XIV ENCONTRO DE RECURSOS HÍDRICOS EM SERGIPE

PREENCHIMENTO DE FALHAS EM SÉRIES TEMPORAIS DE NÍVEIS DE ÁGUAS SUBTERRÂNEAS USANDO MÉTODOS UNIVARIADOS

Rubens Oliveira da Cunha Júnior [1]; *Celme Torres Ferreira da Costa* [2] & *Paulo Renato Alves Firmino* [3]

RESUMO: *O problema dos dados ausentes em séries temporais é comum na modelagem de águas subterrâneas. Diante disso, os métodos de imputação ou de preenchimento de falhas são necessários na etapa de pré-processamento dos dados. Este trabalho avaliou o desempenho de métodos univariados de imputação de séries temporais aplicados a dados de níveis diários de águas subterrâneas. O conjunto de dados consistiu em séries de níveis estáticos medidos em poços monitorados pela RIMAS/CPRM em bacias sedimentares da região semiárida do Nordeste do Brasil. Através de simulações de valores ausentes segundo os conceitos de mecanismos da ausência de dados, métodos como média aritmética, interpolação linear e por spline cúbica, médias móveis e imputação com decomposição sazonal foram aplicados e os resultados comparados segundo o RMSE. Os resultados mostraram que métodos que fazem uso das características da série, como os baseados em decomposição, obtiveram melhores desempenhos. Para baixas porcentagens de falhas, métodos simples, como a interpolação linear obtiveram bom desempenho. Os resultados obtidos neste trabalho contribuem para a solução de problemas relacionados à etapa de pré-processamento de dados na modelagem de águas subterrâneas.*

Palavras-Chave – Dados ausentes; Imputação; Hidrogeologia

INTRODUÇÃO

A modelagem de séries temporais para processos hidrológicos requer dados confiáveis, disponíveis de forma contínua e a longo prazo (KENDA *et al.* 2018). Para a modelagem em águas subterrâneas, a principal fonte de informações são as medidas de poços de monitoramento. Contudo, dados de águas subterrâneas são limitados em muitas regiões, sendo comum a ocorrência de períodos sem informações disponíveis, isto é, com falhas, devido a fatores técnicos, humanos, institucionais ou naturais. Para solucionar tal problema, pode-se recorrer a técnicas para se estimar valores em substituição àqueles em falta, conhecidas como métodos de preenchimento de falhas ou de imputação de dados ausentes (OIKONOMOU *et al.* 2018).

Em especial, o preenchimento de falhas em séries de níveis de águas subterrâneas são um problema complexo. Autores como Zhang *et al.* (2017), Kenda *et al.* (2018), Oikonomou *et al.* (2018) e He *et al.* (2020) têm se dedicado ao desenvolvimento de métodos de imputação específicos para

1) Centro de Ciências Agrárias e da Biodiversidade (CCAB), Universidade Federal do Cariri (UFCA), Ícaro Moreira de Sousa, 63130025, Crato, CE, Brasil, cunhajunior.rubens@gmail.com
2) Centro de Ciências e Tecnologia (CCT), Universidade Federal do Cariri (UFCA), Tenente Raimundo Rocha, 63048080, Juazeiro do Norte, CE, Brasil, celme.torres@ufca.edu.br
3) Centro de Ciências e Tecnologia (CCT), Universidade Federal do Cariri (UFCA), Tenente Raimundo Rocha, 63048080, Juazeiro do Norte, CE, Brasil, paulo.firmino@ufca.edu.br

dados de águas subterrâneas. Diante das dificuldades encontradas, muitos hidrogeologistas também têm adotado estratégias univariadas para a imputação em séries de níveis de águas subterrâneas (SAKIZADEH *et al*. 2019; BRÉDY *et al*. 2020; SAHU *et al*. 2020; MÜLLER *et al*. 2021). Tais técnicas não se baseiam em relações entre covariáveis, mas exploram as características da própria série temporal em estudo (MORITZ *et al*. 2015).

