Revista Brasileira de Climatologia
Brazilian Journal of Climatology

OPEN ACCESS

DOI: 10.5380/abclima

SIMULAÇÃO DE VALORES AUSENTES EM SÉRIES TEMPORAIS DE PRECIPITAÇÃO PARA AVALIAÇÃO DE MÉTODOS DE IMPUTAÇÃO

Missing values simulation in rainfall time series for evaluating imputation methods

Simulación de valores faltantes en series de tiempo de precipitación para la evaluación de métodos de imputación

Rubens Oliveira da Cunha Júnior
Centro de Ciências Agrárias e da Biodiversidade, Universidade Federal do Cariri (UFCA)
E-mail: cunhajunior.rubens@gmail.com

Paulo Renato Alves Firmino
Centro de Ciências e Tecnologia, Universidade Federal do Cariri (UFCA)
E-mail: paulo.firmino@ufca.edu.br

Resumo: Dados ausentes em séries temporais de precipitação são um dos principais problemas em estudos hidrológicos. Neste sentido, as técnicas de preenchimento de falhas constituem uma ferramenta importante para a reconstrução de conjuntos de dados pluviométricos. O objetivo do presente trabalho foi comparar diferentes métodos de preenchimento de falhas em séries mensais de precipitação. Como caso de estudo, foram consideradas séries temporais de 1974 a 2004 de estações pluviométricas localizadas na região do Cariri, Ceará, Brasil. Para a imputação dos valores ausentes, foram aplicados métodos como média aritmética (MA), inverso da potência da distância (IPD), ponderação regional (PR), regressão linear múltipla (RLM) e redes neurais artificiais (RNA). Utilizando os conceitos de mecanismos de ausência de dados, foram realizadas simulações de valores ausentes gerados artificialmente para diferentes porcentagens de falhas, a saber, 10% e 40%. O desempenho dos métodos de imputação foi avaliado por métricas de erro como a raiz do erro quadrático médio (REQM) e o erro absoluto médio (EAM). A sazonalidade do regime pluviométrico das séries também foi considerada. Numericamente, o método RNA obteve as menores médias de REQM e EAM, seguido pelos métodos RLM, PR, MA e IPD. Contudo, os valores médios obtidos por todos os métodos foram semelhantes. Os métodos avaliados foram capazes de estimar com boa precisão os dados faltantes na série pluviométrica estudada.

Palavras-chave: Imputação de dados ausentes. Regressão linear múltipla. Redes neurais artificiais. Hidrologia.

Abstract: Missing data in rainfall time series is one of the main problems in hydrological studies. In this regard, gap-filling techniques are an important tool for reconstructing rainfall data sets. This paper aims to compare different gap-filling methods for monthly rainfall time series. As a case study, time series ranging from 1974 to 2004 from meteorological stations of the Cariri region, Ceará, Brazil, were considered. For the imputation of missing values, methods such as the arithmetic average (AA), inverse distance weighting (IDW), regional weighting (RW), multiple linear regression (MLR), and artificial neural networks (ANN) were applied. Simulation of artificially generated missing values was performed using concepts of missing data mechanisms for different missing values rates, namely, 10% and 40%. The performance of the imputation methods was evaluated by error metrics such as the root mean squared error (RMSE) and mean absolute error (MAE). The seasonality of rainfall patterns was also considered. Numerically, the ANN method achieved the lowest RMSE and MAE averages, followed by the MLR, RW, AA, and IDW methods. However, the average values obtained by all methods were similar. The methods evaluated were able to estimate the missing values in the studied time series with good accuracy.
Keywords: Missing data imputation. Multiple linear regression. Artificial neural networks. Hydrology.

Resumen: Los datos faltantes en series de tiempo de precipitación son un problema en la hidrología. En este sentido, las técnicas de llenado de fallas constituyen una herramienta importante para completar conjuntos de datos de precipitación. El objetivo de este trabajo fue comparar diferentes métodos de llenado de fallas en series de tiempo mensuales de lluvia. Se tomaron como estudio de caso las series de tiempo de estaciones meteorológicas de la región de Cariri, Ceará, Brasil, considerando el período de 1974 a 2004. Se utilizaron los métodos media aritmética (MA), ponderación de distancia inversa (DIP), ponderación regional (PR), regresión múltiple (RLM) y redes neuronales artificiales (RNA) para la imputación. Usando los conceptos de mecanismos de datos faltantes, se realizaron simulaciones de valores perdidos generados artificialmente para diferentes porcentajes de fallas, a saber, 10% y 40%. La evaluación del rendimiento de los métodos de imputación se realizó mediante la raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE). También se consideró la estacionalidad del régimen de lluvias de las series. Numéricamente, el método RNA ha logrado los menores valores medios de RMSE y MAE, seguido por los métodos RLM, PR, MA y DIP. Sin embargo, los valores medios obtenidos por todos los métodos fueron similares. Los métodos evaluados estimaron los datos faltantes en las series de tiempo estudiadas con buena exactitud.
Palabras clave: Imputación de datos faltantes. Regresión múltiple. Redes neuronales artificiales. Hidrología.

Submetido em: 11/10/2021
Aceito para publicação em: 08/06/2022
Publicado em: 10/06/2022

1. INTRODUÇÃO

A precipitação é uma das mais importantes variáveis hidrológicas. O estudo de dados pluviométricos é essencial para a modelagem e previsão de fenômenos hidrológicos, para o correto entendimento das variações climáticas e para a gestão sustentável dos recursos hídricos (MEKIS *et al*., 2018). Contudo, é comum a ocorrência de períodos sem informações nos dados de precipitação, devido a diversos fatores, tais como erros humanos ou falhas nos equipamentos de monitoramento. Esses valores em falta são um importante problema na análise e modelagem de processos hidrológicos (NAGHETTINI; PINTO, 2017). Entre os métodos para lidar com o problema das falhas, abordagens para descartar os registros com falhas reduzem a quantidade de informações disponíveis para a análise. Portanto, pesquisadores têm se dedicado ao desenvolvimento de estratégias para se estimar de maneira razoável valores para preencher tais falhas, também conhecidas como métodos de imputação de valores ausentes (LITTLE; RUBIN, 2019).

Para se avaliar o desempenho de um método de preenchimento de falhas, o procedimento usualmente descrito na literatura consiste em se obter um conjunto de dados completo, isto é, sem falhas, e então realizar simulações a partir de valores ausentes gerados artificialmente (TWALA, 2009; MORITZ *et al*., 2015). Para dados de precipitação, as falhas podem ser estimadas usando os dados de estações meteorológicas vizinhas. Existem diversos métodos para a imputação de dados ausentes em séries de precipitação (KARAMOUZ; NAZIF; FALAHI, 2012). Autores como Brubacher, Oliveira e Guasselli (2020) destacaram em uma revisão de literatura os principais desafios e perspectivas das técnicas de preenchimento de falhas em dados pluviométricos.

Métodos baseados em interpolação espacial, tais como a média aritmética (MA), ponderação regional (PR) e a interpolação pelo inverso da potência da distância (IPD), são os mais tradicionalmente escolhidos para a estimativa de falhas em dados de chuva (TUCCI, 2001; RADI; ZAKARIA; AZMAN, 2015). Estas abordagens têm sido tema de estudos realizados por diversos autores (TEEGAVARAPU; CHANDRAMOULI, 2005; KIM; RYU, 2016; BIELENKI JUNIOR *et al*., 2018; AIEB *et al*., 2019). Além disso, métodos baseados em regressão se destacam entre as técnicas estatísticas mais amplamente usadas para imputação de maneira geral (LIN; TSAI, 2020). Em especial, para o método de regressão linear múltipla (RLM), as informações de

precipitação em uma dada estação são correlacionadas com os dados correspondentes em estações vizinhas através de modelos lineares. A RLM é um dos métodos mais aplicados para preenchimento de falhas em dados de precipitação (EISCHEID *et al*., 2000; GAO *et al*., 2018).

