]	[04]
$e_{s\ max}$	pressão de vapor saturado máxima	[kPa]	[05]
$e_{s\ min}$	pressão de vapor saturado mínima	[kPa]	[06]
e_a	pressão de vapor d'água do ar	[kPa]	[07]
Δ	inclinação da curva de pressão de vapor saturado	[kPa $ºC^{-1}$]	[08]
γ	coeficiente psicrométrico	[kPa $ºC^{-1}$]	[09]
P	Pressão atmosférica calculada	[kPa]	[10]
d_r	inverso da distância relativa terra-sol	[]	[16]
φ	latitude	[rad]	[17]
δ	declinação solar	[rad]	[18]
ω_s	ângulo horário do pôr do sol	[rad]	[19]
R_e	radiação extraterrestre	[MJ m^{-2} dia^{-1}]	[15]
R_{so}	radiação solar céu-claro (sem nuvens)	[MJ m^{-2} dia^{-1}]	[14]
R_{nl}	radiação líquida de ondas longas	[MJ m^{-2} dia^{-1}]	[13]
R_{ns}	radiação líquida solar	[MJ m^{-2} dia^{-1}]	[12]
R_n	radiação líquida	[MJ m^{-2} dia^{-1}]	[11]
ET_o	evapotranspiração de referência	[mm dia^{-1}]	[01]

Fonte: Autoria Própria.

2.4 MECANISMO DE FALTA

O conjunto de dados escolhido é simulado segundo um mecanismos de falta. Os pesquisadores Little e Rubin (2002) descrevem os mecanismos: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Missing Not at Random* (MNAR). Os pesquisadores Garciarena e Santana (2017) constataram uma forte relação entre o método de imputação e o mecanismo de falta.

Considerando a ocorrência de um valor ausente em um fenômeno probabilístico, pode-se dizer que na matriz de dados ***Y*** de ordem *n x p* (*n*

observações e *p* variáveis), existem *m* valores *missings* ($\boldsymbol{Y}_{missing}$) e *np* - *m* valores observados ($\boldsymbol{Y}_{observado}$). Então, para:

a. Mecanismo **MCAR** (CHEN et al., 2017), os valores ausentes ocorrem de forma completamente aleatória, ou seja, os *missings* podem ocorrer em qualquer posição dos dados por meio de uma função probabilística que especifica a mesma probabilidade para cada posição f($\boldsymbol{Y}_{missing}$), não dependendo de $\boldsymbol{Y}_{observado}$ ou de $\boldsymbol{Y}_{missing}$ de outra posição.

b. Mecanismo **MAR** (LI; PARKER, 2014) (JUNNINEN et al., 2004) (EIROLA et al., 2013), o local de ocorrência dos valores ausentes podem ter relação com as variáveis observadas, porém não tem relação outros *missings*, f($\boldsymbol{Y}_{missing}$ / $\boldsymbol{Y}_{observado}$), ou seja, não depende de $\boldsymbol{Y}_{missing}$.

c. Mecanismo **MNAR**, o local de ocorrência dos valores ausentes tem relação outros *missings*, f($\boldsymbol{Y}_{missing}$ /$\boldsymbol{Y}$) depende de $\boldsymbol{Y}_{missing}$ e presumidamente de $\boldsymbol{Y}_{observado}$.

Na literatura tem-se trabalhos analisando o comportamento dos três mecanismos, por exemplo, os trabalhos de Pan et al. (2015); Zhu, He e Liatsis (2012); Khoshgoftaar e Van Hulse (2008) Twala, Jones e Hand (2008).

2.5 TAXA EM FALTA

A maioria das pesquisas examinam o desempenho das técnicas de imputação em diferentes taxas de valores ausentes. O trabalho de Lin e Tsai (2020) faz uma discussão do tema, enquadrando os trabalhos analisados na faixa de 5 a 50% de taxa de valores ausentes. Como esperado, o desempenho dos métodos de imputação diminui com o aumento da taxa de dados ausentes(GARCIARENA; SANTANA, 2017).

2.6 TÉCNICA DE IMPUTAÇÃO

O tratamento dos dados faltantes pode iniciar com a decisão de eliminar ou estimar os valores faltantes (PANDEY, 2020). Para eliminar os valores ausentes tem-se as técnicas de exclusão completa de casos (*listwise deletion*) e a exclusão pareada (*pairwise deletion*). Na exclusão completa de casos, *listwise deletion,* todos os casos que possuem pelo menos um valor ausente são eliminados, o que pode resultar na perda de muitos dados. A exclusão pareada, *pairwise deletion,* apenas as observações com valores ausentes para a variável de interesse são excluídas, o que permite que diferentes subconjuntos de dados sejam utilizados em diferentes análises, dependendo da disponibilidade de dados para cada variável. Embora a exclusão pareada possa ser mais eficiente do que a exclusão completa de casos, ela pode resultar em diferentes tamanhos amostrais para cada análise e afetar a validade das comparações entre variáveis. Essas abordagens são simples e fáceis de implementar, mas podem resultar em perda significativa de informações (VAN BUUREN, 2018).

A imputação refere-se à substituição de dados ausentes por valores estimados. Existem várias maneiras pelas quais os valores ausentes podem ser imputados, dependendo da natureza do problema e dos dados. Dependendo da natureza do problema, as técnicas de imputação podem ser amplamente classificadas como técnicas básicas de imputação que não consideram o tempo, tem-se a substituição por um valor constante, que pode ser alguma medida descritiva de posição (média, mediana ou mais frequente) de cada coluna na qual os valores ausentes estão localizados (MAGNANI, 2004; WEI et al., 2018). Agora para as técnicas básicas que leva em consideração o tempo, série temporal, tem-se as técnicas de substituição pelo valor observado anterior (*forward fill*), substituição pelo valor observado posterior (*back fill*) e interpolação linear (NGUYEN et al., 2020). A técnica interpolação linear é uma técnica de imputação que assume uma relação linear entre os observados e ausentes.

Para os métodos avançados, pode-se classificar em técnicas estatísticas multivariadas ou *machine learning*. As técnicas avançadas de imputação *machine learning*, usam algoritmos de aprendizado de máquina para imputar os valores ausentes em um conjunto de dados. Destaque para *K-nearest neighbor* (PATIL;

JOSHI; TOSHNIWAL, 2010) que considera os valores observados dos vizinhos mais próximos para substituir o valor ausente. Das técnicas estatísticas destaque para aplicação da análise de componentes principais, PCA, em conjunto com o algoritmo NIPALS, *Nonlinear Iterative Partial Least Squares* (WRIGHT, 2017) e com o algoritmo EM, *Expectation-Maximization* (DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, 1977).

Considerando a proposta de aplicação, avaliar a performance das técnicas estatísticas de imputação pela média simples e os algoritmos NIPALS-PCA e EM-PCA, as técnicas baseadas em aprendizado de máquina não foram contempladas no escopo deste capítulo.

2.6.1 IMPUTAÇÃO PELA MÉDIA

O método estatístico considerado padrão é a imputação pela média (GARCÍA-LAENCINA; SANCHO-GÓMEZ; FIGUEIRAS-VIDAL, 2010). Neste caso os valores ausentes de uma variável são preenchidos pelo valor médio dos valores observados desta variável ou com base na média de todo conjunto.

2.6.2 ANÁLISE DOS COMPONENTES PRINCIPAIS

A análise de componentes principais (PCA), a qual foi introduzida por *Karl Pearson* (1901) e está fundamentada em *Hotelling* (1933), visa reduzir a dimensão de um conjunto de dados, explicando a estrutura de variância e covariância de um vetor aleatório, composto de *p*-variáveis aleatórias, por meio da construção de novas variáveis, obtidas pela combinação linear das variáveis originais. Estas combinações lineares são chamadas de componentes principais e não correlacionados entre si (MINGOTI, 2005).

Dado um conjunto, com *p*-variáveis e *n*-amostras, denotado por $\boldsymbol{X} = [x_{ij}]$, com o elemento genérico x_{ij} representa a resposta da *i-ésima* variável (*i = 1, ..., p*) mensurada na *j-ésima* amostra (*j = 1, ..., n*), ou seja, a linha apresenta o vetor das respostas observadas nas mensurações realizadas na *j-ésimo* amostra.

Considerando a padronização dos dados, a matriz $\boldsymbol{X}$ pode ser reescrita pela matriz $\widehat{\boldsymbol{X}}_r$, fazendo a minimização da norma dos mínimos quadrados, dado por $\left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\right\|^2$ (DRAY; JOSSE, 2014).

Com $\widehat{\boldsymbol{X}}_r = \boldsymbol{USV}^T$, solução obtida pela decomposição em valores singulares de $\boldsymbol{X}$ (WRIGHT, 2017). Sendo $\boldsymbol{U}_{(n \times r)}$ e $\boldsymbol{V}_{(p \times r)}$ as matrizes ortonormais, e $\mathbf{S}_{(r \times r)}$ é uma matriz diagonal associada aos *r* autovalores.

Importante notar que para uma matriz ***X*** incompleta, a PCA padrão não pode ser aplicado. A seguir são apresentados dois algoritmos associados a PCA que podem trabalhar com dados ausentes, *EM-PCA* e *NIPALS-PCA*.

2.6.3 NIPALS-PCA

O algoritmo NIPALS (*Nonlinear Iterative Partial Least Squares*) pode ser utilizado para obter as componentes principais com a decomposição $\boldsymbol{X} = \boldsymbol{TP}^T$. Com as colunas de ***T*** chamadas de *scores* e as colunas de ***P*** de *loadings*.

O algoritmo começa com a inicialização de *h = 1* e $\mathbf{X}_h = \boldsymbol{X}$, então prosseguir com os seguintes passos:

1. Escolher $\mathbf{t}_h$ como alguma coluna de $\mathbf{X}_h$
2. Calcular *loadings* $\boldsymbol{p}_h = \frac{\boldsymbol{X}_h^T \boldsymbol{t}_h}{\boldsymbol{t}_h^T \boldsymbol{t}_h}$ e Normalizar $\boldsymbol{p}_h = \frac{\boldsymbol{p}_h}{\sqrt{\boldsymbol{p}_h^T \boldsymbol{p}_h}}$
3. Calcular *scores* $\boldsymbol{t}_h = \frac{\boldsymbol{X}_h^T \boldsymbol{p}_h}{\boldsymbol{p}_h^T \boldsymbol{p}_h}$

Repetir (2) e (3) até convergência para as ***h*** primeiras componentes principais

Faz $\boldsymbol{X}_{h+1} = \boldsymbol{X}_h - \boldsymbol{t}_h \boldsymbol{p}_h^T$. e $\boldsymbol{\lambda}_h = \boldsymbol{t}_h^T \boldsymbol{t}_h$ (autovalor).

Incrementa h = h+1 e repete para a próxima componente principal.

Monta as colunas da matriz ***T*** com $\mathbf{t}_h$ e as colunas de ***P*** com $\mathbf{p}_h$.

O algoritmo NIPALS pode ser modificado para trabalhar com dados ausente, proposta dos pesquisadores Martens e Martens (2001).

Se, para uma determinada variável *k* [coluna de ***X***], um valor ausente for encontrado em ***X*** para um determinado objeto ***i*** [linha de ***X***], então os elementos

correspondentes em $\mathbf{t}_{ih}$ também devem ser ignorados no cálculo dos carregamentos, que para $\boldsymbol{X}$-variável *k* é : $\boldsymbol{p}_{hk} = \frac{\boldsymbol{X}_{k,h-1}\boldsymbol{t}_h^T}{\boldsymbol{t}_h^T\boldsymbol{t}_h}$.

Da mesma forma, se, para uma determinada amostra *i* [linha de $\boldsymbol{X}$)], um valor ausente for encontrado em $\boldsymbol{X}$ para uma determinada variável *k* [coluna de $\boldsymbol{X}$], então os elementos correspondentes em $\mathbf{p}_{kh}$ também devem ser ignorados no cálculo das pontuações, que para a amostra *i* é: $\boldsymbol{t}_{ih} = \frac{\boldsymbol{X}_{i,h-1}\,\boldsymbol{p}_h}{\boldsymbol{p}_h^T\boldsymbol{p}_h}$.

Este algoritmo NIPALS-PCA foi implementado no ambiente computacional *R-Gui*, pacote NIPALS (WRIGHT, 2017).

2.6.4 EM-PCA

O método *iterativo* EM-PCA (KIERS, 1997) fornece os *scores* e loadings minimizando o critério de mínimos quadrados nas entradas observadas, $\|\boldsymbol{W} \circ (\boldsymbol{X} - \widehat{\boldsymbol{X}})\|^2$ com $w_{ij} = 0$ se x_{ij} está faltando e 1 caso contrário e o operador $\circ$ denota produto elementar.

