



Validação da Imputação Múltipla via *Predictive Mean Matching* para Preenchimento de Falhas nos Dados Pluviométricos da Bacia do Médio São Francisco
Validation of Multiple Imputation by Predictive Mean Matching for Pluviometric Data in Middle San Francisco Basin

Laurizio Emanuel Ribeiro Alves¹ & Heliofábio Barros Gomes²

¹Universidade Federal de Alagoas, Instituto de Ciências Atmosféricas, Programa de Pós-Graduação em Meteorologia,
Av. Lourival Melo Mota, s/n, 57072-970, Maceió, AL, Brasil

²Universidade Federal de Alagoas, Instituto de Ciências Atmosféricas,
Av. Lourival Melo Mota, s/n, 57072-970, Maceió, AL, Brasil
E-mails: laurizio.r@gmail.com; heliofab@gmail.com

Recebido em: 09/08/2019 Aprovado em: 21/11/2019

DOI: http://dx.doi.org/10.11137/2020_1_199_206

Resumo

Um dos principais problemas atualmente para analisar longas séries de dados no Brasil é a falta de um banco de dados diários consistentes de estações meteorológicas. Diante disso, o objetivo deste trabalho é avaliar a acurácia do método de imputação múltipla *Predictive Mean Matching* (PMM) no preenchimento de dados faltantes de séries diárias de precipitação para a Bacia Hidrográfica do Médio São Francisco (BMSF). Para isso foram adquiridos dados diários de chuva cedidos pelo Instituto Nacional de Meteorologia (INMET) das estações de Bom Jesus da Lapa (BA), Carinhanha (BA), João Pinheiro (MG), Remanso (BA) e Unai (MG), para o período de 2001 a 2017. Em seguida, dois cenários foram criados com 5% e 15% de falhas para avaliar a precisão do método PMM no preenchimento de dados faltantes. As séries originais, sem falhas, e as preenchidas pelo método adotado apresentaram uma correlação alta ($r > 0,80$), que consiste numa boa relação entre elas. O coeficiente de determinação (R^2) foi de 0,7 (0,6) para 5% (15%) de falhas. Além disso, o Erro Médio Absoluto e o Erro Quadrático Médio foram baixos para todas as estações. Também foi aplicado o teste de Wilcoxon, o qual verificou a boa acurácia na aplicação do método PMM para preencher os dados faltantes de chuva.

Palavras-chave: Chuva; Série sintética; Testes estatísticos

Abstract

One of the main problems currently in analyzing long data sets in Brazil is the lack of a consistent database of daily weather stations. Therefore, the objective of this work is to evaluate the accuracy of the Predictive Mean Marching (PMM) multiple imputation method in filling missing data from daily rainfall series. For this purpose, daily rainfall data from the National Institute of Meteorology (INMET) from Bom Jesus da Lapa (BA), Carinhanha (BA), João Pinheiro (MG), Remanso (BA) and Unai (MG) stations were acquired from the period 2001 to 2017. Then two scenarios were created with 5% and 15% failures so that it was possible to evaluate the accuracy of the PMM method for filling in missing data. The original series, without fail, and the ones completed by the adopted method presented a high correlation ($r > 0.80$), which consists of a good relationship between them, and coefficient of determination (R^2) was 0.7 (0.6) for 5% (15%) failures. In addition, Absolute Mean Error and Mean Square Error were low for all seasons. The Wilcoxon test was also applied, which verified the good accuracy in the application of the PMM method to fill in the missing rainfall data.

Keywords: Rainfall; Synthetic Series; Statistical tests

1 Introdução

Atualmente um dos principais problemas ao analisar longas séries de dados no Brasil é a falta de informações consistentes de estações meteorológicas. De acordo com Bier & Ferraz (2017), essas falhas geralmente estão associadas a problemas nos equipamentos, quando a estação for automática, ou falta de observadores, quando a estação é convencional. Como os dados são comumente utilizados para fins científicos, a necessidade de preencher essas falhas é fundamental para uma melhor avaliação da variável em questão.

Diversas técnicas de preenchimento de dados faltantes foram desenvolvidas e aplicadas eficientemente nos últimos anos para séries temporais meteorológicas mensais ou anuais, tais como: o método de Ponderação Regional (Tucci, 2001; Diaz *et al.*, 2018) e a Regressão Linear (Tucci, 2001; Bier & Ferraz, 2017; Melo *et al.*, 2017). Entretanto, pouco se discute a respeito do preenchimento de séries de dados mais longas, como dados diários por exemplo, pois o preenchimento desses bancos de dados pode acarretar erros grosseiros.