Quanto às inovações neste sentido, merece destaque a utilização de técnicas baseadas em inteligência artificial e aprendizado de máquina (MEKANIK *et al*., 2013). 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). O aprendizado de máquina é um ramo da inteligência artificial amplamente utilizado em diversas áreas em problemas de classificação e regressão. Técnicas como as redes neurais artificiais (RNA) são algoritmos que possuem a capacidade de aprender através de um processo de treinamento e, baseado nisto, generalizar o aprendizado adquirido para prever cenários ou eventos futuros (AWAD; KHANNA, 2015). As RNA têm sido aplicadas para o preenchimento de falhas em dados de precipitação por autores como Kashani e Dinpashoh (2012), Oliveira *et al*. (2013), Correia *et al*. (2016) e Ruezzene *et al*. (2021).

Diante das considerações realizadas, o presente estudo busca investigar a aplicação de diferentes metodologias de preenchimento de falhas em séries temporais de precipitação. Para este fim, foram realizadas simulações de dados ausentes gerados artificialmente. A qualidade e a precisão dos métodos foram avaliadas por meio de métricas de erro conhecidas.

2. MÉTODOS

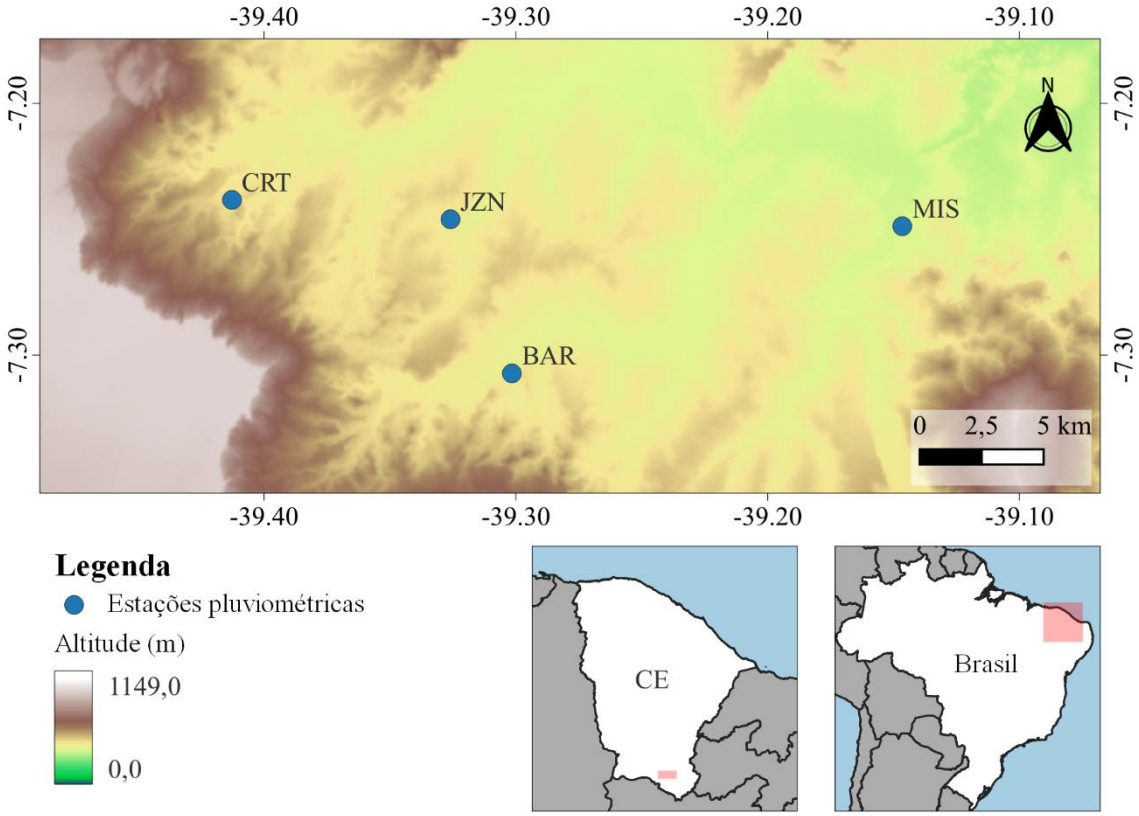
2.1. Caso de estudo e conjunto de dados

O presente estudo foi aplicado na região nordeste do Brasil, no sul do estado do Ceará, compreendendo municípios da Região Metropolitana do Cariri. A região tem clima tropical quente semiárido brando, tropical quente e tropical quente subúmido. O relevo é constituído por dois domínios principais, planalto e depressão, conhecidos como Chapada do Araripe e Vale do Cariri, respectivamente (COGERH, 2009).

Séries de precipitação entre janeiro de 1974 e dezembro de 2004 medidas nas estações pluviométricas de Crato (CRA), Juazeiro do Norte (JZN), Barbalha (BAR) e Missão Velha (MIS) foram consideradas. Os dados foram obtidos a partir da Fundação Cearense de Meteorologia (FUNCEME, 2021). O critério de seleção das estações foi não possuir falhas nos dados durante

o período considerado. A Figura 1 mostra a localização das estações pluviométricas utilizadas no presente estudo.

Figura 1 - Mapa de localização das estações pluviométricas: Barbalha - BAR, Crato - CRA, Juazeiro do Norte - JZN e Missão Velha - MIS, Ceará, Brasil.



Fonte: Elaborado pelos autores (2021).

O regime pluviométrico da região é marcado pela irregularidade interanual e variabilidade espaço-temporal. As chuvas se iniciam na chamada pré-estação chuvosa, referente aos meses de dezembro e janeiro de cada ano, mas se concentram em uma estação chuvosa, compreendida entre os meses de fevereiro a maio (TEIXEIRA, 2003).

2.2. Métodos de imputação para dados de precipitação

Média aritmética simples (MA)

Trata-se do método mais simples comumente utilizado para preenchimento de falhas em dados de precipitação. Os valores são estimados através do cálculo da média aritmética simples dos dados correspondentes nas estações vizinhas, através da Equação (1).

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

em que y_i é o valor ausente estimado, x_i é a precipitação correspondente medida na *i*ª estação mais próxima e *n* é o número de estações consideradas. O método MA é satisfatório para estações uniformemente distribuídas (SATTARI; REZAZADEH-JOUDI; KUSIAK, 2017).

Ponderação regional (PR)

O método da ponderação regional (PR) é um método simplificado utilizado para a imputação de valores ausentes em séries mensais ou anuais. Na sua aplicação, são selecionadas pelo menos 3 estações climaticamente homogêneas e que possuam no mínimo 10 anos de dados. As falhas são preenchidas segundo a Equação (2).

$$y_i = \frac{\bar{y}}{n}\left(\sum_{i=1}^{n} \frac{x_i}{\bar{x}_i}\right), \quad (2)$$

em que $\bar{y}$ é a precipitação média na estação que possui falhas e $\bar{x}_i$ é a precipitação média na *i*ª estação vizinha. Os outros parâmetros são dados conforme a Equação (1) (TUCCI, 2001).

Inverso da potência da distância (IPD)

O método IPD é o mais comumente utilizado para estimar valores ausentes em séries pluviométricas. As falhas são preenchidas através de ponderações dos valores observados em outras estações, segundo a distância entre elas. A Equação (3) mostra o método IPD.

$$y_i = \frac{\sum_{i=1}^{n}\left(\frac{x_i}{d_i^b}\right)}{\sum_{i=1}^{n}\left(\frac{1}{d_i^b}\right)}, \quad (3)$$

em que d_i é a distância entre a estação com falhas e a *i*ª estação vizinha e b é a potência da distância. Os demais parâmetros são definidos iguais aos da Equação (1). No presente trabalho, o método IPD foi aplicado utilizando-se o parâmetro b = 1 (HARMAN; KOSEOGLU; YIGIT, 2016; SATTARI; REZAZADEH-JOUDI; KUSIAK, 2017).

