**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Comparação entre métodos de imputação de dados em
diferentes intensidades amostrais na série homogênea de
precipitação pluvial da ESALQ**

Suelen Cristina Gasparetto

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração:
Estatística e Experimentação Agronômica

**Piracicaba
2019**

Suelen Cristina Gasparetto
Licenciada em Matemática

**Comparação entre métodos de imputação de dados em diferentes
intensidades amostrais na série homogênea de precipitação pluvial da
ESALQ**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:

Profª Drª **SÔNIA MARIA DE STEFANO PIEDADE**

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração:
Estatística e Experimentação Agronômica

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Gasparetto, Suelen Cristina

Comparação entre métodos de imputação de dados em diferentes intensidades amostrais na série homogênea de precipitação pluvial da ESALQ / Suelen Cristina Gasparetto. – versão revisada de acordo com a resolução CoPGr 6018 de 2011. - Piracicaba, 2019.

84 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Chuva 2. Homogeneidade de série climatológica 3. Imputação múltipla 4. *MICE* 5. *PMM* 6. *Random forest* 7. Regressão linear via método *bootstrap* I. Título

DEDICATÓRIA

Dedico...

A todas as mulheres que estiveram a minha volta neste período, e me inspiraram de modo extremamente positivo a ser a mulher que sou hoje, em especial a minha querida mãe, Arlete Gasparetto, a qual foi sempre uma inspiração de representação feminina para mim, sempre com muita força e extrema dedicação.

“A força das mulheres
Está no cheiro de terra molhada
Depois da chuva esperada

A força das mulheres
Está na terra arrasada
Na planície devastada do sertão
Erguendo-se em seu próprio pó ...”

Águas – Cristiane Sobral

AGRADECIMENTOS

Segundo Aurélio (2006) a palavra agradecer significa mostrar-se grato, ou seja, mostrar sentimento experimentado por uma pessoa em relação a alguém que lhe concedeu algum favor, um auxílio ou benefício qualquer, porém neste trabalho tive muita gente que concedeu-me favores, auxílios entre tantas outras formas de ajuda, assim, quero agradecer a todos os colegas, amigos e até mesmo conhecidos que participaram da minha vida durante este período, sou muito grata a todos vocês, infelizmente não conseguirei citar todos os nomes, e por isso peço perdão.

Primeiramente, gostaria de agradecer a Deus, por mais esta oportunidade na minha vida, de ter conseguido encarar mais esta experiência tão válida e tão construtiva como foi o mestrado.

Gostaria de agradecer a minha família sanguínea: minha mãe, meu irmão, tia Ardevia, tio Mario, tia Marize, tio Sírío, minha nona Deleza, meus primos e primas, os quais fizeram diferença neste momento de abalo emocional. Obrigada por vocês sempre me apoiarem, me mostrarem como vocês aprenderam a amar e me mostrarem o quanto eu tenho sorte de ter vocês ao meu lado sempre, por meio de orações, de conselhos, e mensagens.

Sei que pode parecer estranho, mas quero agradecer o meu anjinho de quatro patas, o meu cachorro Cauchy Newton. Este bichinho mostrou muito amor e carinho por mim em todo este período, sempre se deitando ao meu lado quando estava insanamente escrevendo a dissertação, e a cada vez que eu levantava da cadeira ele estava lá, prontamente de pé, esperando de rabo abanando para me seguir aonde quer que eu fosse.

Não poderia deixar de agradecer a família que eu escolhi para ficar comigo neste momento, meus amigos. Bruno Koga, depois que o conheci, soube que tinha um parceiro de risadas, com as nossas *fanfics* sobre os assuntos do departamento, para os estudos, principalmente de Estatística Matemática I, e o mais importante, para os assuntos da vida. Fernando Costa, Kamylla Rodrigues Leandro, Michele Maria Signoretti Penso e Alessandra de Lima Goes, agradeço demais a vocês por todo o companheirismo e toda a ajuda que vocês me deram, vocês me mostraram que eu não estava sozinha nos momentos de desespero, achando que eu poderia ser a única que não estava entendendo nada no começo da nossa jornada neste mestrado, e me mostraram o quanto é importante ter colegas que não precisam ser rivais e sim companheiros.

O meu muito obrigada para a minha orientadora, professora Sônia Maria De Stefano Piedade, por toda sua compreensão, pela aposta que a senhora fez em mim, me dando a ideia para este tema de projeto tão especial.

Um agradecimento especial para o professor Luiz Roberto. Angelocci, o qual me ajudou de forma excepcional, a esta pessoa eu não tenho palavras para definir o tamanho da minha gratidão. Um professor que não tinha motivo algum de ter me ajudado do jeito que me ajudou, sempre me ensinando de forma atenciosa e compassiva.

Agradeço também aos colegas da ESALQ de forma geral Welinton Yoshio Hirai, Eduardo Elias Ribeiro Junior, Valiana Alves Teodoro, Danilo Roberti Alves de Almeida (Pulga) entre outros, vocês estiveram presentes nos momentos de criação e finalização deste projeto, dando ideias interessantes, e considerações construtivas.

Ao Departamento de Ciências Exatas - LCE, em especial aos professores com quem tive a oportunidade de ter aulas, obrigada por todo conhecimento que

adquiri com vocês em especial as professoras Roseli Aparecida Leandro e Renata Alcarde Sermarini, os professores Carlos Tadeu dos Santos Dias, José Silvio Govone e Cristian Marcelo Villegas Lobos e todos os outros. Agradeço aos funcionários do departamento, em especial a Solange de Assis Paes Sabadin e a Luciane Bração, por toda ajuda e informação, um agradecimento especial para Eliana Maria Garcia, funcionária da biblioteca, por toda ajuda com a formatação deste documento.

Gostaria de agradecer também aos meus amigos de fora da ESALQ que compartilharam comigo este período, Vinicius Brasílio Canale, Pablo Gonzalez Pagotto, Suzete Maria Silva Afonso, Danilo Cardoso e Jamal Bakr vocês são muito bons e muito inteligentes, me inspiro em vocês.

Ao apoio financeiro do CNPq para o desenvolvimento e realização desse trabalho.

Durante estes dois anos e meio eu entendo que tive um crescimento muito importante, não só profissional mas como pessoal, logo, eu não poderia de deixar de agradecer a Juliana Danelon, por todos os momentos os quais ela me proporcionou de reflexão e de confronto com as minhas próprias ideias e ideais, sem a ajuda dela eu não seria a pessoa que eu gosto de ser hoje.

Ao Daniel de Almeida Papa, o meu obrigada. Você me fez enxergar como eu poderia ser uma pessoa melhor.

Ao meu pai, por ter me mostrado o quanto a vida é curta e que não podemos perder momentos preciosos por covardia e conformismo.

A vida não é para fracos e indecisos, é necessário ter garra, e acima de tudo coragem. Seguir os desejos e sonhos não é fácil ou cômodo, mas quem não faz isso não vive, mesmo que isso muitas vezes vá contra o que a sociedade nos impõe.

Hoje eu prefiro ir contra as regras do jogo do que deixar que os sonhos e os desejos passem.

SUMÁRIO

RESUMO	10
ABSTRACT	11
LISTA DE FIGURAS	12
LISTA DE TABELAS	14
1. INTRODUÇÃO	15
1.1 Objetivo	19
2. REVISÃO DE LITERATURA	21
2.1 Homogeneidade da série	21
2.1.1 Teste de Pettitt	23
2.1.2 Teste de Buishand	24
2.1.3 Teste Normal Padrão (<i>SNHT</i>)	26
2.1.4 Classificação de homogeneidade	28
2.2 Imputação de dados	29
2.2.1 Generalidades sobre a imputação de dados	29
2.2.2 Mecanismos associados aos dados faltantes	30
2.2.3 Imputação única	32
2.2.4 Imputação múltipla	32
2.2.4.1 <i>PMM (Predictive Mean Matching)</i>	36
2.2.4.2 <i>Random forest</i>	38
2.2.4.3 Regressão linear via o método <i>bootstrap</i>	38
2.3 Comparação estatística dos métodos de imputação	41
2.3.1 Raiz do erro quadrático médio (REQM)	41
2.3.2 Índice de acurácia de Willmott (d)	42
2.3.3 Coeficiente de correlação de Pearson (r)	42
2.3.4 Índice de desempenho (c)	44
2.4 Homogeneidade e imputação de dados climáticos, com ênfase em precipitação no Brasil	45
3. MATERIAL E MÉTODOS	49
3.1 Dados meteorológicos	49
3.2 Tratamento dos dados	51
3.3 Análise descritiva preliminar	53
3.4 Homogeneidade da série	53

3.5 Métodos de imputação	54
3.6 Comparação estatística dos métodos de imputação.....	54
4. RESULTADOS E DISCUSSÃO.....	55
4.1 Análise descritiva preliminar.....	55
4.2 Homogeneidade da série	56
4.3 Métodos de imputação	59
4.3.1 <i>PMM</i>	59
4.3.2 <i>Random forest</i>	61
4.3.3 Regressão linear via método <i>bootstrap</i>	63
4.4 Comparação estatística dos métodos de imputação.....	65
4.4.1 Raiz do erro quadrático médio	65
4.4.2 Índice de acurácia de Willmott (<i>d</i>)	67
4.4.3 Índice de desempenho (<i>c</i>).....	69
5. CONCLUSÕES	75
REFERÊNCIAS.....	77

RESUMO

Comparação entre métodos de imputação de dados em diferentes intensidades amostrais na série homogênea de precipitação pluvial da ESALQ

Problemas frequentes nas análises estatísticas de informações meteorológicas são a ocorrência de dados faltantes e ausência de conhecimento acerca da homogeneidade das informações contidas no banco de dados. O objetivo deste trabalho foi testar e classificar a homogeneidade da série de precipitação pluvial da estação climatológica convencional da ESALQ, no período de 1917 a 1997, e comparar três métodos de imputação de dados, em diferentes intensidades amostrais (5%, 10% e 15%) de informações faltantes, geradas de forma aleatória. Foram utilizados três testes de homogeneidade da série: Pettitt, Buishand e normal padrão. Para o “preenchimento” das informações faltantes, foram comparados três métodos de imputação múltipla: *PMM (Predictive Mean Matching)*, *random forest* e regressão linear via método *bootstrap*, em cada intensidade amostral de informações faltantes. Os métodos foram utilizados por meio do pacote *MICE (Multivariate Imputation by Chained Equations)* do R. A comparação entre cada procedimento de imputação foi feita por meio da raiz do erro quadrático médio, índice de exatidão de Willmott e o índice de desempenho. A série de chuva foi entendida como de classe 1, ou seja, “útil” - Nenhum sinal claro de falta de homogeneidade foi aparente e, o método que resultou em menores valores da raiz quadrada dos erros e maiores índices 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 “Péssimo”.

Palavras-chave: Chuva; Homogeneidade de série climatológica; Imputação múltipla; *MICE*; *PMM*; *Random forest*; Regressão linear via método *bootstrap*

ABSTRACT

Comparison between data imputation methods at different sample intensities in the ESALQ homogeneous rainfall series

Frequent problems in the statistical analyzes of meteorological information are the occurrence of missing data and missing of knowledge about the homogeneity of the information contained in the data base. The objective of this work was to test and classify the homogeneity of the rainfall series of the conventional climatological station of the ESALQ from 1917 to 1997 and to compare three methods of data imputation in different sample intensities (5%, 10% and 15%), of missing data, generated in a random way. Three homogeneity tests were used: Pettitt, Buishand and standard normal. For the "filling" of missing information, three methods of multiple imputation were compared: PMM (Predictive Mean Matching), random forest and linear regression via bootstrap method, in each sampling intensity of missing information. The methods were used by means of the MICE (Multivariate Imputation by Chained Equations) package of R. The comparison of each imputation procedure was done by root mean square error, Willmott's accuracy index and performance index. The rainfall series was understood to be class 1 "useful" - No clear sign of lack of homogeneity was apparent and the method that resulted in smaller values of the square root of the errors and higher indexes was the PMM, in particular the intensity of 10% of missing information. The performance index for the three methods of imputation the data at all missing observation intensities was considered "Terrible".

Keywords: Rainfall; Homogeneity of climatological series; Multiple imputation; MICE; PMM; Random forest; Regression linear by method bootstrap

LISTA DE FIGURAS

Figura 1. Principais etapas da imputação múltipla.....	33
Figura 2. Classificação das correlações de Pearson, segundo Dancey e Reidy (2017)	43
Figura 3. Organograma dos bancos de dados e seu uso nos diferentes métodos de imputação	53
Figura 4. Alturas pluviométricas anuais da estação meteorológica da ESALQ no período de 1917 a 1997.....	55
Figura 5. Variação da altura pluviométrica média mensal da estação meteorológica da ESALQ no período de 1917 a 1997	56
Figura 6. Estatística K_T , do teste de Pettitt para precipitação pluvial em escala mensal da estação meteorológica da ESALQ, nos anos de 1917 a 1997.....	57
Figura 7. Estatística $R/\sqrt{T}$, do teste de Buishand para precipitação pluvial em escala mensal da estação meteorológica da ESALQ, nos anos de 1917 a 1997	58
Figura 8. Estatística T_v , do teste <i>SNHT</i> para precipitação pluvial em escala mensal da estação meteorológica da ESALQ, nos anos de 1917 a 1997.....	58
Figura 9. Gráficos de dispersão da variável precipitação pluvial mensal imputada pelo método <i>PMM</i> versus observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” respectivamente.	60
Figura 10. Gráficos de caixa (<i>boxplot</i>) da variável precipitação pluvial mensal imputada pelo método <i>PMM</i> e a variável observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” respectivamente.	60
Figura 11. Gráficos de distribuição empírica da variável precipitação pluvial mensal imputada pelo método <i>PMM</i> e a variável observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” respectivamente.	61
Figura 12. Gráficos de dispersão da variável precipitação pluvial mensal imputada pelo método <i>random forest</i> versus observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” respectivamente.	61
Figura 13. Gráficos de caixa (<i>boxplot</i>) da variável precipitação pluvial mensal imputada pelo método <i>random forest</i> e a variável observada, ambas em milímetros (mm), para os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” respectivamente.	62

Figura 14. Gráficos de distribuição empírica 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.63

Figura 15. 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.64

Figura 16. 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.64

Figura 17. Gráficos de distribuição empírica da variável precipitação pluvial mensal imputada pelo método de 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.65

Figura 18. Gráfico de linhas 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 de dados faltantes de 5%, 10% e 15%.67

Figura 19. Gráfico de linhas do índice de acurácia de Willmott (d) da precipitação pluvial imputada por meio dos métodos: *PMM*, *random forest* e regressão linear via método *bootstrap*, nas porcentagens de dados faltantes de 5%, 10% e 15%68

Figura 20. Gráfico de linhas do coeficiente de correlação de Pearson (r) da precipitação pluvial imputada por meio dos métodos: *PMM*, *random forest* e regressão linear via método *bootstrap*, nas porcentagens de dados faltantes de 5%, 10% e 15%70

Figura 21. Gráfico de linhas do índice de desempenho (c) da precipitação pluvial imputada por meio dos métodos: *PMM*, *random forest* e regressão linear via método *bootstrap*, nas porcentagens de dados faltantes de 5%, 10% e 15%71

