INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO SERTÃO PERNAMBUCANO
PRÓ-REITORIA DE PESQUISA, INOVAÇÃO E PÓS-GRADUAÇÃO (PROPIP)
CAMPUS SALGUEIRO
PÓS-GRADUAÇÃO LATO SENSU EM RECURSOS HIDRICOS PARA O SEMIÁRIDO

RUBENS OLIVEIRA DA CUNHA JÚNIOR

PREVISÃO DA PRECIPITAÇÃO PLUVIOMÉTRICA NA BACIA DO RIO SALGADO, CEARÁ, USANDO MODELOS DE SÉRIES TEMPORAIS

Salgueiro, PE

11 de setembro de 2023

RUBENS OLIVEIRA DA CUNHA JÚNIOR

PREVISÃO DA PRECIPITAÇÃO PLUVIOMÉTRICA NA BACIA DO RIO SALGADO, CEARÁ, USANDO MODELOS DE SÉRIES TEMPORAIS

Monografia apresentada ao curso de Pós-graduação *Lato Sensu* em Recursos Hídricos para o Semiárido, ofertado pelo Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano, como parte dos requisitos para obtenção do título de Especialista em Recursos Hídricos para o Semiárido.

Orientador: Francisco Dirceu Duarte Arraes

Salgueiro, PE

11 de setembro de 2023

Dados Internacionais de Catalogação na Publicação (CIP)

d0 da Cunha Júnior, Rubens Oliveira.

Previsão da precipitação pluviométrica na bacia do Rio Salgado, Ceará, usando modelos de séries temporais / Rubens Oliveira da Cunha Júnior. - Salgueiro, 2023.
16 f. : il.

Trabalho de Conclusão de Curso (Especialização em Recursos Hídricos) -Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano, Campus Salgueiro, 2023.
Orientação: Prof. Dr. Francisco Dirceu Duarte Arraes.

1. Ciências Agrárias. I. Título.

CDD 630

Gerado automaticamente pelo sistema Geficat, mediante dados fornecidos pelo(a) autor(a)

INSTITUTO FEDERAL DO SERTÃO PERNAMBUCANO

Autarquia criada pela Lei n° 11.892 de 29 de Dezembro de 2008

PÓS GRADUAÇÃO LATO SENSU EM RECURSOS HÍDRICOS PARA O SEMIÁRIDO

O artigo "**Previsão da precipitação pluviométrica na bacia do Rio Salgado, Ceará, usando modelos de séries temporais**", autoria de **Rubens Oliveira da Cunha Júnior**, foi submetida à Banca Examinadora, constituída pela ERHS/IFSertãoPE, como requisito parcial necessário à obtenção do título de Especialista em Recursos Hídricos para o Semiárido, outorgado pelo Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano – IFSertãoPE.

Aprovado em 11 de setembro de 2023.

COMISSÃO EXAMINADORA

Prof. Dr. Francisco Dirceu Duarte Arraes – IFSertãoPE
(Presidente)

Prof. Dr. Paulo Renato Alves Firmino – UFCA
(1º Examinador)

Profa. Me. Raquel Costa da Silva – IFSertãoPE
(2ª Examinadora)

Prof. Dr. Juarez Cassiano de Lima Júnior – UFC
(Suplente)

Profa. Dra. Adriana de Carvalho Figueiredo Rodrigues - IFSertãoPE
(Suplente)

Previsão da precipitação pluviométrica na bacia do Rio Salgado, Ceará, usando modelos de séries temporais

Rubens Oliveira da Cunha Júnior, Francisco Dirceu Duarte Arraes

Resumo

A previsão da precipitação pluviométrica é importante para o planejamento e gestão dos recursos hídricos em bacias hidrográficas. Abordagens baseadas em séries temporais são uma alternativa confiável para a obtenção de estimativas dos valores de precipitação. Este artigo tem como objetivo desenvolver modelos de séries temporais para previsão da precipitação pluviométrica. A área de estudo foi a bacia do Rio Salgado (BRS), no estado do Ceará, Brasil. Dados mensais de 19 estações meteorológicas foram considerados. A precipitação média na região foi obtida pelo método dos Polígonos de Thiessen. O estudo compreendeu um período de 48 anos (1974 a 2022), havendo uma partição em conjuntos de treinamento e teste. Modelos autorregressivos integrados de médias móveis (ARIMA), redes neurais perceptron multicamadas (MLP) e máquinas de aprendizado extremo (ELM) foram usados para a modelagem e previsão. O desempenho dos modelos foi avaliado pela raiz do erro quadrático médio (RMSE), erro absoluto médio (MAE), erro percentual absoluto arco tangente médio (MAAPE) e coeficiente de eficiência Nash-Sutcliffe (NSE). O modelo MLP foi superior em relação aos outros modelos, com RMSE e MAE iguais a 42,88 mm e 29,35 mm, respectivamente. Os modelos foram capazes de prever a precipitação média na bacia do Rio Salgado de forma satisfatória. Apesar dos padrões complexos exibidos pela série, os modelos obtiveram estimativas consistentes no conjunto de teste.

Palavras-chave: Séries temporais. Recursos Hídricos. Semiárido. Inteligência artificial. Redes neurais artificiais.

Abstract

Rainfall forecasting is essential for water resource planning and management in river basins. Time series-based approaches are a reliable alternative for obtaining rainfall estimates. This paper aims to develop time series models for rainfall forecasting. The study area was the Salgado River basin (SRB), Ceará State, Brazil. Monthly data from 19 meteorological stations were used. The average rainfall over the region was estimated using the Thiessen polygon method. The study covered 48 years (1974 to 2022), splitting the data into training and test sets. Autoregressive integrated moving average (ARIMA) models, multilayer perceptron neural networks (MLP), and extreme learning machines (ELM) were used for the modeling and forecasting. The metrics root mean squared error (RMSE), mean average error (MAE), mean arctangent absolute percentage error (MAAPE), and Nash-Sutcliffe efficiency (NSE) evaluated the models' performance. MLP model outperformed the other models, with RMSE and MAE values of 42.88 mm and 29.35 mm, respectively. The models were able to forecast the average rainfall in the Salgado River basin in a proper way. Despite the complex patterns in the time series, the models achieved consistent predictions on the test set.

Keywords: Time series. Water resources. Semiarid. Artificial intelligence. Artificial neural networks.

1 Introdução

A chuva é uma importante variável do ciclo hidrológico, que afeta o meio ambiente e diversas atividades humanas. A nível regional, o estudo da precipitação é fundamental para o desenvolvimento sustentável (Li *et al.*, 2021; Ye *et al.*, 2021; Zhang *et al.*, 2022), sobretudo em

regiões semiáridas, que são mais vulneráveis a mudanças climáticas e cujas chuvas são irregularmente distribuídas espacial e temporalmente (Brito *et al.*, 2021). Diante disso, a estimativa acurada da precipitação é essencial em diversas aplicações, tais como agricultura, gestão de recursos hídricos, monitoramento e controle de desastres, previsão climática, entre outros (Esmaeili, Shabanlou e Saadat, 2021; Wang, H. *et al.*, 2021). Contudo, a previsão da precipitação é um desafio, devido à complexidade de tal fenômeno meteorológico. Os modelos de previsão baseados em processos físicos se fundamentam nos mecanismos físicos dos processos hidrológicos e demandam uma grande quantidade de informações e um profundo conhecimento dos sistemas envolvidos. Por outro lado, modelos baseados em dados são modelos empíricos e relativamente mais simples de se utilizar, uma vez que se baseiam em dados históricas e não requerem informações sobre os processos físicos (He, Guan e Qin, 2015; Ni *et al.*, 2020). Neste contexto, a modelagem e previsão de séries temporais constituem uma alternativa confiável para a solução do problema (Kumar *et al.*, 2021). Séries temporais são coleções de observações realizadas sequencialmente no domínio do tempo. A sua análise permite identificar padrões e prever valores futuros a partir de um histórico conhecido (Box *et al.*, 2015)

