Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Comparação de métodos de imputação para dados de pecuária de precisão

Vivian Aparecida Brancaglioni

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

Piracicaba
2023

Vivian Aparecida Brancaglioni
Licenciada em Matemática

Comparação de métodos de imputação para dados de pecuária de precisão

versão revisada de acordo com a Resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

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

Piracicaba
2023

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

Brancaglioni, Vivian Aparecida
Comparação de métodos de imputação para dados de pecuária de precisão / Vivian Aparecida Brancaglioni. – – versão revisada de acordo com a Resolução CoPGr 6018 de 2011. - - Piracicaba, 2023 .
56 p.

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

1. Métodos de imputação 2. Dados longitudinais 3. Estudo de simulação 4. MICE I. Título.

À minha família.

AGRADECIMENTOS

Agradeço a Deus pelas providências divinas para a conclusão desse trabalho.

Aos meus pais, Antônio e Maria Luiza, e minha irmã Caren, que me deram suporte e apoio nas decisões que precisei tomar ao longo desse árduo caminho. Foram minha força, meu suporte, me deram colo e sempre me lembraram que se orgulham de mim, me ajudando a acreditar que o impossível era possível.

Ao meu noivo Felipe que compartilhou de todas as minhas angústias e me pegou pela mão, com sua calma e generosidade, nunca me deixou desistir ou duvidou de minha capacidade e força. Tenho muita admiração por você.

Ao meu orientador, Professor Carlos Tadeu, que confiou e acreditou em mim até o último instante.

Aos amigos que fiz ao longo do caminho que me auxiliaram nessa etapa tão intensa e desafiadora, seja nas atribuições do trabalho ou em suporte psicológico. Em especial ao meu amigo Welinton que sempre se prontificou a me ajudar.

À @Tech por me ceder os dados para a realização deste trabalho.

Aos professores e aos funcionários do Departamento de Ciências Exatas da Esalq.

Aos membros da banca de qualificação e defesa que gentilmente contribuíram para a melhoria do trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ).

SUMÁRIO

RESUMO

Comparação de métodos de imputação para dados de pecuária de precisão

Durante a condução de um experimento ou pesquisa é comum existir perda de informação, seja por preenchimento incorreto do banco de dados ou por falta de informação para algumas observações de determinada variável. Isso ocorre por motivos que muitas vezes não se sabe definir, dessa forma, o valor que deveria ter sido coletado se configura como valor ausente, tornando o conjunto de dados obtido incompleto. Estudos com a presença de observações ausentes são muito comuns em grande parte das áreas do conhecimento, e com dados obtidos a partir da pecuária da precisão não seria diferente. Dados de pecuária de precisão auxiliam o setor agropecuário a acompanhar, mapear e identificar problemas e buscar soluções. O conjunto de dados utilizado neste trabalho provém da pecuária de precisão, no qual pode-se acompanhar a oscilação de peso de 38 animais, das raças Nelore e Cruzado Britânico, divididos entre macho inteiro e macho castrado. Esses dados foram coletados a partir de um sistema de plataforma de pesagem automática. No entanto, durante as pesagens algumas informações de peso foram perdidas e o objetivo deste trabalho foi comparar o desempenho de quatro métodos de imputação de dados da classe MICE, implementados no *software* R por meio do pacote *mice*: método de média preditiva (PMM), método baseado na regressão linear bayesiana (BLR), árvore de classificação e regressão (CART) e floresta aleatória (RF). Esses métodos foram comparados por meio de quatro critérios, raiz do erro quadrático médio (RMSE), pelo coeficiente de correlação de Pearson, índice de acurácia de Willmott e índice de desempenho. A análise foi conduzida da seguinte forma: primeiro foram removidas as observações com valor de peso faltante do conjunto de dados original, obtendo-se um conjunto completo; e a partir dele foram criados novos bancos com diferentes porcentagens de dados faltantes, 5%, 10% e 15%, removidos aleatoriamente. A partir desses novos cenários obtidos, cada um dos métodos foram aplicados, sendo consideradas 5 e 10 iterações. Pôde-se observar que não houve diferença para as imputações em todos os métodos e cenários com relação a quantidade de iterações. Fixando-se os métodos e comparando as diferentes proporções de dados faltantes, observou-se uma diminuição da variabilidade das medidas que envolvem os critérios de comparação para os diferentes métodos, exceto para o método de floresta aleatória, para maior quantidade de ausências. Quando comparados os métodos, fixando-se os cenários, foi possível observar que o método de árvore de classificação e regressão teve melhor desempenho e o método de floresta aleatória se destacou de forma negativa. Ao aplicar os métodos no conjunto de dados originais, foi observado resultado semelhante, sendo o método CART o mais adequado para substituir os valores faltantes.

Palavras-chave: Métodos de imputação, Dados longitudinais, Imputação múltipla, MICE

ABSTRACT

Comparison of imputation methods for precision livestock data

During experiments or research it is common for information to be lost, either by incorrectly filling out the database or by lack of information for some observations of a particular variable. This occurs for reasons that often cannot be defined so that the value that should have been collected is configured as a missing value, making the data set obtained incomplete. Studies with missing observations are very common in most areas of knowledge, and with data obtained from precision farming, it would be no different. Precision livestock data helps the agricultural sector to track, map, and identify problems and seek solutions. The data set used in this work comes from precision cattle breeding, where it is possible to follow the oscillation of weight of 38 animals, of the Nelore and "Cruzado Britânico"breeds, divided into full male and castrated males. These data were collected from an automatic weighing platform system. However, during the weightings, some weight information was lost and the objective of this work was to compare the performance of four MICE class data imputation methods, implemented in *software* R by means of the *mice* package: predictive mean method (PMM), Bayesian linear regression (BLR) based method, classification and regression tree (CART) and random forest (RF). These methods were compared using four criteria, root mean square error (RMSE), Pearson's correlation coefficient, Willmott's accuracy index, and performance index. The analysis was conducted as follows: first, observations with a missing weight value were removed from the original data set, obtaining a complete set; and from it, new databases were created with different percentages of missing data, 5%, 10%, and 15%. From these new obtained scenarios each of the methods was applied, with 5 and 10 iterations being considered. It could be observed that there was no difference in the imputations in all methods and scenarios regarding the number of iterations. By fixing the methods and comparing the different proportions of missing data, a decrease in the variability of the measures involving the comparison criteria was observed for the different methods, except for the random forest method, for a larger amount of missing data. When comparing the methods, and setting the scenarios, it was possible to observe that the classification and regression tree method performed better, and the random forest method stood out in a negative way. When applying the methods to the original data set, a similar result was observed, with the CART method being the most suitable to replace the missing values.

Keywords: Imputation methods, Longitudinal data, Multiple imputation, MICE

1 INTRODUÇÃO

O avanço tecnológico das últimas décadas possibilitou o desenvolvimento na indústria de tecnologia agropecuária, tornando-se possível a coleta de uma grande quantidade de dados em fazendas do mundo todo (Webber *et al.*, 2019). Diversas ferramentas foram desenvolvidas e destinadas a diferentes finalidades dentro da agropecuária de precisão. A pecuária de precisão fornece meios para uma melhoria do estado de saúde e bem-estar do animal, redução de custos veterinários, melhoria no gerenciamento de resíduos agrícolas, auxiliando à redução de impactos ambientais; e consequentemente contribuindo para uma sustentabilidade ambiental e econômica (Simitzis *et al.*, 2021; Murphy *et al.*, 2021; Tzanidakis *et al.*, 2023). Dentre as diversas finalidades, temos as que visam uma maior acurácia no acompanhamento dos rebanhos, manejos e tratos com o solo.

Laca (2009) apresenta diversas ferramentas utilizadas pela pecuária de precisão, utilizando de diversas informações específicas sobre os animais e a heterogeneidade de ambientes em que estão inseridos, visando otimizar a produção. Dentre as ferramentas apresentadas, os métodos remotos para pesagem do gado têm grande importância, uma vez que o processo tradicional consistiria em retirar o animal da área de pastejo ou confinamento, levá-lo até um brete para que um técnico faça a pesagem do mesmo, tal processo deixa o animal estressado afetando sua produtividade. Estudos mostraram que a simples pesagem manual de um animal resulta em uma redução de 5% a 10% no peso e na produtividade (Ruchay *et al.*, 2022). Já com a pesagem remota, esse processo ocorre de maneira passiva, por meio da entrada voluntária dos animais em uma balança que dá acesso a água, e um sistema interligado entre o brinco do animal e os sensores da balança, que enviam o peso obtido via wireless.

A utilização desse sistema de pesagem permite o acompanhamento simultâneo da curva de ganho de peso dos animais, verificando o seu desenvolvimento individualizado e prevenindo problemas com perdas de produção. No entanto, existem vários fatores que podem levar a perda de informação ou obtenção de medidas incorretas.

Fenômenos naturais e biológicos, como chuva, barro e movimentos bruscos dos animais podem danificar o brinco (chip R-FID, em inglês Radio Frequency IDentification, de leitura dos pesos). Fatores técnicos e econômicos, como por exemplo a compra de material de baixa qualidade a fim de reduzir os custos dentro de uma operação de confinamento, podem resultar em uma porcentagem de falhas maior, levando até mesmo a interrupção da coleta de informações, ocasionando o não envio das pesagens para o sistema central. Essa falta de informação pode se estender por dias ou semanas, até que um novo chip seja reposto nos animais que apresentam falhas.

Além disso, o atraso no cadastramento das TAGS (cadastro único do animal) no início do confinamento do animal causa perdas das informações de pesagens, nesse que é considerado o período de adaptação do animal, em que podem ocorrer grandes variações do ganho de peso e indicativos de quando o animal não se adapta bem ao confinamento. Pode-se também cadastrar TAGS com o mesmo número, confundindo-se a informação entre os animais.

Outro fator de perda de informação ocorre com a queima de todas, ou parte das células de carga da balança dentro da baia, resultando no descarte das pesagens da balança. Também podem ocorrer falhas na antena de captação localizada na baia responsável por enviar as informações de pesagens para o sistema que armazena os dados. Outro fator é a falta de internet, atrasando a transmissão das informações coletadas para o sistema, de modo que até que não sejam atualizadas essas medidas serão analisadas como dados perdidos.

Somados aos problemas intrínsecos ao equipamento, clima, e manejo, pode haver perda de informação devido aos próprios animais, como por exemplo, animais que pulam para fora das baias (fugas) deixando de ser monitorados; e animais que adoecem e são removidos do sistema de pesagem por um período. Devido ao formato livre do sistema, no qual o animal se dirige a balança de forma

independente é comum que no momento da pesagem o animal se posicione apenas com duas de suas patas sobre a balança, ou até mesmo que se capture o peso de animais que sobem juntos na balança, obtendo-se assim informação que não correspondem a realidade. Todos esses fatores resultam na perda de informação ou medidas incorretas de valores importantes nesse processo, gerando bancos de dados incompletos e com valores discrepantes.

A análise de conjuntos de dados com observações perdidas, principalmente em estudos longitudinais, tem sido um grande desafio em diversas áreas do conhecimento (Donders *et al.*, 2006), uma vez que a utilização de métodos inadequados para a análise de dados com observações incompletas pode levar a conclusões equivocadas sobre o fenômeno em estudo.

Tem-se percebido um crescente interesse pelo desenvolvimento de técnicas que contemplem estudos com dados faltantes. Tais técnicas têm como objetivo preencher as lacunas de conjuntos de dados, a fim de possibilitar sua análise como dados completos, e são conhecidas como técnicas de imputação de dados. As imputações podem ser simples, quando apenas um valor é atribuído para cada observação ausente, ou múltipla, quando são gerados mais de um valor para cada valor não mensurado. Deve-se ressaltar que a incerteza relativa à imputação deve ser levada em conta, pois um método de imputação sem critérios pode criar mais problemas do que resolvê-los, distorcendo estimativas, erros padrões e resultados de testes de hipóteses, como descrito por Little e Rubin (2014).

O objetivo deste trabalho foi comparar a eficiência de alguns métodos de imputação múltipla em dados de pesagem de bovinos provenientes da pecuária de precisão. Os métodos aplicados foram o método da média preditiva (PMM), regressão linear bayesiana (BLR), floresta aleatória (RF) e árvore de classificação e regressão (CART), implementados no pacote MICE por meio do *software* R. Os métodos foram avaliados a partir de conjuntos de dados obtidos a partir da retirada aleatória de observações do conjunto de dados reais em diferentes porcentagens, sendo elas, 5%, 10% e 15%. Os critérios utilizados para a comparação dos métodos foram as estatísticas: raiz do erro quadrático médio, coeficiente de correlação de Pearson, coeficiente de acurácia de Willmott e coeficiente de desempenho. A partir da avaliação dos métodos também foi realizada a imputação dos valores ausentes do conjunto de dados originais.

2 REVISÃO DE LITERATURA

2.1 Imputação de dados

Anteriormente à 1970, os problemas que envolviam dados perdidos eram resolvidos por meio da edição desse conjunto de observações, nos quais as observações com valores faltantes eram removidas ou substituídas a partir de dados adicionais que foram observados, sendo apenas em 1976 desenvolvido um método de inferência por meio de dados incompletos (Rubin, 1976; Marwala, 2009).

A partir da década de 90, muitos métodos foram desenvolvidos para estimar dados perdidos, despertando o interesse dos pesquisadores em estudar a sensibilidade da estimativa desses dados no processo de decisão que utiliza essas componentes estimadas. Pesquisas foram conduzidas para determinar novas abordagens em busca de aproximar os valores obtidos a valores reais, como por exemplo, técnicas de inteligência computacional, como as redes neurais e técnicas de otimização, como por exemplo a computação evolucionária (Marwala, 2009).

Antes de abordar os métodos de imputação de dados, é importante entender o padrão e mecanismo de ausência dos dados em questão, uma vez que as técnicas de imputação utilizadas devem manter a relação das observações faltantes e presentes.

2.1.1 Padrão de ausência dos dados

Dados ausentes podem ser caracterizados por vários padrões, que determinam a forma como as unidades faltantes estão distribuídas em um conjuntos de dados (Enders, 2010), ou seja, de maneira geral os padrões descrevem a localização dos dados ausentes.

Little e Rubin (2014) em seu livro apresentam cada um desses padrões. Primeiramente toma-se uma estrutura retangular de dados, com n linhas e p colunas, $\mathbf{Y} = (y_{ij})$ com valores ausentes, em que o valor y_{ij} é o valor da variável j para o elemento i. Define-se a matriz $\mathbf{M} = (m_{ij})$, com n linhas e p colunas, indicadora da ausência e presença dos dados tal que, $m_{ij} = 1$ se y_{ij} é ausente e $m_{ij} = 0$ se y_{ij} é presente. Como o próprio nome indica, a matriz $\mathbf{M}$ define o padrão dos dados ausentes.

Os padrões mais frequentes estão descritos a seguir (Enders, 2010) e representados na Figura 2.1:

Padrão univariado: apresenta a falta de dados isoladamente em uma variável, caso comumente encontrado em estudos experimentais; nas ciências agrárias é conhecido por problema da parcela perdida.

Padrão de não resposta: ocorre quando para algumas variáveis faltam intencionalmente uma grande proporção de observações do conjunto de dados. Pesquisas com esse tipo de delineamento podem reduzir os custos da coleta de informação e a quantidade de respondentes necessária. Em geral tais pesquisas são realizadas por meio de questionários, como o censo e a PNAD.

Padrão monótono: se dá pela perda de informação a partir de um dado momento. Dessa forma, uma vez apresentada resposta ausente para determinado indivíduo, as que se seguem também serão ausentes. É bastante presente em pesquisas clínicas, no qual indivíduos participantes da pesquisa em algum momento deixam ou não podem continuar no estudo devido à alguns fatores.

Padrão geral: também conhecido como padrão arbitrário ou totalmente casual, apresenta dispersão de unidades ausentes por toda a matriz de dados, indicando um comportamento aleatório para os dados ausentes, embora possa existir uma relação entre a falta desses valores. É importante destacar que os métodos de imputação múltipla são eficientes para esta configuração.

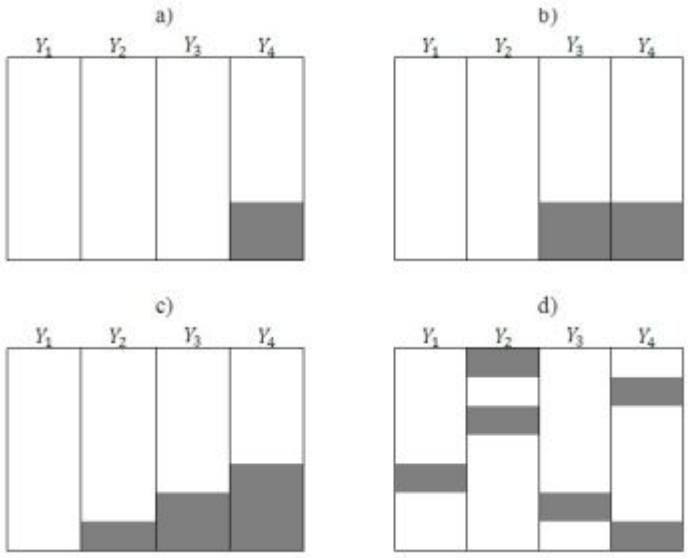


Figura 2.1. Ilustração dos padrões de ausência de dados. Os dados ausentes estão representados em cinza conforme os seguintes padrões: a) Padrão univariado, b) Padrão de não resposta, c) Padrão monótono e d) Padrão geral.

Fonte: Silva (2012)

De acordo com BERGAMO (2007) dentre os padrões apresentados os principais são o monótono e o geral. É importante ressaltar que alguns métodos destinados a análise de dados perdidos aplicam-se a qualquer padrão de dados ausentes, no entanto, outros métodos são restritos a determinados padrões.

2.1.2 Mecanismo de ausência dos dados

De acordo com Rubin (1976) toda observação em um conjunto de dados tem alguma probabilidade de estar ausente. Para ele, existem mecanismos que descrevem diferentes possibilidades nos quais a probabilidade de causa da ausência de valores está, ou não, relacionada aos dados, isto é, descrevem possíveis relações entre as variáveis medidas e a probabilidade de dados ausentes (Enders, 2010). Esses mecanismos não apontam a causalidade da ocorrência de valores perdidos. Os três principais mecanismos que são discutidos na literatura foram descritos a seguir:

Ausência totalmente aleatória (*Missing completely at random* - MCAR): ocorre quando a ausência de dados não tem relação com a variável que possui informações faltantes e nem com outras variáveis envolvidas no estudo, ou seja, a probabilidade do dado faltante é independente dos valores observados e também dos ausentes. Esse processo é considerado como uma ausência puramente aleatória (Enders, 2010). Por exemplo, durante uma pesagem determinada balança ficou sem bateria, alguns dos valores faltaram por conta do acaso.

Ausência aleatória (*Missing at random* - MAR): ocorre de forma aleatória indicando que a perda da observação ocorreu devido à alguma informação contida no conjunto de dados, ou seja, a probabilidade da resposta ausente está relacionada às partes observadas, e é independente a parte dos valores perdidos (Donders *et al.*, 2006). Por exemplo, em uma pesquisa sobre peso, pessoas do sexo feminino tendem a não declarar seu peso.

Ausência não aleatória (*Missing not at random* - MNAR): ocorre quando a falta depende das informações da variável que contém a falta e/ou de outras informações não mensuradas, ou seja, a probabilidade dos valores ausentes depende dos valores que não foram observados (Pedersen *et al.*, 2017). Por exemplo, em uma pesquisa sobre renda, em geral pessoas que possuem renda muito

baixa ou muito alta, tendem a não responder a pesquisa.

2.2 Métodos de imputação de dados

O desenvolvimento de métodos que contemplem a análise de dados perdidos é relativamente recente, de acordo com Little e Rubin (2014), métodos de imputação de dados podem ser agrupados nas seguintes categorias, que não são mutuamente exclusivas:

Procedimentos Baseados em Unidades Completamente Registradas: quando algumas unidades não são registradas para determinadas variáveis, sendo sua análise determinada por simplesmente descartar as unidades observacionais incompletas e analisar apenas unidades completas. Geralmente é de fácil realização e pode ser satisfatória para pequenas quantidades de dados perdidos. No entanto, pode levar a vieses grandes, e geralmente não é muito eficiente, especialmente ao fazer inferências para subpopulações.

