COMPARAÇÃO DAS ÁGUAS DOS RIOS JAGUARI E ATIBAIA NA REGIÃO DE LANÇAMENTO DE EFLUENTE DE INDÚSTRIA PETROQUÍMICA

Eduardo Schneider Bueno de Oliveira

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro - 2016

COMPARAÇÃO DAS ÁGUAS DOS RIOS JAGUARI E ATIBAIA NA REGIÃO DE LANÇAMENTO DE EFLUENTE DE INDÚSTRIA PETROQUÍMICA

Eduardo Schneider Bueno de Oliveira

Orientador: Prof. Dr. **Antonio Carlos Simões Pião**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro - 2016

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Oliveira, Eduardo Schneider Bueno de.
Comparação das águas dos rios Jaguari e Atibaia na região de lançamento de efluente de indústria petroquímica / Eduardo Schneider Bueno de Oliveira. - Botucatu, 2016

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Antonio Carlos Simões Pião
Capes: 10202005

1. Água - Qualidade. 2. Critica de imputação de dados (Estatística). 3. Bootstrap (Estatística). 4. Resíduos industriais - Aspectos ambientais. 5. Jaguari Mirim, Rio, Bacia (MG e SP). 6. Atibaia, Rio (SP).

Palavras-chave: Bootstrap em blocos; Imputação de dados; Qualidade da água.

MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DE EDUARDO SCHNEIDER BUENO DE OLIVEIRA, INTITULADA "COMPARAÇÃO DAS ÁGUAS DOS RIOS JAGUARI E ATIBAIA NA REGIÃO DE LANÇAMENTO DE EFLUENTE DE INDÚSTRIA PETROQUÍMICA ", APRESENTADA AO INSTITUTO DE BIOCIÊNCIAS, UNESP, CAMPUS DE BOTUCATU, SÃO PAULO, EM 03 de fevereiro de 2016 .

APROVADA PELA COMISSÃO JULGADORA:

Prof(a) Dr(a) ANTONIO CARLOS SIMÕES PIÃO
Instituição: IGCE/UNESP
Assinatura:

Prof(a) Dr(a) LÚCIO BORGES ARAÚJO
Instituição: Universidade Federal de Uberlândia
Assinatura:

Prof(a) Dr(a) LUZIA APARECIDA TRINCA
Instituição: IBB/UNESP
Assinatura:

Dedicatória

À Deus, sem Quem nada existiria.
À minha família, sem quem eu não existiria.
À meus verdadeiros amigos, sem os quais
muitos belos momentos não existiriam.

"Necessitamos uns dos outros para sermos nós mesmos." Agostinho de Hipona

Agradecimentos

Agradeço primeiramente à Deus, pelo dom da vida e pela oportunidade de crescer a cada dia.

Agradeço aos meus pais, José Roberto Bueno de Oliveira e Sirlene Habermann Schneider Bueno de Oliveira, pelo carinho e amor e também pelos conselhos, apoio e incentivo.

À todos os demais familiares, pela presença em minha vida, pelo amor e pelos incentivos que sempre me deram. Em especial, agradeço aos meus avós, Yvone Habermann Schneider e Milton Schneider, por todo o carinho e por sempre me incentivarem e também a meu tio Gilberto Habermann Schneider, que teve grande importância em meu processo de formação e mesmo na escolha do curso de Estatística.

Aos amigos, que fazem com que haja mais brilho no dia-a-dia e tornam melhor a caminhada, tanto os companheiros de Mestrado quanto os demais, cuja amizade adquirida ao longo da vida permanece, servindo como grande apoio e motivação. Agradeço em especial aos amigos Murilo Cantoni e Adriane Donegá, por todas as conversas e pelo companheirismo, mesmo que muitas vezes à distância e pela grande importância em meu desenvolvimento, seja ele acadêmico ou pessoal.

Ao meu orientador, o Prof. Dr. Antonio Carlos Simões Pião pela disponibilidade e por ter colaborado com meu crescimento durante esse período, tanto como pessoa quanto no âmbito profissional.

À Sérgio Arciniegas-Alarcón, pela ajuda com o endendimento da parte de Imputação Múltipla e com a pesquisa bibliográfica nessa área.

Aos professores da banca avaliadora deste trabalho, pelo tempo confe-

rido e pelas contribuições valiosas ao trabalho desenvolvido.

Aos professores e funcionários do departamento de Bioestatística da UNESP de Botucatu e do departamento de Estatística, Matemática Aplicada e Computação da UNESP de Rio Claro.

Agradeço também à CAPES, pelo apoio financeiro cedido ao projeto de Mestrado.

Enfim, a todos que tiveram e tem muita importância tanto no processo de crescimento acadêmico e profissional quanto na vida pessoal, meu muito obrigado.

"Em tudo dai graças, porque esta é a vontade de Deus em Cristo Jesus para convosco." 1 Tessalonicenses 5.18

Sumário

Página

LISTA DE FIGURAS **viii**

LISTA DE TABELAS **xi**

RESUMO **xiii**

SUMMARY **xv**

1 INTRODUÇÃO **1**

2 OBJETIVOS **4**

3 REVISÃO BIBLIOGRÁFICA **6**

3.1 Imputação de Dados 6

3.1.1 Distribuição dos dados ausentes 7

3.1.2 Padrões de dados ausentes 8

3.1.3 Mecanismos de dados ausentes 9

3.1.4 Métodos de estimação de dados ausentes 12

3.1.5 Imputação de Dados livre de Distribuição 15

3.1.6 Qualidade do Ajuste da Imputação de Dados 19

3.2 Métodos de Avaliação da Qualidade da Água 22

3.2.1 Índice de Qualidade da Água 22

3.3 Análise de Variância 27

3.4 Bootstrap 30

3.4.1 Definição do tamanho dos blocos 35
3.4.2 Bootstrap em Blocos móveis 36
3.4.3 Bootstrap Circular 37
3.4.4 Bootstrap Estacionário 37
3.4.5 Bootstrap no R 38
3.5 Conceitos de Séries Temporais 39
3.5.1 Estacionariedade 39
3.5.2 Sazonalidade 41

4 MATERIAL E MÉTODOS 42

4.1 Localização da bacia hidrográfica 42
4.1.1 Pontos de Amostragem 43
4.2 Metodologia da Coleta 46
4.2.1 Períodos de Coleta 46
4.2.2 Análises Físicas e Químicas 47
4.2.3 Os Dados do Estudo 48
4.3 Imputação de Dados 50
4.3.1 Padrão e Mecanismo dos Dados Analisados 50
4.3.2 Método de Imputação de Dados 51
4.4 Análise Estatística 54
4.4.1 Bootstrap 55
4.4.2 ANOVA 56

5 RESULTADOS 57

5.1 Imputação de Dados 57
5.2 Análise Estatística 66
5.2.1 Matriz Original 66
5.2.1.1 Análise dos dados originais 67
5.2.1.2 Análise pelo Bootstrap em blocos 72
5.2.2 Matriz Imputada 79

5.2.2.1 Análise dos dados imputados 80
5.2.2.2 Análise pelo Bootstrap em Blocos 82
5.2.2.3 Análise por quinquênios 101

6 DISCUSSÃO **112**

7 CONCLUSÃO **115**

REFERÊNCIAS BIBLIOGRÁFICAS **117**

Lista de Figuras

Página

1 Representação dos padrões de ausência de dados. 9
2 Representação dos mecanismos de ausência de dados. 11
3 Curvas de classificação de subíndices. 25
4 Lógica do Método Bootstrap 31
5 Esquema ilustrativo dos diversos pontos de amostragem. 43
6 Pontos de coleta da água. 45
7 IQA ao longo do período estudado para cada ponto. 69
8 Boxplot das observações mensais para o Ponto 1. 69
9 Boxplot das observações mensais para o Ponto 3. 70
10 Boxplot das observações mensais para o Ponto 5. 71
11 Boxplots das estimativas Bootstrap em blocos móveis. 73
12 Histogramas das estimativas Bootstrap em blocos móveis. 73
13 Boxplots das estimativas Bootstrap estacionário. 74
14 Histogramas das estimativas Bootstrap estacionário. 74
15 Boxplots das estimativas Bootstrap em blocos móveis - Contrastes. . . . 76
16 Histogramas das estimativas Bootstrap em blocos móveis - Contrastes. . 77
17 Boxplots das estimativas Bootstrap estacionário - Contrastes. 77
18 Histogramas das estimativas Bootstrap estacionário - Contrastes. 78
19 IQA para cada imputação por ponto. 81
20 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a primeira imputação. 83

21 Histogramas e Boxplots das estimativas Bootstrap estacionário para a primeira imputação. 83
22 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a segunda imputação. 84
23 Histogramas e Boxplots das estimativas Bootstrap estacionário para a segunda imputação. 85
24 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a terceira imputação. 85
25 Histogramas e Boxplots das estimativas Bootstrap estacionário para a terceira imputação. 86
26 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quarta imputação. 87
27 Histogramas e Boxplots das estimativas Bootstrap estacionário para a quarta imputação. 87
28 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quinta imputação. 88
29 Histogramas e Boxplots das estimativas Bootstrap estacionário para a quinta imputação. 89
30 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a primeira imputação - Contrastes. 93
31 Histogramas e Boxplots das estimativas Bootstrap estacionário para a primeira imputação - Contrastes. 94
32 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a segunda imputação - Contrastes. 94
33 Histogramas e Boxplots das estimativas Bootstrap estacionário para a segunda imputação - Contrastes. 95
34 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a terceira imputação - Contrastes. 95

35 Histogramas e Boxplots das estimativas Bootstrap estacionário para a terceira imputação - Contrastes. 96
36 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quarta imputação - Contrastes. 96
37 Histogramas e Boxplots das estimativas Bootstrap estacionário para a quarta imputação - Contrastes. 97
38 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quinta imputação - Contrastes. 97
39 Histogramas e Boxplots das estimativas Bootstrap estacionário para a quinta imputação - Contrastes. 98
40 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis - Primeiro quinquênio. 103
41 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis - Segundo quinquênio. 103
42 Histogramas e Boxplots das estimativas Bootstrap em blocos móveis - Terceiro quinquênio. 104
43 Histogramas e Boxplots das estimativas Bootstrap estacionário - Primeiro quinquênio 104
44 Histogramas e Boxplots das estimativas Bootstrap estacionário - Segundo quinquênio 105
45 Histogramas e Boxplots das estimativas Bootstrap estacionário - Terceiro quinquênio 105
46 Boxplots dos IQAs por ponto para os quinquênios - Primeira Imputação 106
47 Boxplots dos IQAs por ponto para os quinquênios - Segunda Imputação . 106
48 Boxplots dos IQAs por ponto para os quinquênios - Terceira Imputação . 107
49 Boxplots dos IQAs por ponto para os quinquênios - Quarta Imputação . 107
50 Boxplots dos IQAs por ponto para os quinquênios - Quinta Imputação . 108

Lista de Tabelas

Página

1 Pesos das variáveis do IQA 26
2 Classificação da Qualidade da Água IQA-NSF 27
3 Classificação da Qualidade da Água IQA-CETESB 27
4 Tabela da ANOVA para experimento inteiramente aleatorizado 29
5 Localização geográfica dos pontos de coleta 44
6 Períodos e número de coletas durante a pesquisa 47
7 Proporção de Ausências para os Pontos 49
8 Desempenho para os diferentes métodos para a variável Coliformes . . . 58
9 Desempenho para os diferentes métodos para a variável DBO 58
10 Desempenho para os diferentes métodos para a variável Fósforo 59
11 Desempenho para os diferentes métodos para a variável Nitrogênio 59
12 Desempenho para os diferentes métodos para a variável OD 59
13 Desempenho para os diferentes métodos para a variável pH 60
14 Desempenho para os diferentes métodos para a variável Resíduos 60
15 Desempenho para os diferentes métodos para a variável Turbidez 60
16 Desempenho para os diferentes métodos utilizando os Q_i para o Ponto 1 62
17 Desempenho para os diferentes métodos utilizando os Q_i para o Ponto 3 62
18 Desempenho para os diferentes métodos utilizando os Q_i para o Ponto 5 63
19 Desempenho para os diferentes métodos para o IQA no Ponto 1 64
20 Desempenho para os diferentes métodos para o IQA no Ponto 3 64
21 Desempenho para os diferentes métodos para o IQA no Ponto 5 65

22 Testes das suposições da ANOVA 67
23 Testes formais de Estacionariedade 71
24 Bootstrap para a Matriz Original por Ponto 75
25 Bootstrap para contrastes na Matriz Original 78
26 Testes formais de Estacionariedade 82
27 Bootstrap para a Matriz Imputada - Primeira imputação 89
28 Bootstrap para a Matriz Imputada - Segunda imputação 90
29 Bootstrap para a Matriz Imputada - Terceira imputação 90
30 Bootstrap para a Matriz Imputada - Quarta imputação 90
31 Bootstrap para a Matriz Imputada - Quinta imputação 91
32 Bootstrap para contrastes na Matriz Imputada - Primeira Imputação . . 98
33 Bootstrap para contrastes na Matriz Imputada - Segunda Imputação . . 99
34 Bootstrap para contrastes na Matriz Imputada - Terceira Imputação . . 99
35 Bootstrap para contrastes na Matriz Imputada - Quarta Imputação . . . 99
36 Bootstrap para contrastes na Matriz Imputada - Quinta Imputação . . . 100
37 ANOVA entre quinquênios para o IQA no Ponto 1 - Primeira Imputação 109
38 ANOVA entre quinquênios para o IQA no Ponto 3 - Primeira Imputação 110
39 ANOVA entre quinquênios para o IQA no Ponto 5 - Primeira Imputação 110

COMPARAÇÃO DAS ÁGUAS DOS RIOS JAGUARI E ATIBAIA NA REGIÃO DE LANÇAMENTO DE EFLUENTE DE INDÚSTRIA PETROQUÍMICA

Autor: EDUARDO SCHNEIDER BUENO DE OLIVEIRA
Orientador: Prof. Dr. ANTONIO CARLOS SIMÕES PIÃO

RESUMO

A ação antrópica na natureza é algo muito constante ao longo de toda a história, mas cada vez mais notam-se os efeitos negativos que por vezes ela pode trazer. Verificar esses efeitos, suas implicações, e aquilo que pode ser feito para evitar maiores problemas é de suma importância para a manutenção de nosso planeta em boas condições e consequentemente para a qualidade de vida do ser humano. O presente estudo realiza uma análise da qualidade da água dos Rios Jaguarí e Atibaia, entre os quais há o despejo de resíduos de uma indústria, além da qualidade da água após o processo de utilização pela indústria, antes de sua devolução ao rio. Com isso, pode-se verificar a qualidade do tratamento de resíduo de tal indústria e analisar possíveis efeitos que possa haver na qualidade da água após o despejo dos resíduos no rio. Para isso, com base em dados sobre características físicas, químicas e

microbiológicas da água, são utilizadas técnicas estatísticas adequadas para realizar a análise necessária ao intuito anteriormente exposto. Como os dados possuem dependência entre si, é necessário que sejam utilizados métodos que permitam tal ocorrência, como o Bootstrap em Blocos não paramétrico (Künsch, 1989; Politis & Romano, 1994). Também há a realização de imputação múltipla de dados, uma vez que há diversos meses do estudo com dados ausentes, através da técnica de Imputação de Dados Livre de Distribuição (Bergamo, 2007; Bergamo et al., 2008).

Palavras-Chave: Qualidade da Água; Imputação de Dados; Bootstrap em Blocos.

COMPARISION OF THE WATER FROM RIVERS JAGUARI AND ATIBAIA AT THE REGION OF WASTEWATER RELEASE BY A PETROCHEMICAL INDUSTRY

Author: EDUARDO SCHNEIDER BUENO DE OLIVEIRA
Adviser: Prof. Dr. ANTONIO CARLOS SIMÕES PIÃO

SUMMARY

The anthropic action in nature is a constant factor along the history, but each day the negative effects that it brings can be increasingly seen. Check these effects, its implications and what can be done in order to avoid bigger problems has a great importance to the manteinance of our planet in good conditions and, consequently, to the human being life quality. This study performs an analysis of the water quality of the Jaguari and Atibaia rivers, among which happens the dumping of residuals from a petrochemical industry, as well as of the quality of the water after its utilization process by the industry, before its devolution to the river. Thus, it is possible to verify this industry's residual treatment quality and to analyze possible effects to the water quality after the residual dumping at the river. For this, based on data about fisical, chemical and microbiological characteristics of the water, appropriate statistical techniques are used, aiming to do the necessary analysis to fullfill

the exposed intention. Because of the existence of dependency, methods that allow this ocurrence shall be used, such as the non parametric Blocks Bootstrap (Künsch, 1989; Politis & Romano, 1994). There is also the realization of multiple imputation, using the technique of the Distribution-free Multiple Imputation (Bergamo, 2007; Bergamo et al., 2008), once for some months there are missing data.

Keywords: Water Quality; Data Imputation; Blocks Bootstrap.

1 INTRODUÇÃO

Desde o início da civilização o ser humano procurou assegurar sua subsistência estando em derredor de fontes com recurso de água doce. A água, em sua função fundamental, é insubstituível para saciar a sede e hidratar as plantas. Além disso, também tem afastado do homem resíduos indesejáveis. Contudo, à medida que há aumento populacional, proporcionalmente aumenta a demanda de água e os resíduos de descarte também se acumulam. Dependendo do volume e da qualidade dos resíduos, assim como do tempo que os mesmos permanecem na água, eles comprometem a qualidade da água para consumo do ser humano e toda a biota nela presente, contribuindo para a destruição da cadeia alimentar.

