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

Comparação de método de imputação para dados de precipitação diária

Valiana Alves Teodoro

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

Piracicaba
2019

Valiana Alves Teodoro
Licenciatura em Matemática

Comparação de método de imputação para dados de precipitação diária
versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Profa. Dr. **VITOR AUGUSTO OZAKI**

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

Piracicaba
2019

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

Teodoro, Valiana Alves

Comparação de método de imputação para dados de precipitação
diária / Valiana Alves Teodoro. - - versão revisada de acordo com a
resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

85 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de
Queiroz".

1. Dados faltantes 2. Filtro de Kalman 3. MICE 4. Ponto de grade I.
Título

DEDICATÓRIA

Aos meus pais
Valdete José Teodoro e Eliana Queiroz Alves Teodoro

Ao meu irmão
Marcelo Alves Teodoro

Aos meus avôs (in memória)
José Teodoro Ferreira
Sebastiana Ferreira Teodoro
João Alves Neto
Jovina Garcia Alves

Com amor, dedico!

AGRADECIMENTOS

A Deus pelo dom da vida e por me mostrar que sou mais forte do que eu penso. Entender a vontade de Deus nem sempre é fácil, mas crer que ele está no comando e tem um plano para nossa vida, faz a caminhada valer a pena. Agradeço ao Senhor por estar sempre presente em nossas vidas, olhando e rogando por nós, nos guiando, nos protegendo e por nos conceder a sua benção, hoje e sempre.

Aos amores da minha vida, meus pais, que fazem da minha vida um caminho simples e lindo de se viver. Ao meu pai, uma pessoa calma e honesta. A minha mãe, uma pessoa forte, com caráter e personalidade inabaláveis. De vocês recebi todo o meu alicerce, aprendi como as pequenas coisas podem nos trazer grandes felicidades e que compartilhar é melhor do que guardar as alegrias só para mim. Tenho eterna gratidão a vocês por me permitirem que eu tivesse acesso ao meu bem mais importante, o conhecimento.

Ao irmão Marcelo pelo amor, carinho e proteção. A nossa sorte é que nós dois herdamos o dom que acompanha a família á gerações: ter um coração enorme que nunca guarda rancores. Dedico também a toda minha família e amigos. É por cada um deles que hoje sou quem sou, que tenho força, coragem e sigo meus sonhos. Eu os amo com intensidade e sem prazo de validade.

Tenho um sentimento de missão cumprida e de gratidão a ESALQ e às pessoas que convivi nesses últimos anos. Trata-se de uma vitória muito importante, pois conviver com pessoas com brilho nos olhos, isso é uma mão dupla, porque além de fornecer alguma coisa para eles, você acaba levando também. Tenho uma grande paixão e orgulho de ter estudado na ESALQ, pois esta escola possui uma energia positiva.

Ao Departamento de Estatística e Experimentação Agronômica da ESALQ pelos valiosos ensinamentos e contribuições a minha formação técnica e pessoal, em especial àqueles que foram meus professores, Carlos Tadeu dos Santos Dias, César Gonçalves de Lima, Clarice Garcia Borges Demétrio, Edwin Ortega, Idemauro A. Rodrigues de Lara, Sônia M. De Stefano Piedade e Vitor Augusto Ozaki. E também aos funcionários, em especial a Solange Paes Sabadin, por sua paciência, gentileza e prontidão em ajudar em todos os momentos.

Um agradecimento especial devo ao meu mestre e amigo Vitor Ozaki. Sua profunda sapiência e seu olhar sempre à frente do seu tempo foram fatores fundamentais na minha formação profissional. Exemplo de professor e orientador, suas ideias e críticas (sempre construtivas) foram fundamentais na elaboração desta obra.

Aos professores Adrielle Giaretta Biase, Antônio Carlos Simões Pião, David José Miquelluti, José Silvio Govone, Lúcio Borges de Araújo e Sônia M. De Stefano Piedade por participar da minha banca de doutorado e contribuir com sugestões válidas, críticas construtivas, pertinentes e importantes para meu trabalho.

Aos amigos do Doutorado pela ajuda em momentos de dificuldade e, principalmente, por vossas amizades: Andreia Hilário, Andreza Jardelino, Douglas Batista, Erasnilson Vieira, Hiron Farias, Simone Grego e Talita Tanaka. E também ao grupo de pesquisa GESER, pela amizade e o conhecimento sobre seguro rural compartilhado ao longo destes anos, em especial ao Arthur Lula, Guilhaume Rocha, Gislaine Vieira e Daniel Miquelluti. Pois é galera, no final dá tudo certo!

Sou grato a Micheline, Valcides, Larissa, Rayssa, Karla, Erasnilson, Daniel, Douglas, Jesiane, Mariane, Iabita, Adrielle, Angel, Mari, Ana Paula, Lúcio e Mirian pessoas que me ensinaram o sentido da palavra generosidade e humildade. Esses amigos, por ter enfrentado tantos altos e baixos na vida, me mostrou a importância de ser persistente e de sempre lutar. Para vocês, meus irmãos de coração, sou grata por me proporcionaram a amizade, momentos de alegria e respeito ao próximo. Essas pessoas iluminadas passaram pelo meu caminho que fizeram destes últimos anos um período único na minha vida. Amigos que estarão sempre presentes, apesar da distância.

Finalmente, agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), como processo 140287/2015-9, por todo o suporte financeiro durante a fase de doutorado.

MUITO OBRIGADA!
#GRATIDÃO

*“Neblina na serra, chuva na terra.
Neblina baixa, sol que racha.”*

Autor desconhecido

SUMÁRIO

RESUMO	9
ABSTRACT	10
LISTA DE FIGURAS	11
LISTA DE TABELAS	13
1. INTRODUÇÃO.....	15
1.1 Conceitos e Definições	18
1.1.1. Dados de Precipitação	18
1.1.1.1. Dados Faltantes	20
1.1.2. Mecanismos e Padrão de Dados Faltantes.....	21
Referências	22
2. PREENCHIMENTO DE FALHAS DE PRECIPITAÇÃO PELA IMPUTAÇÃO MÚLTIPLA POR EQUAÇÕES ENCADEADAS UTILIZANDO DADOS EM PONTO DE GRADE	27
RESUMO	27
ABSTRACT	27
2.1. Introdução.....	28
2.2. Revisão da Literatura.....	30
2.3. Metodologia.....	32
2.3.1. MICE	32
2.3.2. Algoritmo	33
2.3.3. Modelo.....	34
2.3.4. Método da Média Preditiva	35
2.3.5. Avaliação do Processo de Imputação	36
2.3.6. Checando o Modelo de Imputação	36
2.4. Descrição dos Dados	37
2.5. Resultados e Discussão.....	41
2.6. Conclusão	55
Referências	56
3. COMPARAÇÃO DE MÉTODOS DE IMPUTAÇÃO NAS ABORGAGENS UNIVARIADA E MÚLTIPLA	61
RESUMO	61
ABSTRACT	61
3.1. Introdução.....	62

3.2. Revisão da Literatura	64
3.2.1. Imputação de Dados.....	64
3.2.2. Imputação de Dados de Precipitação	65
3.3. Metodologia	67
3.4. Mecanismo de Dados Faltantes.....	67
3.5. Métodos de Imputação	67
3.5.1. Filtro de Kalman	67
3.5.2. Persistência.....	68
3.5.3. Média Móvel Ponderada	68
3.5.4. Imputação Clássica	69
3.5.5. Amostra Aleatória	69
3.5.6. Decomposição Sazonal	69
3.5.7. MICE.....	70
3.5.7.1. Algoritmo	71
3.5.7.2. Modelo	71
3.5.7.3. Método da Média Preditiva.....	72
3.6. Avaliação do Processo de Imputação.....	73
3.7. Descrição dos Dados	74
3.8. Resultados e Discussão	77
3.9. Conclusão	82
Referências.....	82

RESUMO

Comparação de método de imputação para dados de precipitação diária

As principais causas da redução da produtividade agrícola são os eventos climáticos, e a variável meteorológica de grande importância para a produção agrícola é a precipitação. Alguns dos problemas das bases de dados meteorológicos são a descontinuidade e dados faltantes. Nesse sentido, os dados de precipitação em ponto de grade (*Gridpoint*), são uma excelente fonte de informações em pesquisas climatológicas. Para superar os problemas de dados faltantes e construir um banco de dados completos é necessário um processo de imputação. Portanto, o objetivo do trabalho foi comparar metodologias de imputação, utilizou abordagens univariada e múltipla, e comparou o desempenho em termos de imputação em diferentes cenários de dados faltantes e utilizou a raiz do erro quadrático médio (RMSE) como métrica. Para séries de precipitação diária que tinham dados faltantes foi realizado a imputação pelo método imputação múltipla por equações encadeadas (MICE), utilizando a informação de mês, ano e precipitação em ponto de grade. Foram utilizados quatro modelos, nos quais a precipitação diária dependia de: mês; mês e ano; precipitação em ponto de grade; mês, ano e precipitação diária em ponto de grade. Utilizou-se a raiz do erro quadrático médio (RMSE) como métrica e para verificar as imputações, analisou-se a semelhança entre os dados observados e os dados imputados pelo Teste de Kolmogorov-Smirnov e pelos gráficos da média e variância das imputações. O modelo com o maior número de variáveis foi escolhido para imputar os dados faltantes das séries de precipitação diária. Nesse trabalho, o uso de dados de precipitação em ponto de grade mostrou ser na imputação de dados de séries de precipitação diária. Para uma série de precipitação diária completa, concentra-se na comparação e avaliação de métodos de imputação nas abordagens univariada e múltipla, para dados de precipitação diária. Na abordagem univariada, utilizou-se diferentes configurações filtro de Kalman, Média Móvel Ponderada e Decomposição Sazonal. Na abordagem múltipla, utilizou-se o método MICE, com diferentes modelos. Os dados faltantes foram estimados em uma série de precipitação diária, em que os dados faltantes foram gerados de maneira aleatória e em trechos e utilizou-se a raiz do erro quadrático médio (RMSE) como métrica. Os resultados identificaram que o método de Filtro de Kalman forneceu os menores valores de RMSE, para todos os cenários de dados faltantes. A aplicação do algoritmo Filtro de Kalman produziu melhores estimativas para os valores diários de precipitação. O Filtro de Kalman pode ser uma importante metodologia para imputação de dados de precipitação diária, garantido uma série temporal completa para análises de vários setores, dentre eles a agricultura.

Palavras-chave: Dados faltantes, Filtro de Kalman, MICE, Ponto de grade

ABSTRACT

Comparison of imputation method for daily precipitation data

The main causes of the reduction of agricultural productivity are the climatic events, and the meteorological variable of great importance for the agricultural production is precipitation. Some of the problems of meteorological databases are discontinuity and missing data. In this sense, grid point precipitation (*Gridpoint*) data is an excellent source of information in climatological research. To overcome missing data problems and build a continuous database, an imputation process is required. Therefore, this work has the objective of comparing two imputation methodologies, using the MICE method and the Kalman filter, and comparing the performance in terms of imputation in different scenarios of missing data, using root mean square error (RMSE) as metric. For series of daily precipitation that had missing data, imputation was carried out by the multiple imputation method by chain equations (MICE), using the information of month, year and precipitation in grid point. Four models were used, in which the daily precipitation depended on: month; month and year; precipitation in grid point; month, year and daily precipitation in grid point. The root mean squared error (RMSE) was used as a metric and to verify imputations, the similarity between the observed data and the data imputed by the Kolmogorov-Smirnov test and the mean and variance imputation graphs were analyzed. The model with the largest number of variables was chosen to impute missing data from the daily precipitation series. In this work, precipitation data in grid point showed the importance and advantages of their use as information in imputation of daily precipitation series data. For a complete daily precipitation series, it focuses on the comparison and evaluation of imputation methods in the univariate and multiple approaches for daily precipitation data. In the univariate approach, we used different Kalman filter configurations, Weighted Moving Average, and Seasonal Decomposition. In the multiple approach, the MICE method was used, with different models. The missing data were estimated in a series of daily precipitation, in which the missing data were generated randomly and in sections, and the root mean square error (RMSE) was used as a metric. The results identified that the Kalman Filter method provided the lowest RMSE values for all missing data scenarios. The application of the Kalman filter algorithm produced better estimates for the daily values of precipitation. The Kalman Filter can be an important methodology for imputation of daily precipitation data, ensuring a complete time series for analysis of several sectors, among them agriculture.

Keywords: Missing data, Kalman filters, MICE, Gripoint

LISTA DE FIGURAS

2.1. Localizações das estações.....	39
2.2. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Bela Vista do Paraíso.....	42
2.3. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Jataizinho.....	43
2.4. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Nova Londrina.....	43
2.5. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Porto Santa Terezina.....	44
2.6. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Santa Cruz do Monte Castelo.....	44
2.7. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Tomazina.....	45
2.8. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Vila Silva Jardim.....	45
2.9. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Bela Vista do Paraíso.....	48
2.10. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Jataizinho.....	48
2.11. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Nova Londrina.....	49
2.12. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Porto Santa Terezina.....	49
2.13. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Santa Cruz do Monte Castelo.....	51
2.14. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Tomazina.....	51
2.15. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Vila Silva Jardim.....	52
2.16. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Bela Vista do Paraíso.....	52

2.17. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Jataizinho.....	53
2.18. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Nova Londrina.....	53
2.19. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Porto Santa Terezina.....	54
2.20. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Santa Cruz do Monte Castelo.....	54
2.21. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Tomazina.....	55
2.22. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Vila Silva Jardim.....	55
3.1. Fluxograma do algoritmo utilizado para comparar a precisão dos métodos de imputação de valores faltantes para a precipitação diária.....	76
3.2. Série temporal de precipitação diária da estação de Porto Eucalipto.	78
3.3. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes.....	79
3.4. Média das RMSE obtidos com diferentes modelos, para 500 dados faltantes.....	75
3.5. Média das RMSE obtidos com diferentes modelos, para 1000 dados faltantes.....	80
3.6. Média das RMSE obtidos com diferentes modelos, para 1500 dados faltantes.....	80
3.7. Média das RMSE obtidos com diferentes modelos, para 2000 dados faltantes.....	81

LISTA DE TABELAS

2.1. Posto pluviométrico utilizadas nesse trabalho.	38
2.2. Intervalos selecionados para latitude e longitude e correlação entre o posto pluviométrico e as grades de $0,50^\circ \times 0,50^\circ$	40
2.3. Modelos utilizadas nas imputações.....	41
2.4. Teste de Kolmogorov-Smirnov comparando a densidade dos dados observados com os dados imputados.	47
3.1. Posto pluviométrico utilizado neste trabalho.	74
3.2. Métodos de imputação utilizados neste trabalho.....	77

1. INTRODUÇÃO

A agricultura é uma das mais antigas e importantes atividades para o ser humano, sendo considerada fonte de matérias-primas, alimentos e energia. No entanto, muitos riscos ameaçam em maior ou menor grau todas as cadeias do agronegócio, basicamente são cinco grandes riscos: risco de produção, mercado, financeiro, legal e humano. Entre esses, os riscos de produção e de mercado apresentam maior impacto na renda, pois não estão sob controle dos produtores rurais (BRISOLARA, 2013).

Basicamente, os riscos de produção estão relacionados às adversidades climáticas, pragas, incêndio, doenças ou uso inadequado de tecnologias (BRISOLARA, 2013). Sendo que as principais causas da redução da produtividade agrícola são os eventos climáticos. Secas, geadas, chuvas excessivas, chuvas de granizo afetam diversas culturas em diferentes estágios de desenvolvimento (OZAKI, 2008). Além disso, a precipitação é considerada a variável meteorológica de grande importância para a produção agrícola (TEODORO et al., 2016)

Ao longo dos anos, várias ferramentas de gerenciamento de riscos foram desenvolvidas para minimizar o impacto do clima na produção agrícola, entre elas, o seguro rural (HARWOOD et al., 1999). Além de minimizar os riscos de perda na agricultura, o seguro rural fornece a recuperação da capacidade financeira do produtor em caso de danos causados por eventos naturais extremos (OZAKI, 2005).

No Brasil, a partir de 2003, iniciou uma participação governamental mais forte no seguro rural com a aprovação da Lei 10.823. Essa Lei, autorizou o Poder Executivo a conceder a subvenção econômica ao prêmio do seguro rural, por meio do qual o Governo Federal arca com parcela dos custos de aquisição da apólice de seguro, tornando-a mais acessível aos produtores rurais (BRASIL, 2003). E em 2005, o Programa de Subvenção ao Prêmio do Seguro Rural (PSR), instituído pelo Decreto nº 5.121/2004, que regulamentou a mencionada Lei, vem auxiliando milhares de produtores a contratar o seguro, como forma de se precaverem contra perdas financeiras decorrentes das adversidades climáticas (BRASIL, 2005).

De acordo com o MAPA (2019), atualmente 25% da área do Brasil é segurada, no qual o número de apólices de seguro rural passou de 21 mil para 63 mil entre os anos de 2006 a 2018, sendo que a maioria das apólices são do estado do Paraná (38%). Em 2018, foram atendidos mais de 15 mil produtores, com mais de 21 milhões de hectares de área segurada, sendo que 98% da área é de grãos no estado do Paraná.

No entanto, o seguro de rural apresenta alguns problemas, entre deles a ausência de séries históricas longas e completas, consequentemente, resultando em altos custos de vendas de apólices de seguro rural e também desencorajando a oferta de produtos pelas companhias seguradoras (COBLE e KNIGHT, 2002; GOODWIN, 2001; QUIGGIN et al., 1993; QULGGIN et al., 1993; SKEES e REED, 1986; SMITH e GOODWIN, 1996b).