Entre as técnicas tradicionais de séries temporais mais usadas, está a família de modelos de Box & Jenkins (Box *et al.*, 2015), que inclui os modelos autorregressivos (AR), de médias móveis (MA), autorregressivos de médias móveis (ARMA) e autorregressivos integrados de médias móveis (ARIMA). Este tipo de modelo estatístico linear é amplamente usado para simulação, modelagem e previsão de séries hidrológicas (Lai e Dzombak, 2020). Além disso, os modelos ARIMA têm sido usados para a previsão da precipitação pluviométrica (Al Balasmeh, Babbar e Karmaker, 2019; Dayal *et al.*, 2019). Por outro lado, os recentes avanços tecnológicos impulsionaram o desenvolvimento de métodos de análise com grande aplicabilidade. Destaca-se o Aprendizado de Máquina, um ramo da Inteligência Artificial que dispõe de um conjunto de algoritmos que possuem a capacidade de aprender através de um processo de treinamento e, baseado no conhecimento adquirido, realizar previsões (Awad e Khanna, 2015). As técnicas baseadas em aprendizado de máquina podem ser usadas para a solução de problemas complexos e não lineares de classificação e regressão. Tais algoritmos têm obtido bom desempenho na modelagem e previsão de séries temporais, inclusive em estudos hidrológicos (Pham *et al.*, 2020; Zhang *et al.*, 2022). As redes neurais artificiais do tipo *perceptron* multicamadas (MLP) são o tipo de rede neural mais popularmente adotado. Em especial, existem estudos meteorológicos que aplicaram as redes MLP para a previsão da precipitação pluviométrica (Dash, Mishra e Panigrahi, 2018; Mishra *et al.*, 2018). Uma alternativa que se destaca pela sua eficiência e modelagem simples é a construção de redes neurais usando o algoritmo de treinamento *extreme learning machine* (ELM) (Huang, Zhu e Siew, 2006). Autores como Ali *et al.* (2018), Zeynoddin *et al.* (2018) e Dash, Mishra e Panigrahi (2019) mostraram o bom desempenho de modelos ELM na previsão da chuva.

Diante das considerações realizadas, este trabalho busca desenvolver e avaliar modelos de previsão de séries temporais de precipitação. Será adotada como caso de estudo uma bacia hidrográfica da região do semiárido brasileiro, a bacia do Rio Salgado (BRS), no estado do Ceará, no Nordeste do Brasil.

2 Materiais e Métodos

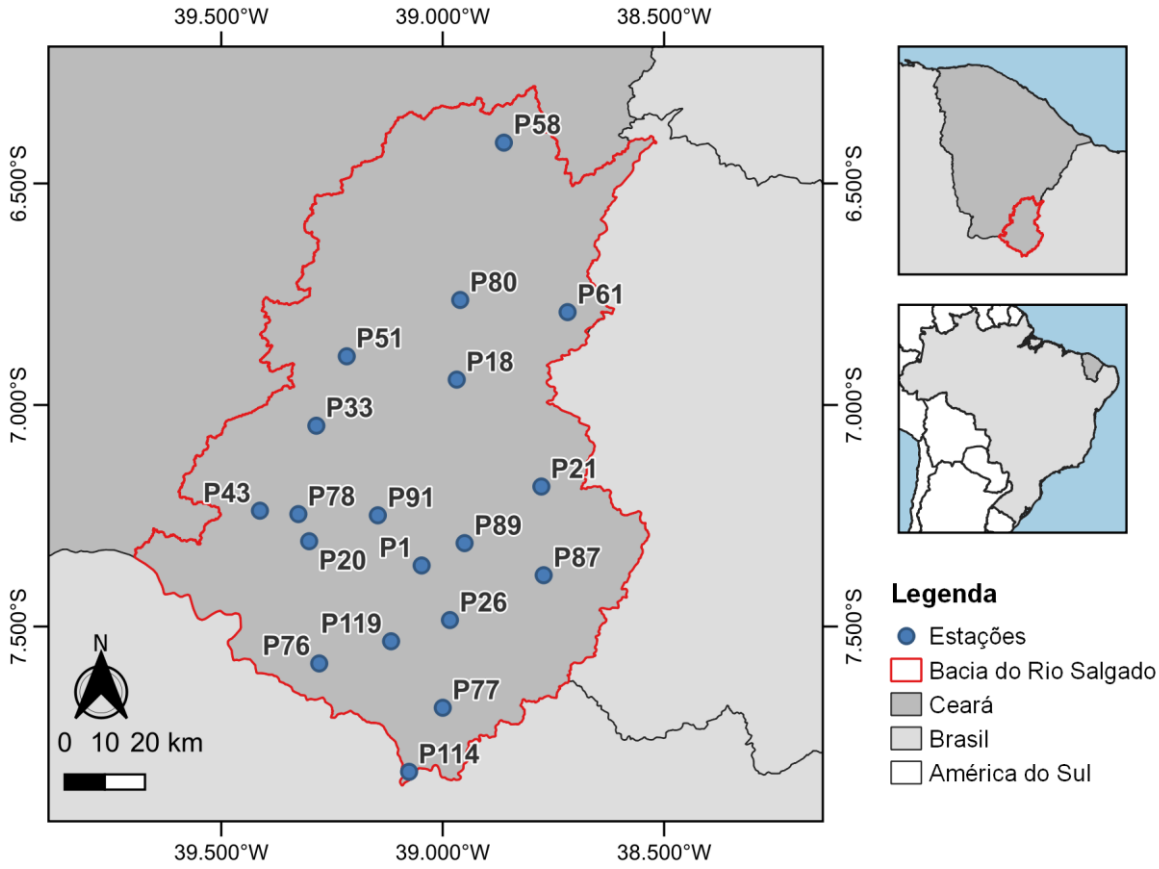
2.1 Área de estudo e conjunto de dados

A área de estudo foi a bacia do Rio Salgado, localizada no estado do Ceará, região Nordeste do Brasil. A bacia do Rio Salgado compreende municípios da Região Metropolitana do Cariri (RMC), sendo uma das mais importantes regiões de planejamento do estado do Ceará,

com uma população estimada superior a 600.000 habitantes (COGERH, 2009). Trata-se de uma área com importância ambiental, social e econômica para a Região Metropolitana do Cariri (RMC) e para o estado do Ceará. O clima é tropical quente semiárido brando, tropical quente e tropical quente sub-úmido. O regime pluviométrico é marcado pela irregularidade interanual e variabilidade temporal e espacial das chuvas. A precipitação média anual na região do Cariri é 700,0 mm. As chuvas se concentram na estação chuvosa, de fevereiro a maio, e nos meses de dezembro e janeiro ocorrem as chuvas de pré-estação (Silva *et al.*, 2021).