Segundo Heidt (2019), que testou alguns métodos de imputação múltipla, os métodos *Predictive Mean Matching* (PMM; Rubin, 1986; Heitjan & Little, 1991) e *Classification and Regression Trees* (CART; Segal, 1988; Burgette & Reiter, 2010) obtiveram os melhores resultados, apresentando bom desempenho em todos os percentis de falhas de dados, com valores próximos aos valores originais, enquanto os demais métodos, *On-the-fly imputation* (OTF; Heidt, 2019) e *Randon Forest* (RF; Breiman, 2001), foram inferiores.

Moura *et al.* (2014) aplicaram o preenchimento de falhas por imputação múltipla utilizando o método PMM em dados de anomalia de altura geopotencial e encontraram ótimos resultados quando comparados os valores preditos e observados. Lopo (2014) aplicou esse método (PMM) para preencher dados faltantes de séries de radiação ultravioleta (UV) e verificou que a variância é homogênea entre os grupos de dados imputados e não imputados, logo, o método foi satisfatório.

Nessa perspectiva, este estudo tem como objetivo avaliar o preenchimento de falhas por imputação múltipla PMM e averiguar o grau de consistência desse método para essa prática, para isso, avaliou-se dados diários de precipitação pluviométrica para a Bacia do Hidrográfica do Médio São Francisco (BMSF). A bacia foi selecionada por estar inserida no semiárido nordestino, ser afetada frequentemente por secas severas e apresentar grande defasagem ou perda de séries históricas de estações meteorológicas (Marengo *et al.*, 2016; Cunha *et al.*, 2017; CBHSF, 2017).

2 Metodologia e Dados

2.1 Área de estudo

A BMSF está inserida em uma das regiões mais secas do Brasil, o semiárido, e, também apresenta o menor desenvolvimento das sub-bacias da Bacia Hidrográfica do Rio São Francisco (BHSF). Cerca de 3,23 milhões de pessoas habitam a bacia (MMA, 2006). A BMSF tem uma área de 402.531km², da qual 57% dessa área é urbanizada, possui densidade demográfica de 8 hab/km² e está localizada entre os estados da Bahia e Minas Gerais (Figura 1).

O clima predominante é o Tropical com inverno seco (Aw) de acordo com a classificação de Köppen (Alvares *et al.*, 2013), com chuva média anual variando entre 600 mm a 1.400 mm, evapotranspiração média anual de 1.300 mm e temperatura média de 24°C. Os meses de novembro a janeiro são os mais chuvosos, contribuindo com 53% da chuva anual, ao passo que o período mais seco ocorre entre junho e agosto (CBHSF, 2004).

Na Figura 2 são apresentadas as normais climatológicas (1981-2010) das estações meteorológicas contidas na BMSF para a média das variáveis: total de chuva, temperatura média do ar e Umidade Relativa (UR). É notória a sazonalidade da chuva, onde se observa um padrão semestral, com mês de dezembro sendo o mais chuvoso (média no período de 209,6 mm) e o mês de julho como o mais seco (média no período de 2,0 mm). Este comportamento semestral também foi observado por Silva *et al.* (2018), que avaliou a climatologia de hidroelétricas brasileiras. Os autores notaram que o padrão de precipitação na região das hidroelétricas de Xingó e Sobradinho, inseridas na BHSF, é semestral, no qual de outubro a março corresponde ao período chuvoso

