unesp

UNIVERSIDADE ESTADUAL PAULISTA

Aplicações do Método de Imputação Múltipla MICE em Dados Longitudinais

Thais Vieira dos Santos

INSTITUTO DE GEOCIÊNCIAS E CIÊNCIAS EXATAS

RIO CLARO

2026

Universidade Estadual Paulista
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Thais Vieira dos Santos

Aplicações do Método de Imputação Múltipla MICE em Dados Longitudinais

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Matemática.

Orientadora
Profa. Dra. Selene Maria Coelho Loibel

Rio Claro/SP
2026

D724a

dos Santos, Thais Vieira
Aplicações do Método de Imputação Múltipla MICE em Dados Longitudinais / Thais Vieira dos Santos. -- Rio Claro, 2026
80 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientadora: Selene Maria Coelho Loibel

1. Multiple imputation (Statistics). 2. Estudos longitudinais. 3. Ausencia de dados (Estatistica). I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Impacto potencial desta pesquisa

Este trabalho apresenta aplicações em estudos longitudinais da Imputação Múltipla com o método intitulado *Multivariate Imputation by Chained Equations* (MICE). Na atual literatura há uma demanda dessas aplicações. Assim, esse estudo pode contribuir uma vez que apresenta soluções para o problema de dados longitudinais incompletos.

Potential impact of this research

This essay presents applications in longitudinal studies of Multiple Imputation with the method entitled Multivariate Imputation by Chained Equations (MICE). In the current literature, there is a demand for these applications. Furthermore, this work can contribute since it presents solutions to the problem of incomplete longitudinal data.

Universidade Estadual Paulista
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Thais Vieira dos Santos

Aplicações do Método de Imputação Múltipla MICE em Dados Longitudinais

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Matemática.

Comissão Examinadora

Prof. Dra. Selene Maria Coelho Loibel
IGCE/UNESP/Rio Claro (SP)

Prof. Dr. José Silvio Govone
IGCE/UNESP/Rio Claro (SP)

Prof. Dr. Marcos Tadeu dos Santos
SEDUC/Rio Claro (SP)

Conceito: Aprovado

Rio Claro (SP), 17 de dezembro de 2025

À minha família.

Agradecimentos

Durante a graduação, a possibilidade de me tornar mestre parecia distante. Vivia um dia de cada vez, esforçando-me para superar as dificuldades que me faziam desacreditar do meu potencial matemático. Encontrei o equilíbrio ao me realizar na educação básica e descobrir a plenitude no ato de ensinar. Com esse propósito, iniciei o mestrado e hoje realizo o sonho de contribuir com a formação de estudantes do ensino básico.

Para a conquista deste título, contei com o amparo divino e o apoio fundamental de pessoas a quem dedico minha profunda gratidão.

Agradeço a Deus, que me concedeu forças para seguir e iluminou minhas escolhas até aqui.

Aos meus pais, Almir e Maria, por todo o apoio, cuidado, paciência e dedicação. Jamais esquecerei as noites e madrugadas em que meu pai me buscava na rodoviária, nem do suporte essencial que ambos me deram para que eu pudesse conciliar o trabalho, os estudos para concursos e o mestrado. Sem o apoio deles, eu certamente não teria suportado uma rotina tão exaustiva.

À minha irmã Nathalia, pelo apoio constante, pela torcida e por ouvir meus desabafos nos momentos difíceis. À minha irmã Cintia, pela torcida sempre presente. À minha avó Lindalva, que mesmo à distância me apoia com suas rezas.

À minha orientadora, Profa. Dra. Selene Maria Coelho Loibel, expresso minha profunda gratidão pela orientação, pela paciência e pelo incentivo constante ao longo de todo este processo. Seus ensinamentos foram fundamentais para a realização deste trabalho e para o meu amadurecimento acadêmico.

Ao PGMAT (Programa de Pós-Graduação em Matemática) e seus docentes, pela excelência no ensino. Agradeço, em especial, à Profa. Dra. Carina, pela paciência e compreensão.

Aos professores da minha banca examinadora: à Profa. Marta, pelas orientações fundamentais no exame de qualificação; ao Prof. Dr. Silvio, por suas valiosas contribuições tanto na qualificação quanto na defesa deste trabalho; e ao Prof. Dr. Marcos, por sua participação e contribuições essenciais na banca de defesa.

Às gestões e colegas das escolas onde atuei e atuo – E.E. Profa. Julieta Farão, E.E. Prof. Arthur Chagas Junior, CEU EMEF Tatiana Belinky, CEU EMEI Profa. Erika de Souza Brito Matos e CIEJA Rolando Boldrin – pela compreensão e apoio à minha jornada dupla entre o magistério e a pesquisa.

À gestão de 2022 da Moradia Estudantil da Unesp de Rio Claro. Agradeço a existência desse espaço, essencial para minha vida acadêmica, onde conheci pessoas valiosas que levo no coração: Nathalia, Valeria e Wilson. Agradeço também aos moradores da casa 12 por toda a acolhida e pela flexibilização do cotidiano diante da minha ida semanal para a realização das disciplinas.

Expresso minha gratidão às minhas amigas Carolina, Mayara, Luciana e Daniele, pela torcida e pela compreensão nos períodos em que estive ausente. Um agradecimento especial à Viviane, por todas as conversas e pelos pertinentes conselhos compartilhados.

Aos colegas Raul e Gabriela, pela contribuição direta neste mestrado.

Por fim, agradeço aos meus colegas de mestrado pelo apoio mútuo, estendendo um agradecimento especial à Renata e à Sara, pela parceria e auxílio.

A vida se mostra a cada passo.
Saikawa Roshi

Resumo

Dados longitudinais estão presentes nos mais variados estudos, sendo comum a presença de dados ausentes nesse contexto. O principal objetivo deste trabalho é mostrar as vantagens do uso da Imputação Múltipla de dados ausentes, com o método denominado *Multivariate Imputation by Chained Equations* (MICE). Duas aplicações com dados longitudinais reais foram realizadas com esse fim. O primeiro estudo contou com um conjunto de dados completos sobre a qualidade da água do Rio Piracicaba e, com a retirada aleatória de uma parte dos dados, foram aplicados os diversos métodos de imputação usuais e a imputação múltipla. O conjunto de dados original e os conjuntos imputados com todos os métodos foram analisados e os resultados comparados. No segundo estudo, sobre parâmetros hematológicos e bioquímicos em serpentes da espécie *Crotalus durissus terrificus*, somente a imputação múltipla foi aplicada para obtenção dos perfis médios das variáveis.

Palavras-chave: Imputação Múltipla. Dados longitudinais. Dados ausentes.

Abstract

Longitudinal data are featured in the most varied studies, and the presence of missing data in this context is common. The main objective of this dissertation is to present the advantages of using Multiple Imputation of missing data, with the method named Multivariate Imputation by Chained Equations (MICE). Two applications with real longitudinal data were made for this purpose. The first study had a complete set of data on the water quality of the Piracicaba River and, with the random removal of a part of the data, the various usual imputation methods and multiple imputation were applied. The original dataset and imputed sets with all methods were analyzed, and the results compared. In the second study, on hematological and biochemical parameters in snakes of the species Crotalus durissus terrificus, only multiple imputation was applied to obtain the average profiles of the variables.

Keywords: Multiple Imputation. Longitudinal data. Missing data.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Em qualquer experimento ou estudo que envolve uma análise estatística é importante conhecer e verificar quais são os tipos de dados que o estudo envolve. Há, então, dados denominados como longitudinais que estão presentes em diversas áreas e com isso, faz-se necessário considerar suas especificidades.

Dados longitudinais podem ser definidos como aqueles que são coletados para acompanhar as mudanças de uma mesma variável ao longo do tempo e tais mudanças podem ser comparadas. Na última década, diversos trabalhos analisaram dados longitudinais, principalmente em contextos das áreas biológicas, como por exemplo o de Schober e Vetter (2018) que apresenta três abordagens para dados longitudinais no contexto de pesquisa médica.

Segundo Cao *et al.* (2022) e Rosato *et al.* (2021) a falta de dados é comum em estudos longitudinais e a presença de dados ausentes pode afetar a validade dos resultados e o poder do estudo. Essa não-resposta se faz presente por diversos fatores, sendo um deles a possibilidade, por exemplo, de informações perdidas ao longo das entrevistas da pesquisa. Outra possibilidade referente a não-resposta acontece quando os participantes da pesquisa se recusam a responder algumas questões.

De acordo com Huque *et al.* (2020) para que não haja resultados viciados, é necessário que métodos de imputação sejam adequados para os dados longitudinais. Nesse contexto, a Imputação Múltipla (IM) surge como uma opção viável e é definida como a geração de múltiplas cópias de conjuntos de dados completos, onde os valores ausentes são substituídos por valores imputados amostrados de sua distribuição preditiva a posteriori, considerando os dados observados.

Muitos trabalhos utilizam esse método como por exemplo, o artigo de Huque *et al.* (2020) no qual foram aplicados sete métodos de Imputação Múltipla (IM) para dados ausentes em configurações longitudinais. Nesse artigo, após a imputação com os diversos métodos, o modelo linear misto foi utilizado na análise final.

Outro trabalho que apresenta esse mesmo contexto é o de Cao *et al.* (2022) que compara quatro métodos de imputação múltipla *Fully Conditional Specification* (FCS). O método FCS gera imputações, iterando densidades condicionais de tal modo que utiliza um modelo de regressão univariado apropriado para cada variável com valor ausente.

O objetivo principal desse trabalho foi o estudo dos métodos de imputação frequentemente utilizados e a verificação das vantagens em se utilizar o método de imputação FCS na maioria das aplicações encontradas na literatura.

O objetivo secundário foi aplicar os diversos métodos de imputação para dois conjuntos de dados reais: dados de qualidade da água doce do Rio Piracicaba e dados de parâmetros hematológicos e bioquímicos em serpentes da espécie *Crotalus durissus terrificus*. A partir dessas aplicaçoes foi possível comparar os métodos.

Neste trabalho de dissertação foi aplicado o método *Fully Conditional Specification* (FCS) em dois tipos de formatos de dados longitudinais: *wide* e *long*. dados longitudinais. A saber: no capítulo 2 foram apresentados os conceitos fundamentais de estatística, o capítulo 3 contém a caracterização dos estudos longitudinais com exemplificação e explanação dos formatos possíveis em que dados longitudinais podem ser organizados, o capítulo 4 apresentou a definição de dados ausentes, com o padrões e mecanismos de perda de dados, no capítulo 5 foi explicitada a imputação de dados ausentes com exemplos de métodos mais simplificados, no capítulo 6 foi apresentada a conceitualização de imputação múltipla e suas vantagens, no capítulo 7 duas aplicações em dados longitudinais foram apresentadas e por fim, seguiu o capítulo de conclusões.

2 Conceitos Fundamentais de Inferência Estatística

A Estatística é um conjunto de técnicas que possibilita, de forma sistemática, organizar, descrever, analisar e interpretar dados provenientes de estudos ou experimentos.

2.1 Inferência Estatística

A Inferência Estatística tem como objetivo construir afirmações sobre uma específica característica da população a partir de informações obtidas de uma parte da mesma. A seguir, serão apresentadas algumas abordagens para a realização da inferência.

2.1.1 Inferência Clássica

Um conceito importante é o de que existe variabilidade de amostra para amostra, ou seja, os dados observados formam apenas um dos conjuntos que poderiam ter sido obtidos nas mesmas circunstâncias. Então, na abordagem clássica da inferência a interpretação e análise dos dados não depende apenas do conjunto particular observado, mas das hipóteses adotadas acerca dos outros conjuntos não observados.

Resumidamente, pode-se dizer que existem três procedimentos muito utilizados ao se fazer inferência estatística: a estimação pontual, que é o mesmo que calcular estimativas de parâmetros; a estimação por intervalos, que consiste em construir um conjunto de valores possíveis para um parâmetro, com uma confiabilidade especificada e os testes de hipóteses que, por exemplo, são métodos para tomar decisões sobre um valor previamente concebido para um parâmetro. Em geral, o desempenho desses procedimentos depende de resultados assintóticos, então quanto maior for o tamanho da amostra ($n \to \infty$) coletada, maior a precisão dos resultados. Além disso, alguns métodos têm restrições com relação ao modelo relacionado com os dados. Há casos, por exemplo, em que só é possível realizar testes de hipóteses se os dados do estudo são distribuídos segundo o modelo normal (ou gaussiano).

2.1.2 Inferência Bayesiana

Na abordagem bayesiana, os parâmetros a serem estimados em um modelo, denotados por θ, que pode ser escalar ou um vetor de parâmetros $\theta = (\theta_1, \theta_2, ...\theta_k)$ são variáveis aleatórias e não valores fixos e desconhecidos como na inferência clássica. Assim, pode-se atribuir modelos probabilísticos tanto para as variáveis aleatórias que compõem uma amostra como para os parâmetros que estão relacionados às distribuições de probabilidade

dessas variáveis. Nesse contexto, os dados apenas entram na função de verossimilhança. Nenhuma probabilidade sobre o espaço amostral, além do valor real da função de verossimilhança, é admitida nas conclusões. Por exemplo, os conceitos de nível de significância e poder do teste que são probabilidades sobre o espaço amostral não são considerados.

Considerar um problema no qual se pretende modelar os dados com uma função densidade de probabilidade com apenas um parâmetro $p(X|\theta)$, sendo que X representa a variável aleatória de interesse e θ é o parâmetro que precisa ser estimado. A seguir apresenta-se um resumo do procedimento básico da inferência bayesiana.

Inicialmente, estuda-se o problema buscando informações de pesquisadores da área ou em estudos anteriores sobre o assunto a fim de estabelecer algumas características para θ (em alguns casos o parâmetro tem interpretação prática direta). Importante observar se ele pode ser considerado discreto ou contínuo, se há restrições no seu domínio, entre outros. Com isso, escolher uma densidade a priori (informativa ou não) de acordo com as informações obtidas. Assumindo θ como uma variável aleatória contínua, a função densidade de probabilidade a priori para θ é denotada por $p(\theta)$.

Para uma amostra de X, com tamanho n, a função de verossimilhança de acordo com o modelo estatístico escolhido anteriormente é dada por:

$$p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta) \tag{2.1}$$

Combina-se a função de verossimilhança e a densidade a priori para obter a densidade de probabilidade a posteriori, denotada por $p(\theta|X)$ como segue:

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{\int\limits_{\Theta} p(\theta)p(X|\theta)d\theta} \propto p(\theta)p(X|\theta) \tag{2.2}$$

Sendo $\propto$, símbolo para proporcionalidade. No caso discreto, a função de probabilidade a posteriori para θ é calculada de forma análoga, apenas substituindo a integral por somatório no denominador de (2.2). Uma vez calculada $p(\theta|X)$, todas as informações sobre o parâmetro θ são obtidas a partir dela, então pode-se calcular a média, a moda, intervalos de credibilidade, fazer gráficos da densidade, testes de hipóteses bayesianos e etc.

Distribuição preditiva

No método de imputação com o algoritmo MICE há o interesse em fazer a previsão de um novo valor de X, denotado por X^*, que é usado para imputar um valor ausente. Para isso pode-se utilizar a distribuição preditiva dada por:

$$p(X^*|X) = \int\limits_{\Theta} p(X^*, \theta|X)d\theta = \int\limits_{\Theta} p(X^*|\theta, X)p(\theta|X)d\theta \tag{2.3}$$

sendo que $p(X^*, \theta|x)$ é a verossimilhança de X^* e $p(\theta|X)$ é a densidade a posteriori dada por (2.2). O método de imputação com o algoritmo MICE está apresentado em detalhes no Capítulo 6.

Amostrador de Gibbs

Seja X uma variável aleatória contínua com função de verossimilhança para uma amostra denotada por $p(X|\theta_1, \theta_2)$. Assumir que θ_1 e θ_2 são variáveis aleatórias contínuas e uma distribuição a priori conjunta para θ_1 e θ_2 é denotada por $p(\theta_1, \theta_2)$. De (2.2) obtém-se a distribuição a posteriori conjunta para θ_1 e θ_2 com:

$$p(\theta_1, \theta_2|X) \propto p(\theta_1, \theta_2)\, p(X|\theta_1, \theta_2) \tag{2.4}$$

A obtenção das distribuições a posteriori marginais para θ_1 e θ_2 pode ser feita com a integração da distribuição a posteriori conjunta como segue:

$$p(\theta_i|X) = \int_{\theta_j} p(\theta_1, \theta_2|X) d\theta_j \quad i, j = 1, 2 \quad \text{e} \quad i \neq j \tag{2.5}$$

Se $p(\theta_1, \theta_2|X)$ não tem solução analítica $p(\theta_1|X)$ e $p(\theta_2|X)$ podem ser obtidas por simulação.

Com o método de simulação de Monte Carlo em Cadeias de Markov (MCMC), denominado de algoritmo Amostrador de Gibbs, é possível gerar valores de θ_1 e θ_2 e a partir das distribuições condicionais $p(\theta_i|\theta_j, X)$, $i, j = 1, 2$ e $i \neq j$ que, em geral, são obtidas facilmente como pode ser visto em Gilks *et al.* (1996).

Algoritmo Amostrador de Gibbs

Passo 1: Atribuir valores iniciais para os parâmetros $\theta_1^{(0)}$ e $\theta_2^{(0)}$ baseados nas informações a priori

Com $t = 1$
Passo 2: Gerar $\theta_1^{(t)}$ com $p(\theta_1|\theta_2^{(t-1)}, X)$ e geral $\theta_2^{(t)}$ com $p(\theta_2|\theta_1^{(t)}, X)$

Repetir o passo 2 para $t = 2, ..., K$.

Os valores gerados formam amostras provenientes das distribuições a posteriori marginais:

Amostra de $\theta_1 : \theta_1^{(1)}, ..., \theta_1^{(K)}$ e amostra de $\theta_2 : \theta_2^{(1)}, ..., \theta_2^{(K)}$

Em geral é necessário considerar um valor alto para K ($K \geq 500$), mas o custo computacional ao executar este algoritmo é baixo. A teoria de MCMC garante a distribuição estacionária dos valores gerados. Mais detalhes podem ser vistos em Gilks *et al.* (1996) e em Casella e George (1992).

3 Caracterização dos Estudos Longitudinais

Neste capítulo são apresentadas as definições de estudos longitudinais, bem como exemplificação e formatos de organização.

3.1 Conceitualização

Em seus estudos para dados longitudinais, Singer *et al.* (2018) apresentam conceitos no contexto de análise univariada.

Segundo os autores, há duas estratégias para a coleta de dados: planejamento transversal e planejamento longitudinal. O primeiro é definido como uma única observação da variável resposta em uma população de interesse. Já a segunda estratégia de coleta consiste em considerar duas ou mais observações da variável resposta em cada unidade amostral.

Estudos longitudinais podem ser vistos como um subconjunto das medidas repetidas e são utilizados em muitos contextos onde se faz necessário modelar comportamentos ao longo de uma escala ordenada. Assim, são analisadas uma ou mais variáveis respostas medidas para uma ou mais populações.

Uma diferenciação necessária de ser realizada é a de que dados longitudinais não são séries históricas, pois as séries de tempo analisam um único indivíduo em muitos instantes, ao passo que dados longitudinais devem lidar com muitas unidades amostrais em um tempo menor do que costuma-se ocorrer em séries históricas.

3.1.1 Exemplo de um conjunto de dados longitudinais com covariáveis

Seja um estudo em que a variável resposta é a nota das provas de matemática durante o ano letivo de alunos do segundo ano de uma determinada escola. O fator X representa um método de ensino (Método A, X=O e Método B, X=1), W representa o sexo (Feminino, W=0 e Masculino, W=1), a variável V indica a idade e a variável Z representa o número de horas dedicadas ao estudo num determinado período. A Tabela 3.1 indica essa representação:

Tabela 3.1: Estrutura para dados longitudinais

			Covariáveis			
Unidade amostral	Resposta	Tempo	X	W	V	Z
1	y_{11}	t_{11}	x_1	w_1	v_1	z_{11}
1	y_{12}	t_{12}	x_1	w_1	v_1	z_{12}
.	.	.	.	.	.	.
1	y_{1m_1}	t_{1m_1}	x_1	w_1	v_1	z_{1m_1}
.	.	.	.	.	.	.
n	y_{n1}	t_{n1}	x_n	w_n	v_n	z_{n1}
n	y_{n2}	t_{n2}	x_n	w_n	v_n	z_{n2}
.	.	.	.	.	.	.
n	y_{nm_n}	t_{nm_n}	x_n	w_n	v_n	z_{nm_n}

Fonte: Singer *et al.*, 2018

3.2 Formatos para organização de dados longitudinais

Os dados longitudinais podem ser organizados em dois formatos: *long* (longo), no qual há uma coluna para todos os registros de cada indivíduo e a variável relativa ao tempo é disposta em uma única coluna, ou *wide* (largo), no qual para cada tempo os registros de cada indivíduo são dispostos em colunas.