LISTA DE TABELAS

Tabela 1. Valores críticos de $R/\sqrt{T}$ para 90%, 95% e 99% de nível de confiança do teste de Buishand em função de T (BUIHAND, 1982)	26
Tabela 2. Critério de interpretação do desempenho dos métodos pelo índice (c) Camargo e Sentelhas (1997).....	44
Tabela 3. Valores da raiz do erro quadrático médio da variável precipitação pluvial imputada por meio dos métodos <i>PMM</i> , <i>random forest</i> e regressão linear via método <i>bootstrap</i> nas porcentagens 5%, 10% e 15% de dados faltantes.....	67
Tabela 4. Valores do índice de acurácia de Willmott (<i>d</i>) da variável precipitação pluvial imputada por meio dos métodos <i>PMM</i> , <i>random forest</i> e regressão linear via método <i>bootstrap</i> nas porcentagens 5%, 10% e 15% de dados faltantes.....	68
Tabela 5. Valores dos coeficientes de correlação de Pearson (<i>r</i>) precipitação pluvial imputada por meio dos métodos <i>PMM</i> , <i>random forest</i> e regressão linear via método <i>bootstrap</i> nas porcentagens 5%, 10% e 15% de dados faltantes.....	70
Tabela 6. Valores do índice de desempenho (c) da precipitação pluvial imputada por meio dos métodos <i>PMM</i> , <i>random forest</i> e regressão linear via método <i>bootstrap</i> nas porcentagens 5%, 10% e 15% de dados faltantes	72

1. INTRODUÇÃO

As informações meteorológicas desempenham um papel crítico na tomada de decisões em várias áreas de atividades humanas, com ênfase nas atividades agrícolas, sendo imprescindível que elas sejam completas e confiáveis. Elas devem consistir em séries homogêneas de dados robustos e contínuos (CARVALHO et al., 2017). Um banco de dados completo possibilita estimar probabilidades de ocorrência de variáveis meteorológicas, como: temperaturas extremas e médias, altura pluviométrica e velocidades extremas do vento (RUIZ-CÁRDENAS; KRAINSKI, 2011). Com a variável chuva em especial, pode-se 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 a temperatura, a caracterização físico-hídrica do solo e a análise de uma série possibilita caracterizar as épocas adequadas ao plantio e colheita para cada região (zoneamento de riscos climáticos) para determinar risco de ocorrência de doenças e pragas e as estratégias de controle. Além disso, séries históricas permitem estudar tendência, ao longo dos anos, da variação de um elemento meteorológico como a quantidade de chuva.

No Brasil a criação de estações meteorológicas ganhou ênfase a partir da última década do século XIX, concentrando-se mais nas Regiões Sul, Sudeste e Nordeste ampliando-se, ainda que numericamente insuficientes, para outras regiões, permitindo o estabelecimento de bancos de dados utilizados nas diferentes especialidades da climatologia. A ESALQ conta com um acervo de observações meteorológicas realizadas dentro do próprio campus. Embora haja relato de medidas no final do século XIX (MORIMONT, 1895), que não foram encontradas, o acervo atualmente disponível consta de observações diárias desde 1917, sendo que recentemente (2017) foram incorporados dados consolidados por mês entre 01/06/1902 e 31/12/1916, descobertos no Museu de Meteorologia da USP. Trata-se de uma das séries mais longas do território brasileiro, entre as cerca de cinquenta com arquivos seculares no país (ANGELOCCI; SENTELHAS, 2017). Além da estação meteorológica servir de apoio didático, as suas observações já serviram ao sistema nacional de previsão do tempo até meados da década de 60. A partir da década de 1950, elas dão grande suporte à pesquisa na ESALQ, sendo parte da produção científica relacionada à análise direta dos dados (CERVELLINI; SALATI;

GODOY, 1966; GOMES; SALATI, 1957; LISKÁ, GILBERTO RODRIGUES; BORTOLINI; SÁFADI; BEIJO, 2011) ou à ajustes dos elementos meteorológicos à distribuições estatísticas (PICCINNI; ANGELOCCI, 1997; SANSIGOLO, 2008 LISKÁ, BORTOLINI; SÁFADI, THELMA; BEIJO, 2011).

Quanto aos estudos de tendência do clima, por exemplo, é fundamental que sejam detectados pontos que indiquem uma falta de coesão entre as observações coletadas, necessitando-se um ajuste dos dados (PETERSON et al., 1992; KYSELÝ; DOMONKOS, 2006). A homogeneidade de uma série, refere-se ao fato de que a variabilidade temporal observada deve ser função apenas das condições de tempo/clima do local, o que pode não ocorrer por vários motivos: mudanças do local da estação (distância e condição topográfica), alterações da cobertura da superfície no seu entorno (cobertura vegetal, construções) e dos protocolos de medida, além de problemas de erros e de calibração dos equipamentos (AGUILAR et al., 2003; MARTÍNEZ et al., 2011; SANTOS, 2012).

A série climatológica aqui estudada, foi obtida nesses mais de 100 anos sob condições que podem levar à falta de homogeneidade, devido a fatores de mudança de local da estação, substituição de equipamentos e de protocolos de medida, além do crescimento físico da cidade de Piracicaba em seu entorno. Ressalta-se que a primeira mudança de local em 1955 da estação meteorológica, foi consequência de ter começado a ficar exposta à presença de grandes prédios e urbanização do campus próximo ao local de instalação. Mesmo o local atual começa a ficar sob efeito de edificações que dele se aproximam, bem como do grande aumento da área urbanizada no entorno do campus.

Muitos estudos já foram feitos testando a homogeneidade das séries climatológicas no mundo. No Brasil, existe a preocupação nos últimos 30 anos em verificar a homogeneidade em séries temporais de localidades em vários estados, principalmente nas regiões sul e sudeste (BACK, 2001; MORTATTI et al., 2004). A verificação de homogeneidade é de fundamental importância para que a análise estatística seja feita corretamente, já que as teorias de probabilidade baseiam-se na hipótese que as séries estudadas sejam homogêneas, isto é, que não apresentem tendências que não sejam única e exclusivamente climatológicas (WANDERLEY et al., 2013). A única informação sobre a possível homogeneidade da série de dados meteorológicos da ESALQ, é encontrada em Assis (1991), mas o autor não deixa

claro como chegou a essa conclusão, pois o tipo de análise por ele feita não explicita um método específico de avaliação de homogeneidade.

Para tanto, no presente estudo serão realizados três testes de homogeneidade da série: de Pettitt, Buishand e normal padrão (*SNHT, standard normal homogeneity test*). Os dois primeiros são mais sensíveis a detecção de falta de homogeneidade no meio de uma série temporal, enquanto o terceiro é mais suscetível a detectar heterogeneidade no fim da série (WIJNGAARD; KLEIN TANK; KÖNNEN, 2003). Os testes serão agrupados de modo quantitativo, e assim poderá ser feita uma classificação adequada da série.

Para resolver problemas de banco de dados com informações faltantes, é indicado o uso de métodos de imputação de dados (CARVALHO et al., 2017). O procedimento mais comum é restringir-se a análise das informações com dados completos nas variáveis, desconsiderando a parte faltante, e a vantagem é a conveniência de trabalhar com um banco de dados completos. Contudo, esse procedimento de exclusão de dados causa complicações para as análises, como estimativas de viés, fazendo com que se possa chegar a conclusões enganosas ou tendenciosas (GANDOLFI, 2016).

Os primeiros métodos de imputação foram: pelo uso de média ou mediana da variável; por meio do vizinho mais próximo, *hot deck*, é um método de imputação no qual um dado faltante é imputado a partir de uma registro similar registrado, este registro é considerado assim possível “doador” para a obtenção dos valores faltantes (FERLIN, 2008); regressão linear e por meio da máxima verossimilhança (GANDOLFI, 2016). Todos eles são métodos de imputação única, o que quer dizer que todos os dados ausentes foram imputados apenas com um único valor, estabelecido por cada método. As técnicas de imputação única têm sérios inconvenientes, visto que a maioria das abordagens produz estimativas de parâmetros tendenciosas e atenuam os erros padrão. Raciocinando em um nível intuitivo, os valores ausentes devem aumentar os erros padrão, já que adicionam outra camada de ruído para as estimativas dos parâmetros (ENDERS, 2010). Para contornar os problemas, Rubin (1987) propôs um novo método: a imputação múltipla.

Os primeiros estudos de imputação de dados aplicados à meteorologia usavam a média do banco de dados completos ou estimações por regressão linear (GARCÍA-PEÑA; ARCINIEGAS-ALARCÓN; BARBIN, 2014). Em estudos mais

recentes, pode-se perceber a atualização dos métodos de imputação, feitos por meio de processos iterativos, como a imputação múltipla. Esses métodos reduzem consideravelmente as desvantagens da imputação única, pois estudos mostram que a imputação múltipla é mais eficiente que técnicas de geoestatística, como krigagem e cokrigagem ordinária (CARVALHO et al., 2017).

Este trabalho testa métodos de imputação múltipla na variável precipitação pluvial em escala mensal, ou seja, a soma das precipitações diárias de cada mês referente.

Os métodos de imputação de dados estão disponíveis nos principais *softwares* estatísticos, de forma paga ou gratuita, como SAS e R (TEAM, 2017). As análises foram feitas por meio do pacote *MICE* do *software* R (gratuito). *MICE* (*Multivariate Imputation by Chained Equation*) é um método de cadeia de Markov Monte Carlo (MCMC- "*Markov chain Monte Carlo*"). De maneira específica, este algoritmo é um amostrador de Gibbs, técnica de simulação Bayesiana, que amostra distribuições condicionais com a finalidade de obter amostras da distribuição conjunta (GANDOLFI, 2016).

O trabalho aplica os métodos estudados na série de dados de precipitação pluvial da estação climatológica da ESALQ - USP. Primeiramente foi feita a classificação adequada da série, utilizando os três testes de homogeneidade. Em seguida são aplicados três métodos de imputação múltipla: *PMM* (*Predictive Mean Matching*), *random forest* e imputação via regressão linear usando método *bootstrap*, em diferentes porcentagens, 5%, 10% e 15% de dados faltantes, ausências provocadas de forma aleatória.

O trabalho tem como preocupação não somente analisar os erros, mas também investigar a acurácia e o desempenho dos modelos de imputação desenvolvidos pela comparação entre valores de altura pluviométrica mensal estimados pelos modelos propostos e os observados pelo uso de índices estatísticos. Serão feitas as comparações por meio da raiz do erro quadrático médio (*REQM*), além dos coeficientes de acurácia de Willmott (*d*) e de desempenho (*c*).

1.1 Objetivo

Este estudo tem como objetivo testar a homogeneidade da série histórica de chuva da ESALQ por meio dos testes de Pettitt, Buishand e normal padrão (*SNHT*), além de comparar três métodos de imputação de dados, com o uso do pacote *MICE* do R: *PMM*, *random forest* e via regressão linear usando *bootstrap*, em diferentes intensidades amostrais, 5%, 10% e 15% de dados faltantes, causados de forma aleatória.

2. REVISÃO DE LITERATURA

Pela importância fundamental da água à vida na Terra, o registro da ocorrência da precipitação pluvial, bem como de sua quantificação e distribuição espacial e temporal, merece atenção especial nas atividades técnico-científicas, principalmente nas aplicações climatológicas às ciências ambientais. Há relatos de sua medição desde a antiguidade principalmente sua aplicação em agricultura (ANGELOCCI; SENTELHAS, 2017), mas a forma atual de observação iniciou-se a cerca de 300 anos, com as medições sistemáticas e formação das primeiras séries históricas de dados meteorológicos na Europa. No Brasil, a observação sistemática de dados climáticos começou a se tornar realidade nas últimas duas décadas do século XIX e, ainda que atualmente o número de estações meteorológicas de superfície não seja o mais adequado para a extensão territorial do país, isso permitiu o estabelecimento de séries climáticas históricas e o seu emprego em múltiplos trabalhos de climatologia básica e da aplicada, sendo que a estatística tem sido uma ferramenta imprescindível neles empregada (SANTOS, 2012; OLIVEIRA; SANTOS, 2017; VIEIRA et al., 2018)

Como ocorre com qualquer outro elemento meteorológico, para a constituição de um banco de dados de altura pluviométrica é necessário medir, processar, corrigir e dar consistência aos dados de maneira mais eficiente possível (WISSMANN et al., 2006). O primeiro passo para a análise correta e a aplicação eficiente desta variável é testar a homogeneidade da série, já que as teorias de probabilidade baseiam-se na hipótese que elas sejam homogêneas, isto é, que não apresentem tendências, logo, são séries que podem apresentar mudanças, porém apenas mudanças periódicas, com um suave acréscimo ou decréscimo nos valores médios no período de registro (BACK, 2001). Após a verificação da homogeneidade da série, pode ser solucionado o problema de dados faltantes, por meio do procedimento de imputação de dados.

2.1 Homogeneidade da série

A aplicação de procedimentos para verificar a homogeneidade de uma série visa analisar tendências, variabilidade dos extremos climáticos e mudanças nas rotinas observacionais, sendo um procedimento associado ao controle de qualidade

dos dados (WIJNGAARD; KLEIN TANK; KÖNNEN, 2003). Segundo Daly *et al.* (2004) o rigoroso controle de qualidade de dados climáticos é provavelmente o aspecto mais importante para garantir um resultado bem sucedido das análises, evitando conclusões equivocadas, implicando também o uso de metadados com informações históricas, as mais completas possíveis, sobre o funcionamento de uma estação meteorológica ou uma rede delas LIMA (2003). O manejo apropriado da coleta de dados nas observações nem sempre é conseguido, além de muitas vezes as informações serem precárias, incorrendo-se em dificuldade de se ter segurança sobre a qualidade dos dados, principalmente em registros de anos passados.

A verificação da homogeneidade de uma série tem sido um aspecto bastante estudado no mundo. (AGUILAR *et al.*, 2003) cita 14 métodos ou técnicas de determinação da homogeneidade de um banco de dados, explicando que esse número é decorrente do fato que cada grupo de pesquisadores desenvolveu uma filosofia diferente quanto aos ajustes de dados que pretendem para cumprir as exigências e os objetivos dos seus trabalhos. Além disso, no desenvolvimento das técnicas, deve-se considerar qual o tipo de elemento meteorológico constitui o interesse (as análises têm sido aplicadas mais à temperatura e/ou à precipitação), em qual extensão temporal e completude dos dados está sendo feita a análise, qual a disponibilidade de metadados, qual é a densidade de estações meteorológicas quando a técnica desenvolvida exige comparações entre diferentes séries, qual é a extensão temporal e espacial dos dados (dependente da parte do mundo onde os dados são obtidos). Tais técnicas são mais aplicáveis quando se trabalha com dados mensais, estacionais ou anuais, uma vez que em alguns estudos pode-se verificar a qualidade e estabilidade de séries históricas nestas grandezas, o uso da escala diária gera estes inconvenientes, que podem causar sérias dificuldades para a análise da série (RUIZ-CÁRDENAS; KRAINSKI, 2011; FERRARI; OZAKI, 2014). O último passo a ser seguido, se necessário, é o ajuste dos dados (“homogeneização”), ou seja, o procedimento para tornar uma série temporal homogênea por uma técnica que remova vieses.

