COMPARAÇÃO DE MÉTODOS TRADICIONAIS E ALTERNATIVOS PARA O PREENCHIMENTO DE SÉRIES TEMPORAIS DE PRECIPITAÇÃO

MIRELLE TAINÁ VIEIRA LIMA[1]; INGRID DE OLIVEIRA CAVALCANTE LIMA[2]; JULIANA PERTILLE; FABRÍCIO DA SILVA TERRA [3]

[1]Universidade Federal de Pelotas (UFPel) – mirellet.vieira@gmail.com
[2] Universidade Federal do Rio Grande (FURG) – ingrid.limaoc@hotmail.com
[3]Universidade Federal de Pelotas (UFPel) – juliana.pertill@gmail.com
[4]Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM) – fabricio.terra@ufvjm.edu.br

1. INTRODUÇÃO

A escassez de dados climatológicos no Brasil é um relevante entrave ao planejamento público e privado, à aplicação de recursos e ao desenvolvimento de pesquisas (RUEZZENE et al., 2021). Além da rede insuficiente de postos de monitoramento, dada a grande extensão territorial e variabilidade espacial climática no país; alguns problemas, como falhas nos dispositivos ou ausência de operadores nas estações, acarretam em erros de medição, inconsistências e dados faltantes. Para contornar a carência de informações, o ramo da hidrologia apresenta diversos métodos para o preenchimento de dados em séries climáticas, os quais são, no geral, dependentes de informações de postos de observação vizinhos. Contudo, nos últimos anos, os métodos de inteligência artificial, voltados à previsão ou imputação de dados em séries temporais, vêm se constituindo como um importante aliado para agregar o preenchimento de dados ambientais (AGUILERA et al., 2020; DIOUF et al., 2022).

Desse modo, o objetivo deste trabalho é comparar dois métodos tradicionais de preenchimento de falhas em dados históricos de chuva, a ponderação regional e a regressão linear, com os métodos de imputação de dados disponíveis nos pacotes do software R, “imputeTS” e “dlookr”.

2. METODOLOGIA

Por meio da Fundação Cearense de Meteorologia e Recursos Hídricos (Funceme), foi obtida a série histórica de precipitação do posto Dom Quintino, localizado no município de Crato-CE, às coordenadas 7,0413 Sul e 39,4712 Oeste. Para a aplicação do método ponderação regional, foram utilizadas as estações de apoio dos postos Caririaçu (a 20,5 km), Crato (a 22,5 km) e Nova Olinda (a 24,5 km), as mais próximas com a série completa, utilizando a seguinte equação:

$$P_x = \frac{1}{n}.\left(\frac{N_x}{N_A}.P_A + \frac{N_x}{N_B}.P_B + \frac{N_x}{N_C}.P_C\right)$$

Em que, P_x é a precipitação a ser estimada no posto X; P_A, P_B e P_C são as precipitações correspondentes ao mês ou ano que se deseja preencher nos postos vizinhos A, B e C; N_A, N_B e N_C são as precipitações médias nos postos vizinhos A, B e C; e N_x é a precipitação média em P_x.

Para a regressão linear, foi utilizada a estação Caririaçu como apoio, em que obteve-se a equação: y = 0.7231x - 4.8649 e coeficiente R^2 = 0.7514, onde y é o