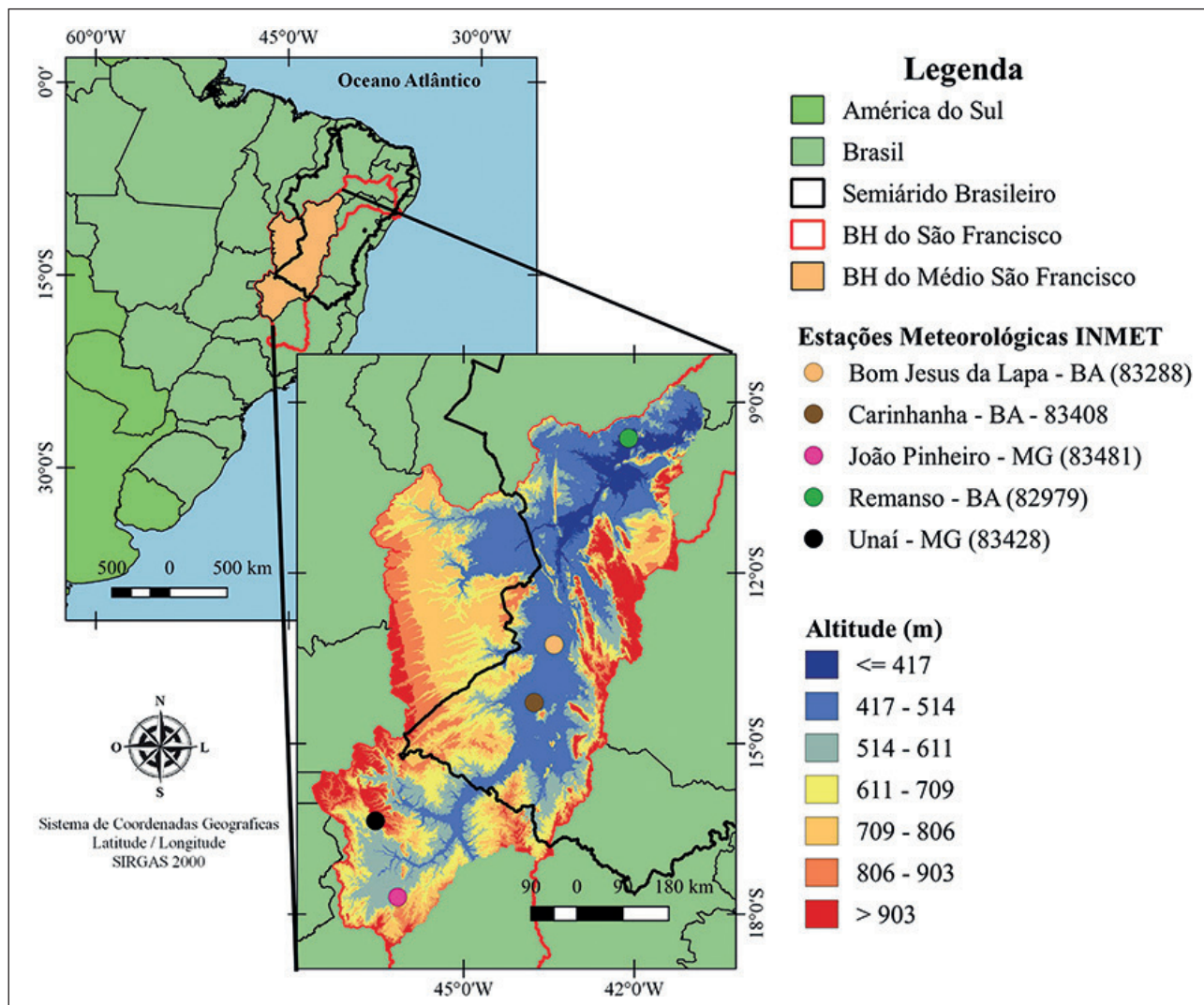


Figura 1 Mapa de localização da BMSF, distribuição espacial das estações meteorológicas e altitude (m).

e de abril a setembro, ao seco. Essa sazonalidade é observada, também, na temperatura média, que apresenta valores mais baixos no período de estiagem e maiores valores no período chuvoso. Com relação a UR esse padrão sazonal não fica muito claro, o que se nota é uma redução da UR ao longo do período seco, com mínima no mês de setembro, e aumento no período chuvoso, com máxima no mês de março.

2.2 Aquisição e Processamento dos Dados de Chuva

Para a análise da chuva da região foram utilizados dados diários de 5 estações meteorológicas convencionais. Suas informações são apresentadas

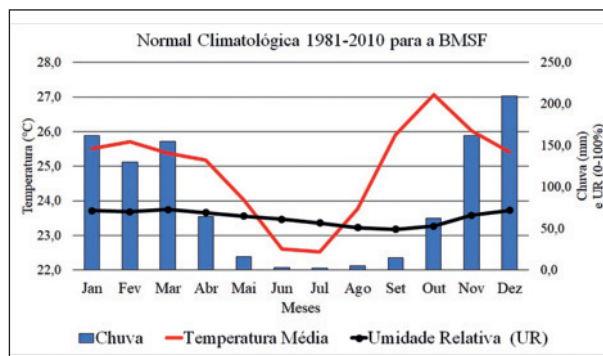


Figura 2 Normais climatológicas (1981-2010) para a BMSF. Linha vermelha, eixo vertical esquerdo: Temperatura média (°C); barras azuis e linha preta, eixo vertical direito: Chuva (mm) e UR (0-100%), respectivamente. Fonte: INMET, adaptado pelos autores.

na Tabela 1, incluindo suas coordenadas espaciais, altitude e percentual de falhas. Os dados diários foram adquiridos através do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET), entre 2001 e 2017, disponível no sítio eletrônico: < <http://www.inmet.gov.br/projetos/rede/pesquisa/inicio.php>>.

Após a aquisição dos dados foi realizada a manipulação e a organização das séries. Em seguida verificou-se a consistência das séries históricas que apresentam baixa defasagem de dados no período 2001-2017 com percentual de falhas inferior a 5% (Tabela 1).

ID	Estação	Lat (°)	Lon (°)	Alt (m)	Percentual de Falhas (%)
1	Bom Jesus da Lapa BA - 83288	-13,26	-43,41	439,96	0,2
2	Carinhanha - BA - 83408	-14,28	-43,76	450,18	4,6
3	João Pinheiro - MG - 83481	-17,70	-46,16	760,36	4,6
4	Remanso - BA - 82979	-9,63	-44,51	500,51	0,3
5	Unai - MG - 83428	-16,36	-46,55	460,00	0,0
				Média	1,94

Tabela 1 Informações das estações meteorológicas.
Fonte: INMET, 2009.