No Brasil existem várias bases de dados meteorológicos, de diferentes instituições (TSUKAHARA et al., 2010). No entanto, há problemas de continuidade, completude, preenchimento de falhas, entre outros. Na etapa de tempo mensal, os problemas de dados são menores do que aqueles com resolução diária (CARVALHO et al., 2017). Sendo que o conjunto de observações de precipitação com longa série temporal são raros devido as mudanças frequentes na localização dos observatórios em localidades próximas, o que resulta em fragmentação ou inconsistência da série de dados (VICENTE-SERRANO et al., 2010).

Os dados do ponto de grade, baseados em conjunto de observações, são uma excelente fonte de informações em pesquisas climatológicas. Esses conjuntos de dados em ponto de grade disponíveis, são baseados em dados observados por estações meteorológicas e em dados de satélite e vêm sendo amplamente utilizados na última década, em especial em trabalhos climatológicos que envolvem análises espaciais, minimizando problemas com limitações de dados faltantes e irregularidade de cobertura espacial (GONCALVES et al., 2009; CARDOSO, QUADRO, 2017). Uma vantagem adicional dos dados de pontos da grade é que longas séries temporais estão disponíveis para muitas variáveis climáticas (SILVA et al., 2007).

Os modelos agrometeorológicos de apoio a decisão agrícola podem gerar resultados comprometedores aos agricultores, se as séries climáticas não forem contínuas (WISSMANN et al., 2006). De acordo com diversos autores, entre os quais Daly et al. (2004), Feng, Hu e Qian (2004) e Vicente-Serrano et al. (2010), para superar esses problemas e para construir um banco de dados contínuo para análise é necessário um processo de imputação e controle de qualidade dos dados de precipitação.

Na década de 70, iniciaram-se várias técnicas estatísticas para substituição dos dados faltantes, estes métodos são conhecidos na literatura como métodos de imputação de dados faltantes ou métodos de substituição de dados faltantes, os quais são classificados como simples e múltiplos. Os métodos de imputação simples consistem em estimar, apenas uma vez, cada valor em falta por algum método estatístico utilizando as informações presentes no conjunto de dados, e assim se obtém-se um banco de dados completo para as análises

(RUBIN 1976, 1978). Os métodos de imputação múltipla surgiram na literatura como tentativa de controlar o viés fornecido pelas estimativas ao se utilizar os métodos estatísticos com o restante dos dados. Estes métodos refletem a incerteza sobre as estimativas dos dados faltantes ao substituir cada valor faltante por um conjunto de valores razoáveis sobre o valor a ser imputado (SILVA, 2012).

As primeiras técnicas de imputação simples são: imputação pela média ou pela mediana da variável, imputação pelo vizinho mais próximo, imputação hot deck, imputação pela regressão linear e imputação por meio da máxima verossimilhança.

Nos métodos de imputação múltipla, um dos destaques foi o método proposto por Rubin (1987), em que cada valor faltante é substituído por um conjunto de valores plausíveis, o que representa a incerteza sobre o valor a ser imputado. A partir disso, duas abordagens gerais para imputar dados multivariados surgiram: modelagem conjunta (JM) (SCHAFER, 1997a) e especificação totalmente condicional (FCS), também conhecido como imputação múltipla por equações encadeadas (MICE) (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

O MICE é uma abordagem flexível e prática para lidar com dados perdidos (WHITE et al., 2011), além disso, uma característica importante dessa técnica é sua capacidade de lidar com diferentes tipos de variáveis (contínua, binária, categórica), porque cada variável é imputada usando seu próprio modelo de imputação (VAN BUUREN, STEF AND OUDSHOORN, 1999; VAN BUUREN e OUDSHOORN, 2000; VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

As técnicas de imputação, citadas acima, depende de correlações entre variáveis (MORITZ e BARTZ-BEIELSTEIN, 2017). Um desafio especial na pesquisa de imputação é a imputação de séries temporais univariada (MORITZ et al., 2015a), para isso algoritmos de imputação de série temporal univariados eficazes precisam empregar as correlações intertemporais (MORITZ e BARTZ-BEIELSTEIN, 2017).

Uma ampla variedade de técnicas de imputação estatística baseadas em ARIMA e modelos estruturais, metodologia de média móvel, e as técnicas simples e complexas de interpolação estão sendo bem-sucedidos na imputação de valores omissos para diferentes tipos de dados (MORITZ e BARTZ-BEIELSTEIN, 2017). Há poucos estudos de imputação de séries temporais univariadas, em que faltam bons artigos de visão geral com comparações de diferentes algoritmos (MORITZ et al., 2015a).

Algumas das técnicas de imputação de séries temporais são: Interpolação, Filtro de Kalman, Persistência, Média Móvel Ponderada, Imputação Clássica, Amostra Aleatória,

Decomposição Sazonal e Sazonalidade por partes. Atualmente, essas técnicas de imputações são exclusivas para séries temporais univariados e inclui várias opções/algoritmos possíveis para sua implementação (DEMIRHAN e RENWICK, 2018; MORITZ e BARTZ-BEIELSTEIN, 2017).

Portanto, esse trabalho tem como objetivo comparar duas metodologias de Imputação, usamos o método MICE e o Filtro de Kalman, comparando-se o desempenho em termos de imputação em diferentes cenários de dados faltantes. Para isso, dois artigos são apresentados nos capítulos seguintes.

No primeiro artigo, organizou-se um banco de dados com os valores de precipitação diária, mês, ano e precipitação diária em ponto de grade. Assim, utilizou-se a metodologia de Imputação Múltipla por Equações Encadeadas, utilizando o método MICE para fazer imputações de valores diários de precipitação diária, criando diferentes séries incompletas, com diferentes preditores. Os resultados obtidos foram comparados, comumente utilizados na estimativa de dados faltantes, para estações climatológicas, que tem dados faltantes, localizadas no Paraná.

No segundo artigo, concentrou-se na comparação e avaliação de métodos de imputação nas abordagens univariada e múltipla, para dados de precipitação. Na abordagem univariada, utilizou-se diferentes configurações filtro de Kalman, Média Móvel Ponderada e Decomposição Sazonal. Na abordagem múltipla, utilizou-se o método MICE, com diferentes modelos. Para os algoritmos MICE, construiu-se um banco de dados com os valores de precipitação diária, mês, ano e precipitação diária em ponto de grade.

1.1 Conceitos e Definições

1.1.1. Dados de Precipitação

O monitoramento da precipitação é uma tarefa indispensável para compreensão das características da região para o desenvolvimento das pesquisas e para diversos setores, como agricultura, florestal, para o planejamento ambiental e também para o estudo das mudanças climáticas, uma vez que referencia a pluviometria regional ao longo do tempo (FRITZSONS et al., 1997; MOL, 2005).

Para monitorar a precipitação, há dois tipos de estações meteorológicas de superfície: as Estações Meteorológicas Convencionais e as Estações Meteorológicas Automáticas. As Estações Meteorológicas Convencionais exigem a presença diária do observador

meteorológico para coleta de dados. A Estação Meteorológica Automática com a coleta de dados totalmente automatizada. Nesse tipo de estação os sensores captam por um sistema de aquisição de dados, possibilitando que o armazenamento e o processamento dos dados sejam informatizados (Galina e Verona).

As estações climatológicas podem ser classificadas como: Estações Climatológicas Principais e Estações Climatológicas Ordinárias. As Estações Climatológicas Principais medem todos os elementos meteorológicos, enquanto as Estações Climatológicas Ordinárias fornecem somente algumas informações meteorológicos. São Estações Climatológicas Ordinárias: Estações Agrometeorologias, Estações Meteorológicas Aeronáuticas e os Postos Pluviométricos. Todas as demais estações com qualidades distintas enquadram-se como Estações Especiais, tais como: Estações Ozonométricas, Micrometeorológicas, Actiométricas, de Radar, de recepção de dados de Satélites, Plataformas automáticas, etc (VIANELLO & ALVES, 1991).

Os postos pluviométricos são destinados à coleta de chuvas (SENTELHAS et al., 2000). Nesses postos, em posições geográficas específicas, a forma mais utilizada para a execução de dados é a construção e a manutenção de uma rede. Os dados de precipitação de postos pluviométricos são coletados periodicamente para verificação da quantidade de precipitação nesse local (CONTI, 2001).

Na década de 60, o surgimento de satélites meteorológicas possibilitando o monitoramento de extensas áreas da superfície terrestre e oceânica, auxiliando na busca de informações sobre áreas de pouca ou nenhuma informação. Os radares meteorológicos conseguem captar a distribuição espacial da precipitação, no entanto, a qualidade dos seus dados deve ser investigada constantemente, pois o equipamento apresenta dificuldades de calibração de um local para outro. Desde então, diversas técnicas de estimativa de precipitação utilizando dados de satélite estão sendo desenvolvidas. A eficácia de algumas técnicas varia de acordo com a sua finalidade e com os tipos de precipitação que se deseja identificar, sendo possível distinguir com certa confiabilidade as situações de chuva e não chuva (SANTOS, J A S, 2012).

O uso da estimativa de precipitação por satélite é uma missão de vários institutos de pesquisas. O *Climate Prediction Center* (CPC) da *National Oceanic and Atmospheric Administration* (NOAA) veem melhorando a análise precipitação diária, para auxiliar os monitoramentos em tempo real e estudos de variabilidade climática. O CPC produz rotineiramente análises de precipitação diária em ponto de grade (*gridpoint*) com diferentes calibres. São monitoramentos em tempo real, sendo essas informações disponibilizadas em

um contexto histórico, dessa forma sendo uma análise aprimorada para a verificação das previsões e estudos do clima (HIGGINS et al. , 2000). Os dados de ponto de grade são séries temporais longas que estão disponíveis para muitas variáveis climáticas. Todos os conjuntos de dados podem ser acessados selecionando em um formato de campos apropriados (TROUET e OLDENBORGH, 2013).

1.1.1.1. Dados Faltantes

A ausência de registros meteorológicos é um problema frequente na maioria das séries climatológicas brasileiras (TSUKAHARA et al., 2010). Existem vários fatores que podem influenciar na ausência ou no erro de dados de uma estação meteorológica, dentre elas são: a ausência do observador, as falhas nos sensores, a calibração dos equipamentos, as falhas nas transmissões dos dados, a manutenção nos sistemas, perda das anotações ou das transcrições dos registros pelos operadores e intervenção de agentes externos (OLIVEIRA et al., 2010b; STRECK et al., 2009; TSUKAHARA et al., 2010).

Os equipamentos de estações meteorológicas são instalados ao ar livre e, conseqüentemente, sujeitos às condições naturais do meio ambiente. A ação de animais próximos aos equipamentos e fortes mudanças no tempo podem atrapalhar as leituras dos sensores, causando falhas no armazenamento dos dados. Ademais, os sensores são aparelhos eletrônicos e estão sujeitos a falhas. Sendo assim, as condições naturais do meio ambiente e problemas técnicos podem interferir na leitura e armazenamento dos dados observados. (HUI et al., 2004; VENTURA, 2012).

Nas estações meteorológicas convencionais, em que o observador anote as medidas captadas pelos equipamentos, pode acontecer problemas, um deles sendo que o observador não realiza alguma leitura do equipamento, em determinado horário ou dia, provocando falhas nos registros dos dados (VENTURA, 2012). Para as estações que as medições são automáticas podem falhar, seja devido à falha do próprio instrumento ou na transmissão dos dados (TARDIVO e BERTI, 2014). Falhas também podem ocorrer por avaria, desligamento de equipamentos, manutenção, calibração, limitações físicas ou fenômenos climáticos (HUI et al., 2004).

1.1.2. Mecanismos e Padrão de Dados Faltantes

Os mecanismos de dados faltantes descrevem possíveis relações entre variáveis medidas e a probabilidade da ausência desses dados. Embora os mecanismos de dados faltantes não ofereçam uma explicação causal para essa ausência, eles representam relações matemáticas genéricas entre os dados e a falta (por exemplo, em um desenho de pesquisa, pode haver um relacionamento sistemático entre o nível de escolaridade e a propensão para falta de dados. Segundo os métodos de imputação desenvolvidos por Rubin (1976) e Rubin (1978), os dados faltantes são estimados de acordo com o mecanismo que gerou a falta e o padrão do seu comportamento.

A ocorrência de dados faltantes em banco de dados geralmente obedece a um mecanismo que aponta as condições de geração dos dados faltantes. A principal terminologia de classificação foi criada por Rubin (1976) e define três mecanismos teóricos gerais extensamente utilizados na literatura:

- Ausência Totalmente Aleatória (Missing Completely at Random - MCAR);
- Ausência Aleatória (Missing at Random - MAR);
- Ausência Não-Aleatórias (Not Missing at Random - NMAR).

A MCAR ocorre quando a os dados faltantes não tem relação com alguma outra variável medida e com a própria variável em falta. Assim, a perda dos dados não está relacionada que não foi medida e nem com variáveis que possam ser medidas no decorrer das medidas, ocorreu simplesmente por conta de fatores inesperados (LITTLE, RA e RUBIN, 2002). Portanto, os pontos de dados faltantes ocorrem inteiramente de forma aleatória (MORITZ et al., 2015a).

A MAR ocorre de forma aleatória indicando que a perda da observação ocorreu devido à alguma informação contida no conjunto de dados, ou seja, os dados faltantes dependem apenas de informações observadas, disponíveis para análise e correlacionadas com a variável que possui dados faltantes. Assim sendo, a MAR ocorre em que a probabilidade de uma variável ter dados faltantes depende apenas das informações disponíveis (LITTLE e RUBIN, 2002; RUBIN, 1976).

Por fim, ambas as suposições são violadas se a probabilidade de falta variar e não for caracterizada por um preditor disponível, essa condição é a MNAR (Rubin, 1976; Little e Rubin, 2002).

Um padrão de dados faltantes refere-se à configuração dos valores observados e faltantes em um conjunto de dados, enquanto os mecanismos de dados faltantes descrevem possíveis relações entre variáveis medidas e a probabilidade de dados faltantes. Observe que um padrão de dados faltantes simplesmente descreve a localização dos "buracos" nos dados e não explica porque os dados estão faltando (ENDERS, 2010).

Referências

BRASIL. Decreto-Lei nº 10.826 de 22 de dezembro 2003 do Ministério da Agricultura, Pecuária e Abastecimento, 2003. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/2003/110.826.htm>. Acesso em: 10 jun. 2019.

BRASIL. Decreto nº 5121 de 29 de junho de 2004 do Ministério de Agricultura, Pecuária e Abastecimento, 2005. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2004/Decreto/D5121.htm>. Acesso em: 10 jun. 2019

BRISOLARA, C. S. Proposições para o desenvolvimento do seguro de receita agrícola no Brasil: p. 239, 2013.

BUUREN, S.; OUDSHOORN, C. Multivariate imputation by Chained equations: MICE V1.0 user's manual. TNO, 2000.

CARDOSO, C. S.; QUADRO, M. F. L. Análise comparativa de dados de precipitação gerados pelo "Climate Prediction Center – CPC" versus dados observados para o Sul do Brasil Camila. Revista Brasileira de Geografia Física, v. 10, n. 4, n. ISSN:1984-2295, p. 1180–1198, 2017.

CARVALHO, J. R. P. DE et al. Model for Multiple Imputation to Estimate Daily Rainfall Data and Filling of Faults Modelo de Imputação Múltipla para Estimar Dados de Precipitação Diária e Preenchimento de Falhas. v. 32, n.4, p. 575–583, 2017.

COBLE, K. H.; KNIGHT, T. O. Crop Insurance as a Tool for Price and Yield Risk Management. In: A Comprehensive Assessment of the Role of Risk in U.S. Agriculture. Boston, MA: Springer US, 2002. p. 445–468.

CONTI, G. N. ESTIMATIVA DA PRECIPITAÇÃO ATRAVÉS DE TÉCNICAS DE SENSORIAMENTO REMOTO: ESTUDO DE CASO PARA O ESTADO DO RIO GRANDE DO SUL. [s.l.] Universidade Federal do Rio Grande do Sul, 2001.

DALY, C. et al. A probabilistic-spatial approach to the quality control of climate observations. Disponível em: <https://www.researchgate.net/publication/252891043_A_probabilistic-spatial_approach_to_the_quality_control_of_climate_observations>. Acesso em: 1 abr. 2019.

DEMIRHAN, H.; RENWICK, Z. Missing value imputation for short to mid-term horizontal solar irradiance data. Applied Energy, v. 225, n. May, p. 998–1012, 2018.

ENDERS, C. K. Applied missing data analysis. [s.l.] Guilford Press, 2010.

FENG, S.; HU, Q. I.; QIAN, W. QUALITY CONTROL OF DAILY METEOROLOGICAL DATA IN CHINA, 1951-2000: A NEW DATASET. INTERNATIONAL JOURNAL OF CLIMATOLOGY Int. J. Climatol, v. 24, p. 853–870, 2004.

FRITZSONS, E. et al. RA'E GA o espaço geográfico em análise. [s.l: s.n.]. v. 23

GONCALVES, L. G. G. et al. The South American Land Data Assimilation System (SALDAS) 5-Yr Retrospective Atmospheric Forcing Datasets. Journal of Hydrometeorology, v. 10, n. 4, p. 999–1010, 2009.

