Aplicação de técnicas de *machine learning* no preenchimento de falhas em séries temporais de precipitação mensal

Machine learning techniques application in filling flaws of monthly precipitation time series

Data de entrada: 29/03/2021
Data de aprovação: 26/10/2022

Guilherme Marques Farias[1*] | Francisco de Assis de Souza Filho[1] | Marco Aurélio Holanda de Castro[1] | David Lopes de Souza[1] | Luis Henrique Magalhães Costa[2]

DOI: https://doi.org/10.36659/dae.2023.058

ORCID ID
Farias GM https://orcid.org/0000-0002-6726-9210
Souza Filho FA https://orcid.org/0000-0001-5989-1731
Castro MAH https://orcid.org/0000-0001-5134-7213
Souza DL https://orcid.org/0000-0001-8041-9511
Costa LHM https://orcid.org/0000-0002-1781-4188

Resumo

O presente trabalho tem como objetivo verificar a eficácia das técnicas de Redes Neurais Artificiais (RNA) e *Random Forest* (RF) no processo de reconstrução de séries temporais de precipitação mensal com falhas. O estudo foi aplicado em séries de estações pluviométricas distribuídas no Estado do Ceará, admitindo que as mesmas apresentam falhas, as quais são corrigidas em função das séries históricas de estações vizinhas. A eficácia das técnicas foi verificada dentro de um processo de validação cruzada. No geral, a *Random Forest* apresentou o melhor desempenho, superando a RNA em número de validações com coeficiente de Nash e Sutcliffe (NSE) superior a 0,75. Nas melhores validações, para ambos os modelos, encontraram-se valores de NSE acima de 0,9 para todas as estações base. O desempenho dos modelos na estação base 2 (EB2), onde obteve-se o melhor desempenho do estudo, apresenta um indicativo de que há uma melhor adaptação dos mesmos a anos com precipitações mais intensas.

Palavras-chave: *Machine Learning*. Séries temporais. Precipitação. Preenchimento de falhas. *Random Forest*.

Abstract

The present work aims to verify the effectiveness of Artificial Neural Networks (ANN) and Random Forest (RF) techniques in the process of reconstructing faulty monthly rainfall time series. The study was applied series of rainfall stations distributed in the State of Ceará, assuming that the same faults presented, as they are corrected in function of the neighboring stations historical series. The techniques effectiveness was verified within a cross-validation process. Overall, Random Forest presented the best performance, surpassing RNA in validations number with Nash and Sutcliffe coefficient (NSE) greater than 0.75. In the best validations, for both models, NSE values above 0.9 were obtained for all base stations. The performance of the models in base station 2 (EB2), where the best performance of the study was obtained, shows that there is a better adaptation of the models to years with more intense rainfall.

Keywords: *Machine Learning. Time series. Precipitation. Fault filling. Random Forest.*

[1] Universidade Federal do Ceará - Fortaleza - Ceará - Brasil.
[2] Universidade Estadual Vale do Acaraú - Sobral - Ceará - Brasil.
* **Autor correspondente:** guilhermemf15@gmail.com.

1 INTRODUÇÃO

Estudos de precipitação são cada vez mais importantes, tendo em vista a necessidade contínua de conhecimento sobre essa variável para o gerenciamento adequado dos recursos hídricos (SANTOS et al., 2009). De acordo com Costa et al. (2013), para estudos que envolvem análises em longo prazo, como previsões climáticas, análises de variações e tendências, faz-se necessário o uso de séries temporais consistentes para que se tenham resultados condizentes com a realidade. Wanderley et al. (2014) destacam que a disponibilidade de uma série de dados contínua e com o mínimo de falhas nos dados observados é fundamental para estabelecer e caracterizar o clima de uma região, proporcionando meios para uma correta avaliação de suas condições.

A lei n° 9.433, que instituiu a Política Nacional de Recursos Hídricos no Brasil, preconiza, entre seus instrumentos, a instituição de um Sistema de Informações sobre Recursos Hídricos; todavia, tal instrumento ainda se mostra muito incipiente (OLIVEIRA et al., 2010). O sistema Hidroweb da Agência Nacional de Águas, apesar de ser uma importante ferramenta para estudos hidrológicos, ainda apresenta muitas falhas diárias, mensais e anuais, o que acaba por inviabilizar a utilização de muitas séries temporais. Depiné et al. (2014) argumentam que as falhas em séries temporais, em previsões ou simulações hidrológicas, dificultam o ajuste de distribuições estatísticas aos dados históricos, reduzindo o desempenho de modelos ou, até mesmo, inviabilizando sua aplicação.

Evidencia-se uma crescente busca na comunidade científica por métodos que corrijam falhas em séries temporais, tornando-as aptas ao uso em diversos estudos. Wanderley et al. (2012) utilizaram *Krigagem* para preencher falhas de dados de precipitação em 63 estações pluviométricas entre os anos de 1965 e 1980, obtendo resultados satisfatórios. Coulibaly e Becker (2009) avaliaram o desempenho de cinco métodos de interpolação espacial (IPD, *Krigagem*, *Krigagem* ordinário, *Krigagem* Universal e Cokrigagem) no preenchimento de falhas de precipitação de 545 estações, tendo a *Krigagem* demonstrado o melhor desempenho entre os demais métodos. Com o intuito de preencher falhas de precipitação mensal, Bier e Ferraz (2017) avaliaram o desempenho de seis métodos, sendo estes: Regressão Linear Múltipla (RLM), Ponderação Regional (PR), Inverso da Potência da Distância (IPD), Método da razão normal (MRN), *Krigagem* Universal (KRG) e Média aritmética simples (MA). Conforme os resultados obtidos, o autor conclui que nenhum método se sobressaiu em relação aos demais.