Para realizar o preenchimento das falhas foi utilizado o pacote do software R denominado *Multivariate Imputation by Chained Equations* (MICE), que consiste em diferentes procedimentos de imputação (Horton & Lipsitz, 2001). Para este estudo foi aplicado o método de imputação múltipla *Predictive Mean Matching* (PMM) para os dados faltantes. Esse método é baseado na metodologia de Moura *et al.* (2014), que destacam que para realizar o preenchimento dos dados faltantes uma nova série de dados sintéticos é gerada. Além do mais, os autores ressaltam que a imputação pode ser aplicada desde que sejam respeitados os critérios de proporção dos dados faltantes (5% e 15%, para uma amostra e população, respectivamente). O método de imputação múltipla PMM é considerado de baixa incerteza, uma vez que combina elementos de regressão, vizinho mais próximo e imputação *hot deck* (técnica paramétricas e não paramétricas). A metodologia é descrita a seguir:

$$Y^{obs} = \{Y_i^{obs} = X_i' B^* : i \in obs(Y)\} \quad (1)$$

Sendo X uma variável sem dados faltantes; Y^{obs} o conjunto de valores observados; $Y_i = X_i' B^*$; e Y_j^{obs} ; e a observação encontrada correspondente ao valor mais próximo de Y_j^* .

O método pode ser calibrado em termos de qualidade de imputação, conforme sua acurácia, concordância e dispersão. Para se verificar a confiabilidade deste método (PMM) se sobressaem, o Erro Quadrático Médio (EQM), o Erro Médio Absoluto (EMA) e a Correlação de Pearson (r), sendo essa última (r) considerada a mais adequada para verificar o desempenho do método de imputação múltipla (Moura *et al.*, 2014).

Para verificar a acurácia da imputação e criação da nova série sintética foram selecionados períodos sem falhas para realizar os testes estatísticos. Para realizar esses testes foi criado um algoritmo para simular falhas, em que foram gerados dois cenários, o primeiro com 5% de falhas, com o propósito de avaliar o método quando há poucas falhas, e o segundo com 15%, quando há maior número de falhas. Essas falhas compreendem a 5% e 15%, visto que o método utilizado recomenda esses valores (Harrel Jr., 2014). As estações de teste selecionadas foram: Bom Jesus da Lapa; Carinhanha – BA; João Pinheiro – MG; Remanso – BA; e Unai – MG.

A primeira fase da verificação consistiu na identificação da intensidade do nível de correlação entre as variáveis. Para isso, foi aplicado o Coeficiente de Correlação Pearson (r). Esse coeficiente mede o grau de ajustamento dos valores em torno de uma reta (Lin, 1989; Correa, 2003), e é dado pela Equação 2.

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}} \quad (2)$$

Temos que, n é o número de observações, x_i é variável independente e y_i é a variável dependente. O coeficiente r varia de -1 a +1, logo, quanto mais próximo da unidade maior correlação entre as variáveis, os valores positivos e

negativos indicam se essa correlação é direta ou indireta, na devida ordem. De acordo com Correa (2003), se estabeleceu critérios para verificar os níveis de correlação, independente se a correlação é direta ou não. Na Tabela 2 a seguir são apresentados os níveis de correlação baseados nesse coeficiente.

Classe	Classe (%)	Nível de Correlação
$0,0 < r < 0,3$	$0 < r < 30$	A correlação é fraca, sem relação entre as variáveis
$0,3 \leq r < 0,6$	$30 \leq r < 60$	A correlação é fraca, porém, pode considerar relativa correlação entre as variáveis
$0,6 \leq r < 1,0$	$60 \leq r < 100$	A correlação é média para forte; a relação entre as variáveis é significativa, o que permite coerência com poucos conflitos na obtenção das conclusões.

Tabela 2 Grau de correlação entre as variáveis dependente e independente baseado em r. Fonte: Correa (2003).

Apenas o coeficiente r não consegue expressar com maior precisão a consistência dos valores imputados. Nessa perspectiva, a aplicação de outros indicadores estatísticos é necessária para um julgamento mais concreto dos valores examinados. Os métodos aplicados para avaliar a consistência dos dados foram o Coeficiente de Determinação (R^2), o EMA e o EQM. R^2 (Equação 3) é a medida de ajuste entre os dados sintéticos (imputados via PMM) e os dados observados. EMA mede o valor médio de erro entre as séries, observada e imputada, dada pela Equação 4. EQM representa o cálculo ponderado dos desvios médio, logo, quanto menor o valor de EQM (Equação 5), melhor o ajuste entre o valor observado e imputado (Camelo *et al.*, 2017a; 2017b; González *et al.*, 2019).

$$R^2 = \frac{\sum_{i=1}^n (imput_i - imput_{médio})(imput_i - obs_{médio})}{\sqrt{\sum_{i=1}^n (imput_i - imput_{médio})^2 \sum_{i=1}^n (obs_i - obs_{médio})^2}} \quad (3)$$

$$EMA = \sum_{i=1}^n |imput_i - obs_i| / n \quad (4)$$

$$EQM = \sum_{i=1}^n (imput_i - obs_i)^2 / n \quad (5)$$

Onde *imput* é o valor da imputação múltipla e *obs* é o valor observado e *n* é número de amostras.