No Estado de São Paulo, embora a condição hidrológica seja razoável, na maior parte de seu território, há uma grande concentração populacional na região leste, com quase dois terços da população do estado nessa região (DAEE et al., 2005). Este fato faz com que haja uma elevada ocupação do solo e da água para abastecimento urbano e industrial, geração de energia elétrica, recreação, irrigação, navegação, entre outros. Com toda essa atividade, também há o recebimento de efluentes sanitários e industriais nas bacias hidrográficas presentes nesses lugares.

A região leste de São Paulo comporta o maior crescimento econômico do país (Vieira, 2009), o qual é estimulado pelas boas condições de escoamento rápido da produção, seja via terrestre ou aérea, além de estar situada na região com o mais bem estruturado esquema de formação de mão de obra especializada. O parque industrial dessa região do estado é diversificado, possuindo também intensa atividade agrícola com grande variedade de cultivares para atender o consumo interno e também o mercado externo, com produtos in natura ou manufaturados.

Esta elevada atividade antrópica tem crescido fortemente nos últimos 40 anos, com a maior industrialização do país, e de alguma forma perturba o meio ambiente. Especialmente nas últimas décadas, a preocupação crescente com a água fez com que o governo, de maneira geral, no âmbito federal, estadual e municipal, assumisse compromissos de controle do uso da água e da sua qualidade (Lemos & Lemos, 2009). Grandes esforços tem sido investidos, a fim de que se estabeleçam parâmetros de qualidade mínima da água para atender às exigências de saúde a nível nacional, comparando-se essa qualidade com os padrões internacionais mais rígidos.

Por outro lado, a implantação de Unidades de Recursos Hidrográficos (UBRHI) e o gereciamento de vinte e duas dessas unidades no Estado paulista, a qual teve início com a Bacia Hidrográfica dos Rios Piracicaba, Capivari e Jundiaí (BH-PCJ ou UGTHI-5) (Comitês-PCJ, 2015c), despertou grande parte da população para a responsabilidade com a qualidade e o uso da água, através de algumas ações tomadas por meio das agências competentes, as quais também fazem planos para salvaguardar a qualidade da água dessa bacia (Comitês-PCJ, 2015a,b).

O crescente aumento populacional que também ocorre na área da BH-PCJ, que em 2013 possuía mais de 5 milhões de habitantes, segundo o relatório de 2014 (Comitês-PCJ, 2014) produzido pela agência PCJ, associado ao uso agrícola e industrial, exige que cada vez mais haja aprimoramento das técnicas de avaliação da qualidade da água, especialmente no que se diz ao consumo humano.

Na elaboração da presente dissertação serão utilizados dados advindos dos estudos sobre a qualidade da água dos Rios Jaguarí e Atibaia em uma área na qual, além do despejo de esgoto advindo do consumo da população das cidades próximas, existe o despejo de resíduos, após tratamento, por uma indústria petrolífera. Os dados foram coletados desde 1999 até 2013, totalizando 11 relatórios técnicos produzidos com colaboração da FUNDUNESP.

Para a análise dos dados serão utilizadas técnicas de imputação de dados para a estimação dos dados faltantes, por falta de análise de alguma variável

específica e principalmente os existentes nas datas entre a elaboração dos relatórios, nas quais não foram coletadas amostras da água para análise. Após isso serão utilizados métodos estatísticos adequados, a fim de obter resultados quanto à avaliação da qualidade da água nos pontos do estudo e realizar comparação entre os mesmos.

2 OBJETIVOS

Este trabalho tem como objetivo principal reunir as informações colhidas na década de 2000 e no início da atual década sobre a qualidade física, química e microbiológica das águas dos Rios Jaguarí e Atibaia, realizando análises estatísticas adequadas e úteis a partir das mesmas.

Visa realizar estudos comparativos das amostras coletadas regularmente das águas dos Rios Jaguarí e Atibaia, tomando como referência os pontos de coleta na montante, na jusante e no local do efluente industrial tratado e descartado no Rio Atibaia, com dados totalizando 195 coletas, sendo que em alguns meses há mais de uma coleta e em outros não há coleta alguma.

A presença de observações em falta complica a análise estatística destes dados. A ocorrência de dados ausentes acontece principalmente devido à conclusão de um projeto e início de outro ao longo dos anos ou devido a alguma variável específica não ser analisada em alguma das coletas realizadas. Estes problemas vêm sendo empecilho na utilização de técnicas de análise estatística multivariadas e demais análises. Diante do problema surgem várias técnicas como solução, uma delas é a técnica de imputação de dados, que será aplicada nos dados (Rubin, 1976, 1987; Enders, 2006).

A técnica de imputação de dados consiste em estimar dados faltantes por estimativas e então, feito isso, os dados completos permitem aplicar uma gama maior de técnicas estatísticas, as quais permitem avaliar a qualidade da água do Rio Atibaia, do Rio Jaguarí e a eficiência do tratamento da indústria.

Através da matriz de dados original e também da imputada, visa-se a realização de análises que permitam a comparação da qualidade da água nos pontos

de coleta, a fim de verificar se existe diferença entre os mesmos. Também é um objetivo do trabalho verificar a situação individual de cada um dos pontos de coleta.

Com estes estudos serão fornecidos subsídios para o Comitê das Bacias Hidrográficas dos Rios Piracicaba, Capivari e Jundiaí, e também para a indústria, no direcionamento de atitudes técnicas ambientalmente corretas, visando diminuir os danos ambientais causados pela atividade antrópica.

3 REVISÃO BIBLIOGRÁFICA

3.1 Imputação de Dados

Muitas vezes, em bancos de dados reais, há ausência em informações de determinadas observações, também chamadas de dados ausentes ou faltantes, que podem afetar negativamente a análise estatística a ser realizada. No caso dos dados de nosso estudo isso ocorre, sendo necessária a correção dessa nuance antes de seguir a análise estatística, dependendo da técnica utilizada.

A imputação de dados é uma técnica que permite a substituição de dados faltantes por valores estimados para eles, a partir das unidades de um conjunto de dados que foram de fato observadas. O objetivo do uso da imputação é estimar os valores ausentes, proporcionando uma matriz de dados completa, a ser posteriormente analisada por métodos estatísticos.

Existem na literatura alguns padrões e mecanismos para os dados ausentes, definidos por Rubin (1976). Através desses padrões é possível ter uma base de qual dos métodos existentes de imputação de dados são mais propícios a serem utilizados em cada caso.

Os padrões de dados ausentes dizem respeito à maneira com que as unidades ausentes se distribuem dentro de um conjunto de dados, porém, não explicam o motivo de os dados estarem ausentes. Os mecanismos de dados ausentes buscam descrever possíveis relações entre as variáveis que foram medidas e a probabilidade de dados faltantes, assim visando informar a causa da ausência, embora não ofereçam de fato uma explicação causal (Enders, 2006).

Para a realização da imputação de dados, existem diversos métodos,

desde os mais simples até os mais complexos, podendo ser classificados em simples e múltiplos. A imputação simples baseia-se na estimação do valor ausente uma única vez para cada valor em falta, já a imputação múltipla traz várias imputações para cada um dos valores ausentes, podendo-se posteriormente utilizar um método, bem como a média ou a mediana, para encontrar uma medida de tendência central de cada valor ausente. A imputação múltipla permite uma melhor estimativa dos valores em falta, além de trazer maiores informações sobre a variabilidade da tendência central dos mesmos, mas dependendo do mecanismo dos dados pode ser impossibilitada.

3.1.1 Distribuição dos dados ausentes

Segundo a teoria proposta em Rubin (1987), os dados completos são representados por $\mathbf{Y}_{comp}$ e podem ser divididos em valores observados $\mathbf{Y}_{obs}$ e valores ausentes $\mathbf{Y}_{aus}$, ou seja, $\mathbf{Y}_{comp} = (\mathbf{Y}_{obs}, \mathbf{Y}_{aus})$

Seja $\mathbf{Y}$ uma matriz de dados n x p, sendo as linhas $(i = 1, 2, \cdots, n)$ as unidades de uma amostra aleatória de determinada distribuição de probabilidade multivariada p-dimensional das variáveis observadas nas colunas $(j = 1, 2, \cdots, p)$. Temos os valores das variáveis respostas para a j-ésima variável agrupados em um vetor $\mathbf{Y}_j^t = (Y_{1j}, Y_{2j}, \cdots, Y_{rj})$. Considerando a variável $j^{'}$ teremos então $\mathbf{Y}_{j^{'}}^t = (Y_{1j^{'}}, Y_{2j^{'}}, \cdots, Y_{nj^{'}}) = (Y_{1j^{'}}, Y_{2j^{'}}, \cdots, Y_{rj^{'}}, Y_{(r+1)j^{'}}, \cdots, Y_{nj^{'}})$.

Nessa variável temos que $\mathbf{Y}_{obs}^t = (Y_{1j^{'}}, Y_{2j^{'}}, \cdots, Y_{rj^{'}})$ corresponde aos valores observados e $\mathbf{Y}_{aus}^t = (Y_{(r+1)j^{'}}, \cdots, Y_{nj^{'}})$ é referente aos valores ausentes. Com isso, o conjunto de dados possui r valores observados e $m = n - r$ valores ausentes. A fim de realizar o estudo do comportamento dos dados ausentes, Rubin cria uma variável indicadora R, a qual fornece uma distribuição de probabilidade através de uma indicadora de falta para cada casela. Com isso, R_{ij} assume o valor 1 caso o dado seja observado e 0 caso seja um dado faltante. Essa distribuição é chamada de distribuição de probabilidade da falta completa (R). Temos então:

$$R_{ij} = \begin{cases} 1, \ se \ Y_{ij} \ é \ observado \\ 0, \ se \ Y_{ij} \ é \ faltante \end{cases}$$

Essa distribuição dependerá da forma com que os dados faltantes se distribuem ao longo da matriz de dados que será trabalhada, sendo importante quando deseja-se verificar a causa da falta de dados, através dos mecanismos de dados ausentes. Por meio de tais mecanismos podemos verificar as relações entre os dados perdidos e a probabilidade de ausência, assim tendo informações sobre o que gerou tais ausências. Os padrões e mecanismos dos dados ausentes serão descritos a seguir.

3.1.2 Padrões de dados ausentes

Baseado em Rubin (1987) e Molenberghs et al. (2014) podemos definir os padrões e mecanismos dos dados ausentes. É importante a definição de tais padrões para a escolha da metodologia mais adequada de imputação a ser utilizada nos dados com os quais deseja-se completar através da mesma.

Os padrões de dados se baseiam nas informações sobre o comportamento dos dados faltantes no que se diz à localização desses valores em falta. A literatura classifica os padrões de dados ausentes em algumas situações específicas.

A primeira situação é a conhecida por padrão univariado (*univariate pattern*) e ocorre quando a falta surge apenas na única variável em estudo, sendo esse o caso de falta ocorrido em experimentos que envolvem delineamento experimental.

O padrão monótomo (*monotome pattern*) ocorre quando em determinado momento o indivíduo deixa a pesquisa ou não pode continuar. São bastante encontrados em pesquisas clínicas quando é preciso que o indivíduo deixe a pesquisa devido à ocorrência de algo que o impede de continuar, bem como, por exemplo, a reação ao medicamento testado.

Também temos o padrão conhecido por padrão geral (*general pattern*) que também pode ser chamado de arbitrário, no qual as posições de ocorrências de perda aparecem arbitrariamente no conjunto de dados. Podemos ver uma descrição

completa em Enders (2006) e Silva (2012).

Por fim, há o padrão de não resposta (*unit nonresponse pattern*), o qual ocorre quando o indivíduo não responde parte de uma pesquisa por algum motivo, sendo esse caso de comum ocorrência em pesquisas realizadas por meio de questionários. São comuns por diversos motivos referentes ao respondente, seja por cansaço, por não entendimento ou pelo mesmo não desejar responder à determinada pergunta. Esses casos podem ocorrer tanto em uma variável específica quanto em mais variáveis, simultaneamente, o que é de especial interesse para o presente estudo e será representado graficamente na Figura 1.

Dentre tais padrões os principais são o geral e o monótomo (Bergamo, 2007). O padrão monótomo é utilizado pela maior parte dos algoritmos de imputação, mas na prática nem sempre isso ocorre, uma vez que as respostas estão muitas vezes associadas a características do próprio indivíduo (Medina & Galván, 2007). A Figura 1 é utilizada em Little & Rubin (2002) a fim de proporcionar uma melhor visualização de tais padrões:

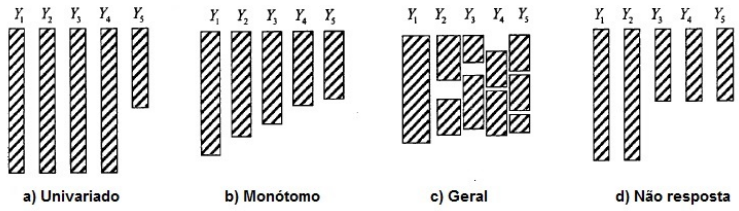


Figura 1: Representação dos padrões de ausência de dados.

3.1.3 Mecanismos de dados ausentes

Quando temos dados ausentes na matriz de dados, o pesquisador deve avaliar o mecanismo que o gerou, a fim de o identificar e considerá-lo na análise dos dados, caso contrário pode haver influência nas inferências estatísticas. Enders (2006); Rubin (1976); Molenberghs et al. (2014) descrevem alguns mecanismos de padrões ausentes.

Para a definição dos mecanismos iremos utilizar as notações $\mathbf{Y}$ como

sendo a matriz completa dos dados, $\mathbf{Y}_{obs}$ como sendo os valores observados de $\mathbf{Y}$, sendo $\mathbf{Y}_{aus}$ os valores ausentes e R é uma variável aleatória associada à probabilidade de ausência.

Primeiramente temos o mecanismo da Ausência totalmente aleatória (*Missing completely at Random)* representado por MCAR. A ausência ocorre dessa maneira quando a probabilidade de dados sobre a variável não é afetada pelas demais variáveis da matriz e nem mesmo pela própria variável em falta. Ela ocorre quando há um fator externo às variáveis em estudo que colabora para a ausência de dados, como, por exemplo, quando plantas são danificadas por fatores fora do controle do pesquisador. Nesse caso existe algum parâmetro ϕ importante para a probabilidade de que $\mathbf{R}$ assuma 0 ou 1, o qual não está relacionado com os dados. Assim temos um mecanismo do tipo:

$$f(\mathbf{R}|\mathbf{Y}, \phi) = f(\mathbf{R}|\phi),\ para\ todo\ \mathbf{Y},\ \phi \tag{1}$$

Há também o mecanismo de Ausência de forma aleatória (*Missing at Random*) representado por MAR. Nesse caso, a ausência de dados depende das informações disponíveis na matriz de dados contendo as variáveis medidas. Sua distribuição de probabilidade indica que a probabilidade da falta completa ($\mathbf{R}$) é dependente da proporção de dados observados, com um parâmetro ϕ relacionando $\mathbf{Y}$ e $\mathbf{R}$. Podemos expressar a distribuição como:

$$f(\mathbf{R}|\mathbf{Y}, \phi) = f(\mathbf{R}|\mathbf{Y}_{obs}, \phi),\ para\ todo\ \mathbf{Y}_{aus},\ \phi \tag{2}$$

Por fim, temos o mecanismo da Ausência não aleatória (*Missing Not at Random*), representado por MNAR. Nesse caso a ausência de dados depende de informações não observadas, da própria variável. Um exemplo da ocorrência do mecanismo MNAR é uma pesquisa onde é perguntando o peso e pessoas com excesso de peso propositalmente deixem em branco essa resposta, causando assim relação entre a ocorrência da ausência e o valor do dado ausente. Sendo assim, a distribuição dos dados em falta depende tanto de $\mathbf{Y}_{obs}$ como de $\mathbf{Y}_{aus}$. Podemos

escrever essa distribuição por:

$$f(\mathbf{R}|\mathbf{Y}, \phi) = f(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{aus}, \phi) \tag{3}$$

Para que os mecanismos de dados ausentes sejam melhor entendidos, Schafer & Graham (2002) os apresenta graficamente baseando-se em um padrão univariado de ausência de dados, como podemos ver na 2. Nessa representação toma-se como X as variáveis totalmente observadas, Y é uma variável que em parte está faltando e Z são as variáveis que causam a falta completa sem relação com X e Y. R representa a indicadora da falta completa. Temos então:

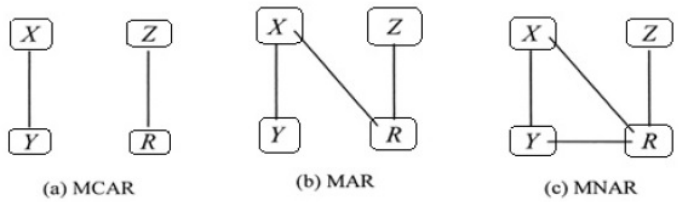


Figura 2: Representação dos mecanismos de ausência de dados.

Na Figura 2 fica claro que em casos de mecanismo MCAR a falta completa, R não tem relação com a variável em falta Y e com a outra variável medida presente no banco de dados, X. Ou seja, a causa do valor ausente está relacionada com fatores aleatórios ou variáveis não medidas. Já no MNAR verificamos que a falta completa, R está relacionada a todos os outros componentes (X, Y e Z).