O formato *wide* é indicado quando a informação sobre todos os indivíduos é registrada em intervalos semelhantes. Desse modo, não há redundância ou repetição.

O formato *long* é indicado para gerar gráficos e análises estatísticas mais complexas. Pesquisadores aplicados frequentemente coletam, armazenam e analisam seus dados no formato *long*. ANOVA e MANOVA são utilizados para medidas repetidas e modelos com equações estruturadas para dados longitudinais assumem o formato *wide*. Técnicas com modelos multiníveis e gráficos estatísticos trabalham com o formato *long*.

De qualquer modo, é conveniente ter os dados organizados e armazenados nos dois formatos segundo van Buuren (2012).

A imputação múltipla para dados longitudinais é feita convenientemente em dados no formato *wide*. Além do fato de que as colunas são ordenadas no tempo, não há nada de especial sobre o problema da imputação.

As representações dos formatos acima citados se dão do seguinte modo de acordo com Cao *et al.* (2022). Seja $Y = \{y_{ijk}\}$ a representação de um conjunto de dados longitudinais multivariados de tal forma que y_{ijk} é o valor da variável Y_{ijk} com $j = 1, ..., p$ representando as variáveis, $i = 1, ..., n$ para os indivíduos, $k = 1, ..., K$ para os tempos. A seguir, nas Tabelas 3.2 e 3.3, apresentam-se os dados nos formatos *wide* e *long*, respectivamente, considerando 3 variáveis Y_1, Y_2 e Y_3 medidas em três instantes de tempo.

Suponha que Y está no formato *wide* e então, sua representação fica armazenada em uma matriz $n \times pK$ com cada linha representando um específico indivíduo e cada coluna correspondendo a uma variável Y_{ijk} medida no tempo t_k, denotada por Y_{jk}.

Tabela 3.2: Tabela no formato *wide*

	Tempo t_1			Tempo t_2			Tempo t_3		
ID	Y_{11}	Y_{21}	Y_{31}	Y_{12}	Y_{22}	Y_{32}	Y_{13}	Y_{23}	Y_{33}
1	y_{111}	y_{121}	y_{131}	y_{112}	y_{122}	y_{132}	y_{113}	y_{123}	y_{133}
2	y_{211}	y_{221}	y_{231}	y_{212}	y_{222}	y_{232}	y_{213}	y_{223}	y_{233}
3	y_{311}	y_{321}	y_{331}	y_{312}	y_{322}	y_{332}	y_{313}	y_{323}	y_{333}

Fonte: Elaboração da autora (2024)

No formato *long*, Y é uma matriz $nK \times (p+1)$, com a última coluna descrevendo os dados de tempo em que a observação é registrada e as demais colunas são para as p variáveis.

Tabela 3.3: Tabela no formato *long*

ID	Y_1	Y_2	Y_3	Tempo
1	y_{111}	y_{121}	y_{131}	t_1
1	y_{112}	y_{122}	y_{132}	t_2
1	y_{113}	y_{123}	y_{133}	t_3
2	y_{211}	y_{221}	y_{231}	t_1
2	y_{212}	y_{222}	y_{232}	t_2
2	y_{213}	y_{223}	y_{233}	t_3
3	y_{311}	y_{321}	y_{331}	t_1
3	y_{312}	y_{322}	y_{332}	t_2
3	y_{313}	y_{323}	y_{333}	t_3

Fonte: Elaboração da autora (2024)

Para simplificar, é definido Y_l, com $l = 1, ..., L$, que é uma coluna no formato *long* ou *wide*. No formato *wide*, $L = pK$ e Y_l é um vetor de tamanho n. No formato *long*, $L = p$ e Y_l é um vetor de tamanho nK.

4 Dados Ausentes

Neste capítulo são apresentadas as definições de padrões e mecanismos de dados ausentes e exemplos.

Ao utilizar métodos de inferência estatística, é esperado que sejam obtidos estimadores eficientes para os parâmetros de interesse. Uma possível dificuldade para esse processo é a existência de dados ausentes uma vez que este fato pode acarretar um vício nos estimadores.

Tais dados podem ser definidos como aqueles que ocorrem quando os valores das variáveis de interesse não são medidos ou registrados para todos os elementos da amostra, Austin *et al.* (2021) .

É importante compreender os possíveis motivos que levaram à falta de dados, sendo que há diversas possibilidades para isso. Algumas das possíveis causas, Brown e Kros (2003) e Lobato (2016).

1. Fatores operacionais como estimativas, erros na entrada dos dados, remoção (acidental) de campos das tabelas, entre outros. Essa causa pode ser exemplificada com uma situação de mau funcionamento de dispositivos de coletas de dados.

2. Recusa na resposta em pesquisa de opinião. Um exemplo possível para esse tópico são estudos na área médica em que pode haver omissão de comportamentos de risco como o consumo de drogas lícitas e/ou ilícitas.

3. Impossibilidade da aplicação de questões. É possível ver essa situação no estudo do desastre de fogos de artifício SE em que houve um acidente que resultou em estresse pós-traumático para muitas pessoas. Foi realizado um ensaio clínico randomizado comparando dois tratamentos para transtornos relacionados à ansiedade. Houve 52 participantes, sendo eles crianças de 4 a 12 anos, adolescentes de 13 a 18 e seus responsáveis. As crianças que tinham menos de 6 anos não preencheram o formulário infantil e, portanto, suas respostas são dados ausentes, van Buuren (2012).

4.1 Padrão e quantificação de dados ausentes

Os dados ausentes podem ser caracterizados em uma série de padrões, que identificam se há ou não um comportamento comum quanto à forma como os dados foram observados. Os principais padrões de ausência de dados são: univariado, monotônico e arbitrário, Lobato (2016).

Na Figura 4.1 vê-se à esquerda o padrão univariado que descreve a situação na qual há dados ausentes apenas em uma das variáveis, por exemplo em Y. Um exemplo deste

padrão é a negligência de um determinado item em um questionário, como por exemplo a informação da renda mensal.

Ao centro, vê-se o padrão monotônico, no qual os dados passam a faltar em determinada variável (Y_2) e, a partir disso, todas as outras variáveis terão valores ausentes. Este comportamento é comum em estudos longitudinais, sendo chamado de *dropout*. Isso pode ocorrer por exemplo quando a observação de Y_j depende da observação de Y_{j-1}, $j = 2, ..., p$.

À direita vê-se terceiro e último padrão, dito arbitrário ou geral, quando variáveis são negligenciadas de forma aleatória no conjunto de casos.

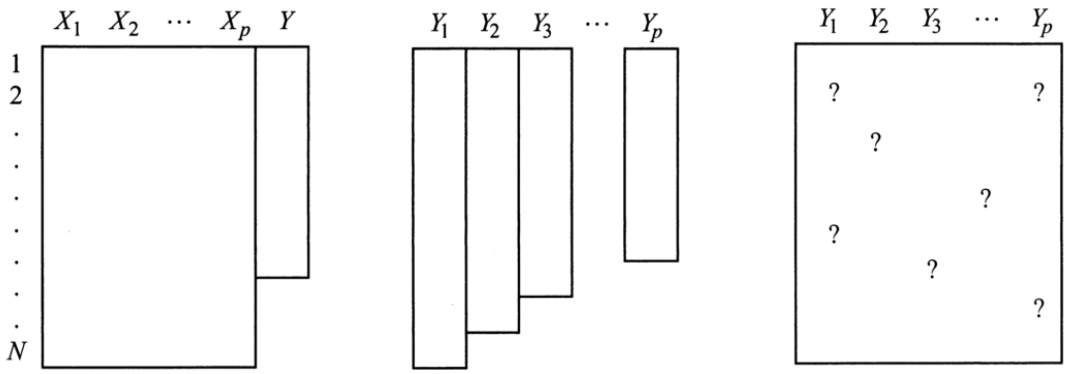

Figura 4.1: Padrões de ausência de dados
Fonte: Lobato, 2016

4.2 Mecanismos que geram dados ausentes

Rubin (1977) classificou o problema dos dados ausentes em três categorias. Dessa forma, cada dado tem uma específica probabilidade de ser ausente e são os mecanismos de dados ausentes que a definem. As seguintes exemplificações dos mecanismos foram baseadas no livro de van Burren (2012) e no trabalho de Mack *et al.* (2018).

1. **MCAR *(Missing Completely at Random)***: Ausência completamente randômica, que significa que todos os dados possuem a mesma probabilidade de serem ausentes. Com isso, não há diferenças sistemáticas entre os participantes com dados ausentes e aqueles com dados completos. Para exemplificar esse mecanismo é necessário considerar que em um estudo houve uma intromissão externa e a causa da ausência não está relacionada com o dado em si. Supor uma aplicação com enfoque nos pesos de um objeto em uma determinada balança durante um período e o equipamento ficou sem bateria.

2. **MAR (*Missing at random*)**: É possível notar que em comparação com o mecanismo anterior, foi retirada a palavra *completely* o que dá a ideia de menor aleatoriedade. No contexto de MAR, a probabilidade de ausência é a mesma somente dentro dos grupos definidos pelos dados observados, ou seja, a ausência de dados está sistematicamente relacionada com os dados observados. Para o exemplo da balança, supor que ela está em uma superfície instável, o que pode produzir mais valores ausentes.

3. **MNAR (*Missing Not at Random*)**: Para esse mecanismo, a probabilidade de ausência varia mediante eventos ou fatores que não são medidos pelo pesquisador. Assim, o mecanismo de ausência não pode ser previsto usando as informações que estão no conjunto de dados observados. Dessa forma, para o exemplo do estudo da balança, seria o equipamento estar precisando de manutenção e não se ter conhecimento disso.

5 Imputação de Dados Ausentes

Neste capítulo são apresentados alguns métodos de imputação, ditos *ad hoc*, ou seja, métodos que não foram devidamente construídos em razão da necessidade de atender a uma demanda específica dos usuários. Esses métodos e seus problemas são exemplificados em uma aplicação na qual comparam-se os seus resultados com o resultado do método de imputação múltipla.

5.1 Histórico das ideias para imputação

O substantivo "imputação" perpassou os séculos tendo alguns significados e em grande parte do tempo foi associado ao contexto de tributação de impostos. No século XIX, era utilizado o termo "rendimento imputado" para designar rendimentos advindos de terrenos. Somente em 1961, no trabalho do US Census Bureau que a palavra imputação foi mencionada com o sentido de preencher dados ausentes.

Os autores Allan e Wishart foram os primeiros a desenvolver um método estatístico para substituir um valor ausente. Na publicação Allan e Wishart (1930), os dados possuíam um único valor ausente e os autores forneceram duas fórmulas para estimar tal valor.

Posterior a isso, Yates (1933) fez a generalização para mais de uma observação ausente, iniciando o caminho para o clássico algoritmo de expectativa-maximização, também chamado de algoritmo EM, que consiste em uma técnica iterativa para dados ausentes, van Buuren (2012). O termo "imputação" foi utilizado no trabalho de 1983 denotado por *Panel in Incomplete Data* e foi no livro de Little e Rubin (1987) que o termo imputação foi estabelecido na literatura estatística.

5.2 Métodos usuais de imputação para valores ausentes

Há diversos métodos de imputação de dados e alguns deles apresentam soluções simplificadas que nem sempre apresentam o êxito esperado. Esses métodos tratam os dados ausentes por meio da simples omissão, seja de casos ou de variáveis, que contenham dados ausentes, editam os dados incompletos para produzir um conjunto de dados completo e são muito utilizados por sua fácil implementação. No entanto, os pesquisadores devem ter cuidado ao utilizá-los porque tem sido mostrado que causam sérios problemas. Alguns autores se referem a esses métodos como inaceitáveis como pode ser visto em Graham *el al* (1997), Graham e Hofer (2000), Little e Schenker (1995) e Shafer e Graham (2002).

Nesse contexto, é possível citar os seguintes métodos: *Listwise deletion, Pairwise deletion,* Imputação pela média, Imputação por regressão e Imputação por regressão estocástica. Além desses, existe outro método muito utilizado para dados longitudinais, denominado como *Last observation carried forward* (LOCF). O método consiste em tomar o último valor observado como substituto dos dados ausentes. A utilização deste método é criticada por vários autores devido aos vícios nos estimadores obtidos, Lachin (2016).

Segue um resumo e aplicações dos métodos citados, baseado principalmente no livro de van Burren (2012). Para exemplificar a aplicação dos diversos métodos, na Tabela 5.1 apresenta-se um conjunto de dados sobre a qualidade do ar que inclui a variável resposta: Y - %Ozônio (porcentagem de ozônio) e 5 covariáveis: X_1 - Rad. Solar (radiação solar), X_2 - Vento e X_3 - Temp. (temperatura), sendo que NA indica dado ausente. A tabela com todos os 153 dados pode ser vista no Apêndice A.

Tabela 5.1: Dados de qualidade do ar

Id	Y : %Ozônio	X_1 :Rad. Solar	X_2 :Vento	X_3 :Temp.
1	12	149	12,6	74
2	18	313	11,5	62
3	NA	NA	14,3	56
4	28	NA	14,9	66
5	NA	194	8,6	69
...	...	...	...	...
153	20	223	11,5	68

Fonte: Adaptação de van Buuren, 2012

Nos métodos de imputação *Listwise deletion* e *Pairwise deletion* o tratamento para dados ausentes se dá por meio da omissão, de casos ou variáveis, que incluam dados ausentes. Assim, é produzido um conjunto de dados completos que pode ser utilizado pela simples execução e conveniência.

5.2.1 *Listwise deletion*

Tal método pode ser chamado de análise de casos completos e consiste em eliminar todos os casos com um ou mais valores ausentes. A vantagem que essa análise oferece é a conveniência.

Se os dados forem MCAR, esse método produz estimativas não viciadas de médias, variâncias e parâmetros de modelos de regressão.

É reconhecido que o método *Listwise deletion* pode fazer com que mais da metade da amostra original seja perdida, ainda mais quando o número de variáveis é grande. É importante ressaltar que há aplicações para algumas áreas, como é o caso das ciências sociais em que os pesquisadores tradicionalmente lidam com dados ausentes e eliminam todas as observações do conjunto de dados em que há ausência de valores em pelo menos uma variável, Pepinsky (2018).

O problema inicia-se quando os dados não são MCAR e o método pode causar grande erro em estimativas de médias, coeficientes de regressão e correlações e isso não representa bem a amostra completa. Little e Rubin (2002), mostraram que o vício na estimativa de

médias cresce com a diferença entre as médias dos casos completos e ausentes e com a proporção de dados ausentes.

É possível que existam contextos em que o método de análise de casos completos seja atrativo, como foi citado com MCAR que produz estimadores não viciados, porém as implicações do uso dependem do contexto dos dados. As opiniões sobre seu uso variam, mas a maioria dos autores considera que as desvantagens superam as vantagens.

Aplicação 1: *Listwise deletion*
Interesse: Ajustar modelos de regressão, via método dos mínimos quadrados ordinários, para verificar a relação entre a % de ozônio no ar e as outras variáveis. Por exemplo, o modelo 1 relaciona a % de ozônio no ar (Y) com o vento (X_2):

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon \qquad \text{com} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Ajuste do modelo 1 com 37 dias excluídos: $\widehat{\beta_0}$ = 96,8729 e $\widehat{\beta_2}$ = -5,5509.

Supor agora que temos interesse em acrescentar a variável radiação solar (X_1), o modelo 2 é dado por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad \text{com} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Ajuste do modelo 2 com 42 dias excluídos: $\widehat{\beta_0}$ = 72,246; $\widehat{\beta_1}$ = 0,1003 e $\widehat{\beta_2}$ = -5,4018.

Os modelos foram ajustados somente com os dados completos. No modelo 2, ao acrescentar a variável radiação solar, a amostra é alterada e o número de valores ausentes aumenta de 37 para 42. Isso traz consequências para a metodologia e interpretação dos resultados, algumas questões são:

1. Pode-se comparar os coeficientes dos dois modelos?
2. Deve se atribuir as mudanças nos coeficientes à mudança de modelo ou à mudança na amostra?
3. Os valores estimados para os coeficientes podem ser generalizados para a população?
4. A amostra tem o tamanho suficiente para detectar os efeitos de interesse?

A simples retirada dos casos (dias) em que essas duas variáveis não foram medidas afeta a análise se a variação no tempo é importante e se quer estabelecer padrões semanais ou ajustar modelos autorregressivos.

5.2.2 *Pairwise deletion*

Também conhecida como análise de casos disponíveis, o método calcula as médias e covariâncias para todos os dados observados. Assim, a média da variável X é baseada em todos os casos que foram observados em X, a média da variável Y é baseada em todos os casos que foram observados em Y e assim por diante. Para calcular a correlação e a covariância, são utilizados os dados observados de X e Y. Os estimadores podem ser viciados se os dados não são do tipo MCAR. Além disso, existem problemas computacionais. A matriz de correlação pode não ser positiva, o que é necessário para a maioria dos métodos multivariados. Correlações fora do intervalo $[-1.1]$ podem ocorrer. Outro problema é que não fica claro qual tamanho de amostra deve ser usado para o cálculo de

erros-padrão. Utilizar a média dos tamanhos de amostra produz erros-padrão subestimados, Little (1992).

Aplicação 2: *Pairwise deletion*

Supor que temos interesse em calcular as médias e matriz de covariâncias para as quatro primeiras variáveis dos dados de qualidade do ar ($Y-$Ozônio, X_1-radiação solar, X_2-vento e X_3-temperatura). Com esse método, esses cálculos são feitos com os valores disponíveis de cada uma das variáveis.

Médias: $\overline{Y}$=42,1293; $\overline{X}_1$=185,9315; $\overline{X}_2$=9,9575 e $\overline{X}_3$=77,8823

Matriz de covariâncias:

$$Cov = \begin{pmatrix} 1088.20 & -70.94 & 1056.58 & 218.52 \\ 1056.58 & 8110.52 & -17.94 & 229.16 \\ -70.94 & -17.94 & 12.41 & -15.27 \\ 218.52 & 229.16 & -15.27 & 89.59 \end{pmatrix}$$

O problema nesta aplicação é que cada um dos cálculos das entradas dessa matriz é feito com a quantidade disponível de dados de cada variável, o que torna impossível fazer comparações e interpretações.

5.2.3 Imputação pela média

Essa imputação não é recomendável pois apesar de ser a forma mais rápida para completar dados ausentes, pelo fato de utilizar um único valor e repeti-lo no lugar de todos os dados ausentes, esse método reduz artificialmente a variância desta variável e diminui a possibilidade de reconhecer as relações entre as variáveis em geral. Além disso, produz estimadores viciados para a média quando os dados não são do tipo MCAR.

Aplicação 3: Imputação pela média

Imputar os valores ausentes das variáveis % Ozonio e radiação solar, pelas suas médias. Na Figura 5.1 estão em azul os dados observados e em vermelho os dados imputados.