GOODWIN, B. K. Agricultural & Applied Economics Association Problems with Market Insurance in Agriculture. Source: American Journal of Agricultural Economics, v. 83, n. 3, p. 643–649, 2001.

HIGGINS, R. W., W. SHI, E. YAROSH, AND R. JOYCE, 2000. NCEP/Climate Prediction Center ATLAS No. 7. Disponível em: <https://www.cpc.ncep.noaa.gov/research_papers/ncep_cpc_atlas/7/toc.html>. Acesso em: 9 jul. 2019.

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

JOY HARWOOD, B. et al. Managing Risk in Farming: Concepts, Research, and Analysis. Abstract. [s.l: s.n.]. Disponível em: <<https://ageconsearch.umn.edu/bitstream/34081/1/ae990774.pdf>>. Acesso em: 3 abr. 2019.

LITTLE, R.; RUBIN, D. Statistical analysis with missing data. [s.l: s.n.].

MAPA. MAPA Indicadores. Disponível em: <<http://indicadores.agricultura.gov.br/atlasdoseguro/index.htm>>. Acesso em: 23 mar. 2019.

MOL, J. Estimativa de Precipitação por meio de Sensoriamento Remoto. Dissertação de Mestrado, p. 103, 2005.

MORITZ, S. et al. Comparison of different Methods for Univariate Time Series Imputation in R. 13 out. 2015.

MORITZ, S.; BARTZ-BEIELSTEIN, T. imputeTS: Time Series Missing Value Imputation version in T. The R Journal, v. XX, p. 1–12, 2017.

OLIVEIRA, L. F. C. DE et al. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. Revista Brasileira de Engenharia Agrícola e Ambiental, v. 14, n. 11, p. 1186–1192, 2010.

OZAKI, V. A. Taxa De Prêmio De Contratos De Seguro Agrícola. p. 347, 2005.

OZAKI, V. A. Em busca de um novo paradigma para o seguro rural no Brasil. *Revista de Economia e Sociologia Rural*, v. 46, n. 1, p. 97–119, mar. 2008.

QUIGGIN, J. C.; KARAGIANNIS, G.; STANTON, J. CROP INSURANCE AND CROP PRODUCTION: AN EMPIRICAL STUDY OF MORAL HAZARD AND ADVERSE SELECTION. *Australian Journal of Agricultural Economics*, v. 37, n. 2, p. 95–113, 1 ago. 1993.

QULGGIN, J.; KAFWGIANNIS, G.; STANTON, J. CROP INSURANCE AND CROP PRODUCTION: AN EMPIRICAL STUDY OF MORAL HAZARD AND ADVERSE SELECTION. *Australian Journal of Agricultural Economics*. [s.l.: s.n.]. Disponível em: <<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8489.1993.tb00531.x>>. Acesso em: 23 mar. 2019.

RUBIN, D. Inference and missing data. *academic.oup.com*, v. 63, n. 3, p. 581–592, 1976.

RUBIN, D. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the survey research methods section of the American Statistical Association*, v. 1, p. 20–34, 1978.

RUBIN, D. B. Multiple imputation for nonresponse in surveys. [s.l.: s.n.].

SANTOS, J. A. S. Análise de Eventos de Precipitação Extrema na Costa Leste do Nordeste Brasileiro Utilizando Conexão de Imagens de Satélites e de Radar Meteorológico. [s.l.] Universidade Federal de Alagoas, 2012.

SCHAFER, J. L. JOSEPH L. Analysis of incomplete multivariate data. [s.l.] Chapman & Hall, 1997.

SENTELHAS, P. C.; PEREIRA, A. R.; ANGELOCCI, L. R. Meteorologia agrícola, 2000. Disponível em: <<https://bdpi.usp.br/item/001073060>>. Acesso em: 5 jul. 2019

SILVA, M. J. C. DA. Imputação múltipla: comparação e eficiência em experimentos multiambientais. [s.l.] Universidade de São Paulo, 2012.

SILVA, V. B. S. et al. An Improved Gridded Historical Daily Precipitation Analysis for Brazil. *Journal of Hydrometeorology*, v. 8, n. 4, p. 847–861, 2007.

SKEES, J. R.; REED, M. R. Rate Making for Farm-Level Crop Insurance: Implications for Adverse Selection. *American Journal of Agricultural Economics*, v. 68, n. 3, p. 653, 1 ago. 1986.

SMITH, V. H.; GOODWIN, B. K. Crop Insurance, Moral Hazard, and Agricultural Chemical Use. *American Journal of Agricultural Economics*, v. 78, n. 2, p. 428, 1 maio 1996.

STRECK, N. A. et al. Associação da variabilidade da precipitação pluvial em Santa Maria com a Oscilação Decadal do Pacífico. *Pesq. agropec. bras.* [s.l.: s.n.]. Disponível em: <<https://ainfo.cnptia.embrapa.br/digital/bitstream/item/38208/1/44n12a01.pdf>>. Acesso em: 9 jul. 2019.

TARDIVO, G.; BERTI, A. The selection of predictors in a regression-based method for gap filling in daily temperature datasets. *International Journal of Climatology*, v. 34, n. 4, p. 1311–1317, 2014.

TEODORO, P. E. et al. Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Mato Grosso do Sul State, Brazil. *Meteorology and Atmospheric Physics*, v. 128, n. 2, p. 197–209, 2016.

TROUET, V.; OLDENBORGH, G. J. VAN. KNMI Climate Explorer: A Web-Based Research Tool for. *Tree-Ring Research*, v. 69, n. 1, p. 3–13, 2013.

TSUKAHARA, R. Y.; JENSEN, T.; CARAMORI, P. H. Utilização de Redes Neurais Artificiais para Preenchimento de Falhas em Séries Horárias de Dados Meteorológicos. *Congresso Brasileiro de Meteorologia*, p. 1–5, 2010.

VAN BUUREN, STEF AND OUDSHOORN, K. Flexible multivariate imputation by MICE. [s.l.] Leiden: TNO, 1999.

VAN BUUREN, S.; GROOTHUIS-OUUDSHOORN, K. Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, v. 45, n. 3, p. 1–67, 2011.

VENTURA, T. M. Preenchimento de Falhas de Dados Micrometeorológicos utilizando técnicas de Inteligência Artificial. [s.l.] Universidade Federal de Mato Grosso, 2012.

VIANELLO & ALVES. *Movimentos Atmosféricos: Aspectos Físicos e Observações*. [s.l.] Meteorologia Básica e Aplicações, 1991.

VICENTE-SERRANO, S. M. et al. A complete daily precipitation database for northeast Spain: Reconstruction, quality control, and homogeneity. *International Journal of Climatology*, v. 30, n. 8, p. 1146–1163, 2010.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, v. 30, n. 4, p. 377–399, 2011.

WISSMANN, J. A. et al. *Varia scientia revista multidisciplinar da Unioeste*. [s.l.] Universidade Estadual do Oeste do Paraná, 2006. v. 6

2. PREENCHIMENTO DE FALHAS DE PRECIPITAÇÃO PELA IMPUTAÇÃO MÚLTIPLA POR EQUAÇÕES ENCADEADAS UTILIZANDO DADOS EM PONTO DE GRADE

RESUMO

Precipitação é a variável meteorológica de grande importância para a produção agrícola. Alguns dos problemas das bases de dados meteorológicos são a descontinuidade e dados faltantes. Nesse sentido, a uso de dados de precipitação em ponto de grade pode ser uma solução. Para superar os problemas de dados faltantes e construir um banco de dados completo é necessário um processo de imputação. Dessa forma, o objetivo desse estudo foi realizar a imputação de séries de precipitação diária pelo método imputação múltipla por equações encadeadas (MICE), utilizando a informação de mês, ano e de estações virtuais. Foram utilizados quatro modelos, em que a precipitação diária dependia de: mês; mês e ano; precipitação diária em ponto de grade; mês, ano e precipitação diária em ponto de grade. Utilizou a raiz do erro quadrático médio (RMSE) como métrica e para checar as imputações, analisou a semelhança entre os dados observados e os dados imputados pelo Teste de Kolmogorov-Smirnov e também pelos gráficos da média e variância das imputações. O modelo com o maior número de variáveis foi escolhido para imputar os dados faltantes das séries de precipitação diária. O uso de dados de precipitação em ponto de grade mostrou a importância e as vantagens de serem utilizados como informação na imputação de dados de séries de precipitação diária.

Palavras-chave: Agricultura; Dados faltantes; MICE; Precipitação diária

ABSTRACT

Precipitation is the most important meteorological variable for agricultural production. Some of the problems of meteorological databases are discontinuity and missing data. In this sense, the use of gridded precipitation may be a solution. To overcome missing data problems and build a continuous database, an imputation process is required. Thus, the objective of this study was to estimate the daily missing values of daily precipitation using the multivariate imputation by chained equations (MICE), as well as information of month, year and gridpoint. Four models were used, in which the daily precipitation depended on: month; month and year; gridded daily precipitation; month, year and gridded daily precipitation. It used the root mean square error (RMSE) as a metric and to check the imputations, analyzed the similarity between the observed data and the data imputed by the Kolmogorov-Smirnov test and the mean and variance plot of the imputations. The model with the largest number of variables was chosen to impute missing data from the daily precipitation series. The use of gridded daily precipitation showed the importance and advantages of being used as information in the imputation of daily precipitation series data.

Keywords: Agriculture; Missing data; MICE; Daily precipitation

2.1. Introdução

A agricultura é uma das atividades mais antigas e importantes para o ser humano, sendo considerada fonte de matérias-primas, alimentos e energia. No entanto, muitos riscos ameaçam em maior ou menor grau todas as cadeias do agropecuária. Ao longo dos anos, os produtores criaram diversas maneiras de administrar seus riscos. Uma delas é a utilização do seguro rural como forma de transferência das consequências do risco (BRISOLARA, 2013). No entanto, o seguro rural apresenta alguns problemas, tais como a ausência de séries históricas longas e completas de clima e produtividade, entre outras, desencorajando portanto a oferta de seguros pelas companhias seguradoras (COBLE e KNIGHT, 2002; GOODWIN, 2001; QUIGGIN et al., 1993; SMITH e GOODWIN, 1996a).

O clima é uma das principais variáveis de impacto na produção agrícola, sendo que a precipitação é considerada a variável meteorológica de grande importância para a produção agrícola e o clima regional (TEODORO et al., 2016). No Brasil existem várias bases de dados meteorológicos de diferentes instituições. No entanto, um dos problemas dessas bases de dados é a descontinuidade, dados faltantes entre outros. Além disso, os maiores problemas são os de séries históricas diárias, considerando que os dados diários reduzem a quantidade de estações com dados confiáveis e séries longas de tempo, limitando a informação a ser utilizada na definição de futuros cenários climáticos (CARVALHO et al., 2017).

Os dados em ponto de grade, baseados em conjunto de observações, são uma excelente fonte de informações em pesquisas climatológicas. Esses conjuntos de dados em ponto de grade disponíveis, são baseados em dados observados por estações meteorológicas e em dados de satélite e vêm sendo amplamente utilizados na última década, em especial em trabalhos climatológicos que envolvem análises espaciais, minimizando problemas com limitações de dados com falhas e irregularidade de cobertura espacial (GONCALVES et al., 2009; CARDOSO e QUADRO, 2017). Uma vantagem adicional dos dados de pontos da grade é que longas séries temporais estão disponíveis para muitas variáveis climáticas (SILVA et al., 2007).

A escassez de séries climáticas contínuas comprometem os modelos agrometeorológicos de apoio a tomada de decisões na agricultura. De acordo com diversos autores, entre os quais Daly et al., (2004), Feng; Hu e Qian, (2004) e Vicente-Serrano et al.,

(2010), para superar esses problemas e construir um banco de dados contínuo é necessário um processo de imputação e controle de qualidade dos dados de precipitação.

Na década de 70, iniciaram-se várias técnicas de métodos de imputação múltipla, um dos destaques foi o método proposto por Rubin (1987), em que cada valor faltante é substituído por um conjunto de valores plausíveis, o que representa a incerteza sobre o valor a ser imputado. A partir disso, duas abordagens gerais para imputar dados multivariados surgiram: modelagem conjunta (JM) (SCHAFER, 1997a) e especificação totalmente condicional (FCS), também conhecido como imputação múltipla por equações encadeadas (MICE) (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

O MICE é uma abordagem flexível e prática para lidar com dados perdidos (WHITE et al., 2011), além disso, uma característica importante dessa técnica é sua capacidade de lidar com diferentes tipos de variáveis (contínua, binária, categórica), porque cada variável é imputada usando seu próprio modelo de imputação (VAN BUUREN, STEF AND OUDSHOORN, 1999; BUUREN; OUDSHOORN, 2000; VAN BUUREN; GROOTHUIS-OUDSHOORN, 2011).

Vários trabalhos vêm utilizando o MICE para imputações de dados, especialmente na área da saúde, no entanto muitos trabalhos se destacam na área da meteorologia. Van Buuren et al. (2006) e Santos e Silva (2013), estudaram a imputação de dados de velocidade do vento e utilizaram as localidades próximas como informações do modelo, enquanto que Wesonga (2015) utilizou o mês, ano e valor máximo de um período de dez dias como informações do modelo para a imputação da velocidade do vento. O estudo de Turrado et al. (2014) comparou métodos de imputação, entre eles o MICE, para irradiação solar. Um trabalho que se destaca na análise de imputação de dados para precipitação diária é o de Carvalho et al. (2017), que comparou o MICE com a Geoestatística.

Entre os vários problemas citados, os dados faltantes nas séries de precipitação diária devem ser destacados. Atualmente, o uso de dados de precipitação baseados em conjunto de observações possibilita uma boa cobertura em nível global, sobretudo em países com uma dimensão territorial extensa como o Brasil. Esses dados permitem o monitoramento contínuo e periódico ao longo do tempo por meio de diversas observações de uma mesma área. Nesse sentido, o objetivo desse estudo foi realizar a imputação de séries de precipitação diária pelo método MICE, utilizando para isso informações de dados em ponto de grade, e também a informações sobre o mês e ano. O algoritmo foi utilizado para fazer a imputação com diferentes modelos. Os resultados obtidos com cada modelo foram comparados e o modelo que apresentar melhor desempenho será utilizado para a estimativa das observações faltantes.

Assim, utiliza-se a metodologia de Imputação Múltipla por Equações Encadeadas, utilizando o método MICE para fazer imputações de valores diários de precipitação diária, criando diferentes séries incompletas, com diferentes modelos. Os resultados obtidos foram comparados, comumente utilizadas na estimativa de valores omissos, para estações climatológicas, que tem dados faltantes, localizadas no Paraná.

O restante do artigo está organizado da seguinte forma: a seção 2.2 apresenta a revisão da literatura. A seção 2.3 apresenta a metodologia utilizada para imputação dos dados de precipitação diária. As descrições dos dados utilizados na pesquisa, bem como, as etapas feitas nas análises estão descritas na seção 2.4. A Seção 2.5 apresenta os resultados e as discussões. Finalmente, a seção 2.6 aborda as observações finais.

2.2. Revisão da Literatura

As técnicas, chamadas de métodos de imputação simples são: a imputação pela média ou pela mediana da variável, a imputação pelo meio do vizinho mais próximo, a imputação hot deck, a imputação pela regressão linear e a imputação por meio da máxima verossimilhança. O fato do dado faltante ser imputado somente uma vez, faz com que a incerteza associada ao procedimento não seja agregada às estimativas geradas pelo banco completo, trazendo uma grande limitação a estes métodos (ENDERS, 2010).

As técnicas de métodos de imputação múltipla surgiram a partir da década de 70. Um método proposto foi a Imputação Múltipla (IM), por Rubin (1978), em que cada valor faltante é substituído por um conjunto de valores plausíveis, que representa a incerteza sobre o valor a ser imputado. Duas abordagens gerais para imputar dados multivariados surgiram: modelagem conjunta (JM) e especificação totalmente condicional (FCS), também conhecido como imputação multivariada por equações encadeadas (MICE).

Schafer (1997) desenvolveu várias técnicas de JM para imputação sob o modelo multivariado normal, o log-linear e o modelo de localização geral. A técnica JM envolve a especificação de distribuições por técnicas de Monte Carlo da cadeia de Markov (MCMC). Esta metodologia é interessante se a distribuição multivariada é uma descrição razoável dos dados.

O FCS especifica o modelo de imputação multivariada, para cada variável incompleta, por um conjunto de densidades condicionais. Partindo de uma imputação inicial, o FCS gera imputações por iterações. O FCS é atraente como alternativa ao JM quando não é

possível encontrar uma distribuição multivariada adequada. A ideia básica do FCS é bastante antiga e foi proposto usando uma variedade de nomes: relaxamento estocástico (KENNICKELL, 1991), imputação variável por variável (BRAND, 1999), regressão de comutação (VAN BUUREN et al., 1999), regressão sequencial (RAGHUNATHAN et al., 2001), amostrador ordenado pseudo-Gibbs (HECKERMAN et al., 2000), MCMC parcialmente incompatível (RUBIN, Donald B., 2003), imputação univariada iterada, FCS (BUUREN, 2007) e atualmente como MICE (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

Van Buuren et al. (2006) estudou a qualidade de métodos de imputação utilizando modelos linear e logística em um conjunto de dados de velocidades de vento na Irlanda e utilizou localidades próximas como informações do modelo. As correlações entre estas estações foram altas, variando de 0,59 a 0,84.

