Revista Brasileira de Climatologia
Brazilian Journal of Climatology

OPEN ACCESS

DOI: 10.5380/abclima

COMPARAÇÃO ENTRE MÉTODOS DE IMPUTAÇÃO DE DADOS EM DIFERENTES INTENSIDADES AMOSTRAIS NA SÉRIE DE PRECIPITAÇÃO PLUVIAL DA ESALQ

COMPARISON BETWEEN DATA IMPUTATION METHODS AT DIFFERENT SAMPLE INTENSITIES AT THE RAINFALL SERIES OF ESALQ

COMPARACIÓN ENTRE MÉTODOS DE INPUTACION DE DATOS EN DIFERENTES INTENSIDADES DE MUESTREO EN LA SERIE DE LLUVIAS DE LA ESALQ

Suelen Cristina Gasparetto
Universidade de São Paulo - ESALQ/USP
suelengasparetto@hotmail.com

Sônia Maria De Stefano Piedade
Universidade de São Paulo - ESALQ/USP
soniamsp@usp.br

Luiz Roberto Angelocci
Universidade de São Paulo - ESALQ/USP
lrangelo@usp.br

Vitor Augusto Ozaki
Universidade de São Paulo - ESALQ/USP
vitorozaki@usp.br

RESUMO: Um problema frequente nas análises estatísticas de informações climatológicas é a ocorrência de dados faltantes, assim, o objetivo deste trabalho foi comparar três métodos de imputação de dados com observações da série de precipitação pluvial de uma estação climatológica convencional, no município de Piracicaba-SP, no período de 1917 a 1997, em diferentes intensidades

amostrais (5%, 10% e 15%) de informações faltantes, geradas de forma aleatória. Para o "preenchimento" dessas informações, foram usados três métodos de imputação múltipla: PMM (*Predictive Mean Matching*), *random forest* e regressão linear, via método *bootstrap*, em cada intensidade amostral de informações faltantes. A comparação entre cada procedimento de imputação foi feita, por meio da raiz do erro quadrático médio, índice de acurácia de Willmott e o índice de desempenho. O método, que resultou em menores valores da raiz quadrada dos erros e maiores índices de desempenho e acurácia, foi o PMM, em especial, na intensidade de 10% de informações faltantes. O índice de desempenho, para os três métodos de imputação de dados, em todas as intensidades de observações faltantes, foi considerado insatisfatório, por isso, é necessária uma atenção maior quando se trata de observações tão variáveis espacialmente e temporalmente quanto as chuvas.

Palavras-chave: Precipitação; Imputação múltipla; MICE; Comparação de métodos.

ABSTRACT: A frequent problem in the statistical analysis of climatological information is the occurrence of missing data. The goal of this work was to compare three methods of imputing data with observations from the conventional serie in the ESALQ pluviometry weather station, from 1917 to 1997, in different sampling intensities (5%, 10% and 15%) of missing data, generated at random. For the "filling in" of the lost data, three multiple imputation methods were used: PMM (Predictive Mean Matching), random forest and linear regression via bootstrap, in each sample intensity of lost data. The methods were used by the MICE package (Multivariate Imputation by Chained Equations) in R. The comparison in each imputation process was made by the root of the mean square error, Willmot's precision index and performance index. The method that resulted in lower values in the square root of the errors and higher levels of accuracy and performance was the PMM, especially in 10% missing data The performance index for the three data imputation methods, in all dissipation intensities, missing was not considered one of the best, for this reason, greater care is needed when it comes to observations as spatially and temporally variable as is rainfall.

Keywords: Rainfall; Multiple imputation; MICE; Comparison of methods.

Resumen: Un problema frecuente en análisis estadísticas de información climatológica es la frecuente falta de datos, el objetivo de este trabajo fue comparar tres métodos de imputación de datos observando la serie de precipitaciones de una estación climatológica convencional, en el municipio de Piracicaba-SP, en el período de 1917 a 1997, con diferentes intensidades de muestreo (5%, 10% y 15%) de información faltante generada aleatoriamente. Para "completar" esta información, fueron utilizados tres métodos de imputación múltiple: PMM (Predictive Mean Matching), bosque aleatorio y regresión lineal, mediante el método bootstrap en cada intensidad de muestreo de la información faltante. Cada procedimiento de inoutación fué comaparado utilizando la raíz del error cuadrático medio, el índice de precisión de Willmott y el índice de rendimiento. El método, que resultó en valores más bajos de la raíz cuadrada de errores e índices más altos de desempeño y precisión, fue el PMM, especialmente a intensidad de 10% de información faltante. El índice de rendimiento, para los tres métodos de imputación de datos en todas las intensidades de las observaciones faltantes se consideró insatisfactorio, por lo que se necesita una mayor atención cuando se trata de observaciones que son tan variables espacial y temporalmente como la lluvia.

Palabras claves: Precipitación; Imputación múltiple; MICE; Comparación de métodos.

Submetido em: 15/10/2020
Aceito para publicação em: 20/11/2021
Publicado em: 22/11/2021

INTRODUÇÃO

A precipitação pluvial é uma variável meteorológica, com grande importância climática, tal que permite analisar os padrões de regime hídrico de uma região, definir locais que tenham características pluviais semelhantes e agrupá-los de forma conveniente em sub-regiões.

Juntamente com outras variáveis, como temperatura, a caracterização físico-hídrica do solo, possibilita caracterizar as épocas adequadas ao plantio e colheita para cada região (zoneamento de riscos climáticos), para determinar risco de ocorrência de doenças, pragas e as estratégias de controle (RICCE *et al.,* 2014; BALDISERA; DALLACORT, 2017). Algumas das principais aplicações na climatologia são a verificação de padrão e magnitude da precipitação (SOUZA *et al.,* 2017; SANTOS *et al.,* 2018).

Este trabalho analisa informações de precipitação da série climatológica da ESALQ – USP, no município de Piracicaba, São Paulo, atualmente com informações de um período superior a 100 anos. Em estudos climatológicos, é importante que se tenha informações de uma série histórica, pois só assim será possível observar tendências de curto e a longo prazo, podendo mostrar características de precipitação pluvial deste período, indicando, desse jeito, o potencial promissor para análises e aplicações futuras (SENA *et al.,* 2012; LEITÃO; CARVALHO, 2021).

Como as informações das variáveis meteorológicas variam de acordo com o lugar e o tempo em que foram medidas, não é difícil de encontrar um banco de dados climáticos incompletos, circunstância que prejudica diretamente as análises das observações, para resolver este problema é indicado o uso de métodos de imputação de dados (SOUZA *et al.,* 2017).

O procedimento mais comum é restringir-se à análise das informações com dados completos nas variáveis, desconsiderando a parte faltante. Contudo esse procedimento de exclusão de dados causa complicações para as análises, como estimativas tendenciosas, fazendo com que se possa chegar a conclusões enganosas (REBOITA *et al.,* 2015; VICENTE *et al.,* 2018).

Desde o século XX são desenvolvidos artifícios para a estimativa de valores faltantes em séries climatológicas, esses recursos acompanharam o avanço de novos métodos matemáticos e o maior poder computacional disponível.

As estratégias precursoras foram: o método da média, da interpolação linear e o das primeiras diferenças, bem como técnicas de geoestatística, como krigagem e cokrigagem ordinária, as quais são denominadas de imputação única, pois o método de preenchimento dessas falhas é feito apenas uma única vez. Porém mostraram diversas limitações, como falta de variabilidade no valor imputado, ou até mesmo problemas com variáveis pouco correlacionadas, produzindo assim estimativas de parâmetros tendenciosas e aumento no valor do erro padrão (GARCÍA-PEÑA; ARCINIEGAS-ALARCÓN; BARBIN, 2014; ALVES; GOMES, 2020).