Procedimentos de Ponderação: modificam os pesos das variáveis em uma tentativa de ajustar os dados ausentes, como se eles fizessem parte da amostra, nas quais as inferências de aleatorização a partir de dados de levantamento amostral sem a informação faltante, em geral medem as unidades amostradas por seus pesos de projeto, que são inversamente proporcionais às suas probabilidades de seleção.

Procedimentos baseados em Imputação: nos quais os valores omissos são preenchidos e os dados completos resultantes são analisados por métodos padrão. Os procedimentos comumente usados para imputação incluem a imputação de *hot deck*, imputação média e imputação de regressão.

Procedimentos baseados em Modelo: os quais a imputação é gerada pela definição de um modelo para os dados observados e com base nas inferências sobre a probabilidade ou distribuição posterior sob esse modelo, com parâmetros estimados por procedimentos como a máxima verossimilhança.

Os métodos apresentados anteriormente, tais como imputações via média, *hot deck*, regressão, última observação localizada e método indicador são exemplos de imputações simples. Destaca-se que a simplicidade de tais métodos está atrelada a limitações importantes, como o fato de não considerar a incerteza da imputação, dessa forma, os desvios padrão calculados nos dados completos não são estimados corretamente (BERGAMO, 2007). A utilização da imputação múltipla (IM) é uma forma de contornar esse problema. Métodos de imputação bayesianos também resolveriam essa questão das incertezas, uma vez que permitem ter a distribuição a posteriori para cada valor faltante.

2.3 Imputação Múltipla (IM)

A IM é única no sentido de que fornece um mecanismo para lidar com a incerteza inerente às próprias imputações. Esse método foi desenvolvido por Donald B. Rubin na década de 70 que observou que a atribuição de um valor (imputação simples) para o valor faltante não poderia estar correta, sendo necessário um modelo para relacionar dados não observados a dados observados, e que, mesmo para certo modelo, os valores imputados por meio da imputação simples não puderam ser calculados com certeza. A solução proposta para este problema foi criar múltiplas imputações que refletiam a incerteza das estimativas dos dados em falta (Van Buuren, 2018).

A ideia fundamental do procedimento, representada pela Figura 2.2, é substituir cada valor ausente por um conjunto de m valores imputados, seguindo os três passos a seguir (BERGAMO, 2007; Silva, 2012):

- Imputação: A partir do banco de dados incompletos, os valores ausentes são preenchidos m vezes, com $m > 1$, gerando m conjuntos completos.
- Análise: Os m conjuntos são analisados por meio de técnicas estatísticas de interesse.

- Combinação ou agrupamento: Os m resultados obtidos são combinados afim de obter uma única inferência dos resultados a serem imputados.

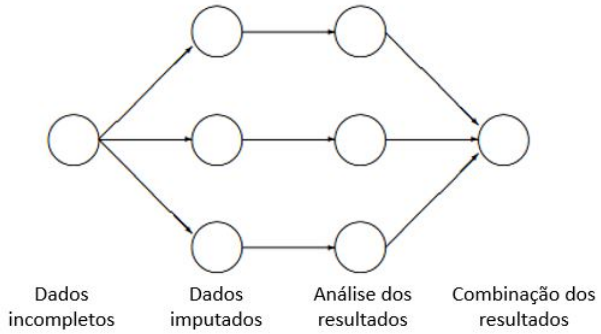


Figura 2.2. Esquema dos principais passos na imputação múltipla.
Fonte: Van Buuren (2018)

Neste trabalho foram aplicados quatro métodos de imputação que têm como base a imputação múltipla, sendo eles: o PMM, BLR, CART e o RF. Todos os métodos foram implementados utilizando o pacote *mice* no *software* R (R Core Team, 2021). O algoritmo que dá nome ao pacote *mice* consiste em um método robusto de imputação múltipla por equações encadeadas. O procedimento consiste em "preencher" (imputar) valores ausentes de um determinado banco de dados fornecido por meio de uma série iterativa de modelos preditivos. Em cada iteração, cada variável especificada no banco de dados é imputada usando as outras variáveis deste mesmo conjunto de dados. Essas iterações devem ser simuladas até que uma determinada convergência possa ter sido alcançada (van Buuren, 2022).

2.4 Métodos de imputação

Os métodos de imputação escolhidos para avaliação neste trabalho partem do mesmo princípio, sendo aplicados a partir de uma regressão linear (Khan e Hoque, 2020), além disso, dois deles pertencem ao conjunto de técnicas atribuídos à *machine learning*, área de estudo que está em ascensão nos últimos anos, sendo eles os métodos CART e RF. A seguir cada um dos métodos será apresentado de forma mais detalhada. Esses métodos foram escolhidos por serem adequados para a imputação de variáveis quantitativas, como é o caso da variável peso que foi imputada neste estudo.

2.4.1 Método da média preditiva (PMM - *Predictive Mean Matching*)

O método PMM foi desenvolvido por Little (1988) e destaca-se por ser uma técnica de imputação múltipla semi-paramétrica que consiste em imputar valores ausentes se baseando no método do vizinho mais próximo, tendo sua distância baseada nos valores esperados das variáveis faltantes condicionadas à covariáveis observadas, combinando assim elementos de regressão linear (Vink *et al.*, 2014). Neste contexto, a variável de interesse será a variável a ser imputada, e as demais variáveis presentes no estudo são as regressoras.

Em resumo, cria-se um modelo preditivo a partir dos valores completos das variáveis relacionadas, que são usados para calcular os valores preditos para os valores de Y observados (Y_{obs}) e Y faltantes (Y_{mis}). A partir do valor predito para o Y faltante, procura-se o valor predito mais próximo a este de um Y_{obs}. Para cada valor ausente o método seleciona um pequeno conjunto de candidatos (3, 5 ou 10 membros) de todos os casos completos que têm valores preditos mais próximo ao valor predito do valor ausente, dos quais, um deles é escolhido aleatoriamente e seu respectivo valor observado é utilizado como

o valor imputado para substituir o valor faltante (Van Buuren, 2018). A variabilidade entre as imputações é gerada por meio dos passos que servem para estimar os parâmetros do modelo e a sua variância, que são repetidos m vezes (Nunes *et al.*, 2009). A seguir será apresentado o algoritmo utilizado para realizar a imputação múltipla dos Y_{mis} por meio do PMM como descrito em Vink *et al.* (2014).

Primeiramente, foram estabelecidas as notações necessárias. Seja Y uma variável contínua incompleta com n unidades amostrais, dividida em dois grupos, Y_{obs} e Y_{mis}, que representam os valores observados e os valores ausentes em Y, respectivamente. Além disso, $X = (X_1, ..., X_p)$ é o conjunto de p covariáveis totalmente observadas, que também serão repartidas em X_{obs} e X_{mis}, que correspondem, aos valores das covariáveis associadas ao grupo de valores observados em Y, e aos valores de covariáveis associadas aos valores ausentes em Y, respectivamente. O número de unidades do grupo de valores observados será representada por n_{obs}, e o número de unidades do grupo composto pelos valores ausentes será representado por n_{mis}.

Algoritmo:

1. Utilizar a regressão linear dos Y_{obs} dado X_{obs} para estimar $\hat{\beta}$, $\hat{\sigma}$ e $\hat{\varepsilon}$ por meio dos mínimos quadrados.
2. Considerar σ^{2*} como $\sigma^{2*} = \frac{\hat{\varepsilon}^t\hat{\varepsilon}}{A}$, em que A é uma variável χ^2 com $n_{obs} - r$ graus de liberdade, sendo r o número de parâmetros do modelo de regressão.
3. Considerar β^* de uma distribuição normal multivariada centrada em $\hat{\beta}$ com matriz de covariância $\sigma^{2*}(X_{obs}^t X_{obs})^{-1}$.
4. Calcular $\hat{Y}_{obs} = X_{obs}\hat{\beta}$ e $\hat{Y}_{mis} = X_{mis}\beta^*$.
5. Para cada $\hat{Y}_{mis,i}$ encontrar $\Delta = |\hat{Y}_{obs,i} - \hat{Y}_{mis,i}|$.
6. Amostrar aleatoriamente um valor dentre $(\Delta^{(1)}, \Delta^{(2)}, \Delta^{(3)})$, em que $\Delta^{(1)}$, $\Delta^{(2)}$ e $\Delta^{(3)}$ são os três menores diferenças em Δ, respectivamente, e tome o correspondente $Y_{obs,i}$, como o valor imputado.
7. Repetir os 6 passos anteriores m vezes, sempre salvando o conjunto de dados completo, sendo m o número de iterações que serão consideradas.

Este método foi selecionado devido a sua facilidade de utilização, uma vez que é o método padrão implementado pelo pacote *mice*. De acordo com a descrição do métodos, é escolhido um pequeno grupo de valores candidatos para substituir o valor faltante, neste trabalho foram utilizados cinco candidatos possíveis para serem escolhidos aleatoriamente e substituir o valor ausente, pois é o *default* da função.

2.4.2 Regressão linear Bayesiana (BLR - *Bayesian Linear Regression*)

Outro método abordado neste trabalho é a imputação por BLR, técnica de imputação baseada em um modelo de regressão, no qual a análise estatística é realizada no contexto de inferência bayesiana. Assim como no método PMM, utiliza-se uma regressão linear múltipla para predizer Y_i de um conjunto de covariáveis X_i. No entanto, neste método a regressão linear é formada com a ajuda de distribuições de probabilidades ao contrário de estimações pontuais (Khan e Hoque, 2020), tornando-se mais vantajoso por capturar toda a variabilidade associada ao valor imputado por meio da distribuição a posteriori.

Desta forma, a variável resposta Y que é a variável a ser imputada, não é avaliada como um único valor, mas sim é extraída de uma distribuição de probabilidades. Portanto, os parâmetros $\boldsymbol{\beta}$ e σ são estimados a partir de uma distribuição *a posteriori* própria, ou seja, o método BLR visa obter as distribuições*a posteriori* para os parâmetros do modelo em vez de encontrar um único melhor valor para eles.

A equação que representa o modelo é descrita a seguir como apresentado em Malaguti e de Faria (2020):

$$Y_{mis} = \beta_0 + \beta_1 X_{mis} + \epsilon \tag{2.1}$$

em que Y_{mis} são os valores imputados, X_{mis} são o conjunto de covariáveis dos dados faltantes, β_0, β_1 e ϵ são retirados aleatoriamente da distribuição *a posteriori* dado os valores observados.

Analogamente ao método PMM, são estimados $\boldsymbol{\beta}$ e σ, no entanto os m valores usados para as imputações são os próprios valores preditos para Y_{mis} gerados por m repetições da estimação de $\boldsymbol{\beta}$ e σ. No pacote *mice* também é possível realizar a imputação por meio desse método, basta alterar o argumento da função para *norm*.

2.4.3 Árvore de classificação e regressão (CART - Classification And Regression Trees)

Proposto por Breiman *et al.* (1984), CART são uma classe de algoritmos de aprendizado de máquinas bastante popular. Os modelos CART buscam preditores e pontos de cortes para esses preditores, de forma que a amostra possa ser dividida em subamostras mais homogêneas. O processo é repetido para cada subamostra, até que seja gerada uma árvore binária.

Como apresentado por Van Buuren (2018) o método possui algumas propriedades que o tornam interessante para imputação, entre essas características está a robustez contra *outliers*, poder lidar com a multicolinearidade e distribuições enviesadas, bem como ser suficientemente flexível para ajustar interações e relações não lineares.

A utilização do CART como método de imputação foi impulsionada por vários autores, seguindo diferentes caminhos. Neste trabalho foi utilizada a abordagem proposta por Burgette e Reiter (2010); Shah *et al.* (2014); Doove *et al.* (2014) que em em seus artigos aprimoraram o uso das árvores para imputação, melhorando suas habilidades em lidar com diferentes relações entre as variáveis e ainda a aplicação em resultados e preditores contínuos e categóricos.

A utilização de árvores para a imputação de valores faltantes é descrita da seguinte forma (Doove *et al.*, 2014):

1. Ajustar uma árvore de classificação ou regressão por particionamento recursivo;
2. Para cada Y_{mis}, encontre o nó terminal de acordo com a árvore ajustada;
3. Faça uma extração aleatória entre os membros no nó e tome o valor observado desse sorteio como a imputação.

O conceito é idêntico ao método PMM, no entanto a "média preditiva" agora é calculada por um modelo de árvore e não mais por um modelo de regressão. Neste caso, a incerteza do parâmetro pode ser incorporada ajustando a árvore em uma amostra aleatorizada (Van Buuren, 2018).

2.4.4 Floresta aleatória (RF - Random Forest)

O método RF é um método de classificação e regressão que realiza a predição de valores por meio de uma combinação de árvores de regressão, em que cada árvore depende dos valores de um vetor aleatório amostrado de forma independente, e com mesma distribuição para todas as árvores de regressão (Breiman, 2001).

De acordo com Shah *et al.* (2014), o RF pode acomodar relações e interações não lineares, seu mecanismo consiste em agregar, por meio de *bootstrap*, várias árvores de regressão para diminuir a

possibilidade de um super ajuste e combina as previsões de muitas árvores para produzir previsões mais precisas.

O processo de construção do algoritmo RF no pacote *mice* é descrito a seguir:

Suponha uma matriz de dados Y, em que Y_j é a j-ésima coluna das variáveis parcialmente observadas (ordenadas para ter números crescentes de valores ausentes para que os modelos sejam construídos com o máximo de informações possível), p é o número de variáveis parcialmente observadas, Y_j^{obs} é o dados observados e Y_j^{mis} são os dados ausentes na j-ésima coluna, e $\dot{Y}$ é a matriz de dados atualmente imputada Y.

1. Para $j = 1, \ldots, p$, preencha as imputações iniciais $\dot{Y}_j^0$ por sorteios aleatórios de Y_j^{obs} e defina uma matriz de dados $\dot{Y}$.

2. Para $j = 1, \ldots, p$, substitua $\dot{Y}_j^0$ da seguinte forma, gerando um conjunto de dados imputado:

   a) Extraia k amostras geradas por *bootstrap* a partir de $\dot{Y}$, restrito aos membros em Y_j^{obs}.

   b) Ajuste uma árvore em cada amostra de *bootstrap* traçada no passo 2a. Isto resulta em k árvores, onde cada árvore tem várias folhas. Cada folha inclui um subconjunto de Y_j^{obs} que serão chamados doadores.

   c) Para os membros em Y_j^{mis}, determinar em que folha irão parar de acordo com as k árvores ajustadas no passo 2b). Isto resulta em k folhas com doadores por membro de Y_j^{mis}.

   d) Para os membros em Y_j^{mis}, pegue todos os doadores juntos das k folhas determinadas no passo 2c) juntos e selecione aleatoriamente um valor Y^{obs} dos doadores. Substitua os valores originalmente ausentes de $\dot{Y}_j^0$ por estes valores de imputação e anexe o valor completo de versão de $\dot{Y}_j$ a $\dot{Y}$ antes de incrementar j.

3. Repetir o passo 2 até ter realizado l (número de iterações) vezes.

4. Repetir os passos 1-3 m vezes, produzindo m conjuntos imputados.

Nota-se que a principal diferença entre os métodos RF e CART é que em RF é construído um conjunto de árvores de decisão para imputar cada um dos valores ausentes, enquanto o método CART cria apenas uma árvore de decisão.

2.5 Critérios de comparação

Para investigar a eficiência dos métodos de imputação citados, foram utilizadas quatro medidas de comparação, que serão descritas nesta seção.

2.5.1 Raiz do erro quadrático médio (RMSE - Root Mean Squared Error)

A raiz do erro quadrático médio é uma das medidas mais utilizadas para previsão numérica. Essa medida representa o desvio-padrão da amostra da diferença entre o valor real e o estimado. A RMSE é dada por

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{n}}, \tag{2.2}$$

em que Y_i é o valor original e Y_i^* é o valor imputado e n é a quantidade de valores da amostra de dados imputados.

Essa medida indica o quão próximo o valor imputado está do valor observado, quanto menor o valor da RMSE mais eficiente será a imputação.

2.5.2 Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson, r, é um método que permite medir a correlação entre duas variáveis. Têm-se a seguinte fórmula

$$r = \frac{\sum_{i=1}^{n} y_i y_i^* - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i^*}{n}}{\sqrt{\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i^2\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^{*2} - \frac{\left(\sum_{i=1}^{n} y_i^{*2}\right)^2}{n}\right)}}, \tag{2.3}$$

sendo y_i o valor original, y_i^* o valor imputado e n a quantidade de valores da amostra de dados imputados. Observa-se que $|r| < 1$, e quanto mais próximo de 1 ou -1, temos uma correlação total positiva ou negativa, respectivamente. No caso de $r = 0$, têm-se que não há correlação entre as variáveis. Segundo Dancey e Reidy (2017), a classificação para esse valor se dá por: $0,10 < |r| \leq 0,30$ (fraco); $0,40 < |r| \leq 0,6$ (moderado); $0,70 < |r| \leq 1,0$ (forte).

2.5.3 Índice de acurácia de Willmott

Esse índice foi proposto por Willmott *et al.* (1985), representado por d e sua fórmula é dada por

$$d = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{\sum_{i=1}^{n} (|y_i^* - \bar{y}| + |y_i - \bar{y}|)^2}, \tag{2.4}$$

em que y_i representa o valor original, y_i^* o valor imputado, $\bar{y}$ a média dos valores removidos e $0 < d < 1$.

O valor 1 indica a concordância perfeita entre os valores imputados e os observados, já o valor 0 denota que não há nenhuma concordância.

2.5.4 Índice de desempenho (c)

O índice de desempenho é dado pelo produto entre o coeficiente de correlação de Pearson (r) e o índice de acurácia de Willmott (d) (Camargo e Sentelhas, 1997). Portanto, têm-se a seguinte fórmula

$$c = r.d \tag{2.5}$$

A Tabela 2.1 apresenta o critério de classificação de desempenho do Índice de desempenho (c) de acordo com Camargo e Sentelhas (1997).

Tabela 2.1. Interpretação qualitativa para o índice de desempenho (c).

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

3 MATERIAL E MÉTODOS

3.1 Banco de dados

O conjunto de dados utilizado neste trabalho é proveniente do estudo conduzido pela @Tech Inovações Tecnológicas para a Agropecuária em parceria com o confinamento e Cooperativa Coplacana, em Piracicaba, São Paulo. O objetivo dos pesquisadores foi monitorar a obtenção de peso de bovinos, até atingir seu peso ótimo. O monitoramento do peso dos animais foi realizado por um sistema de plataformas de pesagem automática instalado à frente do bebedouro dentro da baia de confinamento (Figura 3.1). Quando o animal bebe água, ele passa obrigatoriamente pela plataforma de pesagem, momento em que será identificado via brinco eletrônico e assim pesado. Os dados coletados no momento da pesagem são transmitidos via wireless para um computador. As informações de pesagens são fundamentais para o cálculo da taxa de ganho e o estudo da desaceleração do crescimento animal.

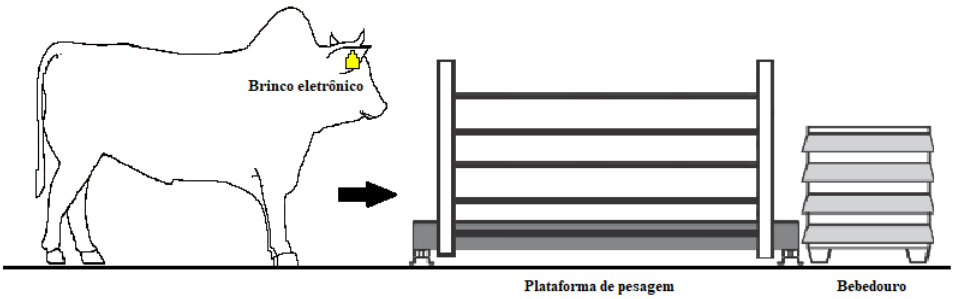


Figura 3.1. Esquema representativo do sistema de plataformas de pesagem automática de cada animal. Fonte: Elaboração Própria (2020)

O estudo utilizou a pesagem de um grupo de 38 animais, das raças Nelore (Nel) e Cruzado Britânico (CB), os quais todos eram machos divididos entre macho inteiro (MI) e castrado (MC). Um resumo da formação do grupo de animais é apresentado na Tabela 3.1.