O estudo de Azur et al. (2011) fornece uma introdução ao método MICE com foco em aspectos práticos e desafios no uso desse método. O estudo analisa dados imputados de forma múltipla e faz uma comparação entre os valores observados e imputados para uso de serviços ambulatoriais e para o consumo de bebidas alcoólicas.

White, Royston e Wood (2011) mostraram como imputar variáveis categóricas e quantitativas, incluindo variáveis assimétricas. Também tem orientações sobre como especificar o modelo de imputação e quantas imputações são necessárias. Descreveu a análise prática de imputações múltiplas de dados, incluindo a construção de modelos e a sua verificação. Foram enfatizadas as limitações dos métodos e discutiu-se as possíveis armadilhas, ilustrando as ideias em conjunto de dados em saúde mental.

Santos e Silva (2013) fizeram estudos com dados de velocidade do vento para 47 estações com 25 anos de informações. Como as estações possuem dados faltantes, utilizou-se o método MICE para preencher os dados de velocidade do vento, no qual para preenchimento das observações faltantes de uma determinada estação utilizou-se de quatro estações meteorológicas geograficamente.

Turrado et al. (2014) avaliou a imputação multivariada de dados de irradiação solar de 100 estações meteorológicas. Os melhores resultados foram obtidos com o método MICE em comparação com outros métodos empregados neste campo, como a Ponderação de Distância Inversa (IDR) e a Regressão Linear Múltipla (MLR). O valor médio da raiz do erro quadrático médio (RMSE) das previsões para o algoritmo MICE foi 13,37%, enquanto que para o MLR foi 28,19% e 31,68% para o IDW.

O estudo de Wesonga (2015) fez imputação de dados para séries de velocidade do vento. As descobertas baseadas no MICE forneceram imputações de dados perdidas confiáveis da velocidade do vento. Nesse trabalho utilizou o mês, ano e valor máximo de um período de dez dias como informações do modelo para o modelo de imputação.

No trabalho de Carvalho et al. (2017), utilizou-se dois bancos de dados de precipitação diária, os registros são de várias estações meteorológicas para duas zonas climáticas homogêneas. Fizeram imputações de duas datas específicas, pelo método MICE e as técnicas de Geoestatística e fizeram a comparação. O modelo MICE foi satisfatório em comparação com a estimativa de precipitação obtida por técnicas Geoestatística.

2.3. Metodologia

A imputação múltipla é o método para banco de dados que possuem observações faltantes, que ocorrem em uma ou mais variáveis (RUBIN, Donald B., 1987, 1996). Esse método é uma técnica que difere substancialmente das abordagens estatísticas convencionais, com diversos modelos de imputação. Para cada variável incompleta, o usuário pode escolher um conjunto de modelos que serão utilizados para a imputação. Esse procedimento é útil para alocar grandes conjuntos de dados contendo centenas de variáveis (RUBIN, DB, 1978).

A imputação múltipla usa a distribuição dos dados observados para estimar um conjunto de valores plausíveis para os dados faltantes. Os componentes aleatórios são incorporados a esses valores estimados para refletir sua incerteza. Vários conjuntos de dados são criados e analisados individualmente, mas de forma idêntica, para obter um conjunto de estimativas de parâmetros. Por fim, as estimativas são combinadas para obter as estimativas globais, as variâncias e os intervalos de confiança (RUBIN, Donald B., 1987).

Duas abordagens gerais para imputar dados multivariados surgiram: modelagem conjunta (JM) e especificação totalmente condicional (FCS), também conhecido como imputação multivariada por equações encadeadas (MICE) (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

2.3.1. MICE

A imputação múltipla por equações encadeadas (MICE) é uma abordagem prática para resolver o problema de um banco de dados com valores faltantes em variáveis, no qual gera

imputações com base em um conjunto de modelos de imputação, um modelo para cada variável com observações faltantes (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011). O MICE também é conhecido como especificação totalmente condicional (BUUREN, 2007) e regressão sequencial multivariada (RAGHUNATHAN et al., 2001).

O MICE é um algoritmo que se inicia com um sorteio aleatório dos dados observados e imputa as observações faltantes, variável por variável. Uma característica importante do MICE é sua capacidade de lidar com diferentes tipos de variáveis (contínua, binária, categórica) porque cada variável é imputada usando seu próprio modelo de imputação (VAN BUUREN, STEF AND OUDSHOORN, 1999).

Inicialmente, todos os valores faltantes são preenchidos por amostragem aleatória simples com reposição dos valores observados. A primeira variável com valores faltantes, é regredida com base nas outras variáveis. Os valores perdidos da primeira variável são substituídos por sorteios da distribuição preditiva posteriori correspondente da primeira variável. Então, a próxima variável com valores faltantes, é regredida em todas as outras variáveis, e utilizando os valores imputados para a primeira variável. Novamente, a segunda variável é substituída por sorteios da distribuição preditiva posteriori da segunda variável. O processo é repetido para todas as variáveis com observações faltantes: isso é chamado de ciclo (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

Para estabilizar os resultados, o procedimento é geralmente repetido por vários ciclos (iterações), para produzir um único conjunto de dados imputado, e todo o procedimento é repetido n vezes para fornecer conjuntos de dados imputados. O número de iterações geralmente pode ser baixo, de 5 a 10 (WHITE et al., 2011).

A descrição formal do algoritmo MICE é descrita a seguir.

2.3.2. Algoritmo

As observações faltantes são substituídas por m conjuntos de valores simulados independentes, extraídos da distribuição preditiva posteriori das observações faltantes condicionado aos dados observados. Assume-se que as observações em Y e $\mathbf{X}$ correspondem a uma amostra aleatória simples de uma população de interesse com distribuição multivariada $P(\mathbf{X}, Y | \boldsymbol{\theta})$, assumindo que a distribuição multivariada de Y e $\mathbf{X}$ são completamente especificada, por um vetor de parâmetros desconhecidos, $\boldsymbol{\theta}$. O modelo

será discutido na seção 2.3.3. O problema é como obter a distribuição multivariada de θ , explícita ou implicitamente. Os passos seguintes são repetidos m vezes:

1. Obtém a distribuição posterior de θ por amostragem iterativa a partir de distribuições condicionais da forma $P(Y | \mathbf{X}, \theta)$.
2. Partindo de um simples sorteio das distribuições marginais observadas, a iteração t -ésima de equações encadeadas é um amostrador Gibbs que sucessivamente chama

$$\theta^{*(t)} : P(\theta | \mathbf{X}, Y^{obs})$$

$$Y^{*(t)} : P(Y | \mathbf{X}, \theta_1^{*(t)})$$

em que $Y^{(t)} = (Y^{obs}, Y^{*(t)})$ é a variável imputada na iteração t . Observa-se que para a imputação de $Y^{*(t-1)}$ somente $Y^{*(t)}$ é considerada, devido ao relacionamento com as variáveis. O θ^* é a estimativa do parâmetro para a t -ésima iteração das equações encadeadas (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

Esse processo é conhecido como imputação adequada porque incorpora todas as fontes de variabilidade e incertezas nos valores imputados, incluindo erros de previsão dos valores individuais e erros de estimação nos coeficientes ajustados do modelo de imputação (RUBIN, Donald B., 1987). O método para criar as m imputações é descrito na seção 2.3.4.

2.3.3. Modelo

Assume-se Y como uma variável com valores faltantes, na qual deseja-se imputar a partir de outras variáveis completas. Seja $\mathbf{X}_k$, com $k = 1, K, l$, sendo l variáveis completas, em que $\mathbf{X} = (X_1, K, X_k)$. Define-se por Y^{obs} as observações observadas, Y^{mis} as observações faltantes e n^{obs} para o número de observações com valores observados de Y .

Um modelo de regressão linear é a escolha mais comum do modelo para imputação de variáveis contínuas normalmente distribuídas:

$$Y | \mathbf{X}, \beta : N(\mathbf{X}\beta, \sigma^2)$$

Seja $\hat{\beta}$ uma estimativa do parâmetro (um vetor com k linha) da adaptação deste modelo das observações de Y . Seja $\mathbf{V}$ uma matriz de covariância estimada de $\hat{\beta}$, $\hat{\sigma}$ é a estimativa da raiz do erro médio quadrático. Em seguida, apresenta os parâmetros de imputação σ^* e β^* a partir da distribuição a posteriori de σ, β (RUBIN, Donald B., 1987).

Para σ^* , é calculado como:

$$\sigma^* = \hat{\sigma} \sqrt{(n^{obs} - k) / g}$$

em que g é uma variável χ^2 com $n_j^{obs} - k$ graus de liberdade. Para β^* , é calculado como:

$$\beta^* = \hat{\beta} + \frac{\sigma^*}{\hat{\sigma}} \mathbf{u}_1 \mathbf{V}^{1/2}$$

em que $\mathbf{u}_1$ é um vetor de k linha independentes e aleatórias de uma distribuição Normal Padrão; $\mathbf{V}^{1/2}$ é uma decomposição de Cholesky de $\mathbf{V}$.

Os valores imputados Y^* para cada observação faltante Y são então obtidos como:

$$Y_i = \mathbf{X}_i \beta^* + u_{2i} \sigma^*$$

em que u_{2i} é um valor aleatórias de uma distribuição Normal Padrão (WHITE et al., 2011).

2.3.4. Método da Média Preditiva

A ideia básica do método de média preditiva de correspondência (PMM - *Predictive Mean Matching*) é usar diferentes modelos de regressão para calcular a observação faltante. Cada conjunto de estimativas dos modelos será usado para gerar um único conjunto de dados imputados (HEITJAN, L F. , LITTLE, 1991). As estimativas dos parâmetros devem ser extraídas de sua distribuição posteriori dos dados, que está descrito na Seção 2.3.2.

A imputação múltipla dos Y^{mis} por meio de PMM é realizado pelo seguinte algoritmo (LITTLE e RODERICK, 1988; VINK et al., 2014):

1. Utiliza-se o modelo descrito na Seção 2.2.3 para calcular um vetor de parâmetros perturbados β^* .
2. Para cada observação faltante de Y com covariáveis $\mathbf{X}$, o procedimento padrão seria a próxima amostra de uma distribuição normal com $\mathbf{x}\beta^*$.
3. Identifique os q menores valores de $\Delta = \left| \mathbf{X}_h \hat{\beta} - \mathbf{X}_i \beta^* \right|$, $h = 1, K, n^{obs}$.
4. Escolhe um valor aleatório de (Δ_1, K, Δ_q) e passa ser valor imputado.
5. Repita os passos 1-4 m vezes, cada vez é salvo o conjunto de dados concluído.

O método PMM executa a imputação múltipla de acordo com a descrição desse algoritmo. Assim, o valor imputado é um valor observado em Y , cuja previsão com os dados observados corresponde de perto à predição perturbada. O método pode ser visto como uma generalização do *bootstrap* bayesiano aproximado, como descrito em RUBIN e SCHENKER (1986) e (RUBIN, Donald B., 1987). Neste trabalho foi usado $q = 3$, que teve um bom desempenho em um estudo de simulação (SCHENKER e TAYLOR, 1996).

2.3.5. Avaliação do Processo de Imputação

Para avaliar os modelos de imputação, várias medidas de desempenho são utilizadas na literatura, cada uma dessas medidas é "dimensionada" na medida em que expressa um erro médio de previsão de modelo nas unidades da variável de interesse. De acordo com Willmott e Matsuura (2005), a raiz do erro quadrático médio (RMSE) é uma das medidas de erro mais usadas na literatura climática e ambiental.

Considere y_i , $i = 1, K, n$, sendo o i -ésimo dado faltante de uma determinada variável, e $y_{imp,i}$ é o i -ésimo valor imputado para o i -ésimo dado faltante de uma determinada variável. A medida RMSE é definida como (CHEN et al., 2017; HYNDMAN, 2016):

$$RMSE(y, y_{imp}) = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{imp,i})^2}{n}}$$

2.3.6. Checando o Modelo de Imputação

Para verificar o modelo de imputação é preciso avaliar a convergência e ainda não há um método claro para determinar se o algoritmo MICE convergiu. O que geralmente é feito plota-se um ou mais parâmetros em relação ao número de iteração. Uma vez que o algoritmo tenha convergido, é importante inspecionar as imputações (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

Para inspecionar as imputações, a tarefa é identificar onde as imputações diferem significativamente dos valores observados. Em ambos os casos, a sinalização compara os valores imputados com os valores observados. Não há testes infalíveis dos pressupostos do procedimento de imputação (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

Baseado em Abayomi, Gelman e Levy (2008) e Raghunathan e Bondarenko (2007), pode-se comparar numericamente as distribuições empíricas dos dados observados e os dados imputados usando o teste de Kolmogorov-Smirnov (KS) para cada variável, levantando em consideração quando encontramos diferenças estatisticamente significativas, ou seja, rejeitam a igualdade para cada conjunto de dados imputado. Os gráficos de densidade empírica são sinalizadores para possíveis problemas com as estimativas imputadas, os gráficos de densidade empírica são representações visuais dos testes.

2.4. Descrição dos Dados

O estado do Paraná está localizado ao norte da região Sul do Brasil, ocupando 199.314 km² e conta com 399 municípios. O estado faz fronteiras com o estado de São Paulo, Santa Catarina, Mato Grosso do Sul e com os países Paraguai e Argentina e é banhado pelo oceano Atlântico. O clima paranaense é predominantemente subtropical com boa distribuição de chuvas anuais e verões suaves. A temperatura média anual fica entre 15 a 20°C e os índices pluviométricos oscilam de 1500 a 2500mm anuais, com aproximadamente, 70% dos dias sem chuvas (IBGE, 2019).

Atualmente, as redes de estações meteorológicas do Paraná são o IAPAR com 22 estações meteorológicas convencionais, o SIMEPAR com 37 estações meteorológicas automáticas, o INMET com 21 estações meteorológicas automáticas, e o INSTITUTO ÁGUAS DO PARANÁ com 348 municípios com posto pluviométrico (IAPAR, 2019).

Para a imputação de dados faltantes, utilizou-se sete séries de precipitação diária, expressas em milímetros (mm), provenientes de postos pluviométricos do estado do Paraná, apresentadas na Tabela 2.1. Os dados são públicos disponibilizados pelo Instituto Águas do Paraná pertencente ao Instituto Agrônomo do Paraná (IAPAR) e compilados pelo Grupo de Gestão em Seguros e Riscos (GESER), da Escola Superior “Luiz de Queiroz” (ESALQ/USP).

O Instituto Águas do Paraná tem 348 municípios com postos pluviométricos (IAPAR, 2019). No banco de dados do GESER tem disponíveis informações de 98 postos do Instituto Águas do Paraná, no período mínimo de 1931 e o período máximo de 2015. São postos com diferentes datas, de início e final, de coleta de informações e uma característica comum na maioria das séries é obter observações faltantes.

Foram selecionados postos que tinham no mínimo 30 anos de informações de precipitação. E uma característica desses postos, com longos períodos, e de possuírem

observações faltantes na série. Assim selecionou-se séries com diversas porcentagens de observações faltantes. Os postos selecionados foram Bela Vista do Paraíso, Jataizinho, Nova Londrina, Porto Santa Terezina, Santa Cruz de Monte Castelo, Tomazina e Vila Silva Jardim.

Cada série teve como data inicial o dia 01 de janeiro de 1979, e a data final é o dia 31 de dezembro, sendo que os anos são diferentes para cada série. Na Tabela 2.1 apresenta-se o nome dos postos enumeradas, o número de observações, a porcentagem de dados faltantes e informações de localização (longitude e latitude) e também o período de cada série.

Na Figura 2.1 apresenta a localização de cada estação. Os postos de Porto Santa Terezina e Vila Silva Jardim localizam nos municípios de Santa Mariana e Paranacity, respectivamente, e os demais postos tem o nome do próprio município.

Para se obter uma variável que auxilie a imputações dos valores faltantes das séries de precipitação, utilizou-se dados de precipitação diária em ponto de grade obtidos a partir de grades igualmente espaçadas com resolução de $0,50^\circ$ latitude $\times$ $0,50^\circ$ longitude, provenientes das séries históricas *KNMI Climate Explorer* (EXPLORER, 2019). As observações se referem a *Climate Prediction Center* (CPC) (CENTER, 2019) de precipitação diária, expressas em milímetros (mm), que compreende o período inicial de janeiro de 1979, o que justifica-se o período inicial das séries utilizadas nesse trabalho. Este produto é oriundo de observações provenientes de pluviômetros, e combina dados de mais de 30.000 estações ao redor do globo utilizando a técnica de Interpolação Ótima (GANDIN, 1965).

Tabela 2.1. Posto pluviométrico utilizadas nesse trabalho.

	Posto Pluviométrico	Número de		Latitude	Longitude	Período
		Observações	Dados Faltantes (%)			
1.	Bela Vista do Paraíso	12842	17.00	-22,95	-51,20	1979-2014
2.	Jataizinho	12419	37.00	-23,25	-50,98	1979-2012
3.	Nova Londrina	13514	11.70	-22,76	-52,98	1979-2015
4.	Porto Santa Terezina	12419	41.00	-23,12	-50,45	1979-2012
5.	Santa Cruz de Monte Castelo	13483	4.40	-22,96	-53,28	1979-2015
6.	Tomazina	13149	0.31	-23,76	-49,95	1979-2014
7.	Vila Silva Jardim	13514	9.30	-22,85	-52,07	1979-2015

Figura 2.1. Localizações das Estações.