No Brasil, vários trabalhos têm empregado algumas dessas técnicas para testar homogeneidade e tendência de séries históricas de clima em várias regiões, com uso de bancos de dados de temperatura e/ou precipitação pluvial (BACK, 2001; SANTOS, 2010; BURIOL *et al.*, 2012; SALVADOR; SANTOS, 2012; FERRARI; OZAKI, 2014). Dentre os métodos utilizados, são citados por exemplo, o de dupla

massa, estratégia que consiste em comparar a precipitação mensal acumulada de um pluviômetro com a média dos pluviômetros localizados em estações vizinhas; teste *Run*, que verifica a aleatoriedade da série, teste de Mann-Kendall, no qual a hipótese nula admite que as observações de uma série temporal ocorrem de forma independente, e identicamente distribuídas. Foram escolhidos para este estudo três testes bastante empregados em trabalhos no exterior e no Brasil sobre a temática deste estudo : teste de Pettitt, Buishand e normal padrão (*SNHT* - “*Standard Normal Homogeneity Test*”) para um único ponto, pois, cada teste tem a sensibilidade em diferentes pontos de mudança ou descontinuidade dentro da série (WIJNGAARD; KLEIN TANK; KÖNNEN, 2003). Além disso, os três podem ser aplicados à série isoladamente, como é o caso do presente estudo e são descritos a seguir.

2.1.1 Teste de Pettitt

Em 1979, foi elaborado o teste de Pettitt. É um teste não paramétrico, o qual utiliza uma versão do teste de Mann-Whitney, que verifica se duas amostras são da mesma população (MORTATTI et al., 2004). O teste de Pettitt é descrito da seguinte forma: a sequência de variáveis aleatórias é dita ter um ponto de mudança (ponto de “quebra”) em A se, x_t para $t = 1, \dots, A$ tem uma função de distribuição acumulada $F_1(x_t)$, e x_t para $t = A + 1, \dots, T$, onde T é a quantidade de observações da série analisada que tem uma função de distribuição acumulada $F_2(x_t)$, onde $F_1(x_t) \neq F_2(x_t)$. Considera-se o problema de testar a hipótese nula de “nenhuma mudança”, $H_0: A = T$, contra a alternativa de “existe mudança”, $H_a: 1 \leq A < T$, usando uma estatística não paramétrica. Não são feitas suposições sobre F_1 e F_2 , exceto que elas são contínuas, o que permite que x_t sejam variáveis Bernoulli e Binomial (A. N. PETTITT, 1979).

O teste de Mann-Whitney considera duas populações, onde não se têm informações sobre a distribuição de cada uma delas, mas é necessário que as observações destas populações sejam medidas em escala ordinal e suas funções de distribuições sejam contínuas. Este teste tem como objetivo verificar se duas amostras aleatórias e independentes são oriundas da mesma população ou não. O equivalente paramétrico a esse teste é o teste t para amostras independentes (BUSSAB; MORETTIN, 2014).

A estatística não paramétrica do teste de Pettitt é definida como:

$$U_{t,T} = \sum_{i=1}^t \sum_{j=t+1}^T \text{sgn}(X_i - X_j), \quad (1)$$

em que $\text{sgn}(x) = 1$ se $x > 0$, 0 se $x = 0$, -1 se $x < 0$.

A estatística $u_{t,T}$ é equivalente a uma estatística de Mann-Whitney conforme descrita no parágrafo anterior.

O ponto de mudança (ponto de “quebra”) da série está localizado em K_T , em que

$$K_T = \max_{1 \leq t < T} |U_{t,T}|, \quad (2)$$

caso em que a estatística K_T seja significativa.

Assim, as probabilidades de significância associadas aos valores K_T são dadas por:

$$\rho_{oa} \cong 2 \exp \left\{ \frac{-6(K_T)^2}{(T^3 + T^2)} \right\}. \quad (3)$$

Esta aproximação é válida com precisão de duas casas decimais para $\rho_{oa} \leq 0,05$.

Os valores críticos de K_T são encontrados pela seguinte equação:

$$K_{crit} = \pm \sqrt{\frac{-\ln(\rho_{oa}/2)(T^3 - T^2)}{6}}. \quad (4)$$

2.1.2 Teste de Buishand

Em 1982, foi desenvolvido o teste de Buishand, o qual tem como objetivo detectar mudança na quantidade média de precipitação por meio de testes de desvios cumulativos (BUISHAND, 1982). Tal teste pode ser descrito da seguinte forma: a sequência $Y_1, \dots, Y_T$, tal que T é a quantidade de observações da série analisada, é dita ter um ponto de mudança (ponto de “quebra”) em m se, Y_i para $i = 1, \dots, m$ tem a mesma esperança ($E[Y_i]$) de Y_i para $i = m + 1, \dots, T$, ou seja:

$$E[Y_i] = \begin{cases} \mu, & i = 1, \dots, m \\ \mu + \Delta, & i = m + 1, \dots, T \end{cases} \quad (5)$$

A variância é:

$$Var[Y_i] = \sigma_Y^2. \quad (6)$$

O modelo assume um salto na média de magnitude Δ depois de m observações, a hipótese nula $H_0: \Delta = 0$ é testada em relação à alternativa, $H_a: \Delta \neq 0$. As distribuições das estatísticas de teste são derivadas para o caso de os Y_i serem estocasticamente independentes e terem uma distribuição normal.

Os testes de homogeneidade podem ser baseados nas somas parciais ajustadas ou desvios cumulativos da média:

$$S_0^* = 0; \quad S_K^* = (Y_i - \bar{Y}) \quad K = 1, \dots, T \quad (7)$$

de modo que, K é a posição da observação que está sendo analisada. Nota-se que $S_n^* = 0$. Para um registro homogêneo, pode-se esperar que os S_K^* 's flutuem em torno de zero, já que não há um padrão sistemático nos desvios do Y_i 's em relação ao seu valor médio $\bar{Y}$.

As somas parciais ajustadas e reescaladas são obtidas dividindo S_K^* 's pelo desvio padrão da amostra:

$$S_K^{**} = \frac{S_K^*}{D_Y} \quad K = 0, \dots, T \quad (8)$$

com

$$D_Y^2 = \frac{\sum_{i=1}^T (Y_i - \bar{Y})^2}{T}. \quad (9)$$

Os valores dos S_K^{**} 's não são influenciados por uma transformação linear dos dados. Por exemplo, se a quantidade de precipitação é expressa em metros em vez de milímetros, os S_K^* 's diminuem em um fator 1000, mas os S_K^{**} 's permanecem inalterados. Por conseguinte, os testes de homogeneidade baseiam-se nas somas parciais ajustadas com reescalamento S_K^{**} :

Uma estatística que pode ser usada para testar a homogeneidade é:

$$R = \max_{0 \leq K \leq T} S_K^{**} - \min_{0 \leq K \leq T} S_K^{**}.$$

(10)

Um gráfico com pontos percentuais da distribuição de R sob a hipótese nula é dada por Wallis e O'Connell (1973). Como, R é a amplitude das somas parciais ajustadas e reescaladas, então $R/\sqrt{T}$ é a amplitude das somas parciais ajustadas e reescaladas da média, uma vez que y_i são estocasticamente independentes e têm uma distribuição normal. Segundo Wallis e O'Connell (1973), no mundo real, onde apenas uma estimativa está disponível e T é um valor pequeno, uma estimativa do fluxo médio e, portanto, do rendimento fixo baseado em um único período de registro é, na melhor das hipóteses, um melhor valor incerto, por este motivo são fornecidos alguns valores críticos de $R/\sqrt{T}$ na Tabela 1.

Tabela 1. Valores críticos de $R/\sqrt{T}$ para 90%, 95% e 99% de nível de confiança do teste de Buishand em função de T (BUISHAND, 1982)

T	Níveis de confiança		
	90%	95%	99%
10	1,21	1,28	1,38
20	1,34	1,43	1,60
30	1,40	1,50	1,70
40	1,42	1,53	1,74
50	1,44	1,55	1,78
100	1,50	1,62	1,86
∞	1,62	1,75	2,00

2.1.3 Teste Normal Padrão (SNHT)

Em 1986, foi elaborado o teste normal padrão, *SNHT* (*standard normal homogeneity test*). Segundo Alexandersson (1986) o ponto de partida deste teste é formar uma série de relações $\{q_i\}$ entre valores da estação e alguns valores de referência regionais, como neste estudo não se fez uso de estações próximas, o valor de referência será a média aritmética da própria série em questão, ou seja, $q_i = \frac{x_i}{\bar{x}}$, onde x_i são as observações da série e $\bar{x}$ é a média aritmética. Como foram usadas as proporções, não é necessário especificar a distribuição da precipitação propriamente dita. O coeficiente de variação de uma série de razões é claramente

menor do que para a série original de precipitação mensal. Isto dá uma redução do ruído nas sequências e ilustra a ideia principal de testar a homogeneidade.

Uma nova sequência padronizada de proporções $\{z_i\}$ é definida desta forma:

$$z_i = \frac{(q_i - \bar{q})}{s_q}, \quad (11)$$

tal que $\bar{q}$ é o valor médio aritmético das proporções $\{q_i\}$ e s_q é o desvio padrão da amostra desta série. Esta nova série $\{z_i\}$ tem, assim, valor médio 0 e desvio padrão 1. Define-se então as hipóteses nula (H_0), e a alternativa (H_1), como:

$$H_0: Z \in N(0,1), \forall i$$

$$H_1: \begin{cases} \text{Para algum } 1 \leq v < T \text{ e } \mu_1 \neq \mu_2 \text{ temos} \\ Z \in N(\mu_1, 1) \text{ para } i \leq v \\ Z \in N(\mu_2, 1) \text{ para } i > v \end{cases}$$

Assim, pode-se assumir que a sequência de proporções é descrita por uma distribuição normal e que a possível quebra é única, sendo ela, apenas um deslocamento do nível médio.

A estatística do teste de *SNHT* é definida como:

$$T_v = v\bar{z}_1^2 + (n - v)\bar{z}_2^2, \quad (12)$$

sendo v a possível quebra de homogeneidade da série, $v = \{1, \dots, T\}$, $\bar{z}_1$ e $\bar{z}_2$ pode ser escrita como:

$$\bar{z}_1 = \frac{1}{v} \sum_{i=1}^v z_i, \quad (13)$$

$$\bar{z}_2 = \frac{1}{(T - v)} \sum_{i=v+1}^T z_i. \quad (14)$$

O valor crítico é:

$$T_0 = \max_{1 \leq v < T} T_v. \quad (15)$$

É importante ressaltar algumas escolhas eventuais deste teste, por exemplo, não é especificada a distribuição da precipitação propriamente dita. Simplesmente foi assumido que as proporções eram variáveis (aproximadas) normais, sendo essa

suposição feita por Alexandersson (1986) mediante a algumas análises de frequência desenvolvidas em seu trabalho. Também não foram supostas teorias sobre o desvio padrão, simplesmente foi assumido que é sempre 1, independentemente das possíveis quebras de homogeneidade. Segundo o citado autor, estas e outras suposições poderiam ser criticadas, mas deve-se lembrar que um teste muito complexo será dificultado em seu uso por sua mera complexidade e que não há sentido elaborar em demasia a matemática quando se pretende usá-la em dados climáticos reais.

2.1.4 Classificação de homogeneidade

Para efeito de classificação da série em relação a sua homogeneidade, foi utilizado o critério proposto por Wijngaard, Klein Tank e Können (2003). Embora os testes descritos anteriormente tenham características semelhantes, eles também são diferentes. Segundo essa fonte citada, os testes de Buishand e Pettitt detectam com relativa facilidade quebras de homogeneidade no meio da série temporal, enquanto o teste *SNHT* é mais sensível a detecção de quebras no começo da série. Foram designadas três classes distintas, tendo por base o número de testes de homogeneidade que rejeitam a hipótese nula para um nível de significância de 5%. Sendo elas:

Classe 1: “útil” - Nenhum sinal claro de falta de homogeneidade é aparente. A série parece ser suficientemente homogênea para análise de tendências e análise de variância. Nenhum teste rejeita a hipótese nula.

Classe 2: “duvidosa” - Os resultados da análise devem ser considerados muito criticamente do ponto de vista da existência de heterogeneidade. Isto ocorre quando um ou dois testes rejeitam a hipótese nula.

Classe 3: “suspeita” - É provável que exista uma heterogeneidade que exceda o nível expresso pelo desvio padrão entre as observações da série em teste. Tal classificação é adotada quando os três testes rejeitam a hipótese nula.

2.2 Imputação de dados

A imputação de dados é um método de substituição de dados faltantes por valores plausíveis. Ela pode ser única, quando apenas um valor é colocado para cada dado ausente, ou múltipla, quando há mais de um valor para cada informação ausente.

2.2.1 Generalidades sobre a imputação de dados

É comum que os bancos de dados utilizados em estudos nas mais variadas áreas de atividade humana apresentem-se com falhas, denominados de dados “faltantes”, “perdidos” ou “ausentes”. Essa incompletude pode ter vários motivos, pois os procedimentos na constituição dos bancos são dependentes das características das áreas científicas nas quais os dados se aplicam (ciências sociais, ciências da saúde, ciências ambientais, etc.). Ignorar os dados ausentes na análise estatística é um procedimento que pode funcionar somente quando o número de ausentes é pequeno e quando ocorrem inteiramente ao acaso (GRACE-MARTIN, 2018; LITTLE, 1988). Normalmente, muitos modelos estatísticos e algoritmos de aprendizado de máquinas dependem de conjuntos completos de dados, salvo alguns algoritmos adaptados aos problemas de falta de dados, que não exigem pré-processamento (BERTSIMAS; ORFANOUDAKI; PAWLOWSKI, 2018). Fora desses casos, a exigência de se trabalhar com banco de dados completos é fundamental para a adequada análise estatística e, assim, a imputação torna-se um procedimento de enorme valia.

A imputação de dados tomou impulso a partir dos trabalhos de D.B. Rubin há cerca de 40 anos, ao introduzir a imputação múltipla (RUBIN, 1977, 1978, 1981). Desde então, a aplicação da imputação é uma técnica de trabalho de bancos de dados que tem atraído enorme atenção. Além do desenvolvimento de abordagens visando a otimização de métodos, a literatura mostra a preocupação que inclui alertas sobre os equívocos e a má utilização de imputação, como discutido nos trabalhos de (RUBIN, 1987, 1996; SCHAFER; GRAHAM, 2002), sendo que o primeiro autor atenta que “a ideia da imputação é ao mesmo tempo sedutora e perigosa”. O desenvolvimento de conceitos estatísticos, de técnicas eletrônicas de

computação com formação de pacotes estatísticos, de algoritmos e o de treinamento de máquinas tem facilitado o trabalho com bases de dados muitas vezes imensas (*big data*, ou também, *megadata*).

Para as séries climatológicas, os primeiros estudos de imputação de dados foram com uso de imputação única, como a utilização de regressão linear simples ou meramente a média das informações presentes no banco de dados (GARCÍA-PEÑA; ARCINIEGAS-ALARCÓN; BARBIN, 2014). Em estudos mais recentes, pode-se perceber o uso de métodos imputação mais sofisticados, como métodos que utilizam recursos de aprendizado de máquinas para o seu desenvolvimento (OLIVEIRA; OLIVEIRA; MONTEIRO, 2017).

