Revista Brasileira de Climatologia
Brazilian Journal of Climatology
OPEN ACCESS
UFGD Editora

Preenchimento de falhas em séries de dados meteorológicos de estações automáticas

Gaps Filling in weather data series from automatic stations

Llenando brechas en series de datos meteorológicos de estaciones automáticas

Ana Rute Batista Pereira
Universidade Federal do Cariri - UFCA
ana.pereira@aluno.ufca.edu.br

Celme Torres Ferreira da Costa
Universidade Federal do Cariri - UFCA
celme.torres@ufca.edu.br

Paulo Renato Alves Firmino
Universidade Federal do Cariri - UFCA
paulo.firmino@ufca.edu.b

Ticiana Marinho de Carvalho Studart
Universidade Federal do Ceará - UFC
ticiana@ufc.br

Carlos Wagner Oliveira
Universidade Federal do Cariri - UFCA
carlos.oliveira@ufca.edu.br

Resumo: Um dos grandes problemas que surgem ao se trabalhar com dados medidos em estações meteorológicas é a quantidade de lacunas encontradas nos bancos de dados. A análise de séries incompletas pode gerar resultados incertos, impactando negativamente a gestão dos recursos hídricos. Com o intuito de solucionar essas falhas, o presente trabalho objetivou realizar a imputação dos valores ausentes, utilizando o método que retorna os menores erros. Os dados utilizados como caso de estudo são referentes à estação meteorológica automática de Iguatu-CE. Para imputação dos valores ausentes foram aplicados métodos como interpolação, média móvel, média, valor ausente decomposto sazonalmente e valor ausente dividido sazonalmente. As simulações de valores ausentes foram realizadas seguindo o esquema de amostragem de ausência aleatória (MAR), gerados para as porcentagens de 10% e 20% de falhas. A qualidade de cada método foi verificada utilizando medidas

de erro, como erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE). Os métodos testados apresentaram bons resultados para o preenchimento dos dados faltantes na série meteorológica sob estudo.

Palavras-chave: Preenchimento de falhas. Dados meteorológicos. Hidrologia estocástica.

Abstract: One of the biggest challenges when working with data collected from meteorological stations is the number of gaps in the databases. The analysis of incomplete series can generate uncertain results, negatively impacting the management of water resources. In order to address these gaps, the present study aimed to perform imputation of missing values using the method that yields the smallest errors. The data used as a case study pertains to the automatic meteorological station in Iguatu-CE. To impute the missing values, methods such as interpolation, moving average, mean, seasonally decomposed missing value, and seasonally divided missing value were applied. Missing value simulations were conducted following the scheme of random absence sampling (MAR), generated for percentages of 10% and 20% of missing data. The efficiency of each method was tested using error measures, mean absolute error (MAE), and root mean square error (RMSE). The tested methods showed good results in filling the gaps in the time series data.

Keywords: Gap filling. Meteorological data. Stochastic hydrology.

Resumen: Uno de los grandes problemas que surgen al trabajar con datos medidos en estaciones meteorológicas es la cantidad de lagunas encontradas en las bases de datos. El análisis de series incompletas puede generar resultados inciertos, impactando negativamente en la gestión de los recursos hídricos. Con el objetivo de solucionar estas deficiencias, el presente trabajo tuvo como objetivo realizar la imputación de los valores faltantes utilizando el método que arroja los errores más bajos. Los datos utilizados como caso de estudio son referentes a la estación meteorológica automática de Iguatu-CE. Para la imputación de los valores faltantes, se aplicaron métodos como la interpolación, el promedio móvil, la media, el valor faltante descompuesto estacionalmente y el valor faltante dividido estacionalmente. Las simulaciones de valores faltantes se realizaron siguiendo el esquema de muestreo de ausencia aleatoria (MAR), generadas para porcentajes de 10% y 20% de fallas. La eficiencia de cada método se probó utilizando medidas de error, error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE). Los métodos probados mostraron buenos resultados para llenar los datos faltantes en la serie.

Palabras clave: Relleno de lagunas. Datos meteorológicos. Hidrología estocástica.

Submetido em: 02/10/2023
Aceito para publicação em: 10/07/2024
Publicado em: 19/07/2024

1. INTRODUÇÃO

Um dos desafios significativos que se apresentam ao lidar com informações coletadas em estações meteorológicas é a quantidade de falhas (períodos sem medição) encontradas nos bancos de dados. As falhas pluviométricas e hidrológicas podem ocorrer devido a problemas técnicos/operacionais, como ausência do observador, falhas instrumentais, quebra na linha de comunicação ou localização geográfica, por exemplo, que podem levar a interpretações incorretas se não forem corrigidas (CORREA *et al.*, 2021). Bier e Ferraz (2017) relatam que a rede de estações meteorológicas no Brasil é muito recente, possuindo poucas estações no país com mais de 100 anos de dados.

Segundo Correa *et al.* (2021), a presença de muitas falhas na série climatológica interfere nos resultados encontrados, gerando problemas de interpretação nos dados. Assim, é importante a utilização de metodologias que sejam capazes de estimar valores que correspondam aos demais valores presentes nas séries meteorológicas de interesse. Dessa forma, para trabalhar com séries contínuas, torna-se necessário que essas falhas sejam preenchidas (OLIVEIRA *et al.*, 2010).

Os dados meteorológicos possuem diversas características diferentes entre si e, como consequência, métodos de preenchimento de falhas podem ter desempenhos variados dependendo do cenário encontrado (ZENERE *et al.*, 2020).

Os recentes avanços computacionais impulsionaram o desenvolvimento de técnicas de análise e otimização com grande aplicabilidade, inclusive no estudo de fenômenos hidrológicos (MACHIWAL; JHA, 2012). Em Correia *et al.* (2016) é descrito o uso de redes neurais artificiais no preenchimento de falhas em séries de precipitação mensal. No trabalho de Mello, Kohls e Oliveira (2017) são utilizados métodos estatísticos para preencher falhas em estações pluviométricas. Cunha Júnior e Firmino (2022) compararam diferentes métodos de preenchimento de falhas em séries mensais de precipitação da região do Cariri-CE. Sabino e Souza (2023) introduziram o programa GapMET, desenvolvido pelos autores, para avaliar a precisão de seus seis métodos de preenchimento de dados nas principais variáveis meteorológicas. Como visto nos trabalhos citados, diferentes métodos podem ser aplicados na imputação de dados faltantes em séries de dados. Desse modo há uma dificuldade para avaliar qual método comporta-se melhor para determinadas variáveis meteorológicas. O

objetivo deste trabalho é avaliar os resultados de cinco técnicas de imputação univariada, incluindo as diferentes metodologias de cada técnica. Utilizando o método que retorna os menores erros, objetiva também realizar a imputação dos valores ausentes para construir séries de dados confiáveis para estudos hidrológicos.