Em pesquisas mais recentes, podem ser percebidas as atualizações das metodologias de preenchimento de dados faltantes, apresentadas por meio de processos realizados inúmeras vezes (BURHANUDDIN; DENI; RAMLI, 2017; CHEN *et al.,* 2019). Estudos mostram que esse tipo de metodologia é mais eficiente que os procedimentos de imputação única, reduzindo consideravelmente as desvantagens desses métodos (CARVALHO *et al.,* 2017). Para contornar os problemas dos métodos de imputação única, Rubin (1987) propôs o método da imputação múltipla, executado por processos iterativos, que consideram as diversas variáveis do banco de dado analisado. Essa técnica mostra-se muito eficiente para a imputação de séries temporais meteorológicas e estão altamente difundidas em estudos recentes (SANTOS *et al.,* 2020; SOUZA *et al.,* 2020). Este trabalho, portanto, tem por objetivo testar os métodos de imputação múltipla na variável precipitação pluvial em escala mensal.

Ademais, o estudo preocupa-se, não somente em analisar os erros, mas também investigar a acurácia e o desempenho dos modelos de imputação. Serão feitas as comparações, por meio da raiz do erro quadrático médio (REQM), além dos coeficientes de acurácia, Willmott e de desempenho.

As análises foram feitas por meio do pacote MICE do software R (R CORE TEAM, 2017). A metodologia MICE (*Multivariate Imputation by Chained Equation*) baseia-se na Cadeia de *Markov* Monte Carlo (MCMC- "*Markov chain Monte Carlo*"), cujo algoritmo é um amostrador de Gibbs, técnica de simulação Bayesiana, que amostra distribuições condicionais com a finalidade de obter amostras da distribuição conjunta (VAN BUUREN; GROOTHUIS-OUDSHOORN, 2011; ESPINOSA; PORTELA; RODRIGUES, 2019).

O estudo possui a seguinte estrutura: na seção 2, apresentam-se a metodologia e a descrição dos dados. Na seção 3, os resultados são apresentados e discutidos. Por fim, na seção 4, apresentam-se as considerações finais.

MATERIAIS E MÉTODOS

Dados coletados

Os dados foram obtidos no posto meteorológico da ESALQ - USP, situado no Campus "Luiz de Queiroz", na cidade de Piracicaba – SP, sendo atualmente composto por duas estações: uma convencional e outra automática (POSTO METEOROLÓGICO "PROFESSOR JESUS MARDEN DOS SANTOS" ESALQ - USP, 2020).

O elemento meteorológico selecionado para esta pesquisa foi a altura pluviométrica, em escala mensal, da estação convencional. O estudo ficou restrito ao período de 1917 a 1997. Depois desse período, os registros foram substituídos por informações de precipitação pluvial registrada na estação automática, cuja medição da altura pluviométrica é feita com outro aparelho denominado pluviômetro de báscula, que faz o registro da variável a cada 15 minutos, enquanto a estação convencional tem como aparelho de medição o pluviômetro de leitura direta, o qual é medido e registrado uma vez ao dia.

As variáveis do banco de dados são: mês, ano e precipitação pluvial em milímetros. Foi verificada a quantidade de informações não disponíveis (NA – *Not available)* desse período, que apresentaram apenas 14 observações, em nível de observação diária, um valor muito baixo para a execução dos métodos de imputação de dados, principalmente para os de imputação múltipla, que se validam por processos iterativos de valores ausentes com valores completos. Por esse motivo, essas 14 observações foram desconsideradas, uma vez que esse valor representa menos que 0,05% da quantidade de observações totais.

Apesar de se ter um número muito baixo de valores faltantes no banco de dados, esta pesquisa se justifica pelo grau de precisão que se pode ter dos valores imputados por meio dos modelos estudados, já que todos os valores que foram retirados de forma aleatória para este estudo eram conhecidos. Após o desenvolvimento dos métodos de imputações puderam ser considerados exatos, tendo, com isso, um resultado ainda mais preciso desses métodos.

O método de análise foi dividido em três diferentes estágios, para cada etapa é feito uma retirada de observações (5%, 10% e 15% respectivamente) e após cada análise, as informações subtraídas são repostas, e um novo procedimento é feito, porém, com outra quantidade de informações retiradas, ou seja, toda retirada é feita com reposição.

A classificação dos dados faltantes é MCAR – *Missing Completly At Random*, pois foram realizadas retiradas aleatórias apenas da variável precipitação, fato que permitiu gerar a substituição dos valores em falta a partir da distribuição dos dados ausentes condicionada aos dados observados.

Análise preliminar

Antes de ser feita a análise de imputação de dados, a série temporal foi submetida a uma verificação preliminar. Tal verificação se resume em observar o valor mínimo e máximo de precipitação pluvial, uma regressão linear simples e o cálculo da média da série, todos submetidos a série dividida em período anual. Já para um entendimento mais específico, foram calculadas as médias mensais de todo o período analisado.

O estudo utilizou-se apenas das variáveis mês, ano e precipitação pluvial, para desenvolver os métodos de imputação propostos a seguir. Essa escolha foi tomada pois as demais variáveis meteorológicas do banco de dados analisado também apresentavam falhas, no mesmo período que a variável precipitação. Salienta-se que, a variável precipitação utilizada, nos métodos de imputação, são valores médios mensais.

Imputação múltipla

O método de imputação múltipla, proposto por Rubin (1987), é um tema que está em grande expansão dentro do estudo de imputação de dados, sendo um algoritmo iterativo com diversas etapas.

(i) Imputação: essa etapa se inicia com o uso de banco com dados faltantes, substituindo-os por valores plausíveis, retirados de uma distribuição especialmente modelada para cada valor faltante. Esse procedimento cria um banco de dados, chamado de banco de dados completo. O processo é feito m vezes e, cada vez que são substituídos os valores em falta, usa-se o mesmo banco de dados faltantes, isto é, todos os bancos de dados completos dessa etapa serão gerados por meio de um único banco de dados.

(ii) Análise: essa segunda etapa visa estimar os parâmetros de interesse, como média, correlação, variância, etc, de cada banco de dados gerado pela etapa anterior,

separadamente. Cada uma dessas m estimações são feitas por meio de análises estatísticas tradicionais.

(iii) Agrupamento: após as conclusões das outras etapas, é feita então a etapa de agrupamento, que tem como objetivo agrupar os resultados, em uma estimativa pontual final, acrescida do desvio-padrão.

Finalmente, as m estimativas geradas, na etapa anterior, podem ser combinadas de uma maneira simples, como proposto por Rubin (1987).

A cada m estimativas geradas na etapa da análise, é obtido um parâmetro Q de interesse, ou seja, $\widehat{Q}_\iota$ tal que i = 1, ..., m sendo a estimativa do i-ésimo parâmetro considerado. A estimativa da média geral do parâmetro de interesse ($\bar{Q}$) será a média das estimativas individuais, assim, essa estimativa é calculada da seguinte forma:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \widehat{Q}_\iota \quad . \tag{1}$$

Para encontrar a média das m variâncias do conjunto dos dados imputados, é necessário fazer o seguinte cálculo:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i \quad , \tag{2}$$

em que U_i é a variância do m-ésimo conjunto de dados imputados.

E a variância entre as imputações:

$$B = \frac{1}{m-1}\sum_{i=1}^{m} \left(\widehat{Q}_\iota - \bar{Q}\right)^2 \quad . \tag{3}$$

Logo a variância combinada será:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \quad , \tag{4}$$

em que $\left(1 + \frac{1}{m}\right)$ é a correção dos números infinitos de imputações.

Em seguida, podem-se realizar testes de hipótese e intervalos de confiança para a média ($\bar{Q}$) por meio de uma distribuição *t-Student*, ou seja:

$$\frac{(\bar{Q} - Q)}{SE} \sim t_{\nu_m} \quad ,$$

em que SE é o valor do desvio-padrão combinado, calculado por meio da raiz da variância combinada com $v_m = (m-1)\left[\frac{1-\bar{U}}{(1-m)^{-1}B}\right]^2$ graus de liberdade, sendo Q o valor da média real da variável de estudo (HUI *et al.,* 2004).

Imputação Multivariada por Equações Encadeadas (MICE)

O método MICE trata-se de um algoritmo que se inicia com um sorteio aleatório dos dados observados e imputa os dados incompletos, variável por variável, em seguida ocorre uma iteração que consiste em um ciclo de todos os elementos da variável de interesse (VAN BUUREN, 2012).

O número de iterações T geralmente pode ser baixo, supondo 5 ou 10, e o algoritmo gera várias imputações executando os passos descritos, a seguir, m vezes de forma paralela (RUBIN, 1987).