Este trabalho avalia o desempenho de métodos univariados de imputação em séries de níveis de águas subterrâneas. Dados de poços monitorados pela Rede Integrada de Monitoramento de Águas Subterrâneas (RIMAS) serão usados. Através de simulações de valores ausentes, o desempenho dos métodos será avaliado segundo métricas de erro bem conhecidas na literatura.

MATERIAIS E MÉTODOS

O presente estudo foi aplicado em diferentes localidades da região Nordeste do Brasil, nos Estados do Ceará e do Piauí, onde estão localizados os poços selecionados. Todos os poços estão instalados em bacias sedimentares na região semiárida do Brasil. A Tabela 1 mostra informações sobre os poços selecionados. O conjunto de dados é formado por 3 séries temporais diárias da evolução dos níveis de água subterrânea medidos em poços da Rede Integrada de Monitoramento de Águas Subterrâneas (RIMAS). A escolha dos poços se baseou em selecionar séries temporais que possuíssem longos períodos de observações diárias sem falhas.

Tabela 1 – Informações dos poços e séries selecionados

Série	Poço	Latitude	Longitude	UF	Aquífero	Bacia	Núm. Obs.	Período Obs.
P1	2200046785	-7,233	-41,912	PI	Cabeças	Parnaíba	2706	Nov/2011 - Mar/2019
P2	2200046856	-6,832	-41,747	PI	Cabeças	Parnaíba	2555	Nov/2011 - Out/2018
P3	2300022909	-7,218	-39,203	CE	Médio	Araripe	1762	Set/2016 - Jun/2021

Para a caracterização das séries temporais e identificação de tendência ou sazonalidade, realizou-se a análise de autocorrelação das séries, através do estudo dos gráficos da Função de Autocorrelação (ACF) (HYNDMAN; ATHANASOPOULOS, 2018).

Para a avaliação do desempenho dos métodos de imputação, foram realizadas simulações de valores ausentes, utilizando os mecanismos de ausência de dados. Little e Rubin (2019) definem três mecanismos que descrevem como as falhas ocorrem. No mecanismo de Ausência Completamente Aleatória (MCAR, *Missing Completely at Random*), os valores ausentes não têm relação com nenhuma variável. A Ausência Aleatória (MAR, *Missing at Random*) ocorre quando as falhas têm relação com outras variáveis. O mecanismo de Ausência Não Aleatória (MNAR, *Missing Not at Random*) considera que as falhas estão relacionadas a outras falhas. O mecanismo MAR é apropriado para dados hidrológicos (JUNNINEN *et al*. 2004), sendo usado para simular falhas causadas por erros em equipamentos durante longos períodos de tempo (BECK *et al*. 2018).

Neste estudo, foram aplicados 5 métodos univariados de imputação para séries temporais, a saber, média aritmética (MA), interpolação linear (LIN) e por Spline (SPL), médias móveis (MM) e imputação com decomposição sazonal (DEC).

O método da média aritmética (MA) é uma técnica simples de imputação, em que as falhas são substituídas pela média das observações disponíveis (ASGHARINIA; PETROSELLI, 2020). O

método das médias móveis (MM) calcula médias ponderadas a partir de um número igual de observações em ambos os lados de um valor ausente central (MORITZ *et al*. 2015).

A interpolação linear (LIN) ajusta uma reta entre os extremos de um intervalo com falhas e permite que os valores ausentes no interior deste intervalo sejam estimados. A imputação por spline cúbica (SPL) é realizada ajustando-se polinômios de grau três (JUNNINEN *et al*. 2004).

A imputação com decomposição (DEC) inicialmente remove a componente de sazonalidade da série, e então realiza uma imputação nas componentes de tendência e de resíduo. Por fim, a componente sazonal é adicionada de volta (MORITZ; BARTZ-BEIELSTEIN, 2017).

