

RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta tese/dissertação será disponibilizado somente a partir de 30/05/2027

At the author's request, the full text of this thesis/dissertation will not be available online until May 30, 2027

CAROLINE PIRES CREMASCO

**PROCEDIMENTOS MULTIVARIADOS PARA IDENTIFICAÇÃO DE *OUTLIERS* E
IMPUTAÇÃO DE DADOS FALTANTES EM SÉRIES TEMPORAIS
METEOROLÓGICAS**

Botucatu

2025

CAROLINE PIRES CREMASCO

**PROCEDIMENTOS MULTIVARIADOS PARA IDENTIFICAÇÃO DE *OUTLIERS* E
IMPUTAÇÃO DE DADOS FALTANTES EM SÉRIES TEMPORAIS
METEOROLÓGICAS**

Tese apresentada à Faculdade de
Ciências Agronômicas da Unesp Câmpus
de Botucatu, para obtenção do título de
Doutora em Engenharia Agrícola.

Orientador: Prof. Dr. Sergio Augusto
Rodrigues

Botucatu

2025

P667p

Pires Cremasco, Caroline

Procedimentos multivariados para identificação de outliers e
imputação de dados faltantes em séries temporais meteorológicas /
Caroline Pires Cremasco. -- Botucatu, 2025

147 p.

Tese (doutorado) - Universidade Estadual Paulista (UNESP),
Faculdade de Ciências Agrônômicas, Botucatu

Orientador: Sergio Augusto Rodrigues

1. análise multivariada. 2. ausencia de dados. 3. valores estranhos.
4. análise de séries temporais. 5. instruções meteorológicas. I. Título.

Impacto potencial desta pesquisa

Esta pesquisa contribui para o aprimoramento da qualidade de dados meteorológicos utilizados em estudos agrometeorológicos, ambientais e agrícolas. Ao comparar métodos multivariados para imputação de dados faltantes e detecção de outliers em séries temporais, o trabalho oferece subsídios para tornar as bases de dados mais confiáveis e reduzir erros em análises, estimativas e tomadas de decisão. Os resultados podem apoiar o uso mais seguro de informações meteorológicas em pesquisas científicas, monitoramento climático, planejamento agrícola e desenvolvimento de sistemas computacionais aplicados à Engenharia Agrícola.

Potential Impact of this research

This research contributes to improving the quality of meteorological data used in agrometeorological, environmental and agricultural studies. By comparing multivariate methods for missing-data imputation and outlier detection in time series, the study provides support for making databases more reliable and reducing errors in analyses, estimates and decision-making processes. The results may support the safer use of meteorological information in scientific research, climate monitoring, agricultural planning and the development of computational systems applied to Agricultural Engineering.

FOLHA DE APROVAÇÃO



UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu



CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: PROCEDIMENTOS MULTIVARIADOS PARA IDENTIFICAÇÃO OUTLIERS E IMPUTAÇÃO DE DADOS FALTANTES EM SÉRIES TEMPORAIS METEOROLÓGICAS

AUTORA: CAROLINE PIRES CREMASCO

ORIENTADOR: SERGIO AUGUSTO RODRIGUES

Aprovada como parte das exigências para obtenção do Título de Doutora em Engenharia Agrícola, pela Comissão Examinadora:

Documento assinado digitalmente
 **SERGIO AUGUSTO RODRIGUES**
Data: 30/05/2025 12:40:53 -0300
Verifique em: <https://validar.jf.gov.br>

Prof. Dr. SERGIO AUGUSTO RODRIGUES (Participação Virtual)
Bioprocessos e Biotecnologia / Faculdade de Ciências Agrômicas - Unesp

Documento assinado digitalmente
 **MAURICIO BRUNO PRADO DA SILVA**
Data: 30/05/2025 11:48:39 -0300
Verifique em: <https://validar.jf.gov.br>

Pós-Doutorando MAURÍCIO BRUNO PRADO DA SILVA (Participação Virtual)
Produção Vegetal / Unesp - FCAT

Documento assinado digitalmente
 **TERESA CRISTINA MARTINS DIAS**
Data: 30/05/2025 13:11:49 -0300
Verifique em: <https://validar.jf.gov.br>

Prof.ª Dr.ª TERESA CRISTINA MARTINS DIAS (Participação Virtual)
Estatística / Universidade de São Carlos

Documento assinado digitalmente
 **VALERIA CRISTINA RODRIGUES SARNIGHAUSEN**
Data: 30/05/2025 15:09:53 -0300
Verifique em: <https://validar.jf.gov.br>

Prof.ª Dr.ª VALERIA CRISTINA RODRIGUES SARNIGHAUSEN (Participação Virtual)
Bioprocessos e Biotecnologia / Faculdade de Ciências Agrômicas de Bc

Documento assinado digitalmente
 **ALEXANDRE DAL PAI**
Data: 30/05/2025 09:11:39 -0300
Verifique em: <https://validar.jf.gov.br>

Prof. Dr. ALEXANDRE DAL PAI (Participação Virtual)
Bioprocessos e Biotecnologia / Faculdade de Ciências Agrômicas - Unesp

Botucatu, 30 de maio de 2025

Aos meus pais Wilson e Clélia, marido Elton,
meus filhos João Pedro e Giovanna,
irmãs Camila e Karine,
dedico.

AGRADECIMENTOS

Aos meus pais, Wilson Cremasco e Clélia Maria Cremasco, por sempre torcerem por mim, incentivarem meus estudos e me transmitirem valores que levarei para toda a vida.

Ao meu marido, Elton Urici, pela compreensão, por sempre acreditar em mim, apoiar-me nas escolhas difíceis e pelo cuidado com nossa família enquanto eu me dedicava a este sonho.