2.2.2 Mecanismos associados aos dados faltantes

A escolha do método de imputação depende da classificação do banco de dados faltantes, ou seja, aquele que contém tanto os valores presentes quanto os valores faltantes. Normalmente o banco de dados obedece um mecanismo que indica as condições dos dados ausentes, podendo ser: como perdas completamente ao acaso (*Missing Completely at Random – MCAR*), perdas ao acaso (*Missing at Random – MAR*) e perdas não-aleatórias (*Not Missing at Random – NMAR*) (RUBIN, 1987).

Para tratar de cada mecanismo e poder assim classificar o banco de dados, considere $\mathbf{Y}$ 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 $\mathbf{Y}$ em dois subconjuntos $\mathbf{Y} = \{\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\}$ em que $\mathbf{Y}_{obs}$ são os dados observados (não-faltantes) e $\mathbf{Y}_{mis}$ são os dados faltantes. Correspondente a matriz de dados $\mathbf{Y}$, existe $\mathbf{R}$, uma matriz indicadora de respostas para cada item de $\mathbf{Y}$, $\mathbf{R}$ possui a mesma dimensão de $\mathbf{Y}$, sendo que os elementos da matriz $\mathbf{R}$ pode ser definido como:

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

A distribuição de R pode depender de $Y = \{Y_{obs}, Y_{mis}\}$, seja por acaso ou pelo planejamento, e essa relação é descrita pelo modelo de dados faltantes (*missing data model*). A expressão geral desse modelo é:

$$\Pr(R|Y_{obs}, Y_{mis}, \theta)$$

tal que θ é o vetor de parâmetros desconhecidos do modelo (LITTLE; RUBIN, 2002).

Os bancos de dados são classificados em três categorias:

a) Com dados faltantes inteiramente ao acaso (MCAR)

Os dados são ditos faltantes inteiramente ao acaso (*MCAR - Missing Completely At Random*) se $\Pr(R = 0|Y_{obs}, Y_{mis}, \theta) = \Pr(R = 0|\theta)$, o que significa que a probabilidade de uma observação não ser registrada não depende de qualquer outra observação, ou seja, uma observação faltante independe das outras observações do conjunto de dados, podendo ser uma observação registrada ou não. O que implica que a probabilidade de ocorrência do dado ausente é a mesma para todos os casos. Quando o mecanismo *MCAR* ocorre, os dados faltantes constituem uma sub amostra aleatória dos dados observados.

b) Com dados faltantes ao acaso (MAR)

Os dados são referidos com perdas ao acaso (*MAR - Missing At Random*), se, $\Pr(R = 0|Y_{obs}, Y_{mis}, \theta) = \Pr(R = 0|Y_{obs}, \theta)$, logo, os dados faltantes dependem apenas das informações registradas, disponíveis para análise e essas informações são correlacionadas com a variável que possui dados faltantes.

c) Com perdas não aleatórias de dados (NMAR)

Os dados são tidos como *NMAR (Not Missing At Random)*, ou seja, com perdas não aleatórias se, $\Pr(R = 0|Y_{obs}, Y_{mis}, \theta) \neq \Pr(R = 0|Y_{obs}, \theta)$. Nesta situação, a probabilidade de dados faltantes depende também das informações não registradas, ou seja, a probabilidade de dados faltantes varia de razões desconhecidas.

Uma vez declarado o tipo de banco de dados, pode-se “preencher” os dados faltantes utilizando o método de imputação mais adequado, levando em consideração o mecanismo que os dados faltantes obedecem.

2.2.3 Imputação única

A imputação única tem como objetivo principal preencher um valor a cada dado em falta, de forma única (RUBIN, 1987). Pode ser executado de uma maneira razoavelmente simples, porém tem desvantagens. As principais delas são: a superestimação e subestimação dos erros padrão das estimativas de precipitação obtidas pelas técnicas de imputação. Isso resulta em intervalos de confiança muito estreitos ou valores muito altos para a estatística do teste, já que os valores são preenchidos apenas uma vez, com o mesmo valor para todas as observações faltantes, desconsiderando a ideia de que outros valores poderiam ser incluídos na estimativa (GROVES et al., 2004).

Existem diversos métodos de imputação única, alguns deles são: pela média, via método do indicador, e exclusão *listwise*. Alguns estudos foram desenvolvidos com este tipo de imputação (NUNES et al., 2010; SALVADOR; SANTOS, 2010; GARCÍA-PEÑA; ARCINIEGAS-ALARCÓN; BARBIN, 2014; GANDOLFI, 2016).

2.2.4 Imputação múltipla

O método de imputação múltipla, proposta por Rubin (1987), é um tema que está em grande expansão dentro do estudo de imputação de dados. Na área de climatologia, estudos mostram que o método de imputação múltipla tem levado a resultados melhores que as técnicas já usadas, como krigagem e cokrigagem ordinária, apresentando um erro quadrático menor e estimativas melhores, quando se divide a área de estudos em zonas climaticamente homogêneas (CARVALHO et al., 2017). É um método que visa estimar dados faltantes, por meio de processos iterativos.

A imputação múltipla resume-se basicamente a três etapas (imputação, análise e agrupamento) (GANDOLFI, 2016), como segue de maneira ilustrada na Figura 1:

Figura 1. Principais etapas da imputação múltipla

Tais etapas são apresentadas a seguir:

(i) Imputação: esta etapa 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 novo banco de dados, chamado de banco de dados completo, sem presença de dados faltantes. O processo é feito m vezes, onde cada vez que são substituídos os valores em falta usa-se o mesmo banco de dados faltantes, ou seja, todos os bancos de dados completos desta etapa serão gerados por meio de um único banco de dados.

Esta primeira etapa exige maior cautela, pois é a partir dela que serão feitas as validações das etapas posteriores (GANDOLFI, 2016). É neste momento que é considerado o mecanismo da ausência de dados. Não é necessário que o modelo utilizado na etapa de imputação seja o mesmo modelo usado na etapa da análise (BARACHO, 2007).

(ii) Análise: esta 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 destas m estimativas serão feitas por meio de análises estatísticas tradicionais.

(iii) Agrupamento: após 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). Esta técnica é conhecida como “Regras de Rubin” (*Rubin rules*), podendo ser usada de forma independente da escolha do método da imputação múltipla (NUNES; KLÜCK; FACHEL, 2009).

A cada m estimativas geradas na etapa da análise é obtido um parâmetro $\bar{Q}$ de interesse, ou seja, $\hat{Q}_i$ tal que $i = 1, \dots, m$ é 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 esta estimativa é calculada da seguinte forma:

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \bar{Q}_i. \quad (16)$$

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

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m U_i, \quad (17)$$

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

E a variância entre as imputações:

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2. \quad (18)$$

Logo a variância combinada será:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B, \quad (19)$$

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

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

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

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

Existem alguns algoritmos para a imputação múltipla em banco de dados. 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 (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).

Os passos do algoritmo *MICE* em dados multivariados segundo Rubin (1987) são descritos a seguir:

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

1. Especificar um modelo de imputação $P(Y_j^{mis} | Y_j^{obs}, Y_{-j}, R)$, para a variável Y_j , e R é a matriz indicadora, já descrita no tópico de mecanismos associados ao banco de dados, tal matriz é a indicadora de resposta da variável Y_j , com $j = 1, 2, \dots, p$.

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

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

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

5. Define-se $Y_{-j}^t = (Y_1^t, \dots, Y_{j-1}^t, Y_{j+1}^{t-1}, \dots, Y_p^{t-1})$ 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 $Y_j^t \sim P(Y_j^{mis} | Y_j^{obs}, Y_{-j}^t, R, \phi_j^t)$.
8. Finaliza-se repetição j .
9. Finaliza-se repetição t .

2.2.4.1 PMM (Predictive Mean Matching)

O *PMM* é um método de imputação múltipla desenvolvida 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.

O método de médias preditivas (*PMM*) tem sido usado em vários trabalhos sobre imputação (KOLLER - MEINFELDER, 2009; NUNES; KLÜCK; FACHEL, 2009; MORRIS; WHITE; ROYSTON, 2014), visto que é um recurso interessante para banco de dados com variáveis semicontínuas, ou seja, variáveis que consistem em uma proporção de respostas com massas pontuais que são fixadas em algum valor e uma distribuição contínua entre as respostas restantes, e tal informação relaxa as suposições de alguns métodos como: imputação pelo método logito ordenado, regressão logística multinomial e pela análise de discriminantes, os quais exigem que a variável seja invariavelmente contínua, discreta, quantitativa ou qualitativa.

Para descrever o *PMM* é necessário que sejam feitas algumas notações preliminares: seja $Y = (Y_{obs}, Y_{mis})$ uma variável semicontínua incompleta com n unidades de amostra, onde Y_{obs} e Y_{mis} denotam os valores observados e os valores de omissão em Y , respectivamente. Além disso, $X = (X_1, \dots, X_j)$ é um conjunto de covariáveis totalmente observadas, onde X_{obs} e X_{mis} correspondem às partes observadas e ausentes em Y . Usamos notação n_{obs} para o número de unidades de amostra com valores observados de X e n_{mis} para o número de unidades de amostra com valores omissos. Finalmente, seja R um indicador de resposta que seja 1 se Y for observado e 0 se Y estiver faltando (descrito anteriormente) Deve-se notar que para esta pesquisa foi restrito ao caso univariado, já que o estudo aqui desenvolvido se utilizou apenas da variável precipitação pluvial para explorar os métodos de imputação,

embora os problemas pudessem ser abordados de forma iterativa para resolver o problema multivariado, sem perda de generalização.

Para imputar valores omissos em $\mathbf{Y}$ e para avaliar variâncias e intervalos de confiança para estimadores com base nos dados imputados, usou-se métodos de imputação múltiplas. Esses métodos podem ser descritos por meio de teorias bayesiana.

A imputação múltipla de $\mathbf{Y}_{mis}$ por meio de *PMM* é realizada pelo seguinte algoritmo:

1. Usa-se a regressão linear de $\mathbf{Y}_{mis}$ dado $\mathbf{X}_{obs}$ para estimar $\hat{\beta}, \hat{\sigma}$ e ε por meio de mínimos quadrados ordinários, com a seguinte regressão linear:

$$\mathbf{Y}_{mis} = \mathbf{X}_{obs}\beta + \varepsilon, \quad (20)$$

em que ε é um erro aleatório normalmente distribuído com esperança zero e variância σ , ou seja, $\varepsilon \sim N(0, \sigma)$.

2. Calcula-se $\sigma^{2*} = \hat{\varepsilon}^T \hat{\varepsilon} / A$, em que A é uma variável aleatória que segue a distribuição χ^2 (qui – quadrado) com $n_{obs} - r$ graus de liberdade, onde r é o número de preditores de $\mathbf{Y}$.

3. Calcula-se β^* de uma distribuição normal multivariada centrada em $\hat{\beta}$ com matriz de covariância $\sigma^{2*}(\mathbf{X}_{obs}^T \mathbf{X}_{obs})^{-1}$

4. Calcula-se $\hat{\mathbf{Y}}_{obs} = \mathbf{X}_{obs}\hat{\beta}$ e $\hat{\mathbf{Y}}_{mis} = \mathbf{X}_{mis}\beta^*$

5. Para cada $\hat{\mathbf{Y}}_{mis,i}$ é encontrada $\Delta = |\hat{\mathbf{Y}}_{obs} - \hat{\mathbf{Y}}_{mis,i}|$, em que $\hat{\mathbf{Y}}_{mis,i}$ é o valor estimado da i -ésima observação faltante no banco de dados

6. Faz-se a amostra aleatória de um valor de $(\Delta^1, \Delta^2, \Delta^3)$, tal que Δ^1, Δ^2 e Δ^3 são os três elementos menores de Δ , respectivamente, e tomar os $\mathbf{Y}_{obs,i}$ correspondentes, por meio da imputação.

7. Repete-se as etapas de 1 a 6 vezes, cada vez salvando o conjunto de dados concluído.

O padrão do *MICE* da função do pacote R executa a imputação múltipla ($m = 5$), m é quantidade de repetições do algoritmo. A função de regressão *mi.pmm* em *mi* também executa a imputação de *PMM* mas calcula $\Delta = \min|\hat{\mathbf{Y}}_{obs} - \hat{\mathbf{Y}}_{mis,i}|$ e seleciona os correspondentes $\mathbf{Y}_{obs,i}$ com a imputação.

2.2.4.2 Random forest

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

A ascensão do uso do método de *random forest* em imputação de dados vem do fato que em métodos tradicionais ocorrem problemas, como incapacidade de lidar com interações complexas, grande quantidade de dados e variáveis mistas (TANG; ISHWARAN, 2017). Segundo Shah *et al.* (2014) o método de *random forest* funciona razoavelmente bem sob ausência artificialmente introduzida em um conjunto de dados complexos.

O método de imputação usando *random forest* usa o procedimento de reamostragens de *bootstrap* para criar várias árvores de regressão, que nada mais são do que árvores de decisão, porém com resultados numéricos e não categóricos. O método de imputação usando *random forest* faz o procedimento de criar árvores de regressão para reduzir o risco de superestimação e combina os resultados de muitas árvores para produzir previsões mais precisas. Este algoritmo tem como objetivo prever com exatidão os valores ausentes individuais, em vez de obter desenhos aleatórios de uma distribuição, de forma que os valores imputados possam levar a estimativas de parâmetros enviesados em modelos estatísticos (SHAH *et al.*, 2014). Este procedimento foi utilizado em diversas áreas como, bioestatística (SOARES, 2014), manejo florestal (RUZA; PAULA; CORTE, 2016), e meteorologia (OLIVEIRA; OLIVEIRA; MONTEIRO, 2017).

2.2.4.3 Regressão linear via o método *bootstrap*

Este 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 que 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). Este método é utilizado em diferentes áreas como agricultura de precisão (DALPOSSO, 2017), engenharia de produção (RIZZO; CYMROT, 2006), e genética (CARLINI-GARCIA et al., 2001), entre outros.

A regressão linear é uma equação que tem por objetivo estimar o valor de uma variável dependente por meio de um conjunto de valores dados por variáveis independentes. Existem dois tipos de regressão linear: simples, em que existe apenas uma variável explicativa (independente) na equação, e múltipla, onde existe mais de uma variável independente.

A imputação de dados via regressão linear é um método que substitui os valores faltantes por valores que são estimados por meio de um modelo de regressão. As regressões utilizadas para imputação podem ser simples ou múltiplas, utilizando uma ou mais variáveis completas presentes na base de dados. Este método tem como pressuposição implícita que as suas variáveis sejam correlacionadas, já que a ideia da imputação por regressão é “completar” os valores faltantes por meio das variáveis completas do banco de dados, sendo assim, só faz sentido este procedimento se as variáveis forem correlacionadas (CASTRO, 2014).

Existem dois métodos de reamostragem com o *bootstrap* utilizados em modelos de regressão: o residual e o de pares. Descreve-se abaixo a caracterização destes métodos conforme (MARTINEZ-ESPINOSA; SANDANIELO; LOUZADA-NETO, 2006).

a) *Bootstrap* residual:

Passo 1) Ajustar um modelo de regressão considerando os dados da amostra original de tamanho n e obter os n resíduos ($\hat{\varepsilon}$) é um erro aleatório normalmente distribuído com esperança zero e variância σ , ou seja, $\hat{\varepsilon} \sim N(0, \sigma)$.