2. METODOLOGIA

2.1. Local de Estudo

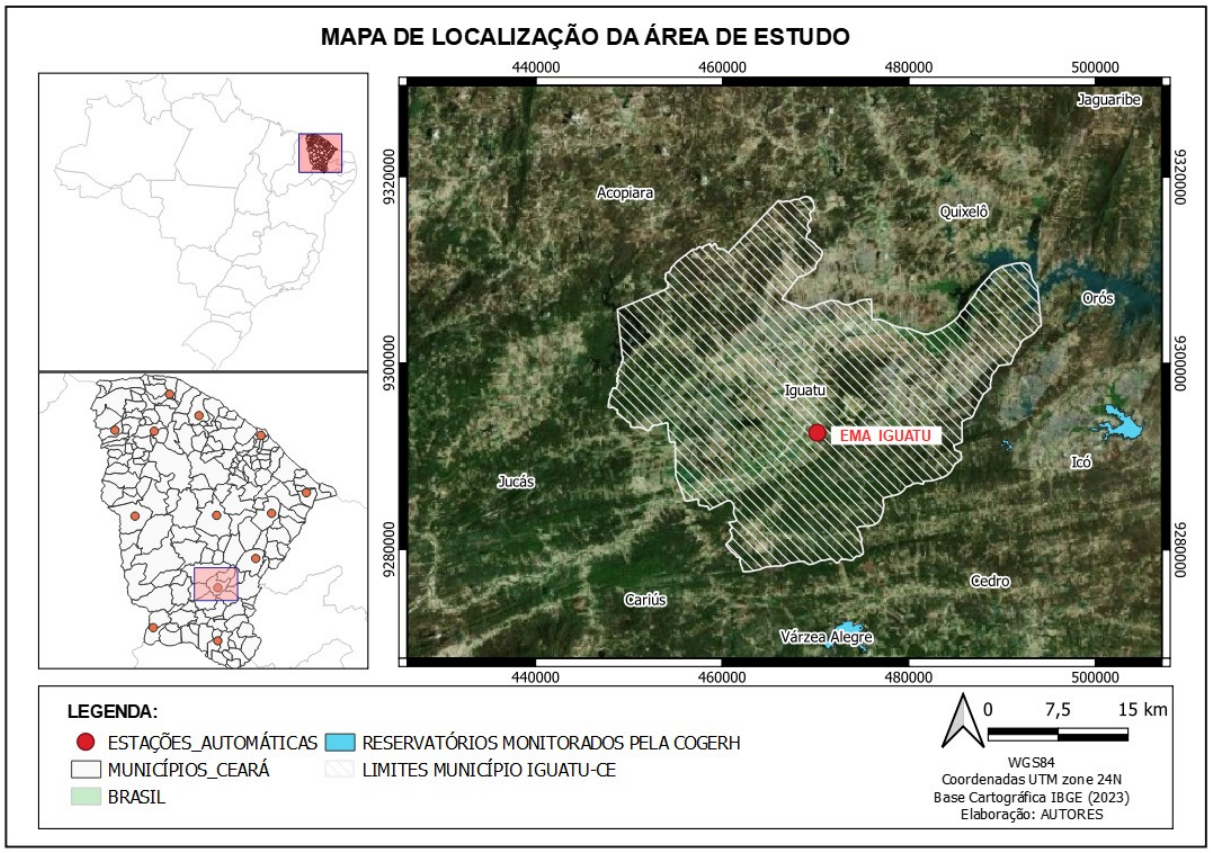
Esta pesquisa foi realizada com dados do município de Iguatu, Centro-Sul do estado do Ceará, localizado na região do Nordeste brasileiro. Segundo dados do IPECE (2017), o município de possui clima Tropical Quente semiárido, pluviosidade média de 806,5 mm/ano, temperatura média entre 26°C e 28°C, tendo seu período chuvoso médio compreendido entre os meses de janeiro a abril. O relevo do município tem como característica principal as depressões sertanejas e o tipo de vegetação predominante é a caatinga.

As estações meteorológicas automáticas (EMA), do INMET coletam, a cada hora, as informações meteorológicas (temperatura, umidade, pressão atmosférica, precipitação, direção e velocidade de vento e radiação) para a área onde estão localizadas. No estado do Ceará existem 13 EMA's monitoradas pelo INMET, nos municípios de Acaraú, Barbalha, Campos Sales, Crateús, Fortaleza, Iguatu, Itapipoca, Jaguaribe, Jaguaruana, Morada Nova, Quixeramobim, Sobral e Tianguá. Na cidade de Iguatu-CE, localiza-se nas coordenadas geográficas -39.27 W e -6.40 S e possui dados horários a partir de 29 de maio de 2007.

Para a análise do melhor método de imputação de dados foi utilizada toda a série histórica, compreendendo o período de maio de 2007 a dezembro de 2022, já para o preenchimento das falhas, optou-se por utilizar os dados de 2014 a 2022, por ser um período com maior quantidade de dados relativos à gestão de recursos hídricos no estado do Ceará, tornando mais relevantes os estudos hidrológicos aplicados a esse período.

Na Figura 1 é possível identificar a localização da Estação Meteorológica Automática de Iguatu-CE e a delimitação da área do município.

Figura 1 - Mapa de Localização da Estação Meteorológica Automática de Iguatu-CE, destacando os limites do referido município e os reservatórios monitorados da Bacia Hidrográfica do Salgado.



Fonte: Autores (2023).

2.2. Métodos de Imputação de dados de séries meteorológicas

Encontrar pacotes capazes de preencher valores ausentes em séries temporais univariadas é desafiador. Essa afirmação baseia-se no fato de que a maioria dos algoritmos de imputação depende de correlações entre atributos, enquanto a imputação de séries temporais univariadas precisa, em vez disso, empregar dependências de tempo. Para realizar esse procedimento, foi utilizado nesse trabalho o pacote *imputeTS* (MORITZ; BART Z-BEIELSTEIN, 2017).

Interpolação

A interpolação, dada pela função "na_interp" do pacote *imputeTS*, pode ser definida como a estimativa de um valor desconhecido a partir de valores conhecidos (DOURADO,

2014). Utilizando o *imputeTS* os valores ausentes podem ser estimados por valores de interpolação linear, interpolação *spline* ou interpolação de Stineman.

A interpolação linear calcula a localização estimada de um objeto por uma linha reta entre dois pontos reais coletados. Conhecido por ser um método simples de ser implementado, possui resultados muito bons para previsão de valores com taxa de mudança constante (GNAUCK, 2004), além de ser utilizado recorrentemente na literatura como parâmetro de comparação em estudos.

A forma da interpolação *spline*, de acordo com Wijesekara e Liyanage (2020), é modelada para n+1 pares de observações $\{(t_i, x_i); i=0, 1, ..., n\}$, interpolando entre todos os pares de observações (t_{i-1}, x_{i-1}) e (t_i, x_i) com polinômios q(t), da forma descrita na Equação 1

(1) $x = q_i\,(t)$, i = 1, 2, ..., n.

A interpolação Stineman é, segundo Turicchi *et al.* (2020), um método de interpolação avançado onde a interpolação ocorre (i) se os valores das ordenadas dos pontos especificados mudam monotonicamente e (ii) as inclinações dos segmentos de linha que unem os pontos especificados mudam monotonicamente.

Média móvel ponderada

Nesta função, "na_ma" - do pacote *imputeTS*, os valores ausentes são substituídos por valores médios móveis. A média nesta implementação é obtida de um número igual de observações em ambos os lados de um valor central. Isso significa que para um valor não disponível (NA, do inglês *Not Available*) na posição i de uma série temporal, as observações de índice i-1, i+1 e i+1, i+2 (assumindo um tamanho de janela de k=2) são usadas para calcular a média. E, no caso de longos intervalos de NA, o algoritmo possui um tamanho de janela semi-adaptável, até que pelo menos 2 valores não-NA estejam presentes (MORITZ; BARTZ-BEIELSTEIN, 2017).

Valores Médios

A média (na_mean, do pacote *imputeTS*), é a técnica de imputação única mais comumente usada, onde os valores ausentes são substituídos pelo valor médio da série temporal (TURICCHI *et al*, 2020). A média ($\bar{x}$) de uma série de valores x_1, x_2, ..., x_n é dada pela Equação 2:

(2) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ,$$

onde n equivale à quantidade de x_i elementos do conjunto.

A função calcula também a mediana, moda, média harmônica ou geométrica sobre todos os valores não-NA e substitui todos os NA's por este valor. Devido à sua fórmula de cálculo, as médias, geométrica e harmônica, não são bem definidas para valores negativos ou valores zero na série de entrada. Moritz e Bartz-Beielstein (2017) ressaltam que usar a média para imputação não é uma escolha ideal e deve ser tratada com muita cautela.

Valor ausente decomposto sazonalmente

O algoritmo "na_seadec", do pacote *imputeTS*, primeiro executa uma decomposição de séries temporais sazonal e de tendência (STL) usando Loess (CLEVELAND *et al.*, 1990). O método STL remove um componente de tendência estimado da série temporal, divide os dados em uma série de subciclos e depois os suaviza usando Loess. Este processo é então repetido até que a convergência na decomposição seja observada. O método STL tem melhor desempenho em circunstâncias onde uma tendência longa está presente, porque o suavizador de Loess é bom na detecção de tendências.