No *KNMI Climate Explorer* (EXPLORER, 2019) na opção de selecionar os campos, foi selecionado o campo diário e escolheu observações de precipitação com pontos de grade mais próximos de $0,5^\circ \times 0,50^\circ$, marcando os intervalos de latitude e longitude para cada posto. Foram feitas *download* diversas séries com diferentes combinações de intervalos e calcula a correlação com a série do posto pluviométrico, a série que teve maior correlação foi escolhida para ser a variável da precipitação diária em ponto de grade. Os valores dos intervalos marcados e selecionados, para latitude e longitude, estão na Tabela 2.2.

Na Tabela 2.2, apresenta as correlações entre a série de precipitação diária do posto com a série de precipitação diária em ponto de grade obtido pelas grades. De acordo com Vedenov e Barnett (2004), fenômenos climáticos, como a chuva tendem a ser razoavelmente localizados, ou seja, mesmo uma distância relativamente pequena entre o posto de medição e o ponto de grade pode resultar em diferenças de valores. Destacando que as correlações, na maioria, foram altas.

Tabela 2.2. Intervalos selecionados para latitude e longitude e correlação entre o posto pluviométrico e as grades de $0,50^\circ \times 0,50^\circ$.

	Posto Pluviométrico	Latitude	Longitude	Correlação
1	Bela Vista do Paraíso	-23,08; -22,81	-51,33; -51,06	0,70
2	Jataizinho	-23,11; -23,38	-51,11; -50,84	0,73
3	Nova Londrina	-22,89; -22,62	-53,11; -52,84	0,47
4	Porto Santa Terezina	-23,25; -22,98	-50,58; -50,31	0,73
5	Santa Cruz de Monte Castelo	-23,09; -22,82	-53,41; -53,14	0,50
6	Tomazina	-23,50; -24,00	-49,50; -50,00	0,65
7	Vila Silva Jardim	-22,98; -22,71	-52,20; -51,93	0,57

Em princípio, para cada posto pluviométrico, montou-se o banco de dados com a variável precipitação diária e as variáveis: mês, ano e precipitação diária em ponto de grade. Somente a variável de precipitação diária tem observações faltantes. Assim considerou-se com Y : precipitação diária; X_1 : mês X_2 : ano X_3 : precipitação diária em ponto de grade.

Etapa 1

Retira os as observações faltantes, assim como os valores correspondentes das demais variáveis. Após isso, o procedimento são os passos a seguir:

1. Retira uma amostra y de Y e dos valores retirados foram considerados como observações faltantes em Y .
2. Calcula as imputações $Y^{*(t)}$, com $t=1, K, m$, das observações faltantes pelo algoritmo MICE, com 5 iterações e $m=5$ imputações pelo método PMM e com um modelo especificado.
3. Calcula raiz do erro quadrático médio $RMSE(y, Y^{*(5)}) = \sqrt{\sum_{i=1}^n (y_i - Y_i^{*(5)})^2 / n}$.
4. Repetir 1200 vezes do 1 ao 3.

No Passo 1, são retiradas de 10%, 20% e 30% de amostras de maneira aleatória, na qual as mesmas amostras para diferentes porcentagens. No Passo 2 são considerados quatro diferentes modelos, apresentados na Tabela 2.3. No passo 3, calcula a RMSE, em que y_i : a i -ésima observações da amostra de Y ; $Y_i^{*(5)}$: a i -ésima valor da imputação, na qual considera a

última imputação e última iteração do MICE e n : número de observações da amostra y_i . Exemplificando, para 10% de observações faltantes, as imputações feitas com o modelo M1, obteve 1200 valores da RMSE.

Etapa 2

Nessa etapa, o procedimento foram os seguintes passos:

1. Calcula as imputações de Y_i^* , pelo algoritmo MICE, com 5 iterações e $m=5$ imputações pelo método PMM e com um modelo especificado.
2. Avalia as densidades dos dados observados com as cinco imputações da última iteração, pelo Teste de Kolmogorov-Smirnov (RAGHUNATHAN, Trivellore e BONDARENKO, 2007) .
3. Faz os gráficos da média e variância por iterações, com o algoritmo MICE, pelo método PMM com 20 iterações e 5 imputações.
4. Faz o gráfico, para cada série, com os valores observados e imputados com o modelo escolhido na análise.

Tabela 2.3. Modelos utilizadas nas imputações.

Nome do Modelo	Variáveis explicativas
M1	X_1 : mês
M2	X_1 : mês e X_2 : ano
M3	X_3 : precipitação diária em ponto de grade
M4	X_1 : mês; X_2 : ano e X_3 : precipitação diária em ponto de grade

2.5. Resultados e Discussão

Nas Figuras 2.2 - 2.8 são apresentados os gráficos com os valores da média da RMSE obtidos com diferentes modelos (M1, M2, M3 e M4), com simulação de 10%, 20% e 30% de dados faltantes. Nota-se que a média das RMSE com o modelo M4 se manteve estável e apresentou os menores valores nas estações.

Na segunda etapa da análise, foi verificado se houve diferença entre os dados observados e os dados imputados, apresentou-se o valor p do Teste de Kolmogorov-Smirnov para todos os modelos, no qual os modelos M3 e M4, apresentou-se em pelo menos umas das imputações, o valor p maior que 0,05, indicando que há uma semelhança entre os dados observados e os dados imputados para as estações desse estudo (Tabela 2.4). Com esse teste foi possível confirmar a semelhança dos dados observados com os dados imputados, para o modelo M4, embora não idênticos.

O MICE faz imputações com um conjunto de iterações. Na Figuras 2.9 – 2.15, são apresentados os gráficos das médias e desvio padrão, sendo que cada linha é uma imputação com 20 iterações. Segundo Van Buuren (2018), a convergência ocorre quando as linhas são intercaladas e também quando é verificada que o desvio padrão entre diferentes sequências não é maior do que o desvio padrão em cada sequência, isso foi possível observar com as imputações da variável precipitação obtidas por MICE com o modelo M4, obtendo convergência nas imputações para precipitação diária.

Figura 2.2. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Bela Vista do Paraíso.

Figura 2.3. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Jataizinho.

Figura 2.4. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Nova Londrina.

Figura 2.5. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Porto Santa Terezina.

Figura 2.6. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Santa Cruz do Monte Castelo.

Figura 2.7. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Tomazina.

Figura 2.8. Média das RMSE obtidos com diferentes modelos, fez 10%, 20% e 30% de observações faltantes, para o posto pluviômetro de Vila Silva Jardim.

Segundo Collins et al. (2001) e Meng (1995), para fazer várias imputações obtendo viés mínimo e certeza máxima, deve-se usar todas as informações possíveis. Este princípio implica que o número de covariáveis deve ser escolhido o máximo possível, assim optou-se pelo modelo M4. Com esse modelo M4, obteve-se os menores valores da média da RMSE e a semelhança entre os dados observados e os dados imputados.

O modelo M4 foi escolhido para imputar as observações faltantes das séries de precipitação diária. Com o uso do algoritmo MICE faz-se 5 imputações e 20 iterações. Dos cinco vetores disponíveis, utiliza-se a quinta imputação para completar as séries de precipitação diária e fazer o gráfico com os valores observados e imputados com o modelo proposto (Figura 2.16- 2.22).

Tabela 2.4. Probabilidade do teste de Kolmogorov-Smirnov comparando a densidade dos dados observados com os dados imputados.

Estação	Modelo	Iteração				
		m=1	m=2	m=3	m=4	m=5
Bela Vista do Paraíso	M1	<0,01	0,08	0,53	0,96	>0,01
	M2	0,20	0,99	0,01	>0,01	>0,01
	M3	0,77	0,99	0,55	0,66	0,98
	M4	<0,01	0,01	0,78	<0,01	0,43
Jataizinho	M1	0,88	<0,01	0,14	0,99	0,63
	M2	0,01	<0,01	<0,01	0,01	0,05
	M3	0,95	0,87	0,88	0,99	0,92
	M4	0,07	<0,01	<0,01	<0,01	0,05
Nova Londrina	M1	<0,01	0,04	<0,01	<0,01	0,47
	M2	0,04	<0,01	<0,01	<0,01	0,86
	M3	0,99	0,56	0,9	0,68	0,99
	M4	0,18	<0,01	0,04	0,04	0,86
Porto Santa Terezina	M1	<0,01	<0,01	<0,01	<0,01	0,97
	M2	0,47	<0,01	<0,01	<0,01	0,01
	M3	<0,01	<0,01	<0,01	0,17	<0,01
	M4	0,17	<0,01	0,42	<0,01	<0,01
Santa Cruz Monte Castelo	M1	0,99	<0,01	0,09	0,86	0,01
	M2	0,53	0,44	0,85	<0,01	0,38
	M3	0,98	0,92	0,99	0,60	0,70
	M4	0,13	0,64	0,81	0,16	0,38
Tomazina	M1	0,66	0,75	0,87	0,05	<0,01
	M2	0,49	0,90	0,39	0,77	0,03
	M3	0,52	0,99	0,99	0,94	0,78
	M4	0,75	0,99	0,98	0,80	0,99
Vila Silva Jardim	M1	0,03	0,45	0,87	0,99	0,88
	M2	0,95	0,13	0,99	0,65	0,30
	M3	0,90	0,97	0,88	0,99	0,87
	M4	0,14	0,84	0,62	0,99	0,30

Figura 2.9. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Bela Vista do Paraíso.

Figura 2.10. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Jataizinho.

Figura 2.11. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Nova Londrina.

Figura 2.12. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Porto Santa Terezinha.

Figura 2.13. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Santa Cruz do Monte Castelo.

Figura 2.14. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Tomazina.

Figura 2.15. Gráfico das imputações, com 5 imputações e 20 iterações, da variável precipitação diária, para o posto pluviômetro de Vila Silva Jardim.

Enquanto Carvalho et al., (2017) comparou a técnica do MICE com a técnica de Krigagem e Co-Krigagem e obteve que o MICE produziu melhores estimativas. Neste trabalho utiliza-se somente a técnica do MICE e testa várias covariáveis que contribuam para a imputação e tendo melhores resultados para a utilização das variáveis mês, ano e precipitação diária em ponto de grade.

Figura 2.16. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Bela Vista do Paraíso.

Figura 2.17. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Jataizinho.

Figura 2.18. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Nova Londrina.

Figura 2.19. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Porto Santa Terezina.

Figura 2.20. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Santa Cruz do Monte Castelo.

Figura 2.21. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Tomazina.

Figura 2.22. Gráfico da série de precipitação diária com os valores imputados, com o modelo M4, para o posto pluviômetro de Vila Silva Jardim.

Nenhum método funciona melhor em todas as situações. O consenso é avaliar a convergência com uma combinação de ferramentas (VAN BUUREN, 2018). Assim, esse estudo mostrou a importância de considerar vários modelos para o método de imputação múltipla para dados faltantes.

2.6. Conclusão

A aplicação da metodologia MICE utilizando um modelo com as variáveis mês, ano e precipitação diária em ponto de grade produziu melhores estimativas para a imputação, das observações faltantes, da precipitação diária dos postos pluviométricos. Atualmente, o uso de dados de precipitação, obtidos por sensores remotos orbitais possibilita boa cobertura em nível global e mostrou a importância e as vantagens de utilizar como informação na imputação de dados de séries de precipitação diária.

Esse fato beneficia o estudo de preenchimento de falhas de precipitação diário, tendo em vista que há poucos estudos na área, devido à dificuldade de obter séries longas de observações diárias.

Nesse sentido, o presente estudo contribui para o avanço da literatura através da aplicação da metodologia MICE para série de precipitação diária e a utilização das

informações de dados de satélites para contribuição da imputação dos valores faltantes de série de fenômenos climáticos.

Teoricamente, a aplicação dessa metodologia na região de estudo (Estado do Paraná) deve incentivar os institutos de pesquisas nas áreas de meteorologia, a utilizar uma metodologia MICE para fazer a imputação dos valores faltantes das séries e assim obter séries completas para melhor compreensão do clima, das variações e alterações climáticas da região.

Ao longo do estudo, foram encontrando alguns problemas limitantes. Entre eles destacam-se as séries de períodos curtos, devido a mudanças de localização ou por desativação do posto pluviométrico. Dessa forma, trabalhou-se somente com sete séries longas de precipitação diária do estado do Paraná.

O avanço da pesquisa na área de imputação de dados e o conjunto de medidas verificadas, poderão tornar o MICE uma importante metodologia para imputação de dados de precipitação diária e garantindo uma série temporal completa para análises de vários setores, dentre eles a agricultura, pois é uma informação de suma importância para os setores de seguro rural, financiamento e planejamento das atividades agrícolas.

Referências

- ABAYOMI, K.; GELMAN, A.; LEVY, M. Diagnostics for multivariate imputations. p. 273–291, 2008.
- AZUR, M. J. et al. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, v. 20, n. 1, p. 40–49, 1 mar. 2011.
- BRAND, J. P. L. Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. [s.l.] University Rotterdam, 1999.
- BRISOLARA, C. S. Proposições para o desenvolvimento do seguro de receita agrícola no Brasil: p. 239, 2013.
- BUUREN, S.; OUDSHOORN, C. Multivariate imputation by Chained equations: MICE V1.0 user's manual. TNO, 2000.
- BUUREN, S. VAN. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, v. 16, n. 3, p. 219–42, 2007.
- CARDOSO, C. S.; QUADRO, M. F. L. Análise comparativa de dados de precipitação gerados pelo “Climate Prediction Center – CPC” versus dados observados para o Sul do Brasil Camila. *Revista Brasileira de Geografia Física*, v. 10, n. 4, n. ISSN:1984-2295, p. 1180–1198, 2017.

CARVALHO, J. R. P. DE et al. Model for Multiple Imputation to Estimate Daily Rainfall Data and Filling of Faults Modelo de Imputação Múltipla para Estimar Dados de Precipitação Diária e Preenchimento de Falhas. v. 32, n.4, p. 575–583, 2017.

CENTER, N. C. P. NOAA's Climate Prediction Center. Disponível em: <<https://www.cpc.ncep.noaa.gov/>>. Acesso em: 8 jun. 2019.

CHEN, C.; TWYLCROSS, J.; GARIBALDI, J. M. A new accuracy measure based on bounded relative error for time series forecasting. PLOS ONE, v. 12, n. 3, p. e0174202, 24 mar. 2017.

COBLE, K. H.; KNIGHT, T. O. Crop Insurance as a Tool for Price and Yield Risk Management. In: A Comprehensive Assessment of the Role of Risk in U.S. Agriculture. Boston, MA: Springer US, 2002. p. 445–468.

COLLINS, L. et al. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, v. 6, n.4, p. 330–351, 2001.

DALY, C. et al. A probabilistic-spatial approach to the quality control of climate observations. Disponível em: <https://www.researchgate.net/publication/252891043_A_probabilistic-spatial_approach_to_the_quality_control_of_climate_observations>. Acesso em: 1 abr. 2019.

ENDERS, C. K. Applied missing data analysis. [s.l.] Guilford Press, 2010.

EXPLORER, K. C. Climate Explorer: Starting point. Disponível em: <<https://climexp.knmi.nl/start.cgi>>. Acesso em: 8 jun. 2019.

FENG, S.; HU, Q. I.; QIAN, W. QUALITY CONTROL OF DAILY METEOROLOGICAL DATA IN CHINA, 1951-2000: A NEW DATASET. INTERNATIONAL JOURNAL OF CLIMATOLOGY Int. J. Climatol, v. 24, p. 853–870, 2004.

GANDIN, L. S. Objective analysis of meteorological fields. Jerusalem: Israel program for scientific translations, 1965.

GONCALVES, L. G. G. et al. The South American Land Data Assimilation System (SALDAS) 5-Yr Retrospective Atmospheric Forcing Datasets. Journal of Hydrometeorology, v. 10, n. 4, p. 999–1010, 2009.

GOODWIN, B. K. Agricultural & Applied Economics Association Problems with Market Insurance in Agriculture. Source: American Journal of Agricultural Economics, v. 83, n. 3, p. 643–649, 2001.

HECKERMAN, D.; CHICKERING, D. M.; KADIE, C. 2. Consistent Dependency Networks. Journal of Machine Learning Research, v. 1, p. 49–75, 2000.

HEITJAN, L F., LITTLE, R. J. A. Multiple imputation for the fatal accident reporting system. Journal of the Royal Statistical Society, v. 40, n.1, p. 13–29, 1991.

HYNDMAN, R. J. forecast: Forecasting functions for time series and linear models. 2016.

IAPAR. Instituto Agrônômico do Paraná - IAPAR. Disponível em: <<http://www.iapar.br/>>. Acesso em: 11 jul. 2019.

IBGE. IBGE | Portal do IBGE. Disponível em: <<https://www.ibge.gov.br/>>. Acesso em: 11 jul. 2019.

KENNICHELL, A. B. Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. . Survey Research Methods Section, p. 1–10, 1991.

LITTLE, R. J. A. Missing-Data Adjustments in Large Surveys. Journal of Business & Economic Statistics, v. 6, n. 3, p. 287–296, jul. 1988.

MENG, X. Multiple imputation with uncongenial sources of input (with discusson). Statistical Science, v. 10, p. 538–573, 1995.

QUIGGIN, J. C.; KARAGIANNIS, G.; STANTON, J. CROP INSURANCE AND CROP PRODUCTION: AN EMPIRICAL STUDY OF MORAL HAZARD AND ADVERSE SELECTION. Australian Journal of Agricultural Economics, v. 37, n. 2, p. 95–113, 1 ago. 1993.