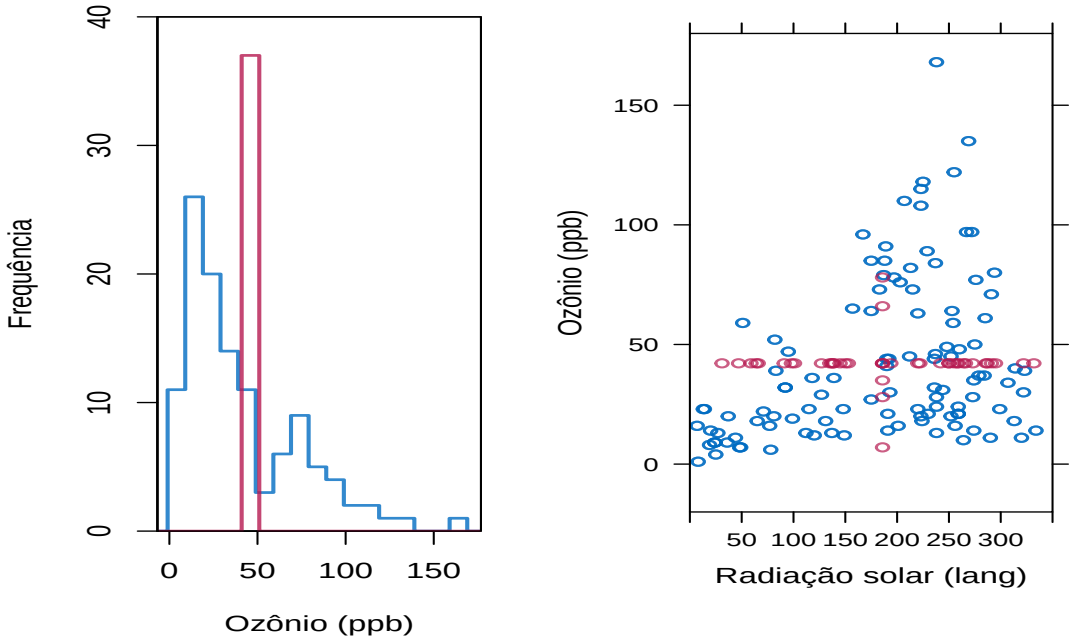

Figura 5.1: Imputação pela média
Fonte: Adaptação de van Buuren, 2012

Na Figura 5.1 do lado esquerdo vemos no gráfico que a imputação cria uma distribuição bimodal. O desvio-padrão para o conjunto imputado é 28,7, muito menor que o desvio-padrão se considerarmos apenas os dados observados, que é 33.

A Figura 5.1 do lado direito mostra que a relação entre a % de ozônio e radiação solar é distorcida pelas imputações. A correlação entre essas variáveis é de 0,35 com os dados observados (em azul somente) e é de 0,3 para o conjunto imputado (em azul e em vermelho juntos).

5.2.4 Imputação por regressão

Tal método incorpora o conhecimento sobre outras variáveis com a ideia de produzir boas imputações. O primeiro passo envolve a construção de um modelo para os dados observados. As predições para os dados ausentes são então calculadas sob este modelo e servem para repor esses dados. Esse método produz estimadores não viciados para a média se os dados são do tipo MCAR e para os coeficientes do modelo de imputação se as variáveis explicativas são completas. No entanto, as estimativas dos coeficientes são viciadas se os dados são do tipo MAR e se os fatores que influenciam a perda de dados fazem parte do modelo. A variabilidade dos dados imputados pode ser subestimada.

Aplicação 4: Imputação por regressão
Supor que será realizada a imputação com o modelo 3, dado por:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad com \;\; \varepsilon \sim N(0, \sigma_\varepsilon^2) \tag{5.1}$$

Primeiro, ajusta-se esse modelo aos dados observados, via método dos mínimos quadrados ordinários, e as predições para os dados ausentes são feitas sob esse modelo ajustado.

Resultados: $\widehat{\beta_0}$=18,5987 e $\widehat{\beta_1}$=0,1272

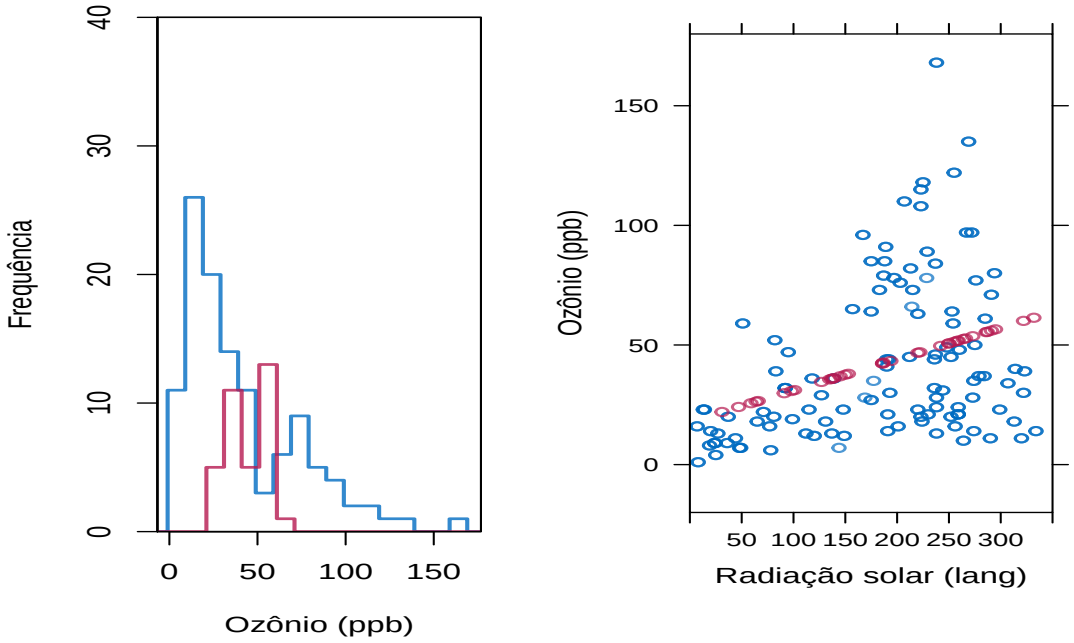

Figura 5.2: Imputação por regressão
Fonte: Adaptação de van Buuren, 2012

A Figura 5.2 do lado esquerdo, mostra que os valores imputados correspondem aos mais prováveis sob esse modelo, no entanto o conjunto de valores imputados (em vermelho) varia menos que o conjunto de dados observados (em azul). Portanto, a distribuição dos dados imputados não reflete a realidade da distribuição dos valores da % de ozônio.

Como pode-se ver na Figura 5.2 do lado direito, a imputação com valores preditos pelo modelo de regressão também tem efeito sobre a correlação entre as variáveis % de ozônio e radiação solar. Considerando somente os valores imputados, essas variáveis têm correlação igual a 1 (um), desde que estão dispostos em linha. Ao passo que se considerarmos os dados observados e os dados imputados conjuntamente, temos correlação 0,39. Os coeficientes da regressão são não viciados se os dados são MAR e se os fatores que influenciam na perda de dados são parte do modelo de regressão. Nesse exemplo, isso corresponde à situação em que a radiação solar explicaria alguma diferença nas probabilidades de perda da variável % de ozônio. Verifica-se que a variabilidade nos dados imputados é subestimada, assim como no caso da imputação pela média.

5.2.5 Imputação por regressão estocástica

Esse método é um refinamento da imputação por regressão que adiciona ruído nas variáveis explicativas do modelo. Primeiro, no caso da regressão linear simples, com apenas uma variável explicativa, são estimados com os dados observados, via método dos mínimos quadrados ordinários, o intercepto (β_0), a inclinação da reta (β_1) e a variância do ruído (σ^2), com no modelo de regressão dado em (5.1). Então são gerados os valores

de Y a serem imputados de acordo com o modelo que pode ser escrito como:

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 X^A + e \tag{5.2}$$

sendo que X^A é o subconjunto com as linhas de X para as quais a variável Y tem valores ausentes e $e \sim N(0, \sigma^2)$. Esse método não resolve todos os problemas, mas a ideia de gerar as imputação a partir do modelo com esse ruído é muito boa e formou a base para técnicas mais avançadas de imputação, van Buuren (2012).

Aplicação 5: Imputação por regressão estocástica

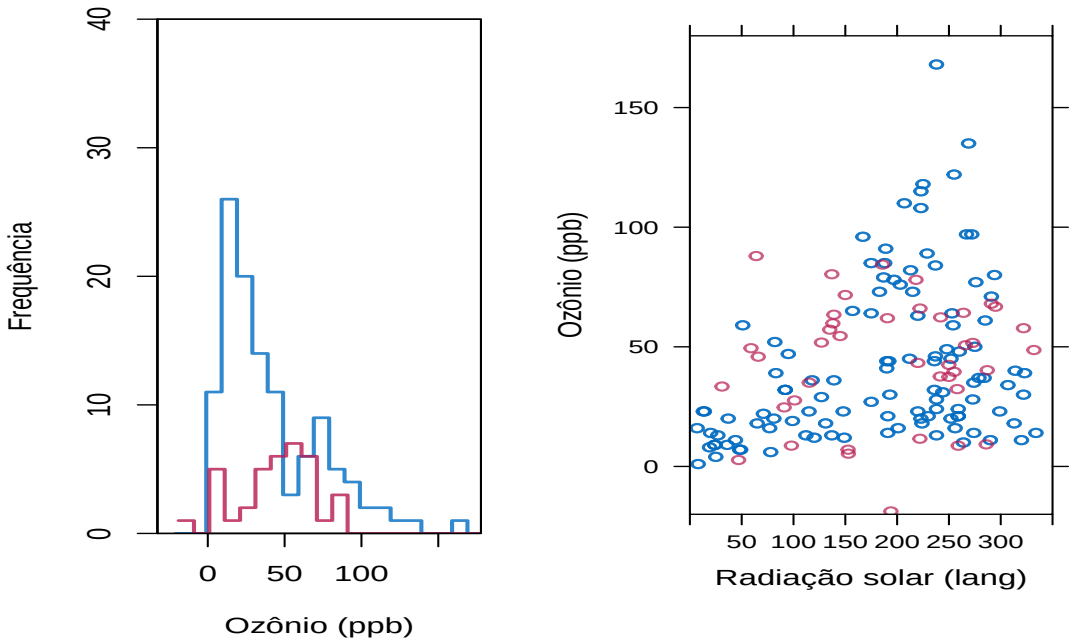

Figura 5.3: Imputação por regressão estocástica
Fonte: Adaptação de Van Buuren, 2012

Observa-se na Figura 5.3 do lado esquerdo, que o método gera alguns valores negativos, sendo que não podem existir % negativas de ozônio. A distribuição para valores altos não é bem coberta e isso ocorre por causa da heterocedasticidade. A variabilidade nos valores de % de ozônio cresce quando a radiação solar atinge o nível de 250 e decresce logo após. Esse fato é devido a um fenômeno genuinamente meteorológico e o modelo de imputação não o considera.

Existem métodos que não se concentram em identificar uma forma de substituir os dados, mas utilizar a informação disponível para preservar as relações entre as variáveis do conjunto de dados completos.

- O método baseado em verossimilhança é um destes, mas requer a especificação de um modelo estatístico para cada análise.
- O algoritmo EM (*Expectation Maximization*) é outro método que tem sido usado, mas a obtenção de erros-padrão usando EM envolve métodos auxiliares tais como o *bootstrap*.

Apenas para efeito de comparação, a seguir apresenta-se o resultado do método de Imputação Múltipla (IM) para os dados de qualidade do ar.

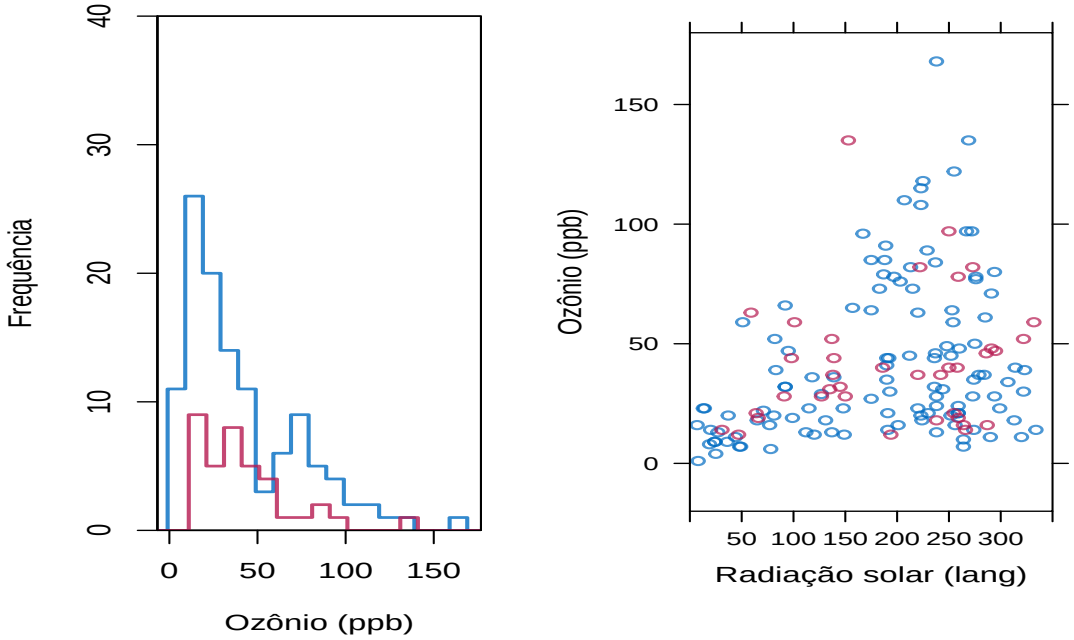

Figura 5.4: Imputação múltipla - Primeiro conjunto completo
Fonte: Adaptação de van Buuren, 2012

Na Figura 5.4 temos, do lado esquerdo, a distribuição da variável % de ozônio para os dados completos e imputados. Podemos observar que a distribuição dos dados observados (em azul) e dos dados imputados (em vermelho) são muito similares.

O problema de gerar valores negativos para a variavel % de ozônio não ocorre com esse método, como no método da regressão estocástica, uma vez que a IM utiliza os valores observados como "doadores" para a imputação dos valores ausentes.

Do lado direito, temos o gráfico de dispersão das variáveis % de ozônio e radiação solar também para os dados completos e imputados e vemos que a IM mantém a heterocedasticidade natural na relação entre as variáveis.

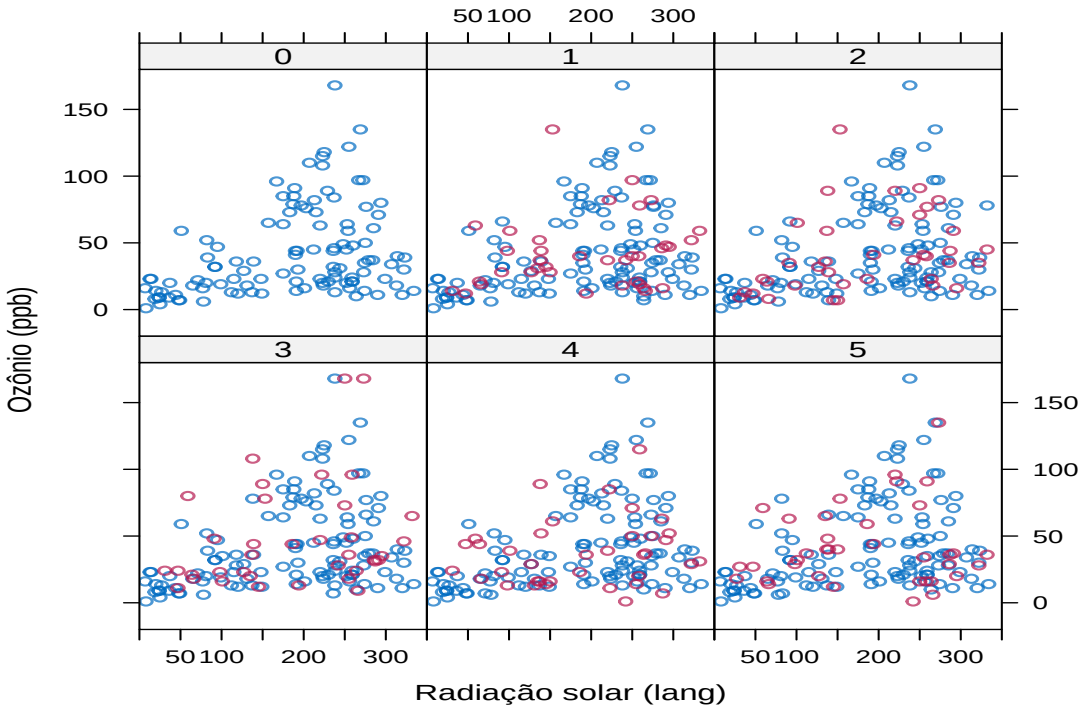

Figura 5.5: Dispersão dos dados observados e de cada um dos 5 conjuntos de dados imputados

Fonte: Adaptação de van Buuren, 2012

A Figura 5.5 apresenta os gráficos de dispersão considerando primeiro somente os dados observados (somente azul) e cinco conjuntos imputados (azul e vermelho). O método de Imputação Múltipla em detalhes e aplicações com dados reais estão apresentados no Capítulo 6.

6 Imputação Múltipla (IM)

Esse capítulo apresenta conceitualização de forma detalhada e a exemplificação da Imputação Múltipla.

6.1 Histórico dos métodos de Imputação Múltipla

Em 1977, Fritz Scheuren trabalhou no projeto articulado com o *Social Security Administration* e o *US Census Bureau* e a imputação múltipla foi desenvolvida como solução para o problema de falta de dados de renda num contexto de pesquisa populacional, denominada como *Current Population Survey (CPS)*, Scheuren *et al.* (1977) e *US Census Bureau* (1961). O *US Census Bureau* estava utilizando um procedimento para imputar e Scheuren notou que a variância não estava sendo calculada adequadamente. Ao levar esse questionamento para Rubin obteve como solução a utilização de múltiplas versões do conjunto de dados completos que ele havia explorado no início da década de 70. O relatório original de 1977 foi publicado em Rubin (2004).

Rubin observou que fazer uma imputação única para um dado ausente poderia, no geral, não ser o ideal. Dessa forma, ele precisava de um modelo para relacionar dados não observados com os dados observados e notou que para um determinado modelo os valores não eram gerados com a precisão adequada. Diante desse fato, criou múltiplas imputações que refletiam a incerteza. Em 1977 lança um trabalho com explicações sobre a execução, Rubin (1977).

A proposta original de Rubin não inclui fórmulas para o cálculo de estimativas combinadas e melhorias do método foram realizadas a partir do final da década de 1980.

A performance da IM em várias situações diferentes tem sido bem estudada e tem se mostrado eficiente por produzir estimadores não viciados que refletem a incerteza associada com a estimação dos dados ausentes, Graham *el al* (1997), Graham e Hofer (2000) e Shafer e Graham (2002).

6.2 Imputação Múltipla

A imputação múltipla (IM) de dados veio como uma alternativa flexível aos métodos baseados em verossimilhança para diversos casos, fornecendo múltiplas opções de imputação para cada valor ausente, Rubin (1986), Rubin (1987), Silva (2012) e van Buuren (2012). É possível definir a imputação múltipla como um procedimento que cria $m > 1$ conjuntos de dados completos, substituindo os valores ausentes por valores plausíveis e cada um desses conjuntos é avaliado pelo método estatístico adequado. Vale ressaltar que esses valores substituídos são extraídos de uma distribuição modelada para cada valor

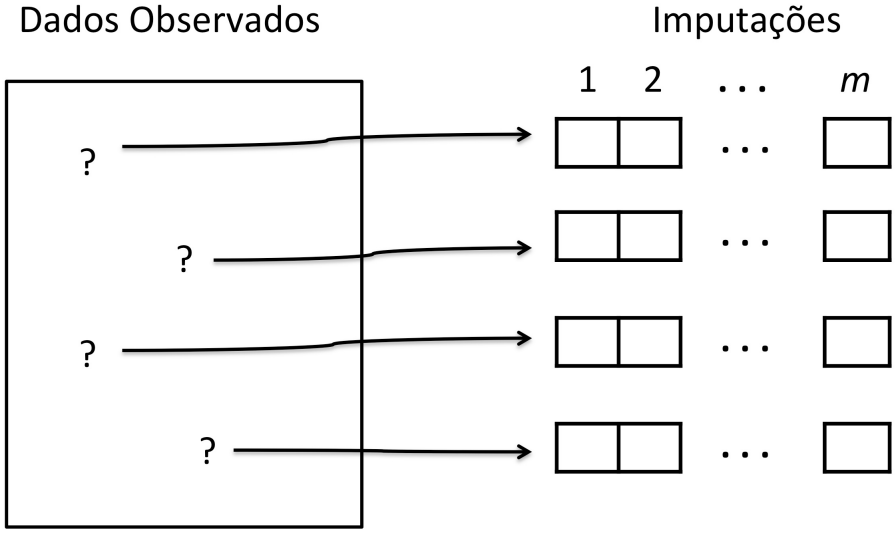

Figura 6.1: Esquema de Imputação Múltipla
Fonte: Adaptação de Lobato (2016)

ausente. Assim, os m resultados são combinados em um único resultado final. A Figura 6.1 mostra um esquema do método de imputação múltipla e o símbolo de interrogação representa um valor ausente, para o qual são gerados m valores.

Seja Y a coleção de dados definida por uma matriz $(n \times p)$ com os valores dos dados nas p-variáveis para todas as n unidades da amostra.