Passo 2) Selecionar uma amostra aleatória de tamanho n , dos resíduos ($\hat{\varepsilon}$) obtidos no passo anterior, utilizando reamostragem com reposição, com probabilidade $\frac{1}{n}$ para cada resíduo selecionado ($\hat{\varepsilon}^{BR}$).

Passo 3) Gerar novos valores de $\tilde{Y}$, por meio da seguinte equação:

$$\tilde{Y}^{BR} = X\tilde{A} + \hat{\varepsilon}^{BR}, \quad (21)$$

em que $\mathbf{x}$ é o vetor dos dados da amostra original, $\tilde{\varepsilon}^{BR}$ são os resíduos obtidos no passo 2, tal que, $\tilde{\mathbf{A}}$ é o vetor de parâmetros do modelo de regressão que foi estimado no passo 1.

Passo 4) Ajustar um modelo de regressão, pelo método dos mínimos quadrados, utilizando o seguinte modelo:

$$\tilde{\mathbf{Y}}^{BR} = \mathbf{X}\tilde{\mathbf{A}}^* + \tilde{\varepsilon}^{BR} \quad (22)$$

para obter as estimativas de $\tilde{\mathbf{A}}^*$ ($\hat{\tilde{\mathbf{A}}}^*$).

Passo 5) Repetir os passos 1 até 4 R vezes, em que R é o número de iterações do processo, geralmente fixado em 999 (DRAPER; SMITH, 1998). Após ter feito as R reamostragens, obtém-se a distribuição empírica do vetor ($\tilde{\mathbf{A}}^*$) e todas as estimativas relacionadas a esta distribuição, bem como média *bootstrap* residual e desvio padrão *bootstrap* residual de cada um dos vetores estimados, além de intervalo de confiança para ($\tilde{\mathbf{A}}^*$).

b) Bootstrap em pares:

Este processo é realizado por reamostragem dos pares originais dos valores (Y_i, x_i') originais, em que Y_i é a i -ésima observação e x_i' é a i -ésima linha da matriz $\mathbf{x}$.

Passo 1) Reamostrar com reposição n pares dos valores originais (Y_i, x_i') , com probabilidade $\frac{1}{n}$ para cada par selecionado, obtendo assim $(\tilde{\mathbf{Y}}^{BP}, \mathbf{X}^{BP})$ pares reamostrados.

Passo 2) Ajustar um modelo de regressão pelo método dos mínimos quadrados, considera-se o seguinte modelo:

$$\tilde{\mathbf{Y}}^{BP} = \mathbf{X}^{BP} \tilde{\mathbf{A}}^{**} + \tilde{\varepsilon}^{BP} \quad (23)$$

este modelo tem objetivo de obter a estimativa de $\tilde{\mathbf{A}}^{**}$ ($\hat{\tilde{\mathbf{A}}}^{**}$).

Passo 3) Repetir os passos anteriores R vezes.

De maneira semelhante ao método *bootstrap* residual, depois de R reamostragens, obtém-se uma distribuição empírica do vetor de parâmetros ($\tilde{\mathbf{A}}^{**}$) e todas as estimativas relacionadas a esta distribuição, bem como média *bootstrap*

residual e desvio padrão *bootstrap* residual de cada um dos vetores estimados, além de intervalo de confiança para ($\tilde{A}^{**}$).

Para este trabalho foi usado o método de *bootstrap* residual para ajustar o modelo de predição dos valores ausentes por meio da distribuição empírica para a etapa de imputação de dados. É possível perceber que para este artifício a variável é ajustada a um modelo de distribuição normal, uma vez que, este modelo foi implementado na primeira etapa do método *MICE*, executado no ambiente R pelo pacote “*MICE*” (BUUREN; GROOTHUIS-OUDSHOORN, 2011). O método de imputação de dados via regressão linear usando *bootstrap* é eficiente e rápido se os resíduos do modelo estiverem próximos da distribuição normal (BUUREN; GROOTHUIS-OUDSHOORN, 2011).

2.3 Comparação estatística 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 este valor mede quanto as observações do modelo de imputação são exatamente iguais aos valores reais. Porém, levar este índice em consideração de forma isolada pode causar interpretações demasiadamente rigorosas, levando a conclusões infundadas do ponto de vista prático. São indicadores comumente utilizados para avaliar o desempenho de métodos de imputação como os testados neste estudo, sendo descritos a seguir. São denominados também, de medidas de diferença.

2.3.1 Raiz do erro quadrático médio (REQM)

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 raiz do erro quadrático médio (*REQM*), são mais usadas na comparação e avaliação de modelos de simulação. Para alguns casos esta medida

substitui os índices baseados em correlação e habilidade como as medidas primordiais de precisão, fato importante, porque, as medidas baseadas em correlação e habilidade não estão consistentemente relacionadas à precisão do modelo, o que sugere que a aplicação dos índices de diferença é efetiva a uma ampla variedade de modelos e dados geofísicos. Esta medida é definida como:

$$REQM = \sqrt{\frac{\sum_{j=1}^n (p_j - o_j)^2}{n}}, \quad (24)$$

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.

2.3.2 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). Este 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. Este valor é definido pela seguinte equação:

$$d = 1 - \frac{\sum_{j=1}^N |d_j|^2}{\sum_{j=1}^N (|p_j - \bar{o}| + |o_j - \bar{o}|)^2}, \quad (25)$$

em que $\bar{o}$ é a média dos valores observados, ou seja, $\bar{o} = \frac{\sum_{j=1}^n o_j}{n}$ e $d_i = p_i - o_i$.

2.3.3 Coeficiente de correlação de Pearson (r)

O coeficiente de correlação de Pearson (r), é um número que nos permite avaliar quanto duas variáveis são correlacionadas. 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 este valor é igual a 1 é dito que ocorre uma correlação positiva, logo, uma variável é proporcionalmente relacionada uma com a outra. Este coeficiente é calculado da seguinte maneira:

$$r = \frac{\left(\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \right)}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) \left(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right)}} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \quad (26)$$

em que X e Y são as variáveis que se deseja saber a correlação, assim, $X = \{x_1, \dots, x_n\}$ e $Y = \{y_1, \dots, y_n\}$, e n é a quantidade de observações das variáveis.

Para interpretar os valores do coeficiente de correlação de Pearson, foram elaboradas algumas classificações: sabe-se que valores extremos (0, -1 e 1) de r são muito difíceis de serem encontrados na realidade, valores -1 e 1 significam uma correlação perfeita, ou seja, indica que pode-se determinar exatamente o escore de uma variável se o escore da outra for conhecido; para o valor de $r = 0$ tem-se que não existe correlação alguma entre as variáveis, ambas as situações são praticamente inviáveis na realidade. Para discutir este coeficiente Dancey e Reidy (2017) apontam uma classificação: intervalos são definidos como força da correlação entre as variáveis, quanto mais próximo de um, a correlação é considerada forte a perfeita, para os valores próximos a zero a correlação é considerada fraca ou correlação zero. Estes intervalos são apresentados por meio da Figura 2:

Perfeita	+1	-1
Forte	+0.9	-0.9
	+0.8	-0.8
	+0.7	-0.7
Moderada	+0.6	-0.6
	+0.5	-0.5
	+0.4	-0.4
Fraca	+0.3	-0.3
	+0.2	-0.2
	+0.1	-0.1
zero	0	

Figura 2. Classificação das correlações de Pearson, segundo Dancey e Reidy (2017)

Vale ressaltar que esta medida pode auxiliar na tomada da decisão, mas não deve ser a base da conclusão do acordo entre os métodos, afinal este coeficiente pode ser enganoso quando avalia-se a acurácia ou exatidão dos valores resultantes do modelo, pois ele evidencia apenas o grau de associação linear entre duas variáveis, e não a concordância entre eles (OLIVEIRA, 2016).

2.3.4 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 estes dois coeficientes da seguinte forma:

$$c = r \cdot d. \quad (27)$$

O critério para interpretar o índice de desempenho c dos métodos estudados em relação ao método padrão é apresentado na Tabela 2.

Tabela 2. Critério de interpretação do desempenho dos métodos pelo índice (c) Camargo e Sentelhas (1997)

Valor de c	Desempenho
$> 0,85$	Ótimo
$0,75 - 0,85$	Muito bom
$0,65 - 0,75$	Bom
$0,60 - 0,65$	Mediano
$0,50 - 0,60$	Sofrível
$0,40 - 0,50$	Mau
$\leq 0,40$	Péssimo

2.4 Homogeneidade e imputação de dados climáticos, com ênfase em precipitação no Brasil

Peterson *et al.* (1992) afirmam que desde a década de 1980 tem havido um grande esforço dos climatologistas para desenvolver técnicas de identificação de falta de homogeneidade em séries climatológicas e, a partir disso, ajustá-las para proceder as análises apropriadas das mesmas e compensar viés decorrente. Esses citados autores e AGUILAR *et al.*, (2003) elaboraram extensas revisões, que discutem os testes de detecção da falta de homogeneidade e, também, procedimentos associados, como análise de fontes introdutoras dessa falta, metodologias diretas e indiretas para testar o grau de homogeneidade e sua aplicação em várias regiões do mundo, controle da qualidade dos dados. Discutem, também, controle de qualidade dos dados. Citam 14 técnicas de detecção de falta de homogeneidade, das quais três serão utilizadas no presente estudo.

No Brasil há estudos desse tipo aplicados à série de precipitação pluvial de Piracicaba de 1917a 1989 (ASSIS; VILLA NOVA, 1995), de Urussanga-SC (BACK, 2001), de 63 estações meteorológicas do Rio Grande do Sul (BURIOL *et al.*, 2012) e para temperaturas extremas de 24 estações de Minas Gerais (SANTOS, 2012).

A climatologia é uma área de estudos na qual a imputação está bem presente na atualidade, acompanhando a explosão de técnicas e métodos a partir de cerca de 40 anos, nas múltiplas áreas das atividades humanas onde a coleta de dados se faz presente; (BERTSIMAS; PAWLOWSKI; ZHUO, 2018) citam 19 métodos enquadrados em 8 categorias. O trabalho de Paulhus e Kohler (1952) sobre uso de interpolação em séries de chuva de estações americanas, enquadrado na categoria de uso de médias, é assumido como o iniciante da imputação de dados em Climatologia. A imputação de dados climatológicos está bem disseminada. No Brasil, podem ser citados os estudos descritos a seguir.

Ruiz-Cárdenas e Krainski (2011), para valores diários de temperaturas extremas e de precipitação de 41 estações do Paraná, usaram abordagens baseadas no algoritmo de maximização da esperança (EM) e metodologia de análise de componentes principais, que demonstraram desempenho superior ao método de coeficiente de correlação modificado com a ponderação do inverso da distância.

García-Peña, Arciniegas-Alarcón e Barbin (2014) utilizaram a imputação na série de dados em escala mensal da estação meteorológica do campus Luiz de Queiroz, usando um conjunto univariado de precipitação mensal de 1917 a 2012 e outro conjunto de matriz multivariada (radiação global, umidade relativa média, amplitude térmica e precipitação), empregando quatro métodos de imputação baseados na decomposição por valores singulares (DVS) e comparação entre eles pela raiz quadrada normalizada do erro quadrático médio, a estatística de similaridade de Proustes e o coeficiente de correlação de Spearman. Para a matriz de precipitação, os autores concluíram que a imputação pela média mensal apresentou melhor desempenho que os métodos baseados na DVS.

Com o interesse de modelar a seca agrícola para a cultura de soja no Paraná, FERRARI; OZAKI (2014) testaram três métodos de imputação da precipitação diária para dados faltantes, num conjunto de dados de 484 estações meteorológicas: o método imputação pela estação vizinha mais próxima, o método de ponderação pelo inverso da distância e o de regressão linear. O segundo desses métodos apresentou os melhores resultados e foi usado para completar o banco de dados nas estações, após o que no passo seguinte foram usados em um processo de controle de qualidade para identificar possíveis erros, para depois ser empregada a distribuição de valores extremos generalizada para se chegar à probabilidade estimada de ocorrência do número máximo com precipitação abaixo de 7 mm, em período diário, entre janeiro e fevereiro, considerado como valor crítico para o enchimento de grãos.

Carvalho *et al.* (2017) utilizaram imputação múltipla com cadeias encadeadas para preenchimento de falhas na escala diária de precipitação de estações meteorológicas em duas zonas homogêneas, uma com 190 estações nos estados de São Paulo e Mato Grosso do Sul e outra com 90 estações em região de clima árido do nordeste brasileiro. Ao compararem os dados de imputação com técnicas krigagem e co-krigagem ordinárias, com altitude como variável auxiliar, verificaram que a imputação múltipla foi melhor que as técnicas geoestatísticas em ambas as regiões.

Oliveira e Santos (2017) utilizaram séries de precipitação relativamente curtas (1964-1978) de 30 localidades na Bahia e usaram a metodologia de Paulhus e Kohler (1952), com uso de dois métodos para preencher falhas, o primeiro com uso da média de três estações próximas e o segundo da média ponderada, em que

a ponderação representa a razão entre a normal anual da chuva na estação alvo e a normal anual de cada estação vizinha. Usaram como indicativo do comportamento dos métodos o valor do que denominam de precisão em relação aos valores medidos. Concluíram que os dois métodos se diferenciaram, mas não usaram especificamente teste estatístico para tal, discutindo as diferenças em termos de uma certa restrição de uso para o primeiro desses métodos, por ser ele aplicável quando a precipitação anual de cada posto vizinho diferiu no máximo 10% daquela da estação alvo.

3. MATERIAL E MÉTODOS

3.1 Dados meteorológicos

Os dados foram obtidos no posto meteorológico da ESALQ - USP, situado no Campus “Luiz de Queiroz”, em dependências do LEB – Departamento de Engenharia de Biosistemas da ESALQ – USP, na cidade de Piracicaba – SP, sendo atualmente composto por duas estações: uma convencional e outra automática. São coletados atualmente os seguintes elementos meteorológicos: precipitação pluvial, temperatura e umidade relativa do ar, evaporação de superfície livre de água (evaporação de tanques), radiação solar e horas de brilho solar, velocidade e direção do vento.

O posto iniciou suas atividades em 1º de junho de 1902, um ano após a inauguração da então Escola Prática Agrícola, hoje a ESALQ (ANGELOCCI; SENTELHAS, 2017), sendo então composto apenas pela estação meteorológica convencional, localizada na latitude de 22°42'S, longitude de 47°38'W e altitude de 546 m, sendo transferida em outubro de 1955 para local 1140 m distante em linha reta, com as coordenadas 22°42' S e 47°37' W, altitude de 566 m. A estação automática foi instalada mais recentemente em 1997. É importante ressaltar que os dados observados encontram-se disponíveis no *site* do LEB (USP, 2018) e tem servido de apoio didático e para inúmeros trabalhos de pesquisa.

O elemento meteorológico selecionado para este estudo foi a altura pluviométrica, em escala mensal, ou seja, foram somadas as precipitações pluviais diárias dos seus respectivos meses. Esse estudo ficou restrito ao período de 1917 a 1997, pois os dados registrados entre 1902 a 1916 foram recentemente descobertos e passam por análise dos responsáveis pelo posto e pela digitalização destas observações. O período de 1998 até o presente momento apresenta uma outra dificuldade, pois com a instalação da estação automática, entendeu-se não ser mais necessário fazer as observações na estação convencional dos dados diários de feriados e finais de semana, sendo assim, essas informações foram “imputadas” pelas observações da estação automática, logo, nos dias que não foram feitas as observações da estação convencional, o *site* foi completado pelas observações correspondentes da estação automática, criando-se no *site* dois arquivos a partir de 1997. O primeiro continuou a série da estação convencional, no qual cerca de

metade dos dias foram introduzido dados de medidas com o pluviômetro de leitura direta e a outra metade com os valores diários obtidos a partir do registro a cada 15 minutos do pluviômetro de báscula.