Aos meus filhos, João Pedro e Giovanna Cremasco Urici, pela paciência, compreensão e por tantas vezes cederem em função deste trabalho.

Ao meu orientador, Prof. Dr. Sergio Augusto Rodrigues, pela orientação, paciência e confiança.

À minha irmã, Profa. Dra. Camila Pires Cremasco Gabriel, e ao meu cunhado, Prof. Dr. Luís Roberto Almeida Gabriel Filho, pelo apoio constante, pelo carinho e pelas palavras de incentivo em todos os momentos.

À minha irmã, Karine Cremasco Aragos, e ao meu cunhado, Rafael Aragos, pelo incentivo, pela torcida e pela presença sempre acolhedora ao longo desta caminhada.

À banca examinadora, pelas valiosas contribuições: aos professores da FCA/UNESP de Botucatu, Prof. Dr. Alexandre Dal Pai e Profa. Dra. Valéria Cristina Rodrigues Sarnighausen, pelo compromisso e sugestões enriquecedoras.

Aos professores externos, Prof. Dr. Maurício Bruno Prado da Silva e Profa. Dra. Teresa Cristina Martins Dias, pela disponibilidade, atenção e observações construtivas.

Ao Dr. Marcus Vinícius Contes Calça e Dr. José Rafael Franco, pelo apoio e colaboração ao longo deste processo.

À Universidade Estadual Paulista “Júlio de Mesquita Filho” (FCA/UNESP), Campus de Botucatu, e ao corpo docente do Programa de Pós-graduação em Engenharia Agrícola, pelas oportunidades de aprendizado, pelo conhecimento transmitido e pela contribuição na minha formação acadêmica, profissional e pessoal.

A todos que contribuíram direta ou indiretamente para a realização deste trabalho, meu sincero agradecimento.

“Nós somos o que fazemos repetidamente. A excelência não é um feito, e sim, um hábito”.

DURANT, W. J. **The story of philosophy**.
Nova Iorque: Pocket Books, 1991.

RESUMO

A qualidade dos dados meteorológicos é fundamental para análises agrometeorológicas confiáveis. Séries temporais frequentemente apresentam falhas de medição (missings) e distorções causadas por valores atípicos (outliers), que comprometem a precisão das estimativas e, consequentemente, a confiabilidade das decisões baseadas nesses dados. O presente trabalho propõe uma abordagem multivariada para a avaliação e comparação de métodos de imputação de missings e detecção de outliers, considerando a estrutura temporal dos dados. No Capítulo 1, são avaliadas sete técnicas de imputação para séries temporais multivariadas: média simples, Análise de Componentes Principais (PCA), Floresta Aleatória, K-Nearest Neighbors (KNN), Imputação Iterativa Dependente do Tempo (TDI), Redes Neurais Recorrentes (RNN) e SAITS. Os métodos são avaliados por MAE, MSE, RMSE, correlação de Pearson e tempo de processamento em diferentes níveis de ausência (5%–35%). O método Floresta Aleatória apresenta os menores erros e maior robustez à elevação do missing, embora com maior custo computacional. O TDI mantém desempenho elevado e menor custo de processamento, configurando o melhor compromisso entre precisão e eficiência. PCA e KNN apresentam desempenho intermediário quando o missing é baixo, mas perdem acurácia à medida que a ausência cresce, especialmente em variáveis com maior variabilidade. RNN e SAITS atingem precisão intermediária, porém com custo computacional mais alto. A média simples apresenta o pior desempenho, por não capturar a variabilidade dos dados. No Capítulo 2, são testados modelos multivariados de detecção de outliers, como PCA Tradicional e suas variações (RPCA, MWPCA), Isolation Forest e modelos baseados em LSTM. As abordagens são validadas com métricas voltadas a dados desbalanceados, como G-Mean, Kappa, F1-Score e MCC. Os métodos RPCA e LSTM se destacam por apresentar desempenho superior na detecção de pontos atípicos, mesmo em conjuntos de dados altamente desbalanceados.

Palavras-chave: análise multivariada; ausência de dados; valores atípicos; séries temporais; instrumentos meteorológicos.

ABSTRACT

The quality of meteorological data is essential for reliable agrometeorological analyses. Time series often present measurement gaps (missings) and distortions caused by outliers, which compromise the accuracy of estimates and, consequently, the reliability of data-driven decisions. This work proposes a multivariate approach for the evaluation and comparison of missing-data imputation methods and outlier detection techniques, considering the temporal structure of the data. In Chapter 1, seven imputation techniques for multivariate time series are evaluated, including Principal Component Analysis (PCA), Random Forest, K-Nearest Neighbors (KNN), Iterative Time-Dependent Imputation (TDI), Recurrent Neural Networks (RNN), and SAITS. These methods are assessed using MAE, MSE, RMSE, Pearson correlation, and processing time under different levels of missingness (5%–35%). TDI maintains high performance with lower computational cost, making it the best accuracy–efficiency trade-off, while Random Forest shows the lowest errors and greater robustness to higher missingness, though at a higher computational cost. PCA and KNN achieve intermediate results when missingness is low but lose accuracy as missing values increase, especially in variables with higher variability. RNN and SAITS reach intermediate accuracy, but with higher computational cost. Simple mean imputation shows the worst performance, as it fails to capture data variability. In Chapter 2, multivariate models for outlier detection are tested, such as Traditional PCA and its variations (RPCA, MWPCA), Isolation Forest, and LSTM-based models. These approaches are validated using metrics designed for imbalanced data, such as G-Mean and Kappa. RPCA and LSTM stand out by achieving superior performance in detecting outliers, even in highly imbalanced datasets.

Keywords: multivariate analysis; missing data; outlier detection; anomaly detection; time series analysis; meteorological instruments.