Regressão linear múltipla (RLM)

Um dos métodos mais utilizados na hidrologia, os modelos de RLM se baseiam nas correlações lineares existentes entre uma variável dependente e mais de uma variável independente. Os dados não devem possuir colinearidade entre as variáveis independentes, isto é, não deve haver correlação entre as variáveis. Para o preenchimento de falhas, um modelo geral de RLM pode ser expresso segundo a Equação (4).

$$y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \quad (4)$$

em que β_0, β_1, ..., β_n são parâmetros da equação linear a serem estimados e x_1, x_2, ..., x_n são as variáveis independentes, referentes às precipitações nas *n* estações consideradas (MAITY, 2018; ASGHARINIA; PETROSELLI, 2020).

Redes neurais artificiais (RNA)

As RNAs são sistemas de processamento baseados no funcionamento do cérebro humano. Tratam-se de modelos com a capacidade de aprender a partir do conhecimento experimentado e, com base nisso, generalizar tal conhecimento adquirido, a fim de realizar previsões de cenários ou eventos futuros. O seu funcionamento se dá através de uma rede de interconexões entre unidades básicas de processamento, análogas aos neurônios. As redes perceptron multicamadas (MLPs, do inglês *Multilayer Perceptron*), amplamente utilizadas em estudos de séries temporais, são formadas por nós interconectados e dispostos em diversas camadas, de entrada, ocultas e de saída, de modo que cada camada se conecta à camada posterior (HAYKIN, 1999; AWAD; KHANNA, 2015).

A determinação da arquitetura da rede MLP é uma etapa fundamental para a aplicação de RNAs. O número de nós da camada de entrada é dado pelo número de variáveis explanatórias do modelo. Entretanto, a determinação do número de nós na camada oculta é um processo de tentativa e erro. Para a previsão um passo adiante, é necessário apenas um nó na camada de saída (ZHANG, 2001; PALIT; POPOVIC, 2005). A Equação (5) expressa matematicamente uma rede MLP com uma camada oculta.

$$y_i = f\left(\sum_{h}^{n_h} w_{ho} \cdot f\left(\sum_{i}^{n_i} w_{ih} x_i + b_h\right) + b_o\right), \quad (5)$$

em que os índices i, h e o se referem às camadas de entrada, oculta e de saída, respectivamente. Por sua vez, y_i e x_i são a entrada e a saída, e n_i e n_h são o número de nós nas camadas de entrada e oculta, respectivamente. Os coeficientes w são os pesos das conexões entre os nós e b são constantes de viés (*bias*). A função f é chamada função de ativação, que permite a aplicação da rede para processos não lineares. Entre as funções de ativação comumente utilizadas, destaca-se a função sigmoide $f(x) = \frac{1}{1+exp(-x)}$ (LEE; LEE; YOON, 2019). O aprendizado de uma RNA se dá através de uma etapa de treinamento. Entre os algoritmos de aprendizado supervisionado, o algoritmo *backpropagation* é amplamente utilizado no treinamento de redes MLP (PALIT; POPOVIC, 2005).

A ideia de se aplicar RNA para preenchimento de falhas consiste em se utilizar todos os registros que não possuam falhas do conjunto de dados para treinar o algoritmo e, baseado nisto, estimar os registros que apresentem falhas (GUPTA; LAM, 1996). Como o desempenho de técnicas baseadas em aprendizado de máquinas é influenciado pela quantidade de dados usada na etapa de treinamento, recomenda-se a utilização de todos dados completos disponíveis para treinar o algoritmo (HONGHAI *et al*., 2005).

2.3. Simulação de valores ausentes

A distribuição dos valores ausentes em um conjunto de dados depende das causas desta ausência. Os mecanismos da ausência de dados descrevem as possíveis relações entre as variáveis e a probabilidade de ausência dos dados. Portanto, compreender os mecanismos da ausência de dados pode ser útil para a correta escolha dos métodos de análise e para determinar meios apropriados para simular valores ausentes (MORITZ *et al*., 2015). Três mecanismos de ausência de dados são descritos na literatura. O mecanismo de Ausência Completamente Aleatória (MCAR, do inglês *Missing Completely at Random*) considera que os valores ausentes não têm relação com nenhuma variável. A Ausência Aleatória (MAR, do inglês *Missing at Random*) ocorre quando as falhas têm relação com outras variáveis, mas é independente da própria variável. O mecanismo de Ausência Não Aleatória (MNAR, do inglês *Missing Not at Random*) considera que os valores ausentes estão relacionados a outros valores ausentes (LITTLE; RUBIN, 2019).

Seja um conjunto de dados com falhas, podem-se definir dois subconjuntos: um subconjunto formado pelos valores observados Y_{obs} e um subconjunto composto pelos registros com pontos de dados ausentes Y_{aus}. A ausência dos dados pode ser considerada como uma variável M, cuja distribuição de probabilidades indica se um determinado ponto é ausente (M = 0) ou observado (M = 1). Portanto, para cada ponto do conjunto de dados, observado ou ausente, existe um valor de M correspondente associado. Baseado nisso, os mecanismos podem ser expressos em termos de distribuições de probabilidades, conforme as Equações (6), (7) e (8) (LITTLE; RUBIN, 2019).

$$MCAR = p(M|Y_{obs}, Y_{aus}) = p(M) \quad (6)$$

$$MAR = p(M|Y_{obs}, Y_{aus}) = p(M|Y_{obs}) \quad (7)$$

$$MNAR = p(M|Y_{obs}, Y_{aus}) = p(M|Y_{obs}, Y_{aus}) \quad (8)$$

A avaliação do desempenho de um método de imputação requer um conjunto de dados completo e a geração de valores ausentes através de simulações (MORITZ *et al*., 2015). O mecanismo MAR é apropriado para simular situações onde as falhas em uma série temporal são causadas devido a falhas em equipamentos durante longos períodos de tempo. Em implementações do mecanismo MAR, a probabilidade de um dado ponto ser ausente depende de os pontos mais próximos também serem ausentes (BECK *et al*., 2018). A suposição do mecanismo MAR é apropriada para dados meteorológicos (JUNNINEN *et al*., 2004; GÓMEZ-CARRACEDO *et al*., 2014; JUNGER; DE LEON, 2015) e de precipitação (AIEB *et al*., 2019).

2.4. Implementação computacional e experimentação

Este trabalho teve como objetivo investigar o desempenho de métodos de preenchimento de falhas em dados de precipitação sob diferentes porcentagens de valores ausentes. Um conjunto de séries temporais completas foi usado e as falhas foram geradas artificialmente por meio de procedimentos de simulação. A qualidade e a precisão dos métodos foram examinadas de acordo com métricas de erro conhecidas.

Todas as análises e simulações foram implementadas usando o ambiente de programação estatística R (TEAM, 2021). O pacote *imputeTestBench* (BECK *et al*., 2018) foi usado para as simulações dos valores ausentes, e as redes neurais artificiais foram treinadas utilizando o pacote *neuralnet* (GÜNTHER; FRITSCH, 2010).

As falhas foram simuladas na série da estação pluviométrica Juazeiro do Norte - JZN e diversos métodos, tais como média aritmética simples (MA), regressão linear múltipla (RLM), ponderação regional (PR), inverso da potência da distância (IPD) e redes neurais artificiais (RNA), foram aplicados para estimar os valores ausentes, usando como variáveis explanatórias os dados nas estações vizinhas. O mecanismo MAR foi adotado para simular as falhas na série JZN. Porcentagens de falhas de 10% e 40% foram usadas, de modo que os valores ausentes foram tomados como blocos contínuos, cujos tamanhos variam de 25% a 100% do total da ausência. Para cada porcentagem, todos os métodos foram testados 50 vezes, totalizando 500 procedimentos de imputação (5 métodos × 2 porcentagens de falhas × 50 repetições).