Como segunda etapa, o algoritmo de imputação selecionado é aplicado nas séries dessazonalizadas. Assim, o algoritmo pode funcionar sem ser afetado por padrões sazonais. Após preencher as lacunas de NA, o componente sazonal é novamente adicionado à série dessazonalizada.

Valor ausente dividido sazonalmente

Essa metodologia divide as séries temporais em temporadas e depois realiza a imputação separadamente para cada um dos conjuntos de dados de séries temporais resultantes (cada um contendo os dados de uma temporada específica). O comando usado é o *"na_seasplit"*. O algoritmo usa internamente imputação de média para séries não sazonais (MORITZ; BART Z-BEIELSTEIN, 2017).

2.3. Simulação de valores ausentes e avaliação do desempenho dos métodos

Todos os passos da metodologia foram implementados na linguagem de programação estatística R (R CORE TEAM, 2023), no Rstudio (RSTUDIO TEAM, 2023). Para definir o melhor método de imputação optou-se pelo uso do pacote *impute Testbench* (BECK et al., 2018), do Rstudio. Além dos métodos padrões de imputação existentes no pacote, ele permite que os usuários incluam metodologias adicionais para comparação.

No teste, o objeto de entrada da função deve ser um conjunto de dados completo, no entanto todas as séries apresentavam lacunas. Como solução, fez-se o uso do pacote *imputeTS* para realizar um pré-processamento (MORITZ; BART Z-BEIELSTEIN, 2017). Sendo assim, foi realizado um preenchimento prévio dos dados com a função *na_seadec (algorithm = "interpolation")* para que fosse possível realizar a escolha do melhor método de imputação em cada uma das variáveis. Posteriormente, as falhas foram simuladas utilizando o mecanismo de ausência aleatória (MAR, do inglês *Missing at Random*), apropriado para simular situações em que as falhas em uma série temporal são causadas devido a falhas em equipamentos durante longos períodos (BECK et al., 2018).

Para realização do presente estudo foram utilizadas nos testes funções de interpolação, média móvel ponderada, média, valor ausente decomposto sazonalmente e valor ausente dividida sazonalmente. Para cada uma das funções, foram feitos os testes considerando as variações de metodologia de cálculo, desse modo, a nomenclatura dos métodos aplicados nessa pesquisa é apresentada no Quadro 1, a seguir:

Quadro 1 – Nomenclatura dos métodos aplicados para imputação dos dados faltantes.

Método	Função
interp_linear	Interpolação linear usando aproximação
interp_stine	Interpolação *Stineman*
interp_spline	Interpolação spline
ma_simple	Média Móvel Simples
ma_linear	Média móvel ponderada linear
ma_exponential	Média móvel ponderada exponencial (escolha padrão)
mean_median	Mediana
mean_mode	Moda
mean_harmonic	Média harmônica
mean_geometric	Média geométrica
seadec_ma	Valor ausente decomposto sazonalmente por média móvel ponderada
seadec_locf	Valor ausente decomposto sazonalmente pela última observação realizada
seadec_interp	Valor ausente decomposto sazonalmente por interpolação
seadec_mean	Valor ausente decomposto sazonalmente por valor médio
seasplit_interp	Valor ausente dividida sazonalmente por interpolação
seasplit_locf	Valor ausente dividida sazonalmente pela última observação realizada
seasplit_mean	Valor ausente dividida sazonalmente por valor médio
seasplit_ma	Valor ausente dividida sazonalmente por média móvel ponderada

Fonte: Autores (2023).

2.4. Avaliação de desempenho

O erro absoluto médio (MAE, do inglês *Mean Absoluto Error*) e a raiz do erro quadrático médio (RMSE, do inglês *Root Mean Squared Error*) medem a magnitude dos erros em um conjunto de estimativas, nas unidades da variável de interesse. Eles são calculados pelas Equações 3 e 4, respectivamente

(3) $$MAE = \frac{\sum_{t=1}^{n} |\hat{y}_t - y_t|}{n}$$

(4) $$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}} \quad ,$$

em que $\hat{y}_t$ representa o valor estimado pelo método para y_t - o valor observado da série no instante de tempo *t*; e *n* é o número de valores faltantes (HODSON, 2022).

Destaca-se também que nesse tipo de trabalho, além das técnicas acima descritas, podem ser utilizados o Erro Médio Relativo, o Vies (Bias) como técnica de análises e avaliações de resultados, assim como técnicas de avaliação do conjunto central dos dados (JUNQUEIRA *et al.*, 2018; OLIVEIRA *et al.*, 2021).

O desempenho dos métodos de imputação foi avaliado para as porcentagens 10% e 20%. As falhas foram simuladas seguindo o esquema de amostragem de ausência aleatória – MAR, onde seleciona observações em blocos contínuos de modo que a probabilidade de seleção para uma única observação depende se uma observação mais próxima no tempo também foi selecionada. Os pontos de dados da série foram amostrados aleatoriamente 50 vezes para cada porcentagem, cujos tamanhos amostrais foram 13.663 e 54.653, para 10% e 20%, respectivamente.

3. RESULTADOS E DISCUSSÃO

3.1. Simulação e preenchimento de falhas

Além das falhas próprias da série, devido a algum tipo de problema na estação automática, percebeu-se que na série de radiação solar (Rs) havia falhas diárias que são lidas como *Not Available* (NA) pelo sistema. Esses valores de radiação representam os dados horários de períodos noturnos, ou seja, dados que não teriam valor significativo para o algoritmo, vide que o objetivo é o preenchimento de falhas de radiação válidas, dessa forma adotou-se o valor igual a zero.

Após essa correção, foi calculado o percentual de dados faltantes em cada série. Como pode-se ver na Tabela 1, os percentuais variam entre 7,05% e 14,92%.

Tabela 1 - Percentual de dados faltantes nos conjuntos de dados meteorológicos da estação automática de Iguatu-CE, após correção da série de radiação, por variável.

Rs	Patm	UR	U2	Temp	Temp. max	Temp. min
7,05%	14,7%	14,92%	14,76%	14,73%	14,92%	14,92%

Nota: Rs- radiação solar; Patm- pressão atmosférica; UR- umidade relativa; U2- velocidade do vento; Temp- temperatura máxima; Temp. max- temperatura máxima; Temp. min- temperatura mínima.

Fonte: Autores (2023).

A Tabela 2 mostra a média aritmética dos valores de erro absoluto médio (MAE, do inglês *mean absolute error*) e raiz do erro quadrático médio (RMSE, do inglês *root mean squared error*), obtidos para cada um dos dez melhores métodos, sob as diferentes porcentagens de falhas, nas variáveis temperatura(°C), temperatura máxima(°C) e temperatura mínima(°C). Os melhores valores estão destacados em negrito na tabela. O significado da nomenclatura dos métodos encontra-se no Quadro 1 (Seção 2.3).

Tabela 2 - Erros (MAE e RMSE) encontrados para cada método de imputação das variáveis de temperatura(°C), temperatura máxima(°C) e temperatura mínima(°C), na estação automática de Iguatu-CE (junho de 2007 a dezembro de 2022). Os melhores valores estão destacados em negrito.