Nos últimos anos, tem-se presenciado avanços significativos no ramo do reconhecimento de padrões não lineares, onde problemas complexos vêm tendo suas soluções garantidas por meio da aplicação de aprendizado de máquina (*Machine Learning*) (LOURENCETTI, 2011). A ideia básica dessas técnicas é prever o comportamento de um determinado fenômeno a partir do reconhecimento de padrões que são ensinados por meio de um processo de treinamento. Costache (2019) afirma que as principais vantagens do uso de *Machine Learning* (ML) residem no seu alto grau de automação e também na fácil identificação de tendências e padrões em um conjunto de dados, além de possibilidade de trabalhar com vários tipos de dados em escala multidimensional. No campo da hidrologia, o uso de ML tem ganhado espaço devido aos bons resultados presenciados na predição de variáveis hidrológicas (OLIVEIRA et al. 2013, DORNELLES et al. 2013; DEPINÉ et al. 2014).

Neste sentido, o presente estudo tem como objetivo verificar o desempenho de duas técnicas de *Machine Learning* (Redes Neurais Artificiais e *Random Forest*) no processo de reconstrução de séries mensais de precipitação que apresentam falhas. O estudo foi aplicado em séries temporais de 25 anos de dados, compreendidos entre os anos de 1991 e 2015, oriundos de medições em

estações pluviométricas distribuídas nas principais bacias hidrográficas do Estado do Ceará.

2 MATERIAIS E MÉTODOS

2.1 Redes neurais artificiais

Uma rede neural artificial (RNA) pode ser entendida como um modelo matemático inspirado no cérebro de organismos inteligentes, que, após um processo de treinamento, adquire conhecimento sobre um determinado padrão e torna-se capaz de efetuar predições. Assim como a estrutura neural do ser humano possui neurônios biológicos, a RNA (Fig. 1a) é composta por um conjunto de unidades computacionais denominadas neurônios artificiais (Fig. 1b). De acordo com Abraham et al. (2019), cada neurônio artificial possui terminais de entrada similares aos dentritos dos neurônios biológicos, que recebem uma informação, computam esse dado e, posteriormente, fornecem uma saída que será propagada para as demais unidades. Ferneda (2006) ressalta que o comportamento das conexões entre os neurônios é definido por meio de pesos atribuídos a cada uma delas, sendo estes valores positivos ou negativos, a depender a finalidade do problema.

Em termos de estrutura, uma RNA convencional é composta por uma camada de entrada, responsável pela apresentação dos dados de entrada a camadas posteriores, uma camada intermediária ou várias, onde são processados os dados de entrada de modo a facilitar a resolução do problema, e a camada de saída, que gera a saída da rede neural. Cada camada é composta por n neurônios, que definem a arquitetura de uma RNA, sendo a quantidade para as camadas de entrada e saída definida pelo número de variáveis de entrada e saída do problema, respectivamente.

O treinamento de uma RNA é feito por uma rotina denominada algoritmo de aprendizado. O processo consiste em fornecer à RNA um conjunto de padrões de entrada, com suas respectivas saídas. Para cada entrada, o algoritmo de aprendizado indica a qualidade da resposta produzida pela RNA por meio da comparação com o resultado esperado, logo o erro entre os dois valores é informado à rede para que sejam feitos ajustes nos pesos das conexões entre os neurônios, a fim de melhorar suas futuras respostas (FERNEDA, 2006). O processo de treinamento persiste até que o erro entre a saída esperada e a saída dada pela RNA atinja um valor mínimo tolerável. Após o treinamento adequado, dado um conjunto de padrões de entrada diferente do que foi utilizado no treinamento, a RNA torna-se capaz de predizer saídas condizentes com valores esperados para o problema.

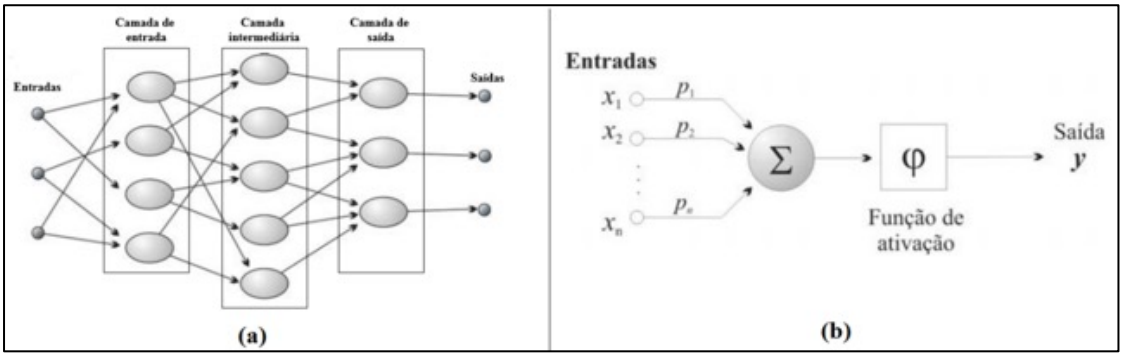

Figura 1 - Representação gráfica de uma rede neural artificial (a) e de um neurônio artificial (b).
Fonte: Ferneda (2006)

2.2 Random Forest

De uma maneira geral, o modelo *Random Forest* (RF) pode ser caracterizado como uma técnica que cria uma série de árvores de decisão de forma aleatória, de modo a formar uma espécie de "floresta" onde cada árvore é utilizada para fornecer um determinado resultado final para o problema.

Em uma RF, as árvores de decisão estabelecem uma estrutura similar a um fluxograma composto por nós (variáveis), onde uma condição é verificada. Caso a condição seja satisfeita, a informação contida no nó segue pelos ramos da árvore até os nós subsequentes, de modo que, ao se alcançar o final da estrutura, obtenha-se um resultado que determina a resposta final da RF. Vale destacar que o resultado final pode ser obtido pela votação do resultado mais relevante dentre os individuais gerados por cada árvore, assim como por meio da média aritmética dos resultados individuais das mesmas. A Fig. 2 apresenta uma ilustração do processo que é feito em uma RF.