Dados de precipitação de 19 estações meteorológicas de monitoramento na bacia do Rio Salgado foram usados. As séries temporais mensais de precipitação medidas em milímetros no período entre 1974 e 2022 foram obtidas do portal na internet da Fundação Cearense de Meteorologia - FUNCEME (http://www.funceme.br). Foram consideradas apenas estações de monitoramento ativas. A Figura 1 mostra a localização da área de estudo, a bacia do Rio Salgado e as 19 estações meteorológicas usadas.

Figura 1 - Mapa de localização da bacia do Rio Salgado e estações meteorológicas selecionadas



Fonte: elaborado pelos autores (2023).

2.1.1 Preparação dos dados

Eventuais valores ausentes nos dados foram preenchidos usando o método univariado da interpolação linear (Moritz *et al.*, 2015). A precipitação média incidente na área da bacia do Rio Salgado foi obtida com o auxílio do método dos Polígonos de Thiessen (Tucci, 2001). A série temporal da precipitação média na bacia do Rio Salgado foi particionada em conjuntos de

treinamento e teste. O conjunto de treinamento foi usado para a construção dos modelos e o conjunto de teste foi usado para verificar a capacidade de generalização dos modelos. O conjunto de teste correspondeu aos 36 meses finais da série de precipitação média e o conjunto de treinamento foi formado pelos demais pontos da série. Autores como Correa e Velho (2020) empregaram estratégias semelhantes para a partição dos dados.

Uma transformação do tipo Mín-Máx para o intervalo $[-0{,}8; 0{,}8]$ foi aplicada ao conjunto de treinamento com o objetivo de diminuir o tempo de treinamento (Aydilek e Arslan, 2013). A Equação (1) mostra a transformação usada.

$$z_t = \frac{y_t - \min(y)}{\max(y) - \min(y)} \times (\text{Max} - \text{Min}) + \text{Min} \quad (1)$$

em que z_t é o valor transformado no instante de tempo t, y_t o correspondente valor original da série, $\min(y)$ e $\max(y)$ são respectivamente os valores mínimo e máximo observados na série y, e Min e Max são os limites mínimo e máximo do novo intervalo, respectivamente (Dash, Mishra e Panigrahi, 2018). No presente estudo, adotou-se Min = $-0{,}8$ e Max = $0{,}8$.

2.2 Técnicas de previsão

2.2.1 Modelos ARIMA

Os modelos autorregressivos (AR) usam os valores passados de uma série temporal e os modelos de médias móveis (MA) usam os erros de previsões passadas para realizar previsões. Pode-se combinar modelos AR e MA, dando origem aos modelos autorregressivos de médias móveis (ARMA). Para a aplicação dos modelos ARMA, a série temporal deve ser estacionária, isto é, as propriedades estatísticas que descrevem seu comportamento devem ser constantes ao longo do tempo. Para atingir a estacionariedade da série, pode-se usar a diferenciação, ou seja, tomar diferenças entre os valores vizinhos da série, por exemplo $y_t - y_{t-1}$. Para uma série não estacionária, pode-se usar os modelos autorregressivos integrados de médias móveis (ARIMA). O modelo ARIMA não sazonal é mostrado na Equação (2).

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

em que y'_t é a série diferenciada, ε_t, ε_{t-1}, …, ε_{t-q} são ruídos brancos, θ_i e ϕ_i são coeficientes e c é a constante da equação de regressão (Chan e Cryer, 2008).

O modelo ARIMA pode ser indicado por $ARIMA(p, d, q)$, em que p é a ordem da parcela autorregressiva, d é o número de diferenciações necessárias à estacionarização, q é a ordem da parcela de médias móveis. Os modelos ARIMA também podem ser usados para séries temporais que possuam sazonalidade, incluindo-se termos sazonais adicionais ao modelo. O modelo ARIMA sazonal pode ser indicado por $ARIMA(p, d, q)(P, D, Q)[m]$, em que P, D e Q são as ordens da parcela sazonal, e m é a sazonalidade da série (Hyndman e Athanasopoulos, 2018).

A seleção da ordem do modelo ARIMA neste estudo recorreu a critérios de informação (Wu, Cheung e Leung, 2020). Diferentes alternativas de modelos foram construídas, e selecionou-se aquele que apresentou o menor valor do Critério de Informação Bayesiano (BIC) (Schwarz, 1978).

2.2.2 Redes neurais artificiais

As redes neurais do tipo *perceptron* multicamadas (MLP) são amplamente utilizadas em problemas de previsão. A sua arquitetura consiste em nós interconectados e dispostos em diversas camadas, de entrada, ocultas e de saída, de modo que cada camada se conecta à camada posterior (Awad e Khanna, 2015; Haykin, 1999). Para a previsão de séries temporais univariadas, costuma-se empregar valores passados da própria série como entradas da rede. Portanto, a MLP pode ser vista como um modelo autorregressivo não linear (Zhang, 2001). O número de nós na camada de entrada é determinado pelo número de valores passados utilizados no modelo autorregressivo. O número de camadas ocultas e de nós nas camadas ocultas pode ser determinado por tentativa e erro ou com o uso de algoritmos de otimização. Para a previsão um passo adiante, é necessário apenas um nó na camada de saída. Uma rede MLP pode ser expressa matematicamente pela Equação (3).

$$y_t = f\left(\sum_{h}^{n_h} \mathrm{w}_{ho} \cdot f\left(\sum_{i}^{ni} \mathrm{w}_{ih} \cdot y_i + v_h\right) + v_o\right) \tag{3}$$

em que os índices i, h e o se referem às camadas de entrada, oculta e de saída, respectivamente, y_i são as entradas, y_t a saída, e n_i e n_h são o número de nós de entrada e na camada oculta, respectivamente. w são os pesos das conexões entre os nós e v são constantes de viés (*bias*). A função f é chamada função de ativação, que permite a aplicação da rede para processos não lineares (Dash, Mishra e Panigrahi, 2019).

Neste estudo, foram usadas redes MLP com uma única camada oculta, função de ativação sigmoide e algoritmo de treinamento *back-propagation* (Riedmiller e Braun, 1993). Quanto à sua arquitetura, a camada de entrada foi formada por valores defasados (*lags*) da série e variáveis binárias auxiliares indicando o mês da observação (variáveis *dummy*), para capturar as variações sazonais da precipitação mensal, como aplicado por Sahoo e Jha (2013). Foram usadas 11 variáveis *dummy* sazonais, pois o 12º mês é capturado quando todas as variáveis *dummy* são iguais a zero. Os coeficientes associados às variáveis *dummy* podem ser interpretados como uma medida do efeito daquela categoria em relação ao mês omitido (Hyndman e Athanasopoulos, 2018).

A seleção dos *lags* se deu com o auxílio da análise da função de autocorrelação (FAC) e função de autocorrelação parcial (FACP) (Kumar *et al.*, 2021). A autocorrelação mede a relação linear entre valores defasados de uma série temporal, y_t e y_{t-k}, em que k é o número da defasagem. Se existe uma correlação entre y_t e y_{t-1}, então também deve haver entre y_{t-1} e y_{t-2}. Contudo, uma possível correlação entre y_t e y_{t-2} deve existir apenas porque ambos têm correlação com y_{t-1}, e não por y_{t-2} ser significante para o modelo. Portanto, para contornar isto, a autocorrelação parcial mede a relação entre y_t e y_{t-k} desconsiderando os efeitos das defasagens $1,2,\ldots,k-1$ (Hyndman e Athanasopoulos, 2018). Para a camada oculta, diferentes valores para o número de nós foram testados.