% Falhas	Método	TEMPERATURA		TEMPERATURA MÁXIMA		TEMPERATURA MÍNIMA	
		MAE (°C)	RMSE (°C)	MAE (°C)	RMSE (°C)	MAE (°C)	RMSE (°C)
10%	interp_linear	0,0390	0,1861	0,0337	0,1616	0,0342	0,1669
	interp_stine	0,0377	0,1816	0,0317	0,1540	0,0323	0,1598
	interp_spline	0,0407	0,1945	0,0329	0,1572	0,0327	0,1616
	ma_linear	0,0758	0,3218	0,0745	0,3112	0,0729	0,3134
	ma_exponential	0,0608	0,2654	0,0585	0,2517	0,0578	0,2552
	seadec_ma	0,0397	0,1896	0,0345	0,1669	0,0360	0,1761
	seadec_locf	0,0471	0,2322	0,0409	0,2032	0,0421	0,2142
	seadec_interp	**0,0369**	**0,1800**	**0,0309**	**0,1516**	**0,0315**	**0,1580**
	seasplit_interp	0,0856	0,3892	0,0817	0,3778	0,0841	0,3851
	seasplit_ma	0,0826	0,3706	0,0793	0,3622	0,0813	0,3665
20%	interp_linear	0,0850	0,2866	0,0744	0,2520	0,0751	0,2613
	interp_stine	0,0814	0,2775	0,0691	0,2375	0,0704	0,2483
	interp_spline	0,0869	0,2950	0,0710	0,2402	0,0710	0,2498
	ma_exponential	0,1292	0,4018	0,1242	0,3818	0,1226	0,3860
	seadec_ma	0,0821	0,2761	0,0715	0,2428	0,0744	0,2578
	seadec_locf	0,0988	0,3422	0,0862	0,3016	0,0891	0,3202
	seadec_interp	**0,0771**	**0,2637**	**0,0650**	**0,2235**	**0,0663**	**0,2353**
	seadec_mean	0,3347	0,9197	0,3379	0,9218	0,3196	0,8741
	seasplit_interp	0,0856	0,5596	0,1663	0,5430	0,1708	0,5521
	seasplit_locf	0,1031	0,6734	0,2011	0,6554	0,2057	0,6637
	seasplit_ma	0,0826	0,5341	0,1613	0,5204	0,1650	0,5263

Nota: MAE – Erro absoluto médio; RMSE - raiz do erro quadrático médio.

Fonte: Autores (2023).

Nas variáveis de temperatura (temperatura, temperatura máxima e temperatura mínima), para ambas as porcentagens de falha (10% e 20%), o método “seadec_interp”, obtido com a função na_seadec (algorithm = “interpolation”), apresentou desempenho superior em comparação com os demais. Considerando os cinco melhores métodos, houve pequenas variações a partir do segundo melhor, mas para todas as condições testadas nessas variáveis destacaram-se o “interp_stine”, “interp_linear”, “interp_spline” e “seadec_ma”. Percebe-se que o comportamento dessas variáveis responde muito bem aos métodos de interpolação.

A Tabela 3 mostra a média aritmética dos valores de MAE e RMSE, obtidos para cada um dos dez melhores métodos, sob as diferentes porcentagens de falhas, nas variáveis pressão atmosférica (kPa), umidade relativa (%), velocidade do vento (m/s) e radiação (MJ/m².dia). Os melhores valores estão destacados em negrito na tabela.

Tabela 3 - Erros (MAE e RMSE) para cada método de imputação das variáveis de pressão atmosférica (kPa), umidade relativa (%), velocidade do vento (m/s) e radiação (MJ/m².dia), na estação automática de Iguatu-CE (junho de 2007 a dezembro de 2022). Os melhores valores estão destacados em negrito.

% Falhas	Método	PRESSÃO		UMIDADE RELATIVA		VELOCIDADE DO VENTO		RADIAÇÃO	
		MAE (kPa)	RMSE (kPa)	MAE (%)	RMSE (%)	MAE (m/s)	RMSE (m/s)	MAE (MJ/m².dia)	RMSE (MJ/m².dia)
10%	interp_linear	0,0016	0,0072	0,1752	0,8805	0,0505	**0,2321**	0,1557	0,7634
	interp_stine	0,0014	0,0062	**0,1687**	**0,8581**	**0,0504**	0,2332	0,1566	0,7691
	interp_spline	**0,0012**	**0,0054**	0,1829	0,9115	0,0592	0,2718	0,1888	0,9266
	ma_simple	0,0061	0,0244	0,4206	1,7714	0,0696	0,2997	0,1652	0,7656
	ma_linear	0,0049	0,0195	0,3449	1,4822	0,0625	0,2719	0,1588	0,7433
	ma_exponential	0,0036	0,0148	0,2763	1,2331	0,0568	0,2508	**0,1552**	**0,7351**
	seadec_ma	0,0021	0,0093	0,1979	0,9879	0,0570	0,2518	0,1608	0,7613
	seadec_locf	0,0032	0,0144	0,2341	1,2147	0,0714	0,3132	0,1934	0,9345
	seadec_interp	0,0014	0,0066	0,1782	0,9144	0,0544	0,2460	0,1641	0,7966
	seasplit_ma	0,0132	0,0516	0,4749	2,1582	0,1198	0,5715	0,2201	0,9571
20%	interp_linear	0,0038	0,0121	0,3803	1,3529	0,1045	**0,3396**	0,3187	1,0863
	interp_stine	0,0032	0,0104	**0,3622**	**1,3098**	**0,1043**	0,3417	0,3212	1,0973
	interp_spline	**0,0026**	**0,0083**	0,3908	1,3904	0,1233	0,4015	0,3941	1,3383
	ma_simple	0,0125	0,0354	0,8649	2,6011	0,1417	0,4303	0,3327	1,0916
	ma_linear	0,0100	0,0286	0,7136	2,1937	0,1276	0,3922	0,3208	1,0606
	ma_exponential	0,0077	0,0223	0,5812	1,8600	0,1168	0,3646	**0,3147**	**1,0507**
	seadec_ma	0,0045	0,0143	0,4068	1,4319	0,1168	0,3643	0,3277	1,0961
	seadec_locf	0,0071	0,0232	0,4904	1,7879	0,1474	0,4590	0,3923	1,3502
	seadec_interp	0,0030	0,0101	0,3704	1,3375	0,1110	0,3558	0,3345	1,1418
	seasplit_ma	0,0264	0,0733	0,9652	3,1130	0,2467	0,8427	0,4493	1,3751

Fonte: Autores (2023).

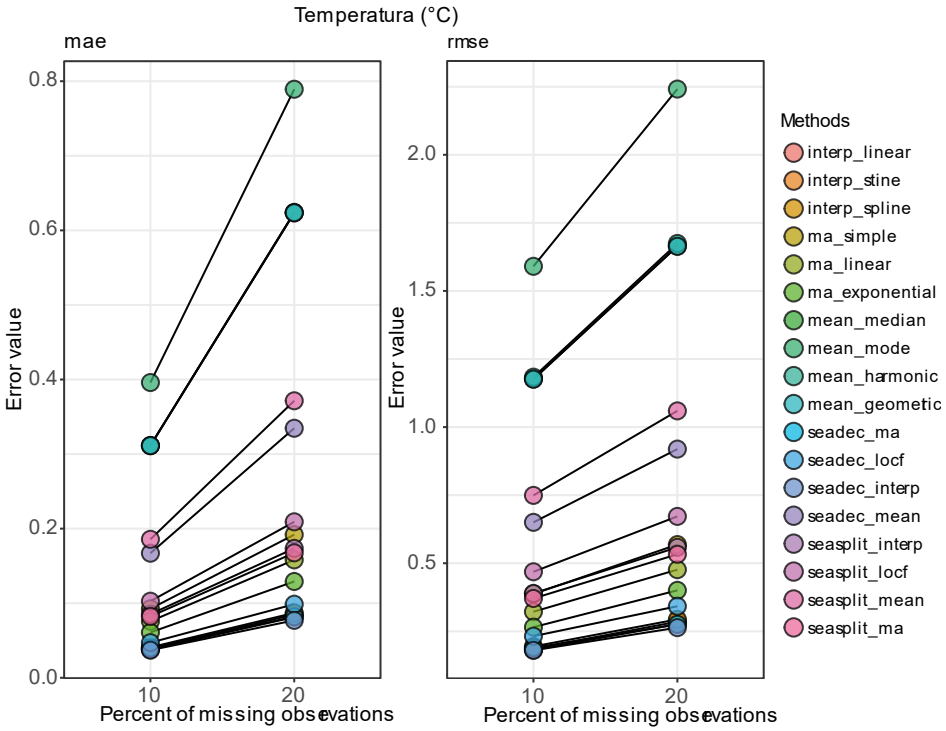
Diferentemente das variáveis de temperatura, os demais conjuntos de dados não apresentaram um padrão em relação ao melhor método de imputação de dados, mas mantiveram um padrão em relação aos percentuais. Os dados de pressão, umidade relativa e velocidade do vento responderam bem aos métodos de interpolação, tiveram melhor êxito com os métodos “interp_spline”, “interp_stine” e “interp_linear”, respectivamente. Já os dados de radiação apresentaram um melhor comportamento com o método de imputação por média móvel exponencial, “ma_exponencial”.