Além disso, aplicou-se nas séries o teste de Shapiro-Wilk para averiguar se a série de dados é paramétrica ou não-paramétrica (Royston, 1992; Silva, 2018). Após definir qual o tipo de série está sendo analisada, realiza-se os testes estatísticos para comparar as amostras. Neste estudo, as séries, segundo o teste de Shapiro-Wilk, são não-paramétricas, pois têm p-valor inferiores a 0,05 ($2,2 \cdot 10^{-16}$) e, por isso, não apresentam distribuição normal. Por conta disso, foi aplicado o teste não-paramétrico de Wilcoxon para amostras pareadas com nível de significância de 0,05 (p-value) (Fix & Hodges Jr., 1955; Silva *et al.*, 2019). A hipótese alternativa H_a testada é que não há diferença significativa entre as médias da chuva observada e da série sintética (imputação PMM). Logo, quando p-value for superior ao nível de significância (0,05), então não há diferença significativa entre as médias, caso contrário (p-value < 0,05) a hipótese nula (H_0) é aceita.

3 Resultados e Discussão

Após a seleção dos períodos sem falhas das estações testes, localizadas na BMSF, foram elaborados dois cenários com 5% e 15% de falhas para que seja possível comparar os dados reais e os dados imputados via PMM. Os resultados obtidos mostram que os valores do coeficiente r em todos os casos foram superiores a 0,80, tanto com o percentil de falhas de 5% quanto para o de 15%; isso mostra que a correlação é forte entre as séries observada e sintética via PMM (Tabela 3). Logo, o preenchimento de falhas de dados de chuva na região é satisfatório. No entanto, apenas essa correlação não garante a precisão do método.

Assim, também, foi realizado um cruzamento dos dados observados e sintéticos (Figuras 3 e 4). A partir desse cruzamento observa-se o alto grau de ajuste entre as séries, com R^2 superior a 0,7 (0,6) para o percentil de falhas de 5% (15%). Das estações testes o melhor ajuste para 5% (15%) de falhas foi observado na estação de Unai – MG com R^2 de 0,9347 (Carinhonha – BA com R^2 de 0,8741) e pior ajuste para a estação de Bom Jesus da Lapa – BA, R^2 de 0,7073, (Unai – MG, R^2 de 0,6735). Observa-se também que os ponto de cruzamento entre os

dados estão bem alinhados com a linha de tendência em vermelho, verificando que os dados estão bem ajustados.

Estações	Bom Jesus da Lapa - BA	Carinhanha BA	João Pinheiro MG	Remanso BA	Unai MG
5% de falhas					
r	0,84	0,93	0,95	0,94	0,97
p-value	0,677	0,879	0,867	0,967	0,750
EMA	0,40	0,28	0,33	0,24	0,25
EQM	4,48	3,12	3,32	2,94	2,41
15% de falhas					
r	0,85	0,93	0,83	0,85	0,82
p-value	0,538	0,471	0,824	0,027	0,743
EMA	0,56	0,50	0,99	0,55	1,05
EQM	7,47	9,00	11,49	8,6	9,55

Tabela 3 Coeficiente de Correlação (r); Teste de Wilcoxon (p-value); EMA (mm) e EQM (mm) via método PMM para percentual de falha de 5% e 15%.

Além disso, aplicou-se o teste de Wilcoxon para verificar com maior grau de confiabilidade a consistência da nova série sintética. No primeiro cenário, 5% de falhas, destacam-se as estações de Remanso (BA) e Carinhanha (BA), que tiveram os maiores valores de p-value, 0,967 e 0,879, respecti-

vamente. Já no segundo cenário, 15% de falhas, resalta-se que a estação de Remanso (BA) apresentou o menor p-value (0,027), o qual é inferior a 0,05, logo, foram observadas diferenças discrepantes entre as séries observada e sintética. A partir dos valores calculados desse teste, a diferença entre as médias das séries observadas e sintéticas não apresentaram diferenças significativas (Tabela 3), pois em 90% dos casos o coeficiente p-value foi superior ao nível de significância de 0,05.