Para a aplicação das RNAs, foram testadas diferentes arquiteturas de redes MLP, em relação às camadas de entrada e oculta. O número de nós na camada de entrada foi avaliado testando-se 7 configurações, referentes às possíveis combinações formadas pelas séries das 3 estações vizinhas à estação Juazeiro do Norte, a saber, CRA, BAR, MIS, CRA+BAR, CRA+MIS, BAR+MIS e CRA+BAR+MIS. Para a camada oculta, foram testadas redes contendo de 1 a 10 nós. As redes foram construídas utilizando-se um nó na camada de saída. A função de ativação utilizada foi a função logística, cuja imagem são valores no intervalo [0;1], para evitar a estimativa de valores de precipitação negativos. O algoritmo de treinamento usado foi o *backpropagation* resiliente (parâmetro “rprop+”) (GÜNTHER; FRITSCH, 2010). O conjunto de dados foi transformado através de uma normalização do tipo Mín-Máx para o intervalo [0,4;0,6], a fim de adequar os dados às entradas da rede (AYDILEK; ARSLAN, 2013). Após a imputação, os dados foram transformados de volta ao intervalo original.

2.5. Avaliação do desempenho

A qualidade dos métodos de imputação de dados ausentes foi avaliada através de parâmetros estatísticos como a raiz do erro quadrático médio (REQM) e o erro absoluto médio (EAM), mostrados nas Equações (9) e (10), respectivamente.

$$REQM = \sqrt{\frac{\sum_{t=1}^{n}(\widehat{y_t} - y_t)^2}{n}}, \tag{9}$$

$$EAM = \frac{\sum_{t=1}^{n}|\widehat{y_t} - y_t|}{n}, \tag{10}$$

em que $\widehat{y_t}$ e y_t são respectivamente o valor estimado por um determinado método e o valor observado correspondente no instante de tempo t e n é o número de valores ausentes (AYDILEK; ARSLAN, 2013; SATTARI *et al*., 2020).

Para cada porcentagem de falhas avaliada (10% e 40%), valores ausentes foram gerados artificialmente em 50 execuções do experimento, segundo o mecanismo de ausência aleatória - MAR. Em cada execução, foi calculada a REQM e o EAM para cada método. Posteriormente, as médias aritméticas dos 50 valores de REQM e EAM obtidos por cada método foram calculadas e comparadas numericamente.

3. RESULTADOS

3.1. Resultados gerais de precipitação

A Tabela 1 mostra medidas descritivas de posição e dispersão das séries utilizadas no presente trabalho. O desvio padrão e a assimetria das distribuições indicam que valores elevados de precipitação ocorreram em determinados períodos do ano. De fato, o máximo mensal precipitado em todas as quatro estações foi igual ou superior a 500 mm, enquanto a maior parte dos valores registrados em cada série foi consideravelmente menor. Medidas como a mediana calculada nas estações evidenciam tais fatos. Esses resultados sugerem que o preenchimento de falhas no conjunto de séries estudadas é uma tarefa complicada.

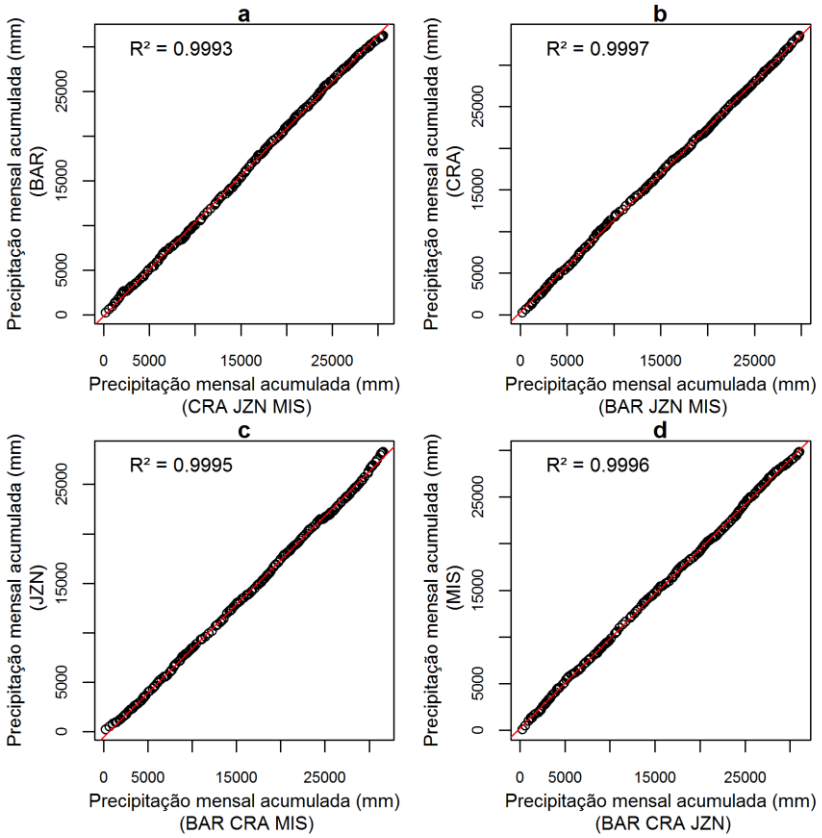
O processo de preenchimento de falhas em dados de precipitação é influenciado por restrições físicas, tais como topografias complexas e a densidade de estações, bem como os padrões de chuva em estações vizinhas (KIM; RYU, 2016). Neste sentido, o método da curva de dupla massa é uma das técnicas mais usadas para verificar a homogeneidade regional e a qualidade dos dados de precipitação (SEARCY; HARDISON, 1960). Essa análise tem sido aplicada por autores como Correia *et al*. (2016). A Figura 2 mostra a curva dupla massa para a série JZN. O ajuste das curvas mostradas nos gráficos e os valores do coeficiente de correlação linear de Pearson próximos de 1 indicam a homogeneidade dos dados de precipitação.

Tabela 1 - Medidas descritivas das séries temporais de precipitação das estações Crato - CRT, Juazeiro do Norte - JZN, Barbalha - BAR e Missão Velha - MIS na área de estudo, Ceará, Brasil (janeiro de 1974 a dezembro de 2004).

Série	CRT	JZN	BAR	MIS
Número de observações	360	360	360	360
Mínimo (mm)	0	0	0	0
1º Quartil (mm)	2,525	0	3,975	0
Mediana (mm)	38,3	25,5	35,7	38,8
Média (mm)	93,158	78,56	86,888	82,85
3º Quartil (mm)	149,6	129,72	144,925	135
Máximo (mm)	500	575,8	533,8	550
Desvio padrão (mm)	114,42	107,45	109,56	107,5
Curtose	1,189	2,809	1,189	1,499
Assimetria	1,362	1,671	1,409	1,435

Fonte: Elaborado pelos autores (2021).

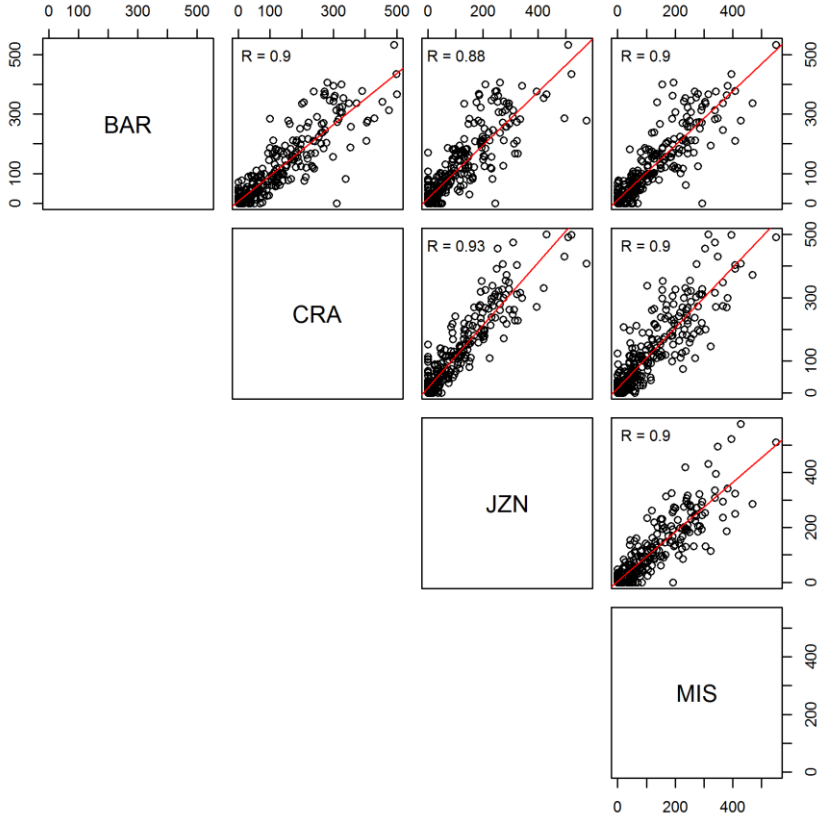
Figura 2 - Gráficos da curva de dupla massa para análise de consistência das séries pluviométricas mensais nas estações da área de estudo: (a) Barbalha - BAR, (b) Crato - CRA, (c) Juazeiro do Norte - JZN e (d) Missão Velha - MIS (janeiro de 1974 a dezembro de 2004).