LISTA DE ILUSTRAÇÕES

Capítulo 1 - Métodos multivariados para imputação de missing em dados multivariados e dependentes do tempo

Figura 1 - Série temporal da temperatura média do ar ($^{\circ}\text{C}$), médias diárias, de dezembro de 2014 a julho de 2024	47
Figura 2 - Série temporal da umidade relativa média do ar (%), médias diárias, de dezembro de 2014 a julho de 2024	48
Figura 3 - Série temporal da precipitação diária (mm), de dezembro de 2014 a julho de 2024	48
Figura 4 - Comparação de desempenho de métodos de imputação de missing para a temperatura do ar média ($^{\circ}\text{C}$), considerando a simulação de 5%	52
Figura 5 - Comparação de desempenho de métodos de imputação de missing para a temperatura do ar média ($^{\circ}\text{C}$), considerando a simulação de 35%	53
Figura 6 - Comparação de desempenho de métodos de imputação de missing para a umidade relativa do ar média (%), considerando a simulação de 5%	54
Figura 7 - Comparação de desempenho de métodos de imputação de missing para a umidade relativa do ar média (%), considerando a simulação de 35%	55
Figura 8 - Boxplot dos Erros RMSE dos Métodos de Imputação	57

Capítulo 2 - Métodos multivariados para detecção de outliers em dados multivariados e dependentes do tempo

Figura 9 - Comparativo da Série Temporal após a etapa de pré-processamento dos dados	83
Figura 10 - Comparativo Distribuição temporal dos outliers nas variáveis meteorológicas, classificados como extremos e fisicamente incoerentes	86
Figura 11 - Comparativo Distribuição temporal dos outliers nas variáveis meteorológicas, outliers inseridos previamente e os detectados pela PCA Robusta Média Móvel	90

Figura 12 - Comparativo Distribuição temporal dos outliers nas variáveis meteorológicas, outliers inseridos previamente e os detectados pelo Isolation Forest	91
Figura 13 - Comparativo Distribuição temporal dos outliers nas variáveis meteorológicas, outliers inseridos previamente e os detectados pelo Boxplot	92
Figura 14 - Identificação dos Falsos Negativos da PCA Robusta Média Móvel através da combinação de Temperatura do Ar (°C) com Pressão Atmosférica (kPa) e Temperatura do Ar (°C) com Umidade Relativa do ar (%) por faixa de hora e estação do ano.....	93
Figura 15 - Identificação dos Falsos Negativos do Isolation Forest através da combinação de Temperatura do Ar (°C) com Pressão Atmosférica (kPa) e Temperatura do Ar (°C) com Umidade Relativa do ar (%) por faixa de hora e estação do ano.....	94
Figura 16 - Identificação dos Falsos Negativos do boxplot através da combinação de Temperatura do Ar (°C) com Pressão Atmosférica (kPa) e Temperatura do Ar (°C) com Umidade Relativa do ar (%) por faixa de hora e estação do ano	95
Figura 17 - Identificação dos Falsos Positivos da PCA Robusta Média Móvel através da combinação de Temperatura do Ar (°C) com Pressão Atmosférica (kPa) e Temperatura do Ar (°C) com Umidade Relativa do ar (%) por faixa de hora e estação do ano.....	96
Figura 18 - Identificação dos Falsos Positivos do Isolation Forest através da combinação de Temperatura do Ar (°C) com Pressão Atmosférica (kPa) e Temperatura do Ar (°C) com Umidade Relativa do ar (%) por faixa de hora e estação do ano.....	98
Figura 19 - Identificação dos Falsos Positivos do boxplot através da combinação de Temperatura do Ar (°C) com Pressão Atmosférica (kPa) e Temperatura do Ar (°C) com Umidade Relativa do ar (%) por faixa de hora e estação do ano	99

LISTA DE TABELAS

Capítulo 1 - Métodos multivariados para imputação de missing em dados multivariados e dependentes do tempo

Tabela 1 - Análise descritiva dos dados diários das variáveis coletadas, considerando o período de dez/14 a julho/24.....	46
Tabela 2 - Análise descritiva dos dados diários das variáveis tratadas, considerando o período de dez/14 a julho/24	47
Tabela 3 - Comparação de desempenho de métodos de imputação de missing para diferentes percentuais de dados ausentes	50

Capítulo 2 - Métodos multivariados para detecção de outliers em dados multivariados e dependentes do tempo

Tabela 4 - Estatísticas descritivas da base tratada e da base com inclusão de outliers.....	84
Tabela 5 - Componentes da Matriz de Confusão dos métodos aplicados	86
Tabela 6 - Desempenho dos métodos de detecção de outliers.....	88

LISTA DE ABREVIATURAS E SIGLAS

EMA	Estação Meteorológica Automática
EM	Expectation-Maximization
GAN	Generative Adversarial Network
G-Mean	Geometric Mean
IF	Isolation Forest
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCC	Coeficiente de Correlação de Matthews
MNAR	Missing Not At Random
MSE	Mean Squared Error
MWPCA	Moving Window Principal Component Analysis
NIPALS	Nonlinear Iterative Partial Least Squares
nRMSE	Normalized Root Mean Squared Error
PCA	Principal Component Analysis
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RPCA	Robust Principal Component Analysis
RSPCA	Robust Sparse Principal Component Analysis
SAITS	Self-Attention-based Imputation for Time Series
SANNI	Self-Attention-based Neural Network Imputation
SVD	Singular Value Decomposition
TDI	Time-Dependent Iterative Imputation
UR	Umidade Relativa