A minimização é alcançada por meio de um procedimento iterativo, os valores ausentes são substituídos por valores aleatórios e, em seguida, o PCA é aplicado no conjunto de dados concluído e os valores ausentes são atualizados pelos valores ajustados $\widehat{\boldsymbol{X}}_S = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ usando um número predefinido de dimensões *S,* o procedimento é repetido até a convergência (DRAY; JOSSE, 2014).

Esse método fornece *scores* para os indivíduos e as variáveis, e uma imputação para os valores faltantes. Uma questão importante diz respeito ao número de dimensões *S* que devem ser definidas no início do algoritmo iterativo.

Os pesquisadores Josse e Husson (2012b) sugeriram métodos baseados na validação cruzada para estimar esse parâmetro a partir de um conjunto de dados incompleto. O método é implementado no ambiente computacional *R-Gui*, função *imputePCA* do pacote *missMDA* (JOSSE; HUSSON, 2012a, 2016).

Maiores detalhes podem ser encontrados nos trabalhos (DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977; JOSSE; HUSSON; PAGÈS, 2009; JOSSE; PAGÈS; HUSSON, 2011; KIERS, 1997; SCHAFER, 1997).

2.7 MÉTODO DE AVALIAÇÃO

Para verificar o desempenho de uma técnica de imputação, pode-se utilizar os métodos de avaliação direta, precisão da classificação (SCIENCEDIRECT, 2023a) e tempo computacional. Na avaliação direta é feita uma comparação entre os valores estimados e observados. A avaliação da precisão da classificação verifica o poder de classificação de um método, aquele que melhor performance tiver frente a um padrão, será considerado melhor método de imputação. Na avaliação do tempo computacional, considera-se de melhor desempenho o algoritmo com menor tempo de imputação.

Nesta seção, o foco foi detalhar a avaliação direta, considerando-se os indicadores estatísticos: coeficiente de correlação de Pearson, erro médio, erro absoluto médio, erro absoluto médio percentual, erro quadrático médio, raiz do erro quadrático médio, raiz do erro quadrático médio percentual, índice de concordância de Willmott e coeficiente de confiança.

2.7.1 COEFICIENTE DE CORRELAÇÃO DE PEARSON

O coeficiente de correlação de Pearson (STANTON, 2001) é o grau de associação entre os valores observados e os valores imputados por um dos métodos avaliados, mede a precisão, o quanto os valores observados e estimados/imputados estão relacionados de maneira linear. Calculado por:

$$r = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} \qquad (20)$$

sendo:

r .....: Coeficiente de Correlação de Pearson

x_i .....: i-ésimo valor observado *(i=1, ..., m)*

$\bar{x}$ .....: média dos valores observados

y_i .....: i-ésimo valor imputado *(i=1, ..., m)*

$\bar{y}$ .....: média dos valores imputados

m .....: número de *missings*

2.7.2 ME – ERRO MÉDIO

O erro médio, *Mean Error*, consiste na média das diferenças entre os valores observados e estimados. Calculado por:

$$ME = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i) \quad (21)$$

sendo:

ME ..... : Erro Médio

x_i ..... : i-ésimo valor observado *(i=1, ..., m)*

y_i ..... : i-ésimo valor imputado *(i=1, ..., m)*

m..... : número de *missings*

2.7.3 MAE – ERRO ABSOLUTO MÉDIO

O erro absoluto médio, *Mean Absolute Error*, (SCIENCEDIRECT, 2023b) consiste na média do módulo das diferenças entre os valores observados e valores estimados. Calculado por:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|x_i - y_i| \quad (22)$$

sendo:

MAE ..... : Erro Absoluto Médio

x_i ..... : i-ésimo valor observado *(i=1, ..., m)*

y_i ..... : i-ésimo valor imputado *(i=1, ..., m)*

m..... : número de *missings*

2.7.4 pMAE – ERRO ABSOLUTO MÉDIO PERCENTUAL

O erro absoluto médio percentual, *Mean Absolute Percentual Error*, consiste na média do módulo da diferença entre os valores observados e estimados, dividido pelo valor observado. Calculado por:

$$pMAE = 100 * \frac{1}{m}\sum_{i=1}^{m}\left|\frac{x_i - y_i}{x_i}\right| \quad (23)$$

sendo:

$pMAE$ .....: Erro Absoluto Médio Percentual

x_i .....: i-ésimo valor observado *(i=1, ..., m)*

y_i .....: i-ésimo valor imputado *(i=1, ..., m)*

m .....: número de *missings*

2.7.5 MSE – ERRO QUADRÁTICO MÉDIO

O erro quadrático médio, *Mean Squared Error*, consiste na média das diferenças ao quadrado entre os valores observados e estimados. Calculado por:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2 \qquad (24)$$

sendo:

MSE .....: Erro Quadrático Médio

x_i .....: i-ésimo valor observado *(i=1, ..., m)*

y_i .....: i-ésimo valor imputado *(i=1, ..., m)*

m .....: número de *missings*

2.7.6 RMSE – RAIZ ERRO QUADRÁTICO MÉDIO

A raiz do erro quadrático médio, *Root Mean Squared Error,* é obtido pela raiz quadrada do erro quadrático médio entre os valores observados e valores estimados. Calculado por:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \qquad (25)$$

sendo:

$RMSE$.....: Raiz do Erro Quadrático Médio

x_i .....: i-ésimo valor observado *(i=1, ..., m)*

y_i .....: i-ésimo valor imputado *(i=1, ..., m)*

m .....: número de *missings*

2.7.7 pRMSE – RAIZ ERRO QUADRÁTICO MÉDIO PERCENTUAL

A raiz do erro quadrático médio percentual, *Root Mean Squared Percentual Error,* é obtido dividindo a raiz quadrada do erro quadrático médio pela média das observações. Facilita na interpretação métrica, erro dado em termos percentual em relação a médias dos dados observados. Para análise do desempenho (JAMIESON; PORTER; WILSON, 1991) utiliza uma escala classificatória para os diferentes intervalos de *pRMSE*: Excelente (pRMSE < 10%); Bom (10% ≤ pRMSE < 20%); Aceitável (20% ≤ pRMSE < 30%) e Pobre (pRMSE ≥ 30). Calculado por:

$$pRMSE = 100 * \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2}}{\frac{1}{m}\sum_{i=1}^{m} x_i} \qquad (26)$$

sendo:

$pRMSE$: .. Raiz do Erro Quadrático Médio Percentual

x_i: ..... i-ésimo valor observado *(i=1, ..., m)*

y_i: ..... i-ésimo valor imputado *(i=1, ..., m)*

m: ..... número de *missings*

2.7.8 ÍNDICE DE CONCORDÂNCIA DE WILLMOTT

O índice de concordância de Willmott (WILLMOTT et al., 1985), representado pela letra “d” é uma métrica que exprime a exatidão, relacionando o afastamento dos valores estimados em relação aos observados, o qual pode ser calculado pela equação 27.

$$d = 1 - \frac{\sum_{i=1}^{m}(x_i - y_i)^2}{\sum_{i=1}^{m}(|x_i - \bar{x}| + |y_i - \bar{x}|)^2} \qquad (27)$$

sendo:

d ..... : Índice de Concordância de Willmott

x_i ..... : i-ésimo valor observado *(i=1, ..., m)*

$\bar{x}$ ..... : média dos valores observados

y_i ..... : i-ésimo valor imputado *(i=1, ..., m)*

m..... : número de *missings*

2.7.9 COEFICIENTE DE CONFIANÇA

O coeficiente de confiança (CAMARGO; SENTELHAS, 1997), *c,* é obtido pelo produto entre o coeficiente de correlação (*r*) e o índice de Willmott (*d*). Os pesquisadores Camargo e Sentelhas (1997), definiram o desempenho de um modelo de estimação, como: Ótimo (c > 0,85); Muito Bom (0,75 < c ≤ 0,85); Bom (0,65 < c ≤ 0,75); Mediano (0,60 < c ≤ 0,65), Sofrível (0,50 < c ≤ 0,60), Mau (c 0,40 < c ≤ 0,50) e Péssimo (c ≤ 0,40). Calculado por:

$$c = r * d \quad (28)$$

sendo:

c .....: Coeficiente de Confiança

r .....: Coeficiente de Correlação de Pearson

d .....: Índice de Concordância de Willmott

2.8 CONSIDERAÇÕES

Neste capítulo foi descrito um planejamento detalhado de simulação em MVI, com foco no banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho. Preparando conceitos para a aplicação apresentada no capítulo três: *COMPARAÇÃO DE ALGORITMOS DE ANÁLISE DE COMPONENTES PRINCIPAIS PARA IMPUTAÇÃO EM DADOS AGROMETEOROLÓGICOS EM ALTA DIMENSÃO E TAMANHO AMOSTRAL REDUZIDO.*

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

ALLEN, R. G. et al. Crop Evapotranspiration - Guidelines for computing crop water requirements. **FAO Irrigation and drainage**, 1998.

BDMEP. **Instituto Nacional de Meteorologia - INMET**. Disponível em: <https://portal.inmet.gov.br/servicos/bdmep-dados-historicos/>. Acesso em: 3 fev. 2023.

CAI, J. B. et al. Simulation of the soil water balance of wheat using daily weather forecast messages to estimate the reference evapotranspiration. **Hydrology and Earth System Sciences**, v. 13, n. 7, p. 1045–1059, 9 jul. 2009.

CAMARGO, Â. P. DE; SENTELHAS, P. C. Avaliação do Desempenho de Diferentes Métodos de Estimativa da Evapotranspiração Potencial no Estado de São Paulo no Brasil. **Revista Brasileira de Agrometeorologia**, v. 5, n. 1, p. 89–97, 1997.

CHEN, X. et al. Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation. **Knowledge-Based Systems**, v. 132, p. 249–262, 15 set. 2017.

CIOS, K. J.; WILLIAM MOORE, G. Uniqueness of medical data mining. **Artificial Intelligence in Medicine**, v. 26, n. 1–2, p. 1–24, 1 set. 2002.

CPTEC. **Centro de Previsão de Tempo e Estudos Climáticos - INPE - Botucatu / SP**. Disponível em: <https://www.cptec.inpe.br/>. Acesso em: 30 jan. 2023.

DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal ofthe Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977.

DOORENBOS, J.; PRUITT, W. O. Crop Water Requirements. **FAO Irrigation and drainage**, v. 24, p. 1–144, 1977.

DRAY, S.; JOSSE, J. Principal component analysis with missing values: a comparative survey of methods. **Plant Ecology**, v. 216, n. 5, p. 657–667, 1 maio 2014.

DUA, D. E GRAFF, C. **Machine Learning Repository UCI**. Disponível em: <http://archive.ics.uci.edu/ml/index.php>. Acesso em: 7 dez. 2022.

EIROLA, E. et al. Distance estimation in numerical data sets with missing values. **Information Sciences**, v. 240, p. 115–128, 10 ago. 2013.

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

FRITSCHEN, L. J. Net and solar radiation relations over irrigated field crops. **Agricultural Meteorology**, v. 4, n. 1, p. 55–62, 1 jan. 1967.

GARCÍA-DIEGO, F. J.; ZARZO, M. Microclimate monitoring by multivariate statistical control: The renaissance frescoes of the Cathedral of Valencia (Spain). **Journal of Cultural Heritage**, v. 11, n. 3, p. 339–344, 2010.

GARCÍA-LAENCINA, P. J.; SANCHO-GÓMEZ, J. L.; FIGUEIRAS-VIDAL, A. R. Pattern classification with missing data: A review. **Neural Computing and Applications**, v. 19, n. 2, p. 263–282, 3 set. 2010.

GARCIARENA, U.; SANTANA, R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. **Expert Systems with Applications**, v. 89, p. 52–65, 15 dez. 2017.

GILLARD, E. **Keeping an eye on Earth's energy budget**. Disponível em: <https://climate.nasa.gov/news/2641/keeping-an-eye-on-earths-energy-budget/>. Acesso em: 14 abr. 2023.

GONSAGA DE CARVALHO, L. et al. Evapotranspiração de referência: uma abordagem atual de diferentes métodos de estimativa. **Pesquisa Agropecuária Tropical**, v. 41, n. 3, p. 456–465, 6 jul. 2011.

HARPER, P. R. A review and comparison of classification algorithms for medical decision making. **Health Policy**, v. 71, n. 3, p. 315–331, 1 mar. 2005.

HART, Q. J. et al. Daily reference evapotranspiration for California using satellite imagery and weather station measurement interpolation. **https://doi-org.ez87.periodicos.capes.gov.br/10.1080/10286600802003500**, v. 26, n. 1, p. 19–33, 2009.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417–441, set. 1933.