Y pode ser dividido em valores observados Y^O e valores que não foram observados Y^A, ou seja, $Y = (Y^O, Y^A)$. Sendo que Y^O contém todos os elementos y_{ij} observados e os valores de Y^A são desconhecidos. Considerar uma quantidade θ que, em geral, somente poderia ser calculada se a totalidade da população fosse observada. Denominamos essa quantidade de parâmetro. Por exemplo θ pode ser a média de uma variável aleatória na população, parâmetros de um modelo de regressão etc. Se estamos interessados em mais que uma quantidade, θ será um vetor.

O objetivo da IM é obter um estimador para θ, denotado por $\widehat{\theta}$, que satisfaça duas propriedades:

1. Seja não viciado, isto é:

$$E(\widehat{\theta}|Y) = \theta \tag{6.1}$$

2. Seja 'confidence valid': Seja U a matriz de variância e covariância estimada de $\widehat{\theta}$. Dizemos que $\widehat{\theta}$ é 'confidence valid' se

$$E(U|Y) \geq V(\widehat{\theta}|Y) \tag{6.2}$$

sendo que $V(\widehat{\theta}|Y)$ é a variância causada pelo processo de amostragem.

Por exemplo, no contexto dos testes de hipóteses, um teste estatístico com nível de significância $\alpha\%$ deve rejeitar a hipótese nula com probabilidade $\alpha\%$ dos casos em que essa hipótese é verdadeira. Se isso ocorre, diz-se que o procedimento é 'confidence valid'.

[1]

A incerteza sobre $\widehat{\theta}$ depende do quanto se conhece sobre os dados ausentes (Y^A) em uma amostra. Se fosse possível recriar perfeitamente esses dados ausentes, poderíamos calcular θ e não haveria necessidade de obter um estimador $\widehat{\theta}$. No entanto, isso não é possível.

Os possíveis valores de θ, dado o conhecimento que se tem dos dados observados Y^O são "capturados" pela distribuição a posteriori $P(\theta|Y^O)$. Essa distribuição pode ser intratável, mas pode-se decompô-la em dois fatores para a solução:

$$P(\theta|Y^O) = \int P(\theta|Y^O, Y^A)P(Y^A|Y^O)dY^A \tag{6.3}$$

sendo que $P(\theta|Y^O, Y^A)$ é a distribuição a posteriori de θ em um conjunto de valores hipoteticamente completo e $P(Y^A|Y^O)$ é a distribuição preditiva, calculada como na equação (2.3).

Supor que $P(Y^A|Y^O)$ seja utilizada para gerar imputações para Y^A, denotados por Y^{*A}. Pode-se utilizar $P(\theta|Y^O, Y^A)$ para calcular θ com o conjunto hipoteticamente completo (Y^O, Y^{*A}). A ideia é repetir esses dois passos até completar todos os dados ausentes.

A equação (6.3) sugere que a distribuição a posteriori verdadeira para θ é a média sobre todas os valores gerados de θ. A importância desse resultado é que a distribuição a posteriori $P(\theta|Y^O)$ pode ser complicada, mas as outras duas distribuições a posteriori $P(\theta|Y^O, Y^A)$ e $P(Y^A|Y^O)$ são mais simples e podem ser utilizadas para a geração das imputações. A média de $\theta|Y^O$ é dada por:

$$E(\theta|Y^O) = E[E(\theta|Y^O, Y^A)\backslash Y^O] \tag{6.4}$$

Este é o valor esperado da média a posteriori de θ sobre todos os valores imputados repetidamente. O resultado sugere seguir esse procedimento para combinar os resultados das m imputações.

Supor que $\widehat{\theta}_l$ é a estimativa da $l - ésima$ imputação, com $l = 1, ..., m$. A estimativa combinada é a média aritmética dos valores:

$$\overline{\theta} = \frac{1}{m}\sum_{l=1}^{m}\widehat{\theta}_l \tag{6.5}$$

A variância de $\theta|Y^O$ é calculada pela soma de dois componentes de variância:

$$V(\theta|Y^O) = E[V(\theta|Y^O, Y^A)\backslash Y^O] + V[E(\theta|Y^O, Y^A)\backslash Y^O] \tag{6.6}$$

sendo que $E[V(\theta|Y^O, Y^A)\backslash Y^O]$ é denominada de variância 'dentro' dos conjuntos completos e $V[E(\theta|Y^O, Y^A)\backslash Y^O]$ denominada de variância 'entre' os conjuntos completos.

Sejam $\overline{U}_\infty$ e B_∞ os componentes de variância 'dentro' e 'entre', considerando $m = \infty$, respectivamente. A variância total será

$$V_\infty(\theta|Y^O) = \overline{U}_\infty + B_\infty \tag{6.7}$$

A equação (6.6) sugere o seguinte procedimento para estimar $V_\infty(\theta|Y^O)$ para m finito:

[1] Na falta de uma boa tradução para o português, optou-se por utilizar a expressão em língua inglesa 'confidence valid'

Calcular

$$\overline{U} = \frac{1}{m}\sum_{l=1}^{m}\overline{U}_l \tag{6.8}$$

sendo $\overline{U}_l$ a matriz de variâncias e covariâncias de $\widehat{\theta}_l$.

O estimador não viciado da variância entre os m conjuntos completos é dado por:

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\widehat{\theta}_l - \overline{\theta})^2 \tag{6.9}$$

Assim, a variância total seria calculada por $V(\theta|Y^O) = \overline{U}+B$, mas é preciso incorporar o fato de que $\overline{\theta}$ é obtido usando m finito e portanto é uma aproximação de $\overline{\theta}_\infty$. Rubin (1987) mostra que a contribuição deste fator para a variância total é sistemática e igual a B_∞/m. Uma vez que B se aproxima de B_∞, pode-se escrever:

$$V(\theta|Y^O) = \overline{U} + B + \frac{B}{m} = \overline{U} + \left(1 + \frac{1}{m}\right)B \tag{6.10}$$

O procedimento para combinar os resultados das m imputações é referido como "Regras de Rubin". Resumindo, a variância total decorre de três fontes:

1. $\overline{U}$: Variância causada pelo fato de que se está utilizando uma amostra e não observando a população inteira. Esta é a medida convencional de variabilidade;
2. B: Variância extra, causada pelo fato de que existem valores ausentes na amostra;
3. B/m: Variância da simulação, causada pelo fato de θ é estimado com m finito.

A adição do último termo é importante para se fazer imputação múltipla com valores baixos de m. As escolhas tradicionais para m são $m = 3, m = 5$ e $m = 10$. Quanto maior m, menor será o efeito do erro de simulação na variância total. Von Hippel (2009) sugere que o número de imputações deve ser similar à porcentagem de dados ausentes. Essa regra se aplica principalmente nos casos em que essa porcentagem é maior que 50%.

Por outro lado, van Buuren (2012) afirma que uma das vantagens da imputação múltipla é que ela pode produzir estimativas não viciadas com intervalos de confiança corretos utilizando um número baixo de conjuntos de dados imputados, mesmo com $m = 2$. A recomendação é que se utilize m entre 3 e 5 quando há uma quantidade moderada de dados ausentes na amostra. Vários autores analisaram a influência de m em aspectos dos resultados e utilizam o argumento de que o investimento necessário para gerar, analisar e armazenar os resultados de um número muito grande de conjuntos imputados, isto é, utilizar $m \to \infty$, não compensa. Schafer (1997) e Schafer e Graham (2002), concluíram que em muitas situações há pouca vantagem no sentido de redução da variabilidade total.

A seguir apresenta-se o imputação pelo método MICE (*Multivariate Imputation by Chained Equations*) que pode também ser creditado como *Fully Conditional Specifications (FCS).*

6.2.1 *MICE*

O algoritmo MICE requer a especificação de um método de imputação univariado para cada variável a ser imputada. Entre os métodos disponíveis no software R destacamos o 'norm', utilizado para regressão linear bayesiana, 'norm.boot' para imputação normal com

bootstrap, o 'logreg' para regressão logística (dados binários) e o 'pmm' que é o método padrão no MICE para variáveis contínuas e que foi utilizado neste trabalho, van Buuren e Groothuis-Oudshoorn (2018) e White *et al.* (2011).

É necessário que o método de imputação leve em conta o procedimento que resultou nos dados ausentes e é de grande relevância que sejam preservadas as relações entre as variáveis e as incertezas sobre essas associações. Seja $Y = (Y_1, ..., Y_p)$ o conjunto de variáveis a ser analisado e Y_j, $j = 1, ..., p$, uma das variáveis incompletas. As partes observadas e não observadas de Y_j são Y_j^O e Y_j^A, respectivamente. Então tem-se $Y^O = (Y_1^O, ..., Y_p^O)$, o conjunto com as partes completas e $Y^A = (Y_1^A, ..., Y_p^A)$, o conjunto com as partes incompletas. O $h-$ésimo conjunto de dados imputados é denotado por Y^h,onde $h = 1, ..., m$. Seja $Y_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_p)$ e seja θ a quantidade de interesse (por exemplo o coeficiente de uma regressão ou um vetor com quaisquer parâmetros).

As estimativas $\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}, ..., \widehat{\theta}^{(m)}$ serão diferentes em cada conjunto imputado e é necessário compreender que essa diferença é devida à incerteza sobre qual valor imputar. Se $(\theta - \widehat{\theta})$ é aproximadamente normal, $N(0, U)$, calcula-se a média e a variância das estimativas:

$$\overline{\theta}_m = \frac{1}{m}\sum_{h=1}^{m}\widehat{\theta}^{(h)} \qquad \overline{U} = \frac{1}{m}\sum_{h=1}^{m}U^{(h)} \tag{6.11}$$

sendo $U^{(h)}$ a variância do $h-$ésimo conjunto de dados imputados. A variância entre as estimativas é dada por:

$$B_m = \frac{1}{m-1}\sum_{h=1}^{m}(\widehat{\theta}^{(h)} - \overline{\theta}_m)^2 \tag{6.12}$$

e finalmente, a variância total de $(\theta - \overline{\theta}_m)$ é calculada por:

$$T_m = \overline{U} + (1 + m^{-1})B_m. \tag{6.13}$$

Considera-se que Y segue uma distribuição $p-$variada $P(Y|\theta)$. Será assumido que a distribuição multivariada de Y seja completamente especificada por θ, um vetor de parâmetros desconhecidos. O problema é como obter a distribuição de θ. O algoritmo MICE obtém a distribuição a posteriori de θ gerando iterativamente das distribuições condicionais da forma

$$P(Y_1|Y_{-1}, \theta_1), ..., P(Y_p|Y_{-p}, \theta_p). \tag{6.14}$$

Os parâmetros $\theta_1, ..., \theta_p$ são específicos de cada distribuição condicional em (6.14) que não são necessariamente o produto de uma fatorização da verdadeira distribuição conjunta $P(Y|\theta)$. A $t-$ésima iteração das equações em cadeia é um amostrador de Gibbs, método de simulação de Monte Carlo em Cadeias de Markov (MCMC), apresentado no capítulo 2, que gera sucessivamente

$$\begin{aligned}
\theta_1^{*(t)} &\sim P(\theta_1|Y_1^O, Y_2^{(t-1)}, ..., Y_p^{(t-1)}) \\
Y_1^{*(t)} &\sim P(Y_1|Y_1^O, Y_2^{(t-1)}, ..., Y_p^{(t-1)}, \theta_1^{*(t)}) \\
&... \\
\theta_p^{*(t)} &\sim P(\theta_p|Y_p^O, Y_1^{(t)}, ..., Y_{p-1}^{(t)})
\end{aligned} \tag{6.15}$$

$$Y_p^{*(t)} \sim P(Y_p|Y_p^O, Y_1^{(t)}, ..., Y_p^{(t)}, \theta_p^{*(t)})$$

sendo que $Y_j^{(t)} = (Y_j^O, Y_j^{*(t)})$ é a $j-$ésima variável imputada na iteração t. Observe que as imputações anteriores $Y_j^{*(t-1)}$ somente entram em $Y_j^{*(t)}$ através das suas relações com as outras variáveis e não diretamente. A convergência portanto é bem rápida, com número de iterações entre 10 e 20, diferentemente de outros métodos de MCMC.

O nome equações em cadeia se refere ao fato de que o algoritmo MICE pode facilmente ser implementado como uma concatenação de procedimentos univariados para preencher conjuntos de dados incompletos. É importante notar que não há necessidade de obter a forma analítica das distribuições a posteriori. No software R, pacote mice 3.16.0, é possível implementar todos os passos da IM com o algoritmo MICE, site https://www.r-project.org.

6.2.2 *Predictive Mean Matching - PMM*

Este é o método padrão no MICE para variáveis contínuas e que foi utilizado neste trabalho. É um método versátil e fácil de se aplicar. É robusto com relação a transformações na variável Y, então imputar $\log(Y)$ frequentemente produz resultado similar a imputar $\exp(Y)$. O método também permite imputar variáveis discretas. O PMM também é robusto em situações em que há dificuldade na especificação de um modelo de imputação e sua performance é considerada melhor que as de alguns métodos teoricamente superiores. Para cada entrada ausente, o método forma um pequeno conjunto de "doadores" candidatos (em geral, são 1, 3 ou 10 membros) escolhidos de todos os casos completos que tem valores preditos próximos ao valor predito da entrada ausente. Um sorteio é feito entre os candidatos, e o valor selecionado dos "doadores" é usado para repor o valor ausente. A suposição feita é que o conjunto de valores ausentes segue a mesma distribuição dos candidatos. As imputações são baseadas nos valores observados, portanto são realistas. Imputações fora do conjunto observado não ocorrem, o que poderia levar a valores irreais (por exemplo, peso negativo). O modelo é implícito, o que significa que não há necessidade de definir um modelo explícito para a distribuição dos valores ausentes.

Várias métricas podem ser definidas para a distância entre os casos. A métrica do método PMM foi proposta por Rubin (1986) e Little (1988). Ela é particularmente útil para aplicações em dados ausentes porque é otimizada para cada variável Y separadamente.

7 Aplicações com dados reais

Neste capítulo são apresentadas duas aplicações com dados reais. Na primeira, um conjunto de dados completo foi utilizado, sendo retirados, aleatoriamente, valores para comparar os resultados dos métodos de imputação com o resultado se não houvesse dados ausentes. Com isso, foram utilizados métodos de imputação de dados ausentes que tiveram seus resultados comparados.

Na segunda, os dados de fato continham dados ausentes e a imputação foi feita apenas pelo método *Fully Conditional Specification* (FCS).

7.1 Comparação de métodos de imputação - Aplicação para dados de qualidade de água doce (Rio Piracicaba - SP)

Nesta aplicação são apresentados métodos e resultados de imputação de dados ausentes para um conjunto de dados de variáveis físicas relacionadas com a qualidade da água doce no Rio Piracicaba-SP, obtidos pelo monitoramento da CETESB (Companhia Ambiental do Estado de São Paulo), Secretaria de Infraestrutura e Meio Ambiente e Governo do Estado de São Paulo.

Todas as informações sobre o problema da qualidade das águas e os dados foram obtidas no site do CETESB (2025).

O objetivo é propor uma metodologia adequada para imputação desse tipo de dados. Para isso foram utilizados dados completos de quatro variáveis físicas. Para exemplificar os métodos de imputação, foram retirados valores nas quatro variáveis e a análise dos dados completos foi comparada com a análise feita com dados imputados.

O monitoramento da qualidade das águas superficiais em corpos d'água doce, como rios e reservatórios é constituído por três redes de amostragem manual e uma rede automática, objetivando um diagnóstico dos usos múltiplos do recurso hídrico. A seguir, um resumo sobre o significado ambiental e sanitário das variáveis físicas de qualidade das águas utilizadas neste trabalho, CETESB (2025).

1. Condutividade

A condutividade é a expressão numérica da capacidade da água conduzir a corrente elétrica. Depende das concentrações iônicas e da temperatura, indica a quantidade de sais existentes na coluna d' água e, portanto, representa uma medida indireta da concentração de poluentes. Em geral, níveis superiores a 100 $\mu S/cm$ indicam ambientes impactados. A condutividade também fornece uma boa indicação das modificações na composição de uma água, especialmente na sua concentração mineral, mas não fornece nenhuma indica-

ção das quantidades relativas dos vários componentes. A condutividade da água aumenta à medida que mais sólidos dissolvidos são adicionados. Altos valores podem indicar características corrosivas da água.

2. Série de Sólidos

Em saneamento, sólidos nas águas correspondem a toda matéria que permanece como resíduo, após evaporação, secagem ou calcinação da amostra a uma temperatura pré-estabelecida durante um tempo fixado. Para o recurso hídrico, os sólidos podem causar danos aos peixes e à vida aquática. Eles podem sedimentar no leito dos rios, destruindo organismos que fornecem alimentos ou, também, danificar os leitos de desova de peixes. Os sólidos podem reter bactérias e resíduos orgânicos no fundo dos rios, promovendo decomposição anaeróbia. Altos teores de sais minerais, particularmente sulfato e cloreto, estão associados á tendência de corrosão em sistemas de distribuição, além de conferir sabor ás águas.

3. Temperatura

Variações de temperatura são parte do regime climático normal e corpos de água naturais apresentam variações sazonais e diurnas, bem como estratificação vertical. A temperatura superficial é influenciada por fatores tais como latitude, altitude, estação do ano, período do dia, taxa de fluxo e profundidade. A elevação da temperatura em um corpo d' água geralmente é provocada por despejos industriais (indústrias canavieiras,por exemplo) e usinas termoelétricas.

A temperatura desempenha um papel crucial no meio aquático, condicionando as influências de uma série de variáveis físico-químicas. Em geral, a medida que a temperatura aumenta, de 0 a 30^oC, viscosidade, tensão superficial, compressibilidade, calor específico, constante de ionização e calor latente de vaporização diminuem, enquanto a condutividade térmica e a pressão de vapor aumentam. Organismos aquáticos possuem limites de tolerância térmica superior e inferior, temperaturas ótimas para crescimento, temperatura preferida em gradientes térmicos e limitações de temperatura para migração, desova e incubação do ovo.

4. Turbidez

A turbidez de uma amostra de água é o grau de atenuação de intensidade que um feixe de luz sofre ao atravessá-la (esta redução dá-se por absorção e espalhamento, uma vez que as partículas que provocam turbidez nas águas são maiores que o comprimento de onda da luz branca), devido a presença de sólidos em suspensão, tais como partículas inorgânicas (areia e argila) e detritos orgânicos, tais como algas e bactérias, plâncton em geral etc.

A erosão das margens dos rios em estações chuvosas, que é intensificada pelo mau uso do solo, é um exemplo de fenômeno que resulta em aumento da turbidez das águas que exige manobras operacionais, tais como alterações nas dosagens de coagulantes e auxiliares, nas Estações de Tratamento de Águas. Este exemplo mostra também o caráter sistêmico da poluição, ocorrendo inter-relações ou transferência de problemas de um ambiente (água, ar ou solo) para outro. Os esgotos domésticos e diversos efluentes industriais também provocam elevações na turbidez das águas.

Alta turbidez reduz a fotossíntese de vegetação enraizada submersa e algas. Esse desenvolvimento reduzido de plantas pode, por sua vez, suprimir a produtividade de peixes. Logo, a turbidez pode influenciar nas comunidades biológicas aquáticas. Além disso, afeta

adversamente os usos doméstico, industrial e recreacional de uma água.

O monitoramento feito pela CETESB inclui muitas outras variáveis para a avaliação da qualidade das águas. Neste trabalho, optou-se por aplicar o método para as quatro variáveis citadas e pretende-se estender a aplicação para as variáveis quantitativas mais relevantes.

7.1.1 Resultados

Nas Tabelas 7.1 e 7.2 estão apresentados os dados completos utilizados neste trabalho que incluem 42 coletas das variáveis: Local (Piracicaba, Limeira, Americana e Santa Maria); Coleta (6 datas de coletas); Condutividade, medida em $\mu S/cm$; Sólidos, medida em mg/L; Temperatura, medida em graus centígrados e Turbidez, medida em UNT.