Tabela 3.1. Quantidade de animais na composição do lote de acordo com a raça e gênero

Raça	Gênero	Quantidade
Nel	MI	14
	MC	13
CB	MI	3
	MC	8

Os animais foram pesados por 85 dias, no período de 20 de junho à 12 de setembro de 2017. Cada animal foi pesado um número aleatório de vezes, uma vez que, o total de pesagens de cada animal depende da quantidade de vezes que ele passa pela plataforma de pesagem no momento em que se direciona ao bebedouro.

Além das variáveis peso, raça e gênero dos animais, o conjunto também apresenta o escore de condição corporal (ecc), que representa o estado nutricional do animal de forma estimada, uma vez que é obtida por meio de avaliação visual e/ou tátil, e consiste em uma importante ferramenta de manejo, pois permite avaliar as reservas energéticas do animal e indicar práticas nutricionais a serem adotadas (Machado *et al.*, 2008). Além disso, o frame de cada animal também compõe o banco de dados, que é a representação numérica do esqueleto do animal; está relacionado com a produtividade do animal, e sua avaliação em bovinos para produção de carne é interessante para a seleção de animais com maior potencial (Mota *et al.*, 2014).

Com relação as informações obtidas sobre o ambiente ao qual os animais estavam confinados, foram informados os valores de temperatura média, umidade relativa média, temperatura máxima e mínima, ocorrência e quantidade de chuvas e dias em que houve manejo dos animais.

Os perfis de temperatura e umidade relativa do ar ao longo do período de estudo, são apresentados nas Figuras 3.2 e 3.3, respectivamente. Por meio dos perfis de peso dos animais (Figura 3.4), é possível observar, de forma exploratória, que o peso dos animais nos meses precedentes ao abate têm um comportamento próximo ao linear. Observando os perfis de temperatura, umidade relativa e do ganho de peso animal, não espera-se que os valores de pesagem do animal tenham sido influenciados por alterações na temperatura e umidade relativa do ar.

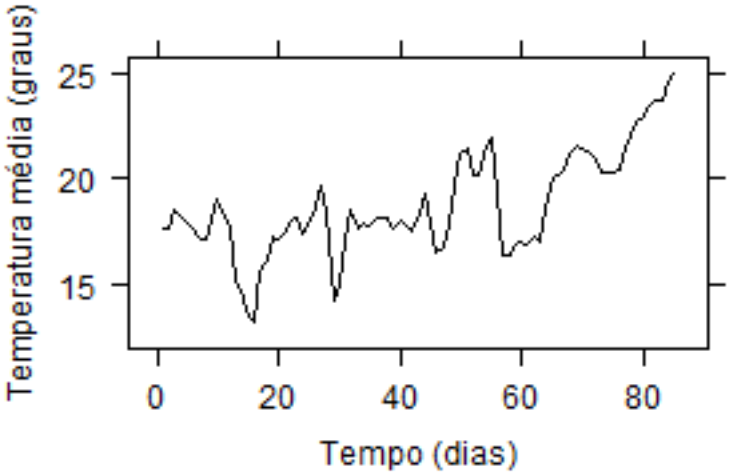


Figura 3.2. Gráfico de perfil da temperatura média durante o período do estudo.

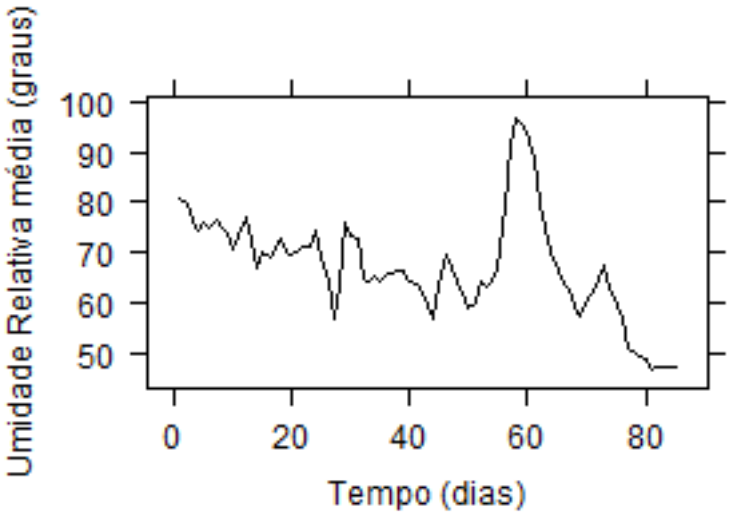


Figura 3.3. Gráfico de perfil da umidade relativa média durante o período do estudo.

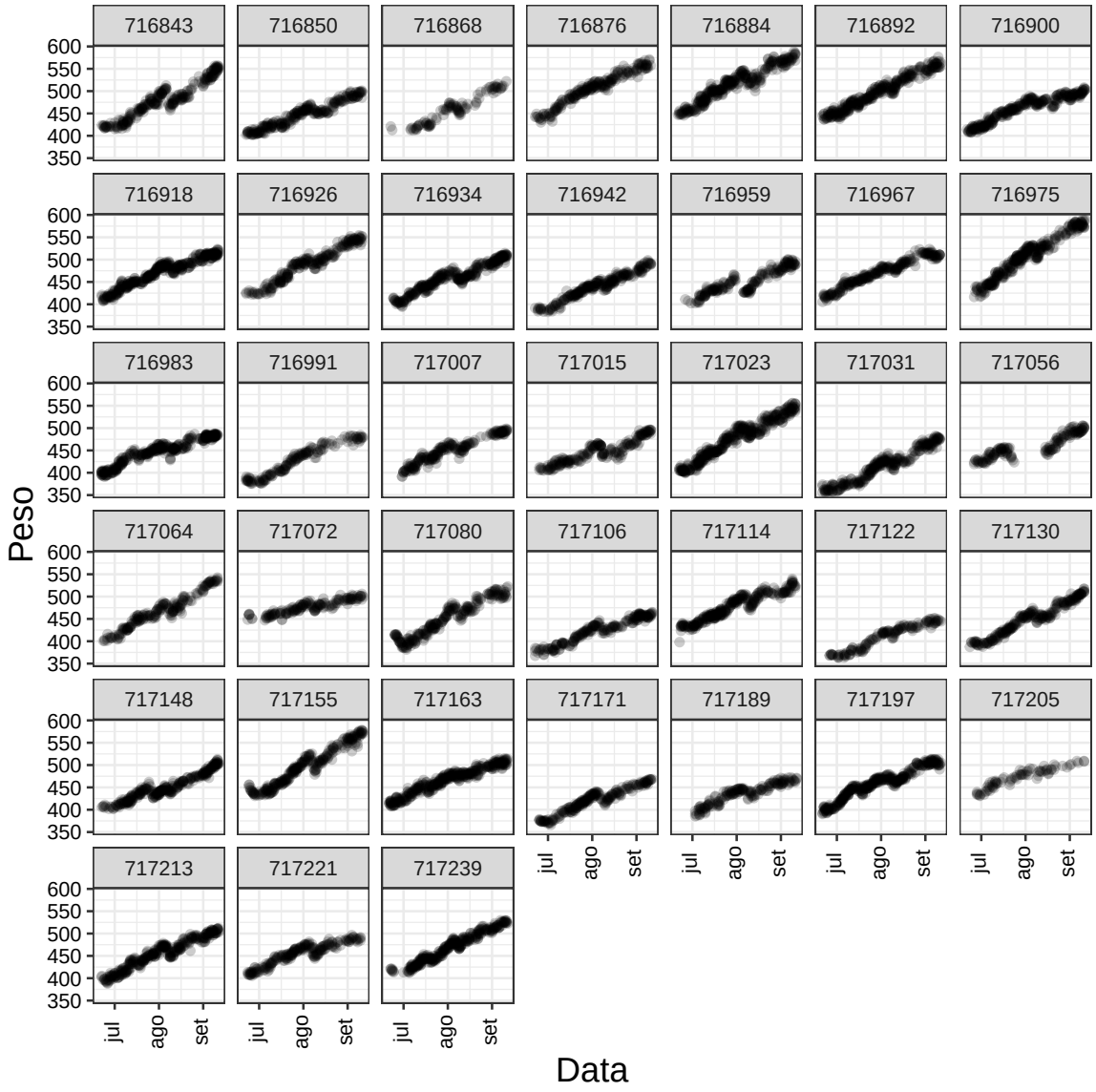


Figura 3.4. Gráfico de perfil de peso de cada um dos animais ao longo do tempo, no período de 20 de junho à 12 de setembro de 2017.

A Figura 3.5 representa a quantidade de valores perdidos para cada um dos animais, que são identificados no gráfico pelo código do dispositivo eletrônico (numeração do Brinco). É possível observar que dois animais tiveram mais de 30 valores perdidos, enquanto que aproximadamente 26% dos animais tiveram menos do que cinco observações perdidas. É importante destacar, que cada valor perdido representa um dia de sem observação, ou seja, naquele dia não foi coletada nenhuma pesagem daquele animal. No banco de dados há 363 falhas, totalizando aproximadamente 3,8% de informação perdida, estes valores que não foram coletados que causam as lacunas apresentadas nos perfis de ganho de peso dos animais, representados pela Figura 3.4.

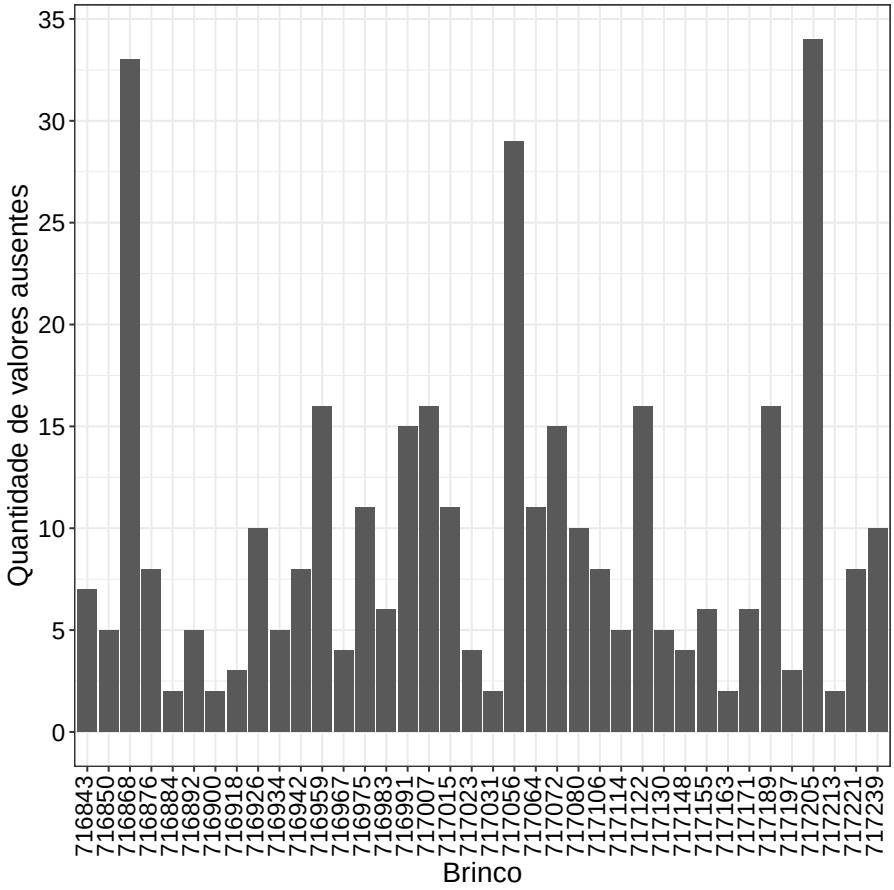


Figura 3.5. Quantidade de dados faltantes por animal, no período de 20 de junho à 12 de setembro de 2017.

A distribuição das informações ausentes ao longo dos dias, é apresentada pela Figura 3.6, na qual quadrados brancos representam valores que foram avaliados, enquanto que os valores não avaliados (NA) estão em vermelho. Por meio desta figura, por exemplo, pode-se ver que o animal identificado pelo código "717064" teve várias observações faltantes, em sequência, durante a condução do estudo. Isso ocorreu devido a uma falha do dispositivo, e após a troca, as pesagens do animal voltaram a ser registradas. Também é possível verificar que durante o período inicial da pesquisa, todos os animais tiveram observações perdidas. Pode-se entender que os valores perdidos neste caso, ocorrem por fatores externos, e não por conta da variável peso ou de qualquer outra variável presente no estudo. Desta forma, adotou-se o mecanismo de perda como MCAR, ou seja, completamente aleatório.

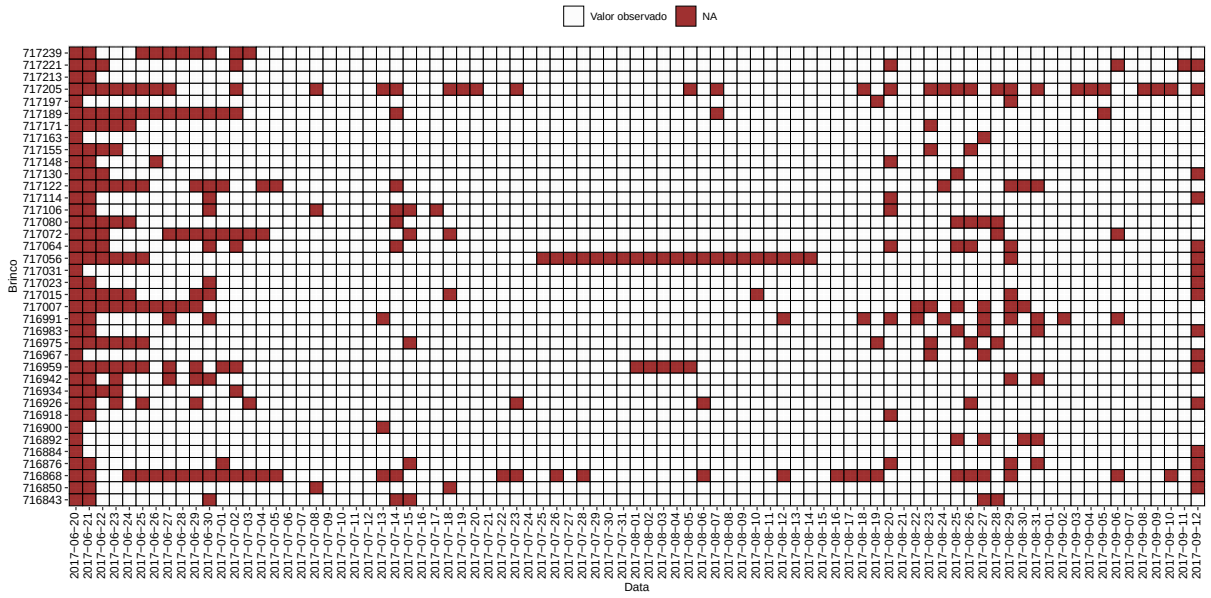


Figura 3.6. Ausência e presença de pesagens por animal ao longo do tempo.

3.2 Estudo segundo retirada aleatória dos dados

Para ser possível realizar a comparação entre os métodos de imputação foi necessário avaliar quão precisa foi a imputação da observação ausente. Desta forma, fez-se necessário tornar conhecido o valor que seria imputado, uma vez que os critérios de comparação escolhidos necessitavam do valor observado e do valor imputado.

Como já foi apresentado na seção 3.1, o conjunto utilizado neste trabalho já possuía valores faltantes, então por meio destas observações não seria possível avaliar a eficiência dos métodos, por meio dos critérios selecionados. A forma encontrada neste trabalho de contornar esse problema, foi remover as observações ausentes do conjunto de dados original, tornando-o completo, e a partir desse novo conjunto criar novos bancos de dados com diferentes porcentagens de valores faltantes (5%, 10% e 15%).

De forma aleatória foram removidas observações com o objetivo de construir 1000 novos bancos de dados, para cada um dos três cenários distintos de ausência de informação para a variável resposta. Os 1000 conjuntos de dados simulados com ausência de 5% das observações foram representados por BI-5. Os novos conjuntos com ausências de 10% e 15% foram denotados por BI-10 e BI-15, respectivamente.

A partir desses novos bancos, foram aplicados os métodos de imputação descritos na seção 2.4 (PMM, BLR, CART e RF) com número de iterações (m) iguais a 5 e 10. Para cada aplicação foram calculadas as medidas utilizadas pelos critérios de comparação descritos na seção 2.5. Esse processo foi executado de acordo com o diagrama apresentado na Figura 3.7.

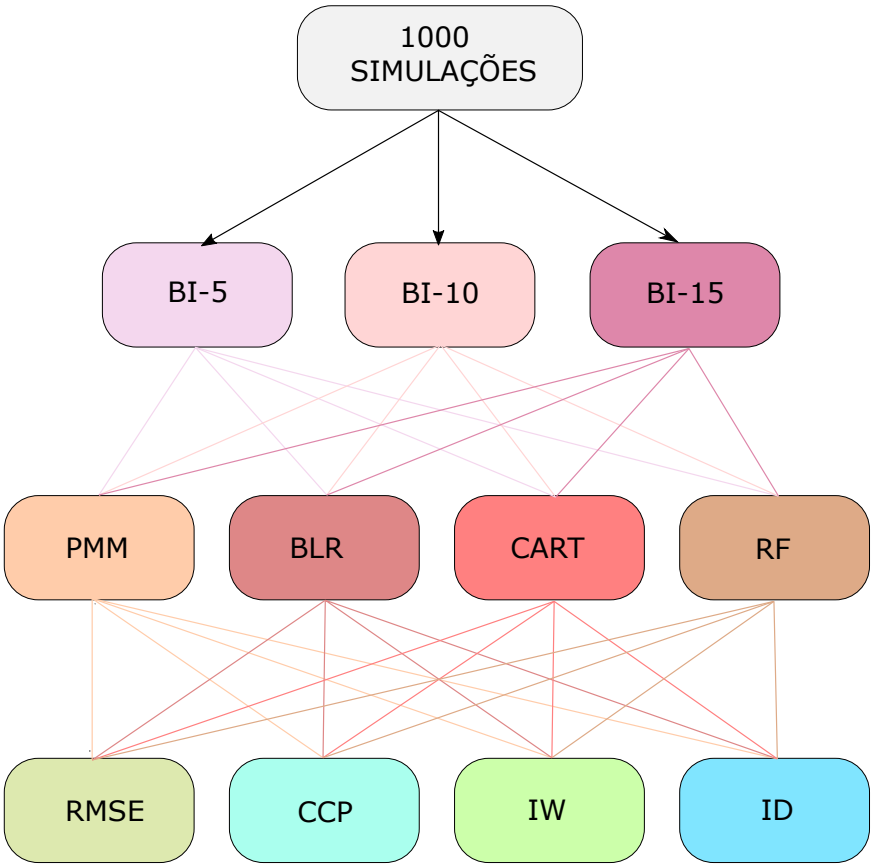


Figura 3.7. Diagrama do processo de imputação (com m iterações) e obtenção dos critérios de comparação. CCP, IW e ID representam Coeficiente de correlação de Pearson, Índice de Willmott e Índice de desempenho, respectivamente. As demais abreviações estão descritas no texto. Esse esquema foi repetido para $m = 5$ e $m = 10$.

Fonte: Elaboração Própria (2023)

As etapas de construção dos novos conjuntos de dados, aplicação dos métodos de imputação e cálculo das medidas utilizadas nos critérios de comparação, foram realizadas por meio de algoritmos implementados no *software* R.