Quando temos dados com o mecanismo MNAR as unidades ausentes são consideradas não ignoráveis e devemos levar em conta o conhecimento prévio que o gerou (Nunes, 2007). A maior parte dos métodos de imputação múltipla foi desenvolvida inicialmente para o mecanismo MCAR e posteriormente para o MAR. Em casos nos quais temos MNAR há a impossibilidade da utilização de diversos métodos (Rubin, 1987; Molenberghs et al., 2014).

3.1.4 Métodos de estimação de dados ausentes

Em diversas áreas do conhecimento há o problema de dados ausentes em bancos de dados. Isso traz ônus para a realização de análises estatísticas, inclusive impedindo que determinadas técnicas sejam utilizadas a menos que tal problema seja solucionado e a matriz de dados a ser analisada esteja completa.

Pensando nisso há duas possíveis soluções para que a matriz com ausência de dados torne-se completa. Podemos tanto desconsiderar as informações ausentes nos dados, assim descartando todas as variáveis para um indivíduo que possua algum dado faltante em pelo menos uma variável, quanto utilizar métodos a fim de estimar os dados faltantes com uma precisão consideravelmente boa.

Na literatura, ao longo do desenvolvimento de tais técnicas durante os anos, nota-se grande número de alternativas para lidar com os dados faltantes. A princípio pensava-se em remover os indivíduos ou variáveis que apresentassem muitos dados em falta, método conhecido como remoção de dados. Esses métodos classificam-se em análises de caso completo e em análise de descarte de casos (Lobo & Kalil, 2006; Silva, 2012). Quando há poucos dados ausentes esse método pode ser eficiente, porém, quando há muitos dados em falta a perca de informações é considerável, e esse método perde sua eficiência.

Os métodos de imputação múltipla de dados começaram a ser desenvolvidos principalmente por Rubin (Rubin, 1976, 1978). Nesses artigos passam a ser sugeridos métodos para substituir os valores em falta por estimativas obtidas a partir de alguma metodologia estatística. Tal substituição dos dados ausentes é conhecida por imputação de dados, a qual pode ser classificada em simples e múltipla. Dezoito anos após essa proposta, o mesmo autor fez uma análise de métodos existentes em Rubin (1996).

A imputação simples consiste em estimar cada valor em falta uma única vez utilizando as demais informações da matriz de dados para suprir a ausência de dados e substituir os valores em falta, obtendo assim uma nova matriz, completa.

Podemos citar dentre os métodos simples os de imputação pela média

dos dados disponíveis para cada variável, a imputação por meio da regressão linear, a imputação pela máxima verossimilhança, a imputação *hot deck*, a pelo vizinho mais próximo, por dados semelhantes, entre outras. Em Rubin (1976) o autor faz uma crítica à utilização de tais métodos de imputação simples, uma vez que os mesmos apresentam aumento no viés das estimativas. Quando possível é interessante utilizar um método de imputação múltipla para corrigir esse problema, mas isso nem sempre é possibilitado devido ao tipo de dados a ser imputado e seu mecanismo de ausência.

A utilização de métodos de imputação múltipla vem como uma solução plausível para esse problema no viés das estimativas presente na imputação simples. Na imputação múltipla é estimado um conjunto de valores razoáveis para o valor a ser imputado e a partir dele utilizada uma metodologia a fim de verificar o valor mais provável para aquele dado ausente. Com isso, a variabilidade é introduzida e os resultados são mais confiáveis que aqueles obtidos pela imputação simples (Jelicic et al., 2009).

Na literatura há a aplicação de diferentes métodos de imputação múltipla. A maior parte dos trabalhos que desejam estudar a eficiência dos métodos partem de uma matriz de dados completa e retiram aleatoriamente uma quantidade dos mesmos. Uma vez feito isso, são utilizados diferentes métodos de imputação, desde os mais simples até os mais complexos, e comparados os resultados obtidos a partir de cada metodologia, como podemos ver em Silva (2012); Arciniegas-Alarcón (2008).

Ayuyev et al. (2009) propõe um método de imputação que se baseia em agrupamento dinâmico em vários tipos de dados. Esse método pode ser aplicado para matrizes de dados com variáveis categóricas, contínuas ou uma mistura delas, o que é de grande utilidade principalmente em casos de pesquisa de opinião pública. O método, definido como *Dynamic Clustering Imputation - DCI* (Imputação Dinâmica por Clusterização) baseia-se na dissimilaridade das medidas entre os indivíduos em um conjunto de dados, utilizando em seguida estas distâncias no algoritmo de agrupamento a fim de identificar casos similares e executar um grupo dinâmico específico

de imputação dos valores em falta.

Há também os métodos que utilizam o algoritmo EM (*Expectation Maximization*) e a metodologia chamada de AMMI (*additive main effects and multiplicative interaction model*) sugeridos por Gauch Jr (1992). A metodologia AMMI é bastante usada em casos de imputação de dados para estudos de genética e leva em conta para a estimação dos dados faltantes tanto os efeitos aditivos quanto os efeitos da interação do genótipo por ambiente.

Para dados no tempo foi desenvolvido um método de imputação por Junger (2002). Nesse método, desenvolvido e aplicado primeiramente em dados meteorológicos referentes à precipitação ao longo do tempo em diferentes locais de coleta de dados diários, o autor utiliza o algoritmo EM e noções de séries temporais. Um dos pressupostos para sua utilização é o de que os dados seguem a distribuição normal.

Um outro método com estimativas precisas dos valores ausentes pode ser encontrado em Bergamo (2007) e Bergamo et al. (2008). Esse método é conhecido por Imputação de Dados Livre de Distribuição (IMDVS) e é uma extensão múltipla do método outrora sugerido por Krzanowski (1988). A grande vantagem desse método é que ele não exige que a distribuição das observações siga uma distribuição de probabilidade específica, bem como a normal. Assim, ela pode ser usada em casos nos quais não há normalidade sem que sua eficiência seja afetada.

Em Arciniegas-Alarcón et al. (2013) podemos encontrar a proposta de um método de imputação baseado em modelos determinísticos. Os autores propõem um método que utiliza a técnica da validação cruzada por vetor através de um esquema iterativo utilizando a decomposição em valores singulares (DVS) da matriz de dados a passar pelo processo de imputação.

É importante estudar os métodos existentes de imputação de dados e a constante tentativa de obter métodos que forneçam estimativas melhores e com menor viés a fim de selecionar o método mais adequado para ser aplicado em cada banco de dados. A escolha do método que se enquadre nos padrões e mecanismos

de ausência de dados é de grande importância para a qualidade da imputação e posteriormente da análise estatística a ser realizada com a matriz de dados completa (Medina & Galván, 2007).

3.1.5 Imputação de Dados livre de Distribuição

Como dito anteriormente, muitos dos métodos de imputação múltipla partem da suposição de normalidade para sua construção. Nem sempre essa suposição é atingida e para tal é necessário o estudo de métodos alternativos para os casos nos quais isso não ocorre.

Alguns métodos de imputação simples podem ser utilizados nesse caso, mas para que a imputação tenha uma boa qualidade é necessário escolher um método que apresente bom desempenho. Métodos como a substituição pela média ou pelo vizinho mais próximo podem ser úteis, mas trazem, por vezes, um viés grande à imputação, uma vez que levam em conta apenas uma medida de posição, no caso da primeira, ou um único dado, no caso da segunda.

Uma possibilidade para a imputação de dados quando falha a suposição de normalidade é a imputação de dados livre de distribuição por meio da decomposição em valores singulares da matriz de dados. Essa técnica de imputação está presente em Krzanowski (1988). Para tal é utilizado o algoritmo EM e a imputação é realizada. Bergamo et al. (2008) desenvolve essa técnica de maneira a formular uma técnica de imputação múltipla baseada na ideia original da imputação simples proposta por Krzanowski (1988), a qual também tem maior desenvolvimento em Perry (2009).

O método desenvolvido por Krzanowski (1988) parte da afirmação de Good (1969), na qual qualquer matriz $\mathbf{Y}_{(n,p)}$, para todo n, p, pode ser decomposta por valor singular como

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^t, \tag{4}$$

em que $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{V}\mathbf{V}^t = \mathbf{I}_p$ e $\mathbf{D} = diag(d_1, \ldots, d_p)$, com $d_1 \geq d_2 \geq \ldots \geq d_p \geq$

Os autovalores não nulos das matrizes $\mathbf{Y}^t\mathbf{Y}$ e $\mathbf{Y}\mathbf{Y}^t$ são os mesmos e os elementos d_i são a raiz quadrada desses autovalores. A i-ésima linha $\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})$ da matriz $\mathbf{V}_{pxp}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\mathbf{Y}^t\mathbf{Y}$. Já a j-ésima coluna $\mathbf{u}_j = (u_{1j}, \ldots, u_{nj})$ da matriz $\mathbf{U}_{nxp}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\mathbf{Y}\mathbf{Y}^t$. A representação elementar da decomposição é dada por 5.

$$Y_{ij} = \sum_{h=1}^{p} u_{ih} d_h v_{jh}. \tag{5}$$

Krzanowski tomou essa representação da decomposição como base para determinar a dimensionalidade de um conjunto de dados multivariados. Se tivermos uma estrutura dos dados essencialmente H-dimensional ($H < p$), podemos tratar a variação na dimensão resultante $p - H$ como ruído aleatório. As principais características dos dados supostamente estarão no espaço dos H primeiros componentes principais. Assim teremos um modelo de H componentes, conforme apresentado em 6, sendo e_{ij} o ruído.

$$Y_{ij} = \sum_{h=1}^{H} u_{ih} d_h v_{jh} + e_{ij}. \tag{6}$$

Supondo o modelo dado pela Equação (6) para um valor de H ($H = 1, 2, \ldots, p-1$), considerando uma única observação y_{ij} ausente na matriz de dados, temos y_{ij} estimado por

$$\widehat{Y}_{ij}^{H} = \sum_{h=1}^{H} u_{ih} d_h v_{jh}, \tag{7}$$

em que devemos estimar u_{ih}, d_h, v_{jh} a partir do restante dos dados. Quanto mais dados disponíveis tivermos, melhor será a estimativa de tais valores. Tomando $\mathbf{Y}^{(-i)}$ como sendo a matriz de dados obtida ao se retirar a i-ésima linha de $\mathbf{Y}$ e $\mathbf{Y}_{(-j)}$ a matriz dos dados obtida com a retirada da j-ésima coluna de $\mathbf{Y}$, a decomposição de valores singulares dessas matrizes fica

$$\mathbf{Y}^{(-i)} = \overline{\mathbf{U}}\ \overline{\mathbf{D}}\ \overline{\mathbf{V}}^t, \overline{\mathbf{U}} = (\overline{u}_{sh}), \overline{\mathbf{V}} = (\overline{v}_{sh}), \overline{\mathbf{D}} = (\overline{d}_1, \ldots, \overline{d}_p) \tag{8}$$

$$\mathbf{Y}_{(-j)} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^t, \widetilde{\mathbf{U}} = (\widetilde{u}_{sh}), \widetilde{\mathbf{V}} = (\widetilde{v}_{sh}), \widetilde{\mathbf{D}} = (\widetilde{d}_1, \ldots, \widetilde{d}_{p-1}) \tag{9}$$

A estimativa na Equação (6) de u_{ih} e v_{jh}, obtida com o máximo de dados de $\mathbf{Y}$, é dada por $\widetilde{u}_{ih}$ e $\overline{v}_{jh}$, respectivamente. Já d_h pode ser estimado por uma combinação de $\overline{d}_h$ e $\widetilde{d}_h$. Uma forma adequada para tal combinação é $\sqrt{\overline{d}_h}\ \sqrt{\widetilde{d}_h}$ e uma estimativa do valor ausente y_{ij} será dada por:

$$\widehat{Y}_{ij}^{H} = \sum_{h=1}^{H} (\widetilde{u}_{ih}\sqrt{\widetilde{d}_h})(\overline{v}_{jh}\sqrt{\overline{d}_h}) \tag{10}$$

Para seguir o preceito da máxima informação possível dos dados, devemos usar o valor mais elevado disponível de H. De 9 esse valor é $p-1$ e assim o valor imputado será dado por

$$\widehat{Y}_{ij} = \sum_{h=1}^{p-1} (\widetilde{u}_{ih}\sqrt{\widetilde{d}_h})(\overline{v}_{jh}\sqrt{\overline{d}_h}) \tag{11}$$

Para utilizar esse processo precisamos de uma estimativa inicial para os valores y_{ij} ausentes. Essas estimativas iniciais são feitas por meio da média $\overline{y}_j$ da j-ésima coluna. Para evitar que a escala ou outros fatores de diferenciação entre as colunas influenciem nas imputações, é recomendado aplicar uma padronização em $\mathbf{Y}$.

Para os valores Y_{ij}, utilizando a matriz já completada com os valores ausentes substituídos pela média $\overline{Y}_j$, calcula-se uma nova média ($\overline{Y}_j^{'}$) e desvio padrão (dp_j) para cada uma das colunas j. Então é aplicada a padronização $Y_{ij}^{'} = \frac{Y_{ij} - \overline{Y}_j^{'}}{dp_j}$. Essa mesma padronização também deve ser feita nas matrizes $\mathbf{Y}^{(-i)}$ e $\mathbf{Y}_{(-j)}$.

As estimativas para cada valor ausente da matriz de dados são recalculadas utilizando a Equação (11) nas matrizes padronizadas. Para cada valor estimado são necessárias duas decomposições, uma para cada i e para cada j necessários. O

processo iterativo é aplicado e continua até que a estabilidade nos valores imputados seja alcançada. Finalmente, agora com a matriz $\mathbf{Y}$ completada, aplica-se uma operação para retorno dos dados à sua escala original. Para tal, se y_{ij}^{c} representa cada valor da matriz $\mathbf{Y}$ completada, é calculada novamente a média de cada coluna j ($\overline{y}_j^{(c)}$) e também o desvio padrão (s_j^c). Assim, cada valor da matriz $\mathbf{Y}$ completada, em sua escala original, é obtido por $y_{ij} = \overline{y}_j^{(c)} + s_j^c y_{ij}^c$.

Uma vez feitos todos esses passos, é obtida a matriz completa através do método de imputação simples por meio da decomposição de valor singular, livre de distribuição e que pode ser aplicada a qualquer matriz de dados numéricos. Em Bergamo et al. (2008) é proposta uma alteração na metodologia desenvolvida por Krzanowski (1988) para transformar a imputação livre de distribuição utilizando a DVS em uma técnica de imputação múltipla. A modificação proposta faz com que seja possível a geração de M diferentes imputações.

O autor propõe que, na primeira etapa da Imputação Múltipla, haja uma mudança nos expoentes dos radicandos $\overline{d}_h$ e $\widetilde{d}_h$ em 11. De uma maneira genérica, se $\sqrt[b]{d^a}$ for representada por meio de uma potência fracionária $d^{\frac{a}{b}}$, o procedimento passa a requerer a mudança no numerador do expoente, tanto de $\widetilde{d}^{\frac{\widetilde{a}}{b}}$ como de $\overline{d}^{\frac{\overline{a}}{b}}$, de maneira que a soma dos expoentes resulte em 1 ($\frac{\widetilde{a}+\overline{a}}{b} = 1$). Como já foi definido anteriormente, Krzanowski (1988) sugere a combinação resultando na forma $\sqrt{\overline{d}_h}\,\sqrt{\widetilde{d}_h}$, a qual admite influências iguais de 8 e 9. Assim, ao variar os expoentes de $\overline{d}_h$ e $\widetilde{d}_h$, admite-se um peso maior para 8 e 9 na estimativa final obtida por 11.

Cada mudança em $\widetilde{a}$, que consequentemente também causará mudança em $\overline{a}$, gera uma nova matriz completada $\mathbf{Y}$. Assim temos um processo de geração de M conjuntos de dados completados, caracterizando a primeira etapa da Imputação Múltipla. No desenvolvimento de sua técnica, Bergamo et al. (2008) parte do pressuposto que o número de imputações vai depender das mudanças nos expoentes. Segundo Rubin (1987) e Schafer & Graham (2002) um número M entre 3 e 5 imputações já é o suficiente para expressar a variabilidade entre as imputações. Com 5 imputações há uma variação entre 40% e 60% nos pesos dados a 8 e 9. O autor

nos informa que, por exemplo, para um denominador fixo como $b = 20$, os valores que $\widetilde{a}$ assume $(8, 9, 10, 11$ e $12)$ e respectivamente os assumidos por $\overline{a}$ $(12, 11, 10, 9$ e $8)$ levam a uma variação $(40\%, 45\%, 50\%, 55\%$ e $60\%)$ nas proporções de 8 e 9 na equação:

$$\widehat{Y}_{ij} = \sum_{h=1}^{p-1} (\widetilde{u}_{ih} \widetilde{d}_h^{\frac{\widetilde{a}}{b}})(\overline{v}_{jh} \overline{d}_h^{\frac{\overline{a}}{b}}) \tag{12}$$

Essa metodologia leva em conta a maior quantidade de dados possível da matriz Y e não depende de nenhuma distribuição da variável resposta, bem como ocorre no método de Imputação Simples de Krzanowski (1988). Essa técnica pode ser aplicada a qualquer matriz numérica, sem suposições de mecanismo dos dados ausentes.