Tabela 7.1: Dados de qualidade da Água - Rio Piracicaba

Local	Coleta	Condutiv.	Sólidos	Temperatura	Turbidez
Piracicaba 1	1	150	290	25,1	170
Piracicaba 1	2	237	232	27,6	100
Piracicaba 1	3	352	188	23,6	11
Piracicaba 1	4	420	278	17,9	6
Piracicaba 1	5	471	362	20,4	5,8
Piracicaba 1	6	277	277	23,6	40
Piracicaba 2	1	156	236	25,1	130
Piracicaba 2	2	261	181	28,3	24
Piracicaba 2	3	377	194	23,7	7
Piracicaba 2	4	493	306	18,5	6,5
Piracicaba 2	5	498	293	21	5,8
Piracicaba 2	6	319	251	25,1	40
Piracicaba 3	1	163	133	23,6	80
Piracicaba 3	2	247	150	26,5	34
Piracicaba 3	3	387	206	22	4,5
Piracicaba 3	4	489	303	17,1	4,5
Piracicaba 3	5	503	301	21,2	4,9
Piracicaba 3	6	299	184	23,5	85
Piracicaba 4	1	158	273	25,2	120
Piracicaba 4	2	262	226	28,4	24
Piracicaba 4	3	383	211	24	5,9
Piracicaba 4	4	489	319	18,7	5,4
Piracicaba 4	5	501	294	21,5	4,7
Piracicaba 4	6	324	218	24,6	20

Fonte: Companhia Ambiental do Estado de São Paulo - CETESP, 2021

Tabela 7.2: Dados de qualidade da Água - Rio Piracicaba

Local	Coleta	Condutiv.	Sólidos	Temperatura	Turbidez
Limeira	1	186	154	24	45
Limeira	2	225	156	25,7	37
Limeira	3	392	306	22,9	8,9
Limeira	4	551	352	17,5	13
Limeira	5	535	331	20,3	11
Limeira	6	439	260	24,2	7,9
Americana	1	135	161	24,2	50
Americana	2	176	145	27,4	22
Americana	3	206	214	23,4	5,7
Americana	4	268	193	17,9	8,1
Americana	5	290	179	20,5	5,1
Americana	6	237	158	24,7	10
Santa Maria	1	224	146	26,1	15
Santa Maria	2	218	124	30	17
Santa Maria	3	196	89	24,8	2
Santa Maria	4	249	174	19,7	11
Santa Maria	5	300	220	20	23
Santa Maria	6	278	182	24,5	9,7

Fonte: Companhia Ambiental do Estado de São Paulo - CETESP, 2021

A Tabela 7.3 apresenta as estatísticas do conjunto de dados completo.

Tabela 7.3: Estatísticas com o conjunto de dados completo

Variável	n	Média	$IC(\mu, 95\%)$	Variância	Dp
Condutiv.	42	317,17	(278,70 ; 355,63)	15239,07	123,45
Sólidos	42	225	(203,70 ; 246,30)	4672,05	68,35
Temperat.	42	23,19	(22,19 ; 24,18)	10,17	3,19
Turbidez	42	29,53	(17,43 ; 41,64)	1508,82	38,84

Resultados da análise com dados completos

Para efeito de comparação de resultados dos métodos de imputação, um modelo de regressão linear múltipla foi ajustado aos conjuntos de dados completo e também, aos conjuntos de dados imputados considerando: imputação por média, os métodos *listwise*, *pairwise* e a imputação múltipla. Com as variáveis denotadas por Y : Condutividade (variável resposta), X_1 : Sólidos, X_2 : Temperatura e X_3 : Turbidez (variáveis explicativas), considerar o modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{7.1}$$

O modelo ajustado para os dados completos resultou em $R^2 = 0,8249$. As variáveis X_1 : Sólidos e X_3 : Turbidez são significantes ao nível de significância de 5% como pode ser visto na Tabela 7.4 que contém as estimativas dos coeficientes, desvio-padrão dos estimadores e o valor do nível descritivo, *valor p*, para cada uma das variáveis explicativas.

Tabela 7.4: Regressão com o conjunto de dados completo

Variável	$\widehat{\beta}$	$dp(\widehat{\beta})$	*valor p*
Intercepto	$172,0443$	$99,2896$	$0,0912$
Sólidos	$1,2169$	$0,1469$	0
Temperatura	$-3,38$	$3,3770$	$0,3093$
Turbidez	$-1,6242$	$0,2294$	0

Para aplicar os métodos de imputação, 23 observações foram retiradas de forma totalmente aleatória, assim considera-se que o mecanismo de ausência dos dados é do tipo MCAR. O padrão das ausências de dados pode ser visto na Figura 7.1. Os números no lado esquerdo indicam o número de linhas, os números do lado direito indicam o número de variáveis nessas linhas, com dados ausentes (em vermelho). Os números abaixo indicam o número total de valores ausentes para cada variável.

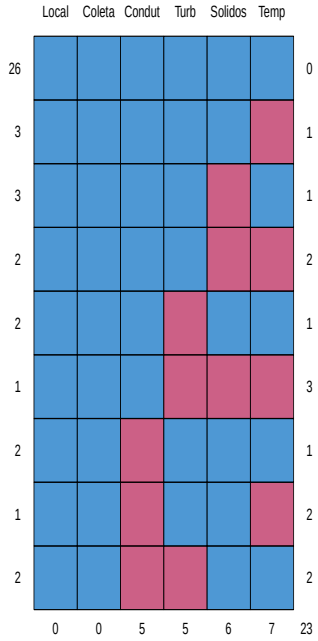

Figura 7.1: Padrão de ausência dos dados
Fonte: Elaborado pela autora (2024)

Resultados da análise com imputação por média

O primeiro método aplicado aos dados com 23 observações ausentes foi a imputação por média. Na Figura 7.2 pode-se ver os valores das médias em vermelho. Na Figura 7.3 pode-ser ver os histogramas dos dados imputados para Condutividade e Sólidos, à esquerda e à direita, respectivamente. Observa-se que como a imputação por média repete o valor da média em todos os valores ausentes, resulta em uma moda artificial na distribuição dessas variáveis além de diminuir artificialmente a variância.

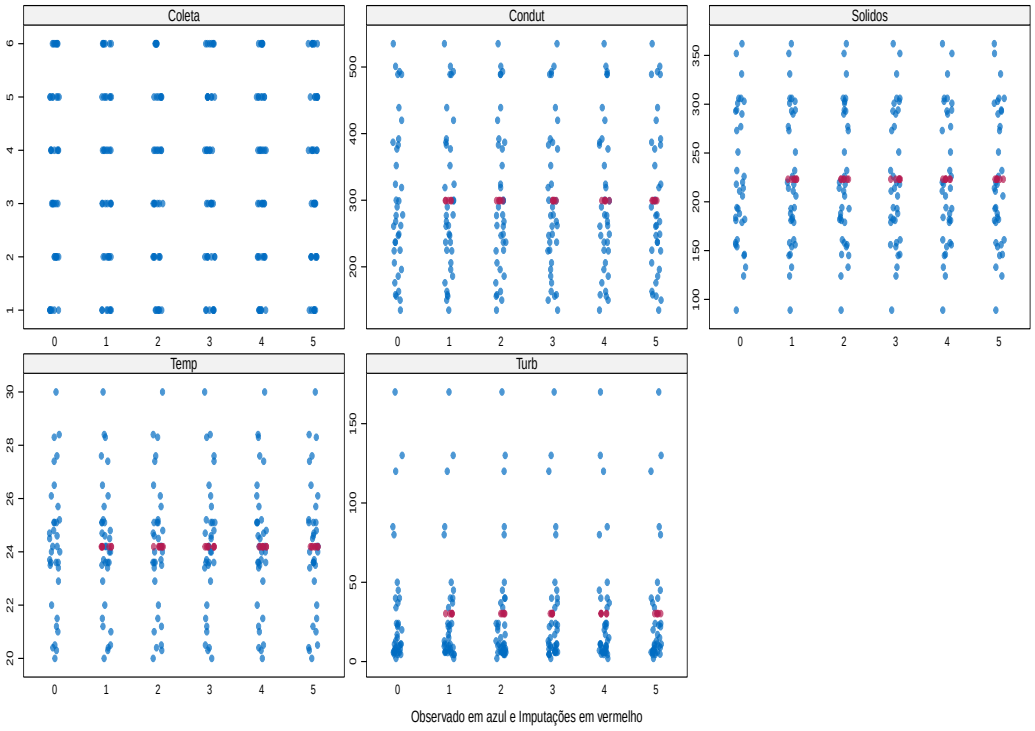

Figura 7.2: Imputação por média - Todas variáveis
Fonte: Elaborado pela autora (2024)

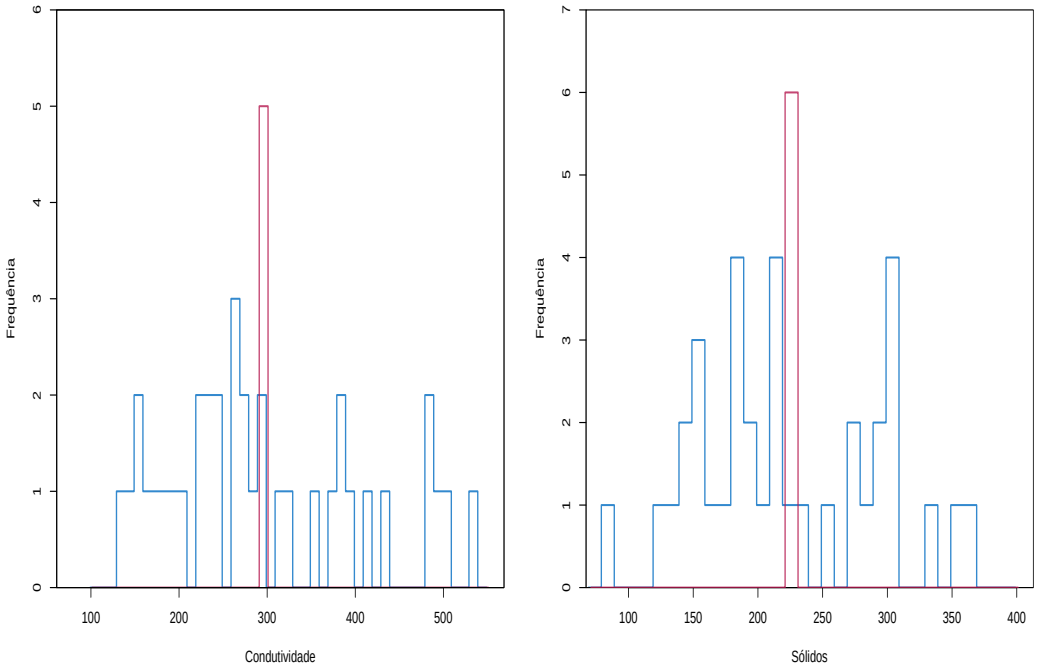

Figura 7.3: Condutividade e Sólidos
Fonte: Elaborado pela autora (2024)

A Tabela 7.5 apresenta as estatísticas do conjunto de dados após a imputação por média.

Tabela 7.5: Estatísticas com a imputação por média

Variável	n	Média	$IC(\mu, 95\%)$	Variância	Dp
Condutividade	42	299,46	(266,53 ; 332,39)	11165,6	105
Sólidos	42	223,14	(203,06 ; 243,22)	4151,5	64,43
Temperatura	42	24,19	(23,5 ; 24,9)	2,92	2,22
Turbidez	42	30,25	(18,88 ; 41,61)	1329,9	36,47

O modelo ajustado para os dados imputados por média resultou em $R^2 = 0,4683$. As variáveis X_1 : Sólidos e X_3 : Turbidez são significantes ao nível de significância de 5% como pode ser visto na Tabela 7.6 que contém as estimativas dos coeficientes, desvio-padrão dos estimadores e o valor do nível descritivo, *valor p*, para cada uma das variáveis explicativas. Na Figura 7.4 temos os gráficos de dispersão e retas ajustadas segundo o modelo descrito na Figura 7.1, com imputação por média.

Tabela 7.6: Regressão com a imputação por média

Variável	$\widehat{\beta}$	$dp(\widehat{\beta})$	*valor p*
Intercepto	305,6955	172,9455	0,0852
Sólidos	0,6949	0,2089	0,002
Temperatura	-4,759	6,1155	0,4413
Turbidez	-1,5266	0,3343	0

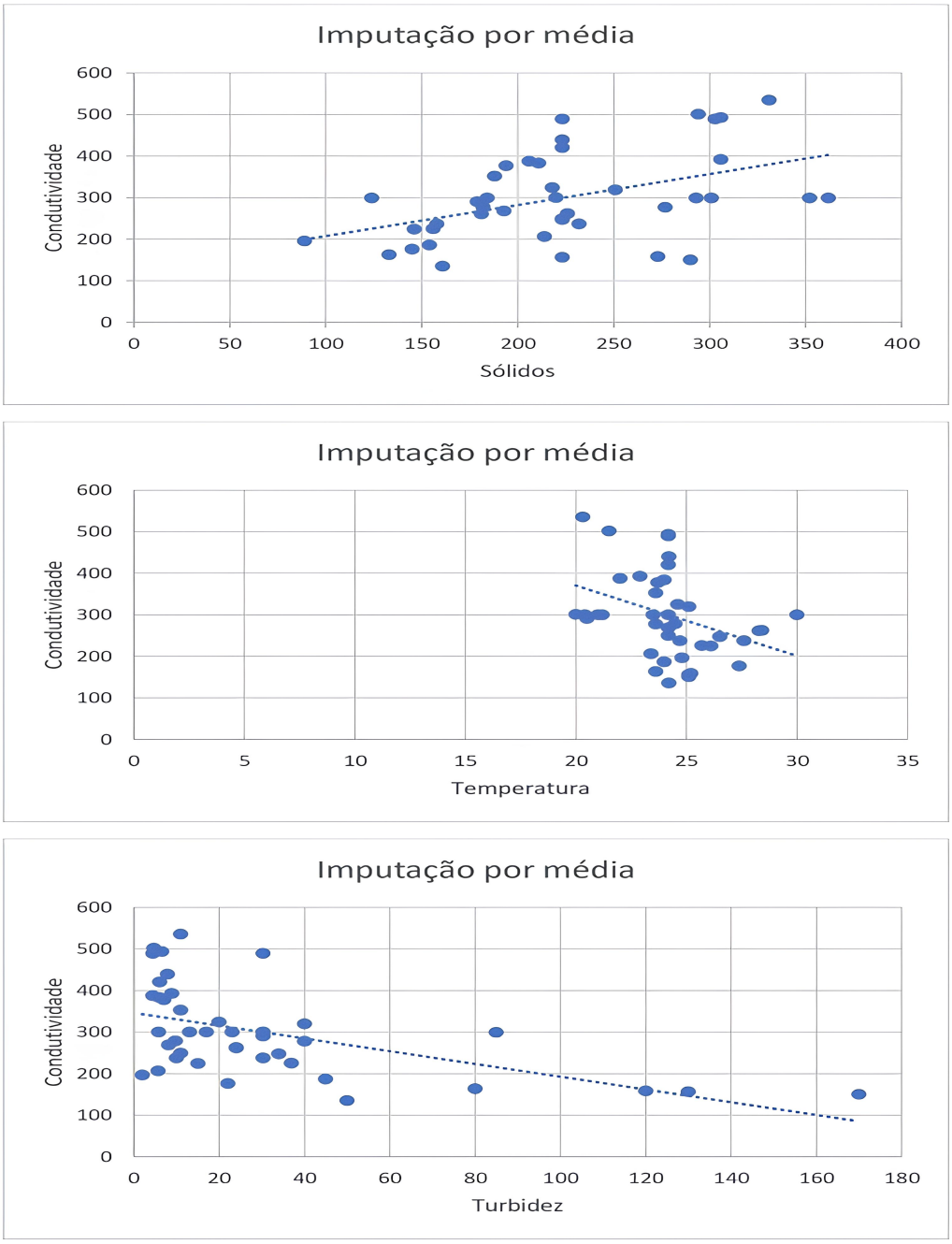

Figura 7.4: Regressão com Imputação por média
Fonte: Elaborado pela autora (2024)

Resultados da análise com *Listwise*

O modelo ajustado para os dados por *Listwise* resultou em $R^2 = 0,7134$. A Tabela 7.7 apresenta as estatísticas do conjunto de dados após a imputação por *Listwise*. As variáveis X_1 : Sólidos e X_3 : Turbidez são significantes ao nível de significância de 5% como pode ser visto na Tabela 7.8 que contém as estimativas dos coeficientes, desvio-padrão dos estimadores e o valor do nível descritivo, *valor p*, para cada uma das variáveis explicativas.

Tabela 7.7: Estatísticas com *Listwise*

Variável	n	Média	$IC(\mu, 95\%)$	Variância	Dp
Condutividade	26	280,88	(238,55 ; 323,22)	10985,95	104,81
Sólidos	26	207,23	(182,89 ; 231,57)	3631,78	60,26
Temperatura	26	24,24	(23,42 ; 25,06)	4,13	2,0318
Turbidez	26	33,67	(17,51 ; 49,83)	1600,52	40

Tabela 7.8: Regressão com *Listwise*

Variável	$\widehat{\beta}$	$dp(\widehat{\beta})$	*valor p*
Intercepto	434,4719	171,4910	0,0189
Sólidos	0,9867	0,2104	0,0001
Temperatura	-12,6566	6,2034	0,0535
Turbidez	-1,5234	0,2954	0

Resultados da análise com *Pairwise*

O modelo ajustado para os dados por *Pairwise* resultou em $R^2 = 0,7393$. A Tabela 7.9 apresenta as estatísticas do conjunto de dados após a imputação por *Pairwise*. As variáveis X_1 : Sólidos e X_3 : Turbidez são significantes ao nível de significância de 5% como pode ser visto na Tabela 7.10 que contém as estimativas dos coeficientes , desvio-padrão dos estimadores e o valor do nível descritivo, *valor p*, para cada uma das variáveis explicativas.

Tabela 7.9: Estatísticas com *Pairwise*

Variável	n	Média	$IC(\mu, 95\%)$	Variância	Dp
Condutividade	37	299,46	(261,86 ; 337,06)	12716,37	112,77
Sólidos	36	223,14	(199,54 ; 246,73)	4863,21	69,74
Temperatura	35	24,19	(23,35 ; 25,03)	5,94	2,44
Turbidez	37	30,25	(17,27 ; 43,22)	1514,66	38,92

Tabela 7.10: Regressão com *Pairwise*

Variável	$\widehat{\beta}$	$dp(\widehat{\beta})$	*valor p*
Intercepto	304,9032	153,8721	0,0578
Sólidos	0,9194	0,1852	0
Temperatura	-6,6984	5,3559	0,2218
Turbidez	-1,6049	0,2897	0

Resultados da análise com IM

Na Figura 7.5 pode-se ver os valores imputados em vermelho. O modelo ajustado para os dados com imputação múltipla resultou em $R^2 = 0,8527$. As variáveis X_1 : Sólidos, X_2 : Temperatura e X_3 : Turbidez são significantes ao nível de significância de 5% como pode ser visto na Tabela 7.11 que contém as estimativas dos coeficientes, desvio-padrão dos estimadores e o valor do nível descritivo, *valor p*, para cada uma das variáveis explicativas.