Além disso, devido ao esforço computacional para gerar os resultados de cada um dos 3000 mil bancos de dados para cada método e critério de comparação utilizado, utilizou-se a estratégia de paralelização dos algoritmos para reduzir o tempo computacional. A criação dos novos bancos de dados foi feita de forma aleatória, e a implementação dos referidos métodos de imputação foi realizada por funções do pacote *mice* (van Buuren, 2022).

Já os valores para os critérios de comparação foram calculados a partir de suas respectivas fórmulas matemáticas.

4 RESULTADOS E DISCUSSÃO

4.1 Análise dos métodos por estudo de simulação

A partir dos resultados individuais de cada uma das análises e cenários obtidos, foram construídos *box plots* para cada um dos métodos, em relação as variações de porcentagens de valores faltantes (5%, 10% e 15%), que foram removidas aleatoriamente, e com 5 e 10 iterações.

O *box plot* ou gráfico de caixas é utilizado para resumir e comparar visualmente grupos de dados. Ele é composto pelos valores mínimos e máximos do intervalo, quartis superior e inferior e pela mediana, que mostram um resumo da distribuição dos dados. Geralmente, é representado por uma caixa retangular usada para indicar as posições dos quartis superior e inferior; o interior dessa caixa indica a área entre os quartis superior e inferior e consiste em 50% da distribuição (Potter *et al.*, 2006). A partir do retângulo, segue-se uma linha até o ponto mais remoto que não exceda o limite superior, dado por $LS = q_3 + (1,5)d_q$, e de forma similar, da parte inferior do retângulo, segue uma linha até o limite inferior, representado por $LI = q_1 - (1,5)d_q$, em que q_1 é o 1° Quartil, q_3 é 3° Quartil, e d_q é a distância interquartil (Bussab e Morettin, 2010). Observações que estiverem fora desses limites são chamados de pontos discrepantes. Por fim a linha localizada no interior das caixas representa a mediana desse conjunto de dados, que também é o 2° Quartil.

As Figuras 4.1, 4.2 e 4.3 representam os *box plots* das medidas calculadas por meio do critério RMSE para os métodos de imputação PMM, BLR, CART e RF, considerando os 1000 conjuntos de dados em cada um dos cenários de porcentagem de ausência dos dados, utilizando 10 iterações. Pode-se observar que os métodos PMM, BLR e CART apresentaram variabilidade muito semelhantes, quando considerados os resultados dentro de cada cenário e entre os cenários.

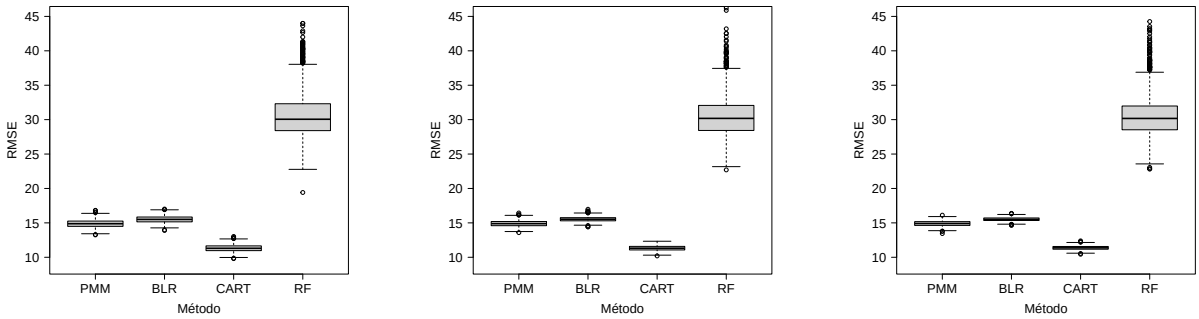
Observa-se na Figura 4.1 que em cada um dos cenários o método CART demonstrou melhor desempenho, apresentando os menores valores para a RMSE (entre 10 e 15, aproximadamente, para todas as porcentagens de ausência de dados), o que indica que os valores imputados via CART foram os que mais se aproximaram dos valores verdadeiros. Em contra partida, o método RF apresentou os maiores valores para a RMSE (entre 20 e 45, para todas as porcentagens de ausência de dados).

Para os demais critérios avaliados, foram obtidos resultados semelhantes, com CART destacando-se positivamente e o RF negativamente. Na Figura 4.2 são apresentados os *box plots* considerando os valores calculados a partir do índice de acurácia de Willmott, enquanto que os métodos CART, PMM e BLR apresentam valores compreendidos entre 0,95 e 1; o método RF tem valores entre 0,6 e 0,95, com maior concentração de valores entre 0,83 e 0,87. Lembrando que, quanto mais próximo de 1, melhor a correspondência com os valores originais.

Na Figura 4.3 tem-se um resultado muito similar, que vale a pena ser discutido uma vez que para o índice de desempenho existe uma tabela de classificação apresentada na seção 2.5.4. De acordo com o apresentado em cada *box plot*, os valores do índice de desempenho, foram superiores a 0,85 em todos os cenários para os métodos CART, PMM e BLR, sendo classificados como ótimo, já para RF os valores estiveram entre 0,2 a 0,8 para todos os cenários de porcentagem de ausência de dados, sendo em sua maioria classificados como mediano e sofrível. Resultado semelhante a esse foi apresentado por Gasparetto *et al.* (2021), que como neste trabalho, compararam três métodos de imputação em diferente intensidades de dados faltantes (5%, 10% e 15%) para dados de precipitação pluvial, em seu estudo observaram que o método PMM foi o que forneceu melhores resultados, também elegendo o método RF como o menos adequado, sendo classificado como péssimo, pelos resultados do índice de desempenho.

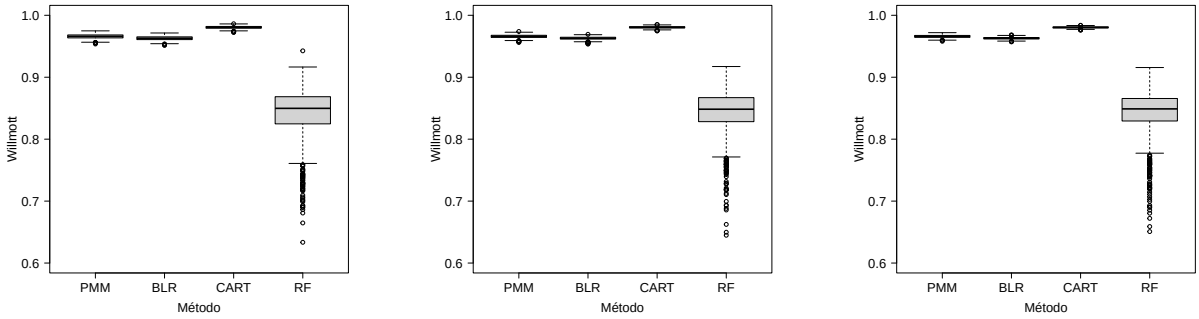
Os resultados obtidos por meio da medida de correlação de Pearson foram apresentados no Apêndice A. No Apêndice B apresentam-se a mesma sequência de *box plots*, mas agora considerando-se 5 iterações para aplicação dos métodos de imputação. Tais resultados não foram apresentados no texto

principal, devido a similaridade da variabilidade apresentada por cada um dos critérios para cada um dos métodos. A comparação dos métodos de acordo com as iterações foi explorada por meio dos resultados apresentados nas Figuras 4.8 à 4.15.



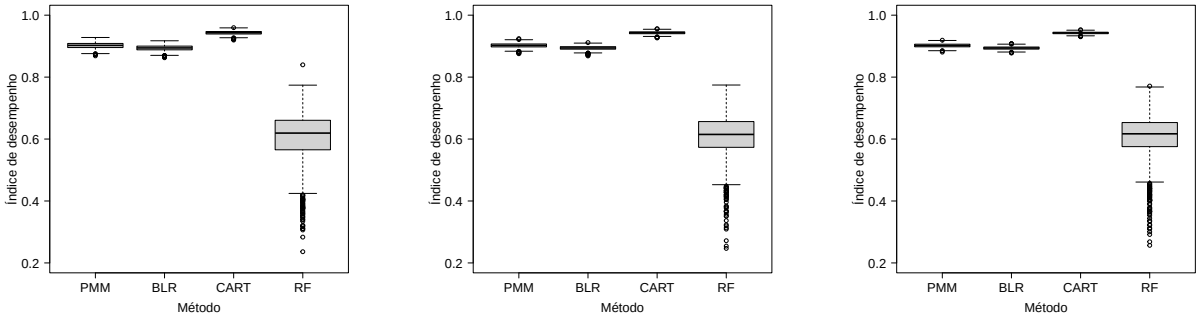
(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura 4.1. *Box plots* para o método de comparação RMSE, considerando diferentes métodos de imputação (com 10 iterações) e porcentagens de valores que foram imputados.



(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura 4.2. *Box plots* para o método de comparação Willmott, considerando diferentes métodos de imputação (com 10 iterações) e porcentagem de imputação.



(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura 4.3. *Box plots* para o método de comparação Índice de Performance, considerando diferentes métodos de imputação (com 10 iterações) e porcentagem de imputação.

A partir dos *box plots* pôde-se fazer uma comparação exploratória para identificar o método que retornou valores imputados mais próximos dos valores reais. Com o intuito de verificar se houve diferença

estatística entre os métodos, foi realizado um teste F feito pela análise de variância para detectar se houve diferença significativa entre os métodos de imputação utilizados, e o teste de comparação de médias de Tukey, para verificar quais métodos apresentavam médias que diferiam entre si.

O teste de Tukey é o mais utilizado para a comparações de médias de um experimento, e consiste em testar os contrastes, entre todos os pares possíveis de combinações de duas médias de tratamento (Banzatto e Kronka, 1992). Ele é baseado na diferença mínima significativa, que depende do valor da amplitude total estudentizada (valor tabelado), do quadrado médio do resíduo e do número de repetições com que a média foi calculada.

Devido a grande quantidade de observações disponíveis, foi necessário realizar a análise de variância e o teste de Tukey a partir de amostras, uma vez que esses testes são sensíveis a valores de repetições muito grandes, levando-os a apontar diferenças significativas entre os métodos para valores extremamente pequenos ($< 0,00001$). Foram selecionados aleatoriamente cinco valores obtidos por meio de cada critério, para cada um dos métodos, totalizando 20 valores em cada um dos cenários que foram testados. O valor 20 foi definido com base no número de observações que forneciam grau de liberdade para o resíduo ser maior do que 12 (Banzatto e Kronka, 1992). A fim de atribuir maior incerteza à análise, foram realizadas 1000 reamostragens aleatórias com reposição.

As conclusões dos testes foram apresentadas por meio de gráficos de barras que mostraram a quantidade de cada uma das letras atribuídas pelo teste de Tukey aos métodos em cada cenário, representando as diferenças significativas apontadas pelo teste. O teste de Tukey atribui letras diferentes para cada um dos métodos que apresenta resultados que diferem significativamente entre si, ao nível de significância de 5%, neste caso. Por exemplo, atribui a letra a para a maior média testada, b para a segunda maior média que se difere significativamente da primeira, seguindo assim até realizar todas as comparações por meio dos contrastes de médias duas a duas.

Na Figura 4.4, considerando as medidas obtidas pelo critério de comparação RMSE, observa-se que, a letra a foi atribuída ao método RF, que apresentou maiores valores para esse critério e que isso ocorre para todas as amostras utilizadas. Ao métodos PMM e BLR, para a maioria das amostras foram atribuídos a letra b, sendo que para BLR, isso ocorreu em 100% das amostras. Já para o método CART em cerca de 80% das amostras foi atribuída a letra c. Observa-se que para um número muito reduzido de amostras, foram apontadas diferenças significativas para os quatro métodos simultaneamente.

Assim, o teste aponta que houve diferença significativa entre o método CART e os demais métodos para mais de 80% das amostras consideradas, para todas intensidades de ausência de dados, 5%, 10% e 15%, ao nível de significância de 5%. O método RF mostrou-se diferente significativamente de todos os métodos, enquanto os métodos PMM e BLR não apresentaram diferença significativa entre si na maioria das amostras, também considerando nível de significância igual a 5%.

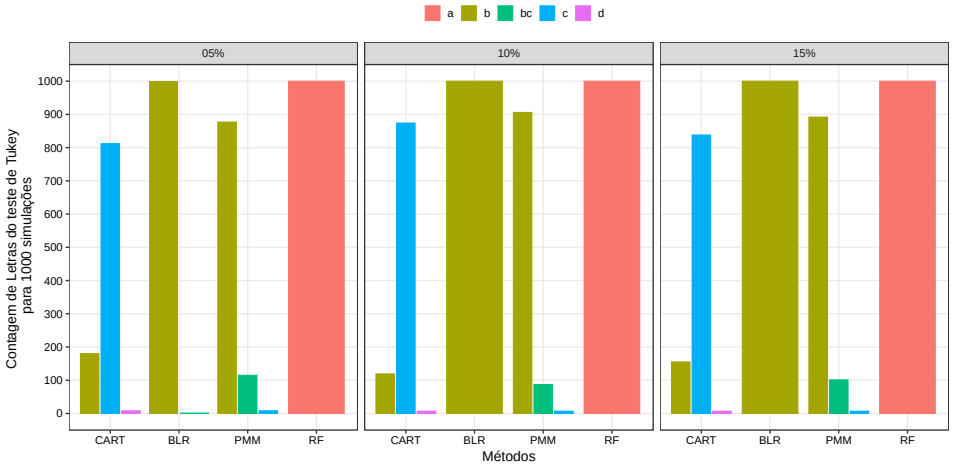


Figura 4.4. Teste de Tukey considerando o critério de comparação RMSE para os diferentes métodos de imputação com 10 iterações e porcentagem de valores imputados.

As Figuras 4.5, 4.6 e 4.7 apresentam os mesmos resultados, mas agora considerando as medidas obtidas para os critérios de Willmott, índice de desempenho e coeficiente de correlação de Pearson, respectivamente. Nestes casos o teste atribuiu a letra *a* em 100% das amostras pra o método CART, e na maioria dos cenários para os métodos BLR e PMM, pois foram os que apresentaram maiores valores para os referidos critérios. Todos os gráficos apontaram a mesma conclusão, de que em cerca de 70% das amostras, para as diferentes porcentagens de ausência de dados, os métodos CART, PMM e BLR não apresentaram diferença significativa entre si, mas suas médias mostraram-se diferentes quando comparadas com as médias obtidas pelo método RF.

Resultados semelhantes a estes foram obtidos considerando-se 5 iterações, os gráficos para comparação localizam-se no Apêndice C.

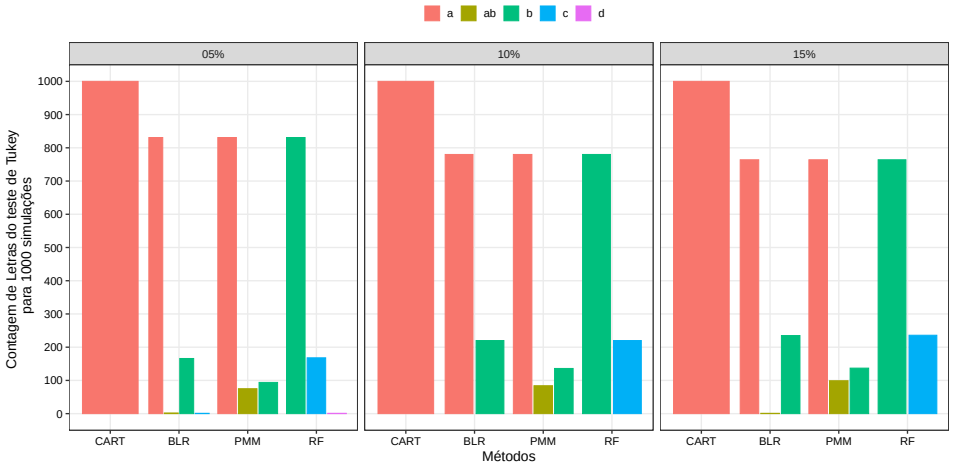


Figura 4.5. Teste de Tukey considerando o critério de comparação Índice de Willmott para os diferentes métodos de imputação com 10 iterações e porcentagem de valores imputados.

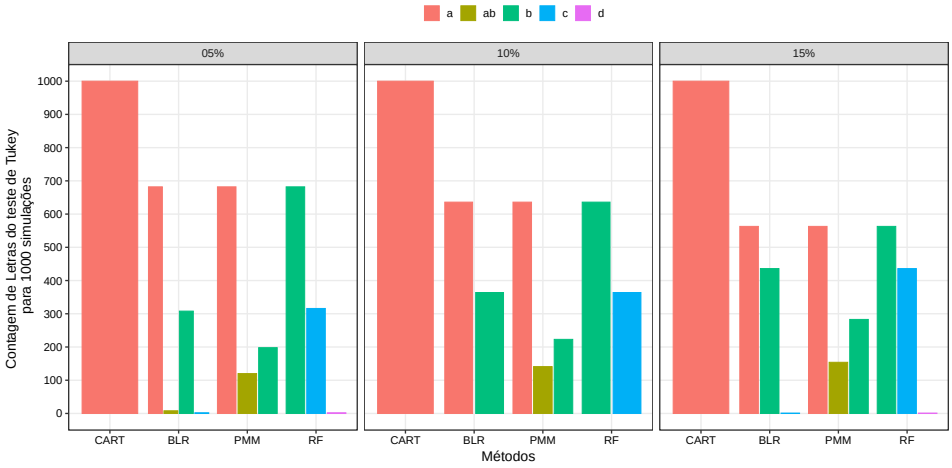


Figura 4.6. Teste de Tukey considerando o critério de comparação Índice de Performance para os diferentes métodos de imputação com 10 iterações e porcentagem de valores imputados.

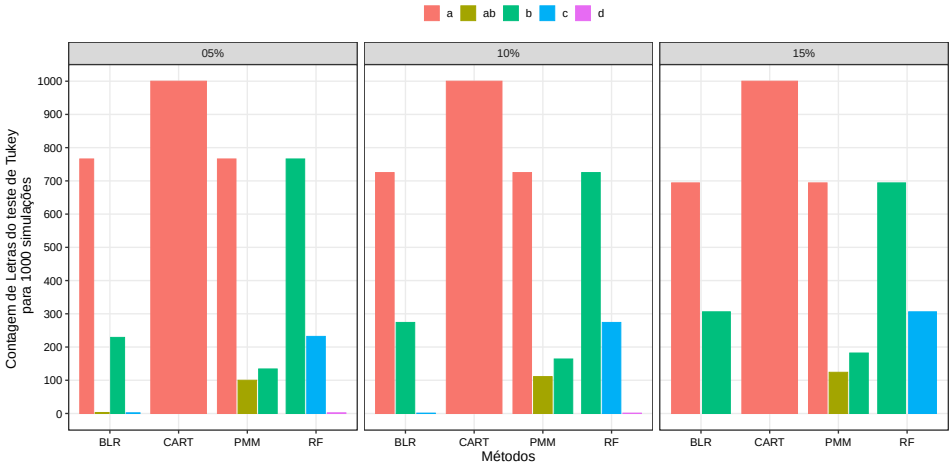
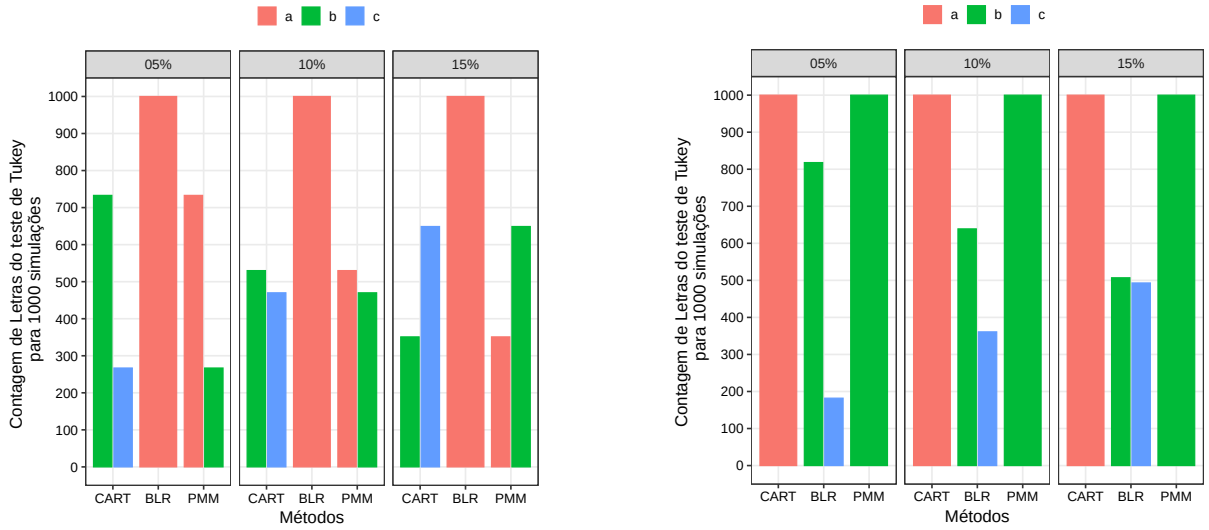


Figura 4.7. Teste de Tukey considerando o critério de comparação Coeficiente de Correlação de Pearson para os diferentes métodos de imputação com 10 iterações e porcentagem de valores imputados.