INMET. **Instituto Nacional de Meteorologia - INMET**. Disponível em: <https://portal.inmet.gov.br/>. Acesso em: 3 fev. 2023.

JAMIESON, P. D.; PORTER, J. R.; WILSON, D. R. A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. **Field Crops Research**, v. 27, n. 4, p. 337–350, 1991.

JOSSE, J.; HUSSON, F. Selecting the number of components in principal component analysis using cross-validation approximations. **Computational Statistics and Data Analysis**, v. 56, n. 6, p. 1869–1879, 2012a.

JOSSE, J.; HUSSON, F. Handling missing values in exploratory multivariate data analysis methods. **Journal de la société française de statistique**, v. 153, n. 2, p. 79–99, 2012b.

JOSSE, J.; HUSSON, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. **Journal of Statistical Software**, v. 70, p. 1–31, 4 abr. 2016.

JOSSE, J.; HUSSON, F.; PAGÈS, J. Gestion des données manquantes en Analyse en Composantes Principales. **Journal de la Société Française de Statistique**, v. 150, p. 2, 2009.

JOSSE, J.; PAGÈS, J.; HUSSON, F. Multiple imputation in principal component analysis. **Advances in Data Analysis and Classification**, v. 5, n. 3, p. 231–246, 2011.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, 1 jun. 2004.

KHOSHGOFTAAR, T. M.; VAN HULSE, J. Imputation techniques for multivariate missingness in software measurement data. **Software Quality Journal**, v. 16, n. 4, p. 563–600, 11 jun. 2008.

KIERS, H. A. L. Weighted least squares fitting using ordinary least squares algorithms. **Psychometrika**, v. 62, n. 2, p. 251–266, 1997.

KISI, O. et al. Modeling reference evapotranspiration using a novel regression-based method: radial basis M5 model tree. **Theoretical and Applied Climatology**, v. 145, n. 1–2, p. 639–659, 1 jul. 2021.

KOOL, D. et al. A review of approaches for evapotranspiration partitioning. **Agricultural and Forest Meteorology**, v. 184, p. 56–70, 15 jan. 2014.

LAKSHMINARAYAN, K.; HARP, S. A.; SAMAD, T. Imputation of Missing Data in Industrial Databases. **Applied Intelligence 1999 11:3**, v. 11, n. 3, p. 259–275, nov. 1999.

LI, Y.; PARKER, L. E. Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks. **Information Fusion**, v. 15, n. 1, p. 64–79, 1 jan. 2014.

LIAO, S. G. et al. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? **BMC bioinformatics**, v. 15, n. 1, 5 nov. 2014.

LIN, W.-C.; TSAI, C.-F. Missing value imputation: a review and analysis of the literature (2006–2017). **Artificial Intelligence Review**, v. 53, p. 1487–1509, 2020.

LITTLE, R. J. A.; RUBIN, D. B. **Single Imputation Methods**. [s.l.] John Wiley & Sons, Ltd, 2002.

MAGNANI, M. Techniques for dealing with missing data in knowledge discovery tasks. 2004.

MARIN, F. R. et al. Revisiting the crop coefficient–reference evapotranspiration procedure for improving irrigation management. **Theoretical and Applied Climatology**, v. 138, n. 3–4, p. 1785–1793, 1 nov. 2019.

MARTENS, H.; MARTENS, M. **Multivariate Analysis of Quality: An Introduction**. Chichester: John Wiley & Sons, 2001.

MARTÍ, P.; GASQUE, M. Ancillary data supply strategies for improvement of temperature-based ETo ANN models. **Agricultural Water Management**, v. 97, n. 7, p. 939–955, 2010.

MARTÍ, P.; ZARZO, M. Multivariate statistical monitoring of ETo: A new approach for estimation in nearby locations using geographical inputs. **Agricultural and Forest Meteorology**, v. 152, n. 1, p. 125–134, 2012.

MCILROY, I. C. Terminology and concepts in natural evaporation. **Agricultural Water Management**, v. 8, n. 1–3, p. 77–98, 1984.

MINGOTI, S. A. **Análise de dados através de Métodos de Estatística Multivariada: uma abordagem Aplicada**. Belo Horizonte: Editora UFMG, 2005.

NGUYEN, M. et al. Predicting Alzheimer's disease progression using deep recurrent neural networks. **NeuroImage**, v. 222, p. 117203, 15 nov. 2020.

PAN, R. et al. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. **Applied Intelligence**, v. 43, n. 3, p. 614–632, 22 out. 2015.

PANDEY, P. **A Guide to Handling Missing values in Python**. Disponível em: <https://www.kaggle.com/code/parulpandey/a-guide-to-handling-missing-values-in-python>. Acesso em: 3 fev. 2023.

PATIL, B. M.; JOSHI, R. C.; TOSHNIWAL, D. Missing value imputation based on k-mean clustering with weighted distance. **Communications in Computer and Information Science**, v. 94 CCIS, n. PART 1, p. 600–609, 2010.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, p. 559–572, 1901.

PEREIRA, A. B. et al. Estimation method of grass net radiation on the determination of potential evapotranspiration. **Meteorological Applications**, v. 21, n. 2, p. 369–375, 1 abr. 2014.

PEREIRA, L. S. et al. Evapotranspiration: Concepts and Future Trends. **Journal of Irrigation and Drainage Engineering**, v. 125, n. 2, p. 45–51, mar. 1999.

RAHMAN, M. G.; ISLAM, M. Z. Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. **Knowledge-Based Systems**, v. 53, p. 51–65, 1 nov. 2013.

SCHAFER, J. L. **Analysis of Incomplete Multivariate Data**. New York: Chapman & Hall/CRC, 1997.

SCIENCEDIRECT. **Classification Accuracy - an overview | ScienceDirect Topics**. Disponível em: <https://www.sciencedirect.com/topics/computer-science/classification-accuracy>. Acesso em: 6 jan. 2023a.

SCIENCEDIRECT. **Mean Absolute Error - an overview | ScienceDirect Topics**. Disponível em: <https://www.sciencedirect.com/topics/computer-science/mean-absolute-error>. Acesso em: 10 jan. 2023b.

SNIRH. **Portal HIDROWEB**. Disponível em: <https://www.snirh.gov.br/hidroweb/apresentacao>. Acesso em: 30 jan. 2023.

STÄDLER, N.; STEKHOVEN, D. J.; BÜHLMANN, P. Pattern alternating maximization algorithm for missing data in large P, small N problems. **Journal of Machine Learning Research**, v. 15, p. 1903–1928, 3 maio 2014.

STANTON, J. M. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. **Journal of Statistics Education**, v. 9, n. 3, 2001.

STEKHOVEN, D. J.; BÜHLMANN, P. MissForest—non-parametric missing value imputation for mixed-type data. **Bioinformatics**, v. 28, n. 1, p. 112–118, 1 jan. 2012.

TERINK, W.; IMMERZEEL, W. W.; DROOGERS, P. Climate change projections of precipitation and reference evapotranspiration for the Middle East and Northern Africa until 2050. **International Journal of Climatology**, v. 33, n. 14, p. 3055–3072, 30 nov. 2013.

TIAN, J. et al. Missing data analyses: A hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering. **Applied Intelligence**, v. 40, n. 2, p. 376–388, 5 mar. 2014.

TROYANSKAYA, O. et al. Missing value estimation methods for DNA microarrays. **Bioinformatics**, v. 17, n. 6, p. 520–525, 1 jun. 2001.

TWALA, B. E. T. H.; JONES, M. C.; HAND, D. J. Good methods for coping with missing data in decision trees. **Pattern Recognition Letters**, v. 29, n. 7, p. 950–956, 1 maio 2008.

VAN BUUREN, S. **Flexible Imputation of Missing Data**. 2nd Editio ed. New York: Chapman and Hall/CRC, 2018.

WANG, J. et al. Development of Monthly Reference Evapotranspiration Machine Learning Models and Mapping of Pakistan—A Comparative Study. **Water 2022, Vol. 14, Page 1666**, v. 14, n. 10, p. 1666, 23 maio 2022.

WEI, R. et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. **Scientific Reports 2018 8:1**, v. 8, n. 1, p. 1–10, 12 jan. 2018.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. **Statistics in Medicine**, v. 30, n. 4, p. 377–399, 2011.

WILLMOTT, C. J. et al. Statistics for the evaluation and comparison of models. **Journal of Geophysical Research**, v. 90, n. C5, p. 8995, 1985.

WRIGHT, K. **The NIPALS algorithm**. Disponível em: <https://cran.r-project.org/web/packages/nipals/vignettes/nipals_algorithm.html>. Acesso em: 14 dez. 2022.

ZHU, B.; HE, C.; LIATSIS, P. A robust missing value imputation method for noisy data. **Applied Intelligence**, v. 36, n. 1, p. 61–74, 2012.

CAPÍTULO 3

COMPARAÇÃO DE ALGORITMOS DE ANÁLISE DE COMPONENTES PRINCIPAIS PARA IMPUTAÇÃO EM DADOS AGROMETEOROLÓGICOS EM ALTA DIMENSÃO E TAMANHO AMOSTRAL REDUZIDO[1]

COMPARISON OF PRINCIPAL COMPONENT ANALYSIS ALGORITHMS FOR IMPUTATION IN AGROMETEOROLOGICAL DATA IN HIGH DIMENSION AND REDUCED SAMPLE SIZE

Valter Cesar de Souza[$&*], **Sérgio Augusto Rodrigues**[$&]

[$]São Paulo State University (Unesp), School of Agriculture, Botucatu, São Paulo, Brasil

* Corresponding author
E-mail: valter.souza@unesp.br

[&]These authors contributed equally to this work.

[1] Capítulo redigido de acordo com as normas do periódico PLOS ONE.

RESUMO

Dados meteorológicos obtidos com acurácia, qualidade e confiabilidade são essenciais para diversas áreas da agronomia, destacando-se estudos sobre a evapotranspiração de referência (ET_o), a qual revela-se importante componente no ciclo hidrológico, no planejamento e gestão de sistemas de irrigação, na modelagem da demanda de água, no monitoramento do stress hídrico, na estimativa do balanço hídrico, em estudos hidrológicos e ambientais. Entretanto, dados registrados no tempo geralmente estão sujeitos a falhas ou erros, como medições ausentes, *missings*. O objetivo deste capítulo foi verificar a performance de procedimentos multivariados alternativos de análise de componentes principais (NIPALS-PCA e EM-PCA) na imputação de dados ausentes em séries temporais de variáveis meteorológicas, considerando bases de dados no cenário de alta dimensão e amostra reduzida, com diferentes percentuais de *missings*. Bases de dados de variáveis meteorológicas, no período de 2011 a 2021, mensurados em 45 estações meteorológicas automáticas da região de São Paulo, Brasil. Foram utilizadas para obtenção de uma base de dados diária de ET_o. Foram simulados cinco cenários de *missings* (10%, 20%, 30%, 40%, 50%). Para cada cenário foi gerado, de forma aleatória, um conjunto de dados (posições) retirados da base de ET_o e na sequência foi feita a imputação pelos procedimentos de NIPALS-PCA, EM-PCA e a imputação simples pela média (IM). Este ciclo foi executado 100 vezes, ao final, foi calculado os indicativos médios de desempenho. Para avaliação do desempenho estatístico foram utilizados os indicativos: coeficiente de correlação entre os valores imputados e observados (r), erro absoluto médio (MAE), erro absoluto médio percentual (pMAE), erro quadrático médio (MSE), raiz do erro quadrático médio percentual (pRMSE), índice de concordância de Willmott (d) e o coeficiente de confiança (c).Considerando o cenário com 10% de *missings*, o NIPALS-PCA obteve o menor *pMAE* (15,4%), seguido da EM-PCA (17,0%), enquanto a IM obteve um *pMAE* igual a 24,7%. No cenário com 50% de *missings* ocorre uma inversão de desempenho, com menor *pMSE* (19,1%) para o algoritmo EM, seguido do NIPALS (19,9%). Para a escala classificação para os diferentes intervalos de *pRMSE*, as abordagens NIPALS-PCA e EM-PCA apresentam bons resultados (10% ≤ pRMSE < 20%) na imputação de valores ausentes. Com destaque para o método NIPALS-PCA nos cenários de 10%, 20% e 30% e EM-PCA para os cenários de 40% e 50%. Com base na classificação do indicativo estatístico da base de validação dos modelos de imputação NIPALS-PCA, EM-PCA e IM, há indicativos que os mesmos são aptos à estimativa de *missings* da evapotranspiração de referência, com destaque para os modelos de imputação PCA nos algoritmos NIPALS e EM.