Nas Figuras 2, 3, 4, 5, 6, 7 e 8, a seguir, são apresentadas de forma gráfica da variação dos valores médios de erro (MAE e RMSE) para cada método de imputação e intervalo de observações ausentes nas variáveis de temperatura (°C), radiação (MJ/m².dia), pressão atmosférica (kPa), temperatura mínima (°C), temperatura máxima (°C), umidade relativa (%), velocidade do vento (m/s) e pressão atmosférica (kPa), respectivamente.

Ao se observar os valores médios calculados para MAE e RMSE, percebeu-se que o aumento na porcentagem de falhas, de 10% para 20%, diminuiu a qualidade e precisão das estimativas de todos os métodos estudados.

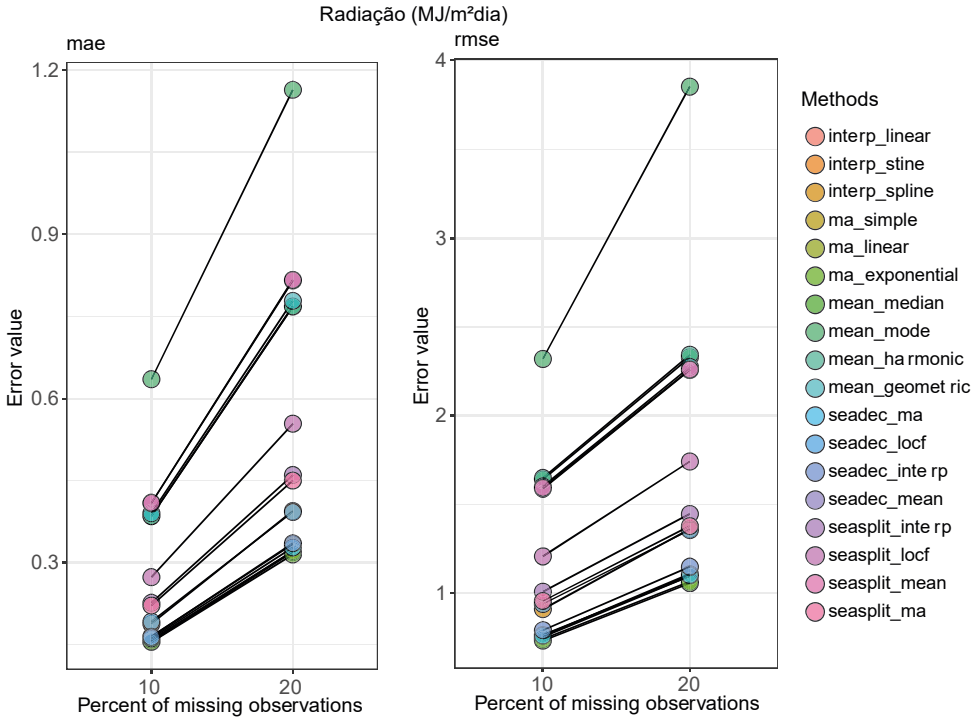
Os métodos que apresentam piores desempenhos são os que fazem a substituição dos valores com base em valores médios da série de dados (“mean_median”, “mean_mode”, “mean_harmonic” e “mean_geometric”). Esse comportamento já era esperado e foi comentado por Moritz e Bartz-Beielstein (2017) ao explicar que usar a média para imputação não é uma escolha ideal e deve ser tratada com muito cuidado.

Figura 2 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável temperatura (°C)



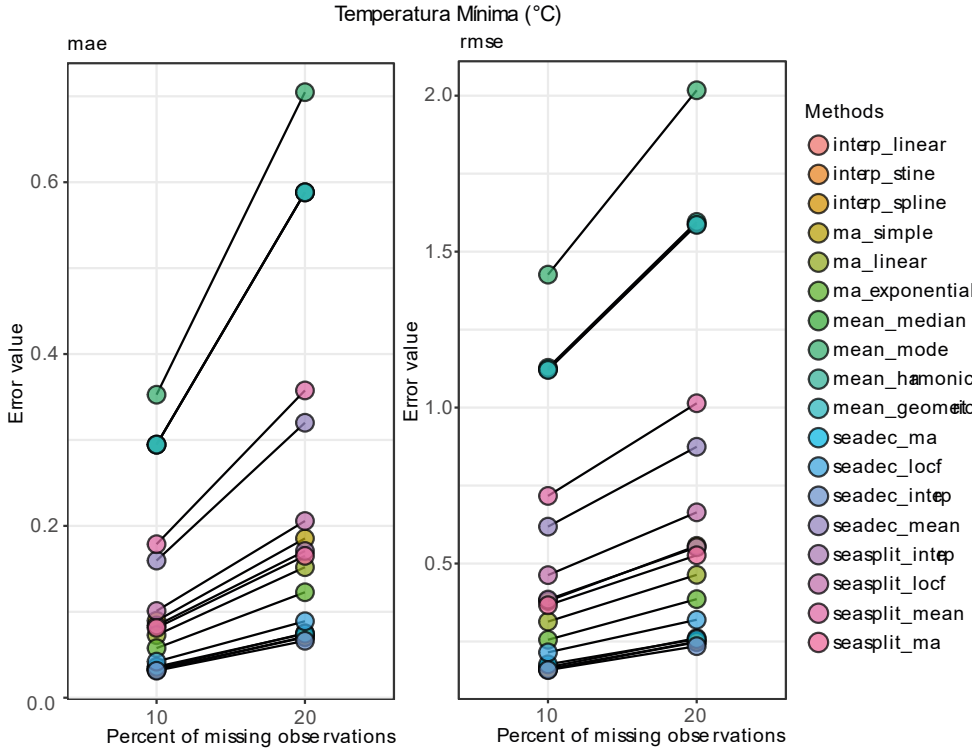
Fonte: Autores (2023).

Figura 3 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável radiação (MJ/m².dia).



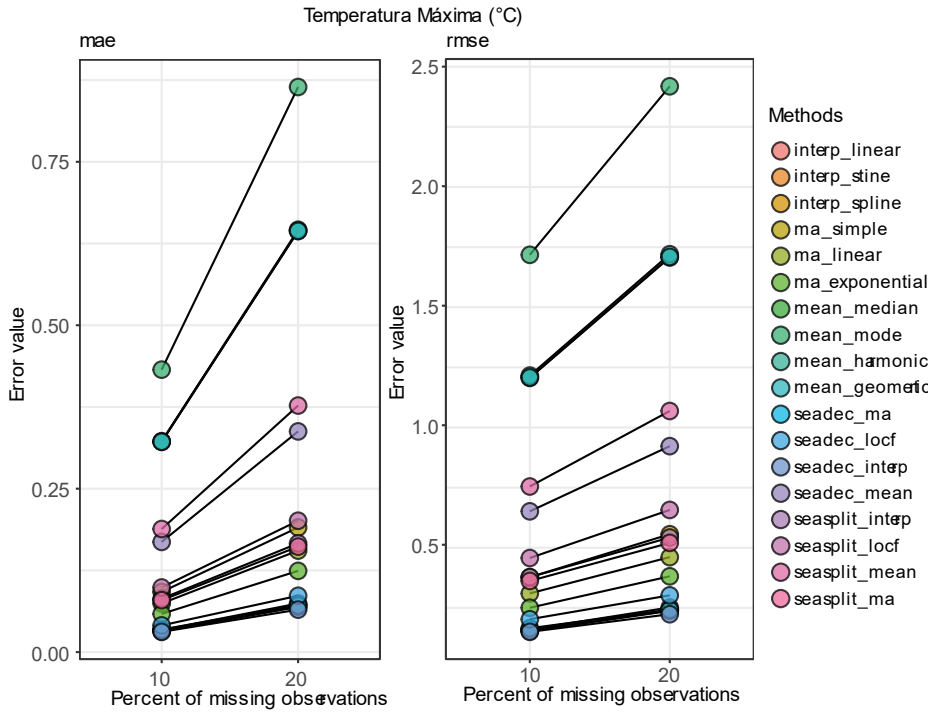
Fonte: Autores (2023).