2.2.3 Extreme Learning Machine

O algoritmo *extreme learning machine* (ELM), foi desenvolvido por Huang, Zhu e Siew (2006) para o treinamento de redes neurais com uma única camada oculta. Diferentemente da rede MLP, na estrutura da rede ELM, os pesos de entradas (conexões entre a camada de entrada e oculta) e os valores das constantes de viés são definidos aleatoriamente. Em seguida, o ELM

determina os pesos da rede por meio de operações matriciais, reduzindo o tempo de treinamento (Xiong, Li e Bao, 2018; Yadav *et al.*, 2017).

Seja um conjunto de treinamento $\{(\mathrm{x}_i, y_i) \mid i = 1,2, \dots, n\}$, em que $\mathrm{x}_i = (y_{i-1}, \dots, y_{i-k})$, uma rede ELM com um número L de nós na camada oculta e função de ativação f pode ser modelada matematicamente pela Equação (4).

$$y_t = \sum_{j=1}^{L} \mathrm{w}_{oj} \cdot f\left(\mathrm{w}_{ij} \cdot \mathrm{x}_t + v_j\right) \tag{4}$$

em que w_{ij} é o vetor de pesos entre a camada de entrada e o nó j da camada oculta, w_{oj} é o vetor de pesos conectando o nó j da camada oculta e o nó de saída, e v_j é a constante de viés do nó oculto j. A Equação (4) pode ser reescrita matricialmente como $\mathrm{H}\eta = \mathrm{Y}$, em que $\eta = [\mathrm{w}_{o1}, \dots, \mathrm{w}_{oL}]^T$, $\mathrm{Y} = [y_1, \dots, y_n]^T$, e H é chamada matriz de saída da camada oculta, mostrada na Equação (5) (Huang, Zhu e Siew, 2006).

$$\mathrm{H} = \begin{bmatrix} f(\mathrm{w}_1 \cdot \mathrm{x}_1 + v_1) & \dots & f(\mathrm{w}_L \cdot \mathrm{x}_1 + v_L) \\ \vdots & \ddots & \vdots \\ f(\mathrm{w}_1 \cdot \mathrm{x}_n + v_1) & \dots & f(\mathrm{w}_L \cdot \mathrm{x}_n + v_L) \end{bmatrix} \tag{5}$$

A determinação dos pesos de saída η é equivalente a solucionar o sistema linear $\mathrm{H}\eta = \mathrm{Y}$. Dessa forma, a solução da Equação (4) é $\hat{\eta} = \mathrm{H}^{+}\mathrm{Y}$, em que H^{+} é a matriz inversa generalizada de Moore-Penrose de H (Huang, Zhu e Siew, 2006). Neste estudo, a seleção da arquitetura da ELM se deu de forma semelhante à usada para a seleção da rede MLP. Foram treinadas 20 redes ELM e o resultado foi obtido pela combinação dos resultados usando a mediana.

2.3 Previsão

Foram realizadas previsões um passo adiante de forma iterativa, em que os valores previstos nos instantes de tempo passados são incorporados às entradas dos modelos, até que todo o horizonte de tempo de previsão seja realizado (Hyndman e Athanasopoulos, 2018). Esta abodagem foi empregada no conjunto de teste, para realizar previsões com um horizonte de 36 meses. Por fim, os valores estimados pelos modelos foram transformados de volta ao intervalo original dos dados usando a operação inversa à Equação (1).

2.4 Avaliação do desempenho

Foram usadas as seguintes métricas para a avaliação do desempenho dos modelos de previsão de séries temporais: Raiz do Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE), Erro Percentual Absoluto Arco tangente Médio (MAAPE) e o Coeficiente de Eficiência de Nash-sutcliffe (NSE). O Critério de Informação Bayesiano (BIC) foi usado para a seleção dos modelos ARIMA. As Equações (6), (7), (8), (9) e (10) mostram o RMSE, MAE, MAAPE, NSE e BIC respectivamente.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{6}$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t| \tag{7}$$

$$\text{MAAPE} = \frac{1}{n}\sum_{t=1}^{n}\arctan\left(\left|\frac{\hat{y}_t - y_t}{y_t}\right|\right) \tag{8}$$

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{9}$$

$$\text{BIC} = n\ln\left[\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}\right] + k\ln(n) \tag{10}$$

em que $\hat{y}_t$ é o valor estimado por um dado modelo e y_t é o valor observado correspondente no instante de tempo t, $\bar{y}$ é a média, n é o número de observações e k é o número de parâmetros estimados pelo modelo.

O RMSE avalia a raiz quadrada da média dos erros quadráticos, enquanto o MAE calcula a média dos erros absolutos. Ambos são dados na mesma unidade de medida dos valores da série temporal em estudo (Dash, Mishra e Panigrahi, 2018). O MAAPE fornece uma medida da precisão do modelo em termos relativos. É uma variação do Erro Percentual Absoluto Médio (MAPE), que supera a limitação do MAPE quando se tem valores iguais a zero (Kim e Kim, 2016). No geral, quanto menores forem os valores de RMSE, MAE e MAAPE, melhores serão as estimativas do modelo. O NSE é usado para avaliar a capacidade de previsão de modelos hidrológicos. Assumindo valores entre $-\infty$ e 1, quanto mais próximo de 1, melhor o ajuste do modelo. Valores de $\text{NSE} = 0$ indicam que as previsões do modelo são tão precisas quanto a média das observações, e caso $\text{NSE} < 0$, a média das observações é um preditor melhor que o modelo (McCuen, Knight e Cutter, 2006). O BIC é um critério usado para a seleção de modelos dentre um conjunto de alternativas de modelos. Quanto menor o valor de BIC, maior a qualidade do modelo (Gheyas e Smith, 2011).

2.5 Implementação computacional

Todas as análises estatísticas foram realizadas em linguagem R (R CORE TEAM, 2023). Para o preenchimento de falhas nas séries temporais, foi usado o pacote *imputeTS* (Moritz e Bartz-Beielstein, 2017). Os modelos ARIMA foram construídos usando o pacote *forecast* (Hyndman e Khandakar, 2008) e as redes neurais MLP e ELM foram construídas com o pacote *nnfor* (Kourentzes, 2019). O cálculo dos polígonos de Thiessen foi realizado através do *software* QGIS (QGIS Development Team, 2022).

3 Resultados e discussão

Este estudo investigou a aplicação de diferentes modelos de séries temporais para a previsão da precipitação na bacia do Rio Salgado, Ceará, Brasil.

3.1 Descrição do conjunto de dados

A Tabela 1 mostra estatísticas descritivas das séries temporais de precipitação usadas. As séries possuem diferentes datas de início e fim, mas considerando todas as séries, o estudo

compreendeu um período de 48 anos, de 1974 a 2022. O número de observações variou entre as séries, por conta dos diferentes períodos de monitoramento. Segundo Phil-Eze (2010), um coeficiente de variação superior a 100% indica uma variabilidade muito alta nos dados. Diante disso, todas as séries apresentaram variabilidade muito alta. A assimetria positiva identificada nas séries sugere que a maior parte dos dados está concentrada abaixo da média. Isto se deve à grande quantidade de valores iguais a zero nas séries, referentes aos meses sem chuva na estação seca do ano em regiões semiáridas.