LISTA DE SÍMBOLOS

$\alpha_{t,j}$	peso de atenção normalizado para a posição j em relação ao tempo t
C_t	estado da célula LSTM no tempo t
$\tilde{C}_t$	vetor candidato do estado da célula
c_t	representação contextual no tempo t
d	distância euclidiana
$e_{t,j}$	escore de atenção entre os tempos t e j
f_k	modelo de previsão na iteração k
f_t	porta de esquecimento da LSTM
h_t	estado oculto no tempo t
i_t	porta de entrada da LSTM
k	número de componentes principais ou de vizinhos considerados
m	número de períodos considerados na imputação
M	máscara de ausência dos dados
N_t	número de instantes de tempo considerados
o_t	porta de saída da LSTM
p	número de variáveis
r	coeficiente de correlação de Pearson
T	número total de árvores
W	matriz de pesos ou autovetores
X	matriz de dados
x_t	valor da variável no instante t
$\hat{x}_t$	valor estimado ou imputado no instante t
y_i	valor real observado
$\hat{y}_i$	valor estimado ou imputado
Z	matriz de escores dos componentes principais
σ	função sigmoide
$\Delta\hat{x}(k)$	mudança média nos valores imputados entre iterações
%	porcentagem
°C	grau Celsius
kPa	quilopascal

SUMÁRIO

INTRODUÇÃO GERAL.....	27
CAPÍTULO 1 - MÉTODOS MULTIVARIADOS PARA IMPUTAÇÃO DE MISSING EM DADOS MULTIVARIADOS E DEPENDENTES DO TEMPO.....	30
1.1 INTRODUÇÃO.....	30
1.2 REVISÃO DE LITERATURA.....	32
1.2.1 PCA interativa para imputação de missing.....	32
1.2.2 Método Floresta Aleatória para imputação de missing.....	33
1.2.3 Método K-Nearest Neighbors (KNN).....	35
1.2.4 Método Imputação Iterativa Dependente do Tempo (TDI).....	36
1.2.5 Redes Neurais Recorrentes com Long Short-Term Memory (LSTM).....	38
1.2.6 Self-Attention-based Imputation for Time Series (SAITS).....	40
1.2.7 Método Média Simples para imputação de missing.....	43
1.3 MATERIAL E MÉTODOS.....	43
1.3.1 Dados do estudo.....	43
1.3.2 Métodos.....	44
1.4 RESULTADOS E DISCUSSÃO.....	46
1.5 CONCLUSÕES.....	58
REFERÊNCIAS	61
CAPÍTULO 2 – MÉTODOS MULTIVARIADOS PARA DETECÇÃO DE OUTLIERS EM DADOS MULTIVARIADOS E DEPENDENTES DO TEMPO.....	65
2.1 INTRODUÇÃO.....	65
2.2 REVISÃO DE LITERATURA.....	67
2.2.1 PCA e suas variantes na detecção de outliers.....	67
2.2.2 Análise de Componentes Principais (PCA).....	67
2.2.3 PCA Média Móvel (MWPCA).....	70
2.2.4 PCA Robusta (RPCA).....	71
2.2.5 Isolation Forest (IF).....	73
2.2.6 Long Short-Term Memory (LSTM).....	75
2.3 MATERIAL E MÉTODOS.....	76
2.3.1 Dados de Estudo.....	76

2.3.2	Métodos.....	77
2.4	RESULTADOS E DISCUSSÕES.....	83
2.5	CONCLUSÕES.....	101
	REFERÊNCIAS	103
	CONSIDERAÇÕES FINAIS.....	106
	REFERÊNCIAS	109
	APÊNDICE A	115
	APÊNDICE B	117
	APÊNDICE C	120
	APÊNDICE D	136

INTRODUÇÃO GERAL

A confiabilidade dos dados é decisiva para qualquer estudo em agrometeorologia e tem um papel fundamental em várias áreas da agronomia para auxiliar na tomada de decisões importantes em setores como agricultura e energia. Na área agrícola os dados meteorológicos desempenham papel importante no monitoramento da saúde das plantas, na estimativa da produção e demanda futuras, além disso, são essenciais na modelagem ambiental para prever eventos extremos como enchentes ou secas. Um desafio nestas investigações é a confiabilidade dos dados ao longo dos anos. Em séries temporais -registros contínuos de informações – podem apresentar falhas seja por falta de dados ou por estes estarem comprometidos.

Essas questões impactam a precisão das análises e comprometem o processo de tomar decisões com base nesses dados coletados através de redes de sensores meteorológicos. Os dados incorretos surgem devido a uma variedade de razões como falhas no *hardware* dos sensores, sinais comprometidos pela degradação ou fonte de alimentação inadequada. Além disso, as mudanças na localização dos dispositivos de medição ou modificações nos processos utilizados pelas estações meteorológicas podem também introduzir erros que distorcem os modelos usados em estudos da agrometeorologia.

A identificação de valores discrepantes, conhecidos como *outliers*, em séries temporais é iniciada por Fox (1972) e desde então, existem diversos métodos estatísticos que procuram garantir qualidade em séries temporais. Os métodos desenvolvidos devem levar em conta a característica do banco de dados sendo classificados como univariados ou multivariados. Um conjunto de dados que descreve uma única informação, como por exemplo, o dispositivo que monitora somente a radiação solar é considerado univariado. Quando se armazena um conjunto de dados com diferentes informações, tem-se dados multivariados, por exemplo, levantamento de dados de vento, temperatura e umidade simultaneamente.

Quando se monitora dados agrometeorológicos, obtém-se medidas de múltiplas variáveis que são registradas ao longo do tempo e em áreas diferentes, o banco de dados é formado por uma matriz multivariada complexa decorrente do grande número de informações. Se o evento precisar ser explicado por muitas variáveis (como por exemplo a evapotranspiração), a análise univariada pode gerar

resultados errôneos, dado que não considera a totalidade das informações fornecidas pelo conjunto das variáveis. Ao aplicar métodos multivariados, todas as relações entre as variáveis são consideradas, permitindo que os efeitos combinados das variáveis resultem em estimativas e interpretações mais precisas.