Figura 4 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável temperatura mínima (°C).



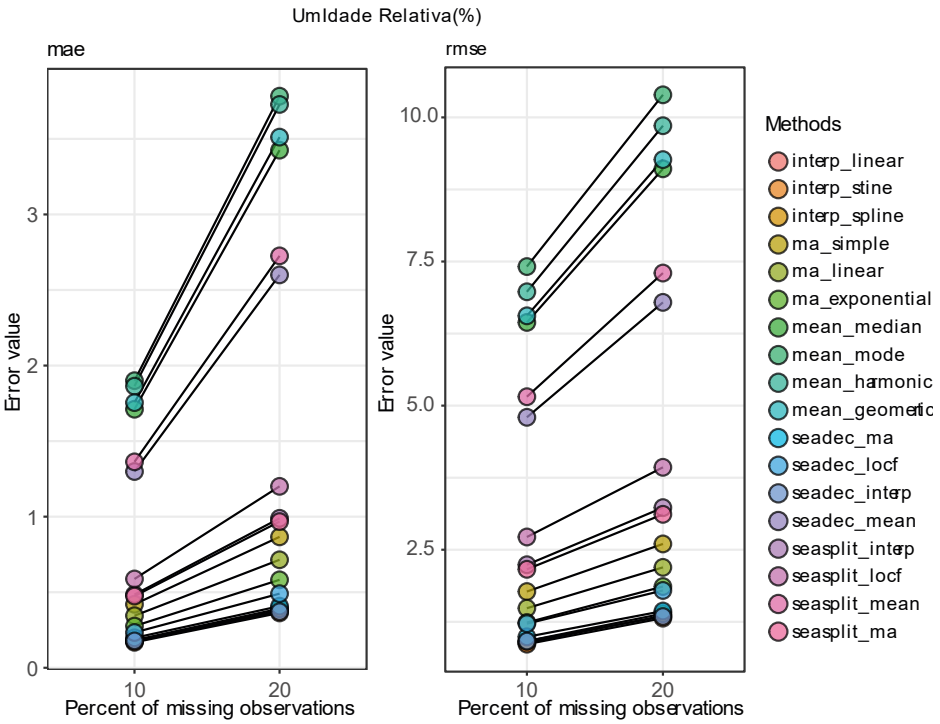
Fonte: Autores (2023).

Figura 5 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável temperatura máxima (°C).



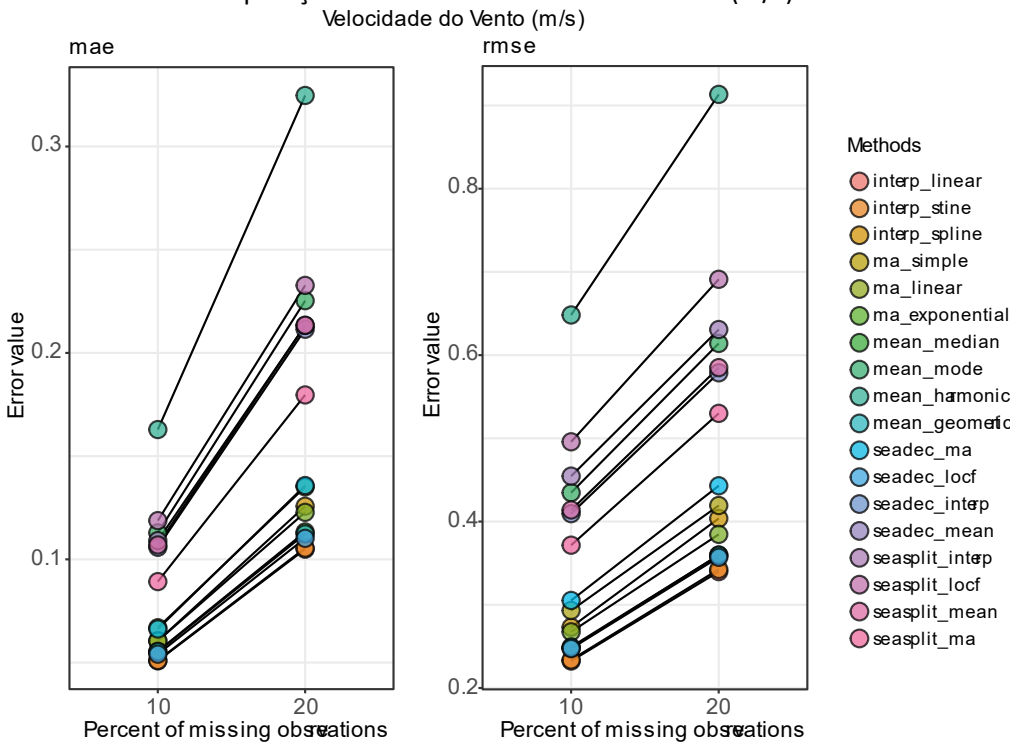
Fonte: Autores (2023).

Figura 6 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável umidade relativa (%).



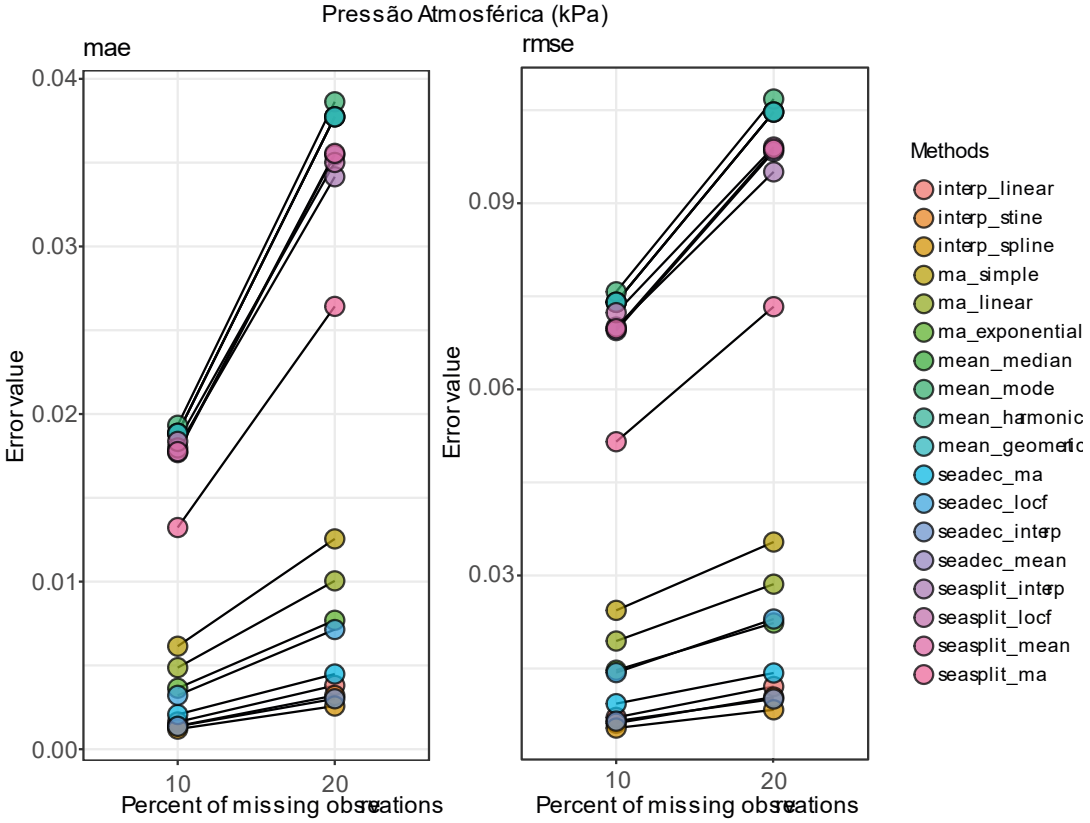
Fonte: Autores (2023).

Figura 7 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável velocidade do vento (m/s).



Fonte: Autores (2023).

Figura 8 - Erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) para cada método de imputação da variável pressão atmosférica (kPa).



Fonte: Autores (2023).

Olhando para todos os métodos de imputação testados, nenhum método pode ser apontado como o melhor globalmente. O desempenho do método é sempre muito dependente das características da série temporal de entrada.