Além da mudança do local, o uso de diferentes modelos de pluviômetros, do tipo de leitura direta (*Ville de Paris* até o início da década de 1970 e HH a seguir daí) e o de cubas basculantes usado na estação automática, podem levar a problemas de heterogeneidade na série. Nas últimas duas décadas, com a disseminação de instrumentos automatizados, há muitos trabalhos de comparação entre medidas simultâneas pelos dois tipos de coleta de dados, a convencional e a automatizada. No Brasil, podem ser citados SENTELHAS et al., 1997; FISCH; SANTOS, 2002; SOUZA; GALVANI; ASSUNÇÃO, 2002; ALMEIDA; SOUZA; ALCANTARA, 2008; ALMEIDA; HERMENEGILDO, 2013; CARVALHO et al., 2017; RIBEIRO; TESTEZLAF; S FERRAREZI, 2018. Não há concordância quanto ao desempenho relativo entre os métodos de medida de chuva, tendo a *OMM - World Meteorological Organization WMO* (2014) feito um alerta quanto aos problemas apresentados pelo pluviômetro de báscula. Em Sentelhas et. al. (1997) foram comparados os dados de 13 meses de chuva medidos por pluviômetro simples, pluviógrafo de boia e de báscula e verificaram que este último apresentou tendência de subestimar a chuva, principalmente com precipitações superiores a 20 mm ao dia, sendo que para o período estudado houve uma diferença negativa de 145,1 mm, importantes para estudos climáticos. Posteriormente, Sentelhas e Caramori (2002) fizeram comparação entre os dados de cerca de 5 anos de dois pluviômetros de báscula com resoluções diferentes (0,1 e 0,2 mm) em relação aos convencionais, atentando para o mesmo problema de subestimação dos de báscula, dependendo inclusive da capacidade desta última, Essa tendência, já observada por outros autores, faz com que a *OMM - World Meteorological Organization WMO* (2014) alerte quanto aos problemas desse tipo de pluviômetro.

Os problemas causados pela mistura de observações de origem convencional e automática desde 1997 fazem parte de projeto atualmente em andamento no Departamento de Engenharia de Biossistemas da ESALQ, de forma a reconstruir a série convencional a partir desse ano para todas as variáveis medidas. As séries de temperatura do ar, radiação solar e horas de brilho solar já têm essa reconstrução praticamente finalizada e no momento trabalha-se com umidade do ar e precipitação.

3.2 Tratamento dos dados

As observações de precipitação pluvial em escala diária foram transferidas do *site* do posto meteorológico da ESALQ. Estas informações foram tabeladas e organizadas pelas variáveis: quantidade de observações, ano, mês, dia e precipitação pluvial em milímetros, do ano de 1917 a 2018, formando o banco de dados denominado “Registro da estação”, o qual contabilizava 36948 observações de precipitação pluvial em escala diária, distribuídas neste período.

Após a seleção e de acordo com a variável ano, o período de 1998 a 2018, foi retirado, devido ao fato de ter sido “preenchido” pela estação automática do posto meteorológico, este período totalizava 7363 observações. Em seguida, foram verificados os *NAs* do período de 1917 a 1997, apenas 14 observações, um valor muito baixo para a execução dos métodos de imputação de dados, principalmente para os de imputação múltipla, que se validam por meio de processos iterativos de valores ausentes com valores completos. Por este motivo, estas 14 observações foram desconsideradas, constituindo o banco de dados denominado “Banco de dados original”, com o total de 29571 informações.

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, uma vez que todos os valores que foram retirados de forma aleatória para este estudo eram conhecidos. Após o desenvolvimento dos métodos imputações pôde ser considerado exato o erro entre os métodos, tendo com isso um resultado ainda mais preciso destes métodos.

Após a exclusão de todos os valores de informações faltantes e períodos tendenciosos, foram somadas todas as precipitações pluviais diárias de acordo com a variável mês a qual cada precipitação estava relacionada, assim, este banco de dados contabilizou 972 informações e foi denominado como “Banco de dados mensais”.

Posteriormente a classificação dos testes de homogeneidade, executados com as 972 informações do “Banco de dados mensais”, foi criado o “Banco de dados homogêneos”.

Retiradas aleatórias da variável precipitação foram feitas, nas intensidades 5%, 10% e 15%, assim, a classificação dos dados faltantes é *MCAR – Missing Completly At Random*, afinal, 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, gerando-se assim, os bancos de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” com o total de 972 informações. Deste total as respectivas intensidades do nome de cada um desses bancos de dados representaram as informações faltantes geradas de maneira aleatória. Por exemplo, o banco de dados “Amostra 5%” tem 972 observações de chuva. Porém, 49 delas, ou seja, 5% eram informações que foram retiradas de maneira aleatória deste banco de dados, e como se tinha uma totalidade de valores conhecidos no “Banco de dados homogêneos”, estas 49 observações constituíram o banco de dados “Retirados 5%”, que eram, totalmente conhecido, fato importante para a comparação entre os métodos de imputação.

De modo análogo foi construído o banco de dados “Amostra 10%” e “Amostra 15%” com a diferença de que os valores omissos eram de 97 e 146, respectivamente para estes bancos de dados, valores que constituem os bancos de dados “Retirados 10%” e “Retirados 15%”. Vale ressaltar que essas retiradas aleatórias são feitas com reposição, para que só assim pudesse ser criado um novo banco de dados, ou seja, após ser feita a comparação entre os métodos de imputação com os bancos de dados “Amostra 5%”, levando em consideração o banco de dados “Retirados 5%”. Os valores usados como referência deste processo, eram “devolvidos” para o “Banco de dados homogêneos” e só então é feito um novo procedimento para a comparação entre os métodos de imputação para “Amostra 10%” levando em consideração os valores pertencentes ao banco de dados “Retirados 10%”. Replicou-se esse procedimento da mesma forma para a intensidade de 15%. É importante esclarecer que todos os bancos de dados “Amostra” tinham o mesmo número de informações (972). A única mudança eram os respectivos valores omissos em cada uma, de acordo com sua denominação.

Foram usados três métodos de imputação de dados e a cada execução foi criado um novo banco de dados denominado “PMM”, “*Random forest*” e “Regressão” em suas respectivas intensidades de informações faltantes (5%, 10% e 15%), cada banco de dados tinha 972 informações e já contava com as precipitações imputadas.

A Figura 3 contém o esquema do tratamento dos dados desenvolvido nesta pesquisa.

Figura 3. Organograma dos bancos de dados e seu uso nos diferentes métodos de imputação

3.3 Análise descritiva preliminar

Foi feita uma análise descritiva das observações contidas no banco de dados original, por meio de construção dos gráficos de precipitações pluviais anuais e mensais, com o objetivo de avaliar como a variável se comportou no período estudado.

3.4 Homogeneidade da série

Os dados de chuva em escala mensal foram submetidos a três testes de homogeneidade: Pettitt, Buishand e normal padrão (*SNHT*), conforme descritos na revisão de literatura, visto que o posto teve uma mudança de local no ano de 1955 e contou também com mudanças no modelo de pluviômetro. Após a execução de cada teste, a série foi classificada de acordo com o número de testes que rejeitam a hipótese nula, logo, a quantidade de testes que rejeitam a hipótese da série ser homogênea define a classe de homogeneidade, então, a série pode ser entendida como “útil”, “duvidosa” ou “suspeita”.

3.5 Métodos de imputação

Foram avaliados neste trabalho três métodos de imputação: *PMM* (*Predictive Mean Matching*), *random forest* e imputação via regressão linear usando método *bootstrap*. Nesta etapa foi utilizado o pacote *Multivariate Imputation by Chained Equation* (*MICE*) no *software R* (BUUREN; GROOTHUIS-OUDSHOORN, 2011).

A escolha dos métodos foi feita de acordo com a praticidade dos usuários de imputação de dados pois, sabe-se que o *PMM* é o *default* do pacote *MICE*, logo, é interessante testá-lo para a imputação de precipitação pluvial. O *random forest* é um método que usa em seu desenvolvimento árvores de decisão, estando relacionado com o assunto de aprendizado de máquina muito discutido atualmente, e cuja evolução dá maior praticidade aos seus usuários. Já a regressão linear via método *bootstrap* foi implementado pelo fato de ser um método que aplica a regressão linear, empregada em vários estudos com imputação de dados, como já foi citado anteriormente.

Foram realizadas 100 simulações para cada banco de dados denominados “Amostra”. Gráficos de dispersão, *boxplot* e de distribuição empírica entre os valores imputados e observados foram construídos para cada um dos métodos pesquisados.

3.6 Comparação estatística dos métodos de imputação

A comparação foi feita por meio do cálculo e análise da raiz do erro quadrático médio (REQM), índices de acurácia de Willmott (d) e de desempenho (c), adotando os intervalos para a classificação de cada indicador de comparação.

4. RESULTADOS E DISCUSSÃO

Os resultados principais deste estudo são que a série histórica de chuva da ESALQ foi classificada como Classe 1, ou seja, nenhum sinal claro de falta de homogeneidade é aparente e, o método de imputação que teve os menores erros, os maiores índices de desenvolvimento e de acurácia foi o *PMM*.

4.1 Análise descritiva preliminar

As análises descritivas dos dados são resultados da soma mensal e soma anual das precipitações. A Figura 4 contém os resultados obtidos por meio dos cálculos de somas anuais das precipitações. O valor médio das precipitações anuais é de 1277,4 mm (linha vermelha), com o valor mínimo de 812 mm no ano de 1921, e o valor máximo de 2018 mm no ano de 1983. Pode-se observar também a linha de tendência crescente (reta azul), uma reta resultante de uma regressão linear simples das informações disponíveis do período estudado (1917 a 1997).

Figura 4. Alturas pluviométricas anuais da estação meteorológica da ESALQ no período de 1917 a 1997

A Figura 5 contém 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 (NETO, 2015), 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 nestes valores, o que é totalmente concordante com estudos de regime hídrico no estado de São Paulo.

Figura 5. Variação da altura pluviométrica média mensal da estação meteorológica da ESALQ no período de 1917 a 1997

4.2 Homogeneidade da série

Os resultados apresentados neste tópico serão das análises exploratórias realizadas a partir dos três testes de homogeneidade empregados neste trabalho. Será iniciado com os resultados do teste de Pettitt, seguido pelo teste de Buishand, teste *SNHT* e finalizando com a classificação de homogeneidade da série.

Para ilustrar a análise exploratória pelo teste de Pettitt é apresentada, na Figura 6, a qual contém o gráfico da estatística K_T referente ao período estudado. Não foi verificado nenhuma quebra da série histórica em questão, já que toda a sua extensão se concentrou entre os limites críticos do intervalo de confiança bilateral de 95% de confiança, estabelecidos pelos pontos críticos da estatística K_T . Sendo os valores $K_{crit} = \pm 17329$, tais limites foram representados na Figura 6 pelas linhas

horizontais. Os meses que mais se aproximaram do limite superior foram setembro de 1981 com $K_T = 17312$, setembro e outubro de 1969 com $K_T = 17099$ e 17096 , respectivamente. O p_{oa} foi de 0,28, valor superior que 0,05, ou seja, para o teste de Pettitt não se rejeita a hipótese nula, e, portanto, não existe mudança, ou ponto de “quebra”, na série analisada neste estudo, ou seja, por este teste há homogeneidade na série.

Figura 6. Estatística K_T , do teste de Pettitt para precipitação pluvial em escala mensal da estação meteorológica da ESALQ, nos anos de 1917 a 1997.

A estatística $R/\sqrt{T}$ pelo teste de Buishand referente ao período estudado é apresentada pela Figura 7. Segundo o resultado deste teste, pode-se dizer que não foi verificada quebra na série histórica em questão, já que toda a sua extensão se concentrou abaixo do valor crítico de $R/\sqrt{T} = 1,75$, valor encontrado na Tabela 1. O p-valor do teste de Buishand é de 0,75, o que significa que o modelo não detectou mudança na quantidade da média de precipitação, ou seja, não se rejeita a hipótese nula ($\Delta = 0$), ou seja, há homogeneidade na série.

Figura 7. Estatística $R/\sqrt{T}$, do teste de Buishand para precipitação pluvial em escala mensal da estação meteorológica da ESALQ, nos anos de 1917 a 1997.

Para o teste *SNHT* foi realizada a análise exploratória de modo análogo aos testes anteriores. Para este teste é apresentada a Figura 8, que contém o gráfico da estatística T_v referente ao período em questão (1917 a 1997). Segundo a análise deste teste, pode-se dizer que não foi verificado quebra na série histórica, já que toda a sua extensão se concentrou abaixo do valor crítico de $T_0 = 5,48$, em que, T_0 é o valor crítico com o índice de 95% de confiança. O p-valor do teste *SNHT* é de 0,45, e como este valor é maior que 0,05, não se rejeita a hipótese nula, o que significa que Z tem uma distribuição normal com média zero e desvio padrão um, apesar do mês de setembro de 1981 ter apresentado um valor de $T_v = 5,47$, valor este muito próximo do valor crítico, ele não é considerado um mês de quebra significativa.

Figura 8. Estatística T_v , do teste *SNHT* para precipitação pluvial em escala mensal da estação meteorológica da ESALQ, nos anos de 1917 a 1997.

Todos os testes de homogeneidade feitos neste estudo aceitaram a hipótese nula com o nível de significância de 5%. Como já foi descrito anteriormente, a série pode ser então, classificada como Classe 1: “útil”, ou seja, nenhum sinal claro de falta de homogeneidade é aparente. A série parece ser suficientemente homogênea para análise de tendências e análise de variância, afinal, nenhum teste rejeitou a hipótese nula. Por meio desta classificação pode ser entendido que o ano de 1955 não teve uma quebra significativa, apesar de se saber que houve uma mudança de local da estação meteorológica estudada, esta mudança não comprometeu significativamente a homogeneidade da série histórica aqui analisada.

4.3 Métodos de imputação

As técnicas de imputação foram avaliadas em diferentes percentuais de dados faltantes (5%, 10% e 15%). Para cada etapa foram feitas 100 repetições, os resultados são uma média aritmética destas simulações. Apenas para efeito de organização, os resultados e discussões das análises exploratórias para os métodos de imputação serão separadas em diferentes intensidades amostrais de dados faltantes em cada banco de dados “Amostra”. Serão apresentados os gráficos de dispersão, gráficos de caixa (*boxplot*) e gráficos de distribuição empírica das variáveis imputadas e observadas para ilustrar os resultados encontrados nesta pesquisa.

4.3.1 PMM

A Figura 9 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 a precipitação imputada é igual a 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 9. 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.

A Figura 10 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. 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, esta estatística da variável imputada é menor que a da variável observada, o que implica dizer que, a variável imputada se concentrou mais entre os valores 0 a 100 mm do que a observada. Estas constatações se reforçam na Figura 11, que contém os gráficos de densidade empírica das variáveis.

Figura 10. 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.