Também avaliou-se a comparação de médias dos diferentes métodos desconsiderando o método RF (Figura 4.8), uma vez que esse apresentou pior desempenho, e ainda precisava-se investigar se havia de fato diferença significativa para o método CART, quando comparado com os métodos PMM e BLR. A partir dessa análise pôde-se observar que, agora para os dois critérios testados (RMSE e índice de Willmott), houve diferença significativa entre os métodos CART e PMM, e CART e BLR, ao nível de significância de 5%, pois para todos os cenários apresentados foram atribuídas letras diferentes ao método CART, em relação aos demais.

Observou-se na Figura 4.8 (a), que para 100% das amostras o teste atribuiu letra *a* para BLR, já para os métodos PMM e CART, pode-se observar que em situações que foram atribuídas a letra *a* para PMM, CART recebeu *b*. Para algumas amostras é possível notar que houve diferença significativa entre os três métodos, ao nível de significância de 5%. Além disso, na Figura 4.8 (b), mostra-se que para todas as amostras, o método CART recebeu a letra *a*, sendo estatisticamente diferente dos demais, também ao nível de 5% de significância.



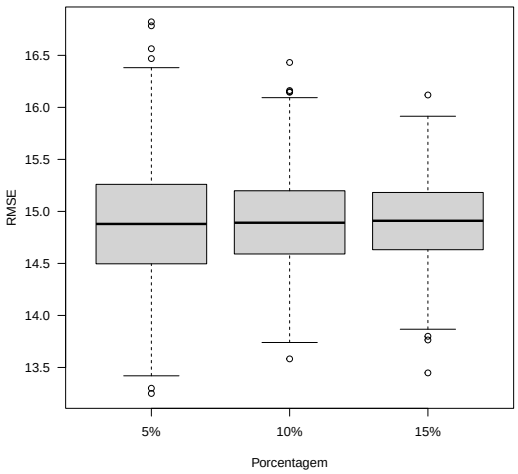
(a) RMSE **(b)** Índice de Willmott

Figura 4.8. Teste de Tukey considerando os critérios de comparação RMSE e Índice de Willmott para os métodos CART, PMM e BLR, com 10 iterações e porcentagem de valores imputados.

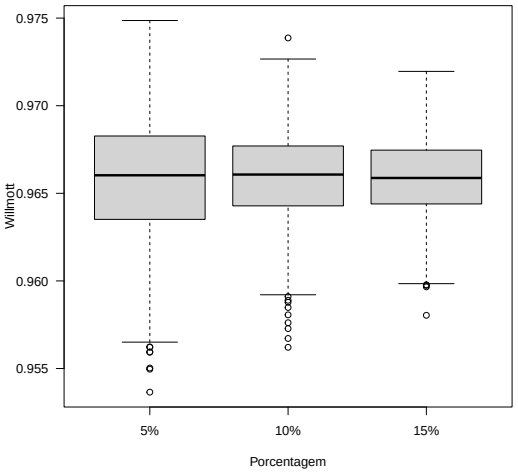
Vale ressaltar que a medida que se aumentou a intensidade de ausência de dados os resultados obtidos para o teste de médias apontaram diferenças significativas, ao nível de significância de 5%, para as médias dos métodos PMM e BLR. No cenário em que haviam 15% de valores faltantes, cerca de 50% das amostras apontaram resultados distintos entre os dois métodos citados (Figura 4.8).

As Figuras 4.9, 4.10, 4.11 e 4.12 apresentam os *box plots*, dos diferentes critérios de comparação para cada um dos métodos de imputação separadamente. Nestas figuras é possível observar o que acontece com as imputações sob diferentes concentrações de dados que devem ser imputados. Um comportamento muito interessante que pôde ser observado, para os métodos CART, PMM e BLR , foi que a medida que o número de valores perdidos aumentava, a variabilidade das medidas calculadas diminuía. Isto mostra que quando aumenta-se a amostra de valores ausentes a variabilidade diminui, esse fato também foi observado por da Cunha Júnior e Firmino (2022), que avaliou métodos de imputação para dados de precipitação para diferentes proporções de falhas, 10% e 40%, e por meio de diversos critérios de comparação, incluindo a RMSE.

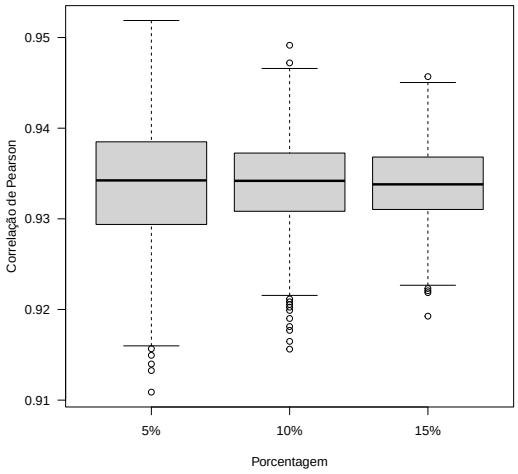
Esse comportamento não foi observado para o método RF, que apresentou uma discreta redução na variabilidade para as intensidades de valores ausentes, apenas em seu limite inferior.



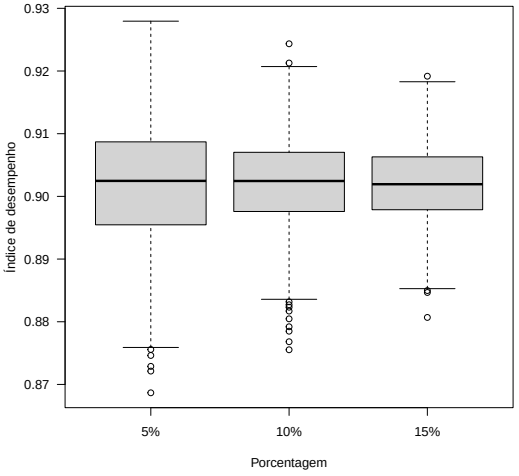
(a) RMSE



(b) Índice de Willmott

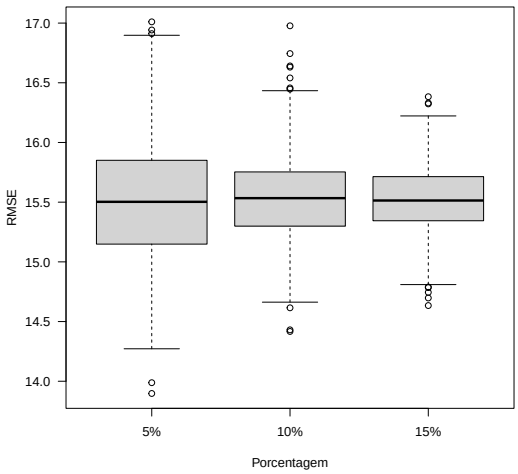


(c) Correlação de Pearson

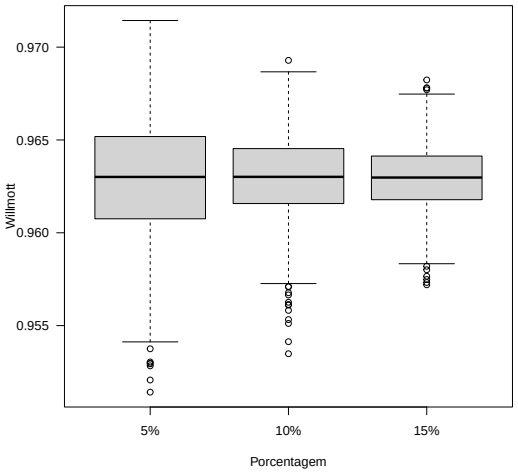


(d) Índice de Performance

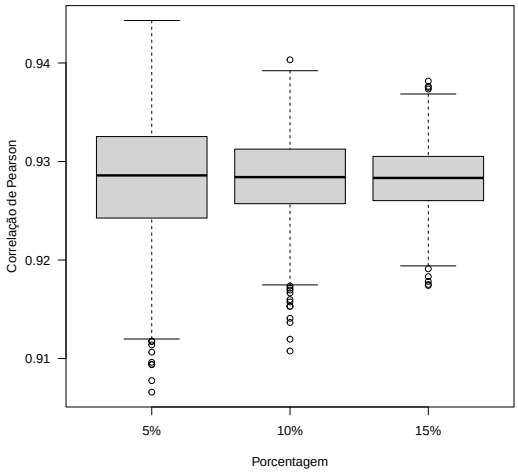
Figura 4.9. *Box plots* para o método de imputação PMM para 10 iterações.



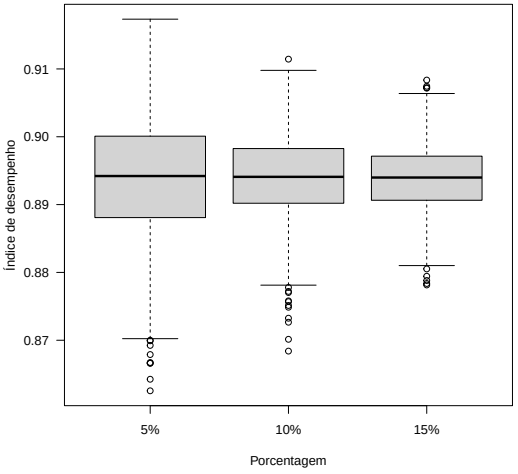
(a) RMSE



(b) Índice de Willmott

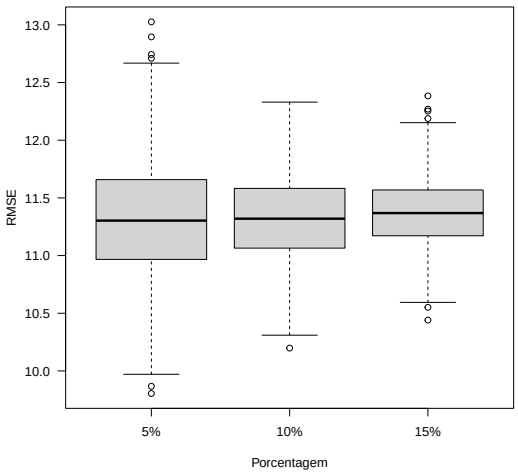


(c) Correlação de Pearson

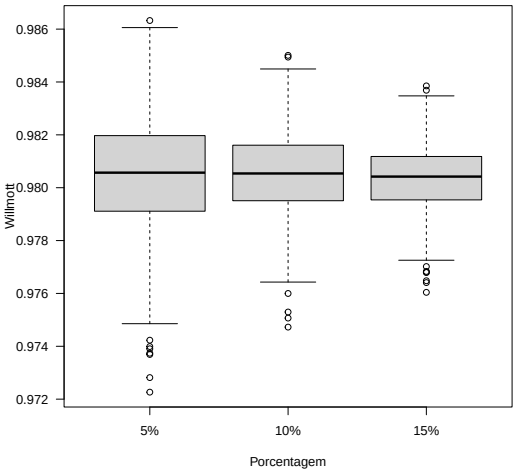


(d) Índice de Performance

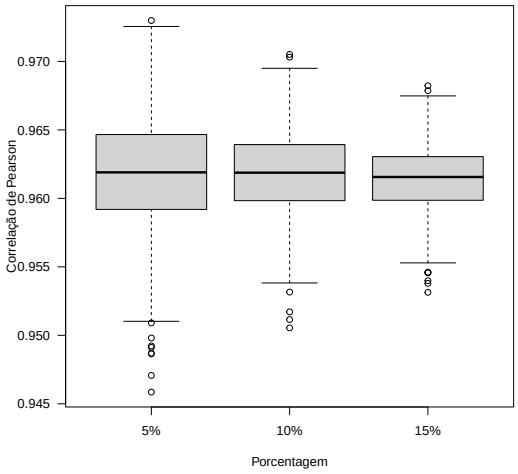
Figura 4.10. *Box plots* para o método de imputação BLR para 10 iterações.



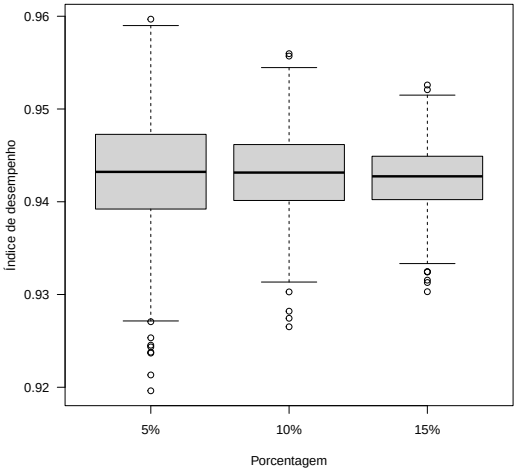
(a) RMSE



(b) Índice de Willmott

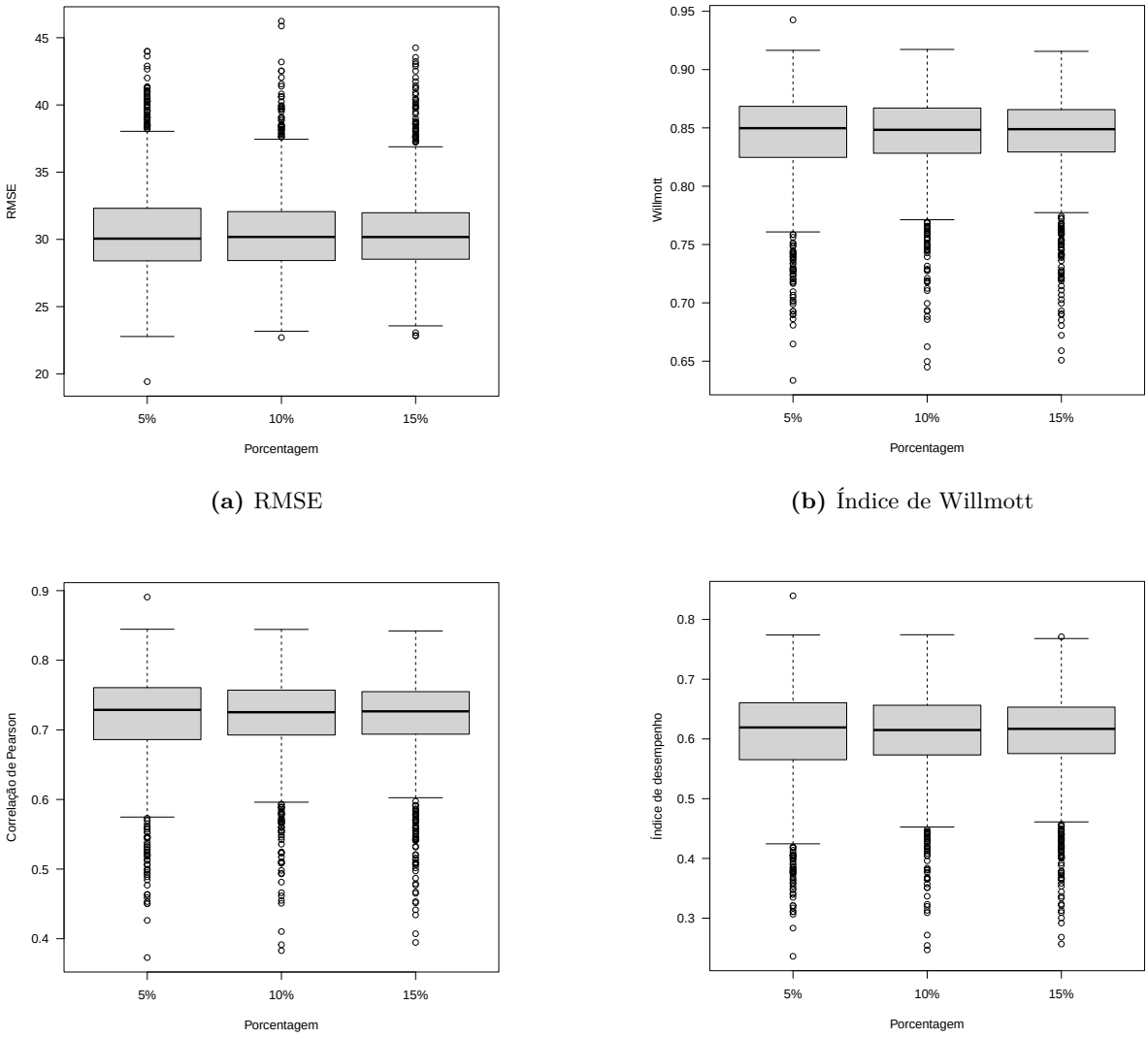


(c) Correlação de Pearson



(d) Índice de Performance

Figura 4.11. *Box plots* para o método de imputação CART para 10 iterações.



(a) RMSE

(b) Índice de Willmott

(c) Correlação de Pearson

(d) Índice de Performance

Figura 4.12. *Box plots* para o método de imputação RF para 10 iterações.

Os mesmos comportamentos foram encontrados com as análises realizadas a partir dos valores imputados considerando-se 5 iterações. As Figuras com seus respectivos comportamentos podem ser visualizados no Apêndice D.

As Figuras 4.13, 4.14, 4.15, 4.16, 4.17, 4.18, 4.19, 4.20, apresentam um resumo dos resultados, enfatizando o efeito das imputações quando se varia o valor de iterações. As figuras mostram que, quando comparados os *box plots* obtidos a partir das medidas de comparação dos métodos de imputação aplicados com 5 e 10 iterações, o comportamento para cada um dos métodos é equivalente entre si, para cada um dos cenários de porcentagem de ausência de dados. Estes gráficos foram produzidos apenas para os critérios da raiz do erro quadrático médio e do índice de Willmott, uma vez que os demais critérios apresentaram a mesma conclusão. Dessa forma, afirma-se que para este conjunto de dados e porcentagens de ausência de valores estudada a aplicação dos métodos para os dois valores de iterações foi similar. Resultado semelhante a esse foi encontrado por Cavalcanti (2021), que avaliou o uso de arquétipos para imputação de dados, comparando-os com métodos de imputação simples e múltipla. A autora também realizou imputações em diferentes proporções de dados ausentes e variando o número de iterações (3, 5 e 10 iterações), verificando que a variabilidade apresentada pelos *box plots* obtidos a partir

das medidas de RMSE e do índice Willmott mantiveram-se as mesmas.