Fonte: Elaborado pelos autores (2021).

A correlação existente entre os dados de estações vizinhas é outro fator importante a se considerar na análise de métodos de imputação (SATTARI *et al*., 2020). Neste estudo, analisou-se a correlação entre as séries. A Figura 3 mostra graficamente a matriz de correlações das séries temporais nas estações estudadas. Os dados de precipitação nas estações vizinhas foram comparados em pares. As linhas diagonais foram obtidas através do ajuste linear entre cada par de estações. Os gráficos e os valores do coeficiente de correlação linear de Pearson variando entre 0,88 e 0,93 sugerem alta correlação linear entre as séries.

Figura 3 - Gráficos de dispersão e valores do coeficiente de Pearson (R) para as séries mensais de precipitação nas estações Barbalha - BAR, Crato - CRA, Juazeiro do Norte - JZN e Missão Velha - MIS (janeiro de 1974 a dezembro de 2004).



Fonte: Elaborado pelos autores (2021).

Diante dos resultados obtidos, percebe-se que todas as estações analisadas no presente estudo estão localizadas em uma região homogênea em termos de precipitação e sob as mesmas condições climáticas. As precipitações registradas nas estações mostram padrões semelhantes, o que viabiliza o preenchimento de falhas.

3.2. Simulação e preenchimento de falhas

Este trabalho avalia o desempenho de métodos de imputação em séries pluviométricas sob diferentes porcentagens de falhas, a saber, 10% e 40%. As falhas foram simuladas assumindo-se o mecanismo de ausência aleatória - MAR, onde os valores ausentes se distribuem como blocos contínuos de comprimentos variados. Os pontos de dados ausentes da série foram amostrados aleatoriamente 50 vezes para cada porcentagem, cujos tamanhos amostrais foram 36 e 144, para 10% e 40%, respectivamente. Os métodos da média aritmética simples (MA), regressão linear múltipla (RLM), ponderação regional (PR), inverso da potência da distância (IPD) e redes neurais artificiais (RNA) foram aplicados, e a qualidade e precisão das estimativas foram avaliadas segundo a raiz do erro quadrático médio (REQM) e o erro absoluto médio (EAM). Para cada repetição do experimento, foram calculados a REQM e o EAM de cada método de imputação.

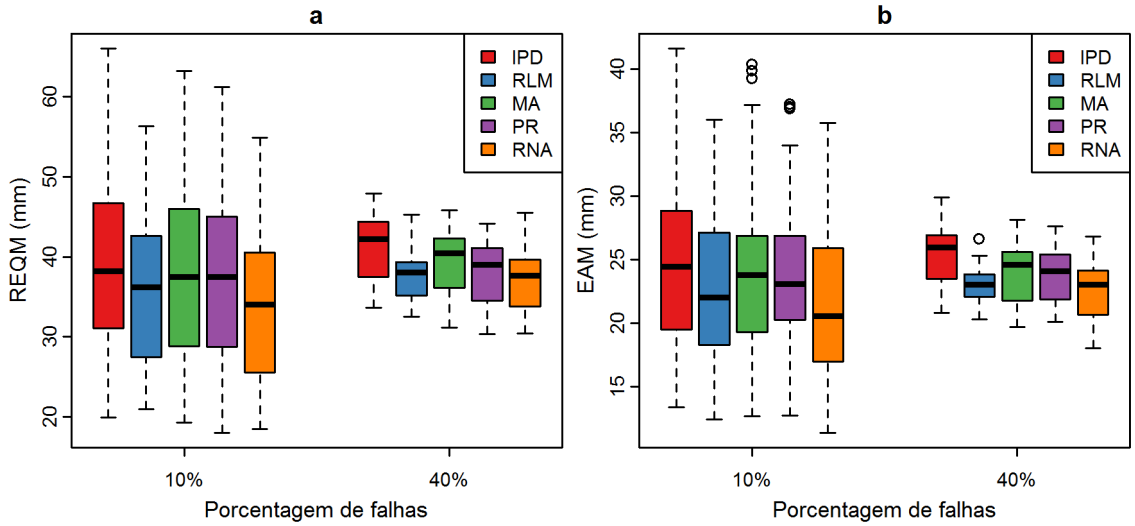
Apesar da correlação linear existente entre as séries analisadas, os modelos de regressão linear múltipla para a série JZN foram construídos tomando-se como variáveis explanatórias apenas as séries CRA e MIS. A série BAR não se mostrou significativa estatisticamente para um nível de significância de 5% e, portanto, foi retirada do modelo.

A Figura 4 mostra os diagramas de caixa com a distribuição dos resultados obtidos pelos métodos para cada métrica avaliada sob as diferentes porcentagens de falhas. Os gráficos permitem comparar a distribuição dos resultados obtidos pelos diferentes métodos.

Todos os métodos apresentaram maior dispersão para a porcentagem de falhas de 10% em comparação aos casos com 40% de falhas. Para 10% de falhas, os valores observados para o REQM possuíram intervalo interquartil (IIQ) variando de 14,73 mm a 16,78 mm. A variabilidade observada diminuiu conforme aumentou-se a porcentagem de falhas, uma vez que o tamanho da amostra aumentou. Com 40% de falhas, foram observados intervalos interquartis de 4,13 mm a 6,5 mm. Fato semelhante foi observado ao se avaliar o EAM. Na experimentação sob uma porcentagem de falhas de 10%, o IIQ variou entre 6,43 mm e 9,01 mm, e em relação à situação com 40% de falhas, foram obtidos valores de EAM com intervalo interquartil variando de 1,78 mm a 3,83 mm. Isso também fica evidente ao analisar o desvio padrão dos dados. A Tabela 2 mostra estatísticas descritivas dos resultados, como a média e desvio-padrão, para todos os métodos, sob diferentes porcentagens de falhas. Foram

calculadas as médias aritméticas dos valores de REQM e EAM obtidos por cada método nas 50 simulações realizadas para cada situação de falhas. Os melhores valores para a média e o desvio-padrão estão destacados em negrito na tabela.

Figura 4 - Diagramas de caixa dos resultados obtidos por cada método de imputação: (a) raiz do erro quadrático médio (REQM) e (b) erro absoluto médio (EAM) sob diferentes porcentagens de falhas (50 repetições).



Fonte: Elaborado pelos autores (2021).

Tabela 2 - Raiz do erro quadrático médio (REQM) e erro absoluto médio (EAM) para cada método de imputação na estação Juazeiro do Norte (janeiro de 1974 a dezembro de 2004) sob 10% e 40% de falhas: média e desvio padrão (mostrado entre parênteses). Os melhores resultados estão destacados em negrito.

Porcentagem de Falhas	Método	REQM (mm)	EAM (mm)
10%	IPD	40,37 (12,09)	25,33 (7,16)
	RLM	36,26 (**9,48**)	22,67 (**5,61**)
	MA	38,55 (11,72)	24,04 (6,91)
	PR	37,34 (11,57)	23,82 (6,25)
	RNA	**34,01** (10,22)	**21,56** (6,03)
40%	IPD	41,26 (4,11)	25,44 (2,36)
	RLM	37,55 (**2,69**)	23,03 (**1,39**)
	MA	39,39 (3,92)	23,91 (2,2)
	PR	37,99 (3,75)	23,7 (1,89)
	RNA	**36,87** (3,41)	**22,41** (2,21)

Fonte: Elaborado pelos autores (2021).