Existem situações mais fáceis de se identificar *outliers* por serem grosseiros ou valores muito extremos. Porém, mudanças mais sutis do processo de medição também podem ser consideradas pontos discrepantes e sua identificação além de complexa, requer métodos de detecção mais avançados. Essas mudanças sutis são muitas vezes difíceis de detectar com métodos tradicionais. Rousseeuw e Hubert (2011) enfatizam que os pontos discrepantes, mesmo que sutis, causam problemas pois influenciam fortemente o resultado, os autores apresentam uma visão geral de métodos robustos e ferramentas para detecção de *outliers* e apresentam procedimentos tanto para dados univariados de baixa dimensão, quanto multivariados que consideram grandes quantidades de dados, em que as dimensões das observações são maiores que o tamanho da amostra. Para estes dados de alta dimensão, os autores apresentam métodos de análise de Classificação e Análise de Componentes Principais (PCA).

Dados agrometeorológicos também apresentam dependência temporal, pois são séries em que as informações são coletadas sequencialmente, ou seja, as observações vizinhas são correlacionadas. Portanto, é necessário modelar essa dependência de forma adequada para garantir a qualidade das análises. De Ketelaere, Hubert e Schmitt (2015) apresentam uma revisão detalhada da literatura sobre dados de alta dimensão e dependentes do tempo e o uso de variantes robustas de PCA, incluindo Análise Dinâmica de Componentes Principais (DPCA), Análise Recursiva de Componentes Principais (RPCA) e Análise de Componentes Principais de Janela Móvel (MWPCA). Essas técnicas abordam tanto a dimensionalidade quanto a dependência temporal e podem ser empregadas para detecção de valores discrepantes. Além disso, Hubert *et al.* (2016) apresentam a abordagem *Robust Sparse Principal Component Analysis* (RSPCA), que é útil para dados de alta dimensão, pois pode analisar ou medir apenas um subconjunto das variáveis originais.

Quanto aos dados faltantes, Rubin (1976) apresenta técnicas estatísticas para preencher os valores ausentes com estimativas plausíveis. As abordagens mais simples fazem esse preenchimento apenas uma vez e não trazem a incerteza associada, o que limita a confiabilidade dos resultados (Enders, 2010). Para contornar

essa limitação, Rubin (1978) desenvolveu a Imputação Múltipla, que insere diferentes valores possíveis para os dados ausentes, permitindo estimar a incerteza envolvida. Josse, Pagès e Husson (2011) analisam um procedimento de imputação baseado na Análise de Componentes Principais (PCA) para gerar vários conjuntos de dados imputados. Dray e Josse (2014) também destacam métodos que utilizam a PCA no tratamento da informação ausente na comparação entre diferentes estratégias teóricas de imputação com essa técnica. O método PCA considera os dados de forma reduzida, indicando a forma mais representativa dos dados originais e é utilizada para diminuir o número de variáveis simplificando o banco de dados o que facilita na interpretação da análise. Esta simplificação é obtida através das componentes principais geradas pela técnica e que explicam a maior porcentagem da variação dos dados. Esta técnica também pode ser utilizada em conjunto de dados de alta dimensão, como em dados meteorológicos, os pontos de dados são esparsos, de modo que todos os pontos são quase que equidistantes uns dos outros. Em outras palavras, o uso de distâncias euclidianas na identificação de outliers torna-se sem sentido (Rousseeuw; Hubert, 2011). O resultado é que o grau de divergência dos pontos de dados é indistinguível um do outro. Por esta razão, os outliers são melhores detectados usando um subespaço local de menor dimensão no qual apenas um subconjunto de recursos pode ser relevante.

Devido a hipótese de ser mais adequado ao banco de dados considerado neste estudo, as variantes da PCA são estudadas por serem técnicas multivariadas que consideram a dimensionalidade e dependência temporal e que podem ser utilizadas tanto na identificação de *outliers*, quanto na imputação de informações *missing*.

Neste estudo os objetivos são:

- i. avaliar e comparar procedimentos multivariados para imputação em séries temporais multivariadas;
- ii. avaliar e comparar procedimentos multivariados para detecção de *outliers*;
- iii. disponibilizar código aberto reproduzível para controle de qualidade de dados.

A tese está organizada em dois capítulos. O Capítulo 1 foca no primeiro e terceiro objetivos e o Capítulo 2 trata dos objetivos dois e três. Ambos os capítulos seguem uma estrutura consistente, com seções dedicadas à introdução, revisão bibliográfica, materiais e métodos, e discussão dos objetivos propostos. Esta organização permite uma abordagem sistemática e coerente dos objetivos da pesquisa, facilitando uma compreensão clara e estruturada dos temas abordados.

CONSIDERAÇÕES FINAIS

Neste estudo, foi apresentada a aplicação de técnicas de imputação de valores ausentes e identificação de *outliers* em séries temporais multivariadas, com base nos dados da estação meteorológica de Botucatu. A imputação e a identificação de *outliers* são procedimentos essenciais em sistemas que automatizam a aquisição de medidas de sinais provenientes de falhas de sensores, ruídos de medição ou eventos atípicos, sendo aplicáveis em monitoramento ambiental e agrometeorológico.

Quanto à imputação de valores ausentes em séries temporais agrometeorológicas, os resultados demonstraram que métodos iterativos como o TDI (*Imputação Temporal Dependente Iterativa*) apresentam bons resultados, oferecendo um bom equilíbrio entre precisão e tempo de processamento, especialmente em cenários com níveis moderados de dados ausentes. O método Floresta Aleatória (*Random Forest*) também foi eficaz, mesmo com altos percentuais de *missings*, devido à sua capacidade de capturar padrões não lineares e interações complexas entre variáveis, porém com um certo custo computacional. Por outro lado, abordagens baseadas em aprendizado profundo, como RNNs e o modelo SAITS, apresentam desempenho intermediário e com custo computacional elevado, sendo mais indicadas quando há maior disponibilidade de recursos computacionais.