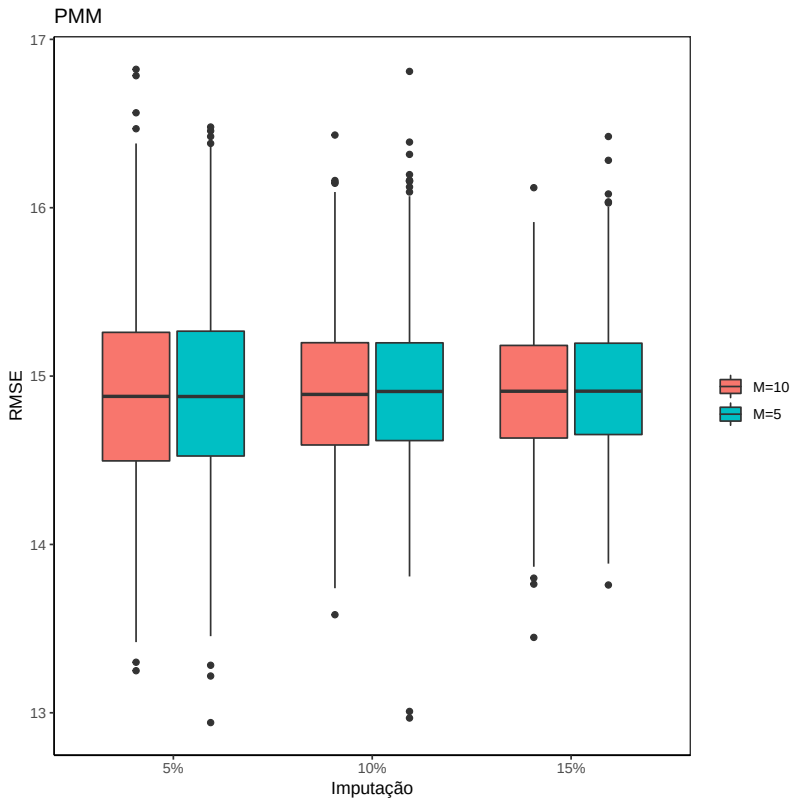


Figura 4.13. *Box plots* dos resultados do critério RMSE para o método de imputação PMM para 5 e 10 iterações.

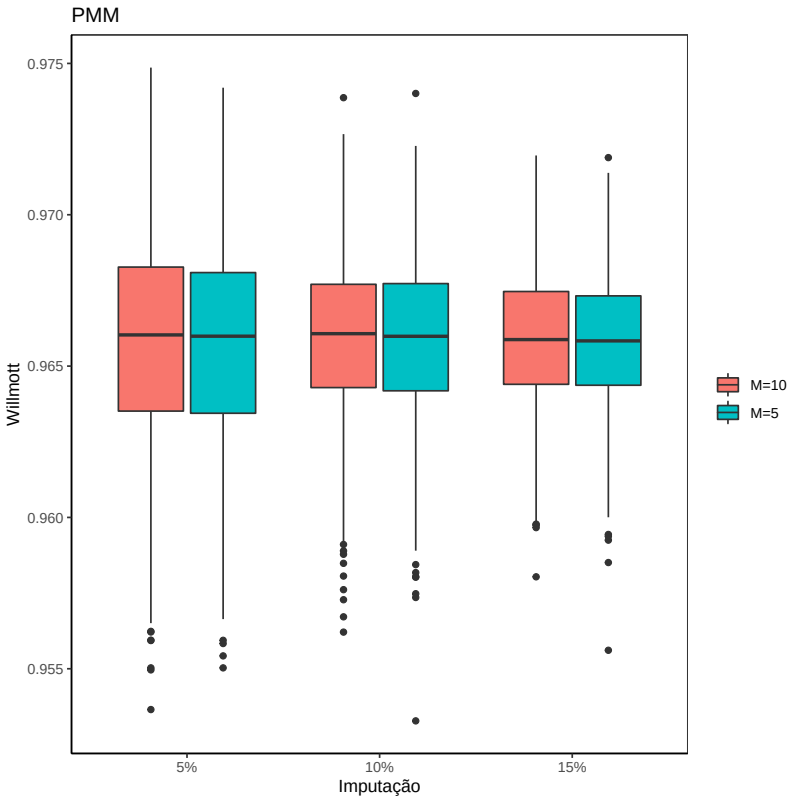


Figura 4.14. *Box plots* dos resultados do critério Índice de Willmott para o método de imputação PMM para 5 e 10 iterações.

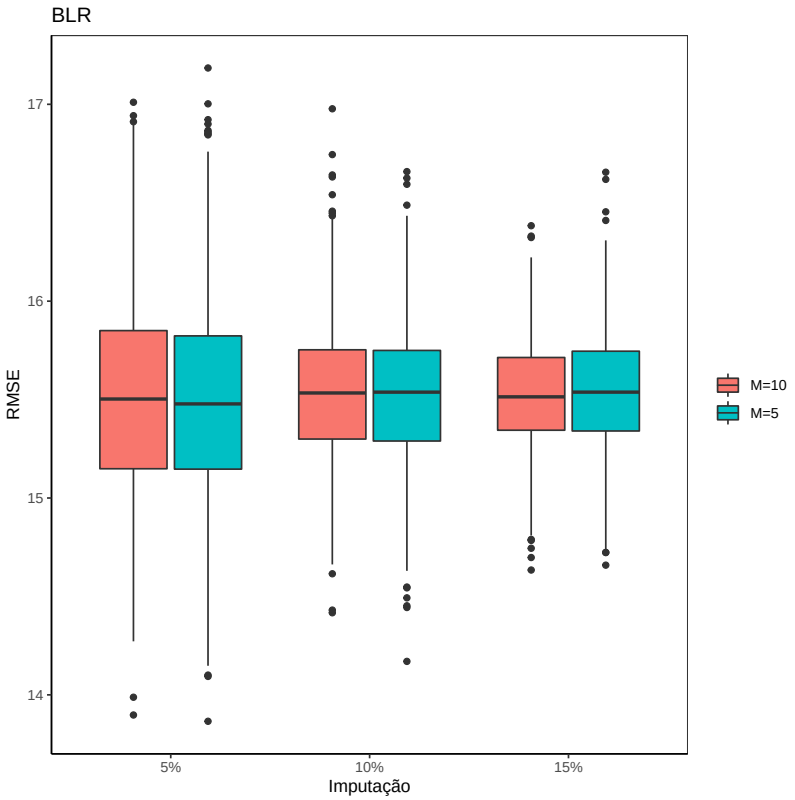


Figura 4.15. *Box plots* dos resultados do critério RMSE para o método de imputação BLR para 5 e 10 iterações.

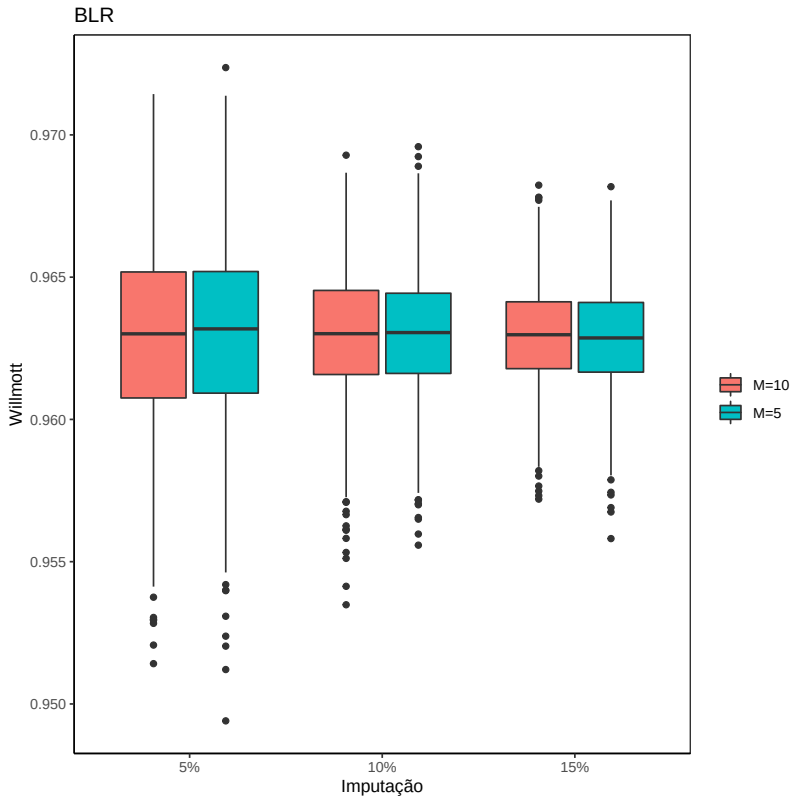


Figura 4.16. *Box plots* dos resultados do critério Índice de Willmott para o método de imputação BLR para 5 e 10 iterações.

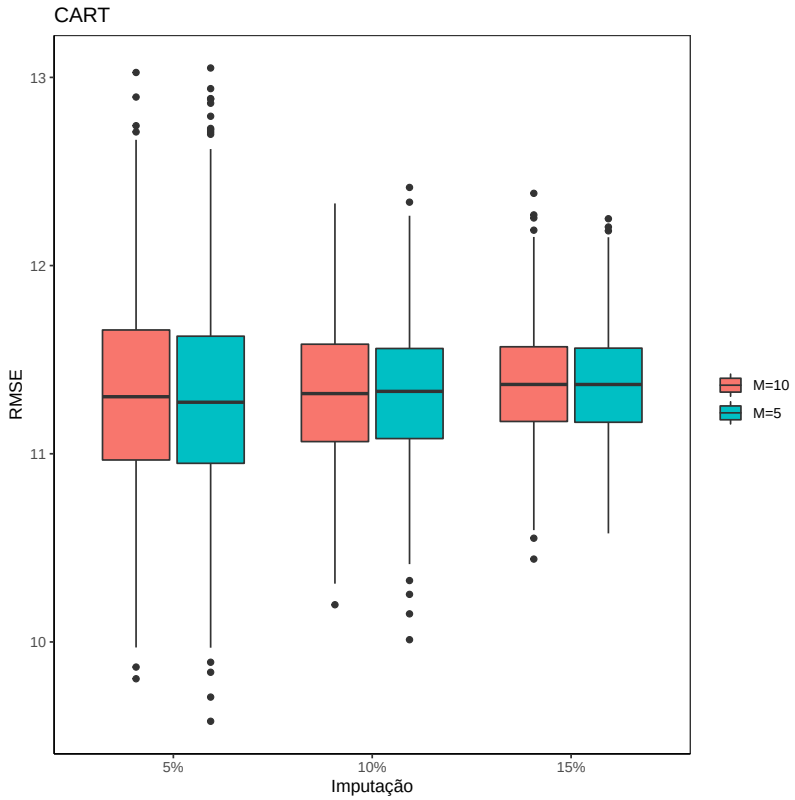


Figura 4.17. *Box plots* dos resultados do critério RMSE para o método de imputação CART para 5 e 10 iterações.

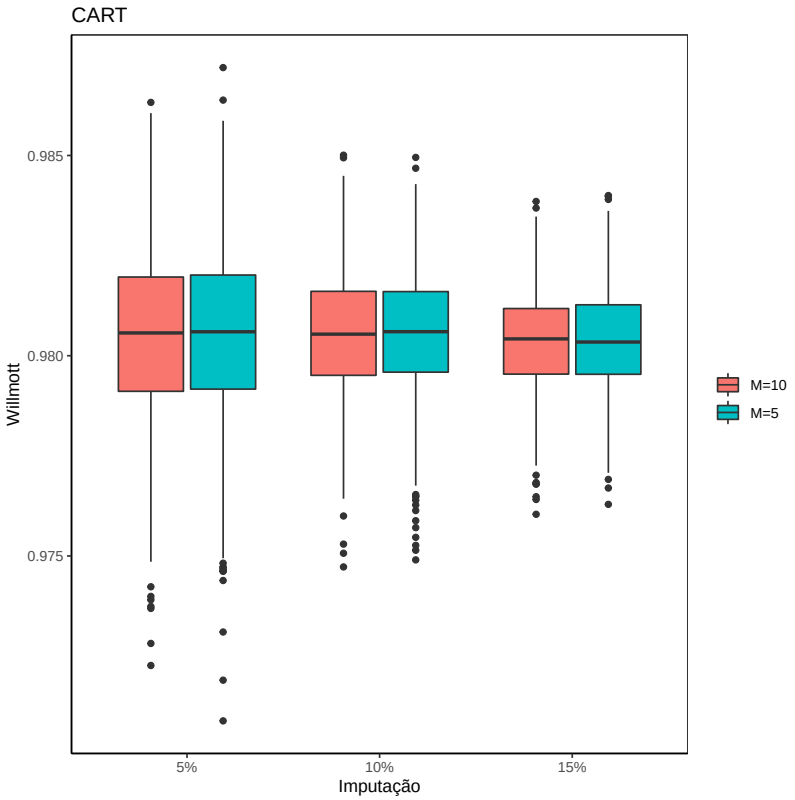


Figura 4.18. *Box plots* dos resultados do critério Índice de Willmott para o método de imputação CART para 5 e 10 iterações.

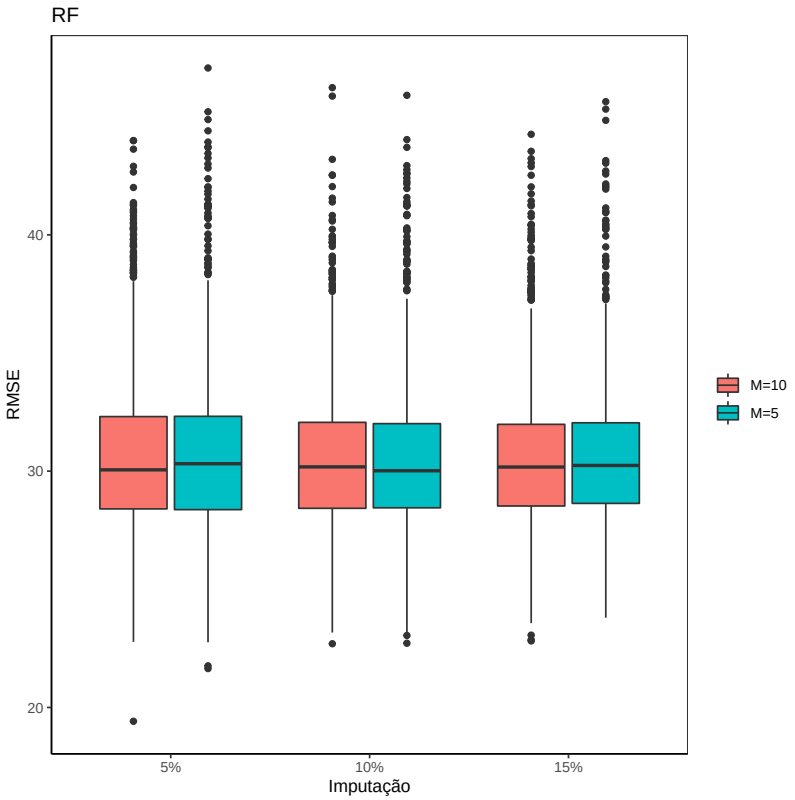


Figura 4.19. *Box plots* dos resultados do critério RMSE para o método de imputação RF para 5 e 10 iterações.

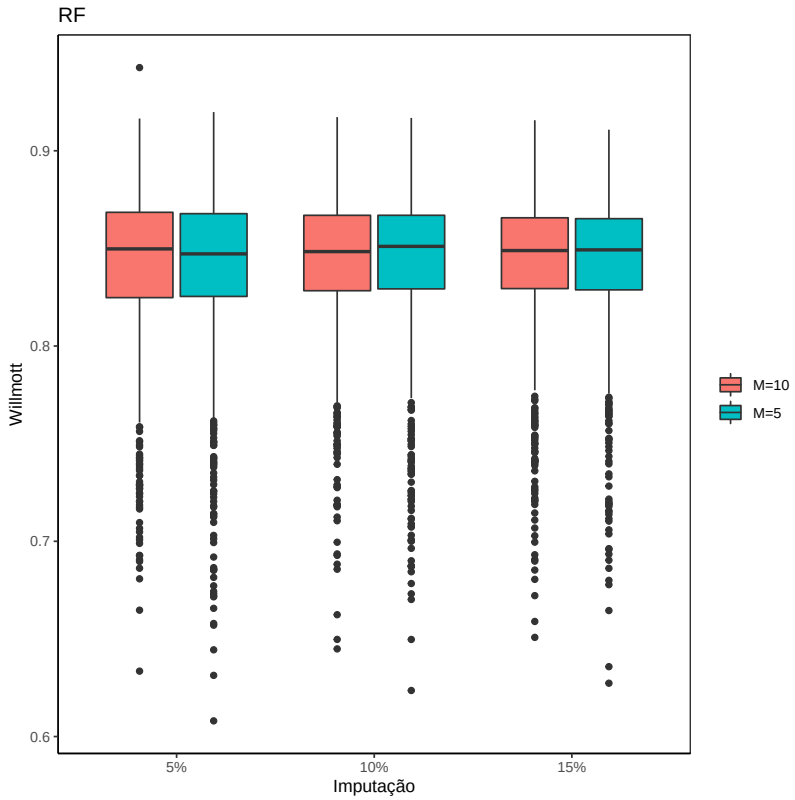


Figura 4.20. *Box plots* dos resultados do critério Índice de Willmott para o método de imputação RF para 5 e 10 iterações.

4.2 Aplicação dos métodos no conjunto de dados reais

A partir do estudo de simulação realizado foi possível observar que para dados desta natureza, com estrutura de dependência no tempo, o método de imputação RF não apresentou resultados satisfatórios. De acordo com os critérios de comparação estabelecidos, os métodos CART, PMM e BLR apresentaram bons resultados, destacando-se o método CART, que apresentou menores valores para RMSE e maiores índices de desempenho e de Willmott para todos os cenários.

Nesta etapa do trabalho, foram preservadas as observações ausentes do banco de dados originais e utilizados os métodos de imputação estudados para completar essas falhas. As imputações foram realizadas utilizando 5 iterações, uma vez que o estudo de simulação demonstrou que não houve melhora nos resultados quando consideradas 10 iterações.

Na Tabela 4.1 é possível observar as estatísticas descritivas calculadas para os valores imputados e para os dados reais sem os valores imputados, apresentadas na última linha da tabela. Percebe-se que o método PMM apresentou a menor média, 434,72, enquanto que a média do método RF foi de 450,09. Os métodos PMM, BLR e CART tiveram médias semelhantes, mas o método CART apresentou menor desvio padrão, entre eles, tendo seus valores imputados mais próximos da média. Todos os métodos tiveram valores de média e mediana menores do que aqueles obtidos para o conjunto de dados original, isso acontece pois como foi apresentado na Figura 3.6, uma parte considerável dos valores perdidos concentraram-se no início do estudo.

Tabela 4.1. Estatística descritiva para os valores imputados pelos diferentes métodos e para os valores reais com as falhas.

Métodos de imputação	Média	Mediana	Mínimo	Máximo	Desvio padrão	CV[1]
PMM ($n = 363$)	434,72	429,50	355,50	586,00	53,05	0,12
BLR ($n = 363$)	439,53	437,13	313,66	586,65	53,04	0,12
CART ($n = 363$)	442,49	436,00	361,00	571,50	44,45	0,10
RF ($n = 363$)	450,09	442,00	359,50	575,50	43,93	0,10
Dados reais ($n = 9604$)	463,41	461	355,5	590	41,01	0,08

[1] Coeficiente de variação

Com o objetivo de visualizar as observações imputadas por cada um dos métodos, a Figura 4.21 foi elaborada. Nesta figura é possível observar os perfis de peso de cada um dos animais, incluindo os valores imputados pelos diferentes métodos. A partir da tendência apresentada pelos dados observados, foi possível avaliar a qualidade dos valores imputados por cada um dos métodos, que foram representados por cores distintas indicadas na legenda.

De uma forma geral, observou-se que as imputações obtidas a partir do método PMM, para determinados animais atribuíram valores subestimados para o peso. Já os métodos BLR e RF atribuíram valores de peso para repor as falhas maiores do que os observados pela tendência do perfil. Os valores imputados pelo método CART foram identificados pela cor verde, e para a maioria dos animais os valores obtidos a partir deste método acompanharam a tendência de crescimento verificada.