Sob 10%, o método RNA apresentou desempenho superior em comparação aos demais métodos, uma vez que obteve valores médios inferiores para REQM e EAM. O método RLM obteve o segundo melhor desempenho, seguido pelos métodos PR, MA e IPD, respectivamente. Os resultados obtidos pelo método RLM foram menos dispersos em comparação aos demais métodos, para ambos REQM e EAM, segundo os valores de desvio padrão calculados.

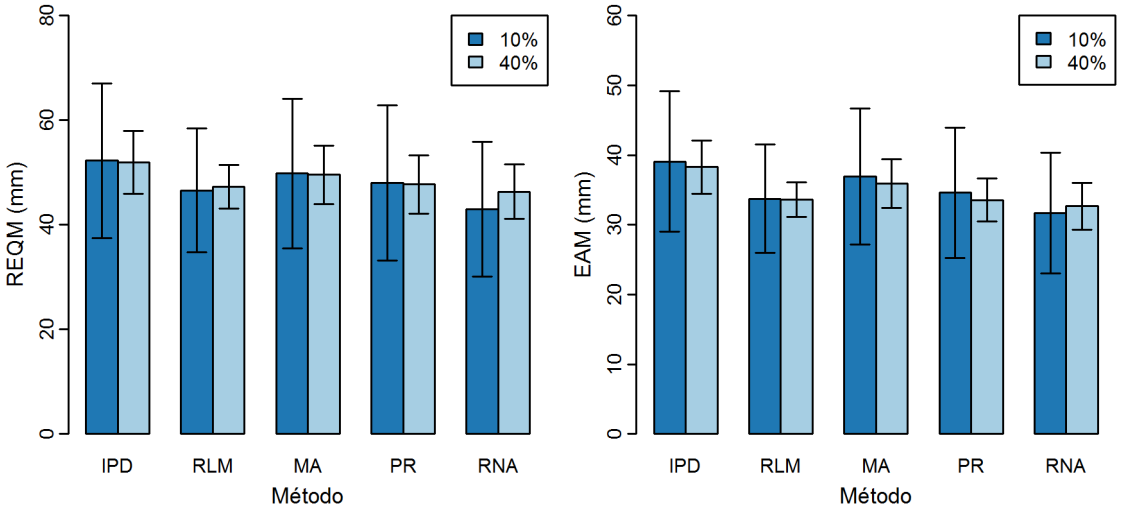
Para os casos em que a porcentagem de falhas foi de 40% de falhas, foram obtidos resultados semelhantes aos da primeira situação. O método RNA foi melhor em comparação aos outros métodos estudados, obtendo os menores valores para REQM e EAM, seguido respectivamente pelos métodos RLM, PR, MA e IPD. O método RLM também apresentou a menor dispersão dos resultados.

Devido às irregularidades interanuais na distribuição pluviométrica na área de estudo, o desempenho dos métodos foi analisado diante dos efeitos sazonais das precipitações. Considerando o regime pluviométrico da região, o presente estudo definiu dois grupos de estações distintos, segundo os quais os métodos foram avaliados: um primeiro grupo, aqui denominado Estação 1, referente aos meses da pré-estação e da estação chuvosa (dezembro a maio), e um segundo grupo, chamado Estação 2, formado pelos meses restantes (junho a novembro). As Figuras 5 e 6 mostram os resultados de cada método considerando as Estações 1 e 2, respectivamente. Os gráficos mostram as médias e os desvios padrão de REQM e EAM calculados a partir das 50 simulações realizadas para cada situação de falhas.

Nos meses das estações chuvosas (Estação 1) da região de estudo, constatou-se que o desempenho dos métodos foi semelhante ao obtido considerando todos os meses (Tabela 3). Para ambas as porcentagens de falhas (10% e 40%), o método RNA foi superior em comparação aos outros métodos, segundo os valores médios de REQM (43,01 mm e 46,36 mm) e EAM (31,69 mm e 32,69 mm). Destaca-se ainda que os resultados observados pelo método RLM também foram menos dispersos em comparação aos demais métodos.

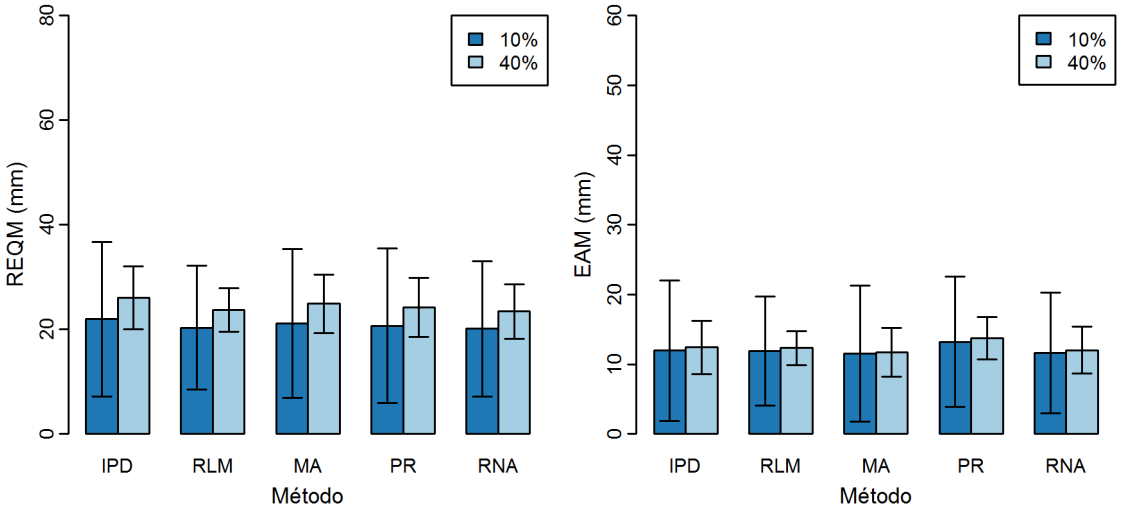
Ao se observar os valores médios calculados para REQM e EAM, percebeu-se que o aumento na porcentagem de falhas, de 10% para 40%, diminuiu a qualidade e precisão das estimativas dos métodos RNA e RLM, enquanto que os desempenhos dos métodos IPD, MA e PR permaneceram iguais ou melhoraram.

Figura 5 - Resultado da imputação (média ± desvio padrão): REQM e EAM sob 10% e 40% de falhas nos meses de dezembro a maio (Estação 1) na série Juazeiro do Norte.



Fonte: Elaborado pelos autores (2021).

Figura 6 - Resultado da imputação (média ± desvio padrão): REQM e EAM sob 10% e 40% de falhas nos meses de junho a novembro (Estação 2) na série Juazeiro do Norte.



Fonte: Elaborado pelos autores (2021).

Em relação aos meses de junho a novembro, correspondentes às estações secas (Estação 2), quando a porcentagem de falhas na série foi de 10%, o método RNA pode ser considerado o melhor em comparação aos demais, segundo a REQM (20,14 mm). O método MA foi superior considerando o EAM (11,52 mm). Para a situação de 40% de falhas, o método RNA obteve melhor desempenho considerando os valores médios de ambos REQM (23,44

mm) e EAM (12,05 mm). Neste caso, o aumento da porcentagem de falhas diminuiu a capacidade de generalização de todos os métodos.