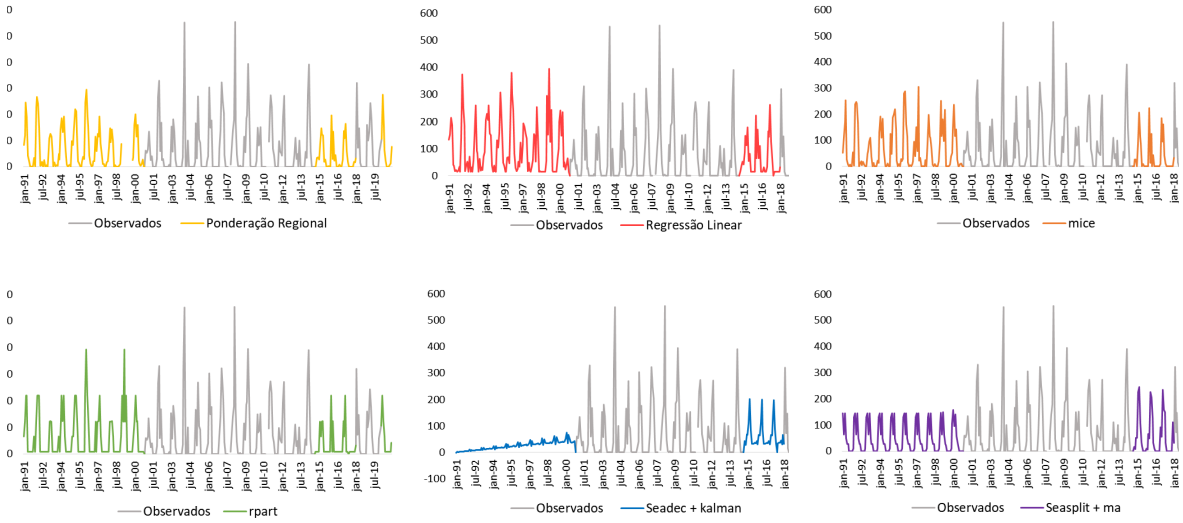
valor mensal previsto para o posto Dom Quintino e x é o valor observado no posto Caririaçu.

Ademais, foram utilizados métodos univariados de imputação de séries temporais, por meio dos pacotes “imputeTS” e “dlookr”, no software estatístico R, para preencher os dados mensais ausentes. Do pacote “imputeTS” foram selecionados o método de decomposição sazonal da função “na_seadec” (Seasonally Decomposed Missing Value Imputation) e o método de divisão sazonal da função “na_seasplit” (Seasonally Splitted Missing Value Imputation); utilizando os algoritmos “kalman” e “ma”, respectivamente (MORITZ; BARTZ-BEIELSTEIN, 2017). Do pacote dlookr, utilizou-se a função “imputate_na”, testando os algoritmos “mice” e “rpart” (RYU, 2019). Realizou-se o testes de raiz unitária Kpss, para verificação da estacionariedade das séries observada e preenchidas, utilizando a função “kpps.test” e p-value igual a 0,1; a qual parte da hipótese nula de que a série é estacionária (KWIATKOWSKI, 1992; TRAPLETTI et al., 2022).

3. RESULTADOS E DISCUSSÃO

O posto Dom Quintino apresenta dados a partir de novembro de 2000, porém, buscou-se estimar os dados de precipitação desde janeiro de 1991 a dezembro de 2020, totalizando o período de uma normal climatológica de 30 anos (Figura 1). No ano de 1999 e nos meses novembro e outubro de 2010 e setembro e outubro de 2007, não foi possível obter valores por ponderação regional devido à ausência de dados nos postos vizinhos, e por ser recomendado no mínimo 3 postos de apoio para a aplicação desse método, conforme BERTONI e TUCCI (2001). Ao todo, a série do posto Dom Quintino apresenta 54,3% de dados faltantes.

Figura 1. Imputação de dados faltantes de precipitação (mm) no posto Dom Quintino entre Janeiro de 1991 e Dezembro de 2020.



Por meio do teste kpps, constatou-se que o regime de precipitação analisado comporta-se de modo estacionário, isto é, a série varia ao longo do tempo, mas de forma que mantém uma média e variância constantes (Tabela 1). Vale ressaltar que quando o kpps com p-value calculado mostra-se inferior ao p-value tabulado

de 0.1 não se rejeita a hipótese nula e reforça a probabilidade da série ser estacionária.

Percebe-se que o "na_seadec" foi o método que mais apresentou dificuldade em deduzir valores faltantes no início da série, de modo que essa imputação reduziu a média histórica. O mice e rpart foram os mais eficientes em modelar a estacionariedade, dentre os métodos alternativos. Dessa forma, deduz-se que o pacote "dlookr" se sobressai em relação ao "imputeTS", sobretudo no período inicial da série, onde conseguiu estimar dados faltantes, mesmo sem dados prévios e obteve resultados semelhantes aos modelos hidrológicos tradicionais.

Em relação aos métodos tradicionais, a ponderação regional foi a que mais se aproximou da série sem preenchimento, em termos de média e variância. Por outro lado, o método da regressão linear estimou os valores mais altos para os dados faltantes.

Tabela 1. Estatística descritiva das séries históricas observada e estimadas.