RAGHUNATHAN, T.; BONDARENKO, I. Diagnostics for Multiple Imputations. Ssrn, n. 2003, p. 1–14, 2007.

RAGHUNATHAN, T. E. et al. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models Key Words: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations. Survey Methodology, v. 27, n. 1, 2001.

RUBIN, D. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. Proceedings of the survey research methods section of the American Statistical Association, v. 1, p. 20–34, 1978.

RUBIN, D. B. Multiple imputation for nonresponse in surveys. [s.l: s.n.].

RUBIN, D. B. Multiple Imputation after 18+ Years. Journal of the American Statistical Association, v. 91, n. 434, p. 473–489, jun. 1996.

RUBIN, D. B. Nested multiple imputation of NMES via partially incompatible MCMC. Statistica Neerlandica, v. 57, n. 1, p. 3–18, 1 fev. 2003.

RUBIN, D. B.; SCHENKER, N. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, v. 81, n. 394, p. 366–374, jun. 1986.

SANTOS, A. T. S. DOS; SILVA, C. M. S. E. Seasonality, Interannual Variability, and Linear Tendency of Wind Speeds in the Northeast Brazil from 1986 to 2011. The Scientific World Journal, v. 2013, p. 1–10, 2013.

SCHAFER, J. L. (JOSEPH L. Analysis of incomplete multivariate data. [s.l.] Chapman & Hall, 1997a.

SCHAFER, J. L. (JOSEPH L. . Analysis of incomplete multivariate data. [s.l.] Chapman & Hall, 1997b.

SCHENKER, N.; TAYLOR, J. M. G. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, v. 22, n. 4, p. 425–446, 1996.

SILVA, V. B. S. et al. An Improved Gridded Historical Daily Precipitation Analysis for Brazil. *Journal of Hydrometeorology*, v. 8, n. 4, p. 847–861, 2007.

SMITH, V. H.; GOODWIN, B. K. Crop Insurance, Moral Hazard, and Agricultural Chemical Use. *American Journal of Agricultural Economics*, v. 78, n. 2, p. 428, maio 1996.

TEODORO, P. E. et al. Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Mato Grosso do Sul State, Brazil. *Meteorology and Atmospheric Physics*, v. 128, n. 2, p. 197–209, 2016.

TURRADO, C. et al. Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions. *Sensors*, v. 14, n. 11, p. 20382–20399, 29 out. 2014.

VAN BUUREN, STEF AND OUDSHOORN, K. Flexible multivariate imputation by MICE. [s.l.] Leiden: TNO, 1999.

VAN BUUREN, S. et al. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, v. 76, n. 12, p. 1049–1064, 2006.

VAN BUUREN, S. Flexible imputation of missing data. [s.l.] Chapman and Hall/CRC, 2018.

VAN BUUREN, S.; BOSHUIZEN, H. C.; KNOOK, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, v. 18, n. 6, p. 681–694, 30 mar. 1999.

VAN BUUREN, S.; GROOTHUIS-OUUDSHOORN, K. Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, v. 45, n. 3, p. 1–67, 2011.

VEDENOV, D. V.; BARNETT, B. J. Efficiency of weather derivatives as primary crop insurance instruments. *Journal of Agricultural and Resource Economics*, v. 29, n. 3, p. 387–403, 2004.

VICENTE-SERRANO, S. M. et al. A complete daily precipitation database for northeast Spain: Reconstruction, quality control, and homogeneity. *International Journal of Climatology*, v. 30, n. 8, p. 1146–1163, 2010.

VINK, G. et al. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, v. 68, n. 1, p. 61–90, 2014.

WESONGA, R. On multivariate imputation and forecasting of decadal wind speed missing data. *SpringerPlus*, v. 4, n. 1, p. 12, 13 dez. 2015.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, v. 30, n. 4, p. 377–399, 2011.

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, p. 79--82, 2005.

3. COMPARAÇÃO DE MÉTODOS DE IMPUTAÇÃO NAS ABORGAGENS UNIVARIADA E MÚLTIPLA

RESUMO

A qualidade e a disponibilidade de dados são essenciais para produzir boas previsões de precipitação diária, uma vez que previsão é uma informação importante para o planejamento das atividades agrícolas. Para isso é importante dispor de dados observacionais de longo prazo, no entanto alguns dos problemas das bases de dados meteorológicos é de dados faltantes, dessa forma é necessário um processo de imputação. Portanto, o objetivo foi comparar e avaliar métodos de imputação, nas abordagens univariada e múltipla, para dados de precipitação diária. Na abordagem univariada, utilizou diferentes configurações filtro de Kalman, Média Móvel Ponderada e Decomposição Sazonal. Na abordagem múltipla, utilizou a imputação múltipla com cadeias encadeadas (MICE), com diferentes modelos. Fez a estimativa de dados faltantes em série de precipitação diária, em que os dados faltantes foram gerados de maneira aleatório e em trechos e utilizou a raiz do erro quadrático médio (RMSE) como métrica. Os resultados identificaram o método de Filtro de Kalman forneceu os menores valores de RMSE, para todos os cenários de dados faltantes. O Filtro de Kalman poderá ser uma importante metodologia para imputação de dados de precipitação diária e garantido uma série temporal completa para análises de vários setores, dentre eles a agricultura.

Palavras-chave: Filtro de Kalman; MICE; Precipitação diária; Séries temporais

ABSTRACT

The quality and availability of data are essential to produce good accuracy of daily precipitation, once the forecast is important information for the planning of agricultural activities. For this it is important to have long-term observational data, however some of the problems of the meteorological databases are missing values, so an imputation process is necessary. Therefore, the objective was to compare and evaluate imputation methods, in the univariate and multiple approaches, for daily precipitation data. In the univariate approach, we used different configurations the Kalman filters, weighted moving average and seasonal decomposition. In the multiple approach, he used the multivariate imputation by chained equations (MICE), with different models. It estimated the missing values in a series of daily precipitation, in which the missing data were generated randomly and in parts and used the root mean square error (RMSE) as a metric. The results identified the Kalman filter method provided the lowest RMSE values for all missing data scenarios. The Kalman filter can become an important methodology for imputation of daily precipitation data and insured a complete time series for analyzes of several sectors, among them agriculture.

Keywords: Kalman filters; MICE; Daily precipitation; Time series

3.1. Introdução

Uma das atividades mais importantes para o Brasil é a agricultura, tanto do ponto de vista econômico quanto social. Ressalta-se que a agricultura é uma atividade amplamente dependente de fatores climáticos e dos recursos naturais. No entanto, os riscos climáticos representam uma ameaça para a produtividade agrícola (OZAKI, 2005).

Para quantificar o risco de eventos e suas consequências, a previsão de ocorrência de fenômenos meteorológicos adversos na agricultura é de suma importância para os setores de seguro rural, financiamento e planejamento das atividades agrícolas (OZAKI, 2008).

Para a identificação do clima de uma região e das variações climáticas é importante dispor de dados observacionais de longo prazo (DALY et al., 2004; FENG et al., 2004). No entanto, vários problemas acontecem nas estações climáticas, como falta de registro de dados de precipitação em estações pluviométricas, devido a problemas com os aparelhos de coleta e a ausência do operador em determinadas épocas, comprometendo a continuidade das informações (OLIVEIRA et al., 2010a).

Para superar esse problema, desde a década de 1980, surgem técnicas estatísticas envolvendo a substituição de dados faltantes por estimativas de valores plausíveis a serem completos em uma série com dados perdidos. Na literatura estatística, essa técnica é conhecida como imputação de dados faltantes e seu uso tem sido generalizado e estendido para outras áreas (CARVALHO et al., 2017; COLLINS et al., 2001; FRASER e YAN, 2007; NUNES et al., 2010; RUBIN, 1996).

Nos métodos de imputação múltipla, um dos destaques foi o método proposto por Rubin (1987), em que cada valor faltante é substituído por um conjunto de valores plausíveis, o que representa a incerteza sobre o valor a ser imputado. A partir disso, surgiram duas abordagens gerais para imputar dados multivariados: modelagem conjunta (JM) (SCHAFER, 1997a) e especificação totalmente condicional (FCS), também conhecida como imputação múltipla por equações encadeadas (MICE) (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

O MICE é uma abordagem flexível e prática para lidar com dados perdidos (WHITE et al., 2011), além disso, uma característica importante dessa técnica é sua capacidade de lidar com diferentes tipos de variáveis (contínua, binária, categórica), porque cada variável é imputada usando seu próprio modelo de imputação (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

As técnicas de imputação múltipla dependem das correlações entre variáveis para estimar valores para os dados faltantes. No entanto, os algoritmos multivariados existentes não podem ser aplicados ou não possuem uma boa execução para séries temporais, pois não possuem mais de um variável (MORITZ et al., 2015b; MORITZ e BARTZ-BEIELSTEIN, 2017).

Algumas das técnicas de imputação de séries temporais são: Interpolação, Filtro de Kalman, Persistência, Média Móvel Ponderada, Imputação Clássica, Amostra Aleatória, Decomposição Sazonal e Sazonalidade por partes. Atualmente, essas técnicas de imputações são exclusivas para séries temporais univariadas e inclui várias opções/algoritmos possíveis para sua implementação (DEMIRHAN e RENWICK, 2018; MORITZ e BARTZ-BEIELSTEIN, 2017). Além disso, esses métodos estatísticos se apresentam como bem-sucedidos na imputação de valores faltantes para diferentes tipos de dados e também de fácil implementação (DEMIRHAN e RENWICK, 2018).

Os métodos de Imputação Clássica, Persistência, Amostra Aleatória são simples e rápidos. Enquanto que os métodos de Interpolação, Filtro de Kalman, Média Móvel Ponderada, Decomposição Sazonal e Sazonalidade por partes são algoritmos mais avançados e precisam de mais tempo de computação. Observando todos os métodos de imputação disponíveis, nenhum algoritmo pode ser apontado como o melhor. O desempenho da imputação é sempre muito dependente das características da série temporal de entrada. Para séries temporais com sazonalidade, em geral os métodos de Filtro de Kalman e Sazonalidade por partes apresentam melhor desempenho. Em geral, para a maioria das séries temporais, um algoritmo de Filtro de Kalman, Interpolação e Decomposição Sazonal produzem os melhores resultados (MORITZ e BARTZ-BEIELSTEIN, 2017).

Todas as abordagens mencionadas, independentemente das suas classificações, dependem da disponibilidade de dados precipitação. Abordagens semelhantes à regressão precisam de variáveis auxiliares para alimentar informações meteorológicas ou espaciais na análise (SHAKYA et al., 2017). Em relação à disponibilidade de dados de precipitação, os dados faltantes são um desafio importante na previsão de precipitação a curto e a longo prazo. É muito comum em dados meteorológicos ter observações faltantes devido aos dispositivos de medição defeituosos em estações meteorológicas (BROOKS et al., 2016). O impacto negativo de dados faltantes é visto com frequência em dados da série temporal. Se a série é observada em intervalos de tempo irregulares, a frequência da série não pode mais ser uniforme devido às observações faltantes e acaba sendo uma série irregularmente observada. Assim, é preciso imputar os dados faltantes, antes de aplicar análise completa. Alternativamente, podemos usar

métodos de computação simples, que são capazes de manipular dados de séries temporais com observações faltantes (PEDRO e COIMBRA, 2015). No entanto, todas as abordagens simples, resultam em perda de eficiência, complicações no manuseio dos dados e viés devido à falta de precisão (BARNARD e MENG, 1999). Outra abordagem é usar métodos estatísticos gerais desenvolvidos para séries temporais irregularmente (NIEBUHR, 2017). A principal desvantagem dessas técnicas é a determinação do horizonte de previsão, que pode não ser claro para séries irregulares. Assim sendo, uma das maneiras mais confiáveis de lidar com valores faltantes em séries temporais é imputá-los usando abordagens estatísticas precisas.

Neste artigo, concentra-se na comparação e avaliação de métodos de imputação nas abordagens univariadas e múltiplas, para dados de precipitação. Na abordagem univariada, utilizou-se diferentes configurações filtro de Kalman, Média Móvel Ponderada e Decomposição Sazonal. Na abordagem múltipla, utilizou-se o método MICE, com diferentes modelos. Portanto, o objetivo desse estudo foi realizar a imputação de séries de precipitação diária pelo método de imputação de univariadas e múltiplas.

O restante do artigo está organizado da seguinte forma: a seção 3.2 apresenta a revisão de literatura. A seção 3.3 apresenta a metodologia utilizada para imputação dos dados de precipitação diária. As descrições dos dados utilizados na pesquisa, bem como, as etapas feitas nas análises estão descritas na seção 3.4. A Seção 3.5 apresenta os resultados e as discussões. Finalmente, a seção 3.6 aborda as considerações finais.

3.2. Revisão da Literatura

3.2.1. Imputação de Dados

No artigo de Santos e Silva (2013) conduziram estudos com dados de velocidade do vento para 47 estações com 25 anos de informações. Devido ao fato das estações possuírem dados faltantes, os autores utilizaram o método MICE para preencher os dados de velocidade do vento, no qual para as observações faltantes de uma determinada estação, utilizou-se de quatro estações meteorológicas geograficamente semelhantes.

Turrado et al. (2014) avaliou a imputação multivariada de dados de irradiação solar de 100 estações meteorológicas. Os melhores resultados foram obtidos com o método MICE em comparação com outros métodos empregados neste campo, como a Ponderação de Distância Inversa (IDR) e a Regressão Linear Múltipla (MLR). O valor médio da raiz do erro quadrático

médio (RMSE) das previsões para o algoritmo MICE foi de 13,37%, enquanto que para o MLR foi de 28,19% e 31,68% para o IDW.

No artigo de Bokde et al. (2018), utilizou-se o algoritmo PSF (*Pattern Sequence Forecasting*) que identifica padrões em dados de séries temporais e prevê valores usando características periódicas das observações. Inicialmente, os autores fizeram uma validação cruzada utilizando o método PSF de 10% a 50%, com intervalos de 10%, de dados faltantes com diferentes tamanhos trechos, sendo que cada análise possuía dois trechos, um no começo e o outro fim da série, com o objetivo de selecionar o tamanho ideal do trecho do início e do trecho do final da série com base na menor raiz do erro quadrático médio (RMSE) e após isso fez-se comparações de métodos de imputações variando a quantidade de dados faltantes de 10 a 60% em intervalos crescentes de 10% nos trechos selecionados no primeiro instante. Os dados utilizados foram: taxa de fluxo de água, velocidade de tráfego e temperatura média mensal.

No artigo de Demirhan e Renwick (2018), fez-se estimativas de dados faltantes de séries de irradiação solar minuto a minuto, hora em hora, diária e semanal. Criou-se dados faltantes em trechos e de maneira aleatória nas séries. Os trechos faltantes foram criados nas partes inicial, central e final da série, com pequenas, médias e grandes trechos de 5%, 10%, 25% e 50% de dados faltantes. De maneira aleatória, foi criado 70% de dados faltantes distribuídos na primeira ou segunda metade da série. Para comparar a precisão das estimativas, utilizou-se dos métodos de Interpolação, Filtro de Kalman, Persistência, Média Móvel Ponderada, Imputação Clássica, Amostra Aleatória, Decomposição Sazonal, Sazonalidade por partes. As interpolações, com opção linear e de Stineman, e filtragem de Kalman, com opção de modelo estrutural e suavização, foram precisas para as séries de minuto a minuto e de hora em hora. E para as séries diária e semanal, o método de Média Móvel Ponderada foi o mais preciso. As métricas comparativas utilizadas foram: erro absoluto médio relativo (MAE), erro escalonado médio (MASE) e raiz do erro quadrático médio (RMSE).

3.2.2. Imputação de Dados de Precipitação

No trabalho de Nascimento et al. (2016), preencheram-se falhas em dados diários da variável meteorológica (precipitação pluvial) a partir da correlação com dados sintéticos adquiridos do CPC (*Climate Prediction Center*). Os dados utilizados foram de precipitação

pluvial da série observada (INMET), com a série sintética (CPC) para sete estações meteorológicas ao longo do rio Amazonas. Para o preenchimento de falhas, os dados do CPC foram observados em quatro pontos próximos da estação em estudo, em seguida escolheu o ponto de melhor correlação para posteriormente completar a série de dados.

Como os dados precipitação pluvial da série observada (INMET) apresentaram períodos com ausência de dados (alguns meses ou até anos consecutivos), foi feita a interpolação por modelos técnicos do CPC (método de análise objetivo operacional).

A análise foi uma regressão linear, no qual a variável independente foi composta dos dados do INMET e a variável dependente foi composta dos dados do CPC. Posteriormente foram feitas as correlações entre as variáveis. A média das correlações foi de 0,79, considerada alta, e conclui-se que para o preenchimento das falhas de precipitação é satisfatório o uso de dados sintéticos adquiridos do CPC.

Mello, Kohls e Oliveira (2017) estudaram métodos para preenchimento de falhas pluviométricas, que foram: ponderação regional, regressão linear simples e múltipla e ponderação regional com base em regressões lineares. Utilizou-se séries mensais completas da região de Joinville (SC) e fez-se uma comparação de dados imputados com dados observados no qual o método. Os resultados mostraram que o método de regressão linear múltipla apresentou os valores mais próximos do real, mostrando-se como o método mais adequado para o preenchimento de falhas na região de estudo.