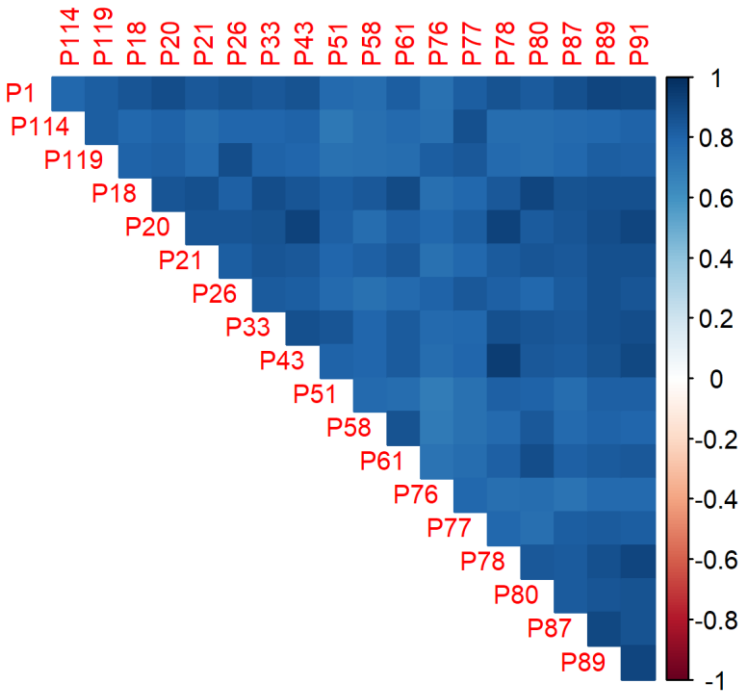
Tabela 1 - Estatísticas descritivas das séries temporais de precipitação mensal na bacia do Rio Salgado. CV (%): coeficiente de variação, n: número de observações, NA (%): valores ausentes.

Série	Mín-Máx	Méd	CV (%)	Assimetria	n	NA (%)	Período
P1	0-540	77.8	138.8	1.6	501	0.00	1981-2022
P114	0-414	48.8	148.3	2.0	501	0.20	1981-2022
P119	0-454.4	67.7	124.0	1.6	525	0.00	1979-2022
P18	0-603.7	78.0	131.7	1.6	585	0.00	1974-2022
P20	0-560.8	88.6	125.4	1.5	585	0.00	1974-2022
P21	0-430.3	68.7	138.5	1.6	585	0.17	1974-2022
P26	0-474.1	76.6	128.5	1.6	585	0.34	1974-2022
P33	0-683	85.6	128.4	1.9	585	0.34	1974-2022
P43	0-531	93.0	123.9	1.4	585	0.00	1974-2022
P51	0-657.4	81.4	140.8	1.9	514	7.59	1979-2021
P58	0-467	64.4	137.5	1.8	585	0.17	1974-2022
P61	0-508.7	77.9	130.8	1.6	537	0.19	1978-2022
P76	0-412.5	57.8	126.7	1.7	525	0.00	1979-2022
P77	0-497	59.1	136.1	1.9	525	0.00	1979-2022
P78	0-575.8	80.2	135.5	1.7	584	0.00	1974-2022
P80	0-570	76.2	132.0	1.7	585	0.00	1974-2022
P87	0-486.5	63.8	133.8	1.7	585	0.00	1974-2022
P89	0-505.3	76.1	127.5	1.5	585	0.00	1974-2022
P91	0-550	83.2	129.6	1.5	585	0.68	1974-2022
PA	0-461.7	72.6	122.6	1.5	480	0.00	1981-2020

Fonte: elaborado pelos autores (2023).

O coeficiente de correlação linear de Pearson mediu a forma como as séries temporais do conjunto de dados estão associadas entre si. A Figura 2 mostra a matriz de correlações na forma de um correlograma. Os valores do coeficiente de Pearson sugerem uma correlação linear e positiva entre as séries. Todas as correlações foram positivas, apresentando valores entre 0.699 e 1. Estes resultados indicam que as estações meteorológicas estão localizadas em uma região homogênea em termos de precipitação e sob as mesmas condições climáticas.

Figura 2 - Correlograma das séries temporais de precipitação mensal medidas nas 19 estações meteorológicas selecionadas na bacia do Rio Salgado

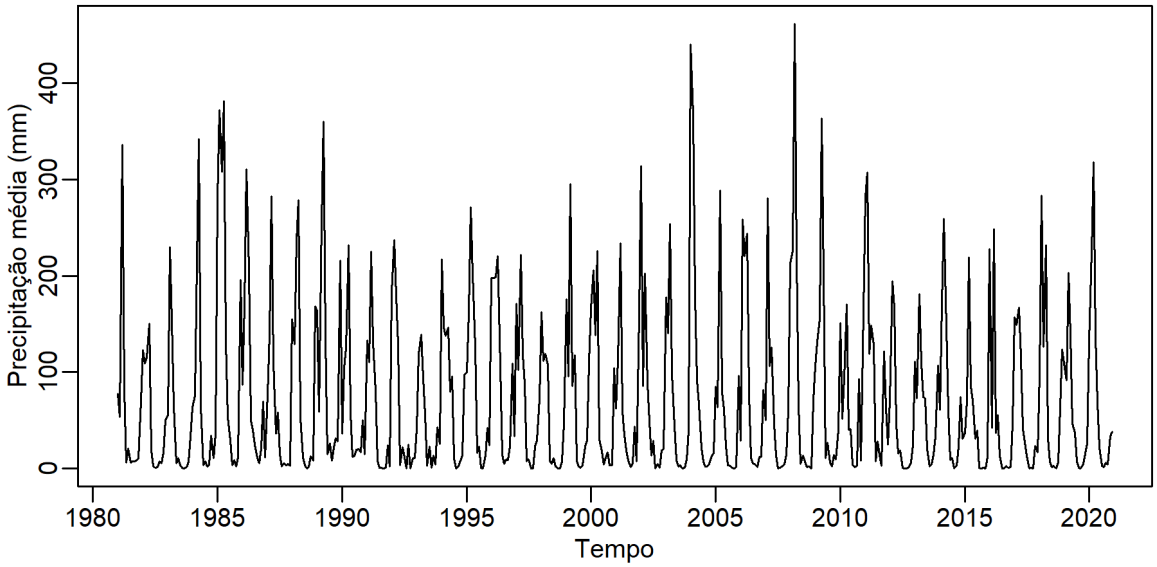


Fonte: elaborado pelos autores (2023).

Calculou-se a precipitação média incidente na área da bacia do Rio Salgado usando a média ponderada das séries temporais medidas as estações de monitoramento consideradas. Para o cálculo, considerou-se o período de 40 anos completos de Janeiro de 1981 a Dezembro de 2020, em que todas as séries possuem observações. Os pesos foram atribuídos segundo o método dos Polígonos de Thiessen. A série temporal da precipitação média foi identificada como PA. As estatísticas descritivas da série PA também estão mostradas na Tabela 1. A Figura 3 mostra o gráfico da série PA.