Bergamo desenvolve um algoritmo no sistema estatístico SAS para implementar sua técnica de imputação. São realizadas 5 imputações diferentes e depois obtida a média entre as imputações e seus erros padrão. O autor dessa metodologia ainda compara os resultados obtidos através de uma matriz de dados completa com algumas caselas retiradas aleatoriamente, a fim de verificar a validade do método.

3.1.6 Qualidade do Ajuste da Imputação de Dados

Para verificar se a estimação de dados ausentes apresenta bons resultados é necessário verificar algumas estatísticas que permitam que isso ocorra. A maior parte dessas estatísticas baseiam-se na comparação de valores observados com valores obtidos através da imputação.

Para casos nos quais são realizadas simulações a partir de dados completos retirados aleatoriamente basta fazer tal procedimento, como pode ser visto, por exemplo, em Arciniegas-Alarcón et al. (2014) e Arciniegas-Alarcón et al. (2011). Para casos nos quais há ausência de dados nos originais, uma possível abordagem é a de tomar apenas as linhas com dados completos e retirar aleatoriamente alguns dados das mesmas a fim de verificar a qualidade do ajuste com base nesses dados.

Algumas das estatísticas que podem ser úteis, as quais são utilizadas nos artigos supracitados, são as chamadas NRMSE (Ching et al., 2010) e PRESS, além da possibilidade do uso da correlação de Spearman. Existe também a medida de acurácia proposta por Bergamo (2007), adaptada de Penny & Jolliffe (1999), chamada de T_{acc}.

A estatística PRESS nada mais é que a soma de quadrados residual obtida através da comparação entre a matriz de dados imputados e a original. Essa estatística será também parte integrante da T_{acc}. Assim:

$$PRESS = \sum_{l=1}^{na} (\widehat{Y}_l - VO_l)^2, \tag{13}$$

com na sendo o número de ausências, $\widehat{Y}_l$ o valor imputado para a ausência l e VO_l o valor original da matriz para a observação l.

Para a NRMSE temos:

$$NRMSE = \frac{\sqrt{média(\mathbf{a}_{imp} - \mathbf{a}_{orig})^2}}{dp(\mathbf{a}_{orig})}, \tag{14}$$

na qual $\mathbf{a}_{imp}$ é o vetor contendo os valores preditos, $\mathbf{a}_{orig}$ é o vetor contendo os valores originais e dp representa o desvio padrão.

Quanto menor o valor da NRMSE melhor a qualidade da imputação.

A T_{acc} pode ser calculada para casos nos quais há ajustes de imputação múltipla, tendo o seguinte procedimento:

$$acc_l = \frac{\sum_{m=1}^{M} (\widehat{Y}_{l(m)} - VO_l)^2}{M - 1}, \tag{15}$$

sendo que M é o número de vezes que o processo de imputação é repetido, VO é o valor original retirado da posição $l = 1, 2, \ldots, na$, na qual na representa o número de valores retirados correspondentes à linha i e coluna j $(i; j)$ dos dados e $\widehat{y}_{ij}$ representa o valor imputado pelo método proposto.

Para obter uma medida geral da acurácia podemos tomar uma média de acc_l. Assim temos:

$$T_{acc} = \frac{\sum_{l=1}^{na} acc_l}{na}. \tag{16}$$

A T_{acc} possui dois componentes, sendo que o primeiro avalia a variância entre as imputações e o segundo representa o viés quadrático médio, assim:

$$T_{acc} = V_E + VQM. \tag{17}$$

em que:

$$V_E = \frac{1}{na} \sum_{l=1}^{na} \frac{\sum_{m=1}^{M} (\widehat{Y}_{l(m)} - \overline{Y}_l)^2}{M-1} \tag{18}$$

e

$$VQM = \frac{1}{na} \sum_{l=1}^{na} \frac{M(\overline{Y}_l - VO_l)^2}{M-1}, \tag{19}$$

com $\overline{Y}_l$) sendo a média das M imputações realizadas e as demais variáveis tendo a mesma interpretação das equações anteriores.

Quanto menor o valor de ambos e consequentemente de T_{acc} melhor o ajuste.

O coeficiente de correlação de Spearman é útil para verificar o quão correlacionados estão os dados imputados com os valores da matriz original. Com isso, quanto maior for o valor obtido, há evidências de que melhor é o ajuste da imputação de dados. Essa medida de correlação é baseada em postos e é um método não paramétrico (Maritz, 1981; Connover, 1980).

Esse coeficiente de correlação pode ser definido por:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \tag{20}$$

na qual d_i=$x_i - y_i$ indica a diferença entre os postos dos pares x_i, y_i das variáveis em estudo.

3.2 Métodos de Avaliação da Qualidade da Água

3.2.1 Índice de Qualidade da Água

Um outro intuito do trabalho é verificar a qualidade da água nos pontos selecionados. Para tal, é importante que seja realizada uma revisão da literatura a fim de selecionar as variáveis úteis e o método de análise a ser utilizado.

Diversas variáveis físicas, químicas e biológicas podem ser utilizadas para trazer informação sobre a qualidade da água. Tanto variáveis referentes a condições químicas e físicas da água quanto dados sobre a fauna e a flora presentes servem como parâmetro para inferir sobre o quão boa ou ruim está a qualidade da água em determinado local.

A fim de sintetizar algumas variáveis úteis e utilizar mais de uma variável ao mesmo tempo para concluir sobre a qualidade da água, existem os índices de qualidade da água, que são diversos e devem ser analisados para verificar-se a adequação dos mesmos em cada estudo realizado. Uma boa revisão da literatura desses índices é fornecida pelo PNMA II (Programa Nacional do Meio Ambiente), podendo ser consultado em CPRH (2013a). Também esse programa fornece um documento que auxilia na seleção de indicadores, dizendo as variáveis dos mesmos e assim permitindo que o pesquisador verifique qual indicador é cabível em sua pesquisa. Esse documento está disponível em CPRH (2013b).

No caso de nosso estudo nos atentaremos de maneira especial para o Índice de Qualidade das Águas (IQA) e nessa parte de revisão de literatura o índice a ser melhor explicado será esse, o qual se mostrou o mais adequado a ser utilizado com base nas variáveis disponíveis.

O IQA é utilizado desde 1975 pela CETESB, conforme pode ser consultado em CPRH (2013a). As informações obtidas através desse índice são utilizadas para o gerenciamento das vinte e duas Unidades de Gerenciamento dos Recursos Hídricos do Estado de São Paulo.

Para análises mais pormenorizadas é indispensável o estudo individual

das variáveis, porém para uma visualização do panorama geral das bacias os índices são de extrema valia, combinando diferentes medidas em um único valor. Além disso, a facilidade de comunicação com o público é maior através de índices como o IQA.

Segundo Derısio (2007) Brown, McClelland, Deininger e Tozer desenvolveram, em 1970, um índice financiado pela National Sanitation Foundation, índice conhecido na literatura brasileira por IQA-NSF (WQI-NSF em inglês), o qual serviu de base para a construção do IQA da CETESB.

Ainda segundo o autor, o IQA-NSF foi construído baseado na técnica de Delphi da Rand Corporation e utilizou um procedimento formal para combinar as opiniões de 142 especialistas. Após os membros da pesquisa responderem a vários questionários, suas respostas foram tabuladas e cada participante recebeu-as de volta para comparar com as respostas dos demais especialistas. Depois dessa etapa eles analisaram as respostas até chegarem a um consenso sobre as variáveis e seus pesos. As variáveis selecionadas foram: Oxigênio Dissolvido, Coliformes Fecais, pH, Demanda Bioquímica de Oxigênio, Temperatura, Fosfato Total, Nitrato, Turbidez e Resíduos Totais.

A partir de trinta e cinco variáveis presentes no começo do estudo, foram selecionadas nove que viriam a compor o IQA-NSF. Para cada variável foi estabelecida uma curva de variação da qualidade da água, que atribuiria um valor para aquela variável de acordo com os parâmetros encontrados na mesma.

A estrutura original, proposta por Brown et al. (1970), é o resultado de uma combinação linear dos subíndices obtidos para cada variável em estudo. A forma aditiva para o cálculo do índice é a seguinte:

$$IQA - NSF = \sum_{i=1}^{n} w_i q_i, \tag{21}$$

na qual: IQA-NSF é um valor entre 0 e 100; w_i é o peso relativo do i-ésimo parâmetro, sendo que $\sum_{i=1}^{n} w_i = 1$; q_i é o valor do subíndice relativo ao i-ésimo parâmetro, vindo da curva para aquela variável; n refere-se ao número de variáveis que entram no cálculo do IQA (nove).

Landwehr & Deininger (1976) propuseram uma forma multiplicativa do indicador. A justificativa para isso é a de que utilizando a forma aditiva eventuais problemas graves são mascarados. Na forma multiplicativa quando um dos subíndices tem um valor extremamente baixo o índice todo tende a ser menor, o que não ocorre em tão grande escala na aditiva. Nessa nova forma de calcular o IQA os pesos passam a ser potências dos sub-índices, conforme a Equação (22).

$$IQA - NSF = \prod_{i=1}^{n} {q_i}^{w_i}, \tag{22}$$

sendo que o significado de cada componente é o mesmo da Equação (21).

A CETESB, baseando-se nesse índice, fez uma adaptação para a análise das águas em solo brasileiro. Novos pesos foram obtidos e ao invés da utilização do nitrato total passou-se a utilizar o Nitrogênio Total na composição do índice. Também é utilizada a quantidade de fósforo total ao invés de fosfato total.

O cálculo do IQA-CETESB segue a Equação (22). A Figura 3 mostra as curvas para cada variável que compõe o IQA, as quais são iguais para ambos os índices, só tendo como alteração o fato de que se fossem relativas ao IQA-NSF deveria-se substituir fósforo por fosfato e nitrogênio total por nitrato. Os índices do IQA-NSF estão disponíveis em Brown et al. (1970) e os da CETESB em CETESB (2013).

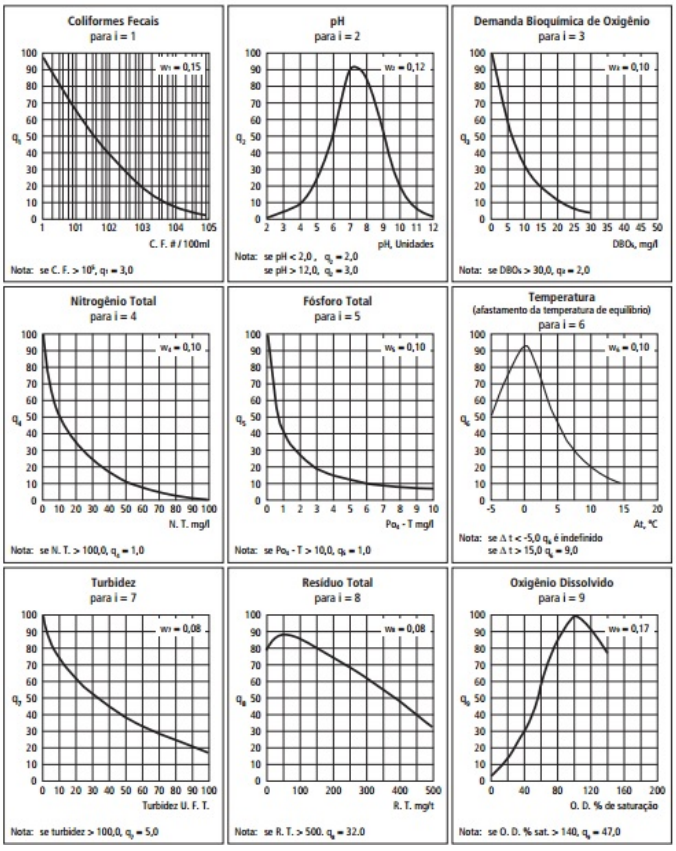


Figura 3: Curvas de classificação de subíndices.

As equações aproximadas dessas curvas podem ser consultadas em Grunitzki et al. (2013). Nesse estudo são informadas as curvas para o IQA-NSF, mas ela vale para ambos, as alterações no índice da CETESB ocorrem nos pesos e na variável nitrogênio, mas não nas curvas para o valor dos parâmetros. No Brasil usualmente utiliza-se o valor máximo para o subíndice relativo à temperatura, pois é considerado que não há grandes diferenças com relação à temperatura de equilíbrio (Grunitzki et al., 2013; Rino et al., 2001).

Na Tabela 6 pode ser visto o valor dos pesos (w_i) para o IQA-CETESB e para o IQA-NSF. Na Tabela 2 pode ser visualizada a classificação da água de acordo com o IQA-NSF e na Tabela 3 de acordo com o IQA-CETESB para o estado de São Paulo.

Tabela 1: Pesos das variáveis do IQA

Variável	IQA-CETESB	IQA-NSF
Oxigênio Dissolvido	0,17	0,17
Coliformes Fecais	0,15	0,16
pH	0,12	0,11
Demanda Bioquímica de Oxigênio	0,10	0,11
Nitrogênio Total (Nitrato Total)	0,10	0,10
Fósforo Total (Fosfato Total)	0,10	0,10
Temperatura	0,10	0,10
Turbidez	0,08	0,08
Resíduo Total	0,08	0,07

Tabela 2: Classificação da Qualidade da Água IQA-NSF

Categoria	Resultado
Excelente	$100 \geq \text{IQA} \geq 90$
Boa	$90 > \text{IQA} \geq 70$
Média	$70 > \text{IQA} \geq 50$
Ruim	$50 > \text{IQA} \geq 25$
Muito Ruim	$25 > \text{IQA} \geq 0$

Tabela 3: Classificação da Qualidade da Água IQA-CETESB

Categoria	Resultado
Ótima	$79 < \text{IQA} \leq 100$
Boa	$51 < \text{IQA} \leq 79$
Regular	$36 < \text{IQA} \leq 51$
Ruim	$19 < \text{IQA} \leq 36$
de Péssima	$0 \leq \text{IQA} \leq 19$

3.3 Análise de Variância

A análise de variância (ANOVA) é bastante utilizada afim de verificar o efeito de dois ou mais níveis de tratamento em um estudo. Há variações da ANOVA, uma vez que existem diferentes tipos de experimentos que podem ser realizados.

Para um experimento com delineamento inteiramente aleatorizado temos que cada observação Y_{ij} pode ser decomposta de acordo com o seguinte modelo (Montgomery et al., 2001; Scheffé, 1999):

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \qquad i = 1, \ldots, I; j = 1, \ldots, J \tag{23}$$

no qual Y_{ij} é a observação da j-ésima repetição do i-ésimo tratamento; μ é a média geral (efeito constante); τ_i é o efeito do i-esimo tratamento; ϵ_{ij} é o erro associado ao i-ésimo tratamento na j-ésima repetição que tem por suposição que $\epsilon_{ij} \overset{IID}{\sim} N(0, \sigma^2)$, com os erros independentes e identicamente distribuídos.

O interesse na aplicação da ANOVA consiste em testar se há diferenças significativas entre as médias dos tratamentos, o que é equivalente a testar as hipóteses:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \ldots = \tau_I = 0 \\ H_1 : \tau_i \neq 0, \textit{para pelo menos um } i \end{cases}$$

Também podemos escrever esses contrastes da seguinte maneira:

$$\begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_I = 0 \\ \mathrm{H}_1 : \mu_i \neq \mu_{i'}, \textit{para pelo menos um } i \textit{ e } i^{'}, \textit{ com } i \neq \ i^{'} \end{cases}$$

sendo $\mu_i = \mu + \tau_i$.

Caso as hipóteses nulas forem verdadeiras, a média para todos os tratamentos serão iguais. A base para a realização da análise de variância está na decomposição da variabilidade total da variável resposta em duas partes, a atribuída aos tratamentos e a atribuída ao erro experimental. Essas variações podem ser medidas através das seguintes somas de quadrados:

$$SQ_{Tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y}_{i.})^2 + \sum_{i=1}^{I}\sum_{j=1}^{j=J}(\overline{y}_{i.} - \overline{y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2 - C \tag{24}$$

em que $y_{i.} = \sum_{j=1}^{J} y_{ij}, y_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}$ e $C = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2}{IJ}$

Tem-se também:

$$SQ_{Trat} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{y}_{i.} - \overline{y}_{..})^2 = \frac{\sum_{i=1}^{I} y_{i.}^2}{J} - C \tag{25}$$

E por fim:

$$SQ_{Res} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \overline{y}_{i.})^2 = SQ_{Tot} - SQ_{Trat} \tag{26}$$

A SQ_{Trat} mede a variação entre os tratamentos diferentes, enquanto a SQ_{Res} mede a variação dentro de cada tratamento, existente devido às repetições dentro do mesmo tratamento.

Podemos a partir dessas somas de quadrado definir a tabela da ANOVA, apresentada na Tabela 4.

Tabela 4: Tabela da ANOVA para experimento inteiramente aleatorizado

Fonte de Variação	Graus de Liberdade	Somas de Quadrados	Quadrados Médios	Valor F
Tratamentos	I-1	SQ_{Trat}	QM_{Trat}	QM_{Trat}/QM_{Res}
Resíduo	I(J-1)	SQ_{Res}	QM_{Res}	
Total	IJ-1	SQ_{Tot}		

em que $QM_{Trat} = SQ_{Trat}/(I-1)$ e $QM_{Res} = SQ_{Res}/[I(J-1)]$.