Considere-se Y uma matriz de dados coletados com m linhas, as quais representam a quantidade de observações, n colunas, que representam as variáveis com $y_{ij} = (y_{i1}, \dots, y_{in})$, em que y_{ij} é o valor da variável j para a observação i. Pode-se separar Y em dois subconjuntos $Y = (Y^{obs}, Y^{mis})$, em que Y^{obs} são os dados observados (não faltantes) e Y^{mis} são os dados faltantes.

Seja Y_j com j = 1, ..., p, uma das variáveis p incompletas, em que $Y = (Y_1, \dots, Y_p)$. As partes observadas e ausentes de Y_j são denotadas por Y_j^{obs} e Y_j^{mis}, respectivamente, então, $Y^{obs} = (Y_1^{obs}, \dots, Y_p^{obs})$ e $Y^{mis} = (Y_1^{mis}, \dots, Y_p^{mis})$ representam os dados observados e ausentes em Y.

Seja $Y_{-j} = (Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_p)$, o conjunto de variáveis $p-1$ em Y, exceto Y_j. Os conjuntos de dados imputados são denotados como $Y^{(h)}$ em que h = 0,..., m. Seja Q a quantidade de interesse científico (por exemplo, um coeficiente de regressão). Na prática, Q é geralmente um vetor multivariado.

O MICE segue os seguintes passos:

1. Especificar um modelo de imputação $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$, para a variável Y_j, e R é uma matriz indicadora com a mesma dimensão de Y. Seus elementos são descritos da seguinte forma:

$$r_{ij} = \begin{cases} 1, se\ y_{ij}\ \text{é}\ observado \\ 0, se\ y_{ij}\ \text{é}\ faltante \end{cases}$$

tal matriz é a indicadora de resposta da variável Y_j , com j = 1, ..., p.

2. Para cada j, começa-se o preenchimento das imputações Y_j^0, por retiradas aleatórias de Y_j^{obs}.

3. Repete-se para t = 1, ..., T, em que T é a quantidade de iterações.

4. Repete-se para j = 1, ..., p, no qual j é a quantidade de variáveis incompletas.

5. Define-se $Y_{-j}^t = \left(Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1}\right)$ como os dados atualmente completos, exceto Y_j.

6. Retira-se $\phi_j^t \sim P(\phi_j^t | Y_j^{obs}, Y_{-j}^t, R)$, em que ϕ_j^t é o parâmetro do modelo de dados faltantes da j-ésima variável incompleta na t-ésima iteração.

7. Retira-se imputações $Y_j^t \sim P\left(Y_j^{mis} | Y_j^{obs}, Y_{-j}^t, R, \phi_j^t\right)$.

8. Finaliza-se a repetição j.

9. Finaliza-se a repetição t.

Os métodos de imputação múltipla escolhidos para este estudo são descritos a seguir e fazem parte dos passos 1 e 2 do algoritmo MICE, descrito anteriormente.

PMM (*Predictive Mean Matching*)

O PMM é um método de imputação múltipla, desenvolvido por Little (1988), tendo como objetivo imputar valores ausentes, por meio do método vizinho mais próximo, com a distância baseada nos valores esperados das variáveis ausentes, condicional às covariáveis observadas, combinando elementos de regressão linear.

Random forest

Random forest é um método de classificação e regressão que usa uma combinação de árvores de regressão para a predição de valores. Árvore de regressão é um diagrama de possíveis resultados escolhidos previamente, em que cada resultado está relacionado com os demais.

Esse diagrama é um processo hierárquico, que começa do maior nível e se divide em menores níveis de possíveis resultados. No *random forest*, cada árvore depende dos valores

de um vetor aleatório, amostrado de forma independente e com a mesma distribuição para todas as árvores de regressão (BREIMAN, 2001).

Regressão linear via método bootstrap

Esse artifício é um processo de imputação múltipla que ajusta um modelo de regressão linear, por meio do processo de *bootstrap* não paramétrico, que é um procedimento de reamostragem e consiste na criação de um novo banco de dados, do mesmo tamanho do original, a partir de um sorteio com reposição de seus elementos (EFRON; TIBSHINRANI, 1993).

Para cada etapa dos métodos de imputação propostos acima, foram feitas 100 repetições. Os resultados que serão apresentados são consequência de médias aritméticas dessas simulações.

Comparação dos métodos de imputação

Para verificar a eficiência dos métodos de imputação, foram utilizados os seguintes indicadores de comparação: raiz do erro quadrático médio (REQM), índice de acurácia de Willmott (d) e o índice de desempenho (c), o qual utiliza, em seu cálculo, o coeficiente de correlação de Pearson (r).

É importante ressaltar que este índice, r, não será analisado de forma isolada, como os outros procedimentos citados, pois, em sua essência, esse valor mede a relação linear entre as observações do modelo de imputação e os valores reais.

Levar esse índice em consideração de forma isolada pode causar interpretações demasiadamente rigorosas, levando a conclusões infundadas do ponto de vista prático, mas, ainda assim, são indicadores comumente utilizados para avaliar o desempenho de métodos de imputação como os testados neste estudo.

REQM (Raiz do Erro Quadrático médio)

A raiz do erro quadrático médio é utilizada como uma medida do erro de previsões. Segundo Willmott *et al.* (1985), as medidas de diferença ou erro, especialmente a REQM, são mais usadas na comparação e avaliação de modelos de simulação.

Para alguns casos, essa medida substitui os índices baseados em correlação e habilidade, como as medidas primordiais de precisão, fato importante, porque as medidas fundamentadas em correlação e habilidade não estão consistentemente relacionadas à precisão do modelo, sugerindo que a aplicação dos índices de diferença é efetiva a uma ampla variedade de modelos e dados geofísicos. Essa medida é definida como:

$$REQM = \sqrt{\frac{\sum_{j=1}^{n}(p_j - o_j)^2}{n}} \quad , \tag{5}$$

em que o_j é o valor observado, p_j é o valor estimado pelo modelo analisado e n é a quantidade de observações da amostra.

Índice de Acurácia de Willmott (d)

O índice de exatidão ou de acurácia de Willmott (d) foi desenvolvido por Cort J. Willmott na década de 80 (WILLMOTT *et al.,* 1985). Esse índice (d) pode variar de 0 a 1, sendo d = 0, uma total discordância entre os valores observados e preditos, enquanto d = 1 indica uma perfeita concordância ou exatidão entre os valores. Esse valor é definido pela seguinte equação:

$$d = 1 - \frac{\sum_{j=1}^{n}|d_j|^2}{\sum_{j=1}^{n}(|p_j - \bar{o}| + |o_j - \bar{o}|)^2} \quad , \tag{6}$$

em que $\bar{o}$ é a média dos valores observados e $d_j = p_i - o_i$.

Coeficiente de correlação de Pearson (r)

O coeficiente de correlação de Pearson (r) é um número que permite avaliar quanto as duas variáveis são correlacionadas, ou seja, mede a relação linear entre as variáveis em análise. É uma medida adimensional, e seu valor varia de -1 a 1, sendo -1 uma correlação negativa, ou seja, quanto mais uma variável aumenta mais a outra diminui.

Quando r = 0, as duas variáveis não são dependentes linearmente uma da outra, porém, quando esse valor é igual a 1, diz-se que ocorre uma correlação positiva, logo uma variável é proporcionalmente relacionada uma com a outra.

Índice de desempenho (c)

O índice de desempenho (c) de um modelo foi desenvolvido por Camargo e Sentelhas (1997), com o intuito de relacionar dois critérios de comparação, o coeficiente de correlação de Pearson (r) e o índice de exatidão de Willmott (d), pois (r) indica quanto as observações preditas pelo modelo estão dispersas, em relação à média e (d) refere-se ao afastamento dos valores preditos em relação aos valores observados. Assim a confiança ou desempenho do modelo reúne esses dois coeficientes da seguinte forma:

$$c = r \times d \quad . \tag{7}$$

Para interpretar esse índice são usados os valores da Tabela 1.

Tabela 1 - Critério de interpretação do desempenho dos métodos pelo índice (c), de Camargo e Sentelhas (1997).