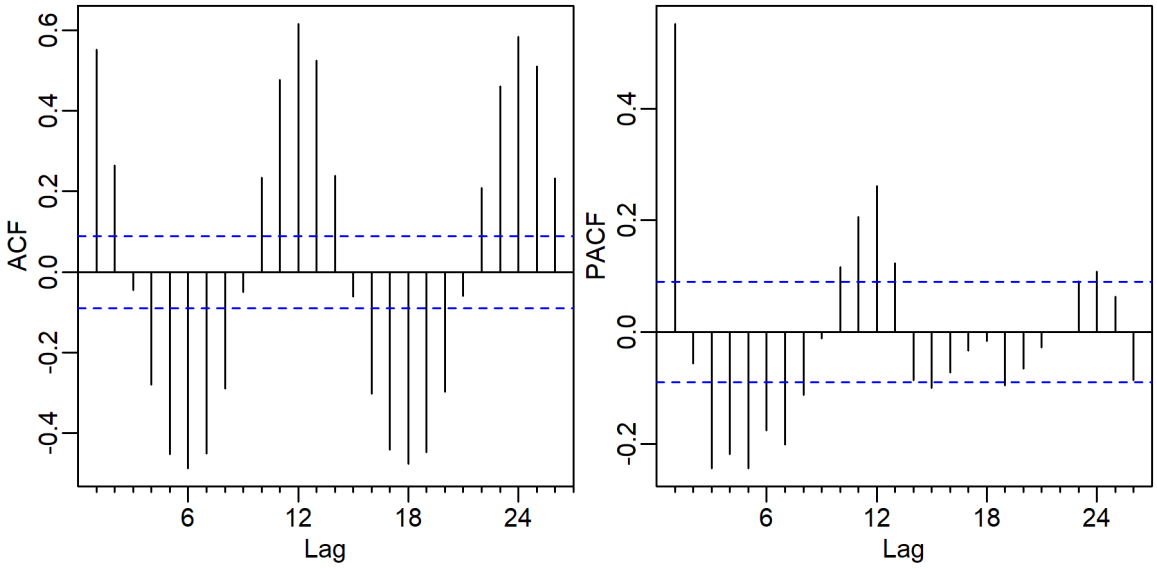
A Figura 4 mostra os gráficos da função de autocorrelação (ACF) e da função de autocorrelação parcial (PACF) da série PA. As linhas horizontais em azul nos gráficos representam os intervalos de 95% de confiança. As oscilações periódicas no valor da autocorrelação mostrada no gráfico da ACF são uma característica de séries com forte padrão sazonal. Este resultado sugere a presença de sazonalidade na série. Quanto ao gráfico da PACF, a maior autocorrelação foi identificada no *lag* 1, havendo ainda significância estatística em outros *lags*. Os *lags* significantes no gráfico PACF são um indicativo de importantes variáveis preditoras para a modelagem da série temporal (Hyndman e Athanasopoulos, 2018).

Figura 3 - Gráfico da série temporal da precipitação média na bacia do Rio Salgado



Fonte: elaborado pelos autores (2023).

Figura 4 - Gráfico da Função de Autocorrelação (ACF) e Função de Autocorrelação Parcial (PACF) da série de precipitação média na bacia do Rio Salgado



Fonte: elaborado pelos autores (2023).

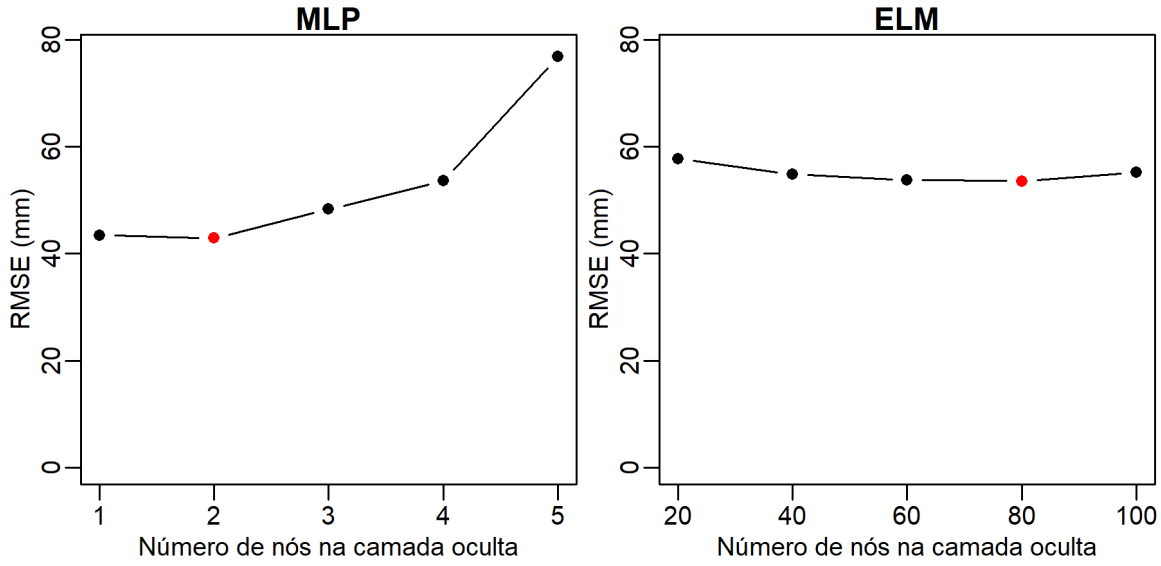
3.2 Construção dos modelos

A série temporal da precipitação média na bacia do Rio Salgado foi usada para a construção dos modelos ARIMA, MLP e ELM. Os parâmetros dos modelos influenciam o seu desempenho. Portanto, neste estudo, diferentes alternativas de modelagem foram investigadas.

O modelo ARIMA selecionado foi ARIMA(2,0,2)(2,1,0)[12], que obteve um valor de BIC igual a -36.3. Quanto às redes MLP e ELM, a camada de entrada foi definida segundo a análise da função de autocorrelação parcial (FACP). A entrada de ambos modelos considerou

9 *lags* com significância estatística, a saber, os *lags* 1, 3, 4, 5, 6, 7, 11, 12, 24. Decidiu-se também por incluir variáveis *dummy* sazonais binárias às entradas das redes, indicando o mês da observação. Para a camada oculta, avaliou-se o emprego de diferentes valores para o número de nós, sendo 1, 2, 3, 4, 5 para a MLP e 20, 40, 60, 80, 100 para a ELM. Todas as alternativas de modelos construídas foram usadas para realizar previsões no conjunto de teste. A rede que obteve o menor valor de RMSE no conjunto de teste foi considerada como o melhor. A Figura 5 mostra a relação entre o número de nós das redes MLP e ELM e o seu desempenho no conjunto de teste segundo o RMSE, com destaque em vermelho para a que atingiu o melhor desempenho.

Figura 5 - Relação entre o número de nós na camada oculta e o desempenho dos modelos ELM e MLP no conjunto de teste segundo o RMSE. Destaque em vermelho para a rede que atingiu o melhor desempenho