Palavras-chave: dados ausentes; PCA; NIPALS-PCA; EM-PCA.

ABSTRACT

Meteorological data obtained with accuracy, quality and reliability are essential for several areas of agronomy, highlighting studies on reference evapotranspiration (ET_o), which proves to be an important component in the hydrological cycle, in the planning and management of irrigation systems, modeling water demand, monitoring water stress, estimating water balance, hydrological and environmental studies. However, data recorded over time are generally subject to flaws or errors, such as missing measurements, missing data. The objective of this chapter was to verify the performance of alternative multivariate principal component analysis procedures (NIPALS-PCA and EM-PCA) in the imputation of missing data in time series of meteorological variables, considering databases in the high dimension scenario and small sample, with different percentages of missing data. Databases of meteorological variables, from 2011 to 2021, measured in 45 automatic meteorological stations in the region of São Paulo, Brazil. They were used to obtain a daily ET_o database. Five missing scenarios were simulated (10%, 20%, 30%, 40%, 50%). For each scenario, a set of data (positions) was randomly generated from the ET_o base and then imputed using the NIPALS-PCA, EM-PCA procedures and simple imputation using the average (IM). This cycle was executed 100 times, at the end, the average performance indicators were calculated. Indicatives were used to evaluate the statistical performance: correlation coefficient between imputed and observed values (r), Mean Absolute Error (MAE), Mean Absolute Percentage Error (pMAE), Mean Squared Error (MSE), Root Mean Squared Percentage Error (pRMSE), Willmott concordance index (d) and the confidence coefficient (c). Considering the scenario with 10% missing, NIPALS-PCA obtained the lowest pMAE (15.4%), followed by EM-PCA (17.0%), while IM obtained a pMAE equal to 24.7%. In the scenario with 50% missing, there is a reversal of performance, with a lower pMSE (19.1%) for the EM algorithm, followed by NIPALS (19.9%). For the rating scale for the different pRMSE intervals, the NIPALS-PCA and EM-PCA approaches show good results (10% ≤ pRMSE < 20%) in imputing missing values. With emphasis on the NIPALS-PCA method in the 10%, 20% and 30% scenarios and EM-PCA for the 40% and 50% scenarios. Based on the classification of the statistical indicator of the validation base of the imputation models NIPALS-PCA, EM-PCA and IM, there are indications that they can estimate missing data for reference evapotranspiration, with emphasis on the imputation models PCA in the NIPALS and EM algorithms.

Keywords: missing data; PCA; NIPALS-PCA; EM-PCA.

3.1 INTRODUÇÃO

A mensuração da evapotranspiração é uma tarefa complexa, devido aos altos custos das técnicas diretas para implantação, operação e manutenção dos equipamentos de medição (ALLEN et al., 2011; RANA; KATERJI, 2000). Como alternativa, métodos indiretos (ONNABI MILANI et al., 2007) são utilizados por meio de equações matemáticas capazes de se ajustarem às condições climáticas locais, necessitando de séries históricas de dados meteorológicos. No entanto, dados registrados no tempo geralmente estão sujeitos a falhas ou erros (HASAN et al., 2021), como medições ausentes, comumente chamados de *missings* (WHITE; ROYSTON; WOOD, 2011).

Entre as diversas alternativas para o preenchimento de *missings*, destaca-se, sobretudo em razão de sua versatilidade e facilidade de aplicação em fenômenos de alta dimensão, possibilitando a interpretação do mesmo em uma dimensão reduzida, o procedimento multivariado de Análise de Componentes Principais (PCA), ferramenta importante de análise exploratória de dados e caracterização da variabilidade espacial (GARCÍA-DIEGO; ZARZO, 2010; JOSSE; HUSSON, 2012; MARTÍ; ZARZO, 2012).

A PCA tradicional é utilizada para análise exploratória e redução da dimensionalidade, podendo também ser utilizada como procedimento de imputação de dados ausentes (DE KETELAERE; HUBERT; SCHMITT, 2015). Destaque para a utilização conjunta com métodos iterativos como o algoritmo NIPALS (DE LA FUENTE; GARCÍA-MUÑOZ; BIEGLER, 2010; ESHGHI, 2014; HOWLEY et al., 2006; PATEL; SIVANATHAN; MHASKAR, 2021; VYAS et al., 2021; YANG et al., 2012), *Nonlinear Iterative Partial Least Squares* (WRIGHT, 2017) e com o algoritmo EM (BUCIOR-KWACZYŃSKA, 2018; MALAN et al., 2020; NILASHI et al., 2022; XIE et al., 2019), *Expectation-Maximization* (DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, 1977).

Utilizando o algoritmo NIPALS em conjunto com PCA, NIPALS-PCA, os pesquisadores Martí e Zarzo (2012) trabalharam com dados de evapotranspiração de referência registrados em 30 estações na costa mediterrânea da Espanha, verificando que NIPALS-PCA apresentou um desempenho melhor ao comparar com métodos de preenchimento baseados em vizinhos mais próximos.

Josse e Husson (JOSSE; HUSSON, 2016) disponibilizam o método EM-PCA no pacote *missMDA* do ambiente computacional *R-Gui*[2] (R CORE TEAM, 2022) o qual possibilita realizar a reconstrução da matriz de dados das variáveis aleatórias associado a um número de componentes pré-determinados, podendo ser utilizada para dados em alta dimensão (DRAY; JOSSE, 2014; JOSSE; HUSSON, 2012; JOSSE; HUSSON; PAGÈS, 2009; JOSSE; PAGÈS; HUSSON, 2011; PODANI et al., 2021).

Os pesquisadores Dray e Josse (2014) apresentaram uma discussão das soluções para estimar valores ausentes por meio da PCA e avaliaram a performance dos métodos de imputação em um conjunto de dados no domínio da ecologia. Estes autores indicaram a necessidade de aplicar as técnicas alternativas de imputação via PCA, avaliando as vantagens e desvantagens frente a diferentes domínios e conjuntos de dados.

Neste contexto, o objetivo deste capítulo foi verificar a performance de procedimentos multivariados alternativos de análise de componentes principais (NIPALS-PCA e EM-PCA) na imputação de dados ausentes em séries temporais de variáveis meteorológicas, considerando bases de dados no cenário de alta dimensão e amostra reduzida, com diferentes percentuais de *missings*.

3.2 MATERIAL E MÉTODOS

Foram utilizadas bases de dados horárias, disponibilizadas pelo Instituto Nacional de Meteorologia (INMET, 2022) de cada variável meteorológica, do período de 1 de janeiro de 2012 a 31 de dezembro de 2021, avaliadas em 45 estações meteorológicas automáticas da região do Estado de São Paulo, Brasil. O Quadro 1 apresenta as informações das estações meteorológicas automáticas utilizadas nesta pesquisa.

Quadro 1 - Informações das estações meteorológicas automáticas

ID	Localidade	Estado	Latitude [°]	Longitude [°]	Altitude [m]
1	ARIRANHA	SP	21°7'59"S	48°50'26"W	525,4
2	AVARE	SP	23°6'6"S	48°56'28"W	776,4

(contínua)

[2] https://www.R-project.org/

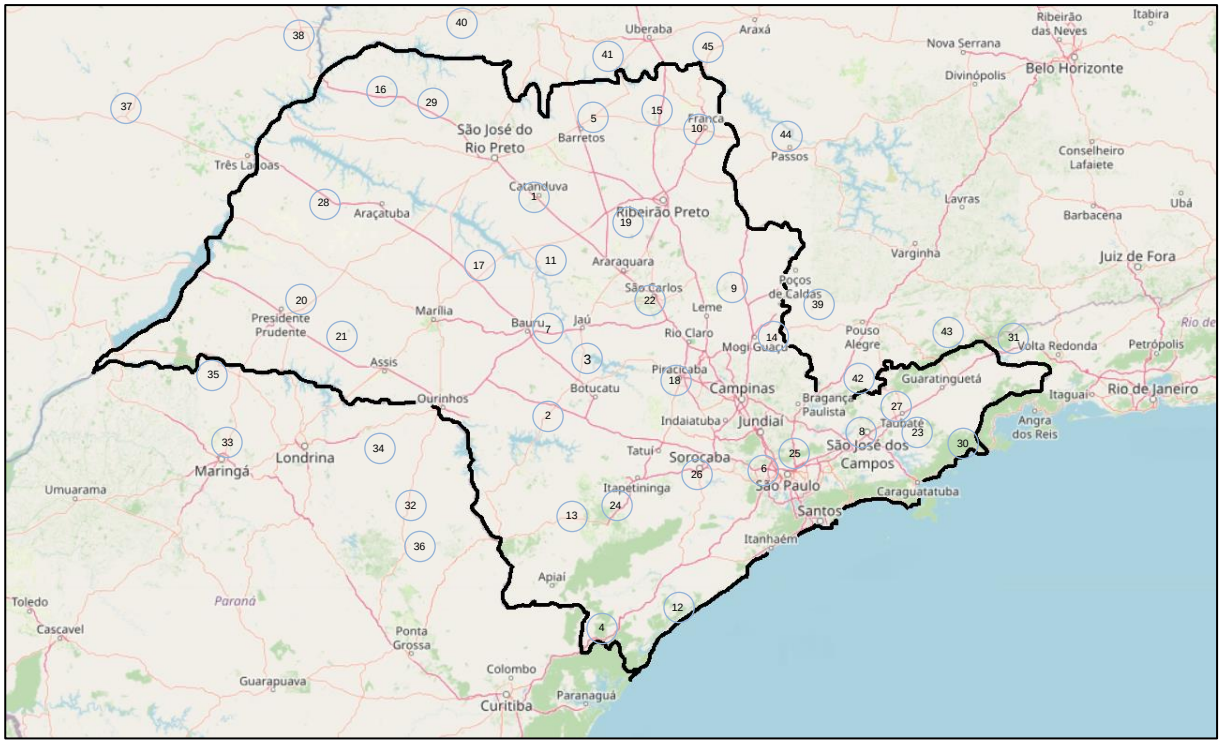
(continuação)

ID	Localidade	Estado	Latitude [°]	Longitude [°]	Altitude [m]
3	BARRA BONITA	SP	22°28'16"S	48°33'27"W	533,7
4	BARRA DO TURVO	SP	24°57'46"S	48°24'59"W	659,9
5	BARRETOS	SP	20°33'33"S	48°32'42"W	534,4
6	BARUERI	SP	23°31'26"S	46°52'10"W	776,5
7	BAURU	SP	22°21'29"S	49°1'44"W	636,2
8	CAMPOS DO JORDAO	SP	22°45'1"S	45°36'14"W	1.663,0
9	CASA BRANCA	SP	21°46'50"S	47°4'31"W	734,2
10	FRANCA	SP	20°35'4"S	47°22'57"W	1.002,8
11	IBITINGA	SP	21°51'20"S	48°47'59"W	496,8
12	IGUAPE	SP	24°40'18"S	47°32'45"W	2,7
13	ITAPEVA	SP	23°58'55"S	48°53'9"W	743,3
14	ITAPIRA	SP	22°24'54"S	46°48'19"W	634,9
15	ITUVERAVA	SP	20°21'35"S	47°46'31"W	610,6
16	JALES	SP	20°9'54"S	50°35'42"W	460,4
17	LINS	SP	21°39'58"S	49°44'5"W	460,7
18	PIRACICABA	SP	22°42'11"S	47°37'24"W	566,5
19	PRADOPOLIS	SP	21°20'18"S	48°6'50"W	540,4
20	PRESIDENTE PRUDENTE	SP	22°7'12"S	51°24'31"W	431,9
21	RANCHARIA	SP	22°22'22"S	50°58'29"W	398,8
22	SAO CARLOS	SP	21°58'49"S	47°53'2"W	859,3
23	SAO LUIS PARAITINGA	SP	23°13'42"S	45°25'1"W	862,3
24	SAO MIGUEL ARCANJO	SP	23°51'7"S	48°9'53"W	675,7
25	SAO PAULO - MIRANTE	SP	23°29'47"S	46°37'12"W	785,6
26	SOROCABA	SP	23°25'34"S	47°35'8"W	609,3
27	TAUBATE	SP	23°2'30"S	45°31'15"W	582,3
28	VALPARAISO	SP	21°19'9"S	50°55'49"W	381,9
29	VOTUPORANGA	SP	20°24'12"S	49°57'58"W	510,4
30	PARATY	RJ	23°13'25"S	44°43'37"W	3,0
31	RESENDE	RJ	22°27'5"S	44°26'42"W	438,8
31	RESENDE	RJ	22°27'5"S	44°26'42"W	438,8
32	JAPIRA	PR	23°46'24"S	50°10'50"W	692,9
33	MARINGA	PR	23°24'19"S	51°55'58"W	548,5
34	NOVA FATIMA	PR	23°24'55"S	50°34'40"W	664,3
35	PARANAPOEMA	PR	22°39'30"S	52°8'4"W	308,7
36	VENTANIA	PR	24°16'49"S	50°12'37"W	1.093,4
37	AGUA CLARA	MS	20°26'40"S	52°52'33"W	323,6
38	PARANAIBA	MS	19°41'44"S	51°10'54"W	408,1
39	CALDAS	MG	21°55'5"S	46°22'59"W	1.077,3
40	CAMPINA VERDE	MG	19°32'21"S	49°31'5"W	559,1
41	CONCEICAO DAS ALAGOAS	MG	19°59'9"S	48°9'6"W	572,5
42	MONTE VERDE	MG	22°51'42"S	46°2'36"W	1.544,9
43	PASSA QUATRO	MG	22°23'45"S	44°57'43"W	1.017,1
44	PASSOS	MG	20°44'43"S	46°38'2"W	781,7
45	SACRAMENTO	MG	19°52'31"S	47°26'3"W	913,1

Fonte: Autoria Própria a partir de dados INMET.