Figura 11. Gráficos de distribuição empírica 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.

4.3.2 *Random forest*

Para análise exploratória do método *random forest* são apresentadas as Figura 12, Figura 13 e Figura 14, respectivamente. A primeira contém os gráficos de dispersão da variável precipitação pluvial mensal imputada *versus* precipitação pluvial observada com 49 (5%), 97 (10%) e 146 (15%) de observações faltantes respectivamente, ambas em milímetros (mm) e a linha diagonal representa a reta $y=x$. Pode-se verificar que para este método as informações imputadas apresentaram um comportamento muito semelhante ao que foi apresentado no método *PMM*, ou seja, é notório um padrão relacionado entre as variáveis.

Figura 12. 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.

Para representar as diferenças entre as distribuições da variável precipitação pluvial imputada pelo método *random forest* e a observada estão contidas nas seguintes figuras: Figura 13 (gráficos de caixa) e Figura 14 (gráficos de distribuição empírica). Nos gráficos *boxplots* das respectivas variáveis, pode-se notar que para o banco de dados “Amostra 5%” o intervalo interquartil da variável imputada é maior do que a variável precipitação observada. Este 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 este cenário a variabilidade das informações observadas é maior do que a da variável imputada.

Figura 13. 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.

A Figura 14 contém os gráficos de distribuição empírica das distribuições das variáveis imputada e observada, para o banco de dados “Amostra 5%”. Pode-se observar que há uma grande quantidade de valores entre 0 a 200 milímetros, enquanto para variável imputada os valores são mais dispersos entre 0 a 300 milímetros. Estes resultados indicam que este método de imputação não teve uma boa aproximação dos dados observados, principalmente para o cenário de 5%, com uma variabilidade maior da variável imputada do que a variável observada.

Figura 14. Gráficos de distribuição empírica 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.

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 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, porém tal estudo se limita a precipitação pluvial em escala diária, escala desconsiderada neste estudo, uma vez que o uso da escala diária gera inconvenientes já citados anteriormente, que podem causar sérias dificuldades para a análise da série (RUIZ-CÁRDENAS; KRAINSKI, 2011), assim, o método *random forest* para este estudo não apresentou uma boa aproximação das informações imputadas com as observadas.

4.3.3 Regressão linear via método *bootstrap*

Os gráficos de dispersão do banco de dados “Amostra 5%”, “Amostra 10%” e “Amostra 15%” da variável precipitação pluvial imputada pelo método da regressão linear via método *bootstrap* versus precipitação observadas são representados na Figura 15. Embora este método tenha imputado valores negativos para precipitação pluvial nas três intensidades amostrais de informações retiradas, valores obviamente discrepantes com a realidade, o método sugere uma concentração maior ao redor da reta $y=x$ do que os métodos já discutidos, ou seja, para as precipitações imputadas maiores ou iguais a zero este método presume assumir valores mais

próximos dos valores observados, pressuposição que será efetivamente válida após a comparação entre os métodos de imputação.

Figura 15. 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.

Para ilustrar a diferença entre as distribuições da variável precipitação imputada pelo método de regressão linear via método *bootstrap* e precipitação observada, as Figuras 16 e 17 são apresentadas. Nota-se que para os cenários de intensidades amostrais 10% e 15% de informações faltantes, tanto o primeiro quartil quanto a mediana tiveram valores próximos entre as variáveis. Deste modo, pode ser dito que 50% das informações estão concentradas em intervalos semelhantes.

Figura 16. 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.

Ao analisar a Figura 17 pode-se perceber com maior clareza a diferença entre as distribuições das variáveis, podendo ressaltar a maior quantidade de valores contidos no intervalo de 0 a 200 mm para variável observada do que para a variável imputada pelo método de regressão linear via método *bootstrap*. Tal evento é decorrente do fato que para este artifício variável imputada é ajustada a distribuição da variável observada. Logo, ao imputar uma grande quantidade de valores entre 0 e 200 mm, a distribuição empírica desta variável acaba estendendo suas imputações para valores menores que zero.

Figura 17. Gráficos de distribuição empírica da variável precipitação pluvial mensal imputada pelo método de 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.

4.4 Comparação estatística 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 e as avaliações dos métodos de imputação foram feitas de acordo com os bancos de dados “Amostra”, nas três intensidades amostrais. Como foram realizadas 100 repetições para cada método nos três cenários amostrais os resultados obtidos são médias aritméticas de cada procedimento.

4.4.1 Raiz do erro quadrático médio

A Figura 18 contém o gráfico de linhas dos resultados da raiz do erro quadrático médio dos três métodos de imputação, em seus respectivos bancos de dados “Amostra”. É importante que os valores deste indicador de comparação sejam

mais próximos de 0, afinal, o valor de *REQM* igual a 0 significa que o estimador prevê observações com precisão perfeita. Todos os valores deste gráfico (Figura 18) podem ser observados por meio da Tabela 3. Percebe-se então que o método de imputação *PMM* indicou os menores valores. Pode-se observar também que para o cenário de 10% de dados faltantes, o resultado foi ainda menor para os três métodos de imputação, ou seja, para o banco de dados “Amostra 10%” os métodos tiveram melhor desempenho para o *REQM*. O método que resultou em valores mais distantes de 0 foi o método *random forest*. Para este indicador de comparação, foi o que apresentou erros maiores, logo, foi o método menos adequado para este estudo.

Os métodos *PMM* e regressão linear via método *bootstrap* além comportamentos semelhantes entre os valores de *REQM*, registraram, valores próximos, o que já era esperado, uma vez que os desenvolvimentos de ambos os métodos de imputação são semelhantes. Eles utilizam o artifício de ajustar uma regressão linear com base nos dados originais e já era esperado também, que entre eles o *PMM* tivesse um desempenho melhor ou igual ao de regressão, por conta que o *PMM* encontra o menor valor entre a diferença do valor imputado estimado e o valor observado. Não foram encontrados fatores que acarretaram um desenvolvimento não satisfatório para o *random forest*, uma vez que, este método tem natureza não-paramétrica, sendo um dos algoritmos de aprendizado de máquina com alta taxa de acurácia.

Figura 18. Gráfico de linhas 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 de dados faltantes de 5%, 10% e 15%.

Tabela 3. 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

Métodos	Porcentagem de Dados Faltantes		
	5%	10%	15%
<i>PMM</i>	88,1	87,2	88,7
<i>Random Forest</i>	104,2	103,2	104,3
Regressão Linear Via Método <i>Bootstrap</i>	88,8	104,3	90,0

4.4.2 Índice de acurácia de Willmott (d)

A Figura 19 contém o gráfico de linhas dos resultados do índice de acurácia de Willmott (d) dos três métodos de imputação, em seus respectivos bancos de dados “Amostra”. É importante ressaltar que os valores deste indicador de comparação sejam mais próximos de 1. Todos os valores representados na Figura 19 estão descritos na Tabela 4. Pode-se perceber 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 5% de falta o método *PMM* e o método de regressão resultaram no mesmo valor 0,71. Assim como no *REQM*, os três métodos de imputação tiveram os mesmos comportamentos entre si, ambos tiveram um desempenho menor no cenário de 5% de dados faltantes, enquanto para o de 10% tiveram o melhor valor 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).

Figura 19. Gráfico de linhas do índice de acurácia de Willmott (*d*) da precipitação pluvial imputada por meio dos métodos: *PMM*, *random forest* e regressão linear via método *bootstrap*, nas porcentagens de dados faltantes de 5%, 10% e 15%

Tabela 4. Valores do índice de acurácia de 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

Métodos	Porcentagem de Dados Faltantes		
	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

4.4.3 Índice de desempenho (c)

O índice de desempenho (c) é obtido por meio do produto do índice d e o coeficiente de correlação de Pearson (r). Como os valores do índice de acurácia de Willmott foram apresentados no tópico anterior, este tópico será reservado para discussão do coeficiente de correlação de Pearson e o próprio índice de desempenho (c).

A Figura 20 contém o gráfico de linhas dos resultados do índice de correlação de Pearson (r) dos três métodos de imputação, em seus respectivos bancos de dados “Amostra”. Os valores próximos de 1 para este indicador de comparação são os valores que indicam uma correlação de maior magnitude. Todos os valores contidos na Figura 20 estão descritos na Tabela 5. O *PMM*, *random forest* e regressão linear via método *bootstrap*, tiveram comportamentos semelhantes aos outros indicadores de comparação, ou seja, menores valores para o banco de dados “Amostra 5%”, um aumento para “Amostra 10%” e uma pequena queda para “Amostra 15%”. O *PMM* teve os maiores coeficientes, resultando em valores iguais apenas com o método de regressão no banco de dados “Amostra 15%” com $r = 0,52$.

Os coeficientes de correlação de Pearson foram considerados de correlação moderada entre as variáveis, para todos os cenários de observações faltantes. Para o método de *random forest* estes valores foram considerados de magnitude moderada, uma vez que se aproximam mais de 0,4 do que 0,3 de acordo com arredondamento. Estes resultados já eram esperados, uma vez que, as Figuras 9, 10 e 15 não indicavam uma linearidade entre as variáveis, ou seja, era possível notar uma baixa dependência linear entre elas.

Figura 20. Gráfico de linhas do coeficiente de correlação de Pearson (r) da precipitação pluvial imputada por meio dos métodos: *PMM*, *random forest* e regressão linear via método *bootstrap*, nas porcentagens de dados faltantes de 5%, 10% e 15%

Tabela 5. Valores dos coeficientes de correlação de Pearson (r) 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

Métodos	Porcentagem de Dados Faltantes		
	5%	10%	15%
<i>PMM</i>	0,53	0,55	0,52
<i>Random Forest</i>	0,37	0,38	0,37
Regressão Linear Via Método <i>Bootstrap</i>	0,51	0,53	0,52

De acordo com a classificação estabelecida para este índice (índice de desempenho) mostrado na Tabela 2, todos os métodos de imputação, em todas as porcentagens de dados faltantes, se classificaram como desempenho “Péssimo”. Dessa forma, eles estão se encaixando no intervalo menor do que 0,4, com exceção do método *PMM* para o banco de dados “Amostra 10%”, que é classificado como mau desempenho. Valores baixos como estes apresentados na Tabela 6, indicam que os desempenhos dos métodos de imputação não foram satisfatórios. Mesmo que os valores do índice de exatidão tenham se revelado altos, o coeficiente de correlação de Pearson indicou o contrário, valores muito baixos. Portanto, pode-se aplicar estes métodos, mas com ressalvas em relação à dependência linear entre a variável imputada e observada.

Figura 21. Gráfico de linhas do índice de desempenho (c) da precipitação pluvial imputada por meio dos métodos: *PMM*, *random forest* e regressão linear via método *bootstrap*, nas porcentagens de dados faltantes de 5%, 10% e 15%

Tabela 6. 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

Métodos	Porcentagem de Dados Faltantes		
	5%	10%	15%
<i>PMM</i>	0,39	0,41	0,38
<i>Random Forest</i>	0,24	0,25	0,24
Regressão Linear Via Método <i>Bootstrap</i>	0,37	0,39	0,37

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

A partir deste trabalho, pode ser dito que a série histórica de chuva da ESALQ foi classificada como Classe 1, ou seja, nenhum sinal claro de falta de homogeneidade foi detectado.

Todos os métodos de imputação, em todas as intensidades, tiveram índices de desempenho baixos, classificados como “Péssimo”. Deve-se atrelar este fato pela alta variabilidade temporal e espacial da variável em estudo, além de se tratar de eventos sequencialmente independentes, fazendo assim, valer ressalva para aplicabilidade deste indicador de comparação.

A estatística de avaliação e até de comparação entre os métodos de imputação necessita de discussão adicional, pois embora outros índices tenham sido usados, os mais comuns são: o erro quadrático médio e o coeficiente de correlação de Pearson, porém alguns estudos (Willmott e Matsuura, 2005) discutem ser mais vantajoso o uso do erro médio absoluto ao invés da raiz do erro quadrático médio, pois este último envolve três passos contra um do erro médio absoluto. 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, como por exemplo García-Peña, Arciniegas-Alarcón e Barbin (2014), que utilizaram a série de precipitação também da ESALQ, porém a pesquisa destes autores se concentrou no período de 1917 a 2012, sem levar em consideração se a série era de fato homogênea e sem se atentar também que a partir de 1997 a série histórica convencional de chuva foi completada pela estação automática, e neste

trabalho concluiu-se que no caso de precipitação, a imputação pela média mensal supera o desempenho de métodos baseados na decomposição de valores singulares, logo, não é recomendado uma comparação entre os estudos, mesmo que sejam feitos com a mesma série de dados de mesma variável, logo, é importante que a validação dos métodos de imputação seja realizada por um conjunto maior de indicadores mais fidedignos possíveis, e para ser realizado o estudo de forma mais completa seria recomendado o emprego de um número maior de métodos de imputação.

Uma meta para trabalhos futuros é estender este estudo envolvendo estações próximas ao posto meteorológico da ESALQ, podendo assim, elaborar pesquisas comparando outros métodos de imputação de dados, como por exemplo os métodos de geoestatística, utilizando o pré-processamento de controle de qualidade e verificação da homogeneidade da série histórica.

5. CONCLUSÕES

Pode-se concluir que a série histórica de precipitação pluvial foi considerada de Classe 1, ou seja, nenhum sinal de falta de homogeneidade foi detectado.

O método que mais se destacou de forma positiva, foi o *PMM*, resultando em menores valores da raiz do erro quadrático médio e os maiores índices de acurácia e desempenho, mas, o índice de desempenho dos três métodos de imputação em todas as intensidades amostrais de dados faltantes foi considerado “Péssimo”.

É recomendado que seja feito o estudo com um número maior de métodos de imputação, inclusive de outras categorias, e também com o número máximo de indicadores de validação dos resultados.

REFERÊNCIAS

A. N. PETTITT. A Non-Parametric Approach to the Change-Point Problem. **Journal of the Royal Statistical Society**, v. 28, p. 126–135, 1979.

AGUILAR, E. et al. Guidance on metadata and homogenization. **Wmo Td**, v. 1186, n. October, p. 53, 2003.

ALEXANDERSSON, H. A HOMOGENEITY TEST APPLIED TO PRECIPITATION. **International Journal of Climatology**, v. 6, p. 661–675, 1986.

ALMEIDA, H. A.; HERMENEGILDO, G. M. S. Comparação dos dados meteorológicos obtidos por estações meteorológicas convencional e automática. **Revista Brasileira de Climatologia**, v. 12, p. 32–47, 2013.

ALMEIDA, H. A.; SOUZA, J. A.; ALCANTARA, H. M. Análise comparativa de dados meteorológicos obtidos por estação convencional e automática no semi-árido paraibano. **Revista Brasileira de Agrometeorologia**, v. 16, p. 58–66, 2008.

ANGELOCCI, L. R.; SENTELHAS, P. C. **Breve história das observações meteorológicas da época de criação da Escola Agrícola Prática aos tempos atuais - A meteorologia como ciência**. Piracicaba: [s.n.].

ASSIS, F. N. DE. **Modelagem da ocorrência e da quantidade de chuva e de dias secos em Piracicaba, SP e Pelotas, RS**. [s.l.] Escola Superior de Agricultura Luiz de Queiroz, 1991.

ASSIS, F. N. DE; VILLA NOVA, N. A. Modelagem da ocorrência e da quantidade de chuva em Piracicaba. **Scientia Agricola**, v. 51, p. 172–182, 1995.