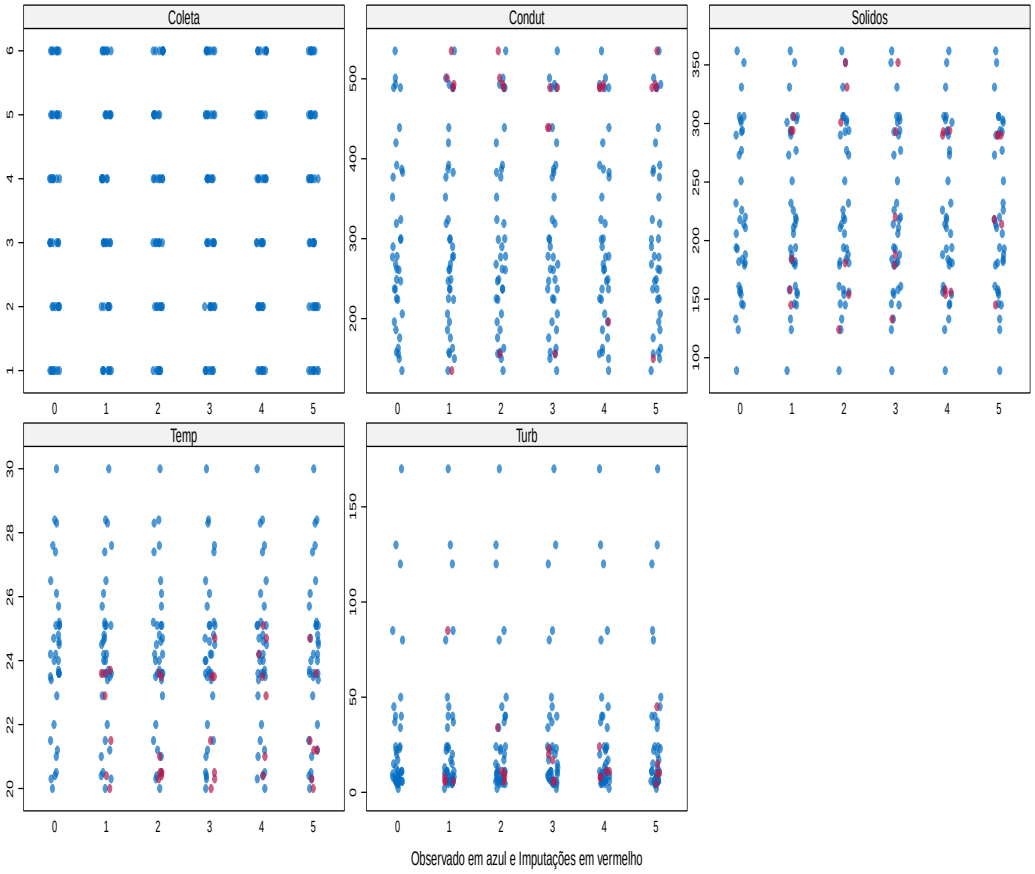

Figura 7.5: Imputação múltipla - Todas variáveis
Fonte: Elaborado pela autora

Tabela 7.11: Regressão com *IM*

Variável	$\widehat{\beta}$	$dp(\widehat{\beta})$	*valor p*
Intercepto	369,2770	104,4439	0,017
Sólidos	1,1252	0,1254	0
Temperatura	-11,0021	3,8637	0,0071
Turbidez	-1,5657	0,2055	0

Comparação dos resultados

Uma forma de comparar a qualidade das estimativas por intervalos para os parâmetros do modelo de regressão é verificar as amplitudes dos intervalos de confiança, obtidos utilizando a normalidade assintótica dos estimadores de máxima verossimilhança. Os intervalos muito amplos não são adequados pois são pouco precisos. Como apresentado nas Tabelas 7.12 até 7.15, observa-se que com o método de imputação por média, as estimativas pontuais dos parâmetros do modelo se distanciam das estimativas do modelo ajustado com dados completos. Nas estimativas por intervalos pode-se constatar que com esse método de estimação temos amplitudes dos intervalos muito maiores do que com os dados completos, o que significa que com esse método os intervalos são menos informativos, ou seja, menos úteis.

Tabela 7.12: Intervalos de Confiança e Amplitudes - Dados Completos

Variável	$\widehat{\beta}$	$IC(\beta, 95\%)$	Amplitude do IC	$R^2 = 82,49\%$
Intercepto	$172,0443$	$(-26,53492; 370,6235)$	$397,15844$	
Sólidos	$1,2169$	$(0,92316; 1,51064)$	$0,58748$	
Temperatura	$-3,48$	$(-10,23398; 3,27398)$	$13,5079$	
Turbidez	$-1,6242$	$(-2,08292; -1,16548)$	$0,91744$	

Tabela 7.13: Intervalos de Confiança e Amplitudes - Dados Imputados pela média

Variável	$\widehat{\beta}$	$IC(\beta, 95\%)$	Amplitude do IC	$R^2 = 46,83\%$
Intercepto	$305,6955$	$(-40,1955; 651,5865)$	$691,782$	
Sólidos	$0,6949$	$(0,2771; 1,1127)$	$0,8356$	
Temperatura	$-4,759$	$(-16,99; 7,472)$	$24,462$	
Turbidez	$-1,5266$	$(-2,1952; -0,858)$	$1,3372$	

Tabela 7.14: Intervalos de Confiança e Amplitudes - Dados Imputados por *Listwise*

Variável	$\widehat{\beta}$	$IC(\beta, 95\%)$	Amplitude do IC	$R^2 = 71,34\%$
Intercepto	$434,4719$	$(91,4899; 777,4539)$	$685,964$	
Sólidos	$0,9867$	$(0,5659; 1,4075)$	$0,8416$	
Temperatura	$-12,6566$	$(-25,0634; -0,2498)$	$24,8136$	
Turbidez	$-1,5234$	$(-2,1142; -0,9326)$	$1,1816$	

Tabela 7.15: Intervalos de Confiança e Amplitudes - Dados Imputados por *Pairwise*

Variável	$\widehat{\beta}$	$IC(\beta, 95\%)$	Amplitude do IC	$R^2 = 71,03\%$
Intercepto	$304,9032$	$(-2,861; 612,6674)$	$615,5284$	
Sólidos	$0,9194$	$(0,549; 1,2898)$	$0,7408$	
Temperatura	$-6,6984$	$(-17,4102; 4,0134)$	$21,4236$	
Turbidez	$-1,6049$	$(-2,1843; -1,0255)$	$1,1588$	

No caso da variável "Sólidos" o intervalo de confiança é $(0,2771; 1,1127)$ e nem sequer inclui o valor da estimativa pontual calculada com os dados completos $(1,2169)$. Além disso o R^2 nesse caso é bem menor do que com dados completos.

Ao comparar os resultados do método *listwise* (Tabela 7.14) com os resultados com dados completos, também observa-se que as estimativas pontuais se distanciam das estimativas do modelo ajustado com dados completos e nas estimativas por intervalos vê-se resultados melhores do que com imputação por média, mas com amplitudes dos intervalos muito grandes também, em comparação com as amplitudes dos intervalos com dados completos.

Com o método *pairwise* (Tabela 7.15) observa-se também grandes amplitudes para os intervalos. Com o método de IM, obtêm-se resultados melhores, nas estimativas pontuais e, principalmente nas estimativas por intervalos dos parâmetros do modelo de regressão. Todos os intervalos contém as estimativas pontuais obtidas com os dados completos e o $R^2 = 82,13\%$ é o mais próximo do obtido com dados completos.

Pode-se concluir que o método de imputação múltipla (Tabela 7.16) fornece resultados mais próximos dos resultados com os dados completos e recomenda-se a sua utilização.

Tabela 7.16: Intervalos de Confiança e Amplitudes - Dados Imputados por IM

Variável	$\widehat{\beta}$	$IC(\beta, 95\%)$	Amplitude do IC	$R^2 = 82,13\%$
Intercepto	369.277	$(150,3892; 588,1648)$	$437,7756$	
Sólidos	$1,1252$	$(0,8744; 1,376)$	$0,5016$	
Temperatura	$-11,0021$	$(-18,7295; -3,2747)$	$15,4548$	
Turbidez	$-1,5657$	$(-1,9767; -1,1547)$	$0,822$	

7.2 Aplicação para dados de serpentes

O Brasil ocupa a terceira posição no quesito países com maior diversidade de répteis, contando com 848 espécies nativas em que 430 são serpentes, sendo que, aproximadamente 70 são peçonhentas inseridas em duas famílias (Viperidae e Elapidae), Costa *et al.* (2021). A família Viperidae é formada por serpentes peçonhentas com dentição solenóglifa (ou seja, possuem presas retráteis no maxilar superior) e glândula de veneno, Roberto (2016). No Brasil, essa família é representada por três gêneros: *Crotalus* (cascavéis), *Bothrops* (jararacas) e *Lachesis* (surucucu).

A manutenção de serpentes em cativeiro é uma alternativa para a obtenção do veneno para pesquisas imunológicas e produção de soro antiofídico, sendo que esse trabalho iniciou-se no Brasil em 1901 com Vital Brazil, no Instituto Butantan que hoje é um dos principais centros produtores de vacinas e soros no Brasil, Cunha (2017).

A aplicação da imputação múltipla foi feita para dados longitudinais de vinte cascavéis (*Crotalus durissus terrificus*) de ambos os sexos e naturalmente parasitadas para a realização do experimento.

Os animais foram divididos em dois grupos. Grupo 1 (G1): composto por 5 machos e 5 fêmeas que foram vermifugados e Grupo 2 (G2): também composto por 5 machos e 5 fêmeas, mas que não receberam vermífugo.

Os animais do G1 e G2 passaram por coletas periódicas de sangue a cada 45 dias. A cada coleta de amostra sanguínea foi retirado um volume de, no máximo, 1% do peso do animal. Uma alíquota de sangue foi acondicionada em tubo de ensaio plástico, sem anticoagulante, para a obtenção de soro para avaliar alguns parâmetros bioquímicos (glicose, fósforo, ácido úrico, colesterol e hemoglobina) e outra alíquota em tubo de ensaio plástico heparinizado para avaliação dos parâmetros hematológicos (contagem total de hemácias, leucócitos e trombócitos).

O método de imputação múltipla foi aplicado para os dados das oito variáveis medidas. O mecanismo de perda considerado foi o MCAR e tais dados foram organizados no formato *long*. A seguir apresenta-se uma breve descrição das Variáveis:

1. Hemácias: Células sanguíneas, responsáveis pela condução de oxigênio para os tecidos.

2. Leucócitos: Conjunto de células sanguíneas, denominadas células brancas. Atuam nos processos inflamatórios e imunológicos.

3. Trombócitos: Células sanguíneas, análogas as "plaquetas" dos mamíferos. Participam dos processos hemostáticos (mecanismo de coagulação) e auxiliam as células brancas em processos inflamatórios e/ou infecciosos.

4. Colesterol: Crucial para a saúde celular, servindo como precursor de hormônios esteroides (como estrogénios) e ácidos biliares, essenciais para a digestão. Além disso, é um componente vital das membranas celulares, conferindo estabilidade e fluidez, e está envolvido no transporte de substâncias e sinalização celular.

5. Hemoglobina: Proteína existente no interior dos eritrócitos e no plasma, cuja principal função é o transporte de oxigênio.

6. Glicose: O papel da glicose no organismo das serpentes é o mesmo que em outros animais: fornecer energia para as funções vitais e processos fisiológicos. Sendo um

monossacarídeo e principal combustível biológico, a glicose é absorvida e convertida em energia utilizável por todas as células. A concentração sanguínea de glicose em répteis varia conforme a espécie, estado nutricional e condições ambientais.

7. Fósforo: Vital para a formação e manutenção dos ossos e dentes, o metabolismo energético (produção de ATP), a composição das membranas celulares (fosfolípidios) e dos ácidos nucleicos (DNA e RNA) e para o equilíbrio ácido-base e osmótico dos fluidos corporais. Essencial para o crescimento e a diferenciação celular, o fósforo atua em todos os ciclos de vida, sendo crucial para a regeneração de tecidos e a manutenção geral da saúde.

8. Ácido úrico (AU): Produto primário final do catabolismo de proteínas. Em serpentes e outros répteis, o ácido úrico serve como um resíduo nitrogenado que é excretado como uma substância pastosa, o que ajuda a conservar água e a proteger o embrião no ovo. Ele é o principal produto final do metabolismo de proteínas nestes animais, sendo fundamental para a sua sobrevivência devido à sua baixa toxicidade e à pouca quantidade de água necessária para a sua eliminação.

7.2.1 Resultados

A imputação múltipla foi feita para as oito variáveis nas coletas com valores ausentes. Esse procedimento foi realizado m vezes, obtendo assim m conjuntos de dados completos com os quais pode-se aplicar os métodos estatísticos adequados para análise dos dados e combinar os resultados em um único resultado final. De acordo com a literatura utiliza-se entre 3 e 10 conjuntos imputados e nessa aplicação optou-se por $m = 5$.

Como o objetivo deste trabalho foi mostrar os métodos de imputação de dados ausentes e principalmente o método de imputação múltipla e suas vantagens, nesta aplicação não foi feita a análise estatística dos dados após a imputação. Os resultados das imputações estão apresentados em tabelas para cada variável, nas quais as médias dos $m = 5$ valores imputados estão destacados em negrito. Para melhor visualização foram feitos gráficos dos perfis médios completos.

Algumas variáveis apresentam alta variabilidade entre os valores dos indivíduos e/ou coletas. Observa-se que essa variabilidade já ocorria nos dados observados, de forma que o método de imputação não aumentou nem diminuiu essas variâncias, como esperado.

7.2.2 Perfis médios do Grupo 1 - Indivíduos tratados com vermífugo

Nas medidas da variável hemácias, ocorreram 5 valores ausentes, nas coletas 2, 3, 4, 5 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas. A Tabela 7.17 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.17: Contagens de hemácias em indivíduos tratados

Hemácias	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	750.000	460.000	380.000	220.000	**524.000**	600.000	510.000
Cdt 9	650.000	680.000	300.000	**490.000**	640.000	530.000	370.000
Cdt 10	510.000	700.000	670.000	580.000	670.000	310.000	430.000
Cdt 11	430.000	360.000	**384.000**	490.000	450.000	630.000	450.000
Cdt 29	690.000	**580.000**	440.000	610.000	320.000	**528.000**	480.000

Na Figura 7.6 são apresentados os perfis médios dos números de hemácias.

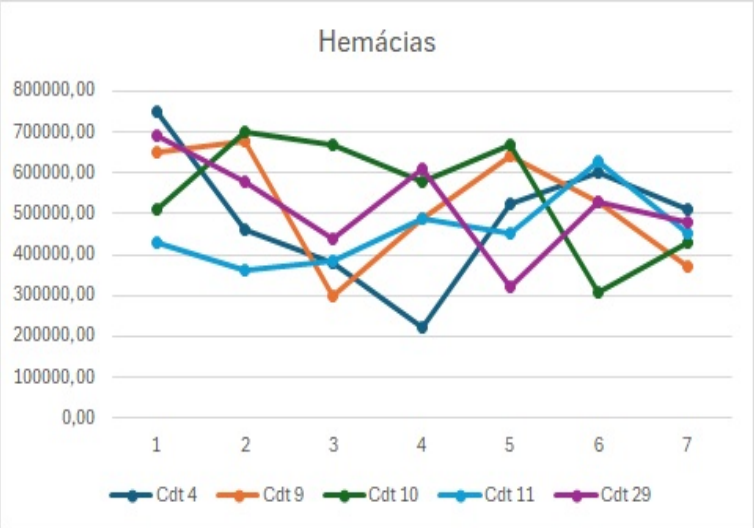

Figura 7.6: Perfis médios - Contagens de hemácias em indivíduos tratados

Nas medidas da variável leucócitos, ocorreram 5 valores ausentes, nas coletas 2, 3, 4, 5 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas. Nota-se que a variabilidade das medidas entre indivíduos é maior nas coletas 1 e 7, nas quais não houveram valores ausentes. Os valores imputados pelo método de imputação múltipla não estão fora do intervalo de valores observados, como espera-se com esse método. A Tabela 7.18 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.18: Contagens de leucócitos em indivíduos tratados

Leucócitos	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	13.500	9.000	4.500	5.500	**9.900**	20.000	18.000
Cdt 9	20.000	15.000	6.500	**6.500**	14.000	16.500	11.500
Cdt 10	5.000	10.000	12.000	6.500	10.000	10.000	8.000
Cdt 11	11.500	3.000	**5.300**	8.000	8.500	13.000	30.000
Cdt 29	14.000	**12.000**	3.500	10.000	13.000	**18.600**	25.000

Na Figura 7.7 são apresentados os perfis médios dos números de leucócitos.

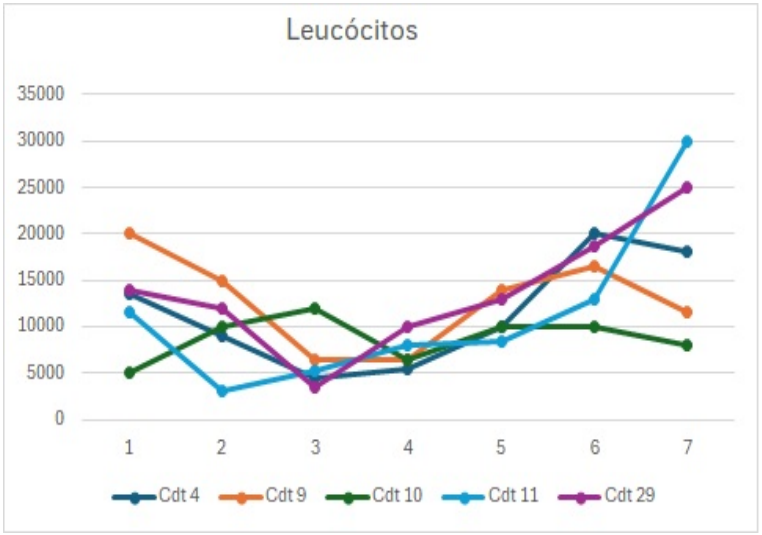

Figura 7.7: Perfis médios - Contagens de leucócitos em indivíduos tratados

Nas medidas da variável trombócitos, ocorreram 5 valores ausentes, nas coletas 2, 3, 4, 5 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas. A Tabela 7.19 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.19: Contagens de trombócitos em indivíduos tratados

Trombócitos	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	7.500	9.000	4.000	3.000	**6.800**	11.500	8.000
Cdt 9	13.000	10.500	5.500	**3.500**	7.000	9.500	9.000
Cdt 10	2.000	8.000	11.000	7.000	6.500	8.500	3.000
Cdt 11	5.500	1.000	**5.600**	3.500	6.500	8.500	17.000
Cdt 29	6.500	**9.000**	5.000	3.500	7.000	**9.300**	15.000

Na Figura 7.8 são apresentados os perfis médios dos números de trombócitos.

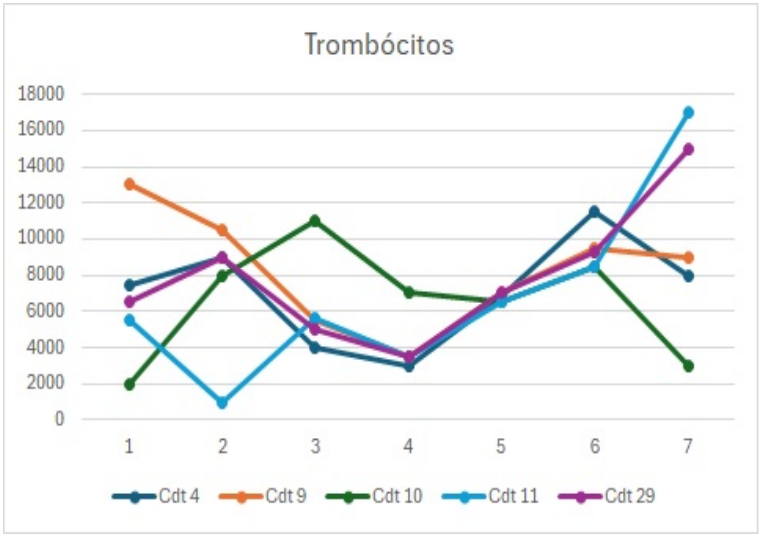

Figura 7.8: Perfis médios - Contagens de trombócitos em indivíduos tratados

Nas medidas da variável hemoglobina, ocorreram apenas 2 valores ausentes, nas coletas 5 e 6. Os indivíduos Cdt9, Cdt10 e Cdt11 tiveram todas as coletas feitas. A Tabela 7.20 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.20: Concentrações de hemoglobina em indivíduos tratados

Hemoglobina	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	7,90	5,00	5,10	4,70	**5,58**	5,10	4,00
Cdt 9	6,90	6,10	5,20	4,40	4,50	5,40	4,70
Cdt 10	10,50	7,60	8,60	9,20	5,50	4,80	1,70
Cdt 11	7,20	5,70	6,80	5,00	6,20	2,90	3,90
Cdt 29	6,60	6,60	8,20	4,50	4,10	**5,16**	4,60

Na Figura 7.9 são apresentados os perfis médios dos níveis de hemoglobina.