Pode-se mostrar que quando y_{ij} são variáveis aleatórias independentes, todos os tratamentos tem variâncias iguais a σ^2 e $Y_{ij} \sim N(\mu_i, \sigma^2)$, então QM_{Trat} $/QM_{Res}$ possui distribuição F com $I-1$ e $I(J-1)$ graus de liberdade.

Caso o $Valor\ F$ para os tratamentos for maior que o valor tabelado da distribuição F, ao nível α de significância, ou seja, $F_{[\alpha,(I-1),I(J-1)]}$, então rejeitamos a hipótese nula. Nesse caso conclui-se que há evidências de diferenças significativas em ao menos um contraste dois a dois entre as médias de tratamentos, ao nível de significância escolhido α. Caso contrário não rejeitamos H_0, ou seja, não existem evidências de diferença significativa, ao nível α entre os níveis de tratamento.

Podem ser adicionadas outras variáveis, blocos, entre outros na análise de variância, seguindo a mesma lógica utilizada na formulação da ANOVA básica apresentada nesse tópico.

3.4 Bootstrap

Métodos usualmente utilizados para comparação de médias, como a Análise de Variância, possuem algumas suposições que nem sempre são de fato verificadas. Quando tais suposições não são válidas para determinados dados, surgem algumas propostas de análises mais adequadas para esses casos.

Entre tais propostas encontram-se os métodos de Estatística Não-Paramétrica (Connover, 1980). Tais métodos não tem por suposição uma determinada distribuição de probabilidade para os dados, porém, mesmo assim supõem, por muitas vezes, independência entre os dados da amostra.

Para casos onde nem mesmo a suposição de independência dos dados amostrais é verificada, o poder do teste, tanto para métodos paramétricos como a ANOVA, quanto para alternativas de estatística não paramétrica é bastante limitado.

Um dos métodos alternativos para análise e infererência estatística, que não é tão dependente de suposições, é o Bootstrap não paramétrico. Esse método é baseado em reamostragem, através da geração de amostras do mesmo tamanho da original a partir da amostragem aleatória simples com reposição.

O Bootstrap foi introduzido primeiramente por Efron (Efron, 1979), inspirando-se em trabalhos anteriores sobre o método chamado de Jackknife (Tukey, 1958; Jaeckel, 1972). O termo "Bootstrap" é originário da frase *"to pull oneself up by one's bootstrap"*, em alusão a uma situação na qual uma pessoa está afundando em um lago e quando acha que tudo está perdido a única solução é imergir por um puxão em um dos cadarços do sapato.

Como a origem do nome do método já sugere, o Bootstrap surge como uma alternativa para casos nos quais os métodos usuais de análise já foram esgotados e não se apresentaram como boas soluções. A realização do Bootstrap é possível

devido ao avanço tecnológico, sendo necessária a utilização de computadores para que ele seja realizado.

A ideia básica da técnica consiste em, através dos dados disponíveis na amostra, reproduzir o mecanismo probabilístico que os gerou. Para casos com a distribuição de probabilidade desconhecida dos dados é possível substituí-la por uma distribuição empírica que possa aproximá-la.

A Figura 4 mostra a lógica do método, e pode ser encontrada em Renno (2011):

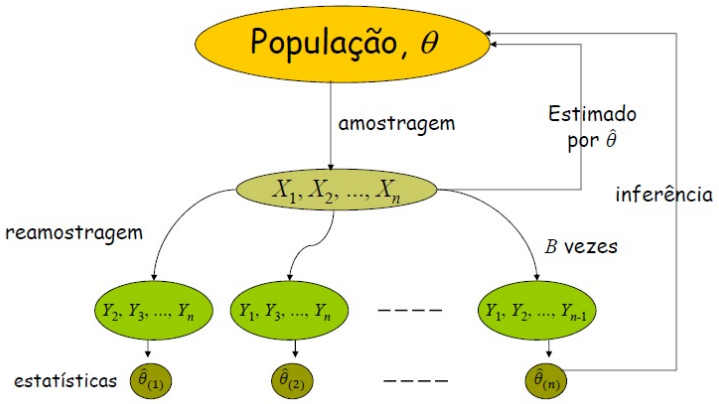


Figura 4: Lógica do Método Bootstrap

Casos da utilização do método para casos com dados dependentes entre si destacam-se principalmente na aplicação em regressão linear simples ou mútipla. Nesses casos para a reamostragem deve-se levar em conta a dependência da variável resposta com relação às preditoras, assim reamostrando o par (para o caso simples) ou o vetor conjunto (para o caso múltiplo) ao mesmo tempo para cada reamostragem (Silva, 1995; Montgomery et al., 2001).

Para a obtenção da estimativa de uma dada estatística através do Bootstrap, é preciso a implementação de um algoritmo que siga alguns passos. Quando temos dados dependentes ou correlacionados que devam ser utilizados em conjunto, como é o caso da regressão, esse algoritmo, no processo de seleção da amostra, deve levar em conta todas as variáveis simultaneamente e não realizar a reamostragem de

maneira separada para cada variável.

Temos assim o seguinte algoritmo:

1. Obter $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, amostra original para cada variável i contendo n observações

2. Construir B reamostras $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \ldots, \mathbf{x}^{*(B)}$, i.i.d. com reposição, de comprimento n cada. Para casos como o da regressão, essas amostras devem conter valores para cada variável i de maneira que a observação j seja a mesma para a seleção de todas as variáveis, simultaneamente. Ou seja, caso seja selecionado x_{i1} é necessário que nas outras variáveis a seleção seja do valor $x_{i'1}$.

3. Calcular para cada amostra $\mathbf{x}^{*(b)}$ a estatística de interesse $\theta^{*(b)} = t(\mathbf{x}^{*(b)})$, $b = 1, 2, \ldots, B$.

4. Através das estatísticas obtidas, podemos obter diversos resultados de interesse (média, variância, intervalo de confiança, entre outros).

Assim, se desejamos obter uma estimativa bootstrap para a média, basta calcular a média a cada reamostragem e ao final das B reamostras calcular a média entre as mesmas. O mesmo vale para a variância, por exemplo, basta calcular a variância de cada reamostra e ao fim a média entre as variâncias obtidas.

Para a obtenção do Intervalo de Confiança, basta avaliar os quantis dos dados reamostrados. Se deseja-se construir um Intervalo de Confiança para a média com $\alpha\%$ de significância, deve-se tomar os quantis $\alpha/2\%$ e $1 - \alpha/2\%$, a menos em casos nos quais há sérios problemas de ausência simetria, nos quais são necessários métodos que corrijam tal ausência Efron & Tibshirani (1993).

Também é possível a obtenção de medidas de precisão do Bootstrap. Para o cálculo da variância bootstrap para uma estatística $\theta^{*(b)}$ obtida a cada uma das reamostras, por exemplo, basta calcular a variância entre as médias obtidas, conforme na Equação (27):

$$\widehat{\sigma}_B^{2\,*} = \frac{1}{B-1}\sum_{i=1}^{B}(\theta^{*(i)} - \overline{\theta^*})^2. \tag{27}$$

Outra medida útil que pode ser obtida é o viés do Bootstrap, o qual pode ser obtido por:

$$\widehat{Vi\acute{e}s}_B^{\,*} = \frac{\sum_{i=1}^{B}\theta^{*(i)}}{B} - \theta(\hat{F}_n), \tag{28}$$

na qual $\theta(\hat{F}_n)$ é a média estimada obtida através da amostra original.

O número B de reamostras para a realização das análises estatísticas depende do intuito do estudo. Efron & Tibshirani (1993) indicam que $B = 200$ dá resultados bastante bons para a estimativa da variância. Relativamente à construção de intervalos de confiança bootstrap o número de reamostras deve ser maior, devendo ser realizadas cerca de $B = 1000$ réplicas.

Esse algoritmo é útil para o caso em que a dependência entre as variáveis em estudo deve ser levada em conta para a construção metodológica da análise, porém ainda tem maior validade em casos nos quais os dados dentro de cada variável sejam independentes e identicamente distribuídos, sendo necessário o uso de uma metodologia alternativa para os casos nos quais há dependência interna nas observações de cada variável.

Uma das alternativas para o uso do Bootstrap para dados dependentes, dentro de cada variável, é o chamado "Bootstrap em blocos" *(Block Bootstrap)*. Essa metodologia tem bastante utilidade em dados de séries temporais e análise de *cluster* (Politis, 2003; Cameron et al., 2008). Em casos de séries temporais que possuem estacionariedade nem sempre o Bootstrap usual produz resultados corretos. Se a dependência for negligenciada podem ocorrer erros consideráveis na estimação (Singh, 1981).

Inicialmente essa técnica foi proposta para séries temporais estacionárias, mas existem trabalhos que mostram que até mesmo para séries que fogem um pouco da estacionariedade mas se aproximam dela, os resultados são bastante bons (Synowiecki, 2007).

A utilização do Bootstrap em blocos em lugar da abordagem paramétrica para modelos de séries temporais é atrativa devido a sua simplicidade de aplicação e boa qualidade dos resultados. Em Vogel & Shalcross (1996) podemos ver uma defesa do uso da metodologia de Bootstrap em blocos para dados de séries temporais em um problema de recursos da água.

Nesse estudo de Vogel e Shallcross é mostrado que até mesmo o bootstrap simples apresenta resultados razoáveis, o que leva a pensar que a introdução da estrutura de correlação dentro da série temporal no Bootstrap produziria excelentes resultados. O uso de blocos móveis faz com que tais resultados tenham uma grande melhoria principalmente na diminuição do erro quadrático médio do estimador. Os autores utilizam para seus testes a técnica de blocos móveis, proposta por Künsch (1989), citando também o Bootstrap circular (Politis & Romano, 1992).

A lógica de utilização do Bootstrap continua sendo a mesma para o caso de blocos. A alteração ocorre principalmente na unidade reamostrada. Enquanto para o Bootstrap usual cada dado é reamostrado individualmente, quando são utilizados blocos a unidade reamostrada é o bloco, assim levando em conta todas as informações em redor de um ponto, as quais são dependentes do mesmo.

Além da metodologia proposta por Künsch (1989), já citada, desenvolvida também por Liu & Singh (1992) e que usa o Bootstrap em blocos com reposição, existem outros métodos para a realização do bootstrap em blocos. Carlstein (1986) propôs o Bootstrap em blocos sem reposição. Politis & Romano (1992) desenvolveram o Bootstrap circular em blocos e Politis & Romano (1994) o Bootstrap estacionário .

O Bootstrap estacionário é menos sensível à especificação do tamanho dos blocos com relação às outras técnicas de Bootstrap em blocos, porém, Nordman & Lahiri (2007) concluem que a variância dos estimadores do Bootstrap estacionário é pelo menos o dobro das demais. Quanto ao viés, Nordman & Lahiri (2007) diz que os resultados para os métodos de Bootstrap em bloco e o método de Bootstrap estacionário são os mesmos.

Cada uma das metodologias apresentadas apresenta suas particularidades. Já considerando o algoritmo do Bootstrap usual anteriormente explicitado como base, um algoritmo geral para implementação do Bootstrap em blocos segue os seguintes passos:

1. Tome a amostra original $X_1, X_2, \ldots, X_N$ e a "envolva" em sequências, isto é, defina novas séries de dados Y_t compostas por elementos de X_t

2. Tome $i_0, i_1, \ldots$, i.i.d. de uma distribuição uniforme no conjunto $1, 2, \ldots, N$. Esses são os pontos iniciais de cada bloco.

3. Tome $l_0, l_1, \ldots$, i.i.d. de alguma distribuição $F_l(.)$ que dependa do parâmetro l. Esses são os tamanhos dos blocos, que podem tanto ser considerados fixos como dependentes de uma distribuição de probabilidade.

4. Construa uma série pseudo-aleatória de Bootstrap $Y_1^*, Y_2^*, \ldots$, reamostrando ao invés das unidades X_t os blocos Y_t construídos a partir de tais unidades.

5. Por fim, basta se focar nos N primeiros pontos de cada série Bootstrap e construir a média aritmética Bootstrap ($\overline{Y}^*_N = N^{-1} \sum_{i=1}^{N} Y_i^*$). A estimativa da variância assintótica da média aritmética é dada por $Var^*(\frac{\overline{Y}^*_N}{\sqrt{N}})$.

6. Com isso temos as estimativas para a média e a variância da média ao longo da realização das reamostragens Bootstrap e podemos utilizá-las da mesma maneira que no Bootstrap usual para obter estimativas.

3.4.1 Definição do tamanho dos blocos

Uma importante etapa para a implementação dos métodos de Bootstrap em blocos é a definição do número de blocos utilizados, o que está relacionado também com o tamanho que os blocos terão.

Na literatura existem diferentes sugestões referentes à escolha do número de blocos utilizados. Hall & Horowitz (1996) utiliza 5 blocos para uma

amostra de tamanho 50 e 10 para uma de tamanho 100. Inoue & Shintani (2003) seleciona o tamanho dos blocos através de procedimentos automáticos, bem como Politis & White (2004) o fazem para o Bootstrap em blocos estacionário e o circular ou de blocos móveis. Em Patton et al. (2009) são corrigidos alguns problemas de Politis & White (2004) e o método é aprimorado.

Os procedimentos automáticos propostos em Politis & White (2004) e Patton et al. (2009) baseiam-se na estrutura de dependência temporal dos dados, na variância bootstrap e em seu erro quadrático médio. Por meio de diversas manipulações algébricas os autores chegam a uma forma de otimizar o procedimento da escolha do tamanho dos blocos minimizando esse erro quadrático médio, sempre levando também em conta a estrutura da dependência temporal.

Há nesses artigos toda a base teórica para a formulação dessa escolha automática, simulações para teste de seu desempenho e também um código da programação para sua implementação, a qual está disponível também no software estatístico R (R Core Team, 2015).

Algumas das maiores críticas ao Bootstrap em blocos são as referentes à escolha do tamanho dos blocos, que em alguns dos métodos é bastante sensível, e também quanto à suposição de independência entre os blocos. Por mais que a utilização de blocos corrija parcialmente o problema de dependência levando em conta tal estrutura, há alguns valores dentro de cada bloco que podem ter alguma dependência com valores de outros blocos (Liu & Singh, 1992; Davison & Hall, 1993).

3.4.2 Bootstrap em Blocos móveis

Como já dito anteriormente, uma das técnicas utilizadas para o uso do Bootstrap em casos de dados dependentes é a desenvolvida por Künsch (1989).

No caso da aplicação dessa técnica, ao invés de se trabalhar com as N observações da amostra original, as mesmas são divididas em blocos de tamanho l. Vamos definir k que faça com que $N \approx kl$, $k \in N$. Para o caso mais simples, sem sobreposição de blocos, teremos que a relação entre o número de blocos (b) e

seu tamanho (l) será dado por $N = bl$. Considerando sobreposições de blocos, essa relação é dada por $b = N - l + 1$, pois haverá elementos da amostra original contidos em mais de um bloco e sem sobreposição por $b = N/l$.

Para o caso sem sobreposição de blocos, a reamostragem é feita com base nos blocos $(Y_1^*, \ldots, Y_l^*), (Y_{l+1}^*, \ldots, Y_{2l}^*), \ldots, (Y_{N-l+1}^*, \ldots, Y_N^*)$. Já quando há sobreposição dos blocos a reamostragem é realizada com base nos blocos $(Y_1^*, \ldots, Y_l^*), (Y_2^*, \ldots, Y_{l+1}^*), \ldots, (Y_{N-l+1}^*, \ldots, Y_N^*)$, os quais são definidos pelo autor como sendo $i.i.d$ e possuindo distribuição $(N - l - 1)^{-1} \sum_{t=0}^{N-l} \delta_{Y_{t+1}, \ldots, Y_{t+l}} = \rho_{Y,N}^l$ na qual δ_y é um centro de massa de $y \in \Re^m$.

Com isso é possível obter a estimativa de estatísticas de interesse, bem como a média, intervalos de confiança e resultados para o desempenho do Bootstrap.

3.4.3 Bootstrap Circular

A técnica proposta por Politis & Romano (1992) é uma variação assintoticamente equivalente do Bootstrap em blocos de Künsch (1989). Esse método é definido inicialmente para o cálculo da média e tem a vantagem de ser sempre centrado em torno da mesma.

Como o nome sugere, a relocação das unidades experimentais nos blocos acontece agora de forma circular. No Bootstrap circular poderão haver N blocos a serem reamostrados, independente do tamanho dos mesmos.

Isso ocorre devido ao método considerar as observações como formando de fato um círculo, assim tomando-as como não tendo um ponto inicial ou um final.

Para esse método tem-se a existência de um método de seleção automática que otimiza a escolha do número de blocos a serem utilizados (Politis & White, 2004; Patton et al., 2009).

3.4.4 Bootstrap Estacionário

Outra técnica bastante interessante para o uso do Bootstrap em casos de dados que não sejam $i.i.d.$ é o Bootstrap Estacionário, proposto por Politis &

Romano (1994).

Para a realização do Bootstrap estacionário, bem como dos outros métodos de Bootstrap em blocos, é necessário a definição do tamanho dos blocos. O tamanho dos blocos do Bootstrap Estacionário não é fixo. Para sua utilização considera-se variáveis aleatórias L_i, $i = 1, 2, \ldots$ como sendo o tamanho de cada bloco.