A Figura 1 mostra a disposição geográfica das estações meteorológicas apresentadas e descritas no Quadro 1.

Figura 1 – Mapa do Estado de São Paulo indicando a localização das 45 estações meteorológicas automáticas



Fonte: Autoria Própria a partir de dados INMET.

Para cada estação foram baixados, do site do Instituto Nacional de Meteorologia, as bases de dados horária, cobrindo o período considerado, em arquivos no formato *.csv*, totalizando quatrocentos e cinquenta arquivos (10 anos x 45 estações). Foi criada uma rotina em ***R*** (R CORE TEAM, 2022) para agregar e gerar os dados diários para as variáveis de interesse: irradiação solar global (R_s, MJ m^{-2} $hora^{-1}$), temperaturas máxima e mínima do ar (T_{max} e T_{min}, °C), umidades relativas máxima e mínima do ar (UR_{max} e UR_{min}, %) e velocidade do vento (u_2, m s^{-1}) medida a 2 metros de altura da superfície. Na sequência, com os valores diários das variáveis (R_s, T_{max}, T_{min}, UR_{max}, UR_{min}, u_2) e utilizando o modelo de Penman-Monteith (ALLEN et al., 1998), recomendado pela Food and Agriculture Organization, no boletim FAO-56, foi obtida uma base de dados diária da evapotranspiração de referência (ET_o) para esta região. Naturalmente, esta base inicial, uma matriz 45 estações (em linhas) por 3653 dias (em colunas), totalizando 164.385 elementos, apresentou dados faltantes, cerca de 9,45%, o que corresponde

a 15.531 dados faltantes no total, que foram integralmente preenchidos pelo *valor médio da coluna* correspondente a posição do dado ausente.

Esta base de dados completa foi utilizada para verificar, por meio de um *script* implementado no ambiente ***R***, o desempenho dos algoritmos NIPALS-PCA (ANDRECUT, 2009; WRIGHT, 2017), EM-PCA (JOSSE; HUSSON, 2016) e imputação pela média das colunas (IM) para o preenchimento de dados faltantes, realizado por meio de simulação para avaliar os métodos de imputação de *missings* na matriz de dados diários completa de ET_o.

O pacote *missMDA*[3] (JOSSE; HUSSON, 2016) do ambiente computacional *R-Gui* disponibiliza as funções para o cálculo do número de componentes (*estim_ncpPCA*)[4] e alguns métodos de imputação, entre eles o EM-PCA (função *imputePCA*)[5]. Para *NIPALS-PCA* foi utilizada a função *nipals*[6] do pacote *nipals*[7] (WRIGHT, 2017) implementado em ***R***.

Para definição do número de componentes a ser utilizada na imputação via PCA foi utilizada a validação cruzada (BRO et al., 2008; JOSSE; HUSSON, 2012) pelo método *kfold* (FUSHIKI, 2011; JUNG, 2017; MORENO-TORRES; SAEZ; HERRERA, 2012). A porcentagem de valores ausentes (*pNA*) é removida e estimada com um modelo EM-PCA usando o intervalo de dimensões [*ncp.min*, *ncp.max*]. Este processo foi repetido *nbsim* vezes. Cada célula é estimada usando a função *imputePCA*, ou seja, usando o algoritmo PCA iterativo (*EM cross-validation*). Para o número de componentes que resultar em um menor erro quadrático médio foi definido como o número de componentes para imputação.

Foram simulados cinco cenários de *missings* (10%, 20%, 30%, 40% e 50%), utilizando o mecanismo *Missing Completely at Random,* MCAR (LITTLE; RUBIN, 2002). Para criar cada cenário de *missings*, foi gerado de forma aleatória um conjunto de dados com algumas posições retiradas da base de dados de ET_o. Este

[3] http://127.0.0.1:17583/library/missMDA/html/00Index.html

[4] http://127.0.0.1:17583/library/missMDA/html/estim_ncpPCA.html

[5] http://127.0.0.1:17583/library/missMDA/html/imputePCA.html

[6] http://127.0.0.1:17583/library/nipals/html/nipals.html

[7] http://127.0.0.1:17583/library/nipals/html/00Index.html

procedimento inicia-se com a geração aleatória das sementes com a função *sample* do pacote básico do ***R***, para cada semente específica (*set.seed*) foi gerado as posições aleatórias dos *missings*, novamente com o comando *sample*, de acordo com uma taxa específica de dados faltantes. Para todas as posições foram substituídos os valores observados pelo valor *"NA"*, ou seja, dado ausente. Para cada cenário de *missings* especificado foram feitas as imputações pelos procedimentos NIPALS-PCA, EM-PCA e IM. Este ciclo foi executado por 100 vezes e, ao final, foi calculado os indicativos médios de desempenho.

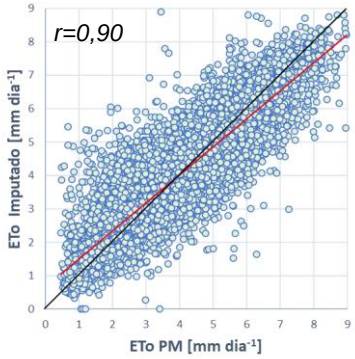
Para avaliação do desempenho estatístico dos procedimentos de imputação NIPALS-PCA, EM-PCA e IM foram utilizados os indicativos: coeficiente de correlação (STANTON, 2001) entre os valores imputados e estimados (*r*), erro absoluto médio – *Mean Absoluto Error (MAE),* erro absoluto médio percentual – *Mean Absoluto Percentual Error (pMAE),* erro quadrático médio – *Mean Squared Error (MSE),* raiz do erro quadrático médio percentual (JAMIESON; PORTER; WILSON, 1991) – *Root Mean Squared Percentual Error (pRMSE),* índice de concordância de Willmott (WILLMOTT et al., 1985) (*d*) e o coeficiente de confiança (CAMARGO; SENTELHAS, 1997) (*c*).

3.3 RESULTADOS E DISCUSSÃO

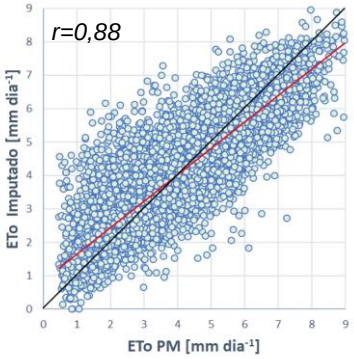
Nos cenários de dados ausentes (10%, 20%, 30%, 40%, 50%) simulados na base de dados de ET_o, utilizou-se a Análise de Componente Principais nos algoritmos NIPALS (com 45 componentes) e EM (com 5 componentes) e a imputação simples pela média (IM) para reconstrução da base de dados e, consequentemente, obter os valores estimados dos *missings* simulados. A Figura 2 (a-o) apresenta a dispersão e o coeficiente de correlação entre os valores imputados (por NIPALS-PCA, EM-PCA e IM) e os valores observados de ET_o de uma simulação. Observa-se que as dispersões dos valores imputados pelos diferentes procedimentos e os valores observados de ET_o, reta em vermelho, nas condições de validação, estão em concordância linear com a reta ideal (45º), em preto, com coeficientes de correlação (r) entre os valores observados e imputados por IM, variando de 0,75 (com 50%, 40% e 30% de *missings*) a 0,76 (com 20 e 10% de *missings*).

Figura 2. Dispersão entre os valores observados de ET_o e os imputados pelos procedimentos NIPALS-PCA, EM-PCA e IM

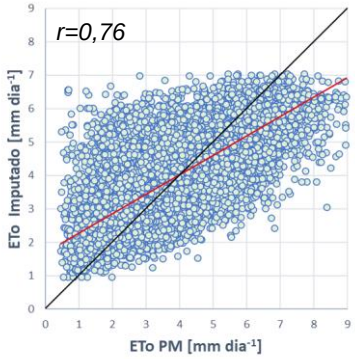
a) NIPALS-PCA (10% missings)



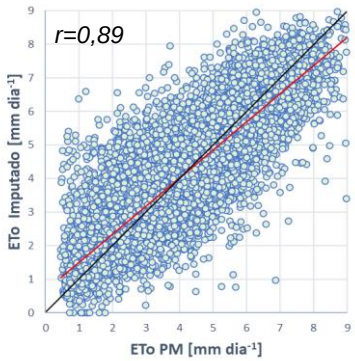
b) EM-PCA (10% missings)



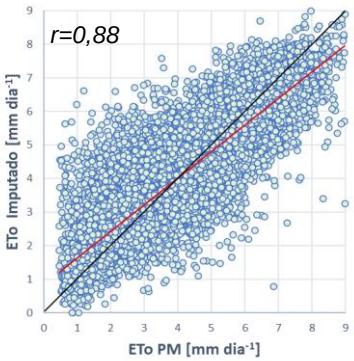
c) IM (10% missings)



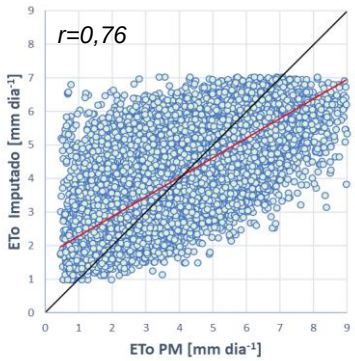
d) NIPALS-PCA (20% missings)



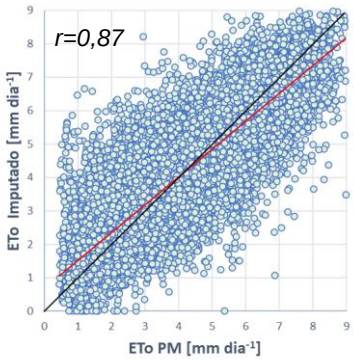
e) EM-PCA (20% missings)



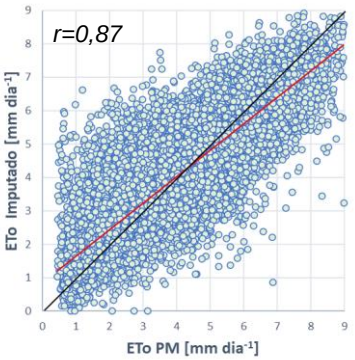
f) IM (20% missings)



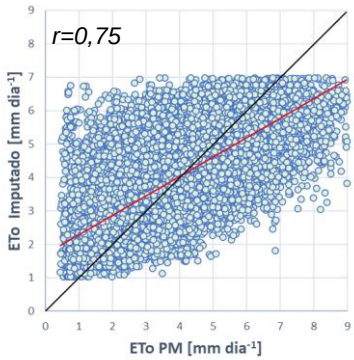
g) NIPALS-PCA (30% missings)



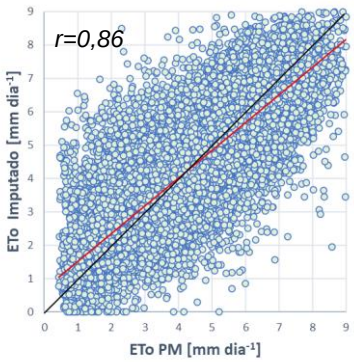
h) EM-PCA (30% missings)



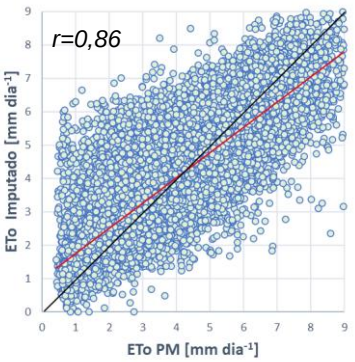
i) IM (30% missings)