Alternativas mais clássicas, como a Análise de Componentes Principais (PCA), demonstraram desempenho intermediários e perdem eficiência conforme a proporção de valores ausentes aumenta.

Quanto à detecção de *outliers*, observou-se que os métodos estatísticos multivariados robustos, especialmente a PCA Robusta janela móvel, mostraram-se eficazes na identificação de padrões atípicos, sobretudo aqueles de natureza sutil ou que envolvem múltiplas variáveis simultaneamente. Os modelos baseados em LSTM também obtiveram resultados satisfatórios, com elevados *F1-Scores* e *G-Mean* na validação, reforçando sua aplicabilidade em séries temporais densas e complexas. Em contrapartida, abordagens mais simples, como o *boxplot* univariado, revelaram elevada taxa de falsos positivos em cenários multivariados e desbalanceados, o que limita sua aplicabilidade em dados reais de sensores ambientais.

O *Isolation Forest*, apesar de sua precisão geral e baixa taxa de falsos positivos, apresentou limitações na detecção de *outliers* incoerentes com base em regras físicas, como os inconsistentes multivariados.

A análise conjunta dos resultados evidencia a importância de considerar tanto as características intrínsecas dos dados (como multicolinearidade e dependência temporal) quanto o contexto físico em que estão inseridos. A detecção eficiente de anomalias em dados agrometeorológicos requer, portanto, o uso de metodologias combinadas, que integrem técnicas estatísticas robustas e filtros baseados em regras de coerência física.

Este estudo contribui para a literatura aplicada ao propor uma análise que garante a reprodutibilidade e pode ser reutilizada, adaptada ou incrementada, uma vez que foi estruturada em formato modular, com código aberto e parâmetros ajustáveis. Essa abordagem permite que o mesmo fluxo de análise seja aplicado em diferentes estações meteorológicas e contextos ambientais, bastando alterar configurações sem necessidade de reescrever o código. O uso de estratégias robustas, como a PCA Robusta com Janela Móvel e modelos baseados em LSTM, mostrou-se particularmente eficaz na detecção de anomalias complexas e na adaptação aos desafios impostos pela variabilidade sazonal e pelas dependências temporais.

REFERÊNCIAS

- AGGARWAL, C. C. Outlier Analysis. 2. ed. **Cham: Springer**, 2017. DOI: 10.1007/978-3-319-47578-3.
- AFZAL, M. *et al.* A robust anomaly detection method using improved PCA for smart manufacturing. **Computers & Industrial Engineering**, v. 154, p. 107130, 2021.
- AHMED, M.; HU, J.; HU, B. A survey of network anomaly detection techniques. **Journal of Network and Computer Applications**, v. 60, p. 19–31, 2016.
- BERTSIMAS, D.; PAWLOWSKI, C.; ZHUO, Y. D. From predictive methods to missing data imputation: an optimization approach. **The Journal of Machine Learning Research**, v. 18, n. 1, p. 7133–7171, 2018. Disponível em: <https://jmlr.org/papers/v18/17-073.html>.
- BLEIDORN, M. T. *et al.* Abordagens metodológicas para imputação de dados faltantes de vazões médias mensais. **Revista Ambiente & Água**, Taubaté, v. 17, n. 2, e2795, 2022. DOI: 10.4136/ambi-agua.2795.
- BLÁZQUEZ-GARCÍA, A.; CONDE, A.; MORI, U.; LOZANO, J. A. A review on outlier/anomaly detection in time series. **Information Fusion**, v. 62, p. 57–83, 2021. DOI: 10.1016/j.inffus.2021.05.008.
- BOTTERMAN, D. *et al.* Application of robust PCA in environmental data with missing values. **Environmental Modelling & Software**, v. 164, p. 105629, 2023.
- BOTTERMAN, H.-L.; ROUSSEL, J.; MORZADEC, T.; JABBARI, A.; BRUNEL, N. Robust PCA for anomaly detection and data imputation in seasonal time series. **arXiv preprint, arXiv:2208.01998**, 2022. Disponível em: <https://arxiv.org/abs/2208.01998>.
- BOUWMANS, T.; ZAHZAH, E.-H. Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance. **Computer Vision and Image Understanding**, v. 122, p. 22–34, 2014. DOI: 10.1016/j.cviu.2013.11.009.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324.
- CANDÈS, E. J.; LI, X.; MA, Y.; WRIGHT, J. Robust Principal Component Analysis? **Journal of the ACM**, v. 58, n. 3, p. 11:1–11:37, 2011. DOI: 10.1145/1970392.1970395.
- CAO, W. *et al.* BRITS: Bidirectional Recurrent Imputation for Time Series. In: Advances in Neural Information Processing Systems (NeurIPS 2018). **Montréal: Curran Associates**, 2018. p. 6775–6785. DOI: 10.48550/arXiv.1805.10572.
- CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? – **Arguments against avoiding RMSE. Geoscientific Model Development**, v. 7, p. 1247–1250, 2014. DOI: 10.5194/gmd-7-1247-2014.

CHAKRABARTI, S. *et al.* Binned data provide better imputation of missing time series data from wearables. **Sensors**, v. 23, n. 3, art. 1454, 2023. DOI: 10.3390/s23031454.

CHE, Z.; PURUSHOTHAM, S.; CHO, K.; SONTAG, D.; LIU, Y. Recurrent neural networks for multivariate time series with missing values. **Scientific Reports**, v. 8, art. 6085, 2018. DOI: 10.1038/s41598-018-24271-9.