A definição de tais variáveis ocorre de maneira pela qual elas sejam i.i.d. e possuam uma distribuição geométrica de probabilidade. Assim, a probabilidade do evento $\{L_i = m\}$ é $(1-p)^{m-1}p$, para $m = 1, 2, \ldots$.

A probabilidade p é obtida a partir de um tamanho esperado pré-estipulado para os blocos. Sabendo que na distribuição geométrica como definida no parágrafo anterior temos que $E(L) = 1/p$ isso implica que uma vez estipulado o valor esperado para o tamanho do bloco, L, basta tomar $p = 1/E(L)$.

O restante do algoritmo segue a mesma lógica dos demais métodos de Bootstrap em blocos e pode ser visto em detalhes em Politis & Romano (1994). Um método de seleção automática para o tamanho esperado dos blocos no Bootstrap Estacionário pode ser visto em Politis & White (2004) e Patton et al. (2009).

Uma vez definido o valor esperado para o tamanho dos blocos, para esse método também é possível o cálculo da média de maneira simples, mas há especificações para a variância. A variância de $\overline{X_N}/\sqrt{N}$ para casos de dependência é definida por:

$$\sigma_N^2 = var(X_1) + 2\sum_{i=1}^{N}\left(1 - \frac{i}{N}\right) cov(X_1, X_{1+i}). \tag{29}$$

3.4.5 Bootstrap no R

Para a implementação do Bootstrap é necessário o uso de softwares que permitam-na. No software estatístico R (R Core Team, 2015) existem algumas funções que trazem a implementação do bootstrap.

Para casos de Bootstrap simples existem os pacotes “boot” (Canty &

Ripley, 2015) e "bootstrap" (Leisch & Tibshirani, 2015) os quais trazem metodologias usuais de Bootstrap. Nesses pacotes existem até mesmo alguns métodos de Bootstrap em blocos com tamanho fixo ou baseados em ditribuições de probabilidade para o tamanho dos blocos através da função "tsboot".

Mais especificamente, também há no R a possibilidade de implementação do Bootstrap através de um pacote criado para a análise de séries temporais chamado "tseries" (Trapletti & Hornki, 2015). Nesse pacote existe uma função chamada "tsbootstrap", a qual permite a relização do Bootstrap para dados estacionários. Nesse pacote é possível realizar o Bootstrap em blocos, de acordo com Künsch (1989) ou o Bootstrap estacionário de acordo com Politis & Romano (1994).

Ainda há a possibilidade da utilização do Bootstrap em blocos usando blocos circulares para dados espaciais no pacote "sperrorest" (Brenning, 2012), através da função "represampling.disc.bootstrap". Esse pacote traz ainda a possibilidade do uso de blocos retangulares, sem sobreposição, pela função "represampling.tile.bootstrap".

Para a determinação ótima do número de blocos segundo o método desenvolvido em Politis & White (2004) e corrigido em Patton et al. (2009) é possível ainda utilizar o pacote "np" (Hayfield & Racine, 2008) através da função "b.star". Essa função pode ser aplicada para dados contínuos e em matrizes de dados de séries temporais.

3.5 Conceitos de Séries Temporais

Como para a aplicação do Bootstrap em blocos é importante o conceito de ausência de sazonalidade e existência de t-estacionariedade (*t-stationarity*), é importante definir tais conceitos.

3.5.1 Estacionariedade

O conceito de estacionariedade baseia-se no desenvolvimento de uma série temporal de maneira aleatória ao redor de uma média constante. Para o caso

do Bootstrap em blocos, basta que haja t-estacionariedade, ou seja, que a série seja *trend stationary*, o que quer dizer que, mesmo que haja não-estacionariedade, ela possa ser removida através da remoção da tendência. Com isso, é necessário que a série não possua sazonalidade.

A rigor, o conceito de estacionariedade estrita é construído da seguinte maneira:

Seja $Z = \{Z_t, t \in T\}$ um processo estocástico. Considerando $F_Z(z_{t_{1+\tau}}, \ldots, z_{t_{k+\tau}})$, como a função de distribuição acumulada conjunta de Z_t para os tempos $t_{1+\tau}, \ldots, t_{k+\tau}$. Então Z_t é dita uma série estritamente (ou fortemente) estacionária, se, para todo k, para todo τ e para todo $t_1, \ldots, t_k$, $F(z_1, \ldots, z_n; t_{1+\tau}, \ldots, t_{n+\tau}) = F(z_1, \ldots, z_n; t_1, \ldots, t_n)$. Ou seja, τ não afeta $F_Z(.)$ e F_Z não é uma função do tempo.

Para que seja considerada a existência de estacionariedade fraca, basta que $\mu(t) = \mu$ e $V(t) = \sigma^2$ para todo $t \in T$, sendo que μ é a média da série e $V(t)$ a variância.

Já o conceito de t-estacionariedade pode ser definido formalmente como (Nelson & Plosser, 1982):

Um processo Z é dito t-estacionário se $Z_t = f(t) + e_t$, sendo t o tempo, f qualquer função que leve dos reais aos reais e e um processo estacionário.

Existem testes de estacionariedade, a maior parte deles baseados no conceito de raíz unitária. Caso não exista raíz unitária a série é considerada estacionária. O conceito de raíz unitária é construído com base na equação característica do processo estocástico. Caso um processo estocástico linear possua o valor 1 como raíz da equação característica do processo ele possuirá raíz unitária (Bhargava, 1986). Ou seja:

Para o processo estocástico Z, supondo que ele possa ser escrito através de um processo autoregressivo de ordem p, temos:

$$z_t = a_1 z_{t-1} + a_2 z_{t-2} + \ldots + a_p z_{t-p} + \epsilon_t \tag{30}$$

com ϵ_t possuindo média 0, variância constante σ^2 e sendo não correlacionado ao longo da série.

Existirá raíz unitária caso $m = 1$ for raíz da equação característica:

$$m^p - m^{p-1}a_1 - m^{p-2}a_2 - \ldots - a_p = 0 \tag{31}$$

Existem alguns testes formais para verificar se uma série possui raíz unitária, bem como o teste de Phillips-Perron (Phillips & Perron, 1988), que tem por hipótese nula a existência de raíz unitária, com a rejeição dessa hipótese evidenciando a existência de estacionariedade.

3.5.2 Sazonalidade

O conceito de sazonalidade é baseado na repetição de padrões em períodos fixados. Assim, por exemplo, caso o período chuvoso apresente influência significativa no padrão existente em determinada variável utilizada para o ajuste de uma série temporal, em detrimento do período seco, existirá sazonalidade. A sazonalidade é comum em dados de precipitação, por exemplo, uma vez que existe um padrão para a quantidade de chuva em cada estação do ano, repetindo-se constantemente esse padrão.

Para verificar a possível existência de sazonalidade, pode-se analisar graficamente a série temporal, bem como verificar os valores obtidos para cada mês ao longo do período de anos estudado, por exemplo. É possível ainda testar formalmente a existência de parâmetros de sazonalidade em modelos desenvolvidos em De Livera et al. (2011), implementados no software R através do pacote "forecast" (Hyndman, 2015).

4 MATERIAL E MÉTODOS

4.1 Localização da bacia hidrográfica

Na constituição do Estado de São Paulo, promulgada em 1989, foram previlegiados os recursos hídricos. O artigo 205 determinou que fosse criado o Sistema Integrado de Gerenciamento dos Recursos Hídricos (Estado de São Paulo, 2014).

A Lei Estadual 7663 de 1991 (Estado de São Paulo, 1991) reconhece a água como sendo um bem público e de valor econômico, com os comitês de Bacia sendo órgãos de adequação de seu uso, levando em conta as características de cada região.

O primeiro Comitê das Bacias Hidrográficas (CBH) instalado foi o CBH-Piracicaba, Capivari e Jundiaí, sendo efetivamente instalado em 18/11/1993, englobando estes três principais rios, os quais envolvem 57 municípios do Estado de São Paulo, cobrindo uma área de $14177,77\ Km^2$ (CBH-PCJ & COBRAPE, 2008). Integrando o CBH-PCJ estão os rios Jaguarí e Atibaia que, juntamente com outros rios afluentes, formam o Rio Piracicaba. O Rio Atibaia sendo um dos formadores do Rio Piracicaba é também um dos afluentes do Tietê, o qual possui $182Km$ de extensão e sofre de um desnível altimétrico de cerca de $2238m$.

O Rio Atibaia percorre 16 municípios do estado de São Paulo, dentre os quais estão: Americana, Campinas, Jaguariúna, Nova Odessa, Paulínia, Valinhos, Vinhedo, Itatiba, Jarinu, Morungaba, Nazará Paulista e Piracicaba. Além disso ele também passa pelo município mineiro de Camanducaia.

O percurso do Rio Atibaia está em uma área de intenso desenvolvimento populacional e também industrial, com isso há intenso uso de suas águas em

tais atividades. Por outro lado, nem todos os municípios têm seus esgotos tratados, o que contribui para o aumento de resíduos que atingem este corpo hídrico.

Na região de Paulínia além do despejo residual de esgotos pela utilização do município de Sumaré e de Paulínia, o Rio Atibaia tem suas águas utilizadas por empresas. Nessa área estão localizadas importantes indústrias químicas e de papel. Neste trecho desagua o Ribeirão Anhumas que há tempos atrás recebia cerca de 40% do esgoto urbano da cidade de Campinas sem qualquer tratamento, esgoto tal que atualmente é tratado e tem sido alvo de um processo de estudos para recuperação ambiental, o que foi o caso, por exemplo, do Projeto Anhumas (Projeto Anhumas, 2006).

Com todos esses contaminantes e poluentes, o Rio Atibaia em seu percurso final torna-se bastante limitado quanto à sua biota macroscópica e também microscópica.

4.1.1 Pontos de Amostragem

As águas dos rios Jaguarí, Atibaia e do efluente industrial foram coletadas em alguns pontos previamente selecionados como podemos ver na Figura 5.

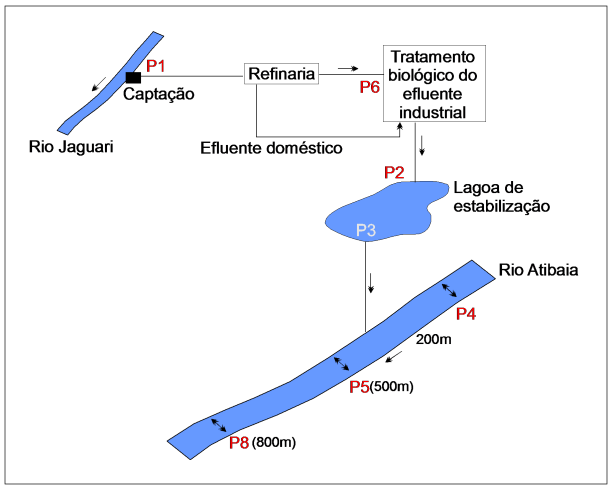


Figura 5: Esquema ilustrativo dos diversos pontos de amostragem.

As coletas foram realizadas durante o período de outubro de 1999 a janeiro de 2015, totalizando 242 amostragens. As localizações em coordenadas são apresentadas na Tabela 5.

Tabela 5: Localização geográfica dos pontos de coleta

Ponto	Local	Longitude (WO) (grau min. seg.)	Latitude (S) (grau min. seg.)
Ponto 1	Rio Jaguarí	22 42 00,0	47 08 06,6
Ponto 2	Entrada da Lagoa de Estabilização	22 44 26,6	47 07 17,5
Ponto 3	Saída da Lagoa de Estabilização	22 44 22,0	47 07 43,0
Ponto 4	Montante da descarga	22 44 43,0	47 07 25,4
Ponto 5	Jusante da descarga do efluente da unidade industrial - 200m	22 44 25,3	47 07 35,2
Ponto 6	Entrada do efluente na estação de tratamento biológico	22 44 18,8	47 07 19,3
Ponto 8	Jusante da descarga do efluente da unidade industrial - 800m	22 44 22,3	47 07 40,8

A localização dos pontos foi obtida através do equipamento *GPS - Garmin eTrex Summit*. O ponto 7 que não encontra-se na tabela foi temporário e não apresentava interesse para essa pesquisa. Ele era referente ao efluente doméstico e o mesmo passou a ser incorporado ao tratamento do efluente da refinaria.

Para visualização dos pontos de coleta serão apresentadas algumas figuras com fotos de cada um dos pontos onde a água foi coletada.



(a) Próximo ao ponto de captação do Rio Jaguarí.



(b) Entrada do efluente tratado na lagoa de estabilização.



(c) Saída do efluente na lagoa de estabilização.



(d) Rio Atibaia, montante da entrada do efluente industrial tratado.



(e) Jusante da entrada do efluente tratado no Rio Atibaia (200m).



(f) Jusante da entrada dos efluentes tratados (800m).

Figura 6: Pontos de coleta da água.

Para a análise de dados serão considerados os seguintes pontos:

- Ponto 1 - Rio Jaguarí : Ponto utilizado a fim de servir como comparação para a qualidade da água no Rio Atibaia, uma vez que o Rio Jaguarí é considerado como um rio de classe 1 (Estado de São Paulo, 1977).

- Ponto 3 - Saída do efluente da lagoa de estabilização : Ponto utilizado a fim de verificar a qualidade da água antes de ser despejada no Rio Atibaia. Na entrada da lagoa de estabilização a água deve estar de fato poluída, sendo tratada para diminuir seu grau de poluição à saída da mesma.

- Ponto 5 - Jusante da descarga do efluente da unidade industrial (200m) : Ponto utilizado a fim de verificar a qualidade das águas do Rio Atibaia logo após o despejo do efluente industrial. Para que o efluente esteja completamente disperso pelo rio o ideal é utilizar o ponto a 800 metros, porém isso não é

possível para todo o período, uma vez que somente mais recentemente ele tem sido coletado. Com isso o ponto 5 torna-se o mais adequado para ser utilizado na análise proposta.

4.2 Metodologia da Coleta

4.2.1 Períodos de Coleta

Durante a execução deste estudo, realizado de 2000 a 2014, foram efetuadas diversas amostragens, atendendo a compromissos firmados por meio de contratos e convênios entre a unidade industrial e a Fundunesp. Há alguns dados para o começo do ano de 2015 e para o final de 1999 mas para a imputação foram utilizados apenas os dados relativos às coletas entre 2000 e 2014.

As coletas foram realizadas preferencialmente às Quartas-Feiras, levando em conta que nos finais de semana o ritmo de operações de algumas indústrias da região é interrompido ou minimizado e isso poderia afetar a análise.

Os períodos a seguir referem-se aos compromissos assumidos e os números de coletas durante cada um deles podendo ser vistos na Tabela 6.

Tabela 6: Períodos e número de coletas durante a pesquisa

Out/1999 a set/2000	23
Jul/2001 a dez/2001	12
Jun/2002 a mar/2003	15
Out/2003 a set/2004	16
Mai/2005 a mar/2006	22
Mai/2006 a mar/2007	22
Ago/2007 a nov/2007	8
Mai/2008 a mar/2009	18
Out/2009 a ago/2010	18
Ago/2010 a jul/2011	18
Out/2011 a ago/2012	18
Ago/2012 a jun/2013	18
Jul/2013 a abr/2014	18
Mai/2014 a jan/2015	16
Total	242

4.2.2 Análises Físicas e Químicas

As amostras foram analisadas quanto aos parâmetros químicos, físicos e biológicos segundo Standard Methods (APHA, 2005; Clesceri et al., 1998) no laboratório de Ensino e Pesquisa de Toxicidade de água do Departamento de Bioquímica e Microbiologia, Instituto de Biociências, Campus de Rio Claro/SP/BR, sendo selecionadas as seguintes variáveis, com seus respectivos métodos de medição e unidade de medida:

- pH: análise com potenciômetro;
- Turbidez: com turbidímetro, em NTU;

- Temperatura: com o termômetro, em graus Celsius;
- Coliformes fecais (E. coli): Analisados pela técnica Coliert, em NMP/100mL;
- Sólidos solúveis e em suspensão (SSS): gravimétrica, em g/L;
- Oxigênio Dissolvido (OD): eletrométrica, em mg O2/L;
- A demanda bioquímica de oxigênio (DBO): eletrométrica, em mg O2/L a 20^{o}C;
- Amônia: eletrométrica, em mg NH3/L;
- Nitrito: colorimétrica, em mg NO2/L;
- Nitrato: colorimétrica, em mg NO3-/L;
- Fósforo: colorimétrica, em mg P/L.

4.2.3 Os Dados do Estudo

Antes de prosseguir, é importante apresentar melhor os dados do estudo e algumas informações úteis para a análise, principalmente no que tange à imputação de dados.

Os dados disponíveis são dados mensais de Janeiro de 2000 até Dezembro de 2014, totalizando assim 15 anos. Há ausência de todas as variáveis nos períodos para os quais não houve coleta da água, mas também é notável a existência de não observação de alguma variável específica mesmo quando há coleta. Principalmente no tocanto à Turbidez, em várias coletas, nas quais outras variáveis foram analisadas, não há análise para a Turbidez.

As análises na água são realizadas e o resultado são os dados nas escalas de medida explicitadas anteriormente. Para a realização da imputação de dados é interessante utilizar os valores dos subíndices q_i que compõem o IQA, uma vez que utilizando tais subíndices há a padronização da escala, que para cada q_i varia agora de 0 a 100 de maneira linear, com a mesma interpretação.