Valor de c	Desempenho
> 0,85	Ótimo
0,76 a 0,85	Muito bom
0,66 a 0,75	Bom
0,61 a 0,65	Mediano
0,51 a 0,60	Sofrível
0,41 a 0,50	Mau
≤ 0,40	Péssimo

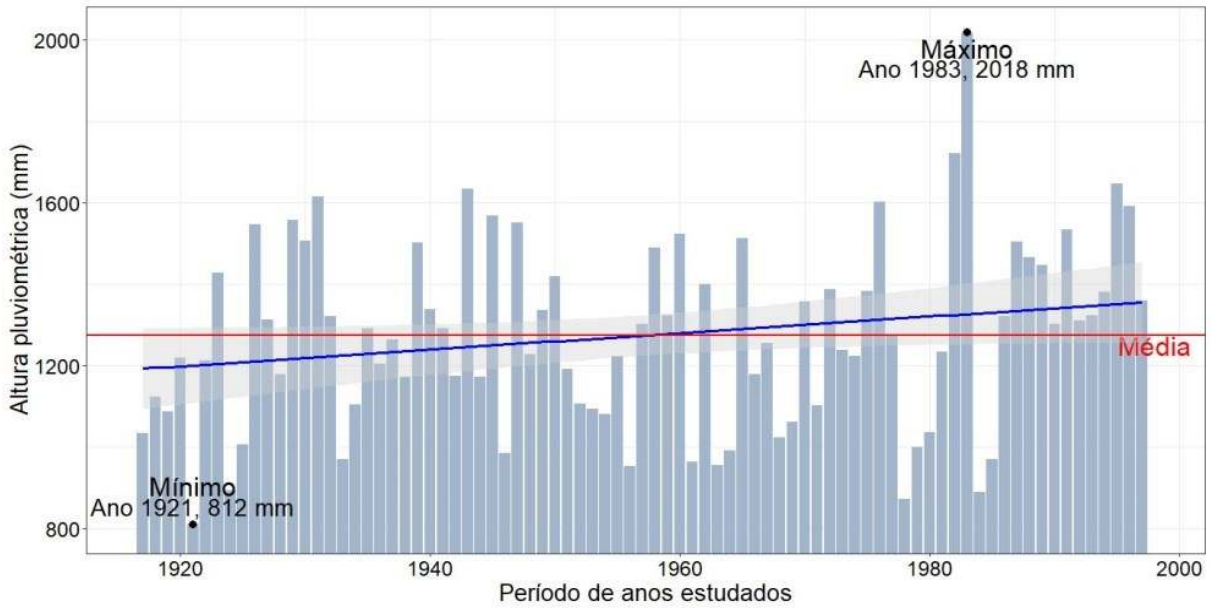
Fonte: Camargos e Sentelhas (1997)

RESULTADOS E DISCUSSÃO

Análise descritiva preliminar

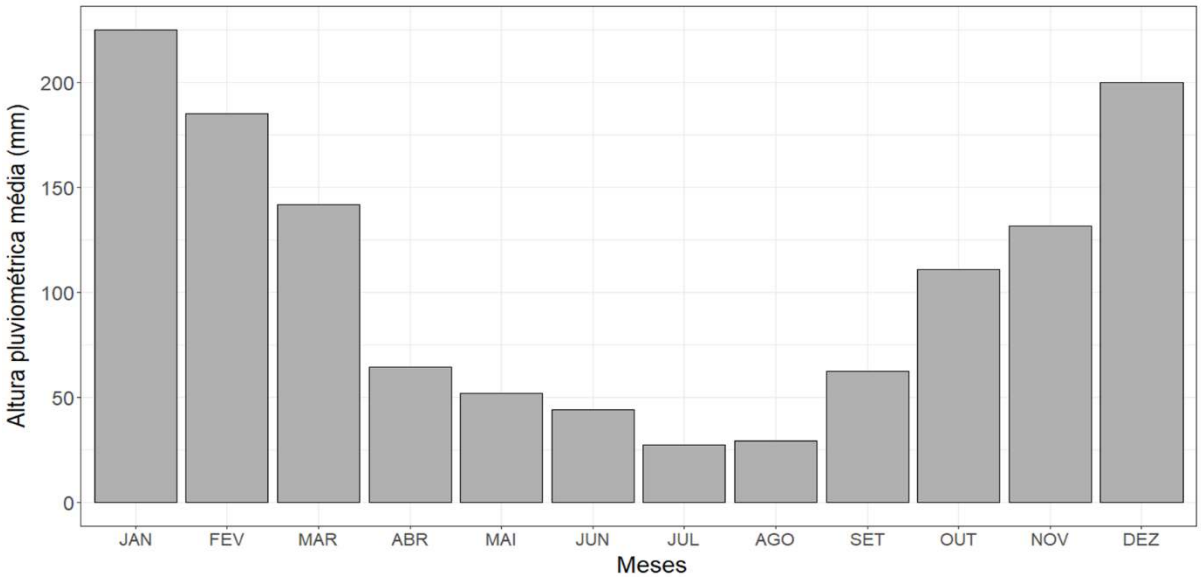
A figura, a seguir, contém os resultados obtidos por meio dos cálculos de somas anuais das precipitações. O valor médio das precipitações anuais é de 1277,4 mm, o valor mínimo é de 812 mm, no ano de 1921 e o valor máximo de 2018 mm no ano de 1983. Uma reta resultante de uma regressão linear simples pode ser observada (em azul), descrita por y = -2703,159 + 2,032x, tal que y representa a variável soma de precipitação anual e x os anos do período estudado. Essa reta auxilia a verificação de que existe uma tendência crescente nas informações dessa série, resultados que concordam com outros estudos das regiões Sul e Sudeste do Brasil (RAIMUNDO; SANSIGOLO; MOLION, 2014; GUEDES; PRIEBE; MANKE, 2019).

Figura 1 - Alturas pluviométricas anuais da estação meteorológica da ESALQ no período de 1917 a 1997.



Fonte: Elaborado pelos autores (2020)

Figura 2 - Variação da altura pluviométrica média mensal da estação meteorológica da ESALQ no período de 1917 a 1997.



Fonte: Elaborado pelos autores (2020)

A Figura 2 mostra os valores médios mensais de altura pluviométrica, observados no posto meteorológico, com distribuição típica dos regimes pluviométricos do estado de São Paulo, ao longo do ano, com valor médio máximo de 225 mm, em janeiro e mínimo de 27 mm, em julho. As estações com maiores volumes de precipitação são: primavera e, principalmente,

o verão, enquanto o outono e o inverno apontam uma queda nesses valores, o que é totalmente concordante com outros estudos de regime hídrico dos estados do Sudeste brasileiro (CONSTANTINO; BRUNINI, 2007; JARDIM; MOURA, 2018).

Imputação de dados

As técnicas de imputação foram avaliadas em diferentes percentuais de dados faltantes (5%, 10% e 15%). Apenas para efeito de organização, os resultados e discussões das análises serão separadas, em diferentes intensidades amostrais de dados faltantes e cada um foi nomeado como “Amostra” seguido da sua intensidade de dados retirados.