O trabalho Cardoso e Quadro (2017), analisou-se o desempenho dos dados de precipitação do CPC para região Sul do Brasil, comparando com dados observados em estações meteorológicas. Foram utilizados dados de precipitação de 81 estações localizadas nos três estados da região Sul do Brasil, disponibilizados pela Agência Nacional de Águas (ANA) e Instituto Nacional de Meteorologia (INMET) e dados diários de precipitação fornecidos pelo CPC, com resolução espacial de $0,50^{\circ} \times 0,50^{\circ}$, no período de 1979 a 2015. A análise estatística dos dados mostrou que o CPC possui bom desempenho em representar a precipitação no Sul do Brasil. Das séries analisadas, as estações possuem percentual de falhas menor que 20% e também superiores a 40% de observações faltantes. Esse trabalho mostrou a importância de uma análise detalhada das séries temporais, já que falhas sucessivas comprometem resultados produzidos em estudos climatológicos que exigem séries contínuas. O estudo de Silva et al. (2007) e Sena et al. (2012) também mostrou que os dados do CPC podem ser utilizados para estudos de precipitação na região Sul do Brasil.

O trabalho de Carvalho et al. (2017) utilizou registros de precipitação diária de várias estações meteorológicas para duas zonas climáticas homogêneas. Foram feitas as imputações

de dados faltantes pelo método MICE e as técnicas de Geoestatística, e fizeram a comparação dessas técnicas. Dentre os quais, o método MICE foi satisfatório em comparação com a estimativa de precipitação obtida por técnicas de Geoestatística.

3.3. Metodologia

3.4. Mecanismo de Dados Faltantes

Na prática, o mecanismo por trás dos dados faltantes não pode é conhecido exatamente. Quando não há nenhum mecanismo específico dos dados faltantes, é chamado Ausência Totalmente Aleatória (MCAR). Quando existem outras variáveis que afetam a existência dos dados faltantes e a probabilidade de ocorrência é independente, esse mecanismo é chamado Ausência Aleatória (MAR). Ao contrário disso, se a probabilidade de ter dados faltantes está associada ao próprio valor, é chamado de Ausência Não-Aleatórias (NMAR) (LITTLE e RUBIN, 2002).

Para o caso da NMAR, precisamos saber o motivo pelo qual os dados estão faltando (MORITZ et al., 2015b). De acordo com Oliveira et al. (2010), para os dados de precipitação diária, podemos associar dados faltantes com problemas com os aparelhos de coleta e a ausência do operador em determinadas épocas. Então, o mecanismo por trás dos dados faltantes de precipitação diária pode ser tomado como MCAR ou MAR, pois de acordo com Rubin (1976), ambos permitem imputar valores perdidos sem conhecer as razões específicas de tê-los.

3.5. Métodos de Imputação

3.5.1. Filtro de Kalman

O filtros de Kalman é aplicados em dois estágios e são baseados principalmente nos modelos de espaço de estados, apresentado a seguir (FUNG, 2006):

$$\begin{cases} x_t = F_t x_{t-1} + \varepsilon_t \\ y_t = H_t x_t + \omega_n \end{cases} \quad (3.1)$$

em que x_t é o vetor de estado do sistema no momento t , y_t é o vetor de medição correspondente no tempo t , F_t é um parâmetro de transição de estado, ε_t é o termo de ruído

do estado aleatório, H_t é um parâmetro de medição, ω_n é o termo de erro de medição. Na primeira etapa, o estado do sistema e variância associada é estimado em relação à primeira expressão na Equação (3.1). Então, no segundo estágio, o estado estimado é atualizado com base em ambas as expressões em Eq (3.1) (DEMIRHAN e RENWICK, 2018). Nesse artigo utilizou-se a implementação usando as opções: ARIMA e modelos estruturais, como modelos de transição.

3.5.2. Persistência

A abordagem de persistência é considerada de ambas as direções para frente e para trás neste estudo. A principal desvantagem dessa abordagem é que não é possível imputar valores faltantes se a observação que será levada para frente ou para trás estiver faltando. Mas, neste caso, os valores faltantes restantes são imputados. Essa abordagem é semelhante àquela dos modelos ingênuos. Consideramos as opções de remoção e substituição com média geral para o restante dos valores faltantes na abordagem de persistência (DEMIRHAN e RENWICK, 2018).

3.5.3. Média Móvel Ponderada

Nesta abordagem, cada um dos valores faltantes é substituído pela média de k observações, o que é chamado de tamanho da janela, de ambos os lados do valor faltante. Seja Y_t a série temporal de interesse. Então, a previsão média ponderada de um passo à frente é definida como na equação a seguir:

$$\hat{Y}_{t+1} = \sum_{i=-k}^k \theta_i Y_{t+1-i}$$

em que $\theta_{-k}, \dots, \theta_k$ são os pesos. Quando a janela completa não estiver disponível devido aos valores faltantes, k será aumentado de forma incremental. Consideramos pesos diferentes para calcular a média: simples e linear. Os pesos lineares são aqueles que diminuem de tal forma que $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots, \frac{1}{K}$, em que o denominador aumenta linearmente (DEMIRHAN e RENWICK, 2018).

3.5.4. Imputação Clássica

Essa é a abordagem mais básica para a imputação de valores faltantes. Todos os valores faltantes são substituídos por média, moda ou mediana dos valores observados. Para substituição com a moda, os valores mais frequentes são usados. Se a frequência de mais de um valor for a mesma, o menor será usado para substituição. Essa abordagem não conta nem com correlação serial na série nem com a sazonalidade (DEMIRHAN e RENWICK, 2018).

3.5.5. Amostra Aleatória

Nesta abordagem, os valores faltantes são substituídos por um conjunto de números aleatórios uniformes gerados dentro do intervalo das séries observadas. A proporção de dados faltantes tem um papel vital nessa abordagem, já que a precisão dessa abordagem diminui com o aumento do número de valores faltantes (DEMIRHAN e RENWICK, 2018).

3.5.6. Decomposição Sazonal

Uma série temporal Y_t , com $t = 1, K, T$, é decomposta em seus componentes como na equação a seguir:

$$Y_t = T_t + S_t + \varepsilon_t$$

em que T_t é o componente da tendência, S_t é o componente sazonal e ε_t é o erro aleatório. A abordagem de decomposição sazonal considera a sazonalidade na série. Primeiro, a sazonalidade é removida da série e a imputação é feita com base no componente de correlação serial restante usando as abordagens descritas em Seções 3.3.2.1 -3.3.2.6. Finalmente, o componente sazonal é adicionado de volta à série. Se não houver sazonalidade na série, essa abordagem converge para um dos métodos de imputação discutidos. Mas na existência de sazonalidade, espera-se que seja melhor por formulários do que por aqueles que não consideram a sazonalidade. Para correr filtro de Kalman sob esta abordagem, usamos o modelo estrutural com suavização. O tamanho da janela é definido como quatro para a implementação do MA. Um modelo estrutural é um caso especial do modelo de espaço de estados dado em Equação (3.1) (DEMIRHAN e RENWICK, 2018). O modelo estrutural básico considerado neste estudo é um modelo com tendência local e um componente sazonal adicional, que é representado como (HYNDMAN, 2016):

$$l_t = l_{t-1} + b_{t-1} + \alpha \varepsilon_t$$

$$b_t = b_{t-1} + \beta \varepsilon_t$$

$$s_t = s_{t-m} + \gamma \varepsilon_t$$

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$$

em que l_t é o termo de nível, b_t é o termo de crescimento, e s_t é o componente sazonal.

3.5.7. MICE

A imputação múltipla por equações encadeadas (MICE) é uma abordagem prática para resolver o problema de um banco de dados com valores faltantes em várias variáveis, que gera imputações com base em um conjunto de modelos de imputação, um modelo para cada variável com observações faltantes (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011). O MICE também é conhecido como especificação totalmente condicional (BUUREN, 2007) e regressão sequencial multivariada (RAGHUNATHAN et al., 2001).

O MICE é um algoritmo que se inicia com um sorteio aleatório dos dados observados e imputa as observações faltantes, variável por variável. Uma característica importante do MICE é sua capacidade de lidar com diferentes tipos de variáveis (contínua, binária, categórica) porque cada variável é imputada usando seu próprio modelo de imputação (VAN BUUREN, STEF AND OUDSHOORN, 1999).

Inicialmente, todos os dados faltantes são preenchidos por amostragem aleatória simples com reposição dos valores observados. A primeira variável com valores faltantes, é regredida com base nas outras variáveis. Os valores perdidos da primeira variável são substituídos por sorteios da distribuição preditiva *posteriori* correspondente da primeira variável. Então, a próxima variável com valores faltantes, é regredida em todas as outras variáveis, e utilizando os valores imputados para a primeira variável. Novamente, a segunda variável é substituída por sorteios da distribuição preditiva *posteriori* da segunda variável. O processo é repetido para todas as variáveis com observações faltantes: isso é chamado de ciclo (VAN BUUREN e GROOTHUIS-OUDSHOORN, 2011).

Para estabilizar os resultados, o procedimento é geralmente repetido por vários ciclos (iterações), para produzir um único conjunto de dados imputado, e todo o procedimento é repetido n vezes para fornecer conjuntos de dados imputados. O número de iterações geralmente pode ser baixo, sendo de 5 a 10 (WHITE et al., 2011).

A descrição formal do algoritmo MICE é descrita a seguir.

3.5.7.1. Algoritmo

As observações faltantes são substituídas por m conjuntos de valores simulados independentes, extraídos da distribuição preditiva *posteriori* das observações faltantes condicionado aos dados observados. Assume-se que as observações em Y e $\mathbf{X}$ correspondem a uma amostra aleatória simples de uma população de interesse com distribuição multivariada $P(\mathbf{X}, Y | \boldsymbol{\theta})$, assumindo que a distribuição multivariada de Y e $\mathbf{X}$ é completamente especificada, por um vetor de parâmetros desconhecidos, $\boldsymbol{\theta}$. O modelo será discutido na seção 3.5.8.2. O problema é como obter a distribuição multivariada de $\boldsymbol{\theta}$, explícita ou implicitamente. Os passos seguintes são repetidos m vezes:

3. Obtém a distribuição *posteriori* de $\boldsymbol{\theta}$ por amostragem iterativa a partir de distribuições condicionais da forma $P(Y | \mathbf{X}, \boldsymbol{\theta})$.
4. Partindo de um simples sorteio das distribuições marginais observadas, a iteração t -ésima de equações encadeadas é um amostrador Gibbs que sucessivamente chama

$$\boldsymbol{\theta}^{*(t)} : P(\boldsymbol{\theta} | \mathbf{X}, Y^{obs})$$

$$Y^{*(t)} : P(Y | \mathbf{X}, \boldsymbol{\theta}_1^{*(t)})$$

em que $Y^{(t)} = (Y^{obs}, Y^{*(t)})$ é a variável imputada na iteração t . Observa-se que para a imputação de $Y^{*(t-1)}$ somente $Y^{*(t)}$ é considerada, devido ao relacionamento com as variáveis. O $\boldsymbol{\theta}^*$ é a estimativa do parâmetro para a t -ésima iteração das equações encadeadas (VAN BUUREN, GROOTHUIS-OUUDSHOORN, 2011).

Esse processo é conhecido como imputação adequada porque incorpora todas as fontes de variabilidade e incerteza nos valores imputados, incluindo erros de previsão dos valores individuais e erros de estimação nos coeficientes ajustados do modelo de imputação (RUBIN, 1987). O método para criar as m imputações é descrito na seção 3.5.8.3.

3.5.7.2. Modelo

Assume-se que Y seja uma variável que tem valores faltantes, na qual deseja-se imputar a partir de outras (co)variáveis completas. Seja X_k , com $k = 1, K, l$, sendo l

covariáveis completas, em que $\mathbf{X} = (X_1, K, X_k)$. Define-se por Y^{obs} são as observações observadas, Y^{mis} são as observações faltantes e n^{obs} para o número de observações com valores observados de Y .

Um modelo de regressão linear é a escolha mais comum do modelo para imputação de variáveis contínuas normalmente distribuídas:

$$Y | \mathbf{X}, \boldsymbol{\beta}: N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$$

Seja $\hat{\boldsymbol{\beta}}$ uma estimativa do parâmetro (um vetor com k linha) da adaptação deste modelo das observações de Y . Seja $\mathbf{V}$ uma matriz de covariância estimada de $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$ é a estimativa da raiz do erro médio quadrático. Em seguida, apresenta os parâmetros de imputação σ^* e $\boldsymbol{\beta}^*$ a partir da distribuição *posteriori* de σ , $\boldsymbol{\beta}$.

Para σ^* , é calculado como:

$$\sigma^* = \hat{\sigma} \sqrt{(n^{obs} - k) / g}$$

em que g é uma variável χ^2 com $n_j^{obs} - k$ graus de liberdade. Para $\boldsymbol{\beta}^*$, é calculado como:

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}} \mathbf{u}_1 \mathbf{V}^{1/2}$$

em que $\mathbf{u}_1$ é um vetor de k linha independentes e aleatórias de uma distribuição Normal; $\mathbf{V}^{1/2}$ é uma decomposição de Cholesky de $\mathbf{V}$.

Os valores imputados Y^* para cada observação faltante Y são então obtidos como:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta}^* + u_{2i} \sigma^*$$

em que $u_{2,i}$ é um valor aleatório da distribuição Normal (WHITE, ROYSTON e WOOD, 2011).

3.5.7.3. Método da Média Preditiva

A ideia básica do método de média preditiva de correspondência (PMM - *Predictive Mean Matching*) é usar diferentes modelos de regressão para calcular a observação faltante. Cada conjunto de estimativas dos modelos será usado para gerar um único conjunto de dados imputados (HEITJAN e LITTLE, 1991). As estimativas dos parâmetros devem ser extraídas de sua distribuição *posteriori* dos dados, que está descrito na Seção 2.3.2.

A imputação múltipla dos Y^{mis} por meio de PMM é realizado pelo seguinte algoritmo (LITTLE, 1988; VINK et al., 2014):

6. Utiliza-se o modelo descrito na Seção 2.2.3 para calcular um vetor de parâmetros perturbados β^* .
7. Para cada observação faltante de Y com covariáveis $\mathbf{X}$, o procedimento padrão seria a próxima amostra de uma distribuição normal com $\mathbf{X}\beta^*$.
8. Identifique os q menores valores de $\Delta = \left| \mathbf{X}_h \hat{\beta} - \mathbf{X}_i \beta^* \right|$, $h = 1, K, n^{obs}$.
9. Escolhe-se um valor aleatório de (Δ_1, K, Δ_q) e passa ser valor imputado.
10. Repita os passos 1-4 m vezes, cada vez é salvo o conjunto de dados concluído.

O método PMM executa a imputação múltipla de acordo com a descrição desse algoritmo. Assim, o valor imputado é um valor observado em Y , cuja previsão com os dados observados corresponde de perto à predição perturbada. O método pode ser visto como uma generalização do *bootstrap* bayesiano aproximado, como descrito em Rubin e Schenker (1986). Nesse trabalho, foi utilizado $q = 3$, que teve um bom desempenho em um estudo de simulação (SCHENKER e TAYLOR, 1996).

3.6. Avaliação do Processo de Imputação

Uma vez que o modelo de imputação tenha sido especificado e as imputações iniciais criadas, é importante verificar o modelo e refinar conforme necessário. Isso pode ser feito de várias maneiras. Várias medidas de erro médio de desempenho de modelos são utilizadas na literatura, cada uma dessas medidas é "dimensionada" na medida em que expressa um erro médio de previsão de modelo nas unidades da variável de interesse. De acordo com Willmott e Matsuura (2005), a raiz do erro quadrático médio (RMSE) é uma das medidas de erro mais usadas na literatura climática e ambiental.

Considere y_i , $i = 1, K, n$, sendo o i -ésimo dado faltante de uma série temporal, e $y_{imp,i}$ é o i -ésimo valor imputado para o i -ésimo dado faltante. A medida RMSE é definida como nas (CHEN et al., 2017; HYNDMAN, 2016)

$$RMSE(y, y_{imp}) = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{imp,i})^2}{n}}$$

3.7. Descrição dos Dados

Para a imputação de dados faltantes, utilizou-se sete séries de precipitação diária, expressas em milímetros (mm), provenientes de postos pluviométricos do estado do Paraná, apresentadas na Tabela 3.1. Os dados são públicos disponibilizados pelo Instituto Águas do Paraná pertencente ao Instituto Agrônômico do Paraná (IAPAR) e compilados pelo Grupo de Estudos em Seguros e Riscos (GESER), da Escola Superior “Luiz de Queiroz” (ESALQ/USP).

Uma característica desses postos é o fato de possuírem observações faltantes na série. E para esse trabalho, optou-se por trabalhar com séries completas, a partir do ano de 1979. Dessa maneira, foi possível selecionar um posto com observações faltantes, somente, nos dias 29 de fevereiro. Assim sendo, esses dias foram retirados da série, tornando a série com os demais dias completos.

O posto pluviométrico selecionado foi de Porto Eucalipto, no município de São Pedro do Paraná, no estado do Paraná. Esse posto foi desativado em 22 de abril de 1998. Portanto, a série teve como data inicial o dia 01 de janeiro de 1979 e a data final 31 de dezembro de 1997, com o número de observações e informações de localização apresentados na Tabela 3.1.

Tabela 3.1. Posto pluviométrico utilizado neste trabalho.