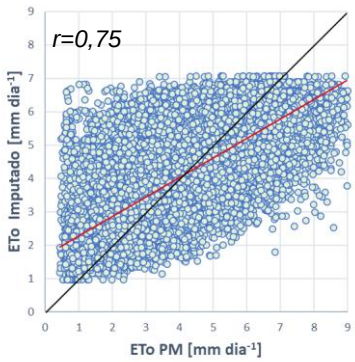
j) NIPALS-PCA (40% missings)



k) EM-PCA (40% missings)



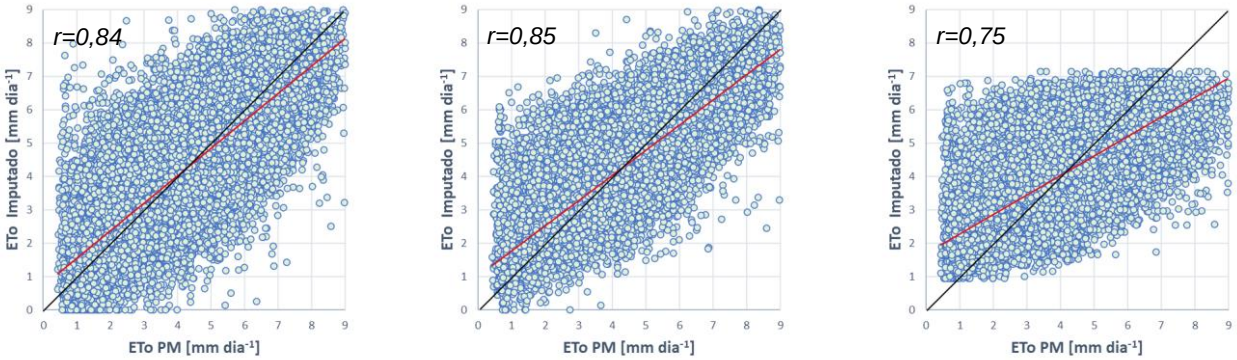
l) IM (40% missings)



(contínua)

(continuação)

m) NIPALS-PCA (50% missings) *n) EM-PCA (50% missings)* *o) IM (50% missings)*



Fonte: Autoria Própria.

Para valores imputados pelo NIPALS-PCA os coeficientes de correlação variaram de 0,84 (com 50% de *missings*) a 0,90 (com 10% *de missings*), enquanto ao considerar os valores imputados pelo EM-PCA os coeficientes de correlação variaram de 0,85 (com 50% de *missings*) a 0,88 (com 10% *de missings*). Independentemente da quantidade de *missings* percebe-se um menor espalhamento e correlações maiores para a NIPALS-PCA, EM-PCA, em média cerca de 0,87, em relação ao método de imputação pela média (IM), com coeficiente de correlação médio de 0,75.

Um resumo descritivo (média, valores mínimo e máximo) dos resultados obtidos pelos indicadores de desempenho nas simulações realizadas para cada cenário e procedimento de imputação pode ser observado na Tabela 1. Considerando os valores máximos e mínimos observados para cada método nos diferentes indicadores, o método EM-PCA apresenta uma menor amplitude em relação NIPALS-PCA, exceto para o caso de 30% *missings*. Isto mostra uma maior precisão do EM-PCA, fato confirmado na Figura 3(f).

Tabela 1 – Indicadores de desempenho nos cenários de dados ausentes (10%, 20%, 30%, 40%, 50%)

Indicadores	% *missings*	NIPALS-PCA	EM-PCA	IM
	10	**0,90** [0,89; 0,91]	**0,88** [0,88; 0,89]	**0,76** [0,75; 0,76]
	20	**0,89** [0,88; 0,89]	**0,88** [0,88; 0,89]	**0,76** [0,75; 0,76]
r	***30***	**0,87** [0,87; 0,88]	**0,87** [0,86; 0,88]	**0,75** [0,75; 0,76]
	40	**0,86** [0,85; 0,86]	**0,86** [0,86; 0,86]	**0,75** [0,75; 0,76]
	50	**0,84** [0,83; 0,84]	**0,85** [0,85; 0,86]	**0,75** [0,75; 0,75]

(contínua)

(continuação)

Indicadores	% *missings*	NIPALS-PCA	EM-PCA	IM
MAE [mm dia^{-1}]	***10***	**0,50** [0,49; 0,51]	**0,53** [0,52; 0,54]	**0,75** [0,73; 0,76]
	20	**0,53** [0,53; 0,54]	**0,54** [0,54; 0,55]	**0, 75** [0,74; 0,76]
	30	**0,57** [0,56; 0,57]	**0,58** [0,55; 0,59]	**0, 75** [0,75; 0,76]
	40	**0,61** [0,60; 0,61]	**0,60** [0,59; 0,60]	**0, 76** [0,75; 0,76]
	50	**0,65** [0,65; 0,66]	**0,60** [0,60; 0,61]	**0, 76** [0,76; 0,77]
pMAE [%]	***10***	**15,44** [14,95; 15,89]	**16,96** [16,53; 17,43]	**24,65** [23,90; 25,43]
	20	**16,40** [16,04; 16,72]	**17,15** [16,83; 17,54]	**24,74** [24,26; 25,23]
	30	**17,46** [17,16; 17,72]	**18,29** [17,18; 18,98]	**24,84** [24,42; 25,23]
	40	**18,59** [18,29; 18,85]	**18,89** [18,63; 19,15]	**24,93** [24,60; 25,26]
	50	**19,92** [19,65; 20,37]	**19,13** [18,91; 19,29]	**25,06** [24,76; 25,28]
MSE [mm^2 dia^{-2}]	***10***	**0,48** [0,45; 0,51]	**0,54** [0,52; 0,56]	**1,07** [1,03; 1,11]
	20	**0,55** [0,53; 0,58]	**0,56** [0,54; 0,58]	**1,07** [1,05; 1,10]
	30	**0,62** [0,60; 0,64]	**0,62** [0,57; 0,66]	**1,08** [1,06; 1,10]
	40	**0,70** [0,68; 0,72]	**0,66** [0,65; 0,67]	**1,08** [1,06; 1,11]
	50	**0,80** [0,78; 0,83]	**0,69** [0,67; 0,69]	**1,09** [1,08; 1,11]
pRMSE [%]	***10***	**17,16** [16,60; 17,60]	**18,18** [17,78; 18,55]	**25,50** [25,03; 26,01]
	20	**18,27** [17,96; 18,70]	**18,41** [18,15; 18,76]	**25,55** [25,24; 25,89]
	30	**19,41** [19,18; 19,65]	**19,46** [18,58; 20,01]	**25,61** [25,41; 25,89]
	40	**20,62** [20,31; 21,01]	**20,03** [19,85; 20,19]	**25,67** [25,45; 25,94]
	50	**22,07** [21,77; 22,41]	**20,30** [20,14; 20,45]	**25,77** [25,58; 25,96]
d	***10***	**0,95 [0,94; 0,95]**	**0,94 [0,93; 0,94]**	**0,85 [0,84; 0,86]**
	20	**0,94** [0,94; 0,94]	**0,94** [0,93; 0,94]	0,**85** [0,85; 0,86]
	30	**0,93** [0,93; 0,93]	**0,93** [0,92; 0,93]	0,**85** [0,85; 0,85]
	40	**0,92** [0,92; 0,93]	**0,92** [0,92; 0,92]	0,**85** [0,85; 0,85]
	50	**0,91** [0,91; 0,92]	**0,92** [0,92; 0,92]	**0,85** [0,85; 0,85]
c	***10***	**0,85** [0,84; 0,86]	**0,83** [0,82; 0,84]	**0,65** [0,64; 0,66]
	20	**0,83** [0,82; 0,84]	**0,82** [0,82; 0,83]	**0,64** [0,64; 0,65]
	30	**0,81** [0,81; 0,82]	**0,80** [0,79; 0,82]	**0,64** [0,64; 0,65]
	40	**0,79** [0,78; 0,80]	**0,79** [0,79; 0,79]	**0,64** [0,64; 0,65]
	50	**0,77** [0,76; 0,77]	**0,79** [0,78; 0,79]	**0,64** [0,64; 0,64]

Fonte: Autoria Própria.
*Média; [Valor mínimo, Valor máximo]

Em relação ao coeficiente de correlação, não se observou diferenças significativas entre os cenários de *missings* para o método IM. Para todos os cenários de *missings*, para IM observou-se um valor de coeficiente de correlação variando de 0,75 a 0,76, exceto para 50% de *missings*, no qual se observou que o valor praticamente não variou entre as simulações realizadas.

Observou-se um gradual acréscimo no *pMAE* e *pRMSE* à medida que são acrescentadas um maior número de valores ausentes, indicando uma piora no desempenho dos procedimentos considerados. Para o cenário com 10% de *missings*, a NIPALS-PCA obteve o menor *pMAE* (15,44%), seguido da EM-PCA (16,96%), enquanto a IM obteve um *pMAE* igual a 24,65%. No cenário com 50% de

missings ocorre uma inversão de desempenho, com menor *pMSE* (19,13%) para o EM-PCA, seguido do NIPALS-PCA (19,92%). Considerando a escala classificação para os diferentes intervalos de *pRMSE*, as abordagens NIPALS-PCA e EM-PCA apresentam bons resultados (10% ≤ pRMSE < 20%) na imputação de valores ausentes. Com destaque para o método NIPALS-PCA nos cenários de 10%, 20% e 30% e EM-PCA para os cenários de 40% e 50%.

Para índice de concordância de Willmott (d), destaque para os métodos NIPALS-PCA e EM-PCA com um índice de concordância de 93% em relação a 85% para IM. Os métodos NIPALS-PCA e EM-PCA apresentaram o valor médio de 0,81 para o coeficiente de confiança (c), considerando a classificação dadas pelos pesquisadores Camargo e Sentelhas (1997), são modelos de estimação *muito bom*, enquanto o método IM que apresentou o valor médio de 0,64 classificado com um modelo de estimação *bom*.

Comparando com os resultados obtidos pelos pesquisadores Martí e Zarzo (MARTÍ; ZARZO, 2012), modelando 30 estações meteorológicas localizadas na região de Valência na Espanha, no período de 2000 a 2007, observa-se resultados menores nos indicadores de desempenho em relação aos encontrados nesta pesquisa, como evidenciados na Tabela 2, para o cenário de 10% de *missings*.

Tabela 2 – Valores de alguns indicativos de desempenho estatístico

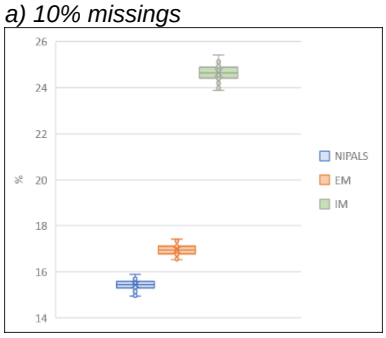
Pesquisadores	**Localidade**	**MSE**	**MAE**	**pMAE**	**r**
		$mm^2\ dia^{-2}$	**$mm\ dia^{-1}$**	**%**	
Martí e Zarzo (2012)	Valência - Espanha	0,11	0,24	9,20	0,98
Presente pesquisa	São Paulo - Brasil	0,48	0,50	15,44	0,90

Fonte: Autoria Própria.

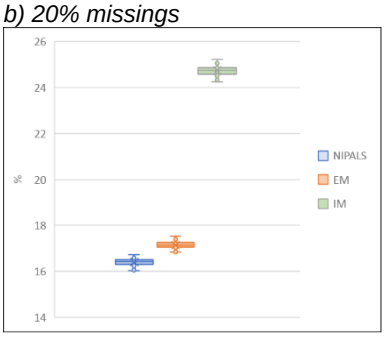
Na Figura 3(a-e), pode-se verificar que os métodos EM-PCA e NIPALS-PCA são similares, enquanto o IM se distancia, apresentando resultados com maiores desvios. Resultado também observado quando considerado uma média de todos os cenários, Figura 3(f). Permite ainda dizer que o método NIPALS-PCA é preferível na maioria dos cenários, exceto 50%, em relação ao EM-PCA, pois apresenta menor erro absoluto médio percentual.