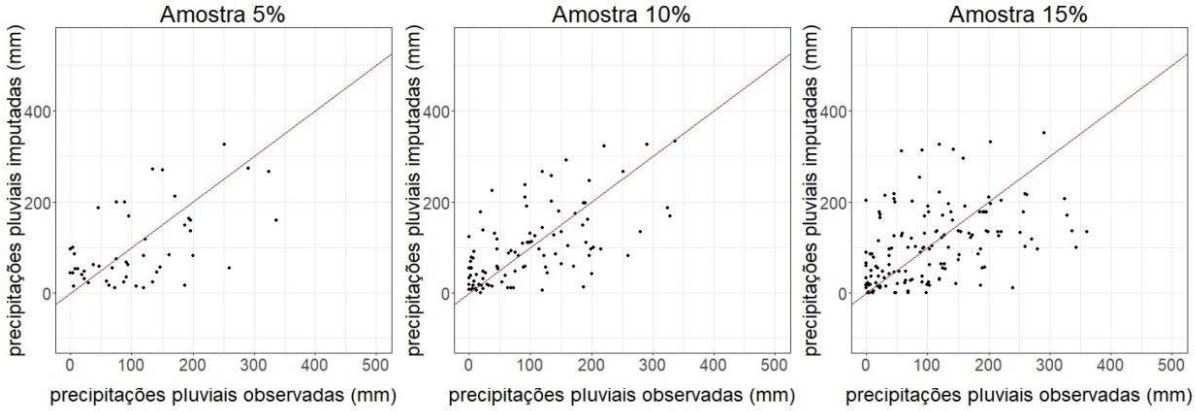
É importante lembrar que o estudo se utiliza apenas das variáveis mês e ano, para desenvolver os métodos de imputação propostos neste trabalho. Essa escolha foi tomada, uma vez que as demais variáveis meteorológicas do banco de dados analisado também apresentavam falhas, no mesmo período que a variável precipitação, ou seja, todos os resultados apresentados, a seguir, são valores médios mensais. Como o estudo se propôs a comparar os métodos de imputação da forma mais próxima da realidade possível, foram excluídas variáveis que também contassem com a presença de dados faltantes.

PMM – Predictive Mean Matching

A Figura 3 contém os gráficos de dispersão da variável precipitação pluvial imputada pelo método PMM versus precipitação pluvial observada, com 49 (5%), 97 (10%) e 146 (15%) de observações faltantes, respectivamente, ambas em milímetros (mm) e escala mensal.

A linha diagonal representa a reta y=x, ou seja, reta que dimensiona quanto à precipitação imputada é igual à precipitação observada. Pode-se verificar que para este método as informações imputadas apresentam uma tendência de dispersão, em torno da reta y=x. Tal comportamento se mantém para os três cenários de informações faltantes apresentados neste estudo.

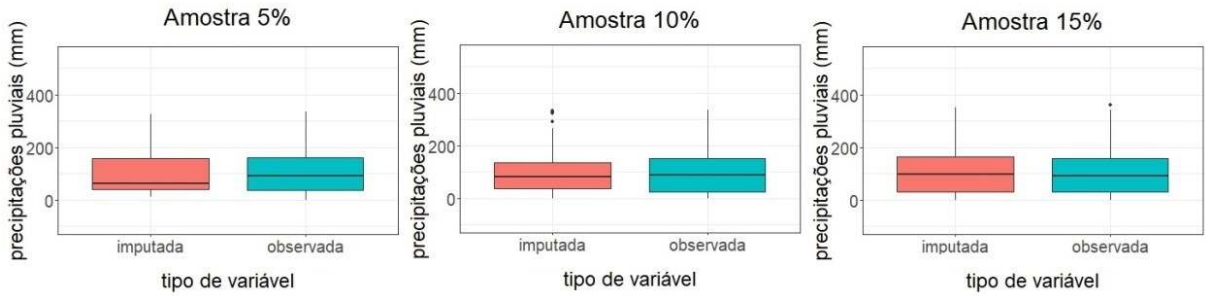
Figura 3 - Gráficos de dispersão da variável precipitação pluvial mensal imputada pelo método PMM versus observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%”, respectivamente.



Fonte: Elaborado pelos autores (2020)

A Figura 4 contém as diferenças entre as distribuições da variável precipitação pluvial imputada pelo método PMM e a observada, por meio dos gráficos de caixa (*boxplot*) das respectivas variáveis.

Figura 4 - Gráficos de caixa (*boxplot*) da variável precipitação pluvial mensal imputada pelo método PMM e a variável observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%”, respectivamente.



Fonte: Elaborado pelos autores (2020)

Pode-se notar que, para as três intensidades amostrais retiradas, tanto o primeiro quanto o terceiro quartil estão com valores muito próximos de uma variável para outra, ou seja, o intervalo interquartílico é muito semelhante entre as variáveis.

O mesmo acontece com a mediana, porém, para o cenário de 5% de falta, essa estatística da variável imputada é menor que a da variável observada, o que implica afirmar que a variável imputada se concentrou mais entre os valores 0 a 100 mm, diferentemente da

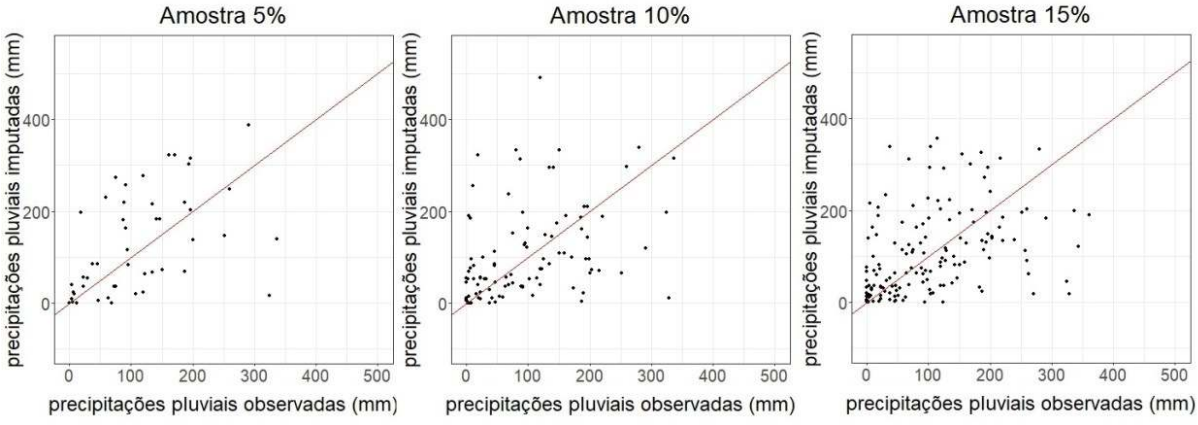
variável observada, que concentrou aproximadamente metade das suas observações nesse intervalo.

Random forest

Para o método *random forest,* é apresentada a seguinte figura (Figura 5), na qual se pode verificar que, para esse método, as informações imputadas apresentaram um comportamento semelhante ao que foi apresentado no método PMM. Contudo o método analisado nesta seção mostra valores bem discrepantes (fora da nuvem de pontos) que não foram notados no mecanismo apresentado anteriormente.

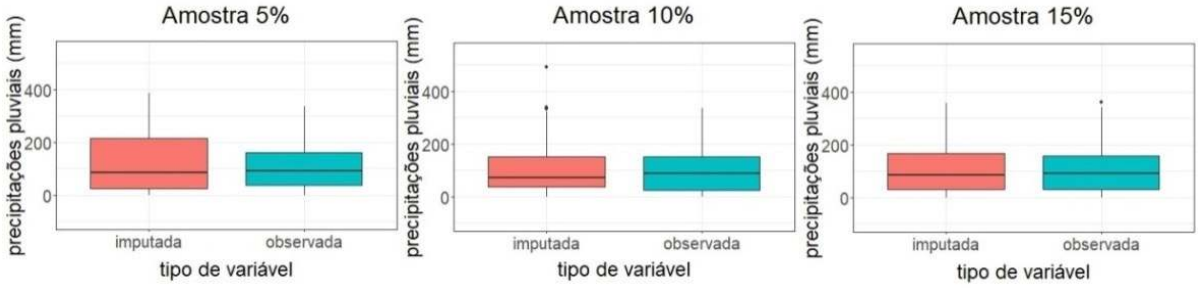
No cenário com 5% dos dados faltantes, um ponto foi imputado com o valor muito próximo de zero, mas o seu valor observado foi de aproximadamente 300 mm; outro ponto intrigante está no cenário de 10% de dados faltantes, no qual observou-se um registro com valor próximo de 150 mm, mas o *random forest* imputou valor muito próximo a 500mm, bem como nesse mesmo cenário, um ponto teve seu valor imputado muito próximo a zero enquanto seu valor observado foi de 350 mm.

Figura 5 - Gráficos de dispersão da variável precipitação pluvial mensal imputada pelo método *random forest* versus observada, ambas em milímetros (mm), para os bancos de dados "Amostra 5%", "Amostra 10%" e "Amostra 15%", respectivamente.