Ademais, o EMA e o EQM foram baixos para as estações de testes tanto para o percentil de falha 5% (EMA < 0,5 mm e EQM < 5,0 mm) quanto para 15% (EMA < 1,5 mm e EQM < 12,0 mm). Nota-se também, que com o aumento no percentual de falhas o método PMM perdeu sua precisão, pois os valores de EMA e EQM são menores para o percentil de falhas de 5%, padrão esse também observado no teste de Wilcoxon. Nessa perspectiva, o preenchimento de falhas por tal método pode ser empregado sem prejuízos na análise da série. Entretanto, deve-se verificar o percentil de falhas, visto que quanto maior for esse valor, menor será a precisão do método.

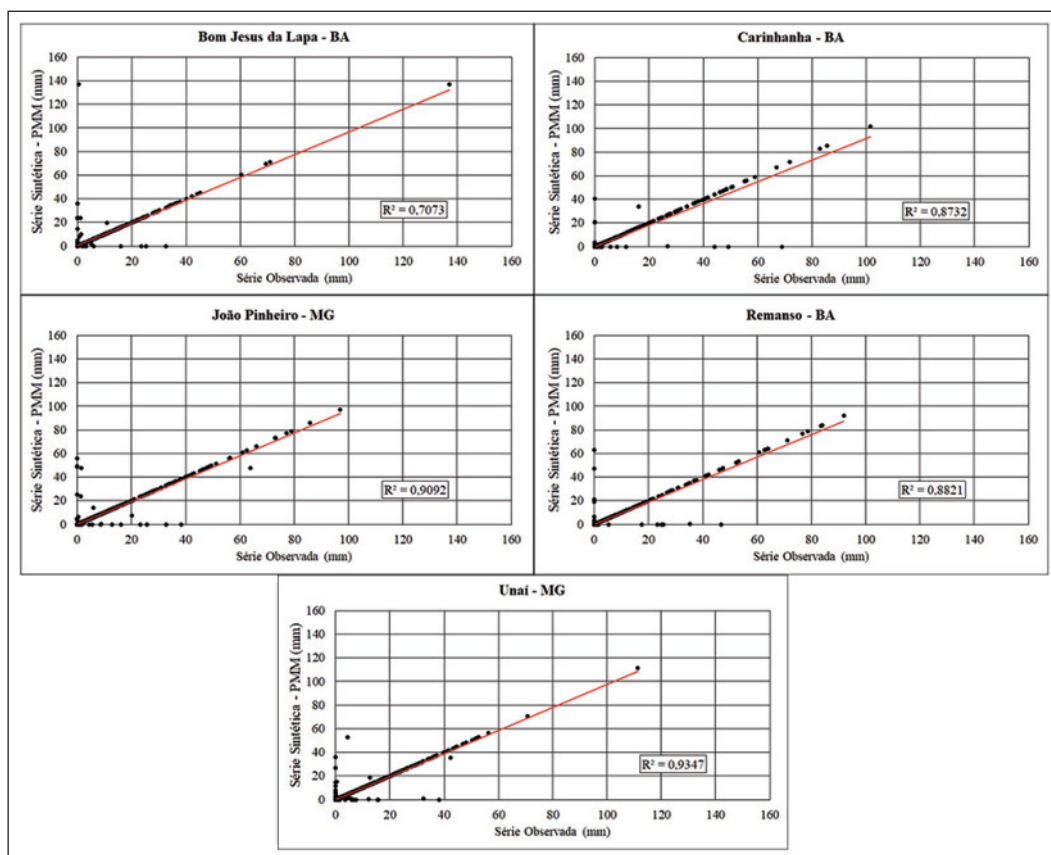


Figura 3 Diagrama de dispersão dos dados observados (mm) e imputados (mm) via PMM (5% de falhas) para as estações teste, para a BMSF.