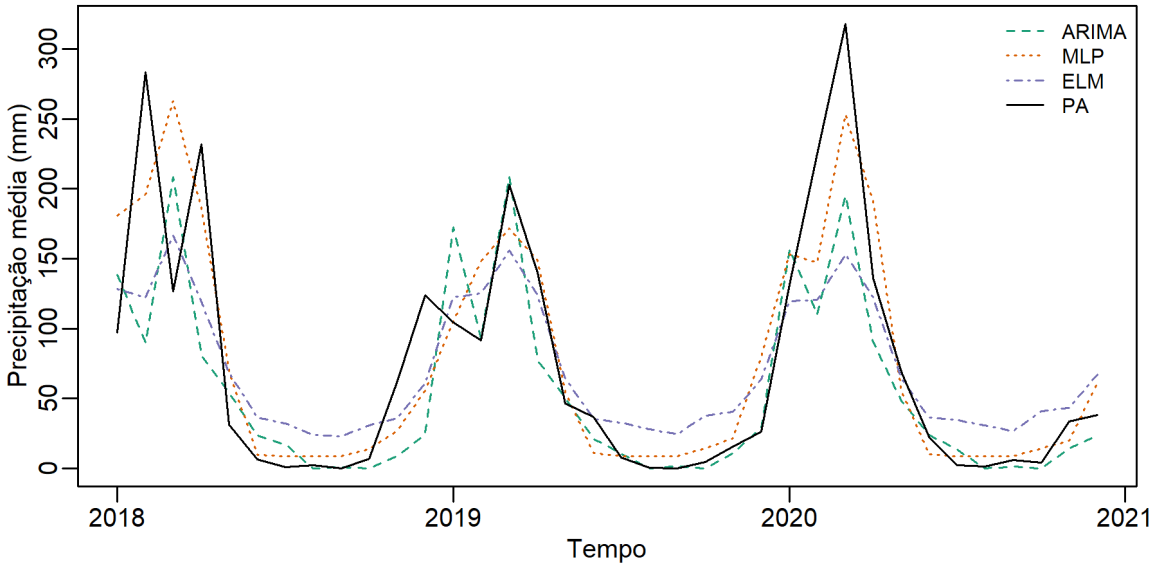
Fonte: elaborado pelos autores (2023).

O desempenho da rede MLP decaiu conforme aumentou-se o número de nós na camada oculta, uma vez que os valores de RMSE aumentaram. O melhor resultado da MLP foi obtido usando 2 nós na camada oculta. Não houve diferenças numéricas no desempenho da rede ELM ao se variar o número de nós na camada oculta. A rede ELM com melhor desempenho foi construída usando 80 nós na camada oculta.

3.3 Previsão da precipitação

Os modelos construídos foram usados para realizar previsões a longo prazo da precipitação média na bacia do Rio Salgado. As previsões foram realizadas considerando um horizonte de tempo que não foi usado na etapa de modelagem, para avaliar a capacidade de generalização dos modelos. O conjunto de testes consistiu nos últimos 3 anos da série temporal PA (36 observações), compreendendo o período de Janeiro de 2018 a Dezembro de 2020. A Figura 6 mostra o gráfico das previsões realizadas pelos modelos ARIMA, MLP e ELM juntamente com os valores originais da série PA no conjunto de teste.

Figura 6 - Gráfico das previsões dos modelos ARIMA, MLP e ELM e dos valores observados da série temporal de precipitação média na bacia do Rio Salgado no conjunto de teste (2018 a 2020)



Fonte: elaborado pelos autores (2023).

Os modelos foram capazes de prever a precipitação média na bacia do Rio Salgado. Apesar dos padrões complexos exibidos pela série, os modelos apresentaram estimativas consistentes no conjunto de teste. Graficamente, nota-se que em alguns anos os modelos não conseguiram capturar os picos nos valores de precipitação. Isto pode estar associado aos valores extremos nos eventos de precipitação na área de estudo. Tal fato também foi observado por autores como Dayal *et al.* (2019). O modelo ELM superestimou os valores de precipitação nas estações secas. No gráfico, as estimativas da rede ELM se distanciaram dos valores originais da série nas estações secas dos anos, diferentemente dos modelos MLP e ARIMA, que obtiveram estimativas mais próximas do esperado.

As medidas RMSE, MAE, MAAPE e NSE foram usadas para mensurar o desempenho dos modelos no conjunto de teste. Os resultados estão mostrados na Tabela 2.

Tabela 2 - Avaliação do desempenho dos modelos no conjunto de teste

Modelo	**RMSE**	**MAE**	**MAAPE**	**NSE**
ARIMA	58,66	34,62	**0,59**	0,54
MLP	**42,88**	**29,35**	0,66	**0,75**
ELM	53,56	38,29	0,78	0,61

Fonte: elaborado pelos autores (2023).

Segundo o RMSE, o melhor modelo foi MLP, seguido por ELM e ARIMA, respectivamente. O melhor resultado segundo o MAE foi obtido pelo modelo MLP, superando nessa ordem ARIMA e ELM. Em relação ao MAAPE, o modelo ARIMA foi superior em relação a MLP e ELM, respectivamente. Segundo o NSE, o modelo que atingiu o melhor

desempenho foi MLP, superando nessa ordem ELM e ARIMA. No geral, este resultado sugere que o modelo MLP foi superior em relação aos outros modelos investigados.

Os resultados obtidos pelos modelos ARIMA e MLP em termos de RMSE e MAE neste estudo foram semelhantes aos achados de Wang, W. *et al.* (2021). Os autores aplicaram, entre outros modelos, ARIMA e MLP para previsão da precipitação mensal na China em um horizonte de 36 meses. Contudo, o modelo ARIMA construído neste trabalho obteve desempenho inferior em termos de NSE em comparação aos resultados de Dayal *et al.* (2019). No referido estudo, a previsão da precipitação mensal usando ARIMA em uma bacia hidrográfica na Índia foi avaliada segundo o NSE com um horizonte de previsão de 12 anos. Por outro lado, os modelos ELM e MLP atingiram resultados relativamente bons em comparação aos achados de Correa e Velho (2020) segundo RMSE e MAE. Usando um horizonte de previsão de 36 meses, os autores destacam também a rede MLP como superior em relação à ELM.

4 Considerações finais

A previsão da precipitação pluviométrica é importante para o planejamento e gestão dos recursos hídricos em bacias hidrográficas. Diferentes técnicas estatísticas e de aprendizado de máquina foram aplicadas para a previsão da precipitação mensal na bacia do Rio Salgado, Brasil. Os modelos investigados foram capazes de realizar previsões dos valores de precipitação na área de estudo de forma satisfatória. A modelagem e previsão das chuvas na área de estudo pode contribuir para o correto entendimento dessa variável e pode fornecer informações para o processo de tomada de decisão na gestão, manejo e conservação de bacias hidrográficas no semiárido. Como limitações deste trabalho, destacam-se o número pequeno de alternativas de modelagem investigadas e a não adoção de técnicas de otimização para a seleção dos parâmetros das redes neurais. Em trabalhos futuros, deve-se investigar a modelagem e previsão de séries temporais diárias de precipitação. Ainda, é interessante avaliar diferentes modelos, além dos aplicados neste estudo.

Referências bibliográficas

AL BALASMEH, O.; BABBAR, R.; KARMAKER, T. Trend analysis and ARIMA modeling for forecasting precipitation pattern in Wadi Shueib catchment area in Jordan. **Arabian Journal of Geosciences**, v. 12, n. 2, p. 27, 7 jan. 2019.

ALI, M.; DEO, R. C.; DOWNS, N. J.; MARASENI, T. Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting. **Atmospheric Research**, v. 213, p. 450–464, 2018.

AWAD, M.; KHANNA, R. **Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers**. [s.l.] Apress open, 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.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time series analysis: forecasting and control**. [s.l.] John Wiley & Sons, 2015.

BRITO, C. S. DE; SILVA, R. M. DA; SANTOS, C. A. G.; BRASIL NETO, R. M.; COELHO, V. H. R. Monitoring meteorological drought in a semiarid region using two long-term satellite-estimated rainfall datasets: A case study of the Piranhas River basin, northeastern Brazil. **Atmospheric Research**, v. 250, p. 105380, 2021.

CHAN, K.; CRYER, J. **Time series analysis with applications in r**. [s.l: s.n.].