Figura 3 – Indicativo estatístico erro absoluto médio percentual

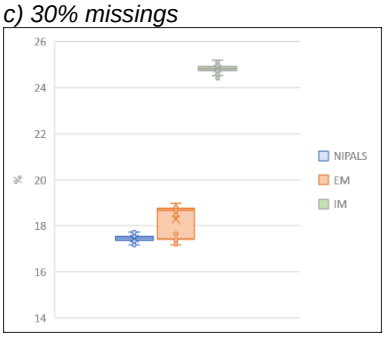
a) 10% missings



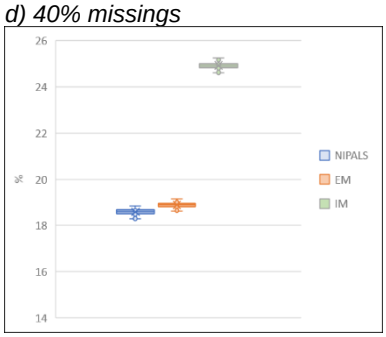
b) 20% missings



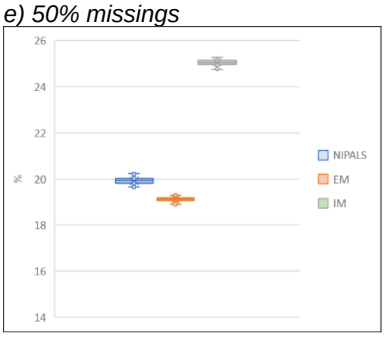
c) 30% missings



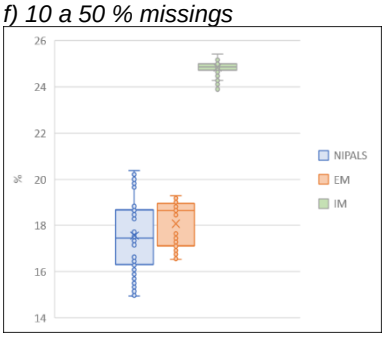
d) 40% missings



e) 50% missings



f) 10 a 50 % missings



Fonte: Autoria Própria.

Os pesquisadores Dray e Josse (2014) fazem uma revisão de alguns métodos de imputação via PCA, aplicados em dados da área de ecologia. Sugerem o uso do EM-PCA em detrimento ao NIPALS-PCA, pelo fato de dificuldade de convergência. Fato não observado neste presente trabalho.

É importante destacar que as simulações feitas neste trabalho utilizaram o mecanismo *Missing Completely at Random* (MCAR) e, portanto, os resultados apresentados podem não ser generalizáveis para as situações nos quais os valores ausentes ocorreram de maneira não aleatória ou tendenciosa. Além disso, este estudo utilizou a média para completar a base inicial, privilegiando este método nas simulação realizadas e, portanto, as diferenças entre os desempenhos dos métodos multivariados via PCA em relação a imputação pela média podem ser maiores.

3.4 CONCLUSÕES

Neste capítulo verificou-se a performance de procedimentos multivariados alternativos de análise de componentes principais, em conjuntos com os algoritmos NIPALS e EM, e a imputação simples pela média (IM) para reconstrução de uma base de dados de evapotranspiração de referência, de alta dimensão e amostra reduzida, e, consequentemente, obtenção dos valores estimados dos *missings* simulados, nos cenários de dados ausentes de 10%, 20%, 30%, 40% e 50%. Para um período de 2012 a 2021, considerando estações meteorológicas automáticas da região de São Paulo, Brasil.

A PCA apresentou-se como uma ferramenta útil para estimativas de valores ausentes na situação em que o tamanho amostral é reduzido em relação ao número de variáveis, pois este estudo focou na imputação de *missings* em uma base de dados de evapotranspiração considerando os dias de medição da ET_o como variáveis correlacionadas (3653 colunas), as quais foram medidas em 45 estações meteorológicas automáticas (linhas).

Na comparação do desempenho estatístico entre as técnicas utilizadas, verificou-se que o melhor desempenho ficou entre NIPALS-PCA e EM-PCA, a depender da percentagem de *missings*, tendo baixo desempenho a imputação simples pelas médias das colunas.

Com base na classificação do indicativo estatístico da base de validação dos modelos de imputação NIPALS-PCA, EM-PCA e IM, há indicativos que os mesmos são aptos à estimativa de *missings* da evapotranspiração de referência, com destaque para os modelos de imputação PCA nos algoritmos NIPALS e EM.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

ALLEN, R. G. et al. Evapotranspiration information reporting: I. Factors governing measurement accuracy. **Agricultural Water Management**, v. 98, n. 6, p. 899–920, 1 abr. 2011.

ANDRECUT, M. Parallel GPU implementation of iterative PCA algorithms. **Journal of computational biology : a journal of computational molecular cell biology**, v. 16, n. 11, p. 1593–1599, 2009.

BRO, R. et al. Cross-validation of component models: A critical look at current methods. **Analytical and Bioanalytical Chemistry**, v. 390, n. 5, p. 1241–1251, 24 mar. 2008.

BUCIOR-KWACZYŃSKA, A. The Possibility of Applying the EM-PCAProcedure to Lake Water. **Polish Journal of Environmental Studies**, v. 27, n. 1, p. 19–30, 2 jan. 2018.

CAI, J. B. et al. Simulation of the soil water balance of wheat using daily weather forecast messages to estimate the reference evapotranspiration. **Hydrology and Earth System Sciences**, v. 13, n. 7, p. 1045–1059, 9 jul. 2009.

CAMARGO, Â. P. DE; SENTELHAS, P. C. Avaliação do Desempenho de Diferentes Métodos de Estimativa da Evapotranspiração Potencial no Estado de São Paulo no Brasil. **Revista Brasileira de Agrometeorologia**, v. 5, n. 1, p. 89–97, 1997.

DE KETELAERE, B.; HUBERT, M.; SCHMITT, E. Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. **Journal of Quality Technology**, v. 47, n. 4, p. 318–335, 2015.

DE LA FUENTE, R. L. N.; GARCÍA-MUÑOZ, S.; BIEGLER, L. T. An efficient nonlinear programming strategy for PCA models with incomplete data sets. **Journal of Chemometrics**, v. 24, n. 6, p. 301–311, 1 jun. 2010.

DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal ofthe Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977.

DRAY, S.; JOSSE, J. Principal component analysis with missing values: a comparative survey of methods. **Plant Ecology**, v. 216, n. 5, p. 657–667, 1 maio 2014.

ESHGHI, P. Dimensionality choice in principal components analysis via cross-validatory methods. **Chemometrics and Intelligent Laboratory Systems**, v. 130, p. 6–13, 15 jan. 2014.

FUSHIKI, T. Estimation of prediction error by using K-fold cross-validation. **Statistics and Computing**, v. 21, n. 2, p. 137–146, 1 abr. 2011.

GARCÍA-DIEGO, F. J.; ZARZO, M. Microclimate monitoring by multivariate statistical control: The renaissance frescoes of the Cathedral of Valencia (Spain). **Journal of Cultural Heritage**, v. 11, n. 3, p. 339–344, 2010.

GONSAGA DE CARVALHO, L. et al. Evapotranspiração de referência: uma abordagem atual de diferentes métodos de estimativa. **Pesquisa Agropecuária Tropical**, v. 41, n. 3, p. 456–465, 6 jul. 2011.

HART, Q. J. et al. Daily reference evapotranspiration for California using satellite imagery and weather station measurement interpolation. **https://doi-org.ez87.periodicos.capes.gov.br/10.1080/10286600802003500**, v. 26, n. 1, p. 19–33, 2009.

HASAN, M. K. et al. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). **Informatics in Medicine Unlocked**, v. 27, p. 1–23, 1 jan. 2021.

HOWLEY, T. et al. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. **Knowledge-Based Systems**, v. 19, n. 5, p. 363–370, set. 2006.

INMET. **Instituto Nacional de Meteorologia**. Disponível em: <https://portal.inmet.gov.br/servicos/bdmep-dados-históricos>. Acesso em: 27 dez. 2022.

JAMIESON, P. D.; PORTER, J. R.; WILSON, D. R. A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. **Field Crops Research**, v. 27, n. 4, p. 337–350, 1991.

JOSSE, J.; HUSSON, F. Selecting the number of components in principal component analysis using cross-validation approximations. **Computational Statistics and Data Analysis**, v. 56, n. 6, p. 1869–1879, 2012.

JOSSE, J.; HUSSON, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. **Journal of Statistical Software**, v. 70, p. 1–31, 4 abr. 2016.

JOSSE, J.; HUSSON, F.; PAGÈS, J. Gestion des données manquantes en Analyse en Composantes Principales. **Journal de la Société Française de Statistique**, v. 150, p. 2, 2009.

JOSSE, J.; PAGÈS, J.; HUSSON, F. Multiple imputation in principal component analysis. **Advances in Data Analysis and Classification**, v. 5, n. 3, p. 231–246, 2011.

JUNG, Y. Multiple predicting K-fold cross-validation for model selection. **https://doi.org/10.1080/10485252.2017.1404598**, v. 30, n. 1, p. 197–215, 2 jan. 2017.

KISI, O. et al. Modeling reference evapotranspiration using a novel regression-based method: radial basis M5 model tree. **Theoretical and Applied Climatology**, v. 145, n. 1–2, p. 639–659, 1 jul. 2021.

LITTLE, R. J. A.; RUBIN, D. B. **Single Imputation Methods**. [s.l.] John Wiley & Sons, Ltd, 2002.

MALAN, L. et al. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. **Nutrition Research**, v. 75, p. 67–76, 1 mar. 2020.

MARIN, F. R. et al. Revisiting the crop coefficient–reference evapotranspiration procedure for improving irrigation management. **Theoretical and Applied Climatology**, v. 138, n. 3–4, p. 1785–1793, 1 nov. 2019.

MARTÍ, P.; ZARZO, M. Multivariate statistical monitoring of ETo: A new approach for estimation in nearby locations using geographical inputs. **Agricultural and Forest Meteorology**, v. 152, n. 1, p. 125–134, 2012.

MORENO-TORRES, J. G.; SAEZ, J. A.; HERRERA, F. Study on the impact of partition-induced dataset shift on k-fold cross-validation. **IEEE Transactions on Neural Networks and Learning Systems**, v. 23, n. 8, p. 1304–1312, 2012.

NILASHI, M. et al. Early Diagnosis of Parkinson's Disease: A Combined Method Using Deep Learning and Neuro-Fuzzy Techniques. **Computational Biology and Chemistry**, p. 107788, fev. 2022.

ONNABI MILANI, A. et al. Evaluating direct and indirect estimation methods of reference evapotranspiration (ETo). 2007.

PATEL, N.; SIVANATHAN, K.; MHASKAR, P. Polymethyl Methacrylate Quality Modeling with Missing Data Using Subspace Based Model Identification. **Processes 2021, Vol. 9, Page 1691**, v. 9, n. 10, p. 1691, 22 set. 2021.

PODANI, J. et al. Principal component analysis of incomplete data – A simple solution to an old problem. **Ecological Informatics**, v. 61, p. 101235, 1 mar. 2021.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**Vienna, AustriaR Foundation for Statistical Computing, , 2022. Disponível em: <https://www.r-project.org/>

RANA, G.; KATERJI, N. Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. **European Journal of Agronomy**, v. 13, n. 2–3, p. 125–153, 1 jul. 2000.

STANTON, J. M. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. **Journal of Statistics Education**, v. 9, n. 3, 2001.

TERINK, W.; IMMERZEEL, W. W.; DROOGERS, P. Climate change projections of precipitation and reference evapotranspiration for the Middle East and Northern Africa until 2050. **International Journal of Climatology**, v. 33, n. 14, p. 3055–3072, 30 nov. 2013.

VYAS, M. et al. State-of-charge prediction of lithium ion battery through multivariate adaptive recursive spline and principal component analysis. **Energy Storage**, v. 3, n. 2, p. e147, 1 abr. 2021.

WANG, J. et al. Development of Monthly Reference Evapotranspiration Machine Learning Models and Mapping of Pakistan—A Comparative Study. **Water 2022, Vol. 14, Page 1666**, v. 14, n. 10, p. 1666, 23 maio 2022.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. **Statistics in Medicine**, v. 30, n. 4, p. 377–399, 2011.

WILLMOTT, C. J. et al. Statistics for the evaluation and comparison of models. **Journal of Geophysical Research**, v. 90, n. C5, p. 8995, 1985.

WRIGHT, K. **The NIPALS algorithm**. Disponível em: <https://cran.r-project.org/web/packages/nipals/vignettes/nipals_algorithm.html>. Acesso em: 14 dez. 2022.

XIE, C. et al. Recovery Method for Missing Sensor Data in Multi-Sensor Based Walking Recognition System. **8th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, CYBER 2018**, p. 558–563, 10 abr. 2019.

YANG, Q. et al. MultiDA: Chemometric software for multivariate data analysis based on Matlab. **Chemometrics and Intelligent Laboratory Systems**, v. 116, p. 1–8, jul. 2012.

CONSIDERAÇÕES FINAIS

Esta pesquisa teve por objetivo aprofundar-se na temática sobre a aplicação da análise de componentes principais na área agrícola, mais especificamente na imputação de dados ausentes em séries temporais de variáveis meteorológicas, considerando bases de dados no cenário de alta dimensão e amostra reduzida, com diferentes percentuais de *missings*.