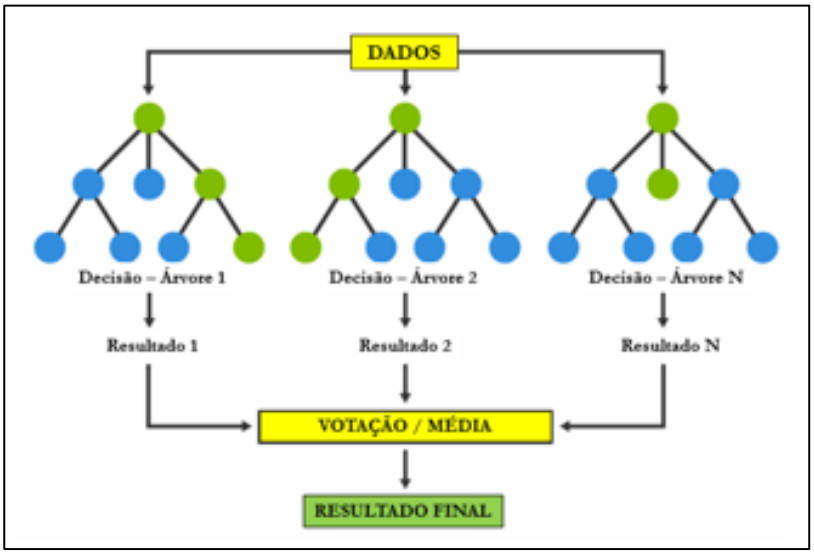

Figura 2 - Representação gráfica de uma Random Forest.

Para a execução de um modelo RF, faz-se necessário, inicialmente, separar o conjunto de dados, sendo uma parte direcionada para efetuar o treinamento do modelo e outra para a validação do mesmo. O processo de criação das árvores de decisão se dá por um mecanismo denominado bagging. Para a criação de cada árvore seleciona-se, de forma aleatória, uma determinada quantidade de amostras de dados do conjunto original de treinamento, sendo estas compostas por valores atribuídos às variáveis utilizadas no problema (XAVIER, 2020). A escolha da variável que será utilizada no primeiro nó da árvore, conforme ilustrado na Fig. 3, dá-se por meio de métodos como entropia ou o Índice de Gini, objetivando captar a variável mais representativa para o conjunto de dados. Nesse caso, são selecionadas, de forma aleatória, duas ou mais variáveis, ficando a critério dos métodos definir qual a mais representativa. Para a escolha da variável utilizada no próximo nó, o processo se repete, excluindo-se as variáveis que já foram selecionadas em nós anteriores, construindo-se a árvore até o último nó. Com a estrutura de árvores criada, é possível apresentar os dados de validação ao modelo RF e efetuar predições, onde cada árvore fornecerá seu resultado individual e este será usado para definir o resultado final (DIDÁTICA TECH, 2019).

Figura 3 - Representação gráfica da escolha de variáveis para a criação das árvores no Random Forest.
Fonte: Didática Tech (2019)

2.3 O processo de preenchimento de falhas

Para o estudo em questão foram utilizadas séries históricas de precipitação média mensal de um banco de dados que contém 1149 estações pluviométricas distribuídas ao longo do território do Estado do Ceará, sendo estas oriundas do site Hidroweb da Agência Nacional de Águas (ANA). As séries históricas apresentam um período de observações de 25 anos, sendo os dados compreendidos entre os anos de 1991 e 2015. Das estações disponíveis, cinco foram consideradas como estações base, nas quais foram admitidas a existência de falhas, que foram preenchidas com o uso das técnicas de Redes Neurais Artificiais e *Random Forest*, tendo como referência as séries históricas de estações vizinhas, denominadas estações secundárias.

Em relação à localização das estações base utilizadas, buscou-se situá-las dentro das principais regiões hidrográficas do Estado do Ceará, de modo a incorporar dentro do processo de previsão por *Machine Learning*, todas as particularidades regionais relacionadas a clima e regime de precipitações, entre outros aspectos que podem influenciar o desempenho dos modelos. Além disso, buscaram-se os locais onde havia uma maior quantidade de séries históricas completas, visto que, após um tratamento preliminar da base de dados, constatou-se que em muitos postos há a detecção de séries com mais de 50% de falhas, o que pode dificultar o processo de previsão por parte dos modelos de ML. Neste sentido, foram atendidas as bacias hidrográficas do Acaraú, da Serra da Ibiapaba, Metropolitana, do Médio Jaguaribe e do Salgado. A classificação das estações base se deu da seguinte forma: EB1 (para a estação localizada no Município de Bela Cruz), EB2 (para a estação localizada no Município de Eusébio), EB3 (para a estação localizada no Reservatório do Castanhão), EB4 (para a estação localizada no Município de Ipu) e EB5 (para a estação localizada no Município de Juazeiro do Norte). A Tabela 1 apresenta as características das estações base e das secundárias utilizadas.

As técnicas de ML foram implementadas na linguagem de programação R, onde há uma infinidade de bibliotecas para desempenhar tal função; no entanto, por questões de eficiência e simplicidade de implementação, optou-se por fazer uso da biblioteca H2O (para Redes Neurais Artificiais) e randomForest (para a técnica de *Random Forest*). Para o preenchimento de falhas foram utilizados dados de séries históricas de postos vizinhos (estações secundárias) à estação base em questão, sendo estes utilizados para treinar os modelos de ML. Com o intuito de obter uma máxima eficiência dentro do processo de predição, foram estabelecidos alguns critérios para o preenchimento de falhas: i) o número mínimo de estações secundárias para o preenchimento é cinco; ii) a distância máxima entre uma estação secundária e a estação base não pode ser superior a 100 km; iii) o conjunto amostral de treinamento e de validação dos modelos de RNA e RF em cada estação é sempre o mesmo, visando a uma comparação idônea entre os métodos utilizados. As localizações das estações base e secundárias ao longo do território cearense podem ser observadas na Fig. 4.