O experimento realizado consiste em 4 etapas: (i) simular valores ausentes gerados artificialmente; (ii) aplicar métodos de imputação; (iii) avaliar o desempenho de cada método; e (iv) comparar os resultados obtidos. Foram adotadas 3 porcentagens de falhas: 5%, 10% e 20%. Para cada porcentagem, foram geradas 100 amostras aleatórias independentes, segundo o mecanismo de ausência aleatória (MAR). As amostras são valores selecionados para serem removidos. Em cada amostra, os métodos MA, LIN, SPL, MM e DEC foram aplicados para estimar os valores ausentes. A partir dos valores removidos e dos valores estimados pelos métodos, foram então calculadas as medidas da Raiz do Erro Quadrático Médio (RMSE) para cada método. Procedendo-se dessa maneira para cada uma das 100 amostras, foram obtidos 100 valores diferentes de RMSE para cada método. Por fim, calculou-se a média dos 100 valores de RMSE e os foram resultados comparados. A Equação 1 mostra o cálculo do RMSE.

$$RMSE = \sqrt{\sum_{t=1}^{n}(o_t - e_t)^2 / n} \quad (1)$$

Em que o_t é o valor observado no instante de tempo t, e_t é o valor correspondente estimado por um método de imputação, e n é número de falhas. As análises foram executadas em linguagem R (R CORE TEAM, 2021). A simulação foi realizada com o pacote *imputeTestbench*. Os métodos MA, LIN e SPL foram aplicados com o pacote *zoo*, e os métodos DEC e MM com o pacote *imputeTS*.

RESULTADOS E DISCUSSÃO

A Figura 1 mostra os gráficos das séries diárias dos níveis nos poços selecionados e os respectivos gráficos da Função de Autocorrelação (ACF). A série P1 mostrou um rebaixamento no nível a longo prazo durante o período considerado. As séries P2 e P3 possuem uma maior estabilidade do nível em torno do seu valor médio. Destaca-se ainda o aumento periódico do nível no primeiro semestre de cada ano, em virtude da recarga durante as estações chuvosas na região.

As linhas horizontais em azul nos gráficos da ACF representam o nível de significância de 5%. Segundo a ACF, a série P1 possui fortes padrões de tendência e sazonalidade. A alta correlação nos *lags* menores, decaindo suavemente, é típico de séries com tendência. Os picos periódicos da correlação nos *lags* próximos a 365 indicam a sazonalidade (HYNDMAN; ATHANASOPOULOS, 2018). A série P2 possui um comportamento predominantemente sazonal, mas pode ser percebido uma fraca tendência; enquanto a série P3 revela um forte padrão sazonal.

A Tabela 3 mostra as médias dos valores de RMSE (medidos em cm) obtidos pelos métodos diante de cada porcentagem de falhas. Os melhores resultados estão destacados em negrito na tabela. Para 5% de falhas, o método LIN obteve melhor desempenho em todas as séries. Isto sugere que para pequenas porcentagens de falhas, métodos mais simples, como a interpolação linear podem ser adequados. Sob 10% de falhas, o método DEC foi superior em relação aos demais nas séries P1, P2 e P3. Resultado semelhante ocorreu para situações com 20% de falhas.