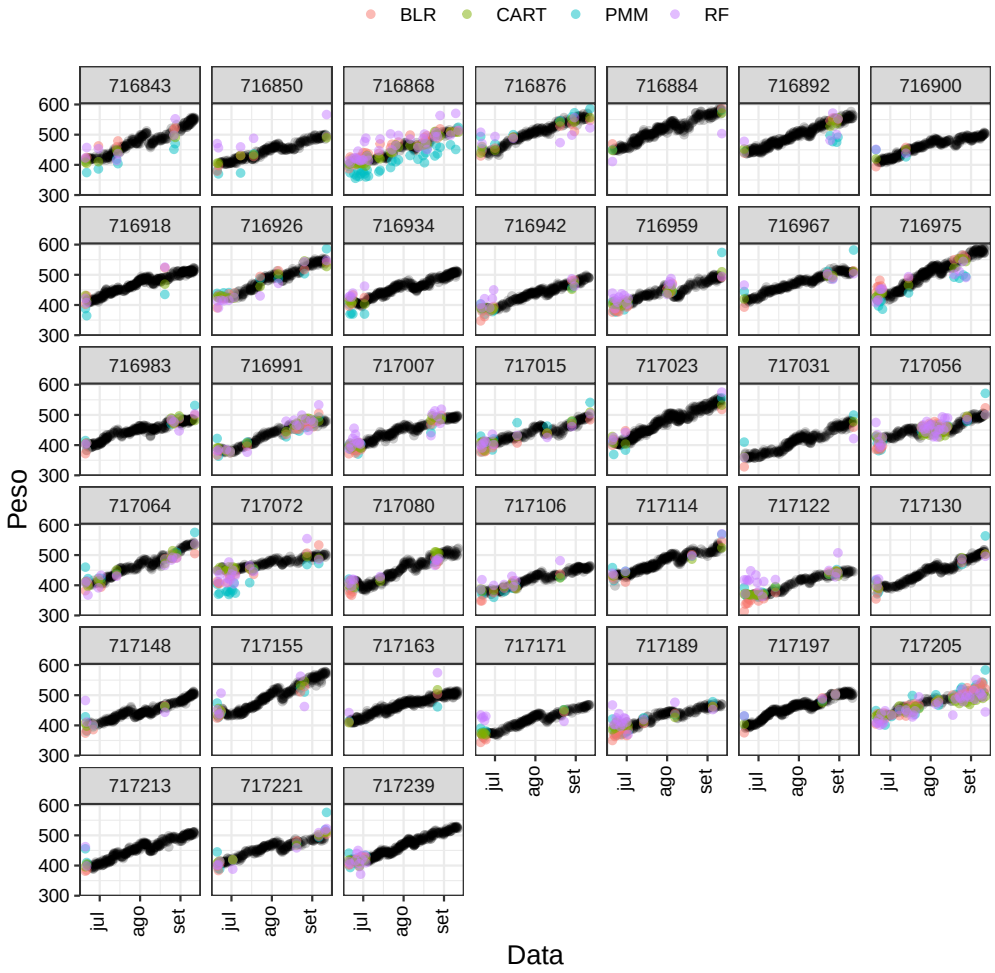


Figura 4.21. Gráfico de perfil do peso de cada animal ao longo do tempo incluindo valores imputados pelos diferentes métodos, com 5 iterações.

Para melhor clareza dos resultados, a Figura 4.22 foi construída, destacando o comportamento do peso e valores imputados de quatro animais que participaram do estudo, identificados pelos códigos 716868, 716975, 717056 e 717122. Esses animais foram escolhidos especialmente por apresentarem diferentes quantidades de valores perdidos e essas ausências ocorrerem em períodos diferentes do estudo.

Para o animal identificado por 716868, apresentou-se valores faltantes para a variável PESO em quase todo o período do experimento. Destacou-se que os valores obtidos pelo método PMM, em azul, foram menores do que os representados pelo curva de peso. O método BLR apresentou valores próximos ao esperado. O método CART apresentou valores que acompanharam a tendência. Já os valores atribuídos pelo método RF, foram os maiores, destoantes da tendência, nota-se que os valores imputados por meio do método no início do experimento foram satisfatórios, no entanto, em outros momentos não apresentou valores adequados.

Os animais representados por 716975 e 717122, apresentaram concentração de falhas no início do estudo e no final do mês de agosto, início do mês de setembro. Observou-se que o único método que apresentou bons resultados para os dois animais, mesmo apresentando padrão de falta de informação semelhantes foi o método CART.

Por fim, o animal com código 717056, continha valores perdidos no início do experimento, além de uma grande quantidade de valores perdidos durante o mês de agosto. Para este animal os valores imputados pelos métodos PMM e BLR foram menores do que a tendência, já o método RF apontou valores abaixo e acima da curva. As falhas do mês de agosto, foram substituídas adequadamente por todos os métodos.

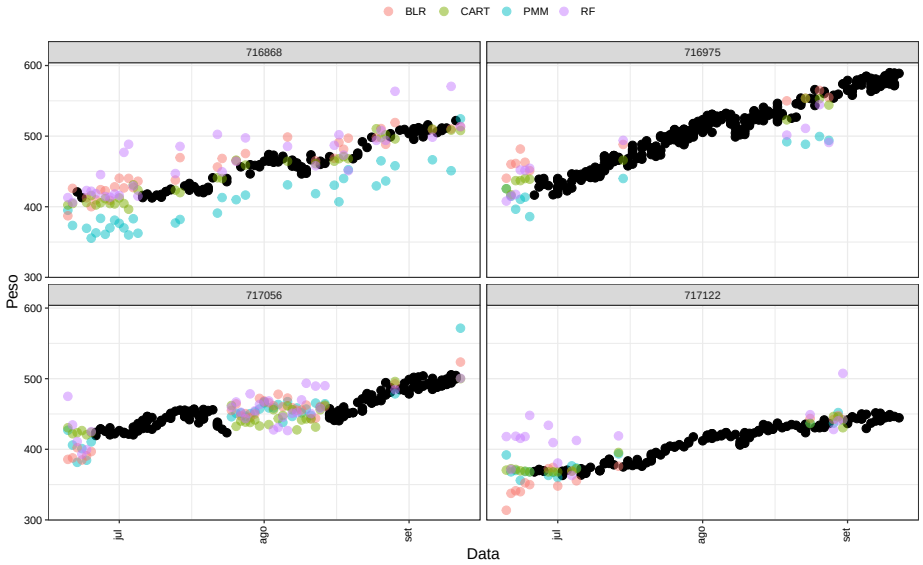


Figura 4.22. Gráfico de perfil do peso de animais selecionados ao longo do tempo incluindo valores imputados pelos diferentes métodos, com 5 iterações.

Foi possível observar que os valores imputados pelo método de árvores de classificação e regressão se destacou positivamente em relação aos demais quando consideradas as falhas originais que deveriam ser completadas, uma vez que, enquanto alguns métodos atribuíram valores que fugiram do perfil do ganho de peso do animal, o método CART preencheu as lacunas de forma satisfatória.

5 CONCLUSÃO

Neste trabalho foram avaliados quatro métodos de imputação, PMM, BLR, CART e RF, por meio de quatro critérios de comparação, RMSE, coeficiente de correlação de Pearson, coeficiente de acurácia de Willmott e coeficiente de desempenho. De acordo com o que foi apresentado concluiu-se que o método de imputação que apresentou maior eficiência foi o CART, pois no estudo de simulação foi o método que apresentou melhores resultados quando submetido aos critérios de comparação, sendo eles os menores valores para RMSE, os maiores valores para o índice de Willmott e de desempenho e maiores correlações também. Em contrapartida, o método RF não teve um bom desempenho para os dados utilizados neste trabalho, que possuem natureza longitudinal.

Ao aplicar os métodos aos dados originais novamente observou-se destaque para o método CART, tendo os valores obtidos por meio deste método preenchendo as lacunas deixadas pelas observações faltantes, encaixando-as na tendência apresentada pelos perfis de peso dos animais. Assim conclui-se que para os dados utilizados neste trabalho o método de imputação mais adequado foi o CART.

Para melhor avaliar os métodos CART e RF, pretende-se submetê-los à diferentes estruturas correlação de dados no tempo e no espaço e analisar seus respectivos desempenhos.

Esta tese, contribuiu para o enriquecimento da literatura da área de imputação, uma vez que, são comuns estudos destinados a dados climáticos, pluviométricos e de melhoramento genético de animais e plantas, mas não há muitos estudos envolvendo dados provenientes da pecuária de precisão, que tem sua base a partir dos dados que são coletados, dessa forma, quanto mais precisas forem as informações obtidas melhor será a tomada de decisão.

REFERÊNCIAS

Banzatto, D. A. e S. d. N. Kronka, 1992 *Experimentação agrícola*, volume 1. Jaboticabal: Funep.

BERGAMO, G., 2007 *Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação. 2007. 89p*, Tese (Doutorado)-Universidade de São Paulo, Piracicaba.

Breiman, L., 2001 Random forests. Machine learning **45**: 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, e C. J. Stone, 1984 Classification and regression trees. belmont, ca: Wadsworth. International Group **432**: 9.

Burgette, L. F. e J. P. Reiter, 2010 Multiple imputation for missing data via sequential regression trees. American journal of epidemiology **172**: 1070–1076.

Bussab, W. d. O. e P. A. Morettin, 2010 *Estatística básica*. Saraiva.

Camargo, A. d. e P. C. Sentelhas, 1997 Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no estado de são paulo, brasil. Revista Brasileira de agrometeorologia **5**: 89–97.

Cavalcanti, P. P., 2021 Archetypal analysis as an imputation method and multivariate data augmentation. University of São Paulo .

da Cunha Júnior, R. O. e P. R. A. Firmino, 2022 Simulação de valores ausentes em séries temporais de precipitação para avaliação de métodos de imputação. Revista Brasileira de Climatologia **30**: 691–714.

Dancey, C. P. e J. Reidy, 2017 *Statistics without maths for psychology*. Pearson London.

Donders, A. R. T., G. J. Van Der Heijden, T. Stijnen, e K. G. Moons, 2006 A gentle introduction to imputation of missing values. Journal of clinical epidemiology **59**: 1087–1091.

Doove, L. L., S. Van Buuren, e E. Dusseldorp, 2014 Recursive partitioning for missing data imputation in the presence of interaction effects. Computational statistics & data analysis **72**: 92–104.

Enders, C. K., 2010 *Applied missing data analysis*. The Guilford press.

Gasparetto, S. C., S. M. D. S. Piedade, L. R. Angelocci, e V. A. Ozaki, 2021 Comparação entre métodos de imputação de dados em diferentes intensidades amostrais na série de precipitação pluvial da esalq. Revista Brasileira de Climatologia **29**: 464–489.

Khan, S. I. e A. S. M. L. Hoque, 2020 Sice: an improved missing data imputation technique. Journal of big Data **7**: 1–21.

Laca, E. A., 2009 Precision livestock production: tools and concepts. Revista brasileira de zootecnia **38**: 123–132.

Little, R. J., 1988 Missing-data adjustments in large surveys. Journal of Business & Economic Statistics **6**: 287–296.

Little, R. J. e D. B. Rubin, 2014 *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Machado, R., R. Corrêa, e M. BERGAMASCHI, 2008 Escore de condição corporal e sua aplicação no manejo reprodutivo de ruminantes. São Carlos, SP: Embrapa Pecuária Sudeste. .

Malaguti, J. G. e L. G. de Faria, 2020 Comparação de métodos de imputação para estatística espacial .

Marwala, T., 2009 *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques*. IGI Global.

Mota, L. F. M., A. V. Pires, T. M. d. A. Mariz, J. d. S. Ribeiro, C. M. Bonafé, *et al.*, 2014 Estrutura corporal (frame size) e influencias no desempenho produtivo de bovinos de corte. UFVJM .

Murphy, D. J., M. D. Murphy, B. O'Brien, e M. O'Donovan, 2021 A review of precision technologies for optimising pasture measurement on irish grassland. Agriculture **11**: 600.

Nunes, L. N., M. M. Klück, e J. M. G. Fachel, 2009 Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Cadernos de Saúde Pública **25**: 268–278.

Pedersen, A. B., E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, e I. Petersen, 2017 Missing data and multiple imputation in clinical epidemiological research. Clinical epidemiology pp. 157–166.

Potter, K., H. Hagen, A. Kerren, e P. Dannenmann, 2006 Methods for presenting statistical information: The box plot. In *VLUDS*, pp. 97–106.

R Core Team, 2021 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. B., 1976 Inference and missing data. Biometrika **63**: 581–592.

Ruchay, A., V. Kober, K. Dorofeev, V. Kolpakov, A. Gladkov, e H. Guo, 2022 Live weight prediction of cattle based on deep regression of rgb-d images. Agriculture **12**: 1794.

Shah, A. D., J. W. Bartlett, J. Carpenter, O. Nicholas, e H. Hemingway, 2014 Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. American journal of epidemiology **179**: 764–774.

Silva, M. J. C. d., 2012 *Imputação múltipla: comparação e eficiência em experimentos multiambientais*, Universidade de São Paulo.

Simitzis, P., C. Tzanidakis, O. Tzamaloukas, e E. Sossidou, 2021 Contribution of precision livestock farming systems to the improvement of welfare status and productivity of dairy animals. Dairy **3**: 12–28.

Tzanidakis, C., O. Tzamaloukas, P. Simitzis, e P. Panagakis, 2023 Precision livestock farming applications (plf) for grazing animals. Agriculture **13**: 288.

Van Buuren, S., 2018 *Flexible imputation of missing data*. Chapman and Hall/CRC.

van Buuren, S., 2022 *mice: Multivariate Imputation by Chained Equations*. R package version 3.15.0.

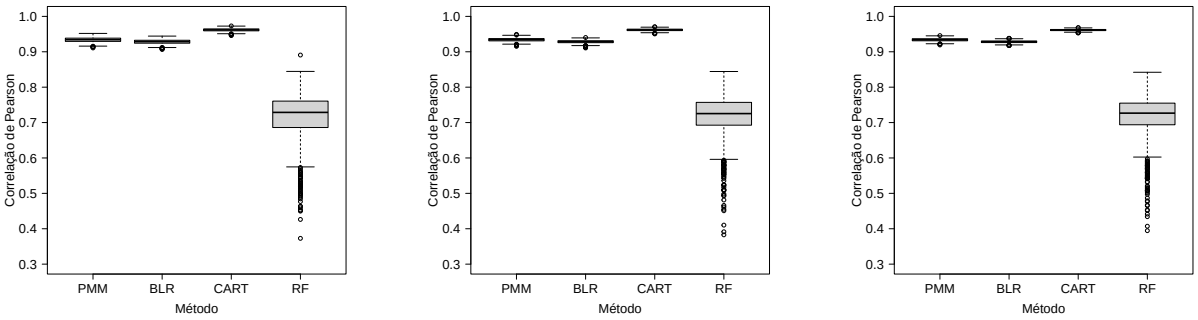
Vink, G., L. E. Frank, J. Pannekoek, e S. Van Buuren, 2014 Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica **68**: 61–90.

Webber, H., V. Heyd, M. Horton, M. Bell, W. Matthews, e A. Chadburn, 2019 Precision farming and archaeology. Archaeological and Anthropological Sciences **11**: 727–734.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'donnell, e C. M. Rowe, 1985 Statistics for the evaluation and comparison of models. Journal of Geophysical Research: Oceans **90**: 8995–9005.

APÊNDICES

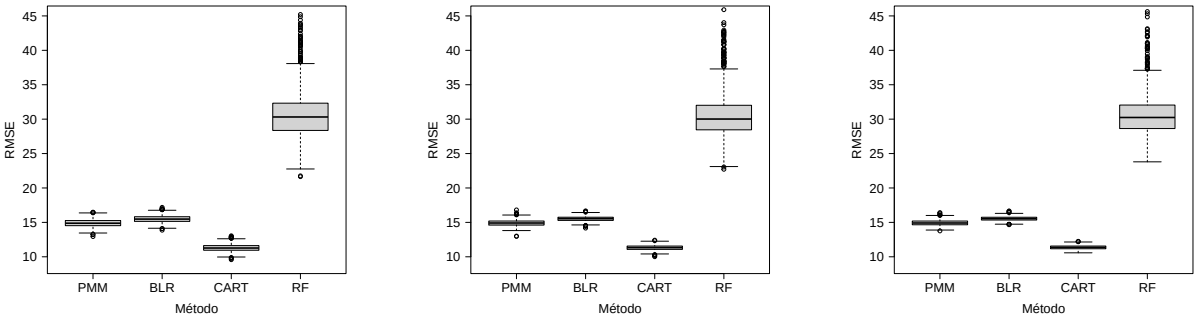
APÊNDICE A - Gráficos para o método coeficiente de correlação de Pearson



(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

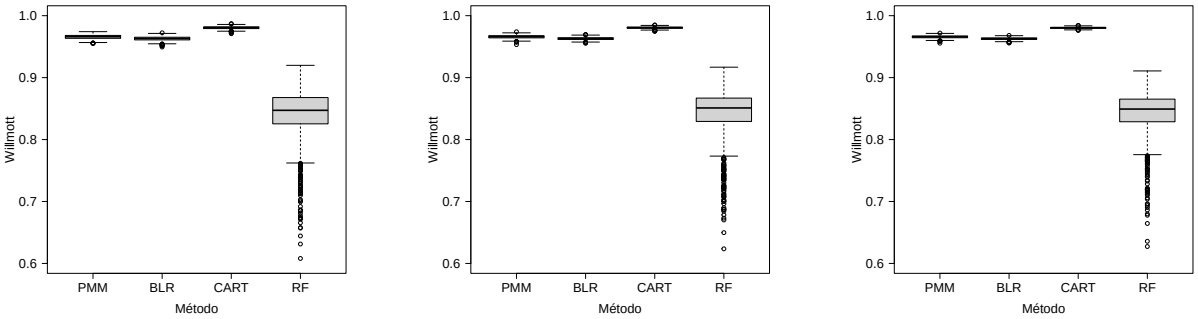
Figura A.1. *Box plots* para o método de comparação coeficiente de correlação de Pearson, considerando diferentes métodos de imputação (com 10 iterações) e porcentagens de valores que foram imputados.

APÊNDICE B - *Box plots* dos critérios de comparação considerando todos os métodos de imputação avaliados com 5 iterações



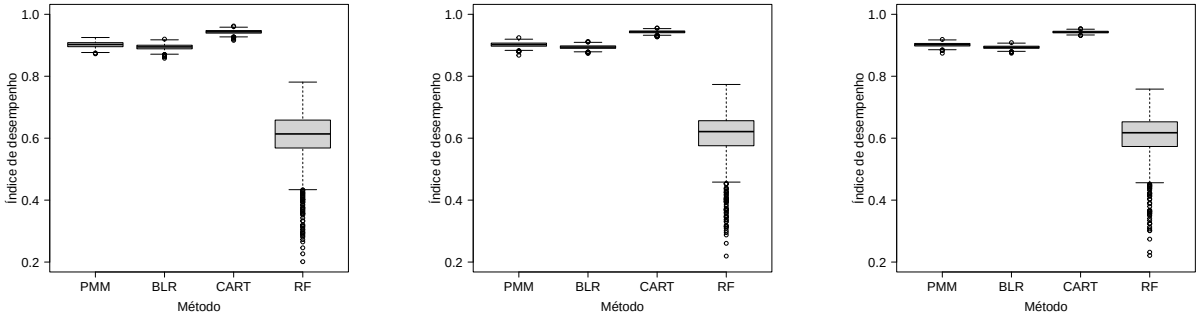
(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura B.2. *Box plots* para o método de comparação RMSE, considerando diferentes métodos de imputação (com 5 iterações) e porcentagem de imputação.



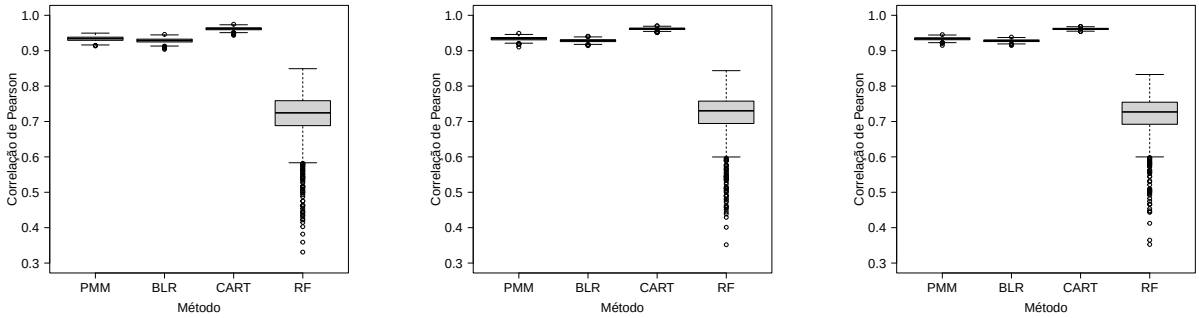
(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura B.3. *Box plots* para o método de comparação Willmott, considerando diferentes métodos de imputação (com 5 iterações) e porcentagem de imputação.