Tabela 1 - Resumo das características das estações base e suas respectivas estações secundárias

Estações base					Estações secundárias			
ID	Código ANA	Nome	Município	Altitude (m)	ID	Município	Distância à estação base (km)	Altitude (m)
EB1	340067	Bela Cruz	Bela Cruz	9,91	ES1	Itarema	32,30	10,92
					ES2	Acaraú	20,26	18,81
					ES3	Aranaú	27,08	0,17
					ES4	Cruz	14,39	26,97
					ES5	Morrinhos	20,18	37,35
EB2	338034	Eusébio	Eusébio	33,26	ES1	Fortaleza	19,03	28,72
					ES2	Maracanaú	14,84	46,59
					ES3	Aquiraz	12,81	23,64
					ES4	Caucaia	26,29	22,66
					ES5	Cascavel	39,26	30,08
EB3	538008	Castanhão	Alto Santo	72,60	ES1	Iracema	41,25	119,57
					ES2	Jaguaribara	6,59	154,71
					ES3	Jaguaretama	44,27	107,85
					ES4	Tabuleiro do Norte	37,94	39,89
					ES5	Potiretama	36,48	170,08
EB4	440078	Ipu	Ipu	245,41	ES1	Croatá	25,75	559,59
					ES2	Ipueiras	25,12	239,20
					ES3	São Benedito	33,14	888,63
					ES4	Hidrolândia	35,77	197,51
					ES5	Pires Ferreira	8,44	197,03
EB5	739065	Juazeiro do Norte	Juazeiro do Norte	397,13	ES1	Santana do Cariri	47,09	535,02
					ES2	Crato	10,50	414,99
					ES3	Porteiras	41,52	505,86
					ES4	Caririaçu	19,95	603,00
					ES5	Barbalha	13,50	442,98

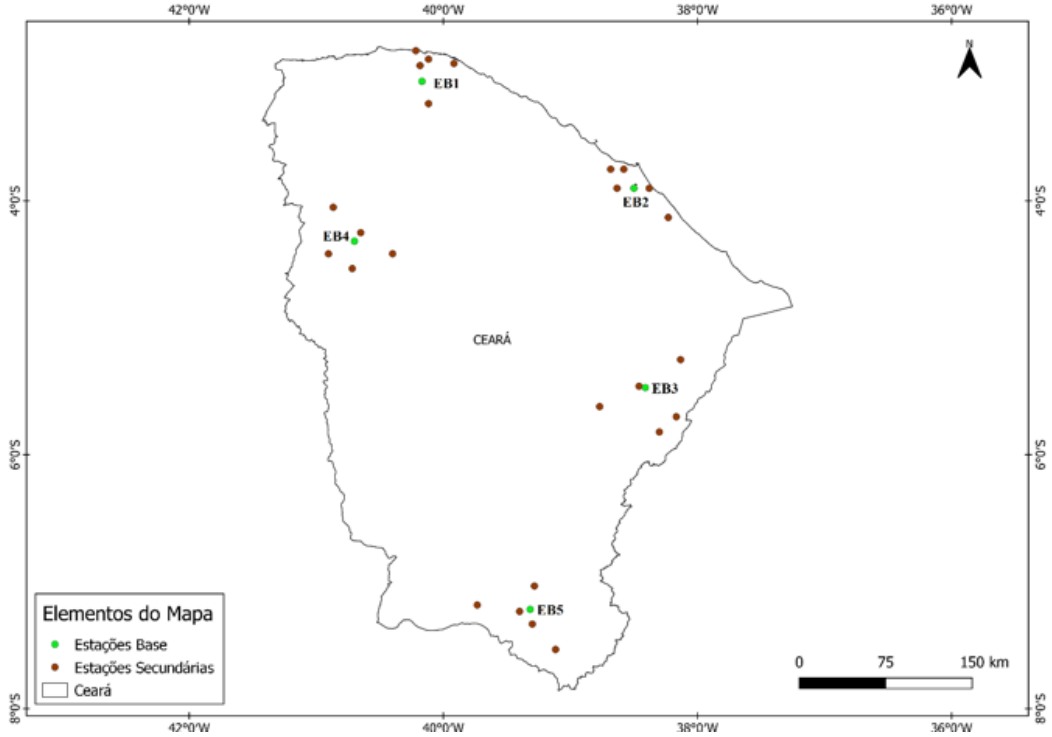

Figura 4 - Localização das estações pluviométricas

O preenchimento de falhas com as técnicas de ML se deu por um processo de validação cruzada, que consiste em separar os dados mensais de um determinado ano, efetuar-se o treinamento com os dados dos demais anos da série histórica e, posteriormente, validar-se o modelo com os dados do ano que foi isolado. Esse procedimento é adotado para todos os anos da série histórica, resultado em um total de 25 validações para cada estação base. Desse modo, é possível compreender a eficácia de predição dos modelos de ML ao longo de toda a série temporal, de modo a identificar possíveis dificuldades de convergência em determinados anos.

Para o treinamento, utilizou-se como dado de entrada uma matriz contendo informações de precipitações médias mensais da estação base e das estações secundárias. O treinamento da RNA foi feito com um algoritmo do tipo *backpropagation*, tendo como função de ativação a *rectifier*. O critério de parada para o treino foi o número de épocas, sendo estas definidas como os ciclos de treinamento, sendo definido o valor de 1000. Para treino da técnica de RF se fez uso de um total de 50 árvores. Os modelos de ML treinados foram utilizados na validação tendo como variáveis de entrada as informações de precipitações médias mensais das estações secundárias, tendo como objetivo predizer as precipitações médias mensais da estação base para o ano avaliado.