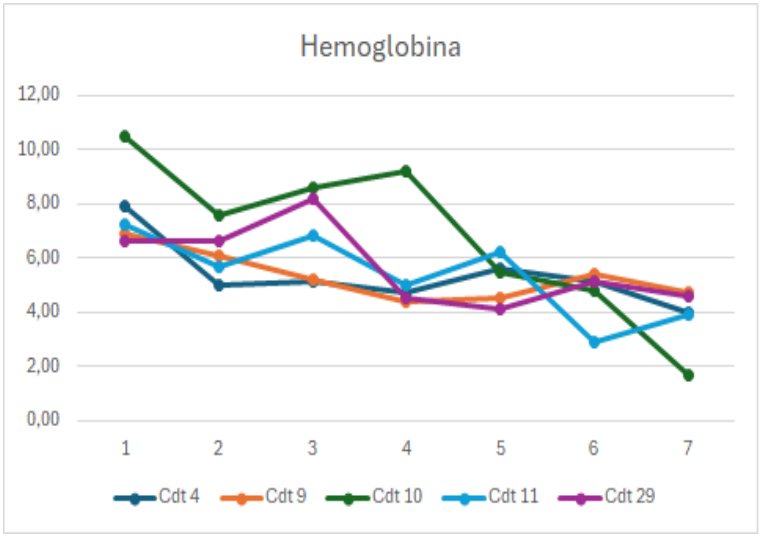

Figura 7.9: Perfis médios - Concentrações de hemoglobina em indivíduos tratados

Nas medidas da variável glicose, ocorreram 4 valores ausentes, nas coletas 3, 4 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas. A Tabela 7.21 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.21: Concentrações de glicose em indivíduos tratados

Glicose	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	36,03	33,53	80,33	**27,31**	23,55	40,91	18,80
Cdt 9	26,56	31,24	51,45	**26,01**	25,51	46,32	40,39
Cdt 10	24,75	29,46	47,62	25,40	27,62	66,20	20,77
Cdt 11	31,51	48,24	**55,69**	29,42	45,83	46,26	16,67
Cdt 29	49,68	39,09	26,21	26,93	57,25	**41,99**	17,82

Na Figura 7.10 são apresentados os perfis médios dos níveis de glicose.

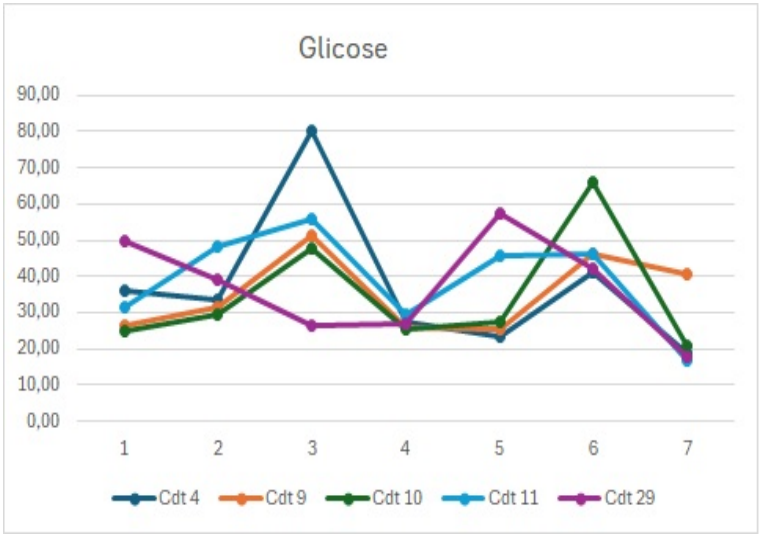

Figura 7.10: Perfis médios - Concentrações de glicose em indivíduos tratados

Nas medidas da variável fósforo, ocorreram 4 valores ausentes, nas coletas 3, 4 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas. A Tabela 7.22 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.22: Concentrações de fósforo em indivíduos tratados

Fósforo	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	3,30	2,90	2,64	**2,19**	2,35	3,46	2,98
Cdt 9	3,50	2,80	1,98	**2,69**	2,59	3,42	1,80
Cdt 10	3,00	1,50	2,09	2,07	2,07	3,10	2,10
Cdt 11	0,50	2,14	**1,72**	2,91	2,45	3,10	1,62
Cdt 29	1,10	1,74	1,28	1,85	2,45	**3,10**	3,31

Na Figura 7.11 são apresentados os perfis médios dos níveis de fósforo.

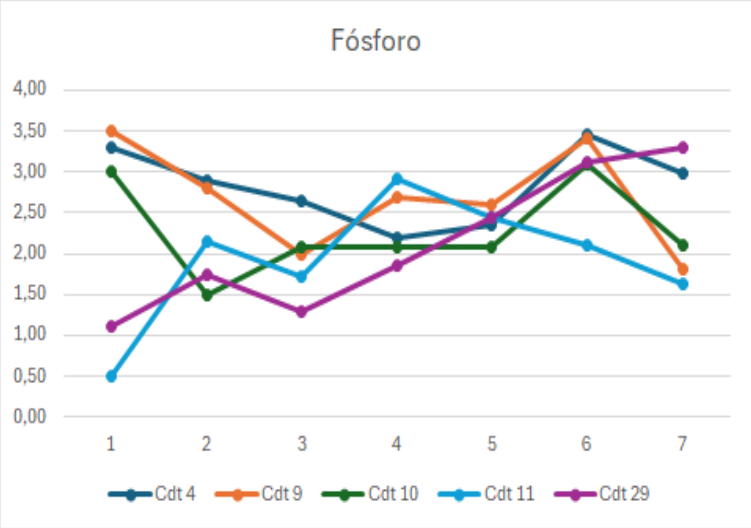

Figura 7.11: Perfis médios - Concentrações de fósforo em indivíduos tratados

Nas medidas da variável ácido úrico, ocorreram 4 valores ausentes, nas coletas 3, 4 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas.A Tabela 7.23 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.23: Concentrações de ácido úrico (AU) em indivíduos tratados

Ácido Úrico	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	1,77	9,36	4,96	**5,80**	0,90	1,70	1,33
Cdt 9	2,07	4,67	2,76	**5,71**	1,06	2,04	2,87
Cdt 10	1,54	4,63	10,78	9,46	0,95	2,73	1,42
Cdt 11	2,79	2,52	**4,31**	0,10	0,89	4,38	2,75
Cdt 29	3,22	1,63	2,17	0,56	1,98	**2,04**	3,00

Na Figura 7.12 são apresentados os perfis médios dos níveis de ácido úrico (AU).

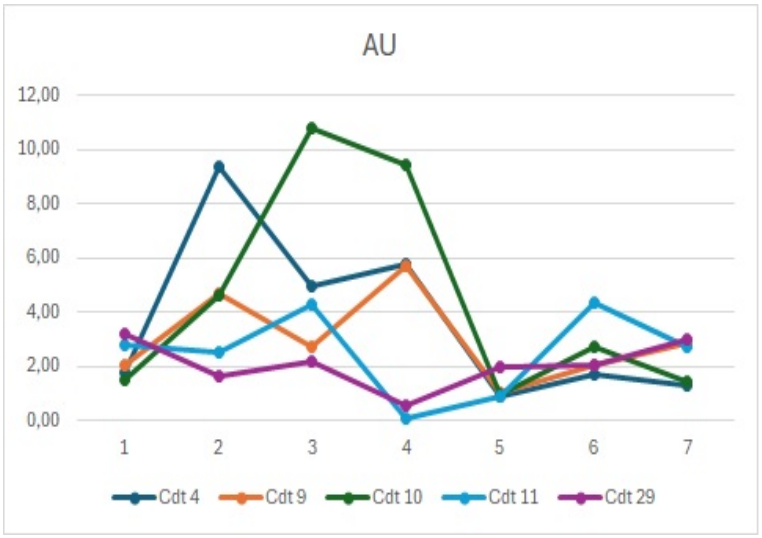

Figura 7.12: Perfis médios - Concentrações de ácido úrico (AU) em indivíduos tratados

Nas medidas da variável colesterol, ocorreram 5 valores ausentes, nas coletas 3, 4 e 6. Somente o indivíduo Cdt10 teve todas as coletas feitas. A Tabela 7.24 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.24: Concentrações de colesterol em indivíduos tratados

Colesterol	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 4	126,67	131,82	**145,81**	**143,74**	201,22	280,19	114,30
Cdt 9	183,90	224,04	173,48	**131,64**	174,44	142,96	94,44
Cdt 10	130,43	122,57	127,02	178,02	286,19	255,35	198,76
Cdt 11	113,60	107,12	**149,84**	67,17	187,53	99,03	243,00
Cdt 29	93,07	15,76	122,25	117,50	325,01	**237,84**	99,18

Na Figura 7.13 são apresentados os perfis médios dos níveis de colesterol.

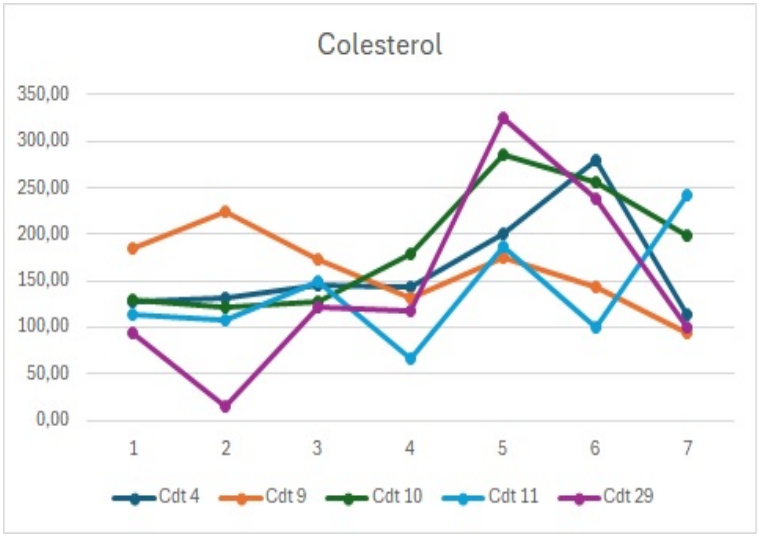

Figura 7.13: Perfis médios - Concentrações de colesterol em indivíduos tratados

7.2.3 Perfis médios do Grupo 2 - Indivíduos não tratados com vermífugo

Nas medidas da variável hemácias, ocorreram 5 valores ausentes, nas coletas 1, 4 e 7. Somente o indivíduo Cdt16 teve todas as coletas feitas. A Tabela 7.25 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito

Tabela 7.25: Contagens de hemácias em indivíduos não tratados

Hemácias	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 2	440.000	390.000	470.000	**546.000**	550.000	570.000	380.000
Cdt 7	870.000	500.000	310.000	810.000	450.000	580.000	380.000
Cdt 16	550.000	490.000	540.000	570.000	590.000	460.000	440.000
Cdt 30	**460.000**	590.000	480.000	490.000	370.000	500.000	**380.000**
Cdt 32	**416.000**	650.000	880.000	630.000	440.000	510.000	**416.000**

Na Figura 7.14 são apresentados os perfis médios dos números de hemácias.

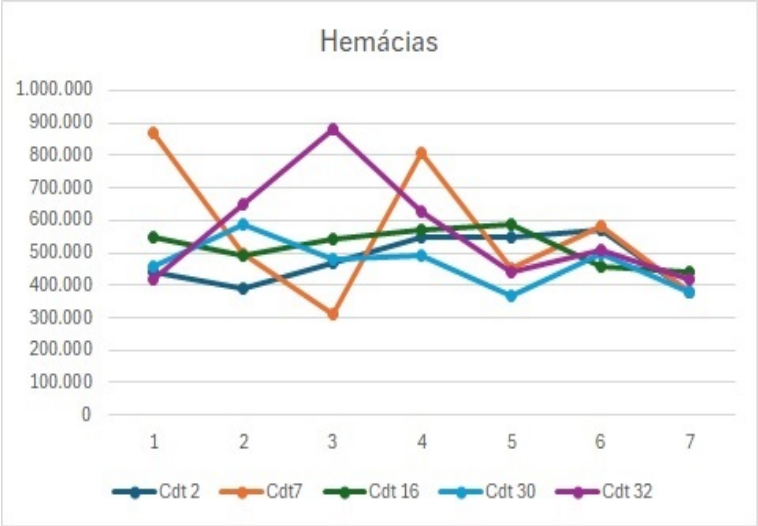

Figura 7.14: Perfis médios - Contagens de hemácias em indivíduos não tratados

Nas medidas da variável leucócitos, ocorreram 3 valores ausentes, nas coletas 1 e 4. Os indivíduos Cdt7 e Cdt16 tiveram todas as coletas feitas. A Tabela 7.26 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.26: Contagens de leucócitos em indivíduos não tratados

Leucócitos	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6
Cdt 2	3.500	5.500	6.000	**10.800**	14.500	13.500
Cdt 7	5.500	3.000	4.000	9.000	9.500	12.500
Cdt 16	6.500	15.500	5.000	9.500	11.000	15.500
Cdt 30	**5.900**	5.000	13.000	12.500	9.500	11.000
Cdt 32	**7.100**	7.500	7.000	10.500	16.000	11.000

Na Figura 7.15 são apresentados os perfis médios dos números de leucócitos.

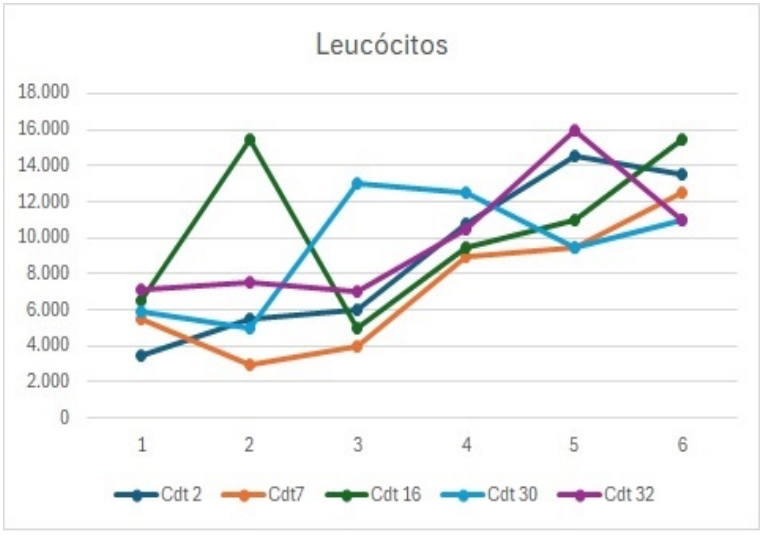

Figura 7.15: Perfis médios - Contagens de leucócitos em indivíduos não tratados

Nas medidas da variável trombócitos, ocorreram 3 valores ausentes, nas coletas 1 e 4. Os indivíduos Cdt7 e Cdt16 tiveram todas as coletas feitas. A Tabela 7.27 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.27: Contagens de trombócitos em indivíduos não tratados

Trombócitos	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6
Cdt 2	6.500	6.000	3.500	**4.700**	13.000	12.000
Cdt 7	4.500	6.500	4.000	2.500	6.000	9.000
Cdt 16	3.500	7.500	5.500	6.000	4.500	12.000
Cdt 30	**5.100**	2.500	5.000	6.500	6.000	8.000
Cdt 32	**4.100**	5.500	4.000	5.000	6.000	6.000

Na Figura 7.16 são apresentados os perfis médios dos números de trombócitos.

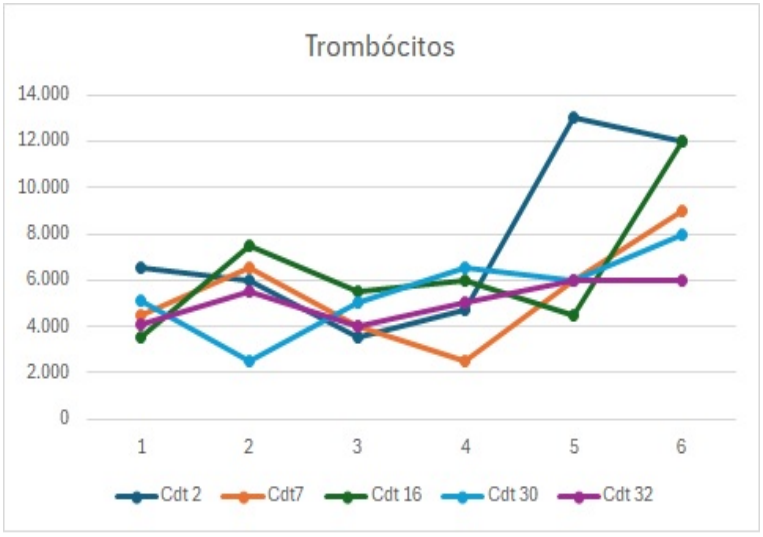

Figura 7.16: Perfis médios - Contagens de trombócitos em indivíduos não tratados

Nas medidas da variável hemoglobina, ocorreram 3 valores ausentes, nas coletas 4 e 7. Os indivíduos Cdt7 e Cdt16 tiveram todas as coletas feitas. A Tabela 7.28 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.28: Concentrações de hemoglobina em indivíduos não tratados

Hemoglobina	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt2	5,7	6,3	7,2	**5,04**	6,4	4,9	4
Cdt7	12,5	5,3	7,1	5,2	4,4	4,7	3,9
Cdt16	8,4	5	5,7	4,3	4,6	2,4	2,9
Cdt30	6,1	10,1	5,5	5,4	3,4	5	**3,52**
Cdt32	11	10,9	7,5	7,4	4,4	3,7	**3,54**

Na Figura 7.17 são apresentados os perfis médios dos níveis de hemoglobina.

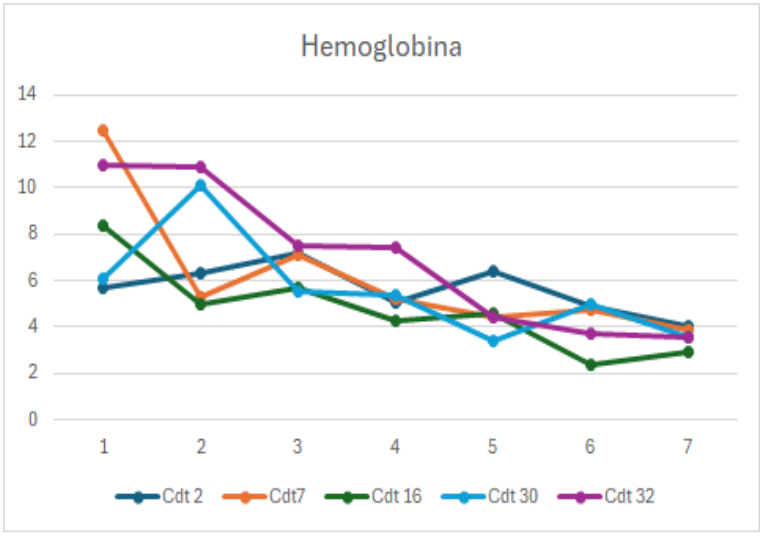

Figura 7.17: Perfis médios - Concentrações de hemoglobina em indivíduos não tratados

Nas medidas da variável glicose, ocorreram 3 valores ausentes, nas coletas 3 e 7. Os indivíduos Cdt2 e Cdt16 tiveram todas as coletas feitas. A Tabela 7.29 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.29: Concentrações de glicose em indivíduos não tratados

Glicose	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 2	17,89	64,89	22,98	22,33	91,30	47,72	28,10
Cdt 7	28,93	31,81	**26,44**	25,57	44,06	43,94	18,91
Cdt 16	28,54	55,11	18,52	20,04	45,22	39,46	9,14
Cdt 30	40,78	19,98	31,98	49,94	57,87	38,87	**16,72**
Cdt 32	42,61	21,33	40,73	94,13	54,97	37	**22,47**

Na Figura 7.18 são apresentados os perfis médios dos níveis de glicose.

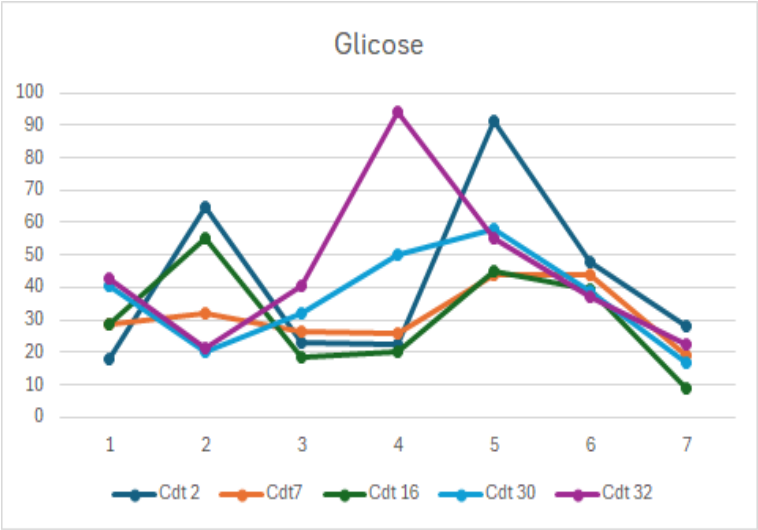

Figura 7.18: Perfis médios - Concentrações de glicose em indivíduos não tratados

Nas medidas da variável fósforo, ocorreram 4 valores ausentes, nas coletas 2, 3 e 7. Os indivíduos Cdt2 e Cdt16 tiveram todas as coletas feitas. A Tabela 7.30 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.30: Concentrações de fósforo em indivíduos não tratados

Fósforo	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 2	3,40	3,10	2,30	1,86	3,56	2	2,61
Cdt 7	3,70	2,60	**1,86**	1,93	1,93	2,60	1,76
Cdt 16	4,20	2,10	0,34	1,89	1,89	1,90	4,18
Cdt 30	1,85	**1,90**	1,96	2,39	2,39	4,16	**2,92**
Cdt 32	1,82	1,20	2,29	2,66	2,66	3,79	**2,89**

Na Figura 7.19 são apresentados os perfis médios dos níveis de fósforo.