Fonte: Elaborado pelos autores (2020)

Figura 6 - Gráficos de caixa (*boxplot*) da variável precipitação pluvial mensal imputada pelo método *random forest* e a variável observada, ambas em milímetros (mm), para os bancos de dados "Amostra 5%", "Amostra 10%" e "Amostra 15%", respectivamente.



Fonte: Elaborado pelos autores (2020)

Nos gráficos *boxplots* (Figura 6), pode-se notar que, para o caso de 5% de observações faltantes ("Amostra 5%"), o intervalo interquartil da variável imputada é maior que a variável precipitação observada. Esse comportamento se repete para o banco de dados "Amostra 15%", porém em menor proporção. Já para o banco de dados "Amostra 10%", o intervalo interquartil da variável observada é menor que a variável imputada, ou seja, para esse cenário a variabilidade das informações observadas é maior que a da variável imputada e, como foi notado no gráfico de dispersão, no cenário de 10% de dados faltantes, a informação imputada com o valor de 500 mm se mostrou como um possível *outlier*, bem como os pontos imputados com valor maior de 250 mm.

Segundo Oliveira, Oliveira e Monteiro (2017), o método *random forest* exibe bons ajustes de valores imputados, em relação aos observados, mostrando grande potencial de uso para a imputação de dados ausentes ou na geração de séries temporais de dados meteorológicos em quadrículas, regiões subdivididas em formatos de grade retangular, ou regiões sem a presença de dados medidos. Tal estudo, entretanto, se limita à precipitação pluvial em escala diária, escala desconsiderada neste estudo, pois, além de não ser tão expressiva na prática, pode causar sérias dificuldades para a análise da série (RUIZ-CÁRDENAS; KRAINSKI, 2011).

Para este estudo o método *random forest* não apresentou uma boa aproximação das informações imputadas com as observadas, apesar de ser um método bastante eficiente, para a geração de dados meteorológicos. As variáveis usadas como preditoras, nos modelos de imputação (mês e ano), não são fortemente correlacionadas com a variável precipitação,

motivo plausível para a falta de maior eficácia do método, como exposto no trabalho de Mital et al. (2020).

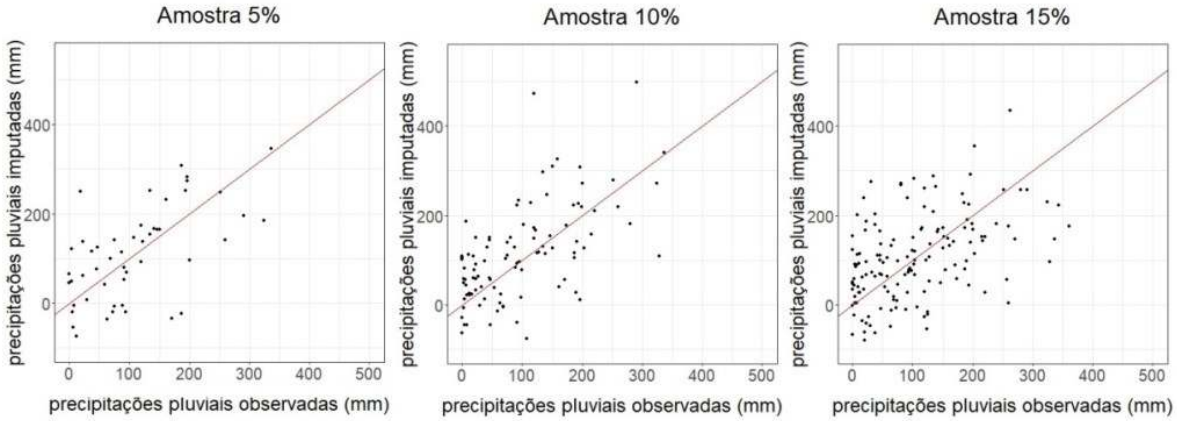
Regressão linear via método *bootstrap*

A Figura 7 mostra os gráficos de dispersão da variável precipitação pluvial imputada pelo método regressão linear via método *bootstrap* versus precipitação pluvial observada. Embora esse método tenha imputado valores negativos, para essa variável, nas três intensidades amostrais de informações retiradas, valores obviamente discrepantes com a realidade, pontos intrigantes também foram notados nesse método.

Valores imputados de forma negativa, como no cenário de 10% de dados ausentes, são claramente pontos fora da realidade para essa variável, além disso, dois valores se sobressaem, uma observação que teve o seu valor imputado por volta de 500 mm enquanto seu valor real era próximo a 150 mm; e outro ponto imputado pelo valor de 500 mm, porém seu valor original é de aproximadamente 300 mm, já no ambiente com 15% dos dados faltantes, pode ser notada a observação com seu valor original de 250 mm aproximadamente e teve um valor imputado de pouco menos de 450 mm.

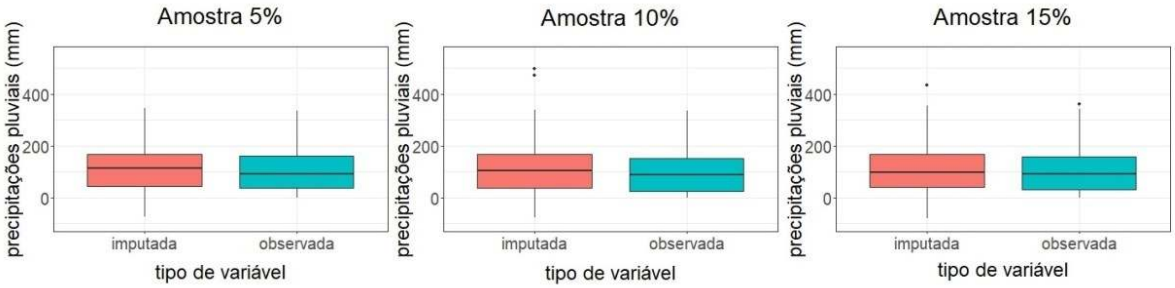
Mesmo com todos esses apontamentos, o método sugere uma concentração maior, ao redor da reta y=x, que os métodos já discutidos, ou seja, para as precipitações imputadas maiores ou iguais a zero, esse método presume assumir valores mais próximos aos valores observados, pressuposição que será efetivamente válida após a comparação entre os métodos de imputação.

Figura 7 - Gráficos de dispersão da variável precipitação pluvial mensal imputada pelo método de regressão linear via método *bootstrap* versus observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%”, respectivamente.



Fonte: Elaborado pelos autores (2020)

Figura 8 - Gráficos de caixa (*boxplot*) da variável precipitação pluvial mensal imputada pelo método regressão linear via método *bootstrap* e a variável observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%", respectivamente.



Fonte: Elaborado pelos autores (2020)

Nota-se que, para os cenários de intensidades amostrais 10% e 15% de informações faltantes (Figura 8), tanto o primeiro quartil quanto a mediana tiveram valores próximos entre as variáveis, assim, 50% das informações estão concentradas em intervalos semelhantes, além de mostrar os pontos discrepantes, que já foram identificados, ao observar os gráficos de dispersão. As duas observações imputadas com valores superiores a 400 mm caracterizam, de fato, possíveis *outliers*, o mesmo acontece com a observação contida no cenário de 15% de dados faltantes imputada com valor aproximado de 450 mm.

Comparação dos métodos de imputação

Para verificar a eficiência dos métodos propostos na estimação dos valores faltantes, foram obtidos valores das seguintes medidas de diferença: REQM, d e c, respectivamente. As

avaliações dos métodos de imputação foram feitas nas três intensidades amostrais de dados faltantes.

REQM – Raiz do erro quadrático médio

Na Tabela 2, percebe-se que o método de imputação PMM indicou os melhores resultados, ou seja, os menores valores de erro entre os três métodos analisados e, para o cenário de 10% de dados faltantes, o valor foi ainda menor. Segundo Alves e Gomes (2020), o método da Imputação Múltipla via PMM (*Predictive Mean Matching*) se mostrou satisfatório, para os objetivos de preenchimento de falhas nos dados pluviométricos, corroborando com os resultados desta pesquisa. Neste estudo, também, conseguem ser notados resultados ainda mais promissores para o cenário de 10% de dados faltantes.