Embora não apresentem um padrão, a pequena variação entre os valores de erro dos melhores métodos justifica a utilização do método "seadec_interp" para a imputação de todos os conjuntos de dados das variáveis faltantes, visando simplificar o algoritmo computacional para imputação de todas as séries de dados testados.

3.2. Séries históricas de dados meteorológicos com imputação dos valores ausentes

A Tabela 4 apresenta um resumo estatístico-descritivo das variáveis meteorológicas que compõe o banco de dados. Foram estudadas a radiação solar acumulada diária, a distribuição média diária da pressão atmosférica, a média diária dos percentuais de umidade relativa, média diária da velocidade do vento e temperaturas médias, máximas e mínimas, diárias.

Tabela 4 - Estatísticas descritivas para as variáveis meteorológicas extraídas da estação automática de Iguatu-CE, no período de 2014-2022.

	Mínimo	Média	Mediana	Máximo	Coeficiente de Variação
Radiação (MJ/m²dia)	2,51	13,40	13,73	17,51	0,16
Pressão atmosférica (kPa)	98,15	98,70	98,68	99,29	0,00
Umidade relativa (%)	27,48	57,81	55,08	96,00	0,24
Velocidade do vento (m/s)	0,47	2,56	2,55	5,45	0,33
Temperatura (°C)	22,08	27,95	27,84	34,37	0,06
Temperatura mínima (°C)	17,10	23,37	23,00	32,35	0,10
Temperatura máxima (°C)	25,93	33,45	33,30	40,40	0,08

Fonte: Autores (2023).

A radiação média acumulada diariamente é de 13,396 MJ/m²dia, chegando ao valor máximo de 17,513 MJ/m²dia, nos dias com maior incidência solar. Em relação à pressão

atmosférica, por mais que a estação meteorológica seja fixa em um local, os valores apresentam pequenas oscilações, pois estão sujeitas a variações horárias (máximos e mínimos) e à diferença entre as estações do ano. Essa pequena variação é refletida no valor do coeficiente de variação, que é o mais baixo dentre as variáveis estudadas.

Quanto à umidade relativa, a média do período é de 59,97% e o menor valor registrado foi de 27.48%. Vale ressaltar que os valores de umidade relativa do ar apresentados são referentes à média diária, compreendendo as 24 horas do dia. A velocidade média do vento é a variável que apresenta maior coeficiente de variação, ou seja, é a que mais varia em relação à média da amostra.

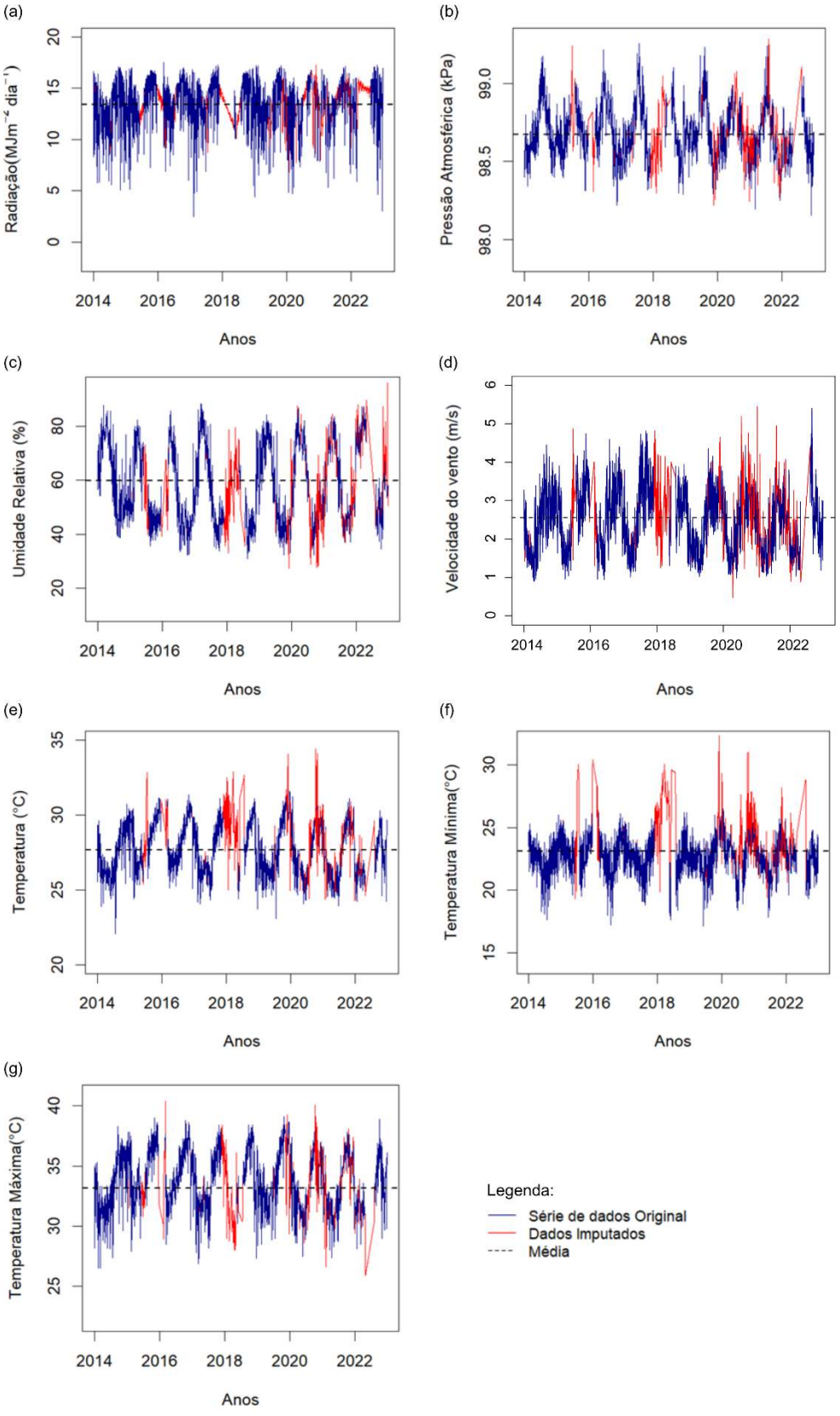
A temperatura média para o período 2014-2022 é de 27,68 °C, sendo o dia mais frio registrado com uma média diária de 22.08°C, e o mais quente com a média diária igual a 34,37°C. Entretanto, quando se trata de picos máximos e mínimos de temperatura, a mínima registrada no período de estudo foi de 15,7°C, enquanto a máxima foi de 40,4 °C.

A Figura 9 ilustra a distribuição diária das variáveis considerando o período de tempo de 2014 a 2022, para ilustrar o padrão observado em cada parâmetro meteorológico, medido na região em estudo. A série de dados original está representada pela cor azul e os trechos que apresentavam falhas e tiveram seus valores imputados estão destacados na cor vermelha. A linha tracejada preta, identifica o valor da média de cada série.

As séries históricas possuem um comportamento sazonal. Podendo ser visualizada de forma mais clara na umidade relativa, Figura 9 (c), e nas temperaturas média e máxima, Figura 9 (e) e Figura 9 (g). Essa característica, atrelada ao clima semiárido no qual se encontra o local de estudo, se justifica pela alternância entre os períodos chuvosos e secos. Nos períodos chuvosos, mais concentrados no início de cada ano, as temperaturas tendem a cair, já quando começa o período de estiagem é normal que elas voltem a aumentar.

Os gráficos apresentados na Figura 9 ressaltam o bom funcionamento do preenchimento de dados ausentes, na maior parte da série. Nos últimos dias do mês de abril, nos meses de maio, junho, julho e alguns dias do mês de agosto de 2022, a estação meteorológica não registrou nenhuma medição para as variáveis, isso afetou o funcionamento do preenchimento de falhas. É possível perceber que, nesse período, o algoritmo não consegue acompanhar o comportamento da série original, preenchendo os dados de forma linear.

Figura 9 - Série histórica dos dados meteorológicos coletados na EMA de Iguatu-CE, com valores faltantes imputados.