Datas	Obs.	Pond. Regional	Reg. Linear	dlookr		imputeTS	
				mice	rpart	na_ seadec	na_ seasplit
Média	66,96	64,08	78,22	63,46	62,99	52,36	59,74
Variância	3,48	3,36	5,03	3,33	3,31	2,75	3,13
kpps	0,02	0,08	0,36	0,07	0,09	0,92	0,10

RUEZZENE et al. (2021) compararam os métodos tradicionais: razão normal, ponderação de distância inversa e regressão múltipla com o método de inteligência artificial de redes neurais, para a imputação de dados de chuva, e obtiveram que as redes neurais e a regressão múltipla apresentaram os melhores resultados, após a validação das previsões. DIOUF et al. (2022), avaliando o desempenho de cinco métodos de imputação: missForest, k-nn, ppca, mice and imputeTS, em dados meteorológicos do Senegal, concluíram que com baixos percentuais de dados faltantes, o desempenho é quase o mesmo para os todos os métodos de imputação analisados. Contudo, obtiveram que, o método de imputação de séries univariadas "imputeTS" foi o mais bem sucedido para reconstruir dados de séries temporais de precipitação. Quanto à função "imputate_na" do "dlookr", provavelmente por ser um pacote recente, não foram encontrados estudos utilizando-o na imputação de dados em séries temporais de precipitação. AGUILERA et al. (2022) compararam as técnicas de machine learning: STK, PMM e RF para simular diferentes porcentagens e padrões de dados ausentes, incluindo faltas extremas (> 90%), em um grande conjunto de dados diários extraído de 112 pluviômetros no período 1975–2017. Esses autores evidenciaram a importância de agregar à análise a verificação da consistência dos dados, dada a ocorrência de superestimação. Mas de modo geral, consideraram encorajadores os resultados obtidos através da aplicação dessas técnicas à falta extrema, tendo em vista que a previsão diária é um processo bastante delicado. Bem como, ressaltaram a necessidade de métodos que levem em conta a sazonalidade da precipitação, a fim de se obter estimativas mais confiáveis.

4. CONCLUSÕES

O presente trabalho comparou métodos tradicionais (ponderação regional e regressão linear) e alternativos ("na_seadec", "na_seasplit" e "imputate_na") para o preenchimento de falhas em séries temporais de precipitação. Os métodos alternativos apresentam vantagem em relação aos modelos hidrológicos tradicionais quando não há estações próximas e com disponibilidade dos dados necessários, visto que tratam-se de modelos univariados. No entanto, as funções "na_seadec" e "na_seasplit", do pacote "imputeTS", requerem dados anteriores às falhas para conseguir identificar a sazonalidade da série e imputar os dados faltantes coerentemente. Enquanto o pacote "dlookr", por meio da função "imputate_na" com os algoritmos "mice" e "rpart", apresentaram melhor desempenho em manter as características de estacionariedade e sazonalidade da série.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AGUILERA, H.; GUARDIOLA-ALBERT, C.; SERRANO-HIDALGO, C. Estimating extremely large amounts of missing precipitation data. **Journal of Hydroinformatics**, v. 22, n. 3, p. 578-592, 2020.

BERTONI, J. C.; TUCCI, C. E. M. Precipitação. In.: **Hidrologia: ciência e aplicação**, Org. Carlos E. M. Tucci, 2ª ed., Porto Alegre: Ed. Universidade/UFRGS: ABRH, 2001.

FUNCEME. Fundação Cearense de Meteorologia e Recursos Hídricos. Disponível em:http://www.funceme.br/produtos/script/chuvas/Download_de_series_historicas/DownloadChuvasPublico.php. Acesso em: Julho de 2022.

KWIATKOWSKI, D.; PHILLIPS, P.C.B.; SCHMIDT, P.; SHIN, Y. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?, **Journal of Econometrics**, 54, 159–178, 1992.

MORITZ, S.; BARTZ-BEIELSTEIN, T. imputeTS: time series missing value imputation in R. **The R Journal**, v. 9, n. 1, p. 207, 2017.

RUEZZENE, C. B.; DE 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.

RYU, Choonghyun. dlookr: Tools for Data Diagnosis, Exploration. **Transformation**, p. 352, 2019.

DIOUF, S.; DÈME, H.; DÈME, A. Imputation methods for missing values: the case of Senegalese meteorological data. **HAL Open Science**, 2022.

TRAPLETTI, A.; HORNIK, K.; LEBARON, B. Package 'tseries', 2022.