No processo de validação, as predições feitas pelas técnicas de ML são comparadas com os dados de precipitação medidos nas estações base em que foram admitidas falhas para o ano a ser validado. Para isso, fez-se uso do coeficiente de eficiência proposto por Nash e Sutcliffe (1970), representado pela Eq. 1. O valor do coeficiente de Nash e Sutcliffe (NSE) varia entre menos infinito e 1, sendo o valor de 1 o indicador de ajuste perfeito entre os dados preditos na modelagem e os valores reais medidos .

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (1)$$

Na referida equação, os valores de O referem-se aos dados de precipitação média mensal medidos nas estações base, S diz respeito aos valores de precipitação média mensal preditos pelos modelos de ML, $\overline{O}$ é a média das precipitações mensais medidas e n é o número de meses do ano avaliado na validação.

3 RESULTADOS E DISCUSSÕES

O processo de treinamento das duas técnicas de ML se mostrou eficiente, sendo constatados, em todos os anos avaliados, um valor de NSE acima de 0,75 para todas as estações base utilizadas neste estudo.

Em uma análise geral, as validações realizadas para a EB2 apresentaram resultados superiores às validações das demais estações base. Das 25 validações feitas com RNA, apenas o ano de 2013 apresentou NSE abaixo de 0,75, obtendo-se o valor de 0,739. Logo, 96% das validações apresentaram resultados satisfatórios, conforme mostra a Tabela 2. Nas validações feitas com RF obteve-se um resultado ainda mais expressivo, uma vez que 100% das validações apresentaram NSE acima de 0,75. Conforme dados apresentados na Tabela 3, a melhor validação para a EB2 obtida com RNA ocorreu para o ano de 2001, resultando em um NSE de 0,979, o que pode ser considerado um bom ajuste de acordo com a interpretação de Pour, Wahab e Shahid (2020). Nas validações feitas com RF, o melhor ajuste ocorreu no ano de 2003, com um NSE igual a 0,982. Os dados presentes na Tabela 2 evidenciam a robustez dos modelos de ML

aplicados ao problema proposto, uma vez que, mesmo no cenário com o menor número de validações com NSE acima de 0,75, que ocorreu para a EB4, mais da metade das validações se enquadraram no referido critério.

De forma gráfica, é possível visualizar na Fig. 5 o bom ajuste entre os dados reais (medidos) e os preditos pelas técnicas de ML para a EB2, sendo a melhor validação da RNA apresentada na Fig. 5A e a melhor validação da RF na Fig. 5B. Em ambos os casos, percebe-se que os dados medidos e preditos estão relativamente próximos à linha identidade, a qual indica o ponto de ajuste perfeito.

Tabela 2 - Percentual de validações com NSE acima de 0,75 para as estações base.

Estações base	NSE acima de 0,75	
	RNA	RF
EB1	76%	84%
EB2	96%	100%
EB3	76%	88%
EB4	72%	84%
EB5	88%	88%

Tabela 3 - Resumo das melhores validações para as estações base

Estações base	Ano de melhor validação		Valores de NSE	
	RNA	RF	RNA	RF
EB1	2008	2015	0,931	0,966
EB2	2001	2003	0,979	0,982
EB3	2014	2012	0,939	0,974
EB4	1991	2004	0,977	0,954
EB5	1995	1992	0,954	0,980

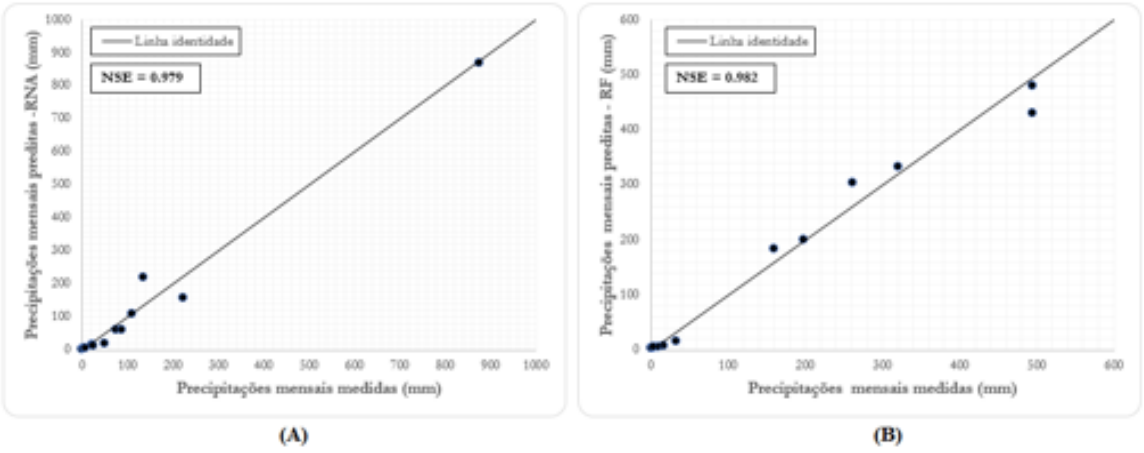

Figura 5 - Resultado da validação para a estação base do Eusébio (EB2): (A) melhor validação com RNA (ano de 2001); (B) melhor validação com RF (ano de 2003).

Em relação às características hidrológicas evidenciadas por meio da série histórica da EB2, apresentada na Fig. 6, destaca-se que as precipitações totais anuais foram, em todos os anos, as maiores encontradas em relação às demais estações base. Tal fato pode ter ligação com o bom ajuste das técnicas de ML, indicando que, para o problema proposto, os modelos usados se adaptam melhor a regiões em que o regime de precipitações é mais intenso. Na Fig. 6 percebe-se que os modelos conseguiram refletir com precisão a maioria dos pontos de pico, assim como os pontos de precipitações mínimas, dando-se destaque ao ajuste para o mês de abril de 2001, onde se estabeleceu o maior valor de precipitação da série histórica, sendo este de 875mm. Mesmo com o valor destoante dos demais, os modelos conseguiram ser bem representativos.