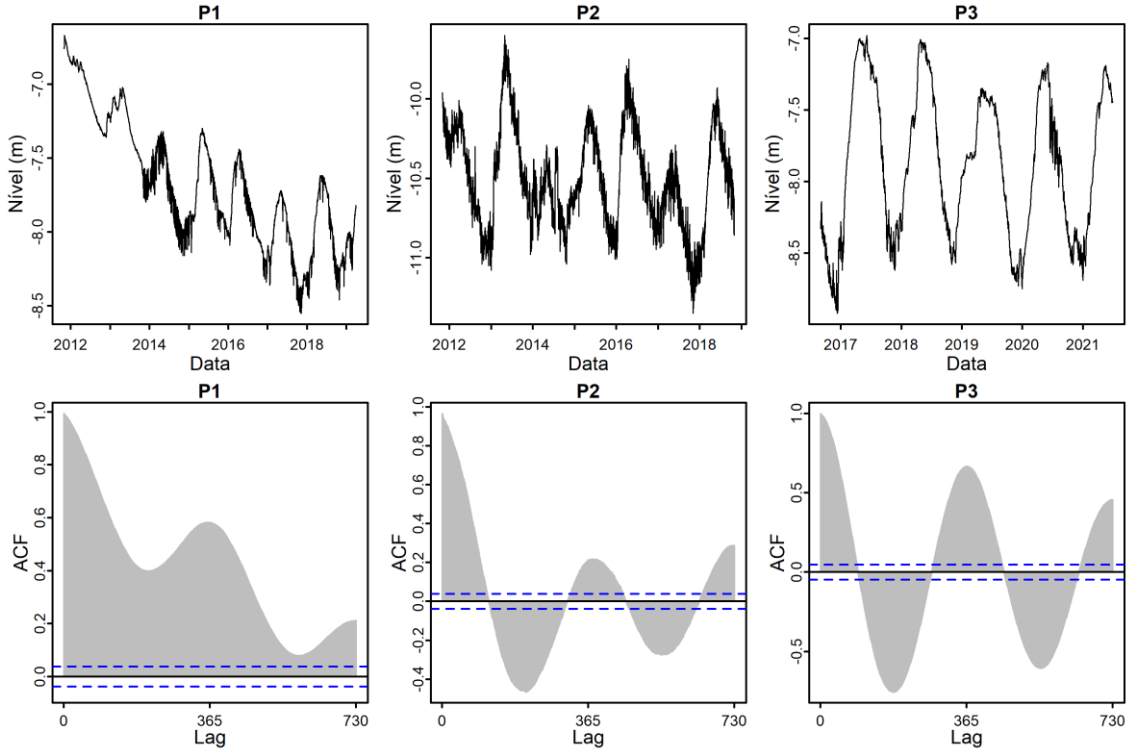

Figura 1 – Gráficos das séries de níveis diários de águas subterrânea e da Função de Autocorrelação (ACF)

Tabela 2 – Resultados das imputações: RMSE médio (medido em cm) obtido em 100 simulações.

Melhores resultados destacados em negrito

Falha	Método	P1	P2	P3
5%	MA	8,33	6,96	11,21
	LIN	**1,53**	**2,60**	**1,84**
	SPL	13,71	22,97	6,02
	MM	2,07	3,20	3,05
	DEC	1,62	2,72	2,06
10%	MA	12,25	9,54	16,66
	LIN	3,52	5,14	5,00
	SPL	46,23	63,11	16,98
	MM	4,69	6,44	7,99
	DEC	**2,98**	**4,62**	**4,39**
20%	MA	18,68	14,35	23,09
	LIN	9,80	13,08	15,70
	SPL	86,65	211,00	65,54
	MM	10,54	13,86	18,45
	DEC	**5,71**	**8,88**	**8,82**

Os métodos DEC e LIN atingiram melhores resultados, seguidos por MM, MA e SPL. A imputação pela média aritmética (MA) subestima o desvio-padrão das observações e reduz a variabilidade dos dados. Quanto ao método SPL, a qualidade da imputação depende da natureza da série em estudo. Por sua vez, o método MM é influenciado pelo comprimento da falha. Em geral, o desempenho de todos os métodos foi influenciado pela porcentagem de falhas, pois os valores médios

do RMSE aumentaram conforme aumentou-se a taxa de falhas. Estes resultados estão de acordo com os achados de Junninen *et al*. (2004) e Moritz *et al*. (2015).

O desempenho de métodos baseados em decomposição, como o DEC, depende do cálculo das componentes da série. As características da série e o comprimento das falhas são fatores que influenciam a decomposição e, por consequência, a qualidade da imputação. Séries de níveis águas subterrâneas exibem padrões complexos, pois as dinâmicas dos lençóis freáticos são influenciadas por diversos fatores, como precipitação, temperatura, entre outros (BRÉDY *et al*. 2020). Diante da complexidade das séries, justifica-se o melhor resultado do método DEC na maioria dos casos, frente a técnicas mais simples. Métodos de imputação baseados em decomposição têm sido aplicados por hidrogeologistas, como Sakizadeh *et al*. (2019) e Müller *et al*. (2021).

Apesar do uso de séries diárias neste estudo, para a imputação em séries de níveis de águas subterrâneas, os métodos analisados podem ser aplicados a dados com diferentes frequências, tais como horária (BRÉDY *et al*. 2020), mensal (SAKIZADEH *et al*. 2019; ASGHARINIA; PETROSELLI, 2020) e diária (SAHU *et al*. 2020; MÜLLER *et al*. 2021).