4. DISCUSSÃO

A série pluviométrica na estação Juazeiro do Norte (JZN) apresentou uma média mensal de 78,56 mm, considerando o período de janeiro de 1974 a dezembro de 2004. Neste mesmo período, a média mensal observada considerando apenas os meses da pré-estação e da estação chuvosa foi de 146,09 mm, enquanto que, ao se considerar apenas os meses da estação seca, a média mensal foi de 11,02 mm. Isso evidencia a diferença no regime pluviométrico da região em decorrência da estação do ano (FUNCEME, 2019). Desse modo, espera-se que na Estação 1 se concentrem os valores altos ou extremos de precipitação, e que na Estação 2 se observem os valores mais baixos ou sem precipitação. Portanto, uma vez que as chuvas na região estão concentradas sobretudo nos meses da pré-estação e da estação chuvosa, os valores médios de REQM e EAM obtidos para a Estação 1 são superiores em comparação aos obtidos para a Estação 2.

O desempenho dos métodos é influenciado pelo regime pluviométrico da região, evidenciado pela sazonalidade das séries. Por conta disso, percebem-se diferenças ao se comparar os resultados das três situações consideradas neste trabalho: todos os meses do ano, os meses da estação chuvosa (Estação 1) e os meses da estação seca (Estação 2). Valores baixos, altos ou extremos de precipitação foram incluídos nas amostragens aleatórias realizadas nas simulações. Considerando o caso geral (todos os meses do ano), cujos resultados são mostrados na Figura 4 e na Tabela 2, os valores de REQM e EAM obtidos se apresentaram em faixas de variação semelhantes aos encontrados por Sattari *et al*. (2020). Em estudos conduzidos em uma região de clima mediterrâneo na Turquia, os autores compararam alguns métodos baseados em aprendizado de máquina para o preenchimento de falhas em séries de precipitação sob 10% de falhas em condições de experimentação semelhantes às do presente trabalho.

Os métodos estudados obtiveram resultados semelhantes ao se comparar as porcentagens de 10% e 40% de falhas, tanto de uma maneira geral, como ao se avaliar às variabilidades interanuais na região de estudo. Segundo as métricas avaliadas, o desempenho

dos métodos tende a se estabilizar em torno de um valor central que independe da porcentagem de falhas na série. Isto é, o desempenho dos métodos não depende da porcentagem de falhas. Tal fato também foi observado por autores como Junninen *et al*. (2004) e Aieb *et al*. (2019) ao avaliarem o desempenho de métodos multivariados de preenchimento de falhas. Os autores destacam ainda que técnicas de imputação baseadas em métodos univariados são dependentes da porcentagem de falhas, diferentemente dos métodos multivariados.

A capacidade de generalização de técnicas baseadas em aprendizado de máquina, como os modelos de RNA, é influenciada pelo tamanho do conjunto de treinamento. Assim, conforme aumenta-se a porcentagem de falhas, menor o número de pontos de dados disponível para a construção do modelo. Isso também foi observado na aplicação de redes MLP para preenchimento de falhas por Junninen *et al*. (2004) e Correia *et al*. (2016). Apesar da correlação linear entre as estações, há uma superioridade do método RNA em relação ao método RLM para este conjunto de dados e área de estudo. Resultados semelhantes foram obtidos para regiões de climas semiárido e árido em estudos realizados no Irã por Kashani e Dinpashoh (2012). Os autores apontam que as redes MLP foram superiores na imputação em dados pluviométricos em comparação aos outros métodos avaliados, inclusive o método RLM. De uma maneira geral, os métodos RNA e RLM foram superiores às técnicas tradicionais. Tal fato está em concordância com os achados de Ruezzene *et al*. (2021).

Os resultados do presente trabalho apontam ainda que o método RLM foi superior em comparação aos métodos tradicionais (MA, IPD e PR) para todas as situações de falhas avaliadas. Esses resultados estão de acordo com os obtidos por Fernandez (2007), Kashani e Dinpashoh (2012) e Mello, Kohls e Oliveira (2017). De fato, o método RLM apresenta bom desempenho em situações com homogeneidade nos dados em termos de precipitação, conforme apontam Sattari *et al*. (2020).

Quanto aos métodos tradicionais, o método PR apresentou os melhores resultados, superando os métodos MA e IPD. No trabalho de Bier e Ferraz (2017), foram usados alguns métodos de preenchimento de falhas semelhantes aos adotados neste trabalho. Os autores destacam que o método PR apresentou as melhores estimativas em média em comparação à métodos como o IPD, MA e RLM, apesar de os resultados obtidos terem sido bastante próximos. De fato, apesar do método RNA ter obtido resultados melhores em comparação aos

demais na maioria dos casos analisados, é importante destacar que todos os métodos apresentaram resultados próximos em relação às métricas utilizadas no presente estudo.

5. CONSIDERAÇÕES FINAIS

O presente estudo comparou o desempenho de diversos métodos de imputação em dados de precipitação, sob diferentes porcentagens de falhas. Os valores médios obtidos pelos métodos foram semelhantes. O método RNA obteve as menores médias de REQM e EAM, seguido pelos métodos RLM, PR, MA e IPD.

Os métodos avaliados foram capazes de estimar com uma boa precisão os valores ausentes na série pluviométrica estudada para situações com baixas e altas porcentagens de falhas. O percurso metodológico adotado, baseado nos conceitos de mecanismos de ausência de dados e simulações, pode ser aplicado em estudos semelhantes em diferentes regiões, independentemente do regime pluviométrico e das condições climatológicas. A metodologia utilizada permite a seleção da técnica mais apropriada entre diversos métodos sob diferentes situações de falhas, segundo as métricas de erro adotadas.

Como limitações, destacam-se o uso de apenas uma estação pluviométrica para as análises, e o número relativamente baixo de estações vizinhas utilizadas. Contudo, dada a homogeneidade pluviométrica das estações e a correlação linear existente entre as séries, pode-se esperar resultados semelhantes aos obtidos ao se analisar outras estações localizadas na mesma região. Destaca-se ainda como limitação o número de medidas de desempenho adotado para avaliar os métodos. Apesar de a REQM e o EAM serem duas das métricas de erro mais comumente utilizadas em estudos similares, a aplicação de uma maior diversidade de medidas de desempenho pode contribuir para a escolha do método mais adequado.

Em trabalhos futuros, a aplicação de métodos univariados de imputação de valores ausentes em séries temporais de precipitação deverá ser investigada. Além disso, deverão ser avaliadas abordagens combinadas para o preenchimento de falhas, que façam uso tanto de técnicas multivariadas, tais como os métodos avaliados neste estudo, quanto de métodos univariados do estado da arte das técnicas de imputação de valores ausentes.

REFERÊNCIAS

AIEB, A.; MADANI, K.; SCARPA, M.; BONACCORSO, B.; LEFSIH, K. A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. **Heliyon**, v. 5, n. 2, 2019.

ASGHARINIA, S.; PETROSELLI, A. A comparison of statistical methods for evaluating missing data of monitoring wells in the Kazeroun Plain, Fars Province, Iran. **Groundwater for Sustainable Development**, v. 10, p. 100294, 2020.

AWAD, M.; KHANNA, R. **Efficient learning machines:** theories, concepts, and applications for engineers and system designers. Springer Nature, 2015.

AYDILEK, I. B.; ARSLAN, A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. **Information Sciences**, v. 233, p. 25-35, 2013.

BECK, M. W.; BOKDE, N.; ASENCIO-CORTÉS, G.; KULAT, K. R package imputetestbench to compare imputation methods for univariate time series. **The R journal**, v. 10, n. 1, p. 218, 2018.

BIELENKI JUNIOR, C.; SANTOS, F. M. D.; POVINELLI, S. C. S.; MAUAD, F. F. Alternative methodology to gap filling for generation of monthly rainfall series with GIS approach. **RBRH**, v. 23, 2018.

BIER, A. A.; FERRAZ, S. E. T. Comparação de metodologias de preenchimento de falhas em dados meteorológicos para estações no Sul do Brasil. **Revista Brasileira de Meteorologia**, v. 32, p. 215-226, 2017.

BRUBACHER, J. P.; OLIVEIRA, G. G.; GUASSELLI, L. A. Preenchimento de Falhas e Espacialização de Dados Pluviométricos: Desafios e Perspectivas. **Revista Brasileira de Meteorologia**, v. 35, p. 615-629, 2020.