Tal padronização torna mais plausível a utilização de métodos como a regressão linear múltipla para casos nos quais alguma variável que compõe o IQA não seja observada porém outras o sejam. Se alguém desejar saber qual o valor na escala original do dado imputado, basta verificar o valor equivalente nas curvas do IQA, que constam na Figura 3. Uma vez que os valores de q_i sejam imputados, para o cálculo do IQA basta aplicar a Equação (22).

É interessante verificar ainda a proporção de ausência de cada variável em meses nos quais as outras foram observadas e também a proporção de ausências multivariadas, a qual ocorre nos meses sem coleta da água. Para as ausências multivariadas, todas as variáveis possuem o mesmo número de não observações, sendo que dos 180 meses do período, há não coleta em 46 meses, resultando assim em uma proporção de cerca de 25.55% de ausência para esse padrão. Essa situação vale para todos os pontos do estudo.

Quanto a ausências por não haver análise de determinada variável, ocorridas nos demais 134 meses temos situações diferentes nos pontos selecionados para o estudo, temos assim:

Tabela 7: Proporção de Ausências para os Pontos

Variável	P1	P3	P5
E. coli	0%	0%	1.49%
pH	0%	0%	1.49%
Nt	2.23%	2.23%	3.73%
Pt	13.43%	11.19%	11.19%
Turbidez	15.67%	15.67%	17.16%
O.D.	2.23%	2.23%	4.48%
DBO	0%	0%	1.49%
S. Totais	2.23%	0%	3.73%

Podemos notar na tabela 7 que algumas variáveis apresentam uma proporção razoavelmente grande de ausência. Quando falta uma das variáveis que compõem o IQA o índice não pode ser calculado e assim o número de ausências torna-se extremamente grande, sendo muito maior que apenas aquelas causadas pela falta de coleta de dados.

Para o Ponto 1 há apenas 95 meses para os quais há a análise de todas as variáveis que compõem o IQA. Sendo assim, dos 180 meses do período, os quais já estava reduzidos a 134 meses devido à ausência de dados quando não foi realizada a coleta, passa a existir uma proporção ainda maior de ausência na variável final analisada (IQA), contendo quase 50% de ausência de observações, principalmente devido à falta de análise da turbidez ser bastante constante ao longo do estudo. Para o Ponto 3 há 98 observações completas e para o Ponto 5 há 96.

4.3 Imputação de Dados

4.3.1 Padrão e Mecanismo dos Dados Analisados

Para a realização da imputação de dados é importante a definição da maneira pela qual os dados se comportam, ou seja, seu padrão e mecanismo. Principalmente o mecanismo de ausência de dados é essencial, uma vez que a maior parte dos métodos são bem definidos apenas para mecanismos do tipo MAR e MCAR, conforme explicado anteriormente no Capítulo 3.

Levando em conta os dados como sendo observações das variáveis em estudo em três pontos diferentes no mesmo mês, temos que os dados em análise apresentam dois tipos de padrão de dados, ambos com o mesmo mecanismo.

Primeiramente há a perca de informação de maneira aleatória em apenas uma das variáveis para algumas datas de coleta. Isso ocorre devido à impossibilidade de coletar a água em algum dos pontos em determinados meses no dia estipulado para coleta, o que usualmente se dá por fatores fora do controle dos pesquisadores, bem como condições meteorológicas que impedem a entrada no rio para

a coleta da água.

Quando isso ocorre temos a observação da variável em dois dos pontos mas não no terceiro. O padrão de perda de dados é geral (arbitrário) e o mecanismo é MCAR, uma vez que a ausência dos dados é unicamente causada por fatores externos à variável em si.

Existe ainda a perca de informação para todos os pontos simultaneamente em determinados meses. A ocorrência dessa ausência de dados se deve ao fato de em determinados períodos não haver coleta da água nos rios devido ao vencimento dos compromissos assumidos para tal e demora de alguns meses para sua renovação.

Na ocorrência desse tipo de ausência de dados o mecanismo também é MCAR, uma vez que ela se dá unicamente por não haver um contrato vigente que garanta a coleta dos mesmos, sendo completamente independente da variável coletada em si. O padrão de dados agora é o de não resposta multivariada.

Já se considerarmos a entrada de dados como sendo todas as variáveis por ponto, desconsiderando-se os meses para os quais não houve coleta de água, levando em conta apenas aquelas para os quais não há análise de alguma variável específica que compõe o IQA, o padrão de ausência é geral (arbitrário). O mecanismo de ausências vale da mesma maneira que para a entrada de dados por pontos, uma vez que a origem dos dados é a mesma.

4.3.2 Método de Imputação de Dados

Para a análise estatística foram utilizados dois métodos diferentes, a fim de verificar o resultado obtido em cada um deles.

Para uma primeira análise foram utilizados os dados com exclusão dos meses nos quais há ausência de dados. A utilização apenas dos dados completos é o que caracteriza a análise de descarte de casos ou análise de caso completo, através da remoção de dados nos quais uma das observações para determinada variável seja ausente.

Também foram analisados os dados após a realização do método de

imputação múltipla proposto por Bergamo (2007) e Bergamo et al. (2008). A escolha desse método de imputação se dá pela ausência de normalidade dos dados em questão, a qual é pressuposta por diversos métodos de imputação. Além disso também será utilizada a imputação pela média e, para os casos nos quais falte apenas uma das variáveis que compõem o IQA, a imputação do valor do índice sem imputar os subíndices, redefinindo os pesos de maneira a sua soma ser 1 mesmo na ausência de alguma variável.

Para a realização da imputação utilizou-se duas maneiras diferentes de entrada de dados na matriz utilizada. Primeiramente a entrada dos dados foi realizada por variável, em cada ponto. Nessa abordagem a matriz tem o seguinte formato:

$$\mathbf{Y}=\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & y_{n3} \end{bmatrix}$$

Para esse caso, y_{ij} é a variável em questão na data i e em determinado ponto j, sendo $j = 1$ referente ao Ponto 1, $j = 2$ equivalente ao Ponto 3 e $j = 3$ equivalente ao Ponto 5. Ou seja, as colunas contêm observações da mesma variável em diferentes pontos, sendo as linhas os meses consecutivos.

O problema da entrada de dados dessa maneira é que o número de ausências múltiplas em todas as variáveis será ainda maior. Além dos casos para os quais não há coleta de água, existem outras ocasiões nas quais, apesar da coleta ser realizada, uma das variáveis que compõem o IQA não o são. Entrando com os dados por variável em cada ponto há casos para os quais existe ausência nos três pontos para determinada variável, mesmo que haja coleta de água nesse caso.

Outra possível maneira de entrada de dados, que corrige esse problema, mas que é inviável para determinados métodos, é a de utilizar uma matriz com os dados para um mesmo ponto em diferentes variáveis. Assim teremos uma matriz do tipo:

$$\mathbf{Y}=\begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} & y_{16} & y_{17} & y_{18} \\ y_{21} & y_{22} & y_{23} & y_{24} & y_{25} & y_{26} & y_{27} & y_{28} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & y_{n4} & y_{n5} & y_{n6} & y_{n7} & y_{n8} \end{bmatrix}$$

Agora y_{ij} representa a observação da variável j, em um ponto específico, na data i. Ou seja, as colunas possuem as diferentes variáveis que compõem o IQA em um mesmo ponto, sendo as linhas os meses consecutivos.

Com isso é corrigido o problema anteriormente apontado para alguns períodos, pois agora há, nos meses nos quais há coleta da água, casos para os quais uma variável específica não é observada mas outras o são, podendo as mesmas serem utiliadas na previsão da variável ausente naquele período específico. O problema de quando há ausência da coleta da água persiste, havendo diversos períodos para os quais há ausência nas oito variáveis utilizadas no cálculo do IQA.

Como o interesse maior é a análise do IQA como um todo, é interessante também comparar a qualidade de imputação obtida através da estimação dos valores ausentes dos subíndices com os resultados que se encontram ao estimarmos somente o valor do IQA, sem estimar os subíndices faltantes. Isso pode ser feito, no caso para o qual há ausência de apenas uma variável ou até mesmo algumas.

Uma possibilidade para tal é utilizar apenas as variáveis observadas para calcular o IQA, recalculando os pesos de cada um dos subíndices q_i no cálculo final do Índice de Qualidade da Água. Nessa abordagem, caso a variável ausente tenha um valor de peso $w_i = x$, com sua ausência a soma dos demais pesos (S_{-i}) resultaria em $S_{-i} = 1 - x$. Assim, a fim de que por meio de uma padronização essa soma resulte em um e a proporção de pesos de cada subíndice seja mantida, podemos tomar os pesos padronizados ($wp_{i'}$) das demais variáveis i' como sendo da forma $wp_{i'} = w_{i'}/(1 - x)$, sendo $w_{i'}$ o peso original para aquele subíndice. Essa abordagem é uma das que surgem mais naturalmente como possível solução do problema.

Uma vez imputados os dados, foram calculadas algumas estatísticas de interesse para avaliação da qualidade do ajuste. A fim de haver um método

simples para meios de comparação da qualidade das técnicas utilizadas, cada variável imputada foi comparada com os resultados obtidos para tais estatísticas com base em uma imputação pela média da variável em questão.

Para a implementação dos métodos de imputação, tanto o de Bergamo quanto o da média, foram utilizados programas computacionais escritos na linguagem IML do software SAS 9.3 (SAS Institute INC, 2011). Para a verificação das estatísticas de qualidade do ajuste utilizou-se tanto o software SAS como o software R (R Core Team, 2015).

4.4 Análise Estatística

Para a análise estatística foi realizada a comparação entre a qualidade das águas nos rios e na lagoa de estabilização. A variável em estudo é o Índice de Qualidade da Água (IQA) conforme proposto pela CETESB. Uma vez que o banco de dados possui todas as variáveis necessárias para o cálculo do IQA, isso é possível.

Foram considerados análises de diferentes informações. Primeiramente temos a análise dos pontos de coleta selecionados (Ponto 1, Ponto 3 e Ponto 5), conforme justificativa anterior, a fim de verificar a condição da qualidade da água nos diferentes locais e a possível diferença entre eles. Também é interessante analisar se ao longo dos anos há diferença na qualidade das águas. Para tal foram comparados os quinquênios (2000-2004, 2005-2009, 2010-2014).

As análises foram feitas iniciando pela realização das mesmas na matriz de dados originais e levando em conta apenas os dados completos, deixando de lado, além dos meses nos quais não houve coleta da água, aqueles que não possuem a análise de alguma das variáveis. Após a realização dessa análise são apresentadas análises dos dados com os valores imputados, possuindo assim informações sobre mais meses. O total de dados será assim igual ao total de coletas existentes.

Há problemas importantes na utilização de métodos usuais de análise. A abordagem de séries temporais apresenta dificuldades devido à ausência de coletas ao longo de todo o período, sendo que há diversos intervalos de tempo sem

observações de nenhuma variável componente do IQA, o que inviabiliza até mesmo a imputação de dados para tais meses. Mesmo a utilização da ANOVA a fim de comparar as médias, levando em conta todos os meses do estudo, não é adequada, uma vez que, mesmo que a falta de normalidade existente nos subíndices q_i possa não existir no IQA, há o problema da existência de dependência entre as observações, o que inviabiliza o uso desse método de análise.

4.4.1 Bootstrap

A técnica do Bootstrap em blocos foi considerada para a obtenção de estimativas das estatísticas para cada ponto e período e para a comparação entre os mesmos. Foram realizadas análises para os dados com análise de casos completos originais e após isso para os dados imputados.

Primeiramente foram calculadas a média, a variância Bootstrap e obtidos intervalos de confiança a nível de significância de 5% para cada um dos pontos e também para cada um dos períodos, separadamente. Feito isso, a fim de obter comparações entre os pontos e posteriormente os períodos foram criados vetores de contraste.

Nesses contrastes, o primeiro vetor é definido como a subtração do IQA no Ponto 3 com relação ao IQA no Ponto 1 (IQAP1-IQAP3), o segundo como a subtração do IQA no Ponto 5 com relação ao no Ponto 1 (IQAP1-IQAP5) e por fim temos o vetor da subtração do IQA no Ponto 5 com relação ao Ponto 3 (IQAP3-IQAP5). Isso pode ser feito para os meses nos quais há possibilidade do cálculo do IQA em todos os pontos, para o caso da matriz original e para todos os meses do estudo após a realização da imputação.

Com isso torna-se possível a obtenção das médias, variâncias e intervalos de confiança para comparação entre os pontos. Caso o valor zero esteja contido nos intervalos de confiança a conclusão é de que não há diferença estatística significativa entre os pontos comparados, caso o zero não esteja contido há diferença significativa. Para os pontos, a fim de obter intervalos de confiança para o IQA de

cada ponto individualmente, foram utilizados os quantis 2.5% e 97.5%.

A fim de realizar intervalos referentes a cada contraste e manter o nível de significância para os três pontos os quantis considerados foram os equivalentes ao grau de significância 1.8% para a comparação entre pontos ($0.982^3 = 0.947$), para que o grau de significância geral aproxime-se de 5%. Esse método de correção do grau de confiança é conhecido como método de Sidak (Sidak, 1967).

Como os dados apresentam auto correlação interna, as técnicas de Bootstrap em blocos serão utilizadas. Os resultados serão obtidos pelo Bootstrap em blocos proposto por Künsch (1989) e no Bootstrap estacionário proposto por Politis & Romano (1994).

A análise foi realizada através do software R através do pacote "tseries". Para tal será usada a função "tsbootstrap" para obtenção das estimativas, considerando o tamanho dos blocos de acordo com a função "b.star", com o tamanho do bloco para o Bootstrap circular sendo considerado para o Bootstrap em blocos, uma vez que são bastante similares, e o tamanho esperado dos blocos para o Bootstrap estacionário sendo tomado do cálculo do mesmo nessa função. Os comandos utilizados estão disponíveis nos Apêndices.

4.4.2 ANOVA

Para a comparação de mudanças ao longo do tempo, considerando-se como unidade de comparação os valores de IQA nos quinquênios do estudo, foi utilizada a técnica da ANOVA. Para que as suposições sejam verificadas é necessário uma mudança na variável resposta, uma vez que, utilizando-se os dados mês a mês, existe dependência entre os mesmos, fazendo a suposição de normalidade dos resíduos falhar.

Utilizando-se a média de alguns meses consecutivos como variável resposta esse problema pode ser corrigido, devendo haver análise para cada um dos pontos sobre quantos meses devem ser considerados para tal. Maiores informações sobre como isso foi realizado encontram-se no capítulo de Resultados.

5 RESULTADOS

5.1 Imputação de Dados

Nessa seção serão apresentados os resultados referentes às estatísticas do ajuste de imputação de dados para o método de imputação múltipla proposto por Bérgamo e para o método de imputação pela média. A parte de análise estatística tanto com os dados completos já existentes quanto para os dados imputados será apresentada em 5.2.

Tais estatísticas foram calculadas com base em simulações para os dados completos existentes levando em conta o mesmo padrão de ausência de dados da matriz original quanto à porcentagem de valores ausentes e seu padrão de ausência.

As estatísticas avaliadas são a NMRSE, o PRESS e o coeficiente de correlação de Spearman, além da segunda parte do T_{acc} que pode ser adaptado para casos de imputação simples. Através dessas estatísticas podemos observar se a imputação através do método proposto apresenta resultados melhores que a imputação simples pela média ou não.

É interessante salientar, antes de prosseguir com tais análises, que a variabilidade interna do método de Bergamo foi praticamente inexistente, utilizando-se a entrada de dados para uma variável nos diferentes pontos. Como há ausências em todos os pontos por diversas vezes, todas as imputações resultam no mesmo valor quando isso ocorre. A variabilidade somente existe para casos nos quais há a observação em dois pontos e a ausência em um, o que ocorre muito raramente.

Ainda é notável que caso a análise seja feita com a entrada de todas as variáveis por pontos, caso sejam utilizados os valores absolutos das mesmas, o

que causa grande diferença de escala entre as variáveis e valores extremos bastante discrepantes em algumas delas, o método de Bergamo não funciona bem. Nesse caso sua implementação resulta em valores negativos para algumas imputações, o que não é coerente para os dados. Ao utilizar os valores, conforme proposto, já de acordo com os subíndices de qualidade q_i que compõem o IQA, esse problema é corrigido.

Nas Tabelas $8 - 15$ são apresentadas as estatísticas para a média e o método de Bergamo, no que se refere à análise variável por variável, além de uma coluna para a diferença entre ambos.