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

CORREA, C.; VELHO, H. C. Observing the Existence of Low-Frequency Variability in Monthly Rainfall Data at Southeastern Brazil using R Package Tools – Neural Networks and Wavelet. **Brazilian Journal of Geophysics**, v. 38, n. 2, 2020.

DASH, Y.; MISHRA, S. K.; PANIGRAHI, B. K. Rainfall prediction for the Kerala state of India using artificial intelligence approaches. **Computers & Electrical Engineering**, v. 70, p. 66–73, 2018.

___. Predictability assessment of northeast monsoon rainfall in India using sea surface temperature anomaly through statistical and machine learning techniques. **Environmetrics**, v. 30, n. 4, p. e2533, 2019.

DAYAL, D.; SWAIN, S.; GAUTAM, A. K.; PALMATE, S. S.; PANDEY, A.; MISHRA, S. K. Development of ARIMA Model for Monthly Rainfall Forecasting over an Indian River Basin. *Em*: **World Environmental and Water Resources Congress 2019**. [s.l: s.n.]. p. 264–271.

ESMAEILI, F.; SHABANLOU, S.; SAADAT, M. A wavelet-outlier robust extreme learning machine for rainfall forecasting in Ardabil City, Iran. **Earth Science Informatics**, v. 14, n. 4, p. 2087–2100, 1 dez. 2021.

GHEYAS, I. A.; SMITH, L. S. A novel neural network ensemble architecture for time series forecasting. **Neurocomputing**, v. 74, n. 18, p. 3855–3864, 2011.

HAYKIN, S. **Neural Networks: A comprehensive Foundation**. [s.l.] Prentice Hall, 1999.

HE, X.; GUAN, H.; QIN, J. A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. **Journal of Hydrology**, v. 527, p. 88–100, 2015.

HUANG, G. B.; ZHU, Q. Y.; SIEW, C. K. Extreme learning machine: Theory and applications. **Neurocomputing**, v. 70, n. 1, p. 489–501, 2006.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. Melbourne, Australia: OTexts, 2018.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, v. 26, n. 3, p. 1–22, 2008.

KIM, S.; KIM, H. A new metric of absolute percentage error for intermittent demand forecasts. **International Journal of Forecasting**, v. 32, n. 3, p. 669–679, 2016.

KOURENTZES, N. **nnfor: Time Series Forecasting with Neural Networks**. [s.l: s.n.].

KUMAR, R.; SINGH, M. P.; ROY, B.; SHAHID, A. H. A Comparative Assessment of Metaheuristic Optimized Extreme Learning Machine and Deep Neural Network in Multi-Step-Ahead Long-term Rainfall Prediction for All-Indian Regions. **Water Resources Management**, v. 35, n. 6, p. 1927–1960, 1 abr. 2021.

LAI, Y.; DZOMBAK, D. A. Use of the autoregressive integrated moving average (ARIMA) model to forecast near-term regional temperature and precipitation. **Weather and Forecasting**, v. 35, n. 3, p. 959–976, 2020.

LI, H.; HE, Y.; YANG, H.; WEI, Y.; LI, S.; XU, J. Rainfall prediction using optimally pruned extreme learning machines. **Natural Hazards**, v. 108, n. 1, p. 799–817, 1 ago. 2021.

MCCUEN, R. H.; KNIGHT, Z.; CUTTER, A. G. Evaluation of the Nash–Sutcliffe Efficiency Index. **Journal of Hydrologic Engineering**, v. 11, n. 6, p. 597–602, 2006.

MISHRA, N.; SONI, H. K.; SHARMA, S.; UPADHYAY, A. Development and analysis of artificial neural network models for rainfall prediction by using time-series data. **International Journal of Intelligent Systems and Applications**, v. 12, n. 1, p. 16, 2018.

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

NI, L.; WANG, D.; SINGH, V. P.; WU, J.; WANG, Y.; TAO, Y.; ZHANG, J. Streamflow and rainfall forecasting by two long short-term memory-based models. **Journal of Hydrology**, v. 583, p. 124296, 2020.

PHAM, B. T.; LE, L. M.; LE, T.-T.; BUI, K.-T. T.; LE, V. M.; LY, H.-B.; PRAKASH, I. Development of advanced artificial intelligence models for daily rainfall prediction. **Atmospheric Research**, v. 237, p. 104845, 2020.

PHIL-EZE, P. Variability of soil properties related to vegetation cover in a tropical rainforest landscape. **Journal of Geography and Regional planning**, v. 3, n. 7, p. 177, 2010.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System**. [s.l.] Open Source Geospatial Foundation, 2022.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2023.

RIEDMILLER, M.; BRAUN, H. **A direct adaptive method for faster backpropagation learning: The RPROP algorithm**IEEE international conference on neural networks. **Anais**...IEEE, 1993

SAHOO, S.; JHA, M. K. Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. **Hydrogeology Journal**, v. 21, n. 8, p. 1865, 2013.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, p. 461–464, 1978.

SILVA, M. I.; GONÇALVES, A. M. L.; LOPES, W. A.; LIMA, M. T. V.; COSTA, C. T. F.; PARIS, M.; FIRMINO, P. R. A.; DE PAULA FILHO, F. J. Assessment of groundwater quality in a Brazilian semiarid basin using an integration of GIS, water quality index and multivariate statistical techniques. **Journal of Hydrology**, v. 598, p. 126346, 2021.

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

WANG, H.; WANG, W.; DU, Y.; XU, D. Examining the Applicability of Wavelet Packet Decomposition on Different Forecasting Models in Annual Rainfall Prediction. **Water**, v. 13, n. 15, 2021.

WANG, W.; DU, Y.; CHAU, K.; CHEN, H.; LIU, C.; MA, Q. A Comparison of BPNN, GMDH, and ARIMA for Monthly Rainfall Forecasting Based on Wavelet Packet Decomposition. **Water**, v. 13, n. 20, 2021.

WU, H.; CHEUNG, S. F.; LEUNG, S. O. Simple use of BIC to Assess Model Selection Uncertainty: An Illustration using Mediation and Moderation Models. **Multivariate Behavioral Research**, v. 55, n. 1, p. 1–16, 2020.

XIONG, T.; LI, C.; BAO, Y. Seasonal forecasting of agricultural commodity price using a hybrid STL and ELM method: Evidence from the vegetable market in China. **Neurocomputing**, v. 275, p. 2831–2844, 2018.

YADAV, B.; CH, S.; MATHUR, S.; ADAMOWSKI, J. Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction. **Journal of Water and Land Development**, v. 32, p. 103–112, 2017.

YE, L.; JABBAR, S. F.; ABDUL ZAHRA, M. M.; TAN, M. L. Bayesian Regularized Neural Network Model Development for Predicting Daily Rainfall from Sea Level Pressure Data: Investigation on Solving Complex Hydrology Problem. **Complexity**, v. 2021, p. 6631564, 1 abr. 2021.

ZEYNODDIN, M.; BONAKDARI, H.; AZARI, A.; EBTEHAJ, I.; GHARABAGHI, B.; RIAHI MADAVAR, H. Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate. **Journal of Environmental Management**, v. 222, p. 190–206, 2018.

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

ZHANG, X.; ZHAO, D.; WANG, T.; WU, X.; DUAN, B. A novel rainfall prediction model based on CEEMDAN-PSO-ELM coupled model. **Water Supply**, v. 22, n. 4, p. 4531–4543, fev. 2022.