No outro extremo, destaca-se a superioridade da RF na predição para o ano onde contatou-se a menor precipitação total anual (ano de 2013) na EB2, obtendo-se um NSE de 0,846, enquanto a predição feita com RNA retornou um NSE de 0,739, conforme mostra a Tabela 4.

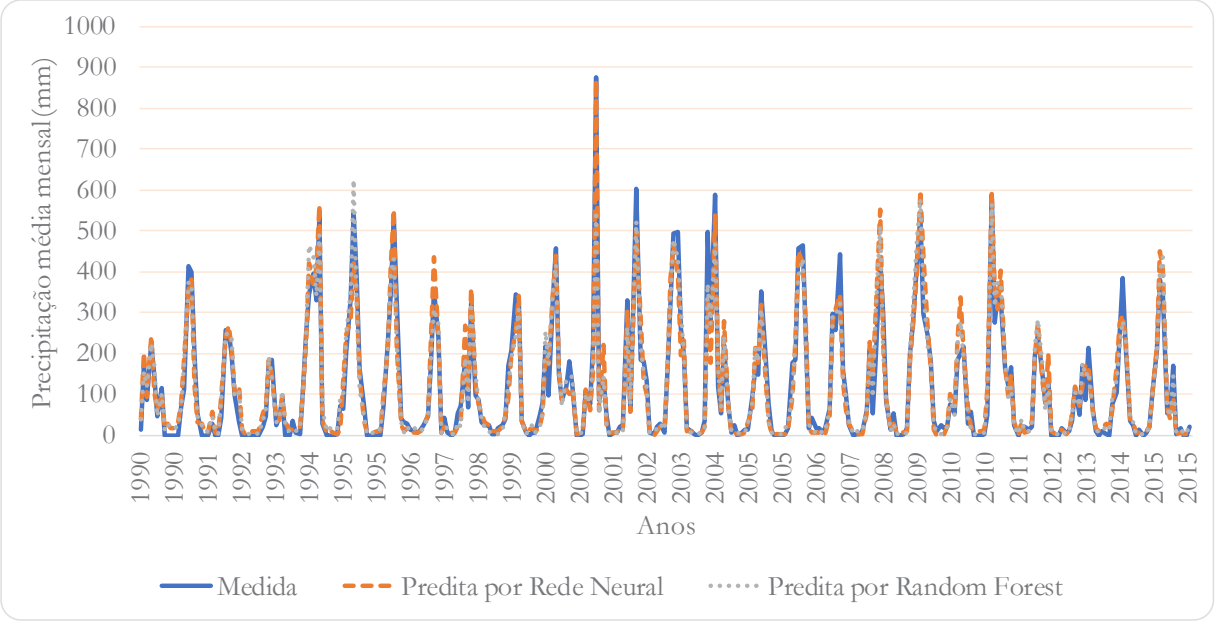

Figura 6 - Série temporal com os valores medidos e preditos pelas técnicas de Machine Learning para a EB2.

Tabela 4 - Resumo das validações feitas nos anos de menor precipitação anual total.

Estações base	Ano de menor precipitação anual	Precipitação total anual (mm)	Valores de NSE	
			RNA	RF
EB1	2012	301	0,655	0,703
EB2	2013	740	0,739	0,846
EB3	1993	192	0,591	0,859
EB4	1998	333	0,665	0,635
EB5	2012	520	0,881	0,900

Ao se analisar os resultados apresentados no processo de validação cruzada para o preenchimento de falhas na EB1, percebe-se, a exemplo do ocorrido para a EB2, que o desempenho das técnicas de ML evidencia uma boa capacidade preditiva para o problema, tendo em vista que no ano que apresentou melhor validação, tanto com RNA como com RF, obteve-se um valor de NSE acima de 0,9. De acordo com a Tabela 3, nas validações feitas com RNA, o processo feito para o ano de 2008 apresentou melhor desempenho, com um NSE de 0,931. Para o mesmo ano, a técnica de RF também apresentou um resultado satisfatório, com um NSE de 0,91. Todavia, o melhor desempenho da RF se deu na validação para o ano de 2015, alcançando-se um NSE igual a 0,966, sendo esta considerada a melhor validação para a EB1. Para o ano em questão, a RNA também apresentou um NSE elevado, com o valor de 0,912.

Em termos de validações satisfatórias (com NSE acima de 0,75), destaca-se que o procedimento com RNA alcançou um total de 76% na EB1, enquanto o modelo RF mostrou-se superior, com 84% das validações alcançando NSE superior a 0,75. Avaliando a Tabela 4, percebe-se que o ano de menor precipitação total anual na EB1 foi 2012, observando-se o valor de 301mm. Nesse caso, o desempenho da RF mostrou-se superior ao da RNA, fato constatado também para os anos de menor precipitação total anual das demais estações base, à exceção do ano de 1998 da EB4, onde os dois modelos mostraram desempenho relativamente próximo.

Nas validações da EB3, identificou-se um desempenho semelhante à EB1, com o modelo de RF mostrando-se superior à RNA quando se observa o total de validações com NSE acima de 0,75. Nesse caso, do total de 25 validações, a RF apresentou apenas três abaixo de tal critério, enquanto a RNA apresentou seis validações em tal condição. O melhor desempenho da RNA se deu para a validação do ano de 2014, apresentando um NSE de 0,939. O ano de 2012 apresentou a melhor validação com a RF, obtendo-se o valor de 0,974 para o NSE.