Fonte: Autores (2023).

4. CONSIDERAÇÕES FINAIS

Neste estudo, foram comparados diversos métodos de imputação em dados meteorológicos, considerando duas diferentes taxas de falhas. Os resultados médios alcançados por esses métodos demonstraram similaridade. O método “seadec_interp” obteve as menores médias de MAE e RMSE para as três variáveis de temperatura. Os dados de pressão atmosférica, umidade relativa e velocidade do vento tiveram melhor êxito com os métodos “interp_spline”, “interp_stine” e “interp_linear”, respectivamente. Já os dados de radiação apresentaram um melhor comportamento com o método “ma_exponencial”.

Os métodos analisados demonstraram uma habilidade consistente na estimativa dos valores ausentes na série de dados meteorológicos, para as duas situações de falhas investigadas. A abordagem metodológica utilizada, baseada nos princípios dos mecanismos de dados faltantes e simulações, pode ser empregada em pesquisas similares realizadas em diversas localidades, independentemente das variações climáticas. A metodologia empregada possibilita a escolha da técnica mais adequada dentre várias opções, considerando diferentes cenários de falhas, com base nas métricas de erro utilizadas.

Em pesquisas futuras, deve-se explorar a aplicação de métodos multivariados para imputação de dados faltantes em séries temporais meteorológicas. Além disso, poderão ser investigadas abordagens combinadas, univariadas e multivariadas, para o preenchimento de dados ausentes, assim como também, ampliar a gama de medidas de desempenho para enriquecer a escolha do método.

REFERÊNCIAS

BECK, Marcus W.; BOKDE, Neeraj Dhanraj; ASENCIO-CORTÉS, Gualberto; KULAT, K.D. R Package imputeTestbench to Compare Imputation Methods for Univariate Time Series. **The R Journal**, v.10, p.1-16, 2018.

BIER, Anderson Augusto; FERRAZ, Simone Erotildes Teleginski. Comparação de Metodologias de Preenchimento de Falhas em Dados Meteorológicos para Estações no Sul do Brasil. **Revista Brasileira de Meteorologia**, v. 32, n. 2, p. 215-226, 2017. Disponível em: https://doi.org/10.1590/0102-77863220008.

CLEVELAND, Robert B.; CLEVELAND, William S.; MCRAE, Jean E.; TERPENNING, Irma. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. **Journal of Official Statistics**, v. 6, n. 1, p. 3-73, 1990.

CORREA, Marcele de Jesus; LIMA, Kellen Carla; SILVA, Jonathan Mota da; MEDEIROS, Gilvandro César. Filling of faults in climatological air temperature series in brazilian state capitals from 1980 to 2017. **Revista Brasileira De Climatologia**, v. 29, p. 251–272, 2021. Disponível em: https://ojs.ufgd.edu.br/index.php/rbclima/article/view/15176.

CORREIA, Tamíres Partélli; DOHLER, Rafael Esteves; DAMBROZ, Carlos Silva; BINOTI, Daniel Henrique Breda. Aplicação de Redes Neurais Artificiais no Preenchimento de Falhas de Precipitação Mensal na Região Serrana do Espírito Santo. **Geociências**, v. 35, n. 4, p.560-567, 2016.

CUNHA JÚNIOR, Rubens Oliveira da; FIRMINO, Paulo Renato Alves. Simulação de valores ausentes em séries temporais de precipitação para avaliação de métodos de imputação. **Revista Brasileira De Climatologia**, v.30, n.18, p.691–714. 2022. Disponível em: https://doi.org/10.55761/abclima.v30i18.15243.

DOURADO, Wesley Barbosa. **Avaliação de técnicas de interpolação de imagens digitais**. 2014. 141 f. Dissertação (Mestrado em Matemática Aplicada e Computacional) - Universidade Estadual Paulista, Presidente Prudente. 2014. Disponível em: https://repositorio.unesp.br/server/api/core/bitstreams/a4a2c492-32c3-4a27-8573-d5f543980736/content. Acesso em: 27 set. 2023.

GNAUCK, Albrecht. Interpolation and approximation of water quality time series and process identification**. Analytical and bioanalytical chemistry**, v. 380, p. 484-492, 2004. Disponível em: https://doi.org/10.1007/s00216-004-2799-3.

HODSON, Timothy O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. **Geoscientific Model Development**, v. 15, n. 14, p. 5481-5487, 2022. Disponível em: https://doi.org/10.5194/gmd-15-5481-2022, 2022.

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ – IPECE. **Perfil Municipal 2017:** Iguatu. Fortaleza: IPECE, 2017.

JUNQUEIRA, Rubens; AMORIM, Jhones da Silva; OLIVEIRA, Alisson Souza de. Comparação entre diferentes metodologias para preenchimento de falhas em dados pluviométricos. **Sustentare**, v.2, n.1, p. 198-210, 2018.

MACHIWAL, Deepesh; JHA, Madan Kumar Jha. **Hydrologic time series analysis:** theory and practice. Springer Science & Business Media, 2012.

MELLO, Yara Rúbia de, KOHLS, Werner; OLIVEIRA, Therezinha Maria Novais de. Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas. **Boletim de Geografia**, v. 35, n.1, p.112-121, 2017. Disponível em: https://doi.org/10.4025/bolgeogr.v35i1.30893.

MORITZ, Steffen; BARTZ-BEIELSTEIN, Thomas. imputeTS: Time Series Missing Value Imputation in R. **R Journal**, v. 9, n. 1, p. 207, 2017.

OLIVEIRA, Luiz F. C. de; FIOREZE, Ana P.; MEDEIROS, Antonio M. M.; SILVA, Mellissa A. S. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação

pluvial anual. **Revista Brasileira De Engenharia Agrícola e Ambiental**, v. 14, n. 11, p. 1186-1192, 2010. Disponível em: https://doi.org/10.1590/S1415-43662010001100008.

OLIVEIRA, Thiago Alves; SANCHES, Fabio de Oliveira; FERREIRA, Cássia de Castro Martins. (2021). Aplicação e avaliação de técnicas para o preenchimento de falhas de dados pluviométricos em anos habituais, secos e chuvosos. **ENTRE-LUGAR**, v.12, n. 24, p. 301–320. Disponível em: https://doi.org/10.30612/rel.v12i24.15137.

R CORE TEAM. **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2023. URL: <https://www.R-project.org/>.

RSTUDIO TEAM. **RStudio:** Integrated Development for R. RStudio, PBC, Boston, MA, 2023. URL: http://www.rstudio.com.

SABINO, Marlus; SOUZA, Adilson P. de. Gap-filling meteorological data series using the GapMET software in the state of Mato Grosso, Brazil. **Revista Brasileira De Engenharia Agrícola e Ambiental**, v.27, n.2, p.149–156, 2023. Disponível em: https://doi.org/10.1590/1807-1929/agriambi.v27n2p149-156.

TURICCHI, Jake; O'DRISCOLL, Ruairi; FINLAYSON, Graham; DUARTE, Cristiana; PALMEIRA, A. L.; LARSEN, Sofus. C.; HEITMANN, Berit L.; STUBBS, R James. Data imputation and body weight variability calculation using linear and nonlinear methods in data collected from digital smart scales: simulation and validation study. **JMIR Mhealth Uhealth**, v. 8, n. 9, p. e17977, 2020. Disponível em: https://doi.org/10.2196/17977.

WIJESEKARA, Lakmini, LIYANAGE, Liwan. Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index. In: ARAI, K., KAPOOR, S., BHATIA, R. (eds) **Advances in Information and Communication:** Proceedings of the 2020 Future of Information and Communication Conference (FICC). Springer International Publishing, vol 1130, p. 257-269, 2020. Disponível em: https://doi.org/10.1007/978-3-030-39442-4_20.

ZENERE, Pedro Vinicius S.; VENTURA, Thiago M.; GOMES, Raphael S. R.; RODRIGUES, Thiago R. Uso de árvore de decisão para escolha de método de preenchimento de falhas em dados meteorológicos. In: BRAZILIAN E-SCIENCE WORKSHOP (BRESCI), 14, 2020, Cuiabá. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, p. 89-96, 2020.