CHETVIORKIN, I. Advanced techniques in handling time series with missing values. **Pattern Recognition Letters**, v. 32, n. 1, p. 119–125, 2011.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 6, p. 1–13, 2020.

DU, Y.; CÔTÉ, D.; LIU, Y. SAITS: Self-Attention-based Imputation for Time Series. Expert **Systems with Applications**, v. 219, art. 119619, 2023. DOI: 10.1016/j.eswa.2023.119619.

GERACI, M.; FARCOMENI, A. Principal Component Analysis in the presence of missing data. In: NAIK, G. R. (ed.). **Advances in Principal Component Analysis – Research and Development**. Singapore: Springer, 2018. p. 47–70. DOI: 10.1007/978-981-10-6704-4_3.

GREFF, K.; SRIVASTAVA, R. K.; KOUTNÍK, J.; STEUNEBRINK, B. R.; SCHMIDHUBER, J. LSTM: A search space odyssey. **arXiv preprint, arXiv:1503.04069**, 2015. Disponível em: <https://arxiv.org/abs/1503.04069>.

HARIRI, S.; KIND, M.; BRUNNER, A. Extended Isolation Forest. IEEE Transactions on **Knowledge and Data Engineering**, v. 33, n. 4, p. 1479–1492, 2021. DOI: 10.1109/TKDE.2019.2947676.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: data mining, inference, and prediction. 2. ed. **New York: Springer**, 2009. DOI: 10.1007/978-0-387-84858-7.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

HONG, S.; LYNN, H. S. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, interactions, and non-linear relations. **BMC Medical Research Methodology**, v. 20, n. 1, 2020. DOI: 10.1186/s12874-020-01080-1.

HOSSAIN, M.; AHMED, M. Time-dependent iterative imputation for missing values in time series data. **Procedia Computer Science**, v. 198, p. 385–392, 2022.

HUBERT, M.; ROUSSEEUW, P. J.; VANDEN BRANDEN, K. ROBPCA: A new approach to robust principal component analysis. **Journal of Chemometrics**, v. 19, n. 5–7, p. 145–163, 2005. DOI: 10.1002/cem.938.

- HUBERT, M.; ROUSSEEUW, P. J.; VANDEN BRANDEN, K. ROBPCA: a new approach to robust principal component analysis. **Technometrics**, v. 47, n. 1, p. 64–79, 2005. DOI: 10.1198/004017004000000563.
- HUBERT, M.; ROUSSEEUW, P. J.; VERDONCK, T. Robust PCA for skewed data and its outlier map. **Technometrics**, v. 51, n. 4, p. 393–404, 2009. DOI: 10.1198/TECH.2009.07145.
- HUNDMAN, K.; KONDO, K.; LINGENFELTER, B.; SODERSTROM, T. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: **Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. p. 387–395, 2018. DOI: 10.1145/3219819.3219845.
- INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). Rede SONDA – Sistema de Organização Nacional de Dados Ambientais. São José dos Campos: **INPE**, 2023. Disponível em: <http://sonda.ccst.inpe.br/>.
- JACKSON, J. E.; MUDHOLKAR, G. S. Control procedures for residuals associated with principal component analysis. **Technometrics**, v. 21, n. 3, p. 341–349, 1979. DOI: 10.1080/00401706.1979.10489779.
- JÄGER, K.; ALLIGNOL, A.; BEYERSMANN, J. Handling missing longitudinal data: The value of imputation methods. **Statistical Methods in Medical Research**, v. 30, n. 5, p. 1181–1198, 2021.
- JOHN, S. E.; EKPENYONG, M. S.; NWORU, E. A. A comparative study of PCA-based missing data imputation methods in time series. **International Journal of Statistics and Applications**, v. 9, n. 6, p. 311–318, 2019.
- JOLLIFFE, I. T. Principal Component Analysis. 2. ed. **New York: Springer**, 2002.
- JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A**, v. 374, 2016, art. 20150202. DOI: 10.1098/rsta.2015.0202.
- JUNNINEN, H.; NISKA, H.; TUPPURAINEN, K.; RUUSKANEN, J.; KOLEHMAINEN, M. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, 2004. DOI: 10.1016/j.atmosenv.2004.02.026.
- KRAMER, M. A. Nonlinear principal component analysis using autoassociative neural networks. **AIChE Journal**, v. 37, n. 2, p. 233–243, 1991. DOI: 10.1002/aic.690370209.
- KU, W.; STORER, R. H.; GEORGAKIS, C. Disturbance Detection and Isolation by Dynamic Principal Component Analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 30, n. 1, p. 179–196, 1995. DOI: 10.1016/0169-7439(95)00076-3.
- LI, C.; SUN, Y.; ZHANG, Z. Comparative study of imputation methods for missing meteorological time series data. **Atmospheric Research**, v. 240, p. 104928, 2020. DOI: 10.1016/j.atmosres.2020.104928.

LIEN, J.; DO, L.; NGUYEN, A. A comparative study of data imputation methods in wearable devices. **Procedia Computer Science**, v. 218, p. 118–125, 2023.

LIM, B.; ZOHREN, S. Time-series forecasting with deep learning: a survey. **Philosophical Transactions of the Royal Society A**, v. 379, n. 2194, 2021. DOI: 10.1098/rsta.2020.0209.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation Forest. In: 2008 Eighth IEEE International **Conference on Data Mining (ICDM)**. p. 413–422, 2008. DOI: 10.1109/ICDM.2008.17.

LIPTON, Z. C.; BERKOWITZ, J.; ELKAN, C. A critical review of recurrent neural networks for sequence learning. **arXiv preprint, arXiv:1506.00019**, 2015. Disponível em: <https://arxiv.org/abs/1506.00019>.