COGERH. **Plano de Monitoramento e Gestão dos Aquíferos da Bacia do Araripe**: Estado do Ceará. Fortaleza: Companhia de Gestão dos Recursos Hídricos - COGERH, CE, 2009.

CORREIA, T. P.; DOHLER, R. E.; DAMBROZ, C. S.; BINOTI, D. H. B. Aplicação de redes neurais artificiais no preenchimento de falhas de precipitação mensal na região serrana do Espírito Santo. **Geociências (São Paulo)**, v. 35, n. 4, p. 560-567, 2016.

EISCHEID, J. K.; PASTERIS, P. A.; DIAZ, H. F.; PLANTICO, M. S.; LOTT, N. J. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. **Journal of Applied Meteorology**, v. 39, n. 9, p. 1580-1591, 2000.

FERNANDEZ, M. N. **Preenchimento de falhas em séries temporais**. Universidade Federal do Rio Grande – FURG. Curso de Pós-Graduação em Engenharia Oceânica. Dissertação de Mestrado, 2007.

FUNCEME. **Fundação Cearense de Meteorologia - FUNCEME**. 2019. Pré-Estação: entenda o período que antecede a quadra chuvosa do Ceará. Disponível em: http://www.funceme.br/?p=5963. Acesso em: 02 de ago. de 2021.

FUNCEME. **Fundação Cearense de Meteorologia - FUNCEME**. 2021. Página inicial. Disponível em: http://www.funceme.br. Acessado em: 02 de ago. de 2021.

GAO, Y.; MERZ, C.; LISCHEID, G.; SCHNEIDER, M. A review on missing hydrological data processing. **Environmental earth sciences**, v. 77, n. 2, p. 1-12, 2018.

GÓMEZ-CARRACEDO, M. P.; ANDRADE, J. M.; LÓPEZ-MAHÍA, P.; MUNIATEGUI, S.; PRADA, D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. **Chemometrics and Intelligent Laboratory Systems**, v. 134, p. 23-33, 2014.

GÜNTHER, F.; FRITSCH, S. Neuralnet: training of neural networks. **R J.**, v. 2, n. 1, p. 30, 2010.

GUPTA, A.; LAM, M. S. Estimating missing values using neural networks. **Journal of the Operational Research Society**, v. 47, n. 2, p. 229-238, 1996.

HAYKIN, S. **Neural Networks**: A comprehensive foundation. Prentice Hall, 1999.

HONGHAI, F.; GUOSHUN, C.; CHENG, Y.; BINGRU, Y.; YUMEI, C. A SVM regression based approach to filling in missing values. In: **International Conference on Knowledge-Based and Intelligent Information and Engineering Systems**. Springer, Berlin, Heidelberg, 2005. p. 581-587.

HARMAN, B. I.; KOSEOGLU, H.; YIGIT, C. O. Performance evaluation of IDW, Kriging and multiquadric interpolation methods in producing noise mapping: A case study at the city of Isparta, Turkey. **Applied Acoustics**, v.112, p.147-157, 2016.

JUNGER, W. L.; DE LEON, A. P. Imputation of missing data in time series for air pollutants. **Atmospheric Environment**, v. 102, p. 96-104, 2015.

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.

KARAMOUZ, M.; NAZIF, S.; FALAHI, M. **Hydrology and hydroclimatology:** principles and applications. CRC Press, 2012.

KASHANI, M. H.; DINPASHOH, Y. Evaluation of efficiency of different estimation methods for missing climatological data. **Stochastic Environmental Research and Risk Assessment**, v. 26, n. 1, p. 59-71, 2012.

KIM, J.; RYU, J. H. A Heuristic Gap Filling Method for Daily Precipitation Series. **Water Resources Management**, v. 30, n. 7, p. 2275-2294, 2016.

LEE, S.; LEE, K. K.; YOON, H. Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors. **Hydrogeology Journal**, v. 27, n. 2, p. 567-579, 2019.

LIN, W. C.; TSAI, C. F. Missing value imputation: a review and analysis of the literature (2006–2017). **Artificial Intelligence Review**, v. 53, n. 2, p. 1487-1509, 2020.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. John Wiley & Sons, 2019.

MACHIWAL, D.; JHA, M. K. **Hydrologic time series analysis:** theory and practice. Springer Science & Business Media, 2012.

MAITY, R. **Statistical methods in hydrology and hydroclimatology**. Springer, 2018.

MEKANIK, F.; IMTEAZ, M. A.; GATO-TRINIDAD, S.; ELMAHDI, A. Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. **Journal of Hydrology**, v. 503, p. 11-21, 2013.

MEKIS, E.; DONALDSON, N.; REID, J.; ZUCCONI, A; HOOVER, J.; LI, Q.; NITU, R.; MELO, S. An overview of surface-based precipitation observations at environment and climate change Canada. **Atmosphere-Ocean**, v. 56, n. 2, p. 71-95, 2018.

MELLO, Y. R.; KOHLS, W.; OLIVEIRA, T. M. N. Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas. **Boletim de geografia**, v. 35, n. 1, p. 112-121, 2017.

MORITZ, S.; SARDÁ, A.; BARTZ-BEIELSTEIN, T.; ZAEFFERER, M.; STORK, J. Comparison of different methods for univariate time series imputation in R. **arXiv preprint arXiv:1510.03924**, 2015.

NAGHETTINI, M.; PINTO, E. J. A. **Hidrologia estatística**. Belo Horizonte: CPRM, 2017.

OLIVEIRA, G. G.; PEDROLLO, O. C.; CASTRO, N. M. R.; BRAVO, J. M. Simulações hidrológicas com diferentes proporções de área controlada na bacia hidrográfica. **Rev. Bras. Recur. Hídricos**, v. 18, n. 3, p. 193-204, 2013.

PALIT, A. K.; POPOVIC, D. **Computational Intelligence in Time Series Forecasting**: Theory and engineering applications. Springer, 2005.

RADI, N. F. A.; ZAKARIA, R.; AZMAN, M. A. Z. Estimation of missing rainfall data using spatial interpolation and imputation methods. In: **AIP conference proceedings**. American Institute of Physics, 2015. p. 42-48.

RUEZZENE, C. B.; MIRANDA, R. B.; TECH, A. R. B.; MAUAD, F. F. Preenchimento de falhas em dados de precipitação através de métodos tradionais e por inteligência artificial. **Revista Brasileira de Climatologia**. v. 29, p. 177-204, 2021.

SATTARI, M. T.; REZAZADEH-JOUDI, A.; KUSIAK, A. Assessment of different methods for estimation of missing data in precipitation studies. **Hydrology Research**, v. 48, n. 4, p. 1032-1044, 2017.

SATTARI, M. T.; FALSAFIAN, K.; IRVEM, A.; QASEM, S. N. Potential of kernel and tree-based machine-learning models for estimating missing data of rainfall. **Engineering Applications of Computational Fluid Mechanics**, v. 14, n. 1, p. 1078-1094, 2020.

SEARCY, J. K.; HARDISON, C. H. **Double-mass curves**. US Government Printing Office, 1960.

TEAM, R. Core. R: A language and environment for statistical computing. 2021.

TEEGAVARAPU, R. S. V.; CHANDRAMOULI, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. **Journal of hydrology**, v. 312, n. 1-4, p. 191-206, 2005.

TEIXEIRA, F. J. C. **Modelos de gerenciamento de recursos hídricos**: análises e proposta de aperfeiçoamento do sistema do Ceará. Dissertação (Mestrado em Recursos Hídricos) - Universidade Federal do Ceará, Fortaleza, 2003.

TUCCI, C. E. M. **Hidrologia**: ciência e aplicação. Porto Alegre: Ed. UFRGS, 2001.

TWALA, B. An empirical comparison of techniques for handling incomplete data using decision trees. **Applied Artificial Intelligence**, v. 23, n. 5, p. 373-405, 2009.

ZHANG, G. P. An investigation of neural networks for linear time-series forecasting. **Computers Operations Research**, v. 28, n. 12, p. 1183–1202, 2001.