Estação	Observações	Latitude	Longitude
Porto Eucalipto	6935	-22,75	-53,13

Para se obter uma variável que auxilie a imputações dos valores faltantes das séries de precipitação, utilizou-se dados de precipitação diária em ponto de grade obtidos a partir de grades igualmente espaçadas com resolução de $0,50^\circ$ latitude $\times$ $0,50^\circ$ longitude, provenientes das séries históricas *KNMI Climate Explorer* (EXPLORER, 2019). As observações se referem a *Climate Prediction Center* (CPC) (CENTER, 2019) de precipitação diária, expressas em milímetros (mm), que compreendem o período inicial de janeiro de 1979, justificando o período inicial das séries utilizadas nesse trabalho. Este produto é proveniente de observações de pluviômetros, e combina dados de mais de 30.000 estações ao redor do globo utilizando a técnica de Interpolação Ótima (GANDIN, 1965).

A padronização dos períodos entre os dados observados e do CPC são fundamentais para fins comparativos na análise proposta, por isso, a série de precipitação diária tem início

em 01 de janeiro de 1979. Justificando o período inicial utilizado neste trabalho, como também o uso do ponto de grade.

No KNMI Climate Explorer (EXPLORER, 2019), foi selecionado o campo diário e escolheu-se observações de precipitação com pontos de grade mais próximos de $0,50^\circ \times 0,50^\circ$, marcando os intervalos de latitude e longitude para cada posto. Foram feitos os *downloads* de diversas séries com diferentes combinações de intervalos e calculou-se a correlação com a série do posto pluviométrico, a série que teve maior correlação foi escolhida como a covariável da precipitação em ponto de grade, obtendo uma correlação alta de 0,60. A grade selecionada foi de latitude de -22,50 a -23 e longitude de -53 a -53,50.

Como a série de precipitação diária está completa, dados faltantes são criados e, em seguida, os valores faltantes artificialmente são imputados usando os métodos de imputação de interesse, esses métodos são apresentados na Tabela 3.2. Para os métodos K1 a K5, utilizou-se a série de precipitação diária como banco de dados. Para os métodos M1 a M4, montou-se o banco de dados com a variável precipitação diária e as variáveis: mês, ano e precipitação diária em ponto de grade. Consequentemente, a entrada dos dados é diferente, mas a análise é baseada na média da RMSE, como descrito na Figura 3.1.

A precisão dos métodos de imputação para a dados faltantes de precipitação diária é comparada para determinar o método mais preciso para a imputação de dados faltantes de precipitação diária. Finalmente, as medidas de erro são calculadas para comparação. Detalhes da implementação são explicados a seguir e um fluxograma é apresentado na Figura 3.1.

Criou-se dados faltantes de maneira aleatória e em trechos. De maneira aleatória, fez-se 10%, 20% e 30% de dados faltantes. Em trechos, fez-se 500, 1000, 1500 e 2000 de dados faltantes, com intervalo de 10 a 60, espaçados em 10 em 10. Calculou-se a RMSE para diferentes métodos que são apresentados na Tabela 3.2.

Figura 3.1. Fluxograma do algoritmo utilizado para comparar a precisão dos métodos de imputação de valores faltantes para a precipitação diária.

Para o algoritmo aleatório, as seguintes etapas são executadas:

1. Cria, aleatoriamente, 1200 amostras com 10%, 20% e 30% de dados faltantes;
2. Aplica os métodos de imputação;
3. Calcula a RMSE.

Para o algoritmo em trechos, as seguintes etapas são executadas:

1. Cria, trechos de 500, 1000, 1500 e 2000 dados faltantes, espaçados de 10, 20, 30, 40, 50, 60 em 60;
2. Aplica o algoritmo de imputação;
3. Calcula a RMSE.

Para analisar a eficiência dos métodos K1 a M4, a medida da RMSE foi escolhida como uma medida de desempenho.

Na prática, não há informações específicas sobre a quantidade típica de dados perdidos em séries de precipitação diária. A precisão dos métodos de imputação de valores faltantes é altamente relacionada à informação disponível nas partes completas (observadas) da série. Especialmente, o comprimento das lacunas na série é altamente influente na precisão da imputação. Em grandes lacunas, a precisão da imputação diminui até o final do intervalo.

Portanto, criamos pequenas, médias e grandes lacunas (trechos) usando uma inadimplência 500, 1000, 1500 e 2000 dados faltantes.

Tabela 3.2 Métodos de imputação utilizados neste trabalho.

Representação	Método	Opção
K1	Filtro de Kalman	ARIMA
K2	Filtro de Kalman	Modelo Estrutural
K3	Média Móvel Ponderada	Linear
K4	Média Móvel Ponderada	Simples
K5	Decomposição Sazonal	-
M1	MICE	Modelo com a variável mês
M2	MICE	Modelo com as variáveis mês e ano
M3	MICE	Modelo com a variável precipitação diária em ponto de grade
M4	MICE	Modelo com as variáveis mês, ano e precipitação diária em ponto de grade

3.8. Resultados e Discussão

Os resultados foram comparações entre os métodos de imputação em uma abordagem univariada e múltipla. A série utilizada foi a de precipitação diária, obtidas de uma estação meteorológica do estado do Paraná, apresentada na Figura 3.2. Especificamente, o conjunto de dados é caracterizado por ter variação periódica e padrões sazonais que são comuns em processos variantes no tempo. Para demonstrar a precisão da imputação, gerou-se dados faltantes aleatórios e em trechos, comparando-se os métodos usando a média da raiz do erro quadrático médio (RMSE).

Figura 3.2. Série temporal de precipitação diária da estação de Porto Eucalipto.

Dados faltantes aleatórios foram gerados para quantidades de 10%, 20% e 30% de dados faltantes no conjunto de dados e calculou-se a RMSE para cada método e apresenta-se na Figura 3.3. Os valores imputados usando os métodos MICE com modelos com a variável mês (M1) e com o modelo com as variáveis mês e ano (M2) tiveram os maiores valores da RMSE, enquanto que MICE com o modelo de precipitação diária em ponto de grade (M3) e matriz preditora mês, ano e precipitação diária em ponto de grade (M4) tiveram melhores resultados. Para os métodos de imputação na abordagem univariada, os algoritmos de imputação por Filtro de Kalman (K1 e K2) superaram todos os métodos com os menores valores da RMSE.

Figura 3.3. Média das RMSE obtidos com diferentes modelos, fazendo 10%, 20% e 30% de observações faltantes.

Para os dados faltantes em trechos, gerou-se para quantidades de 500, 1000 e 1500 de dados faltantes no conjunto de dados, para espaços de 10, 20, 30, 40, 50 e 60 e calculou-se a RMSE para cada método (Figura 3.4). Os valores imputados usando os métodos MICE tiveram os maiores valores da RMSE, enquanto os algoritmos de Filtro de Kalman (K1 e K2) superaram todos os métodos.

Figura 3.4. Média das RMSE obtidos com diferentes modelos, para 500 dados faltantes.

Figura 3.5. Média das RMSE obtidos com diferentes modelos, para 1000 dados faltantes.

Figura 3.6. Média das RMSE obtidos com diferentes modelos, para 1500 dados faltantes.

Figura 3.7. Média das RMSE obtidos com diferentes modelos, para 2000 dados faltantes.

Os resultados nas Figuras 3.3, 3.4 e 3.7, para os métodos K1 e K2, fornecem valores estáveis da RMSE, sendo menores que 12. Na Figura 3.5 e na Figura 3.6 para o método K1, os resultados fornecem valores estáveis da RMSE, enquanto para o método K2 tem uma variação dos valores da RMSE, mas em ambos, os métodos os valores são menores que 12.

Para a série de precipitação diária, o Filtro de Kalman foi capaz de fornecer os menores valores da RMSE comparados com a imputação MICE, com diferentes cenários de dados faltantes. De acordo com Moritz e Bartz-Beielstein (2017), de todos os métodos de imputação disponíveis, não existe um método global como sendo o melhor. O desempenho da imputação depende das características da série temporal de estudo e geralmente, o Filtro de Kalman é considerado uma boa abordagem para a imputação de dados univariados, como observado neste trabalho.

No artigo de Bokde et al. (2018), inicialmente, foi feita uma validação cruzada utilizando um método de imputação de 10% a 50%, com intervalos de 10%, de dados faltantes com diferentes tamanhos trechos, sendo que cada análise possuía dois trechos, um no começo e o outro fim da série, para selecionar o tamanho ideal do trecho do início e do trecho do final da série com base na menor RMSE e após isso fez-se comparações de métodos de imputações variando a quantidade de dados faltantes de 10 a 60% em intervalos crescentes de 10% nos trechos selecionado no primeiro instante. Nesse trabalho optou-se por fazer dados faltantes aleatórios e em trechos. Para os dados faltantes aleatórios, gerou-se 10%, 20% e 30% de dados faltantes. Para os dados faltantes em trechos, gerou-se intervalos de 500, 1000, 1500 e

2000 de dados faltantes com diferentes espaços. Enquanto que nesse trabalho obteve valores estáveis da RMSE para diferentes cenários de trechos e também diferentes valores de dados faltantes.

3.9. Conclusão

A aplicação do Filtro de Kalman produziu melhores estimativas para os valores diários de precipitação. Esse método provou ser uma técnica versátil e adaptável a diferentes estações do ano e deve ser considerado como uma alternativa para preencher lacunas em séries temporais de precipitação.

O presente estudo contribui para o avanço da literatura através da aplicação da imputação de séries temporais para dados de precipitação diária e a para contribuição da imputação dos valores faltantes de séries de fenômenos climáticos. Esse fato tem repercussão benéfica no estudo de preenchimento de falhas de precipitação diário, tendo em vista que há poucos estudos na área, devido à dificuldade de obter séries longas de observações diárias.

O avanço da pesquisa na área de imputação de dados e o conjunto de medidas verificadas poderá tornar o Filtro de Kalman uma importante metodologia para imputação de dados de precipitação diária, garantindo uma série temporal completa para análises de vários setores, dentre eles a agricultura.

A previsão de ocorrência de fenômenos meteorológicos é extremamente relevante para agricultura, pois é uma informação de suma importância para o seguro rural, financiamento e planejamento das atividades agrícolas. Assim sendo, para séries temporais com dados faltantes, é de grande importância a imputação dos dados, utilizando abordagens estatísticas precisas.

Referências

- BARNARD, J.; MENG, X.-L. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, v. 8, n. 1, p. 17–36, 2 fev. 1999.
- BOKDE, N. et al. A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern Recognition Letters*, v. 116, p. 88–96, 2018.
- BROOKS, M. J.; VON BACKSTRÖM, T. W.; VAN DYK, E. E. Performance characteristics of a perforated shadow band in the presence of cloud. *Solar Energy*, v. 139, p. 533–546, 1 dez. 2016.

BUUREN, S. VAN. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, v. 16, n. 3, p. 219–42, 2007.

CARDOSO, C. S.; QUADRO, M. F. L. Análise comparativa de dados de precipitação gerados pelo “Climate Prediction Center – CPC” versus dados observados para o Sul do Brasil Camila. *Revista Brasileira de Geografia Física*, v. 10, n. 4, n. ISSN:1984-2295, p. 1180–1198, 2017.

CARVALHO, J. R. P. DE et al. Model for Multiple Imputation to Estimate Daily Rainfall Data and Filling of Faults Modelo de Imputação Múltipla para Estimar Dados de Precipitação Diária e Preenchimento de Falhas. v. 32, n.4, p. 575–583, 2017.

CENTER, N. C. P. NOAA’s Climate Prediction Center. Disponível em: <<https://www.cpc.ncep.noaa.gov/>>. Acesso em: 8 jun. 2019.

CHEN, C.; TWYXCROSS, J.; GARIBALDI, J. M. A new accuracy measure based on bounded relative error for time series forecasting. *PLOS ONE*, v. 12, n. 3, p. e0174202, 24 mar. 2017.

COLLINS, L. et al. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, v. 6, n.4, p. 330–351, 2001.

DALY, C. et al. A probabilistic-spatial approach to the quality control of climate observations. Disponível em: <https://www.researchgate.net/publication/252891043_A_probabilistic-spatial_approach_to_the_quality_control_of_climate_observations>. Acesso em: 1 abr. 2019.

DEMIRHAN, H.; RENWICK, Z. Missing value imputation for short to mid-term horizontal solar irradiance data. *Applied Energy*, v. 225, n. May, p. 998–1012, 2018.

EXPLORER, K. C. Climate Explorer: Starting point. Disponível em: <<https://climexp.knmi.nl/start.cgi>>. Acesso em: 8 jun. 2019.

FENG, S.; HU, Q. I.; QIAN, W. QUALITY CONTROL OF DAILY METEOROLOGICAL DATA IN CHINA, 1951-2000: A NEW DATASET. *INTERNATIONAL JOURNAL OF CLIMATOLOGY Int. J. Climatol*, v. 24, p. 853–870, 2004.

FRASER, G.; YAN, R. Guided multiple imputation of missing data: Using a subsample to strengthen the missing-at-random assumption. *Epidemiology*, v. 18, n. 2, p. 246–252, 2007.

FUNG, D. Methods for the estimation of missing values in time series. 2006.

GANDIN, L. S. Objective analysis of meteorological fields. Jerusalem: Israel program for scientific translations, 1965.

GNAUCK, A. Interpolation and approximation of water quality time series and process identification. *Analytical and Bioanalytical Chemistry*, v. 380, n. 3, p. 484–492, 9 out. 2004.

HEITJAN, L F. , LITTLE, R. J. A. Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society*, v. 40, n.1, p. 13–29, 1991.

HYNDMAN, R. J. forecast: Forecasting functions for time series and linear models. 2016.

LITTLE, R. J. A. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, v. 6, n. 3, p. 287–296, jul. 1988.

LITTLE, R.; RUBIN, D. Statistical analysis with missing data. [s.l: s.n.].

MELLO, Y. R. DE; KOHLS, W.; OLIVEIRA, T. M. N. DE. Uso De Diferentes Métodos Para O Preenchimento De Falhas Em Estações Pluviométricas. *Boletim de Geografia*, v. 35, n. 1, p. 112, 2017.

MORITZ, S. et al. Comparison of different Methods for Univariate Time Series Imputation in R. 13 out. 2015.

MORITZ, S.; BARTZ-BEIELSTEIN, T. imputeTS: Time Series Missing Value Imputation version in T. *The R Journal*, v. XX, p. 1–12, 2017.

NASCIMENTO, T. S. DO et al. Preenchimento De Falhas Em Banco De Dados Pluviométricos Com Base Em Dados Do Cpc (Climate Prediction Center): Estudo De Caso Do Rio Solimões-Amazonas. *Revista Brasileira de Climatologia*, v. 7, n. January, 2016.

NIEBUHR, T. Irregularly observed time series – some asymptotics and the block bootstrap. *Statistics*, v. 51, n. 5, p. 1118–1131, 3 set. 2017.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. *Revista Brasileira de Epidemiologia*, v. 13, n. 4, p. 596–606, 2010.

OLIVEIRA, L. F. C. DE et al. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 14, n. 11, p. 1186–1192, 2010.

OZAKI, V. A. Taxa De Prêmio De Contratos De Seguro Agrícola. p. 347, 2005.

OZAKI, V. A. Em busca de um novo paradigma para o seguro rural no Brasil. *Revista de Economia e Sociologia Rural*, v. 46, n. 1, p. 97–119, mar. 2008.

PEDRO, H. T. C.; COIMBRA, C. F. M. Short-term irradiance forecastability for various solar micro-climates. *Solar Energy*, v. 122, p. 587–602, 1 dez. 2015.

RAGHUNATHAN, T. E. et al. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models Key Words: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations. *Survey Methodology*, v. 27, n. 1, 2001.

RUBIN, D. Inference and missing data. *academic.oup.com*, v. 63, n. 3, p. 581–592, 1976.

RUBIN, D. B. Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, v. 4, n. 1, p. 87–94, jan. 1986.

RUBIN, D. B. Multiple imputation for nonresponse in surveys. [s.l: s.n.].

RUBIN, D. B. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, v. 91, n. 434, p. 473–489, jun. 1996.

RUBIN, D. B.; SCHENKER, N. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, v. 81, n. 394, p. 366–374, jun. 1986.

SANTOS, A. T. S. DOS; SILVA, C. M. S. E. Seasonality, Interannual Variability, and Linear Tendency of Wind Speeds in the Northeast Brazil from 1986 to 2011. *The Scientific World Journal*, v. 2013, p. 1–10, 2013.

SCHAFER, J. L. ; JOSEPH L. Analysis of incomplete multivariate data. [s.l.] Chapman & Hall, 1997.

SCHENKER, N.; TAYLOR, J. M. G. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, v. 22, n. 4, p. 425–446, 1996.

SHAKYA, A. et al. Solar Irradiance Forecasting in Remote Microgrids Using Markov Switching Model. *IEEE Transactions on Sustainable Energy*, v. 8, n. 3, p. 895–905, jul. 2017.

TURRADO, C. et al. Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions. *Sensors*, v. 14, n. 11, p. 20382–20399, 29 out. 2014.

VAN BUUREN, STEF AND OUDSHOORN, K. Flexible multivariate imputation by MICE. [s.l.] Leiden: TNO, 1999.

VAN BUUREN, S.; GROOTHUIS-OUUDSHOORN, K. Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, v. 45, n. 3, p. 1–67, 2011.

VILLIERS, J. DE. Mathematics of Approximation. Paris: Atlantis Press, 2012.

VINK, G. et al. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, v. 68, n. 1, p. 61–90, 2014.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, v. 30, n. 4, p. 377–399, 2011.

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, p. 79–82, 2005.