AURÉLIO. **Mini dicionário da língua portuguesa**. 6^a edição ed. Curitiba: [s.n.].

BACK, A. J. Aplicação de análise estatística para identificação de tendências climáticas. **Pesquisa Agropecuária Brasileira**, v. 36, n. 5, p. 717–726, 2001.

BARACHO, S. M. L. Tratamento de Dados Ausentes em Estudos Longitudinais. p. 1–24, 2007.

BERTSIMAS, D.; ORFANOUDAKI, A.; PAWLOWSKI, C. Imputation of Clinical Covariates in Time Series. **Workshop at NeurIPS**, p. 4–9, 2018.

BERTSIMAS, D.; PAWLOWSKI, C.; ZHUO, Y. D. From Predictive Methods to Missing Data Imputation : An Optimization Approach. **Journal of Machine Learning Research**, v. 18, p. 1–39, 2018.

BREIMAN, L. **RANDOM FORESTS**. Berkeley: [s.n.].

BUISHAND, T. A. Some methods for testing the homogeneity of rainfall records. **Jornal of Hydrology**, v. 58, p. 11–27, 1982.

BURIOL, G. A. et al. Homogeneidade dos totais mensais de precipitação pluvial no Estado do Rio Grande do Sul. **Pesquisa Agropecuária Gaúcha**, v. 18, p. 145–157, 2012.

BUSSAB, W. D. O.; MORETTIN, P. A. **Estatística básica**. 8ª edição ed. São Paulo: Editora Saraiva, 2014.

BUUREN, S. VAN. **Flexible Imputation of Missing Data**. Chapman & ed. [s.l: s.n.].

BUUREN, S. VAN; GROOTHUIS-OUDSHOORN, K. **mice** : Multivariate Imputation by Chained Equations in R. **Journal of Statistical Software**, v. 45, n. 3, 2011.

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**, v. 5, n. May 1997, p. 89–97, 1997.

CARLINI-GARCIA, L. A. et al. Método bootstrap aplicados em níveis de reamostragem na estimação de parâmetros genéticos populacionais. **Scientia Agricola**, v. 58, n. 4, p. 785–793, 2001.

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**, v. 32, p. 575–583, 2017.

CASTRO, I. Q. **Uma Aplicação de Métodos de Imputação no Estudo de Fatores Associados ao Baixo Peso ao Nascer**. [s.l.] Universidade Federal de Juiz de Fora, 2014.

CERVELLINI, A.; SALATI, E.; GODOY, H. Estimativa da Distribuição da energia solar no estado de São Paulo. **Boletim Científico do Instituto Agrônomo do Estado de São Paulo**, v. 25, n. 1, p. 31–39, 1966.

DALPOSSO, G. H. **Método Bootstrap na agricultura de precisão**. [s.l.] UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ CAMPUS CASCAVEL, 2017.

DALY, C. et al. A probabilistic-spatial approach to the quality control of climate observations. **Bulletin of the American Meteorological Society**, n. January 2004, p. 5005–5012, 2004.

DANCEY, C.; REIDY, J. **Statistics without Maths for Psychology**. 7th editio ed. [s.l.] Pearson, 2017.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. [s.l: s.n.].

EFRON, B.; TIBSHINRANI, R. J. **An Introduction to the Bootstrap**. [s.l: s.n.].

ENDERS, C. K. **Applied Missing Data Analysis**. The Guilfo ed. New York: [s.n.].

FERLIN, C. **Imputação Multivariada: Uma Abordagem em Cascata**. [s.l.] Universidade Federal do Rio de Janeiro, 2008.

FERRARI, G. T. **Imputação de dados pluviométricos e sua aplicação na modelagem de eventos extremos de seca agrícola**. [s.l.] Universidade de São Paulo - Escola Superior de Agricultura “Luiz de Queiroz”, 2011.

FERRARI, G. T.; OZAKI, V. Missing data imputation of climate datasets: implications to modeling extreme drought events. **Revista Brasileira de Meteorologia**, v. 29, n. 1, p. 21–28, 2014.

FISCH, G. F.; SANTOS, J. M. **Comparação entre observações Meteorológicas convencionais e automáticas na região do Vale do Paraíba**. [s.l: s.n.].

GANDOLFI, M. **IMPUTAÇÃO MÚLTIPLAVIA ALGORÍTMO MICE E MÉTODO IMLD**. [s.l.] Universidade Estadual de Maringá, 2016.

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**, v. 29, n. 4, p. 527–536, 2014.

GOMES, F. P.; SALATI, E. As interações num Grupo de Experimentos de Adubação da Cana-de-Açúcar. **Revista de Agricultura**, v. 32, p. 261–270, 1957.

GRACE-MARTIN, K. **Seven ways to make up data: common methods to imputing missing data. The Analysis Factor.**

GROVES, R. M. et al. **Survey Methodology.** [s.l: s.n.].

KOLLER - MEINFELDER, F. **Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching.** [s.l: s.n.].

KYSELÝ, J.; DOMONKOS, P. Recent increase in persistence of atmospheric circulation over Europe: Comparison with long-term variations since 1881. **International Journal of Climatology**, v. 26, n. 4, p. 461–483, 2006.

LIMA, J. G. DE S. **Gerenciamento de Dados Climatológicos Heterogêneos para Aplicações em Agricultura.** [s.l.] UNICAMP - Universidade Estadual de Campinas, 2003.

LISKA, GILBERTO RODRIGUES; BORTOLINI, J.; SÁFADI, T.; BEIJO, L. A. Estimativas de velocidade máxima de vento em Piracicaba – SP via séries temporais e teoria de valores extremos. **Revista Brasileira de Biometria**, v. 31, p. 295–309, 2011.

LITTLE, R. J. A. Missing-Data Adjustments in Large Surveys Missing-Data Adjustments in Large Surveys. **Journal of Business & Economic Statistics**6, v. 6, n. July 2013, p. 287–296, 1988.

LITTLE, R. J. A.; RUBIN, D. B. **STATISTICAL ANALYSIS WITH MISSING.** [s.l: s.n.].

MARTINEZ-ESPINOSA, M.; SANDANIELO, V. L. M.; LOUZADA-NETO, F. O método de bootstrap para o estudo de dados de fadiga dos materiais. **Revista de Matemática e Estatística**, v. 24, n. 2, p. 37–50, 2006.

MARTÍNEZ, M. D. et al. Response to the comments on “time trends of daily maximum and minimum temperatures in Catalonia (NE Spain) for the period 1975-2004”. **International Journal of Climatology**, v. 31, n. 1, p. 153–157, 2011.

MORIMONT, L. E. Relatório da Fazenda São João da Montanha em Piracicaba de 1894. **Typographia Paulista**, p. 70, 1895.

MORRIS, T. P.; WHITE, I. R.; ROYSTON, P. Tuning multiple imputation by predictive mean matching and local residual draws. **Medical Research Methodology**, v. 14, p. 1–13, 2014.

MORTATTI, J. et al. Hidrologia dos rios Tietê e Piracicaba: series temporais de vazão e hidrogramas de cheia. **Revista de Ciência & Tecnologia**, v. 12, n. 23, p. 55–67, 2004.

NETO, J. L. S. A TENDÊNCIA DAS CHUVAS NO ESTADO DE SÃO PAULO : uma contribuição aos estudos sobre as mudanças climáticas regionais nos últimos 100 anos A TENDÊNCIA DAS CHUVAS NO ESTADO DE SÃO PAULO : uma contribuição aos estudos sobre as mudanças climáticas regionais no. n. October, 2015.

NUNES, L.; KLÜCK, M.; FACHEL, J. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos Multiple imputations for missing data: a simulation with epidemiological data. **Cadernos de Saúde Pública**, v. 25, n. 2, p. 268–278, 2009.

NUNES, L. N. et al. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. **Rev Bras Epidemiol**, v. 13, n. 4, p. 596–606, 2010.

OLIVEIRA, E. A. DE. **Métodos para análise de concordância: estudo de simulação e aplicação a dados de evapotranspiração**. [s.l.] Universidade de São Paulo Escola Superior de Agricultura “Luiz de Queiroz” Métodos, 2016.

OLIVEIRA, H. L. C. DE; OLIVEIRA, S. R. DE M.; MONTEIRO, J. E. B. DE A. GERAÇÃO DE SÉRIES TEMPORAIS DE DADOS METEOROLÓGICOS UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA. **11º Congresso Interinstitucional de Iniciação Científica**, p. 262–267, ago. 2017.

OLIVEIRA, P. R. S.; SANTOS, R. L. O USO DOS MÉTODOS DE PAULHUS & KOHLER (1952) PARA PREENCHIMENTO DE LACUNAS EM SÉRIES HISTÓRICAS DE PRECIPITAÇÕES PLUVIOMÉTRICAS NO TERRITÓRIO DE IDENTIDADE PORTAL DO SERTÃO (TIPS). Paulo Roberto Santana Oliveira 1; Rosângela Leal Santos 2 ; 2017.

OMM - WORLD METEOROLOGICAL ORGANIZATION WMO. **Guide to Meteorological Instruments and Methods of Observation**Genebra, 2014.

PAULHUS, J. L. H.; KOHLER, M. A. Interpolation of Missing Precipitation Records. **Monthly Weather Review**, v. 80, n. 8, p. 129–133, 1952.

PETERSON, W. T. et al. Anchovy spawning in relation to the biomass and the replenishment rate of their copepod prey on the western Agulhas bank. **South African Journal of Marine Science**, v. 12, n. 1, p. 487–500, 1992.

PICCINNI, M. R. D.; ANGELOCCI, L. R. Probabilidades de precipitação de intensidade máxima para Piracicaba, SP. **Anais ... Piracicaba: ESALQUIUSP - Sociedade Brasileira de Agrometeorologia**, p. 128–130, 1997.

RIBEIRO, M. D.; TESTEZLAF, R.; S FERRAREZI, R. Validação de parâmetros operacionais para o manejo de mesas de subirrigação. **Horticultura Brasileira**, v. 35, p. 549–556, 2018.

RIZZO, A. L. T.; CYMROT, R. Utilização da técnica de reamostragem bootstrap em aplicação na engenharia de produção. **X Encontro Latino Americano de Iniciação Científica**, p. 488–491, 2006.

RUBIN, D. B. Assignment to treatment group on the basis of a covariate. **Journal of Educational Statistics**, v. 2, p. 1–26, 1977.

RUBIN, D. B. Bayesian inference for casual effects: the role of randomization. **The Annals of Statistics**, v. 15, p. 34–58, 1978.

RUBIN, D. B. The bayesian bootstrap. **The Annals of Statistics**, v. 9, p. 130–134, 1981.

RUBIN, D. B. **Multiple imputation for nonresponse in surveys**. [s.l.] John Wiley & Sons, 1987.

RUBIN, D. B. Multiple Imputation after 18 + Years. v. 91, p. 473–489, 1996.

RUIZ-CÁRDENAS, R.; KRAINSKI, E. T. **PREENCHIMENTO DE FALHA EM BANCOS DE DADOS METEOROLÓGICOS DIÁRIOS: UMA COMPARAÇÃO DE ABORDAGENS**. Guarapari: [s.n.].

RUZA, M. S.; PAULA, A.; CORTE, D. **USO DO LiDAR PARA AVALIAÇÃO DA INFLUÊNCIA TOPOGRÁFICA NA ALTURA MÉDIA E ÁREA BASAL DE TALHÕES DE Pinus taeda L.** [s.l.] Universidade Federal do Parana, 2016.

SALVADOR, M. DE A.; SANTOS, L. S. F. C. DOS. Análise da precipitação na cidade de São Paulo no período 1961-2009. **Revista Brasileira de Climatologia**, p. 7–20, 2010.

SANSIGOLO, C. A. Distribuições de extremos de precipitação diária, temperatura máxima e mínima e velocidade do vento em Piracicaba, SP (1917-2006). **Revista Brasileira de Meteorologia**, v. 23, n. 3, p. 341–346, 2008.

SANTOS, R. S. DOS. **HOMOGENEIDADE E RECONSTRUÇÃO DE SÉRIES CLIMATOLÓGICAS PARA LOCALIDADES NO ESTADO DE MINAS GERAIS**. [s.l.] Universidade Federal de Viçosa, 2012.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: Our view of the state of the art. **Psychological Methods**, v. 7, p. 147–177, 2002.

SENTELHAS, P. C. et al. Análise comparativa de dados meteorológicos obtidos por estações convencional e automática. **Revista Brasileira de Agrometeorologia**, v. 5, p. 211–215, 1997.

SENTELHAS, P. C.; CARAMORI, P. H. Inconsistências na medida da chuva com pluviômetros de báscula, utilizados em estações meteorológicas automáticas. **Revista Brasileira de Agrometeorologia**, v. 10, p. 301–304, 2002.

SHAH, A. D. et al. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. **American Journal of Epidemiology**, v. 179, n. 6, p. 764–774, 2014.

SOARES, P. **Insights into tuberculosis: a survival analysis of time to recurrence**. [s.l.] Universidade de Lisboa - Faculdade de ciências, 2014.

SOUZA, I. A.; GALVANI, E.; ASSUNÇÃO, H. F. Estudo comparativo entre elementos meteorológicos monitorados por estações convencional e automática na região de Maringá, estado do Paraná. **Acta Scientiarum Technology**, v. 25, p. 203–207, 2002.

TANG, F.; ISHWARAN, H. Random forest missing data algorithms. **Statistical Analysis and Data Mining**, v. 10, n. 6, p. 363–377, 2017.

TEAM, R. C. **R: A Language and Environment for Statistical Computing** Vienna, Austria, 2017. Disponível em: <<http://www.r-project.org/>>

USP, L.-E.-. **Posto Meteorológico “Professor Jesus Marden dos Santos” ESALQ - USP**. Disponível em: <<http://www.leb.esalq.usp.br/posto/index.html>>. Acesso em: 13 fev. 2018.

VIEIRA, S. A. et al. METODOLOGIA DE IMPUTAÇÃO DE DAOS HIDROMETEOROLÓGICOS PARA ANÁLISE DE SÉRIES HISTÓRICAS - BACIA DO RIO DOS SINOS, RS, BRASIL. **Revista Brasileira de Climatologia**, v. 23, p. 189–204, 2018.

WALLIS, J. R.; O'CONNELL, P. E. Firm reservoir yield—how reliable are historic hydrological records? **Hydrological Sciences Bulletin**, v. 18, n. 3, p. 347–365, 1973.

WANDERLEY, H. S. et al. Variabilidade da precipitação no Sertão do São Francisco , estado de Alagoas. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 17, n. 31, p. 790–795, 2013.

WIJNGAARD, J. B.; KLEIN TANK, A. M. G.; KÖNNEN, G. P. Homogeneity of 20th century European daily temperature and precipitation series. **International Journal of Climatology**, v. 23, n. 6, p. 679–692, 2003.

WILLMOTT, C. et al. Statistics for the Evaluation and Comparison of Models. **Journal of Geophysical Research**, v. 90, n. 5, p. 8995–9005, 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**, v. 30, n. 1, p. 79–82, 2005.

WISSMANN, J. A. et al. Ferramenta computacional para análise de consistência de dados pluviométricos. **Revista Varia Scientia**, v. 06, n. 11, p. 99–106, 2006.