Apesar do bom desempenho no preenchimento de falhas da EB3, houve dificuldade das duas técnicas empregadas na validação do ano de 2010, onde se verificou uma das menores precipitações anuais totais, sendo esta da ordem de 366mm. Nesse caso, percebeu-se uma tendência de superestimativa dos modelos de ML para o ano em questão. Tal constatação sugere um reforço da tese de que as técnicas de ML possuem uma melhor adaptação em anos chuvosos, conforme evidenciado na EB2. Todavia, na maioria dos casos, a RF mostrou-se mais robusta que a RNA diante de validações em anos com baixo índice de precipitações.

Ao se avaliar os dados apresentados na Tabela 4, percebe-se que no ano de 1993 da EB3 houve a menor precipitação total anual dentre as demais estações base, com o valor de 192mm. Na validação do referido ano, a RNA apresentou certa dificuldade em predizer os dados, obtendo-se um NSE de 0,591. Por outro lado, a RF apresentou uma diferença expressiva de desempenho, com NSE de 0,859.

Para as EB4 e EB5, as melhores validações apresentaram resultados semelhantes às estações base discutidas anteriormente, ou seja, valores de NSE acima de 0,9. No caso, a melhor validação encontrada para a EB4 com RNA ocorreu no ano de 1991 com NSE igual a 0,977. Com a RF, obteve-se a melhor validação no ano 2004, com NSE de 0,954. Em relação à EB5, constatou-se o melhor desempenho com RNA no ano de 1995, com NSE igual a 0,954. A melhor validação com RF ocorreu no ano de 1992, apresentando-se um NSE de 0,98.

Na avaliação geral das validações, a RF mostrou um desempenho melhor do que o modelo de RNA na EB5, enquanto na EB4 os resultados se mostraram relativamente próximos. Verificou-se, para a EB4, que 72% das validações feitas com RNA apresentaram NSE acima de 0,75, enquanto 84% das validações feitas com RF se enquadraram no referido critério. Na EB5, ambas as técnicas de ML conseguiram alcançar o percentual de 88% das validações com NSE acima de 0,75.

Apesar do bom desempenho das técnicas no geral, destaca-se que dentre as validações feitas, o ano de 2005 da série histórica da EB4 apresentou o menor valor de NSE constatado no estudo, sendo este de 0,258 utilizando RNA. Na EB5, também se verificou um baixo desempenho da RF, ocorrido no ano de 1997, com valor de NSE igual a 0,304. Os dados em questão estão listados na Tabela 5, onde se evidenciam os anos onde se obtiveram os piores desempenhos dos modelos para cada estação base. De acordo com Pour, Wahab e Shahid (2020), valores de NSE abaixo de 0,5 indicam que o desempenho do modelo usado não foi adequado, sendo a média dos dados medidos um melhor parâmetro para efeito comparativo.

Tabela 5 - Resumo das piores validações para as estações base

Estações base	Ano de melhor validação		Valores de NSE	
	RNA	RF	RNA	RF
EB1	1994	2002	0,529	0,554
EB2	2013	2001	0,739	0,820
EB3	2010	2010	0,513	0,435
EB4	2005	1998	0,258	0,635
EB5	1994	1997	0,603	0,304

Ao compararem-se os resultados obtidos no estudo em questão com valores apresentados na literatura, percebe-se que as técnicas de *Machine Learning* vêm apresentando robustez na predição de variáveis hidrológicas. Ao utilizar-se de RNA para o preenchimento de falhas em séries temporais de precipitação mensal para quatro estações pluviométricas, Correia et al. (2016) obtiveram valores de NSE acima de 0,8. Depiné et al. (2014) preencheram falhas em séries de dados pluviométricos de nove estações por meio de RNA, obtendo valores superiores a 0,9 para o NSE. Apesar de importantes, estudos apresentados na literatura muitas vezes se limitam a verificar a eficiência de apenas um modelo de ML; todavia, conforme os resultados apresentados neste estudo, há diferenças significativas entre a capacidade de predição dos mesmos, em função de características hidrológicas da área estudada, fato que pode ser crucial na escolha de um método quando se demanda um grau de precisão mais apurado.

4 CONCLUSÕES

O processo de correção de dados de precipitação em séries históricas para o Estado do Ceará, por meio de técnicas de *Machine Learning* possibilitou estimar, de forma satisfatória, tal variável, em função de precipitações observadas em séries temporais oriundas de estações pluviométricas vizinhas.

As técnicas de *Machine Learning* avaliadas neste estudo apresentaram um bom desempenho no processo de preenchimento de falhas em séries temporais de precipitação, tendo em vista que nas melhores validações constatadas para cada estação base, sempre se obtiveram valores de NSE superiores a 0,9. Na avaliação geral, constatou-se que a técnica que apresentou o melhor desempenho foi a RF, visto que, no preenchimento de todas as estações base, a mesma igualou ou superou a RNA em número de validações com NSE superior a 0,75. Em relação ao desempenho nos pontos extremos da série temporal, ao se fazer uma análise sobre a precipitação total anual, a RF apresentou um desempenho superior à RNA tanto para os anos em que o total precipitado anual foi máximo como para anos onde o mesmo foi mínimo, o que sugere uma melhor capacidade de generalização do modelo RF.

As melhores validações apresentadas neste estudo foram para a EB2. Para a estação em questão, das 25 validações com RNA, apenas 1 apresentou NSE abaixo de 0,75. Em relação à RF, não se verificou nenhum valor de NSE abaixo de 0,75. Tal constatação pode ter ligação com o fato de que as precipitações totais anuais da série histórica da estação em questão foram superiores aos totais anuais precipitados das demais estações, sugerindo que as técnicas de ML apresentam melhor desempenho para anos chuvosos. O fato em questão é reforçado em função de as duas técnicas de ML não terem apresentado um bom resultado na validação do ano de 2010, para a EB3, onde constatou-se um dos menores valores de precipitação total anual dentre todas as estações base.

Os menores valores de NSE apresentados nas validações foram para as EB4 e EB5. Para o ano de 2005 da EB4, a RNA obteve-se um NSE de 0,258, enquanto no ano de 1997 da EB5 constatou-se um NSE igual a 0,304 com o modelo de RF.