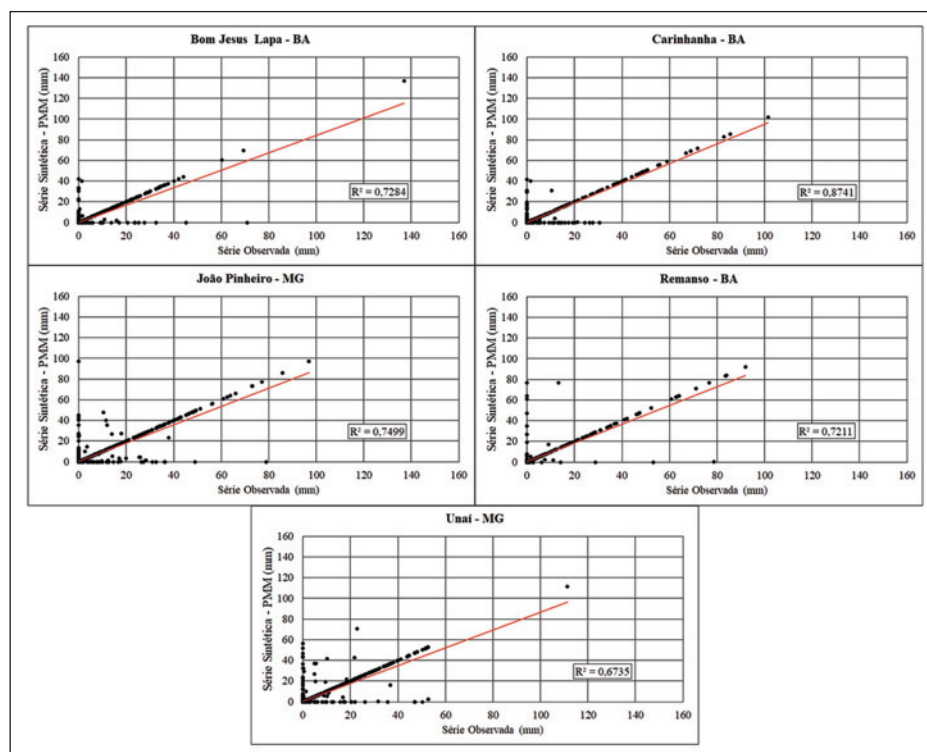


Figura 4 Diagrama de dispersão dos dados observados (mm) e imputados (mm) via PMM (15% de falhas) para as estações teste, para a BMSF.

4 Conclusões

Com a perspectiva de avaliar o preenchimento de falhas via imputação múltipla PMM para dados de chuva na BMSF, avaliou-se dois cenários, os quais o primeiro apresenta 5% de falhas e o segundo 15%. A confiabilidade do método adotado se baseia na alta correlação entre os dados medidos e estimados/imputados, $r > 0,8$, bem como o R^2 que foram superiores a 0,8 para 5% e 0,7 para 15% de falhas.

Ademais, aplicou-se o Teste de Wilcoxon para série não-paramétricas, onde o grau de confiabilidade/semelhança entre os dados observados e sintéticos foi alto (p -valor $> 0,05$), em 90% dos casos. Além disso, o EMA e EQM foram baixos para os diferentes percentis de falhas. Isso confirma que o método pode ser uma boa alternativa para o preenchimento de falhas em séries de dados.

Logo, conclui-se que o preenchimento de falhas é satisfatório para os objetivos do estudo e pode ser replicado em diferenças locais inseridas nos limites e áreas subjacentes da bacia. Para a aplicação em âmbito nacional do método PMM é necessário realizar mais validações em diferentes áreas do território

brasileiro com diferentes características de clima, vegetação e relevo. Além disso, a implementação de novas variáveis meteorológicas seria essencial para consolidar a aplicação do método para o preenchimento de dados faltantes.

5 Agradecimentos

O primeiro autor agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão de bolsa Mestrado processo 167798/2017-0. O segundo autor agradece pelo apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)/Modelagem.

6 Referências

- Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Gonçalves, J.L.M. & Sparovek, G. 2013. Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, 22(6): 711-728.
- Bier, A.A. & Ferraz, S.E.T. 2017. Comparação de Metodologias de Preenchimento de Falhas em Dados Meteorológicos para Estações no Sul do Brasil. *Revista Brasileira de Meteorologia*, 32(2): 215-226.
- Breiman, L. 2001. Random Forests. *Machine learning*, 45(1): 5-32.

Validação da Imputação Múltipla via Predictive Mean Matching para
Preenchimento de Falhas nos Dados Pluviométricos da Bacia do Médio São Francisco