O algoritmo que resultou em valores mais distantes de 0 foi o *random forest*. Para esse indicador de comparação, foi o que apresentou erros maiores, logo foi o método menos adequado para este estudo.

Tabela 2 - Valores da raiz do erro quadrático médio da variável precipitação pluvial imputada, por meio dos métodos PMM, *random forest* e regressão linear via método *bootstrap,* nas porcentagens 5%, 10% e 15% de dados faltantes.

	Porcentagem de dados faltantes		
Métodos	5%	10%	15%
PMM	88,1	87,2	88,7
Random forest	104,2	103,2	104,3
Regressão linear via método *bootstrap*	88,8	104,3	90,0

Fonte: Elaborado pelos autores (2020)

Índice de Acurácia de Willmott (d)

A tabela, a seguir (Tabela 3), contém os resultados do índice de acurácia de Willmott (d) dos três métodos de imputação e seus respectivos valores faltantes. É importante ressaltar que quanto melhor o método de imputação mais o valor desse indicador de comparação é próximo de 1.

Segundo Silva *et al.* (2019), o índice de acurácia de Willmott analisado em imputações múltiplas de dados de radiação solar via algoritmo MICE, utilizando métodos de médias

preditivas ponderadas, apresentou valores superiores a 0,9, vale destacar que tal metodologia é um caso específico da metodologia do PMM, circunstância concordante com os resultados apresentados a seguir, pois percebe-se que o método PMM apresentou melhores índices, tendo como o valor mais próximo de 1 o valor 0,73 no cenário de 10% de dados faltantes e, para o cenário de 15%, o método PMM e de regressão resultaram no mesmo valor, 0,71.

Assim como no REQM, os três métodos de imputação tiveram os mesmos comportamentos entre si, tiveram um desempenho menor no cenário de 5% de dados faltantes, enquanto para o de 10%, tiveram os melhores valores e, para 15%, tiveram uma pequena queda, com exceção do método *random forest* o qual permaneceu com o mesmo valor do índice de 10% (0,62).

Tabela 3 - Valores do índice de acurácia Willmott (d) da variável precipitação pluvial imputada, por meio dos métodos PMM, *random forest* e regressão linear via método *bootstrap,* nas porcentagens 5%, 10% e 15% de dados faltantes.

	Porcentagem de dados faltantes		
Métodos	5%	10%	15%
PMM	0,71	0,73	0,71
Random forest	0,61	0,62	0,62
Regressão linear via método *bootstrap*	0,70	0,72	0,71

Fonte: Elaborado pelos autores (2020)

Índice de desempenho (c)

De acordo com a classificação estabelecida para este índice (índice de desempenho), mostrado na Tabela 1, todos os métodos de imputação, em todas as porcentagens de dados faltantes, classificaram-se como desempenho “Péssimo”, exceto, o método PMM no cenário de 10% de dados faltantes, que se classificou como “Mau”, resultados que são possíveis constatar ao observar a Tabela 4.

Mesmo que os valores do índice de exatidão tenham se revelado altos, o coeficiente de correlação de Pearson indicou o contrário, valores muito baixos. Ressalta-se que o único cenário que se classificou diferente do pior índice de desempenho, método PMM com 10% de observações faltantes, o qual teve classificação “Mau”, ainda sim, não é considerado um resultado satisfatório, visto que ele se encontra no extremo do limite inferior desse intervalo

de classificação. Portanto pode-se aplicar esses métodos, mas com ressalvas em relação à dependência linear entre a variável imputada e a observada.

Tabela 4 - Valores do índice de desempenho (c) da precipitação pluvial imputada, por meio dos métodos PMM, *random forest* e regressão linear via método *bootstrap,* nas porcentagens 5%, 10% e 15% de dados faltantes.

	Porcentagem de dados faltantes		
Métodos	5%	10%	15%
PMM	0,39	0,41	0,38
Random forest	0,24	0,25	0,24
Regressão linear via método *bootstrap*	0,37	0,39	0,37

Fonte: Elaborado pelos autores (2020)

CONSIDERAÇÕES FINAIS

Ao analisar os três indicadores de comparação de forma geral, pode-se afirmar que o PMM foi o melhor, uma vez que esse método de imputação resultou em menores erros quadráticos médios e maiores índices de acurácia e desempenho. Pode ser destacado positivamente o cenário de 10% de informações faltantes, o qual teve os melhores resultados entre os métodos de imputação de dados.

Todos os demais métodos de imputação, em todas as intensidades, tiveram índices de desempenho baixos, classificados como “Péssimo”. Pode-se atrelar esse fato à alta variabilidade temporal e espacial da variável em estudo, além de se tratar de eventos sequencialmente independentes, afinal, cada observação é altura média de precipitação pluvial correspondente a cada mês do período estudado estes aspectos interferem de fato no desempenho dos métodos de imputação, sendo que eles são desenvolvidos com base nas observações já existentes no banco de dados.

Para este estudo, utilizaram-se informações apenas de uma estação meteorológica. Possivelmente, ao associar os dados de outras estações próximas à estação estudada, os resultados poderiam ter sido melhores.

Algumas pesquisas discutem outras estatísticas para a avaliação dos erros das metodologias de imputação (GARCÍA-PEÑA; ARCINIEGAS-ALARCÓN; BARBIN, 2014; CAMELO; LUCIO; LEAL JÚNIOR, 2017; ALVES; GOMES, 2020). Os mais difundidos, para esse tipo estudo,

são: o erro quadrático médio (EQM) e o erro médio absoluto (EMA). No entanto Willmott e Matsuura (2005) apontam uma relativa vantagem do erro médio absoluto, em vez da raiz do erro quadrático médio, que envolve três passos contra apenas um do EMA, ou seja, diminuindo a fonte de erros para os cálculos, logo, em estudos futuros, será considerado também o EMA como método de avaliação de erro dos métodos de imputação.

Ao se empregar indicadores que não representam fidedignamente o comportamento dos métodos de imputação, introduz-se a dificuldade de comparação entre os trabalhos de imputação, que utilizaram a série de precipitação.

Este trabalho tem como principal diferença a imputação de chuva, em uma única estação meteorológica. Vale ressaltar que, muitas vezes, não é possível obter informações robustas e sem falhas de estações vizinhas, em períodos tão longos como os inferidos por este estudo. Por essa razão, pesquisas que analisam métodos de imputação, em variáveis meteorológicas obtidas de outras estações, reservam-se a períodos de tempo inferiores a 60 anos para análise (CHEN *et al.,* 2019; ESPINOSA; PORTELA; RODRIGUES, 2019; SOUZA *et al.,* 2020).

Este estudo traz à luz que, mesmo os métodos mais sofisticados, como o *random forest,* ou métodos extremamente tradicionais na área, como regressão linear, tem suas fragilidades, mostrando-se necessária uma escolha melhor de variáveis explicativas para os métodos de imputação, bem como a melhor escolha de modelos para esses métodos (OLIVEIRA; OLIVEIRA; MONTEIRO, 2017; CHEN *et al.,* 2019; ALVES; GOMES, 2020).

Portanto este estudo será aprimorado, inserindo outras variáveis climatológicas, que podem descrever melhor o comportamento da precipitação pluvial, como radiação solar, umidade relativa do ar e temperatura. Serão testados também outros métodos de imputação, que possivelmente terão como base modelos de ajuste espaço-temporal, por exemplo, modelos da classe autoregressivos e de média móvel espaço-temporais (STARMA)(SAHA *et al.,* 2020; CHEN *et al.,* 2021).

REFERÊNCIAS

ALVES, L. E. R.; GOMES, H. B. Validação da imputação múltipla via Predictive Mean Matching para preenchimento de falhas nos dados pluviométricos da Bacia do Médio São Francisco. **Anuário do Instituto de Geociências,** Rio de Janeiro, v. 43, n. 1, p. 199–206, 2020.

BALDISERA, R. S.; DALLACORT, R. Influência das variáveis climáticas: declinação solar,

fotoperíodo e irradiação no topo da atmosfera, em regiões agricultáveis do Brasil. **Revista de Ciências Agro-Ambientais,** Alta Floresta, v. 15, n. 2, p. 109–115, 2017.