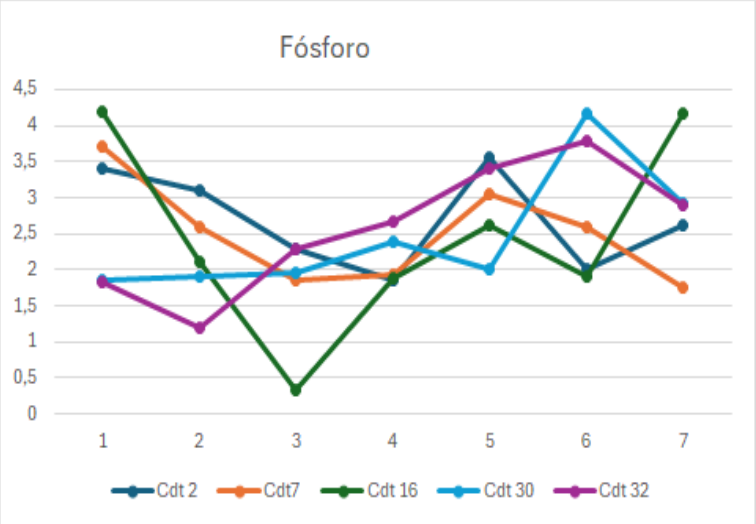

Figura 7.19: Perfis médios - Concentrações de fósforo em indivíduos não tratados

Nas medidas da variável ácido úrico, ocorreram 3 valores ausentes, nas coletas 3 e 7. Os indivíduos Cdt2 e Cdt16 tiveram todas as coletas feitas. A Tabela 7.31 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.31: Concentrações de ácido úrico (AU) em indivíduos não tratados

AU	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 2	5,20	1,73	1,65	0,15	1,22	1,41	1,80
Cdt 7	1,66	3	**3,30**	0,18	1,69	1,69	2,40
Cdt 16	1,17	3,96	2,55	1,57	0,82	1,84	3,46
Cdt 30	3,66	3,42	1,85	3,25	2,73	1,81	**2,25**
Cdt 32	4,33	4,76	5,14	4,60	3,31	1,20	**2,79**

Na Figura 7.20 são apresentados os perfis médios dos níveis de ácido úrico (AU).

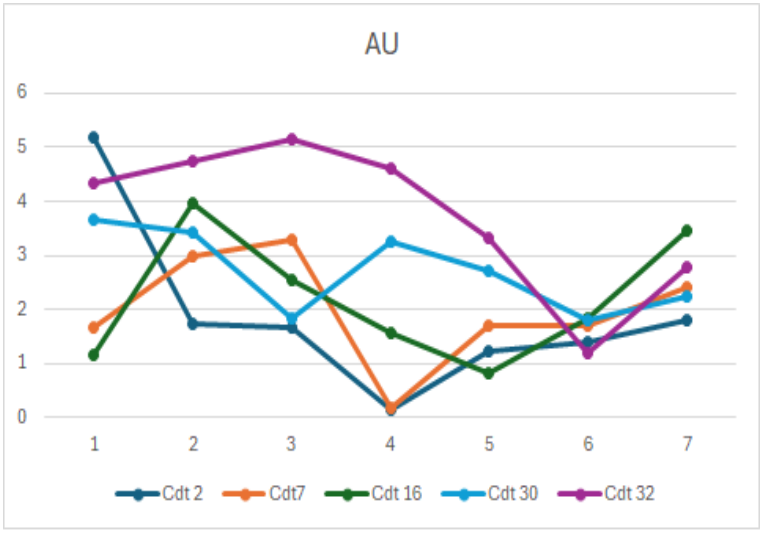

Figura 7.20: Perfis médios - Concentrações de ácido úrico (AU) em indivíduos não tratados

Nas medidas da variável colesterol, ocorreram 3 valores ausentes, nas coletas 3 e 7. Os indivíduos Cdt2 e Cdt16 tiveram todas as coletas feitas.A Tabela 7.32 apresenta a combinação dos 5 conjuntos imputados sendo as médias desses conjuntos apresentadas em negrito.

Tabela 7.32: Concentrações de colesterol em indivíduos não tratados

Colesterol	Coleta 1	Coleta 2	Coleta 3	Coleta 4	Coleta 5	Coleta 6	Coleta 7
Cdt 2	68,22	103,50	98,76	131,86	134,12	120,93	152,18
Cdt 7	161,99	164,65	**128,65**	207,87	329,67	172,85	208,07
Cdt 16	87,81	60,17	162,80	135,68	239,11	181,05	166,71
Cdt 30	97,89	106,31	127,02	235,92	94,29	64,32	**166,26**
Cdt 32	115,17	131,63	149,42	311,77	75,10	91	**185,71**

Na Figura 7.21 são apresentados os perfis médios dos níveis de colesterol.

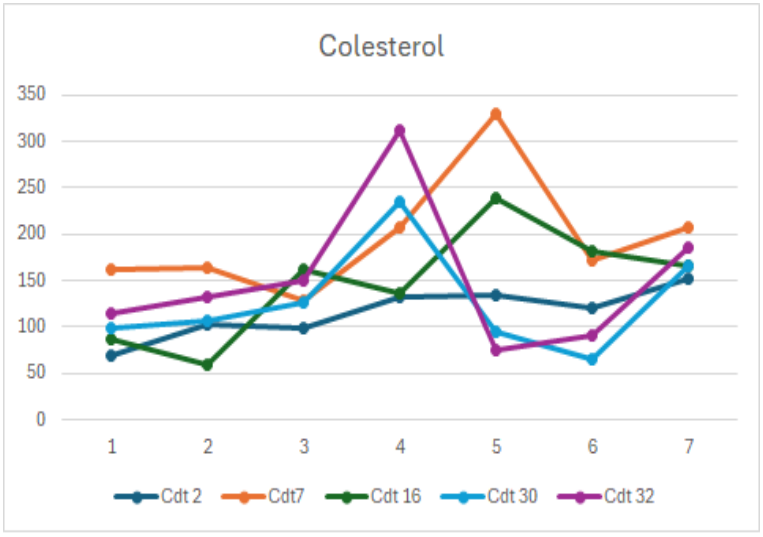

Figura 7.21: Perfis médios - Concentrações de colesterol em indivíduos não tratados

8 Conclusões

Neste trabalho foram apresentados os diversos métodos de imputação de dados ausentes, desde os mais simples até o método de imputação múltipla. Foram aplicados os métodos a três conjuntos de dados. A primeira aplicação foi inteiramente reproduzida com todos os cálculos refeitos no software R, baseando-se nos dados de qualidade do ar de um exemplo do livro "*Flexible Imputation of Missing Data*". Esse exemplo mostra os problemas quando se utiliza os métodos mais simples: ocorre a falsa diminuição ou aumento da variabilidade total, há a imputação de valores fora do suporte das variáveis, há vícios nos estimadores, entre outras desvantagens.

Na segunda aplicação, sobre qualidade da água do Rio Piracicaba, um conjunto de dados completo foi utilizado para a comparação dos métodos de imputação. Foram retirados aleatoriamente alguns valores em cada variável e em seguida esses valores foram imputados pelos diversos métodos. Um modelo de regressão foi ajustado para todos os conjuntos de dados: o original (completo) e os imputados. Os resultados foram comparados e verificou-se que o método de imputação múltipla foi o que forneceu resultados mais próximos dos resultados com os dados originais (completos).

Para a terceira aplicação, com dados longitudinais sobre parâmetros hematológicos e bioquímicos em serpentes, ocorreu somente a imputação múltipla sendo aplicada com o algoritmo MICE. Os dados originais apresentavam vários valores ausentes tornando a análise de perfis inviável uma vez que o plano dos pesquisadores seria coletar dados de apenas 20 serpentes, divididas em dois grupos, em 7 ocasiões. Na realidade, vários dados foram perdidos. Percebe-se que há variabilidade grande entre as medidas dos indivíduos, e entre coletas considerando os dados originais. Após a imputação, com cinco conjuntos de dados imputados, vemos que as médias dos valores imputados estão de acordo com os intervalos de valores observados e não há aumento nem diminuição da variabilidade dos dados. Considerando os resultados das três aplicações pode-se concluir que o método de imputação múltipla se mostrou eficaz e com boas propriedades, possibilitando a análise estatística posterior confiável. Esse método é de simples implementação no software R e deve ser considerado quando a ausência de dados não pode ser desprezada.

Referências

[1] ALLAN, F. E.; WISHART, J. A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, [S. l.], v. 20, n. 3, p. 399–406, 1930.

[2] AUSTIN, P. C. et al. Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*, [S. l.], v. 37, n. 9, p. 1322–1331, 2021.

[3] BROWN, M.; KROS, J. F. Data mining and the impact of missing data. *Industrial Management & Data Systems*, [S. l.], v. 103, n. 8, p. 611–621, 2003.

[4] CAO, Y. et al. Review and evaluation of imputation methods for multivariate longitudinal data with mixed-type incomplete variables. *Statistics in Medicine*, [S. l.], v. 41, n. 30, p. 5844–5876, 2022.

[5] CASELLA, G.; GEORGE, E. I. Explaining the gibbs sampler. *The American Statistician*, [S. l.], v. 46, n. 3, p. 167–174, 1992.

[6] United States Bureau of the Census. *Census of Manufactures 1958*. [S.l.]: Washington, D.C., 1961.

[7] CETESB. *Relatórios de Qualidade das Águas Interiores do Estado de São Paulo*. São Paulo: [s.n.], 2025. Disponível em: <https://cetesb.sp.gov.br/aguas-interiores/publicacoes-e-relatorios/>. Acesso em: 16 abr. 2025.

[8] COSTA, H. C.; GUEDES, T. B.; BERNILS, R. S. Lista de répteis do brasil: padrões e tendências. *Herpetologia Brasileira*, [S. l.], p. 110–279, 2021.

[9] CUNHA, L. E. R. Soros antiofídicos: história, evolução e futuro. *Journal Health NPEPS*, [S. l.], v. 2, n. 1, p. 1–4, 2017.

[10] GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. *Markov Chain Monte Carlo in Practice*. London: Chapman-Hall, 1996.

[11] GRAHAM, J. W.; HOFER, S. M.; PICCININ, A. M. K. Analysis with missing data in prevention research. In: BRYANT, M.; WINDLE, M.; WEST, S. (Ed.). *The science of prevention: methodological advances from alcohol and substance abuse research*. Washington, D.C.: American Psychological Association, 1997. p. 325–366.

[12] GRAHAM, J. W.; HOFER, S. M. Multiple imputation in multivariate research. In: LITTLE, T. D.; SCHNABEL, K. U.; BAUMERT, J. (Ed.). *Modeling longitudinal and multilevel data*. [S. l.]: Psychology Press, 2000. p. 189–204.

[13] HUQUE, M. H. et al. Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biometrical Journal*, [S. l.], v. 62, n. 2, p. 444–466, 2020.

[14] LACHIN, J. M. Fallacies of last observation carried forward analyses. *Clinical Trials*, London, v. 13, n. 2, p. 161–168, 2016.

[15] LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data*. New York: John Wiley & Sons, 1987.

[16] LITTLE, R. J. A. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, [S. l.], v. 6, n. 3, p. 287–296, 1988.

[17] LITTLE, R. J. A. Regression with missing x's: A review. *Journal of the American Statistical Association*, [S. l.], v. 87, n. 420, p. 1227–1237, 1992.

[18] LITTLE, R. J. A.; SCHENKER, N. Missing data. In: ARMINGER, G.; CLOGG, C. C.; SOBEL, D. B. (Ed.). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Boston, MA: Springer US, 1995. p. 39–75.

[19] LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data*. 2. ed. New York: John Wiley & Sons, 2002.

[20] LOBATO, F. M. F. *Estratégias evolucionárias para otimização no tratamento de dados ausentes por imputação múltipla de dados*. Tese (Tese (Doutorado)) — Universidade Federal do Pará, Belém, 2016.

[21] MACK, C.; SU, Z.; WEITRICH, D. *Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide*. [S. l.]: [s. n.], 2018.

[22] PEPINSKY, T. B. A note on listwise deletion versus multiple imputation. *Political Analysis*, [S. l.], v. 26, n. 4, p. 480–488, 2018.

[23] ROBERTO, I. J. *Répteis (Testudines, Squamata, Crocodylia) da Reserva Biológica de Pedra Talhada*. [S. l.]: Melgarejo, 2015.

[24] ROSATO, R. et al. Missing data in longitudinal studies: comparison of multiple imputation methods in a real clinical setting. *Journal of Evaluation in Clinical Practice*, [S. l.], v. 27, n. 1, p. 34–41, 2021.

[25] R Core Team. *The R Project for Statistical Computing*. [S. l.], 2026. Disponível em: <https://www.r-project.org/>. Acesso em: 1 mar. 2026.

[26] RUBIN, D. B. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, [S. l.], v. 72, n. 359, p. 538–543, 1977.

[27] RUBIN, D. B. Adjustments in large using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, [S. l.], v. 4, n. 1, p. 87–94, 1986.

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

[29] RUBIN, D. B. The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, [S. l.], v. 58, n. 4, p. 298–302, 2004.

[30] SCHAFER, J. L. *Analysis of incomplete multivariate data*. London: Chapman-Hall, 1997.

[31] SCHAFER, J. L.; GRAHAM, J. W. Missing data: Our view of the state of the art. *Psychological Methods*, [S. l.], v. 7, n. 2, p. 147–177, 2002.

[32] SCHEUREN, F. *Projeto articulado com o Social Security Administration e o US Census Bureau*. [S. l.]: [s. n.], 1977.

[33] SCHOBER, P.; VETTER, T. R. Repeated measures designs and analysis of longitudinal data: If at first you do not succeed try, try again. *Anesthesia and Analgesia*, [S. l.], v. 127, n. 2, 2018.

[34] SILVA, M. J. C. *Imputação múltipla comparação e eficiência em experimentos multiambientais*. Tese (Tese (Doutorado)) — Universidade de São Paulo, Piracicaba, 2012.

[35] SINGER, J. M.; NOBRE, J. S.; ROCHA, F. M. M. *Análise de dados longitudinais*. [S. l.]: [s. n.], 2018.

[36] BUUREN, S. V.; GROOTHUIS-OUDSHOORN, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, [S. l.], v. 45, n. 3, 2011.

[37] BUUREN, S. V. *Flexible Imputation of Missing Data*. [S. l.]: Chapman & Hall/CRC, 2012.

[38] HIPPEL, P. T. V. *How to impute interactions, squares and other transformed variables*. London: Chapman-Hall, 2009.

[39] WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, [S. l.], v. 40, n. 4, p. 377–399, 2011.

[40] YATES, F. The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, [S. l.], v. 1, n. 2, p. 129–142, 1933.

A Apêndice A

A.1 Dados de qualidade do ar

Tabela A.1: Dados de qualidade de ar

Obs.	**%Ozônio**	**Rad. solar**	**Vento**	**Temp**
1	41	190	7,4	67
2	36	118	8	72
3	12	149	12,6	74
4	18	313	11,5	62
5	NA	NA	14,3	56
6	28	NA	14,9	66
7	23	299	8,6	65
8	19	99	13,8	59
9	8	19	20,1	61
10	NA	194	8,6	69
11	7	NA	6,9	74
12	16	256	9,7	69
13	11	290	9,2	66
14	14	274	10,9	68
15	18	65	13,2	58
16	14	334	11,5	64
17	34	307	12	66
18	6	78	18,4	57
19	30	322	11,5	68
20	11	44	9,7	62
21	1	8	9,7	59
22	11	320	16,6	73
23	4	25	9,7	61
24	32	92	12	61
25	NA	66	16,6	57
26	NA	266	14,9	58
27	NA	NA	8	57
28	23	13	12	67
29	45	252	14,9	81
30	115	223	5,7	79
31	37	279	7,4	76
32	NA	286	8,6	78
33	NA	287	9,7	74
34	NA	242	16,1	67

Obs.	**%Ozônio**	**Rad. solar**	**Vento**	**Temp**
35	NA	186	9,2	84
36	NA	220	8,6	85
37	NA	264	14,3	79
38	29	127	9,7	82
39	NA	273	6,9	87
40	71	291	13,8	90
41	39	323	11,5	87
42	NA	259	10,9	93
43	NA	250	9,2	92
44	23	148	8	82
45	NA	332	13,8	80
46	NA	322	11,5	79
47	21	191	14,9	77
48	37	284	20,7	72
49	20	37	9,2	65
50	12	120	11,5	73
51	13	137	10,3	76
52	NA	150	6,3	77
53	NA	59	1,7	76
54	NA	91	4,6	76
55	NA	250	6,3	76
56	NA	135	8	75
57	NA	127	8	78
58	NA	47	10,3	73
59	NA	98	11,5	80
60	NA	31	14,9	77
61	NA	138	8	83
62	135	269	4,1	84
63	49	248	9,2	85
64	32	236	9,2	81
65	NA	101	10,9	84
66	64	175	4,6	83
67	40	314	10,9	83
68	77	276	5,1	88
69	97	267	6,3	92
70	97	272	5,7	92
71	85	175	7,4	89
72	NA	139	8,6	82
73	10	264	14,3	73
74	27	175	14,9	81
75	NA	291	14,9	91
76	7	48	14,3	80
77	48	260	6,9	81
78	35	274	10,3	82
79	61	285	6,3	84
80	79	187	5,1	87
81	63	220	11,5	85
82	16	7	6,9	74
83	NA	258	9,7	81
84	NA	295	11,5	82
85	80	294	8,6	86
86	108	223	8	85

Obs.	**%Ozônio**	**Rad. solar**	**Vento**	**Temp**
87	20	81	8,6	82
88	52	82	12	86
89	82	213	7,4	88
90	50	275	7,4	86
91	64	253	7,4	83
92	59	254	9,2	81
93	39	83	6,9	81
94	9	24	13,8	81
95	16	77	7,4	82
96	78	NA	6,9	86
97	35	NA	7,4	85
98	66	NA	4,6	87
99	122	255	4	89
100	89	229	10,3	90
101	110	207	8	90
102	NA	222	8,6	92
103	NA	137	11,5	86
104	44	192	11,5	86
105	28	273	11,5	82
106	65	157	9,7	80
107	NA	64	11,5	79
108	22	71	10,3	77
109	59	51	6,3	79
110	23	115	7,4	76
111	31	244	10,9	78
112	44	190	10,3	78
113	21	259	15,5	77
114	9	36	14,3	72
115	NA	255	12,6	75
116	45	212	9,7	79
117	168	238	3,4	81
118	73	215	8	86
119	NA	153	5,7	88
120	76	203	9,7	97
121	118	225	2,3	94
122	84	237	6,3	96
123	85	188	6,3	94
124	96	167	6,9	91
125	78	197	5,1	92
126	73	183	2,8	93
127	91	189	4,6	93
128	47	95	7,4	87
129	32	92	15,5	84
130	20	252	10,9	80
131	23	220	10,3	78
132	21	230	10,9	75
133	24	259	9,7	73
134	44	236	14,9	81
135	21	259	15,5	76
136	28	238	6,3	77
137	9	24	10,9	71
138	13	112	11,5	71

Obs.	**%Ozônio**	**Rad. solar**	**Vento**	**Temp**
139	46	237	6,9	78
140	18	224	13,8	67
141	13	27	10,3	76
142	24	238	10,3	68
143	16	201	8	82
144	13	238	12,6	64
145	23	14	9,2	71
146	36	139	10,3	81
147	7	49	10,3	69
148	14	20	16,6	63
149	30	193	6,9	70
150	NA	145	13,2	77
151	14	191	14,3	75
152	18	131	8	76
153	20	223	11,5	68