Laurizio Emanuel Ribeiro Alves & Heliofábio Barros Gomes

- Burgette, L.F. & Reiter, J.P. 2010. Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9): 1070-1076.
- Camelo, H.N.; Lúcio, P.S. & Leal Júnior, J.V. 2017a. Modelagem da velocidade do vento usando metodologias ARIMA, Holt-Winters e RNA na previsão de geração eólica no Nordeste Brasileiro. *Revista Brasileira de Climatologia*, 21(1): 449-466.
- Camelo, H.N.; Lúcio, P.S.; Leal Júnior, J.B.V. & Carvalho, P. C.M. 2017b. Previsão de velocidade do vento no nordeste brasileiro através do modelo ARIMAX. *Revista Ciência e Tecnologia*, 20(36): 9-17.
- CBHSF. 2004. Companhia das Bacias Hidrográficas do São Francisco. *Plano de recursos hídricos da Bacia Hidrográfica do Rio São Francisco*. Disponível em: < http://www.saofrancisco.cbh.gov.br/_docs/planos/PlanoDecenaldeRecursosHidricos.pdf>. Acessado em: 29 mai. 2019.
- CBHSF. 2017. Companhia das Bacias Hidrográficas do São Francisco. *Afluentes do Médio São Francisco (SF9) – Minas Gerais*. Disponível em: < <http://cbhsaofrancisco.org.br/2017/comites-de-afluentes/cbh-afluentes-mineiros-do-medio-sao-francisco-minas-gerais/>>. Acessado em: 08 de ago. de 2018 e 12 de ago. 2019.
- Correa, S.M.B.B. 2003. *Probabilidade e Estatística*: 2. Ed. Belo Horizonte: PUC Minas Virtual, 116p.
- Cunha, A.P.M.A.; Brito, S.S.B.; Rossato, L.; Alvalá, R.C.S.; Carvalho, M.A.; Zeri, M.; Cunningham, C.; Maciel, A.P.R.; Andrade, E.S. & Vieira, R.M.S.P. 2017. Avaliação de indicador para o monitoramento dos impactos da seca em áreas de pastagens no semiárido do Brasil. *Revista Brasileira de Cartografia*, 69(1): 89-106.
- Diaz, C.C.F.; Pereira, J.A.S. & Nóbrega, R.S. 2018. Comparação de dados estimados por dois diferentes métodos para o preenchimento de falhas de precipitação pluvial na Bacia do Rio Pajeú, Pernambuco, Brasil. *Revista Brasileira de Climatologia*, 22(1): 324-339.
- Fix, E. & Hodges Jr., J.L. 1955. Significance probabilities of the Wilcoxon test. *The Annals of Mathematical Statistics*, 26(2): 301-312.
- González, A.R.; Kjaersgaard, J.; Trooien, T.; Reta-Sanchez, D.G.; Sánchez-Duarte, J.I.; Preciado-Rangel, P. & Fortis-Hernández, M. 2019. Comparison of Leaf Area Index, Surface Temperature, and Actual Evapotranspiration Estimate Using the METRIC Model and In Situ Measurements. *Sensors*, 19(8): 1-21
- Harrel Jr., F.E. 2014. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Switzerland, Springer, 581p.
- Heidt, K. 2019. *Comparison of Imputation Methods for Mixed Data Missing at Random*. Program Mathematic Science, East Tennessee State University, Dissertação de Mestrado, 79p.
- Heitjan, D.F. & Little, R.J.A. 1991. Multiple imputation for the fatal accident reporting system. *Applied Statistics*, 40(1): 13-29.
- Horton, N.J. & Lipsitz, S.R. 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3): 244-254.
- Lin, L.I.K. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1): 255-268.
- Lopo, A.B. 2014. *Modelagem Estocástica do Índice de Radiação Ultravioleta na Costa Leste do Nordeste do Brasil*. Programa de Pós-Graduação em Ciências Climáticas, Universidade Federal do Rio Grande do Norte, Tese de Doutorado, 112p.
- Marengo, J.A.; Cunha, A.P. & Alves, L.M.A. 2016. A seca de 2012-15 no semiárido do Nordeste do Brasil no contexto histórico. *Climanálise (online)*, 3(1): 49-54.
- Melo, Y.R.; Kohls, W.; Oliveira, T.M.N. 2017. Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas. *Boletim de Geografia* 35(1): 112-121.
- MMA. 2006. Ministério do Meio Ambiente. *Caderno da Região Hidrográfica do São Francisco*. Disponível em: <http://www.mma.gov.br/estruturas/161/_publicacao/161_publicacao03032011023538.pdf>. Acessado em: 27 mai. 2018.
- Moura, R.G.; Aravéquia, J.A. & Lopo, A.B. 2014. Preenchimento de falhas em dados de correlação de Anomalia da altura geopotencial (500hPa). *Ciência e Natura*, 36(Ed. Especial II): 503-509.
- Royston, P. 1992. Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, 2(3): 117-119.
- Rubin, D.B. 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1): 87-94.
- Segal, M.R. 1988. Regression Trees for Censored Data. *Biometrics*, 44(1): 35-47.
- Silva, B.H.L. 2018. *Comparação entre estimativas de altura de eucalipto realizadas pelo MOTI e pelo clinômetro HA-GLOF*. Departamento de Engenharia Florestal, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Trabalho de Conclusão de Curso em Engenharia Florestal, 31p.
- Silva, T.A.; Sanches, S.K.R. & Igarashi, D.C.C. 2019. Contingências passivas e teoria dos prospectos: análise dos efeitos em indicadores financeiros de empresas de construção civil e construção pesada. *Revista Evidenciação Contábil & Finanças*, 7(1): 39-57.
- Silva, W.L.; Xavier, L.N.R.; Maccira, M.E.P. & Rotunno, O.C. 2018. Climatological and hydrological patterns and verified trends in precipitation and streamflow in the basins of Brazilian hydroelectric plants. *Theoretical and Applied Climatology*, 137(1-2): 353-371.
- Tucci, C.E.M. 2001. *Hidrologia: ciência e aplicação*. Porto Alegre, UFRGS: ABRH. 944p.