Para isso, buscou-se comparar o desempenho de procedimentos multivariados alternativos de análise de componentes principais na imputação de dados ausentes em séries temporais de evapotranspiração, bases de dados em que o número de variáveis supera o número de amostras com diferentes percentuais de *missings*, ou seja, amostras independentes de 45 estações meteorológicas (linhas) com 3653 observações diárias correlacionadas (colunas). Com uma descrição detalhada do planejamento para estudos de simulação em MVI, com foco no banco e tipo de dados, mecanismo e taxa de falta, técnica de imputação e método de avaliação de desempenho. E uma análise bibliométrica no tema *Missing Value Imputation*. O resultado esperado foi a publicação de artigos em revistas científicas de impacto na área de ciências agrárias.

Por meio de uma análise bibliométrica, utilizando a base dados *Web Of Science*, considerando o período de 1940 a 2022, foram levantados em 2 de janeiro de 2023, uma amostra de 19.745 trabalhos. Destas publicações, aproximadamente 70% foram publicados na última década. Dos países envolvidos, destaque para Estados Unidos, China e Inglaterra em termos de universidades, pesquisadores e periódicos de maior impacto na área analisada, MVI. A distribuição temporal da ocorrência das palavras-chave, caracterizada pelo início com o *Algoritmo EM*, passando por *Imputação Múltipla* e atingindo *Aprendizado de Máquina*, técnicas de Inteligência Artificial.

Verificou-se a performance de procedimentos multivariados alternativos de análise de componentes principais, em conjunto com os algoritmos NIPALS e EM, e a imputação simples pela média (IM) para reconstrução de uma base de dados de evapotranspiração, de alta dimensão e amostra reduzida, e, consequentemente, obtenção dos valores estimados dos *missings* simulados, nos cenários de dados ausentes de 10%, 20%, 30%, 40% e 50%. Para um período de 2012 a 2021,

considerando quarenta e cinco estações meteorológicas automáticas da região de São Paulo, Brasil. O mecanismo de falta utilizado foi o *Missing Completely at Random*, como atenção, os resultados das simulações apresentados podem não ser generalizáveis para as situações nos quais os valores ausentes ocorreram de maneira não aleatória ou tendenciosa.

Este estudo gerou oportunidades de pesquisa e inovação na área agrícola com as divulgações dos resultados em congressos e revistas científicas nacionais e internacionais, além de contribuir com a disseminação e ampliação da utilização adequada dos procedimentos multivariados de análise de dados agronômicos.

REFERÊNCIAS

AHN, J. et al. The high-dimension, low-sample-size geometric representation holds under mild conditions. **Biometrika**, v. 94, n. 3, p. 760–766, 2007.

ALLEN, R. G. et al. Evapotranspiration information reporting: I. Factors governing measurement accuracy. **Agricultural Water Management**, v. 98, n. 6, p. 899–920, 1 abr. 2011.

AYESHA, S.; HANIF, M. K.; TALIB, R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. **Information Fusion**, v. 59, n. January, p. 44–58, 2020.

BUCIOR-KWACZYŃSKA, A. The Possibility of Applying the EM-PCAProcedure to Lake Water. **Polish Journal of Environmental Studies**, v. 27, n. 1, p. 19–30, 2 jan. 2018.

CAI, J. B. et al. Simulation of the soil water balance of wheat using daily weather forecast messages to estimate the reference evapotranspiration. **Hydrology and Earth System Sciences**, v. 13, n. 7, p. 1045–1059, 9 jul. 2009.

CALANCA, P. Weather Forecasting Applications in Agriculture. **Encyclopedia of Agriculture and Food Systems**, p. 437–449, 1 jan. 2014.

CHEN, Y.; WIESEL, A.; HERO, A. O. Robust shrinkage estimation of high-dimensional covariance matrices. **IEEE Transactions on Signal Processing**, v. 59, n. 9, p. 4097–4107, 2011.

DE KETELAERE, B.; HUBERT, M.; SCHMITT, E. Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. **Journal of Quality Technology**, v. 47, n. 4, p. 318–335, 2015.

DE LA FUENTE, R. L. N.; GARCÍA-MUÑOZ, S.; BIEGLER, L. T. An efficient nonlinear programming strategy for PCA models with incomplete data sets. **Journal of Chemometrics**, v. 24, n. 6, p. 301–311, 1 jun. 2010.

DEMPSTER, A. P.; LAIRD, N. M. ; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal ofthe Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977.

ESHGHI, P. Dimensionality choice in principal components analysis via cross-validatory methods. **Chemometrics and Intelligent Laboratory Systems**, v. 130, p. 6–13, 15 jan. 2014.

ESTÉVEZ, J.; GAVILÁN, P.; GIRÁLDEZ, J. V. Guidelines on validation procedures for meteorological data from automatic weather stations. **Journal of Hydrology**, v. 402, n. 1–2, p. 144–154, 13 maio 2011.

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

GARCÍA-DIEGO, F. J.; ZARZO, M. Microclimate monitoring by multivariate statistical control: The renaissance frescoes of the Cathedral of Valencia (Spain). **Journal of Cultural Heritage**, v. 11, n. 3, p. 339–344, 2010.

GONSAGA DE CARVALHO, L. et al. Evapotranspiração de referência: uma abordagem atual de diferentes métodos de estimativa. **Pesquisa Agropecuária Tropical**, v. 41, n. 3, p. 456–465, 6 jul. 2011.

HART, Q. J. et al. Daily reference evapotranspiration for California using satellite imagery and weather station measurement interpolation. **https://doi-org.ez87.periodicos.capes.gov.br/10.1080/10286600802003500**, v. 26, n. 1, p. 19–33, 2009.

HASAN, M. K. et al. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). **Informatics in Medicine Unlocked**, v. 27, p. 1–23, 1 jan. 2021.

HONG, S. et al. Sample size in factor analysis. **Psychological Methods**, v. 4, n. 1, p. 84–99, 1999.

HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology**, v. 103, n. 1–2, p. 137–157, 1 jun. 2000.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417–441, set. 1933.

HOWLEY, T. et al. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. **Knowledge-Based Systems**, v. 19, n. 5, p. 363–370, set. 2006.

HOYLE, D. C. Automatic PCA dimension selection for high dimensional data and small sample sizes. **Journal of Machine Learning Research**, v. 9, p. 2733–2759, 2008.

JOHNSTONE, I. M.; PAUL, D. PCA in High Dimensions: An Orientation. **Proceedings of the IEEE**, v. 106, n. 8, p. 1277–1292, 2018.

JOSSE, J.; HUSSON, F. Selecting the number of components in principal component analysis using cross-validation approximations. **Computational Statistics and Data Analysis**, v. 56, n. 6, p. 1869–1879, 2012.

JUNG, S.; SEN, A.; MARRON, J. S. Boundary behavior in High Dimension, Low Sample Size asymptotics of PCA. **Journal of Multivariate Analysis**, v. 109, p. 190–203, 2012.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, 1 jun. 2004.

KISI, O. et al. Modeling reference evapotranspiration using a novel regression-based method: radial basis M5 model tree. **Theoretical and Applied Climatology**, v. 145, n. 1–2, p. 639–659, 1 jul. 2021.

MALAN, L. et al. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. **Nutrition Research**, v. 75, p. 67–76, 1 mar. 2020.

MARIN, F. R. et al. Revisiting the crop coefficient–reference evapotranspiration procedure for improving irrigation management. **Theoretical and Applied Climatology**, v. 138, n. 3–4, p. 1785–1793, 1 nov. 2019.

MARTÍ, P.; GASQUE, M. Ancillary data supply strategies for improvement of temperature-based ETo ANN models. **Agricultural Water Management**, v. 97, n. 7, p. 939–955, 2010.

MARTÍ, P.; ZARZO, M. Multivariate statistical monitoring of ETo: A new approach for estimation in nearby locations using geographical inputs. **Agricultural and Forest Meteorology**, v. 152, n. 1, p. 125–134, 2012.

MINGOTI, S. A. **Análise de dados através de Métodos de Estatística Multivariada: uma abordagem Aplicada**. Belo Horizonte: Editora UFMG, 2005.

MULLER, K. E. et al. Limitations of High Dimension , Low Sample Size Principal Components for Gaussian Data. **Journal of the American Statistical Association**, n. February, 2008.

NILASHI, M. et al. Early Diagnosis of Parkinson's Disease: A Combined Method Using Deep Learning and Neuro-Fuzzy Techniques. **Computational Biology and Chemistry**, p. 107788, fev. 2022.

ONNABI MILANI, A. et al. Evaluating direct and indirect estimation methods of reference evapotranspiration (ETo). 2007.

PATEL, N.; SIVANATHAN, K.; MHASKAR, P. Polymethyl Methacrylate Quality Modeling with Missing Data Using Subspace Based Model Identification. **Processes 2021, Vol. 9, Page 1691**, v. 9, n. 10, p. 1691, 22 set. 2021.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, p. 559–572, 1901.

RANA, G.; KATERJI, N. Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. **European Journal of Agronomy**, v. 13, n. 2–3, p. 125–153, 1 jul. 2000.

ROTH, P. L. MISSING DATA: A CONCEPTUAL REVIEW FOR APPLIED PSYCHOLOGISTS. **Personnel Psychology**, v. 47, n. 3, p. 537–560, 1 set. 1994.

SHAUKAT, S. S.; RAO, T. A.; KHAN, M. A. Impact of sample size on principal component analysis ordination of an environmental data set: Effects on eigenstructure. **Ekologia Bratislava**, v. 35, n. 2, p. 173–190, 2016.

SHEN, D. et al. the Statistics and Mathematics of High. v. 26, p. 1747–1770, 2016.

SHEN, D.; SHEN, H.; MARRON, J. S. Consistency of sparse PCA in High Dimension, Low Sample Size contexts. **Journal of Multivariate Analysis**, v. 115, p. 317–333, 2013.

SIDDIQUI, K. Heuristics for sample size determination in multivariate statistical techniques. **World Applied Sciences Journal**, v. 27, n. 2, p. 285–287, 2013.

STIGTER, C. J. From basic agrometeorological science to agrometeorological services and information for agricultural decision makers: A simple conceptual and diagnostic framework. **Agricultural and Forest Meteorology**, v. 142, n. 2–4, p. 91–95, 12 fev. 2007.

STRIKE, K.; EMAM, K. EL; MADHAVJI, N. Software cost estimation with incomplete data. **IEEE Transactions on Software Engineering**, v. 27, n. 10, p. 890–908, 2001.

TAKLE, E. S. Agricultural Meteorology and Climatology. **Encyclopedia of Atmospheric Sciences: Second Edition**, p. 92–97, 1 jan. 2015.

TERINK, W.; IMMERZEEL, W. W.; DROOGERS, P. Climate change projections of precipitation and reference evapotranspiration for the Middle East and Northern Africa until 2050. **International Journal of Climatology**, v. 33, n. 14, p. 3055–3072, 30 nov. 2013.

VYAS, M. et al. State-of-charge prediction of lithium ion battery through multivariate adaptive recursive spline and principal component analysis. **Energy Storage**, v. 3, n. 2, p. e147, 1 abr. 2021.

WANG, J. et al. Development of Monthly Reference Evapotranspiration Machine Learning Models and Mapping of Pakistan—A Comparative Study. **Water 2022, Vol. 14, Page 1666**, v. 14, n. 10, p. 1666, 23 maio 2022.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. **Statistics in Medicine**, v. 30, n. 4, p. 377–399, 2011.

WRIGHT, K. **The NIPALS algorithm**. Disponível em: <https://cran.r-project.org/web/packages/nipals/vignettes/nipals_algorithm.html>. Acesso em: 14 dez. 2022.

XIE, C. et al. Recovery Method for Missing Sensor Data in Multi-Sensor Based Walking Recognition System. **8th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, CYBER 2018**, p. 558–563, 10 abr. 2019.

YANG, Q. et al. MultiDA: Chemometric software for multivariate data analysis based on Matlab. **Chemometrics and Intelligent Laboratory Systems**, v. 116, p. 1–8, jul. 2012.

YATA, K.; AOSHIMA, M. Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. **Journal of Multivariate Analysis**, v. 101, n. 9, p. 2060–2077, 2010.

YATA, K.; AOSHIMA, M. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. **Journal of Multivariate Analysis**, v. 105, n. 1, p. 193–215, 2012.

YOZGATLIGIL, C. et al. Comparison of missing value imputation methods in time series: The case of Turkish meteorological data. **Theoretical and Applied Climatology**, v. 112, n. 1–2, p. 143–167, 2013.