MORITZ, S.; BARTZ-BEIELSTEIN, T. imputeTS: Time Series Missing Value Imputation in R. **The R Journal**, v. 9, n. 1, p. 207–218, 2017. DOI: 10.32614/RJ-2017-009.

NOY, A.; SHAMIR, O. Time-aware iterative imputation of clinical time series. In: **Proceedings of the 36th AAAI Conference on Artificial Intelligence**. p. 1123–1131, 2023.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 11, p. 559–572, 1901.

PYTHON SOFTWARE FOUNDATION. Python Language Reference, version 3.12. 2024. Disponível em: <https://docs.python.org/3.12/>.

QIN, Rui; WANG, Xudong; ZHANG, Li; WU, Xun; ZHANG, Wei. ImputeGAN: Generative adversarial network for multivariate time series imputation. **Entropy**, v. 25, n. 1, art. 137, 2023. DOI: 10.3390/e25010137.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2024. Disponível em: <https://www.R-project.org/>

RSTUDIO TEAM. RStudio: Integrated Development Environment for R. Boston, MA: Posit Software, PBC, 2024. Disponível em: <https://posit.co/products/open-source/rstudio/>

RODGERS, J. L.; NICEWANDER, W. A. Thirteen ways to look at the correlation coefficient. **The American Statistician**, v. 42, n. 1, p. 59–66, 1988. DOI: 10.1080/00031305.1988.10475524.

RODRIGUEZ, J. J.; MARK, A. G. Time series missing data imputation with machine learning methods. **Data Mining and Knowledge Discovery**, v. 25, p. 565–592, 2011.

RODRIGUEZ, P.; WOHLBERG, B. Incremental principal component pursuit for video background modeling. **Journal of Mathematical Imaging and Vision**, v. 55, n. 1, p. 1–18, 2016. DOI: 10.1007/s10851-015-0610-z.

- RUBIN, D. B. Inference and missing data. **Biometrika**, v. 63, n. 3, p. 581–592, 1976.
- SAAD, A.; MAHMOOD, A.; AHMED, M. Improving accuracy of missing data imputation using LSTM networks. **Procedia Computer Science**, v. 192, p. 2834–2842, 2021.
- SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **PLOS ONE**, v. 10, n. 3, e0118432, 2015. DOI: 10.1371/journal.pone.0118432.
- 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.
- SOUZA, V. C. de; RODRIGUES, S. A.; GABRIEL FILHO, L. R. A. Comparação de algoritmos de análise de componentes principais para imputação em dados agrometeorológicos em alta dimensão e tamanho de amostra reduzido. **PLOS ONE**, v. 19, n. 12, p. e0315574, 2024. DOI: 10.1371/journal.pone.0315574.
- STEKHOVEN, D. J.; BÜHLMANN, P. MissForest — non-parametric missing value imputation for mixed-type data. **Bioinformatics**, v. 28, n. 1, p. 112–118, 2012. DOI: 10.1093/bioinformatics/btr597.
- SUN, Y. *et al.* Cost-sensitive boosting for classification of imbalanced data. **Pattern Recognition**, v. 40, n. 12, p. 3358–3378, 2007.
- VAN ROSSUM, M. C. *et al.* Missing data imputation techniques for wireless continuous vital signs monitoring. **Journal of Clinical Monitoring and Computing**, 2023, p. 1–14. DOI: 10.1007/s10877-023-00975-w.
- VASWANI, A. *et al.* Attention is all you need. In: **Advances in Neural Information Processing Systems 30 (NIPS 2017)**. 2017. DOI: 10.48550/arXiv.1706.03762.
- WANG, J.; LIU, H.; CHEN, X. Evaluation of missing data imputation methods for daily meteorological data. **Atmospheric Environment**, v. 213, p. 57–70, 2019. DOI: 10.1016/j.atmosenv.2019.05.049.
- WANG, X.; KRUGER, U.; IRWIN, G. Process monitoring approach using fast moving window PCA. **Industrial & Engineering Chemistry Research**, v. 44, n. 15, p. 5691–5702, 2005. DOI: 10.1021/ie048873f.
- WU, Z.; ZHANG, L.; LIN, L. Hybrid deep learning models for time series missing value imputation. **Expert Systems with Applications**, v. 165, p. 113957, 2021.
- YAN, P. *et al.* An imputation-enhanced hybrid deep learning approach for traffic volume prediction in urban networks: A case study in Dresden. **Data Science for Transportation**, v. 6, art. 22, 2024. DOI: 10.1007/s42421-024-00104-2.
- ZAINUDDIN, A. *et al.* Time series data and recent imputation techniques for missing data: a review. In: **INTERNATIONAL CONFERENCE ON GREEN ENERGY, COMPUTING AND SUSTAINABLE TECHNOLOGY (GECOST)**. IEEE, 2022. p. 26–31.
- ZAINUDDIN, N.; CHONG, K. W.; KADIR, E. A.; MANSOR, M. S. B. Outlier detection in environmental time series data using robust PCA with moving window.

Environmental **Modelling & Software**, v. 148, p. 105278, 2022. DOI: 10.1016/j.envsoft.2022.105278.

ZHANG, S.; KONG, D.; WANG, Y. LSTM-based missing data imputation for environmental time series. **Neurocomputing**, v. 410, p. 460–470, 2020.

ZHANG, W.; CHEN, J.; LI, X. Comparative analysis of imputation methods in environmental time series. **Environmental Modelling & Software**, v. 144, p. 105121, 2021.

ZOU, H.; HASTIE, T.; TIBSHIRANI, R. Sparse principal component analysis. **Journal of Computational and Graphical Statistics**, v. 15, n. 2, p. 265–286, 2006. DOI: 10.1198/106186006X113430.