5 AGRADECIMENTOS

Agradeço à CAPES pela concessão de bolsa de estudos, a qual possibilitou o desenvolvimento desta pesquisa científica.

6 CONTRIBUIÇÃO DOS AUTORES

Conceitualização: Farias, GM e Souza Filho, FA; **Metodologia:** Farias, GM, Souza Filho, FA, De Castro, MA e Costa, LHM; **Investigação:** Farias, GM,

Souza Filho, FA, De Castro, MA e Souza, DL; **Interpretação dos resultados:** Farias, GM, Souza Filho, FA e De Castro, MA; **Redação da primeira versão:** Farias, GM e Souza, DL; **Revisões:** Farias, GM e Costa, LHM.

7 REFERÊNCIAS

ABRAHAM, E. R.; REIS, J. G. M.; TOLOI, R. C.; SOUZA, A. E.; COLOSSETTI, A. P. Estimativa da produção da soja brasileira utilizando redes neurais artificiais. **Agrarian**, v. 12, n. 44, p. 261–271, 2019. https://doi.org/10.30612/agrarian.v12i44.9209

BIER, A.A; FERRAZ, S.E.T. Comparação de metodologias de preenchimento de falhas em dados meteorológicos para estações no Sul do Brasil. **Rev. Bras. Met**, v. 32, n. 2, p. 215-226, 2017. https://doi.org/10.1590/0102-77863220008

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

COSTA, M.N. M.; BECKER, C.T.; BRITO, J.I.B. Análise das séries temporais de precipitação do semiárido paraibano em um período de 100 anos - 1911 a 2010. **Rev. Bras. Geogr. Física**, v. 6, n. 4, p. 680-696, 2013. https://doi.org/10.26848/rbgf.v6i4.233058

COSTACHE, R. Flood Susceptibility Assessment by Using Bivariate Statistics and Machine Learning Models - A Useful Tool for Flood Risk Management. **Water Resource Management.** Vol. 33, p. 3239–3256, 2019. https://doi.org/10.1007/s11269-019-02301-z

COULIBALY, M.; BECKER, S. Spatial interpolation of annual precipitation in South Africa - Comparison and evaluation of methods. **J. Wat. Inter.**, v. 32, n. 3, p. 494-502, 2009. https://doi.org/10.1080/02508060708692227

DEPINÉ. H.; CASTRO, N.M.R.; PINHEIRO, A.; PEDROLLO, O. O. Preenchimento de falhas de dados horários de precipitação utilizando redes neurais artificiais. **Rev. Bras. Rec. Hídr.**, v. 19, n. 1, p. 51-63, 2014. https://doi.org/10.21168/rbrh.v19n1.p51-63

DIDÁTICA TECH. **O que é e como funciona o algoritmo RandomForest**. São Paulo, 2019. Disponível em: < https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/> Acesso em: 07 set. 2022.

DORNELLES, F.; GOLDENFUM, J.A.; PEDROLLO, O.C. Artificial neural network methods applied to forecasting river levels. **Rev. Bras. Rec. Hídr.**, v. 18, n. 4, p. 45-54, 2013. https://doi.org/10.21168/rbrh.v18n4.p45-54

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**. Vol. 35, n.1, p. 25 – 30, 2006. https://doi.org/10.1590/S0100-19652006000100003

LOURENCETTI, F. H. **Estudo da reprodução do comportamento hidráulico de sistemas de abastecimento de água via redes neurais artificiais (RNAs)**. Dissertação (Mestrado em Hidráulica e Saneamento) – Escola de Engenharia de São Carlos, Universidade de São Paulo. São Carlos, p. 177, 2011.

NASH, J.E.; SUTCLIFFE, J.V. River flow forecasting through conceptual models, Part I - A discussion of principles. **Journal of Hydrology**. v. 10, 1970, p. 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

OLIVEIRA, L.F.C.; FIORENZE, A.P. ; MEDEIROS, A.M.M.; SILVA, M.A.S. Comparação de Metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. **Rev. Bras. Eng. Agríc. Ambient.**, v. 14, n. 11, p. 1186-1192, 2010. https://doi.org/10.21168/rbrh.v18n3.p193-204

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.

POUR, S.H; WAHAB, A.K.A.; SHAHID, S. Physical-empirical models for prediction of seasonal rainfall extremes of Peninsular Malaysia. **Atmospheric Research**, v.233, 2020. https://doi.org/10.1016/j.atmosres.2019.104720

SANTOS, G. G.; FIGUEIREDO, C. C.; OLIVEIRA, L. F. C.; GRIEBELER, N. P. Intensidade-duração frequência de chuvas para o Estado de Mato Grosso do Sul. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.13, p.899-905, 2009. https://doi.org/10.1590/S1415-43662009000700012

WANDERLEY, H.S.; AMORIM, R.F.C.; CARVALHO, F.O. Variabilidade espacial e preenchimento de falhas de dados pluviométricos para o estado de Alagoas. **Rev. Bras. Met.**, v. 27, n. 3, p. 347-354, 2012. https://doi.org/10.1590/S0102-77862012000300009

WANDERLEY, H. S.; AMORIM, R. F. C.; CARVALHO, F. O. Interpolação espacial de dados médios mensais pluviométricos com redes neurais artificiais. **Revista Brasileira de Meteorologia**, v.29, p. 389-396, 2014. https://doi.org/10.1590/0102-77862013063

XAVIER, L.C.P.; DA SILVA, S.M.O.; CARVALHO, T.M.N.; FILHO, J.D.P.; FILHO, F.A.S. Use of Machine Learning in Evaluation of Drought Perception in Irrigated Agriculture: The Case of an Irrigated Perimeter in Brazil. **Water**, v.12, 2020. https://doi.org/10.3390/w12061546