(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura B.4. *Box plots* para o método de comparação Índice de desempenho, considerando diferentes métodos de imputação (com 5 iterações) e porcentagem de imputação.



(a) Imputação de 5% **(b)** Imputação de 10% **(c)** Imputação de 15%

Figura B.5. *Box plots* para o método de comparação coeficiente de correlação de Pearson, considerando diferentes métodos de imputação (com 5 iterações) e porcentagens de valores que foram imputados.

APÊNDICE C - Histogramas com os resultados de contagem das letras fornecidas pelo teste de Tukey a partir dos critérios aplicados neste estudo (com 5 iterações)

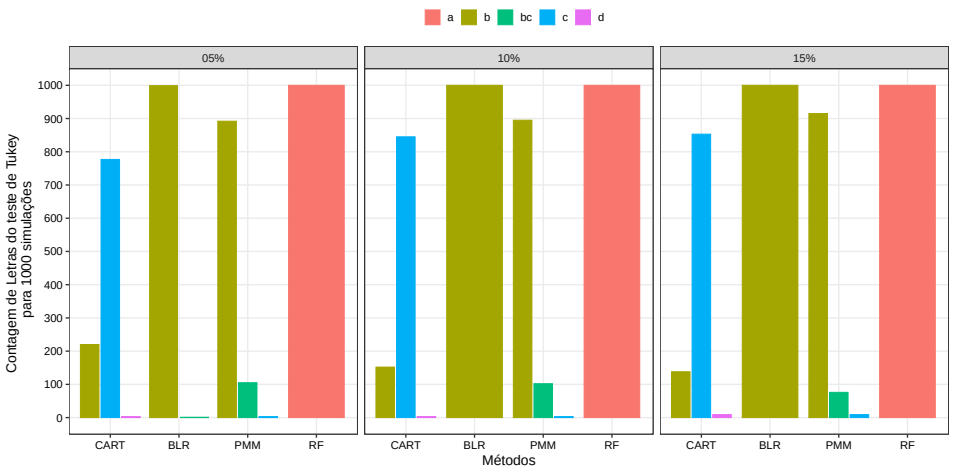


Figura C.6. Teste de Tukey considerando o critério de comparação RMSE para os diferentes métodos de imputação com 5 iterações.

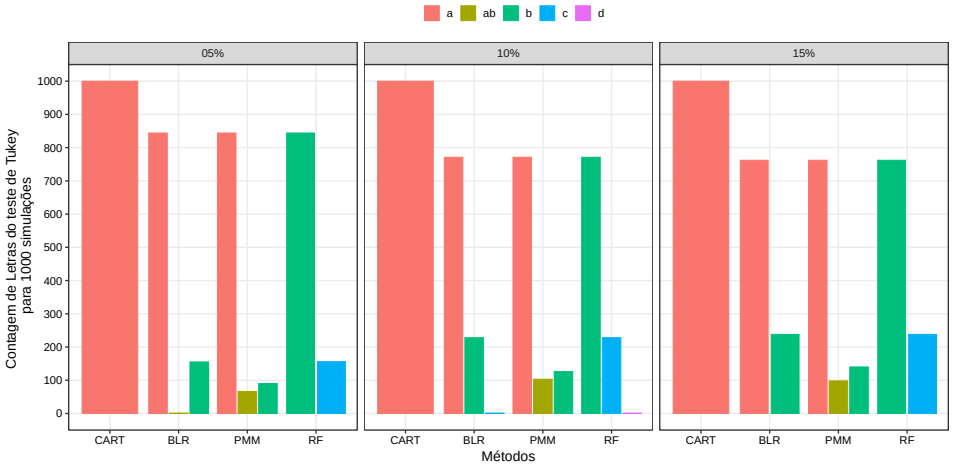


Figura C.7. Teste de Tukey considerando o critério de comparação Índice de Willmott para os diferentes métodos de imputação com 5 iterações.

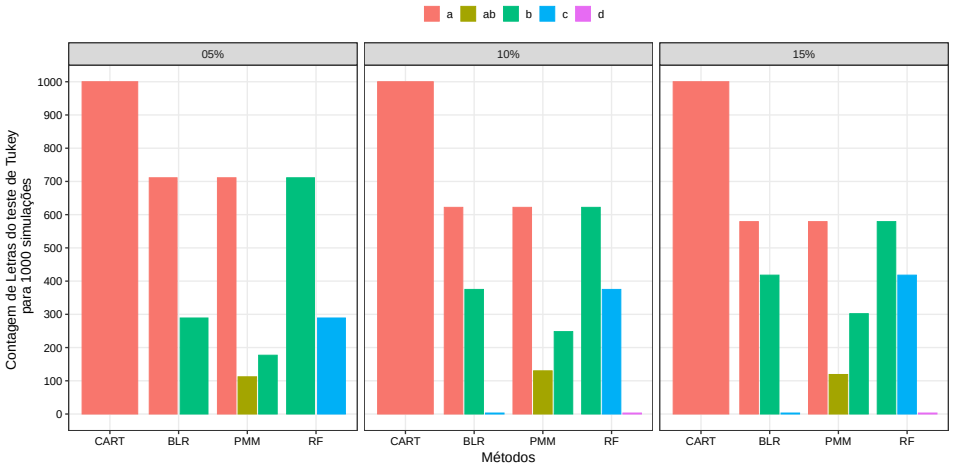


Figura C.8. Teste de Tukey considerando o critério de comparação Índice de Performance para os diferentes métodos de imputação com 5 iterações.

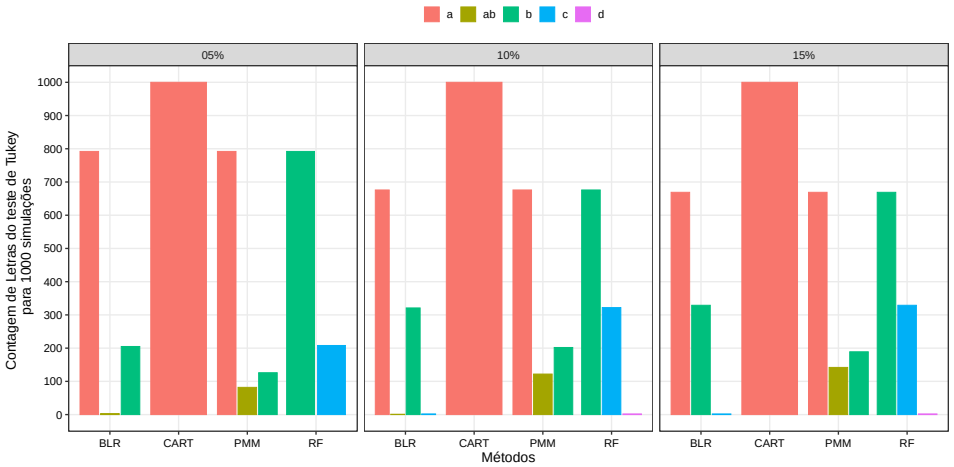
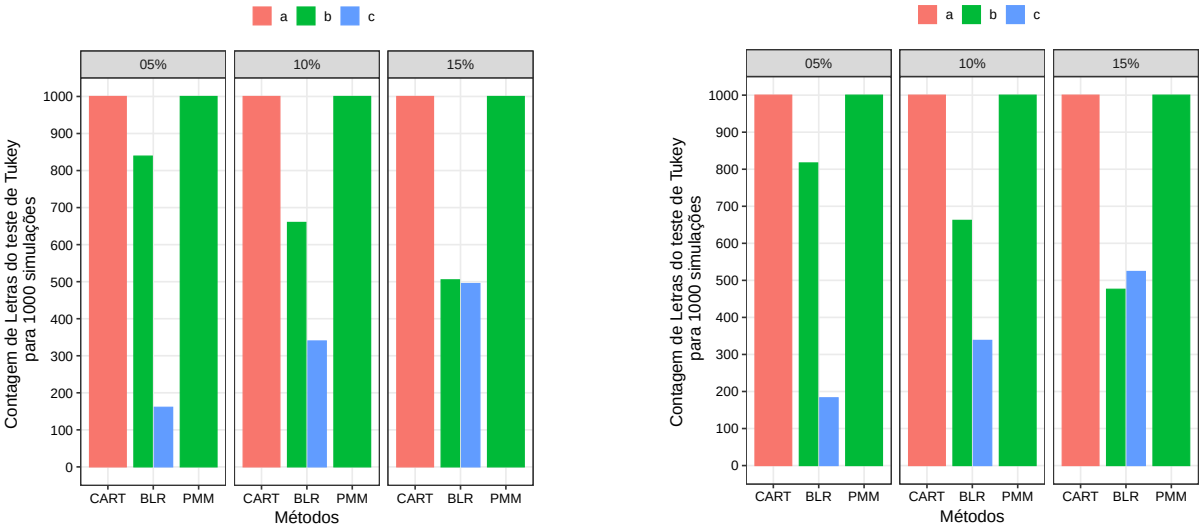


Figura C.9. Teste de Tukey considerando o critério de comparação Coeficiente de Correlação de Pearson para os diferentes métodos de imputação com 5 iterações.

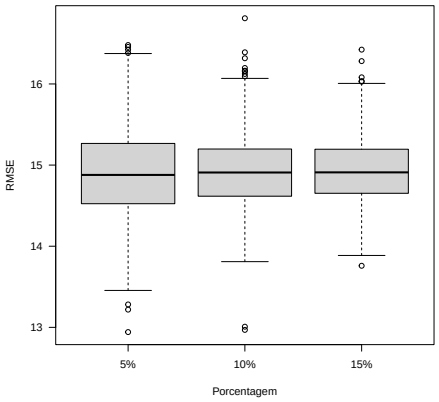


(a) RMSE

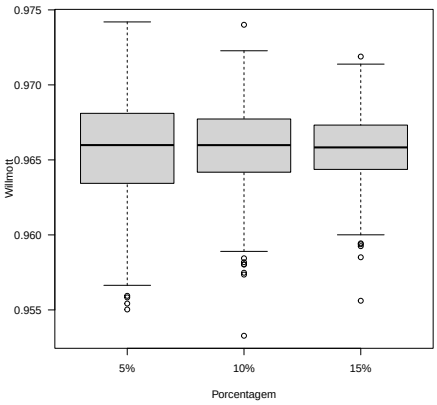
(b) Índice de Willmott

Figura C.10. Teste de Tukey considerando os critérios de comparação RMSE e Índice de Willmott para os métodos CART, PMM e BLR, com 5 iterações.

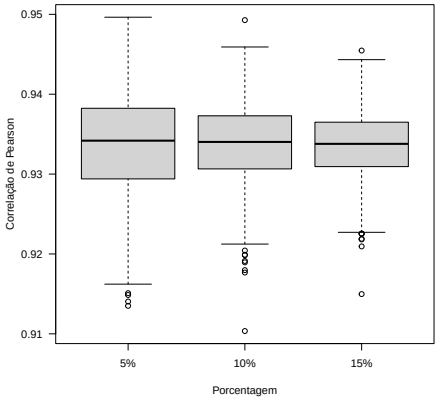
APÊNDICE D- Box plot para todos os métodos de imputação (com 5 iterações) avaliando os diferentes cenários de valores faltantes



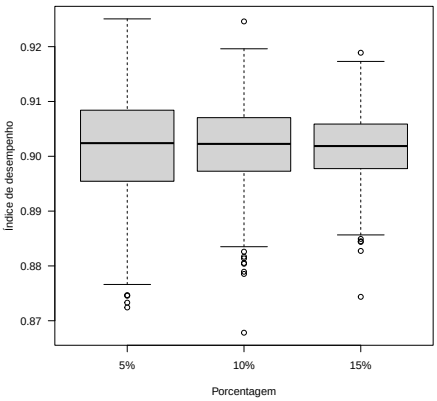
(a) RMSE



(b) Índice de Willmott

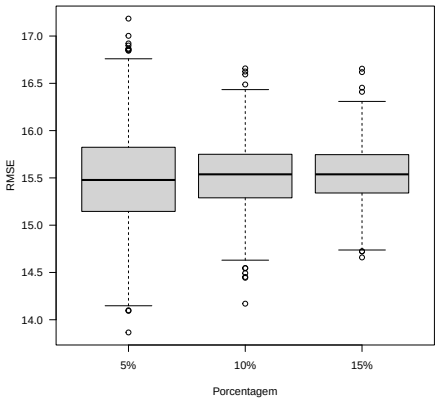


(c) Correlação de Pearson

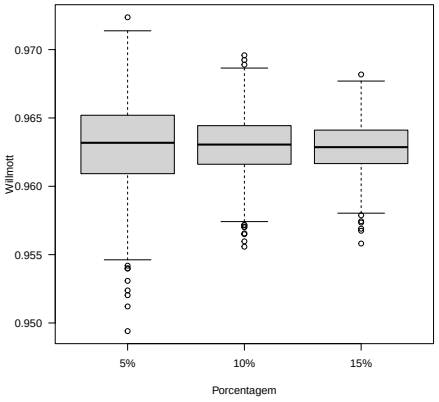


(d) Índice de Performance

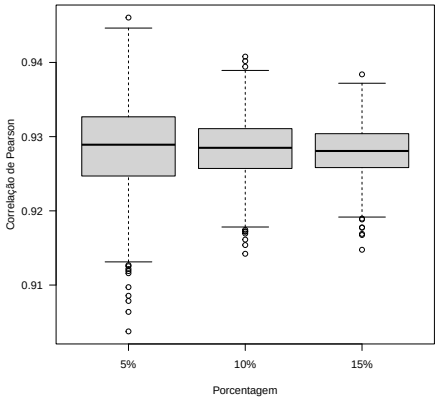
Figura D.11. *Box plots* para o método de imputação PMM para 5 iterações.



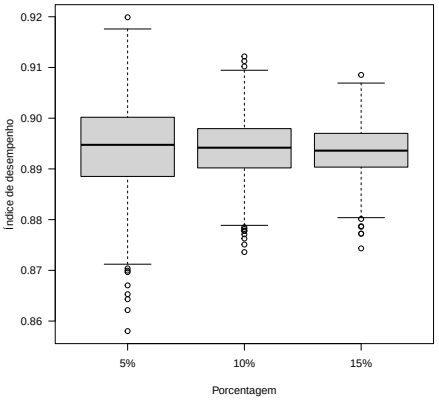
(a) RMSE



(b) Índice de Willmott

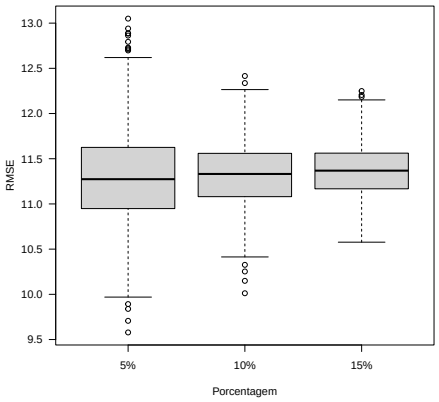


(c) Correlação de Pearson

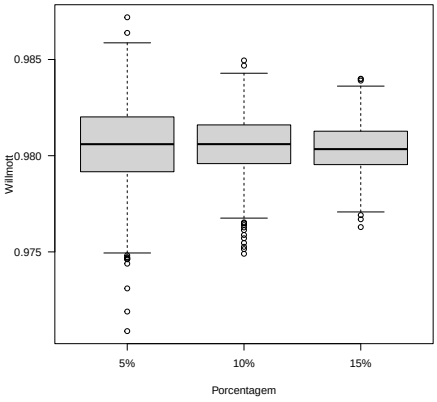


(d) Índice de Performance

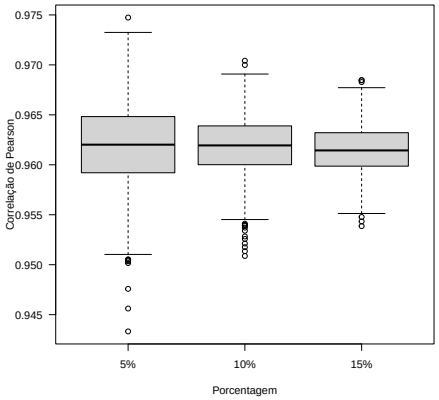
Figura D.12. *Box plots* para o método de imputação BLR para 5 iterações.



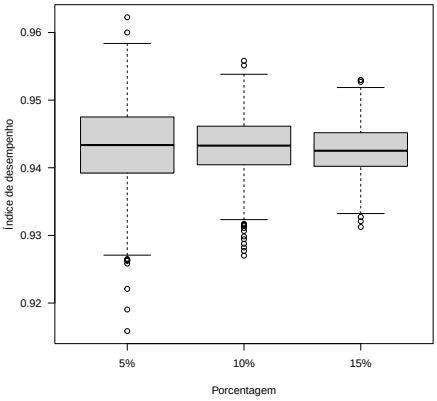
(a) RMSE



(b) Índice de Willmott

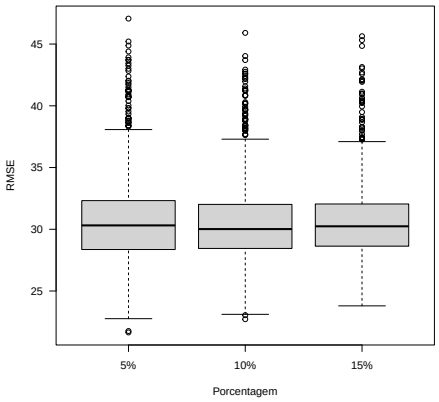


(c) Correlação de Pearson

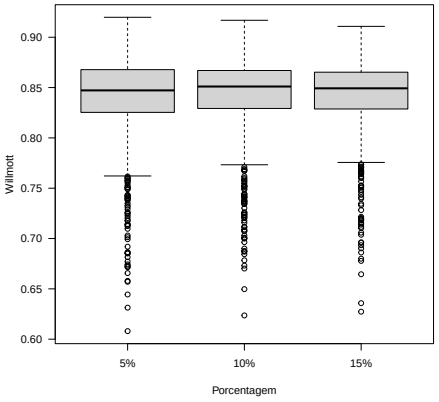


(d) Índice de Performance

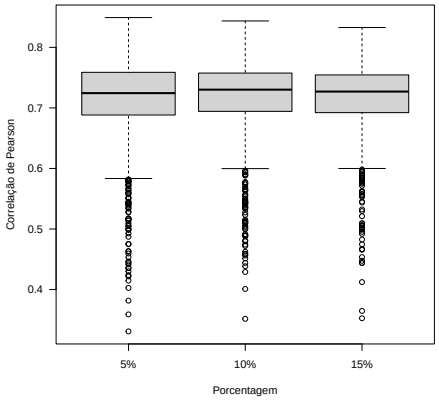
Figura D.13. *Box plots* para o método de imputação CART para 5 iterações.



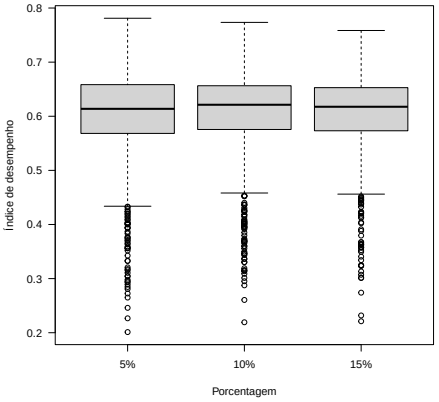
(a) RMSE



(b) Índice de Willmott



(c) Correlação de Pearson



(d) Índice de Performance

Figura D.14. *Box plots* para o método de imputação RF para 5 iterações.