BREIMAN, L. Random forests. **Machine Learning,** Boca Raton, v. 45, n. 1, p. 5–32, 2001.

BURHANUDDIN, S. N. Z.; DENI, S. M.; RAMLI, N. M. Normal ratio in multiple imputation based on bootstrapped sample for rainfall data with missingness. **International Journal of GEOMATE,** New York, v. 13, n. 36, p. 131–137, ago. 2017.

CAMARGO, Â. P. de; SENTELHAS, P. C. Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo, Brasil. **Revista Brasileira de Agrometeorologia,** Santa Maria, v. 5, n. 6, p. 89–97, jan. 1997.

CAMELO, H. do N.; LUCIO, P. S.; LEAL JÚNIOR, J. V. Modelagem da velocidade do vento usando metodologias ARIMA, Holt-Winters e RNA na previsão de geração eólica no nordeste brasileiro. **Revista Brasileira de Climatologia***,* v. 21, n. 13, p. 449–466, jul./dez. 2017.

CARVALHO, J. R. P. de *et al.* Modelo de imputação múltipla para estimar dados de precipitação diária e preenchimento de falhas. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 32, n. 4, p. 575–583, out./dez. 2017.

CHEN, H. *et al.* A spatiotemporal estimation method for hourly rainfall based on F-SVD in the recommender system. **Environmental Modelling & Software,** Oxford, v. 144, p. 1-11, out. 2021.

CHEN, L. *et al.* Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. **Journal of Hydrology,** [Amsterdam], v. 572, p. 449–460, mai. 2019.

CONSTANTINO, G. C.; BRUNINI, O. Caracterização do regime de evapotanspiração real, em escala decendial, no Estado de São Paulo. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 22, n. 1, p. 75–82, abr. 2007.

EFRON, B.; TIBSHINRANI, R. J. **An introduction to the Bootstrap.** New York: Chapman & Hall, p. 452, 1993.

ESPINOSA, L. A.; PORTELA, M. M.; RODRIGUES, R. Spatio-temporal variability of droughts over past 80 years in Madeira Island. **Journal of Hydrology:** regional studies, [Amsterdam], v. 25, p. 1-18, out. 2019.

GARCÍA-PEÑA, M.; ARCINIEGAS-ALARCÓN, S.; BARBIN, D. Imputação de dados climáticos utilizando a decomposição por valores singulares: uma comparação empírica. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 29, n. 4, p. 527–536, dez. 2014.

GUEDES, H. A. S.; PRIEBE, P. dos S.; MANKE, E. B. Tendências em séries temporais de precipitação no norte do estado do Rio Grande do Sul, Brasil. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 34, n. 2, p. 283–291, abr./jun. 2019.

HUI, D. *et al.* Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. **Agricultural and Forest Meteorology,** [Amsterdam], v. 121, n. 1–2, p. 93–111, jan. 2004.

JARDIM, C. H.; MOURA, F. P. de. Variações dos totais de chuvas e temperatura do ar na bacia do Rio Pandeiros, norte do estado de Minas Gerais - Brasil: articulação com fatores de diferentes níveis escalares em áreas de transição climática de cerrado para semiárido. **Revista Brasileira de Climatologia***,* n. 14, p. 172–189, nov. 2018.

LEITÃO, N.; CARVALHO, L. Considerações sobre a construção e análise de séries longas de precipitação em Cernache do Bonjardim, Sertã, Portugal, 1916/2021. **Revista Pensar Acadêmico,** Manhuaçu, v. 19, n. 3, p. 654–685, set./dez. 2021.

LITTLE, R. J. A. Missing-data adjustments in large surveys missing-data ad ustments in large surveys. **Journal of Business & Economic Statistics,** [Washington, DC], v. 6, n. 3, p. 287–296, jul. 1988.

MITAL, U. *et al.* Sequential imputation of missing spatio-temporal precipitation data using random forests. **Frontiers in Water,** Lausanne, v. 2, n. 20, p. 1–15, 2020.

OLIVEIRA, H. L. C. de; OLIVEIRA, S. R. de M.; MONTEIRO, J. E. B. de A. Geração de séries temporais de dados meteorológicos utilizando algoritmo de aprendizado de máquina. *In:* CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 11., 2017, Campinas. **Anais** [...]. Campinas: Embrapa, p. 262–267, 2017.

POSTO METEOROLÓGICO "PROFESSOR JESUS MARDEN DOS SANTOS" ESALQ - USP. Tempo em Piracicaba. **Posto Meteorológico,** Piracicaba, 2020. Disponível em: http://www.leb.esalq.usp.br/posto/index.html. Acesso em: 27 ago. 2020.

RAIMUNDO, C. do C.; SANSIGOLO, C. A.; MOLION, L. C. B. Tendências das classes de precipitação na Região Metropolitana de São Paulo. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 29, n. 3, p. 397–408, set. 2014.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Disponível em: http://www.r-project.org/. Acesso em: 25 jul. 2021.

REBOITA, M. S. *et al.* Aspectos climáticos do estado de Minas Gerais. **Revista Brasileira de Climatologia***,* v. 17, n. 11, p. 206–226, jul./dez. 2015.

RICCE, W. da S. *et al.* Zoneamento agroclimático da cultura do abacaxizeiro no Estado do Paraná. **Semina:** ciências agrarias, Londrina, v. 35, n. 4, p. 2337–2346, 2014.

RUBIN, D. B. **Multiple imputation for nonresponse in surveys.** New York: John Wiley & Sons, p. 258, 1987.

RUIZ-CÁRDENAS, R.; KRAINSKI, E. T. Preenchimento de falhas em bancos de dados meteorológicos diários: uma comparação de abordagens. *In:* CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA, 17., 2011, Guarapari. **Anais** [...]. Guarapari: SESC, p. 1-5, 2011.

SAHA, A. *et al.* A hybrid spatio-temporal modelling: an application to space-time rainfall forecasting. **Theoretical and Applied Climatology,** Wien, v. 142, n. 3–4, p. 1271–1282, set. 2020.

SANTOS, R. S. dos *et al.* Caracterização de extremos mensais de precipitação em Cacoal (RO).

Revista Brasileira de Climatologia, v. 22, n. 14, p. 267–268, jan./jun. 2018.

SANTOS, T. V. dos *et al.* Estimativa da evapotranspiração na mesorregião do baixo São Francisco. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 35, p. 981–993, out./dez. 2020.

SENA, J. P. de O. *et al.* Comparação entre dados de chuva derivados do Climate Prediction Center e observados para a região do Cariri Paraibano. **Revista Brasileira de Geografia Física,** Pernambuco, v. 5, n. 2, p. 412–420, 2012.

SILVA, D. S. B. S. *et al.* Imputação de dados diários de radiação solar global via ambiente R. **Enciclopédia Biosfera,** Goiânia, v. 16, n. 29, p. 957–969, 2019.

SOUZA, E. B. de *et al.* Padrões climatológicos e tendências da precipitação nos regimes chuvoso e seco da Amazônia Oriental. **Revista Brasileira de Climatologia***,* v. 21, n. 13, p. 81–93, jul./dez. 2017.

SOUZA, E. de O. *et al.* Estimativa e espacialização da erosividade em mesorregiões climáticas no estado de Alagoas. **Revista Brasileira de Meteorologia,** São José dos Campos, v. 35, p. 769–783, out./dez. 2020.

VAN BUUREN, S. **Flexible imputation of missing data.** New York: Chapman & Hall, p. 316, 2012.

VAN BUUREN, S.; GROOTHUIS-OUDSHOORN, K. MICE: Multivariate Imputation by Chained Equations in R. **Journal of Statistical Software,** [California], v. 45, n. 3, p. 1–68, dez. 2011.

VICENTE, M. R. *et al.* Evapotranspiração de referência utilizando o método FAO Penman-Monteith com dados faltantes. **Global Science and Technology,** Rio Verde, v. 11, n. 3, p. 217–228, set./dez. 2018.

WILLMOTT, C. J. *et al.* Statistics for the evaluation and comparison of models. **Journal of Geophysical Research,** Hoboken, v. 90, n. C5, p. 8995–9005, set. 1985.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate Research,** Oldendorf/Luhe, v. 30, p. 79–82, dez. 2005.