CONCLUSÕES

Métodos de imputação que fazem uso das características da série, como a decomposição, obtiveram melhores desempenhos na maioria dos casos segundo o RMSE médio. Para baixas porcentagens de falhas, métodos simples, tais como a interpolação linear foram superiores. As simulações realizadas possibilitaram a escolha do método de mais apropriado para cada caso. Os resultados obtidos contribuem para a solução de problemas relacionados à etapa de pré-processamento de dados na modelagem de águas subterrâneas.

AGRADECIMENTOS

O primeiro autor agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudos concedida durante a execução deste estudo.

REFERÊNCIAS

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

BECK, M. W.; BOKDE, N.; ASENCIO-CORTÉS, G.; KULAT, K. (2018). "*R package imputetestbench to compare imputation methods for univariate time series"*. The R Journal 10(1), pp. 218.

BRÉDY, J.; GALLICHAND, J.; CELICOURT, P.; GUMIERE, S. J. (2020). "*Water table depth forecasting in cranberry fields using two decision-tree-modeling approaches"*. Agricultural Water Management 233, pp. 106090.

HE, L.; CHEN, S.; LIANG, Y.; HOU, M.; CHEN, J. (2020). "*Infilling the missing values of groundwater level using time and space series: case of Nantong City, east coast of China"*. Earth Science Informatics 13(4), pp. 1445–1459.

HYNDMAN, R. J.; ATHANASOPOULOS, G. (2018). *Forecasting: principles and practice*. Melbourne, Australia: OTexts. Disponível em: https://www.OTexts.com/fpp2. Acesso em: fev. 2021.

JUNNINEN, H.; NISKA, H.; TUPPURAINEN, K.; RUUSKANEN, J.; KOLEHMAINEN, M. (2004). "*Methods for imputation of missing values in air quality data sets"*. Atmospheric Environment 38(18), pp. 2895-2907.

KENDA, K.; KOPRIVEC, F.; MLADENIĆ, D. (2018). "*Optimal missing value estimation algorithm for groundwater levels"*. Multidisciplinary Digital Publishing Institute Proceedings 2(11), pp. 698.

LITTLE, R.; RUBIN, D. (2018). *Statistical Analysis with Missing Data*. Wiley.

MORITZ, S.; SARDÁ, A.; BARTZ-BEIELSTEIN, T.; ZAEFFERER, M.; STORK, J. (2015). "*Comparison of diferente methods for univariate time series imputation in r"*. arXivpreprint arXiv:1510.03924.

MORITZ, S.; BARTZ-BEIELSTEIN, T. (2017). "*imputeTS: time series missing value imputation in R"*. The R Journal 9(1), pp. 207–218.

MÜLLER, J.; PARK, J.; SAHU, R.; VARADHARAJAN, C.; ARORA, B.; FAYBISHENKO, B.; AGARWAL, D. (2021). "*Surrogate optimization of deep neural networks for groundwater predictions*". Journal of Global Optimization 81(1), pp. 203-231.

OIKONOMOU, P. D.; ALZRAIEE, A. H.; KARAVITIS, C. A.; WASKOM, R. M. (2018). "*A novel framework for filling data gaps in groundwater level observations"*. Advances in Water Resources 119, pp. 111–124.

R CORE TEAM. (2021). *R: A language and environment for statistical computing*.

SAHU, R. K.; MÜLLER, J.; PARK, J.; VARADHARAJAN, C.; ARORA, B.; FAYBISHENKO, B.; AGARWAL, D. (2020). "*Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network"*. Frontiers in Water, Frontiers 2, pp. 46.

SAKIZADEH, M.; MOHAMED, M. M.; KLAMMLER, H. (2019). "*Trend analysis and spatial prediction of groundwater levels using time series forecasting and a novel spatio-temporal method"*. Water Resources Management 33(4), pp. 1425–1437.

ZHANG, Z.; YANG, X.; LI, H.; LI, W.; YAN, H.; SHI, F. (2017). "*Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level"*. Journal of Hydrology 553, pp. 384–397.