Tabela 8: Desempenho para os diferentes métodos para a variável Coliformes

Estatística	Média	Bergamo	Diferença
NRMSE	0.938	0.981	-0.043
SPEARMAN	0.527	0.372	155
PRESS	15592.752	37548.371	-21955.619
Tacc 2	187.413	451.302	-263.889

Tabela 9: Desempenho para os diferentes métodos para a variável DBO

Estatística	Média	Bergamo	Diferença
NRMSE	0.889	0.891	0.002
SPEARMAN	0.464	0.452	0.012
PRESS	56425.273	56696.908	-271.635
Tacc 2	678.188	681.453	-3.265

Tabela 10: Desempenho para os diferentes métodos para a variável Fósforo

Estatística	Média	Bergamo	Diferença
NRMSE	0.899	0.916	-0.017
SPEARMAN	0.362	0.338	0.024
PRESS	36444.144	37883.096	-1438.952
Tacc 2	382.817	397.932	-15.115

Tabela 11: Desempenho para os diferentes métodos para a variável Nitrogênio

Estatística	Média	Bergamo	Diferença
NRMSE	0.863	0.862	0.001
SPEARMAN	0.552	0.547	0.005
PRESS	44337.249	44214.465	122.784
Tacc 2	532.900	531.424	1.476

Tabela 12: Desempenho para os diferentes métodos para a variável OD

Estatística	Média	Bergamo	Diferença
NRMSE	0.991	0.990	0.001
SPEARMAN	0.155	0.159	-0.004
PRESS	58077.972	57987.781	90.191
Tacc 2	678.481	677.427	1.054

Tabela 13: Desempenho para os diferentes métodos para a variável pH

Estatística	Média	Bergamo	Diferença
NRMSE	0.661	0.659	0.002
SPEARMAN	0.805	0.804	0.010
PRESS	4738.711	4715.318	23.393
Tacc 2	56.956	56.674	0.282

Tabela 14: Desempenho para os diferentes métodos para a variável Resíduos

Estatística	Média	Bergamo	Diferença
NRMSE	0.711	0.761	-0.050
SPEARMAN	0.713	0.727	-0.014
PRESS	21634.239	24775.563	-3141.324
Tacc 2	260.027	297.783	-37.756

Tabela 15: Desempenho para os diferentes métodos para a variável Turbidez

Estatística	Média	Bergamo	Diferença
NRMSE	0.964	0.957	0.007
SPEARMAN	0.257	0.289	-0.032
PRESS	43342.159	42734.553	607.606
Tacc 2	433.422	427.345	6.077

Para quase todas as variáveis no estudo nota-se que não são obtidos bons resultados para a imputação pelo método de Bergamo. Em alguns casos ela é até mesmo inferior à imputação pela média, possuindo menor coeficiente de correlação de Spearman, maior NRMSE, PRESS e T_{acc}. Os subíndices dos resíduos e do pH são os que apresentaram melhor resultado quanto às estatísticas avaliadas. Um dos motivos para que isso ocorra é que dentro dessas variáveis há menor variabilidade.

Quanto à variabilidade entre imputações, ela é bastante pequena. A parte referente à variabilidade entre imputações do T_{acc} é sempre próxima a zero. Isso ocorre porque, devido ao padrão de ausência, quando a variável não é observada nos três pontos, ela sempre apresentará o mesmo valor em cada uma das imputações. A exceção a isso são os raros casos nos quais há ausência em apenas um ponto, mas mesmo para eles há variabilidade bastante baixa entre as imputações. Isso dá evidências de que a imputação múltipla não traz muitas informações que melhorem a análise em relação à simples.

Além disso, como os valores se repetem para os três pontos, nas linhas em que há ausência simultânea para todos os pontos, isso pode trazer certo viés ao serem realizadas determinadas comparações posteriores utilizando esses dados imputados. Principalmente quanto à comparação entre períodos, o fato de, independetemente do mês, haverem valores imputados iguais faz com que essa análise seja prejudicada. É notável ainda que isso ocorre em diversas datas, mesmo quando há coleta de água mas não há a análise de uma determinada variável, uma vez que, quando isso ocorre, não há tal análise para nenhum dos três pontos. Com isso há ainda mais casos de ausência para os três pontos para determinada variável, não se limitando essa ocorrência para quando não há coleta de dados.

Com isso, realizar a imputação dessa maneira não traz bons resultados, mesmo utilizando um método que não suponha normalidade dos dados.

Quanto à imputação realizada por pontos ao invés de por variável, considerando-se agora, primeiramente, o caso no qual há ausência de alguma variável porém outras são observadas, ou seja, casos nos quais há coleta de água mas não há análise de determinado parâmetro, obtêm-se os resultados apresentados nas Tabelas $16 - 18$, levando em conta a imputação de todos os subíndices ausentes.

Tabela 16: Desempenho para os diferentes métodos utilizando os Q_i para o Ponto 1

Estatística	Média	Bergamo	Diferença
NRMSE	0.986	1.150	-0.164
SPEARMAN	0.371	0.332	0.039
PRESS	12355.708	16818.365	-4462.657
Tacc 1	-	1.271	-
Tacc 2	482.645	618.322	-135.677

Tabela 17: Desempenho para os diferentes métodos utilizando os Q_i para o Ponto 3

Estatística	Média	Bergamo	Diferença
NRMSE	0.662	0.798	-0.136
SPEARMAN	0.739	0.599	0.140
PRESS	6342.973	9236.935	-2893.962
Tacc 1	-	62.776	-
Tacc 2	247.772	339.593	-91.821

Tabela 18: Desempenho para os diferentes métodos utilizando os Q_i para o Ponto 5

Estatística	Média	Bergamo	Diferença
NRMSE	0.709	0.620	0.089
SPEARMAN	0.732	0.827	-0.095
PRESS	13064.346	9997.302	3067.044
Tacc 1	-	0.831	-
Tacc 2	510.326	367.548	142.778

Nota-se que para o Ponto 1 a imputação não apresentou resultados tão bons, principalmente no que diz respeito ao método de Bergamo, porém, para os demais pontos, há uma grande melhoria utilizando os subíndices q_i de cada ponto como entrada para a realização da imputação de dados. Com excessão do Ponto 3, a variabilidade entre as imputações, expressa nas tabelas através do Tacc 1, não é muito grande. O Ponto 5 teve resultados bastante melhores para o método de Bergamo do que os obtidos através da imputação pela média.

Para essa maneira de entrada de dados, considerando-se apenas as datas nas quais há ausência de alguma variável específica componente do IQA, não há o problema de repetição de valores, uma vez que não há ausência múltipla em todas as variáveis para determinado ponto. A ausência múltipla ocorre nos casos para os quais não há coleta de dados, persistindo nesses casos a dificuldade de imputação.

Vale ainda ressaltar que métodos como a imputação pela regressão não apresentam bons resultados nem mesmo para essa maneira de entrada de dados. Além dos subíndices q_i não possuírem normalidade, eles foram selecionados de maneira que fossem complementares na formulação do IQA, o que faz com que haja baixa correlação entre os mesmos, tornando a previsão de um subíndice por meio dos outros bastante inviável.

Como o intuito final é a comparação do IQA, vale verificar os resultados obtidos através do cálculo do IQA considerando somente as variáveis presentes no estudo, com recálculo dos pesos, comparando-os com os resultados do índice ob-

tidos após a realização da imputação para os subíndices que o compõem. Caso, em média, a imputação pelo método do Bergamo resulte em melhores resultados para o IQA, é bastante interessante utilizar o método de imputação múltipla para a análise posterior. Já se o método de Bergamo apresentar valores finais de IQA menos verossímeis, por mais que haja a presença da variabilidade ao utilizar tal método, há evidências de que ele não seja mais adequado para essa análise específica se comparado à imputação simples resultante do recálculo dos pesos e obtenção do índice final a partir das variáveis existentes no banco de dados.

Calculando as estatísticas de análise considerando-se agora o valor do IQA, calculado após a imputação de seus subíndices, no caso da média e do método de Bergamo, e considerando-se recálculo dos pesos no caso do método dos pesos, tem-se os resultados apresentados nas Tabelas $19 - 21$, para cada ponto.

Tabela 19: Desempenho para os diferentes métodos para o IQA no Ponto 1

Estatística	Média	Bergamo	Pesos
NRMSE	0.167	0.190	0.192
SPEARMAN	0.984	0.983	0.977
PRESS	100.676	129.661	131.761
Tacc 2	3.932	5.066	5.147

Tabela 20: Desempenho para os diferentes métodos para o IQA no Ponto 3

Estatística	Média	Bergamo	Pesos
NRMSE	0.108	0.122	0.102
SPEARMAN	0.993	0.994	0.995
PRESS	42.465	54.419	129.997
Tacc 2	1.712	2.061	5.242

Tabela 21: Desempenho para os diferentes métodos para o IQA no Ponto 5

Estatística	Média	Bergamo	Pesos
NRMSE	0.466	0.323	0.440
SPEARMAN	0.810	0.936	0.905
PRESS	252.021	121.329	1146.699
Tacc 2	9.000	4.333	40.953

Nota-se que o método do recálculo dos pesos foi o menos efetivo em todos os pontos. No Ponto 1 e no Ponto 3 há bastante semelhança entre o método de Bergamo e o da imputação pela média nas estatísticas de qualidade de ajuste. Porém, é importante notar que o método de Bergamo resulta, nesse caso, em valores diferentes para os subíndices q_i em cada ocasião, fazendo com que a contribuição que esse valor tem no cálculo final do IQA não seja sempre igual, como ocorre no caso da média. Já no Ponto 5 a imputação pelo método de Bergamo foi bastante superior aos outros dois métodos.

Com isso, é interessante utilizar a imputação múltipla pelo método de Bergamo, realizada com relação aos subíndices que compõem o IQA, a fim de imputar valores necessários ao cálculo do índice final, cuja análise é o intuito do estudo, principalmente por ele permitir também permitir avaliar a variabilidade nas imputações realizadas.

A aplicação desse método é efetiva apenas no caso para o qual não há observação de alguma variável que compõe o IQA mas as outras são observadas. No caso de ausência completa, de todas as variáveis, os resultados serão sempre os mesmos, como o foram na entrada de dados por variável. Isso torna inviável a imputação de dados para os meses nos quais não houve coleta da água. Para a imputação de dados com um padrão de ausência múltipla dessa maneira ainda há certa lacuna a ser preenchida nos métodos de imputação. Mesmo aqueles que utilizam a metodologia de *splines*, por exemplo, não são tão efetivas devido a ausência de dados ser longa em alguns dos períodos nos quais não houve convênio, o qual

possibilita a coleta da água.

Com isso, as análises estatísticas serão realizadas, primeiramente, para a matriz original de dados e posteriormente para a imputação nos casos em que há coleta de água mas não há informações sobre alguma das variáveis componentes do IQA. Nos meses nos quais não houve coleta de água, pela inviabilidade dos métodos, a imputação não será realizada.

Caso houvesse o interesse em estudar cada variável separadamente e não o índice como um todo, também seria interessante o desenvolvimento de novos métodos de imputação que permitissem uma melhor previsão dos mesmos.

5.2 Análise Estatística

5.2.1 Matriz Original

Utilizando a matriz de dados original apenas com os meses que possuem todas as variáveis necessárias ao cálculo do IQA temos uma grande redução na quantia de dados.

Para meses nos quais não houve coleta das águas há ausência de todas as variáveis necessárias ao cálculo do IQA, assim impossibilitando a utilização de tais períodos nessa análise. Também para alguns dos meses nos quais houve coleta das águas há ausência de alguma variável específica necessária ao cálculo do IQA (principalmente turbidez e fósforo), assim tornando necessário que os períodos no qual isso ocorre também sejam descartados quando realizando-se a análise com os dados completos sem imputação.

Nesse caso os dados são desbalanceados, não só entre os períodos secos ou chuvosos, mas até mesmo entre os pontos. Para determinados pontos não há coleta de dados em algumas ocasiões devido à condições climatológicas contrárias.

Para a análise dos dados o IQA calculado para cada ponto e estação será considerado com duas casas decimais.

Ao realizar a ANOVA com dois fatores (ponto e período), é possível

notar falhas nas suposições do modelo, as quais desviabilizam sua utilização. A normalidade é verificada, porém existe heterogeneidade da variância e também dependência dos resíduos.

Para a verificação das suposições foram utilizados os testes de Shapiro Wilk para a normalidade, de Breusch-Pagan para a homogeneidade e de Durbin Watson para a independência dos resíduos. A Tabela 22 mostra o resultado dos testes formais para cada uma dessas suposições.

Tabela 22: Testes das suposições da ANOVA

Suposição	Estatística de Teste	Valor-p	Resultado
Normalidade	W=0.9953	0.522	Não rejeita normalidade
Homogeneidade	BP=32.183	4.79×10^{-7}	Rejeita homogeneidade
Independência	D-W=0.7304	$< 2 \times 10^{-16}$	Rejeita independência

Com isso a utilização da Análise de Variância não é a mais adequada. Como os dados apresentam dependência, não sendo i.i.d., até mesmo a utilização de métodos não paramétricos não é adequada. Apesar delas não suporem distribuições, a falta de independência diminui o poder do teste e tais técnicas não são as mais indicadas.

Sendo assim, foi utilizado o Bootstrap em blocos a fim de comparar as médias para os pontos e períodos. A metodologia de Bootstrap utilizada será a que leva em conta a dependência dos dados analisados. Através disso podem ser obtidos intervalos de confiança que permitem a comparação estatística das médias em cada caso analisado.

5.2.1.1 Análise dos dados originais

Antes de proceder a análise, para a utilização do Bootstrap, os dados foram balanceados. Aqueles dados para os quais há observações apenas para uma variável foram excluídos da análise, levando-se em conta apenas os meses nos quais

há observações para todos os pontos do estudo. Com isso, os dados disponíveis totalizam 93 coletas.

Primeiramente foram obtidos resultados para cada um dos pontos, procedidos da análise dos contrastes que permitem a comparação da existência ou não de diferenças estatisticamente significantes entre os mesmos quanto ao IQA.

Antes de partir para a análise dos resultados em si, a fim de verificar se a aplicação do Bootstrap em blocos tem sua maior validade, é imporante verificar se a série não possui sazonalidade, sendo ao menos t-estacionária. Há resultados interessantes mesmo para casos nos quais não há estacionariedade, como foi estudado na revisão bibliográfica, mas a propriedade da t-estacionariedade e de não haver sazonalidade garante a melhor eficácia do método.

Como há diversas ausências, para os testes formais, a série será considerada levando em conta apenas os dados existentes, verificando se dentro dos mesmos tais propriedades são verificadas.

Os resultados das observações da matriz original quanto ao IQA, ao longo do tempo, podem ser observados na Figura 7. A mediana e a distribuição das observações divididas por mês podem ser vistas nas Figuras $8 - 10$.

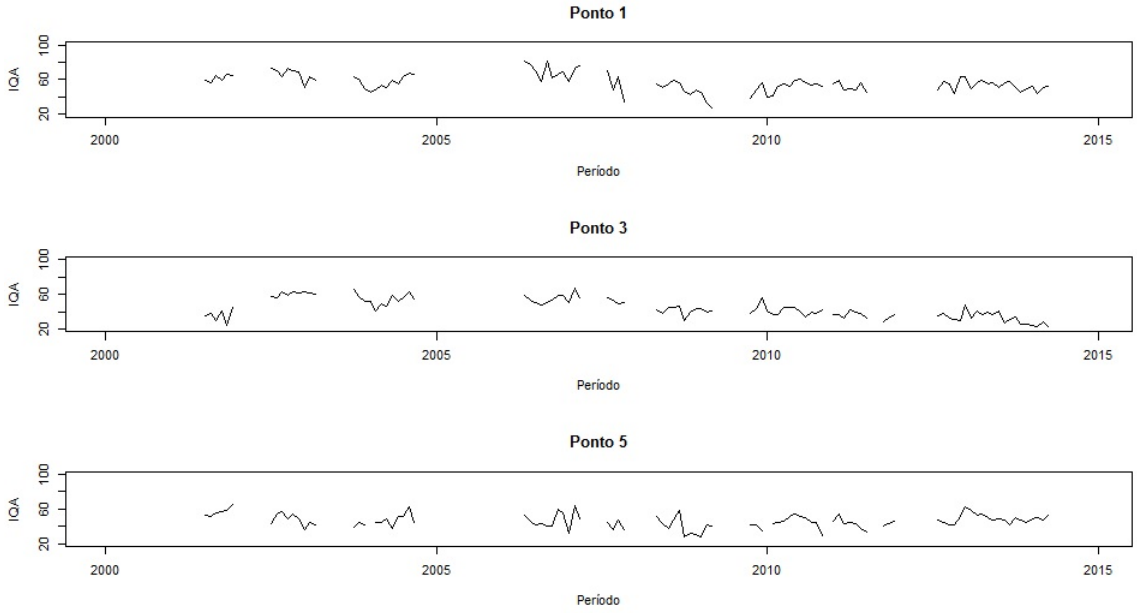


Figura 7: IQA ao longo do período estudado para cada ponto.

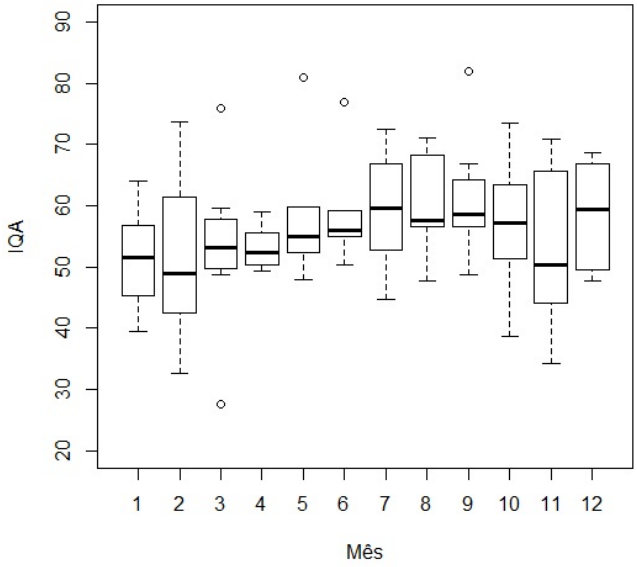


Figura 8: Boxplot das observações mensais para o Ponto 1.

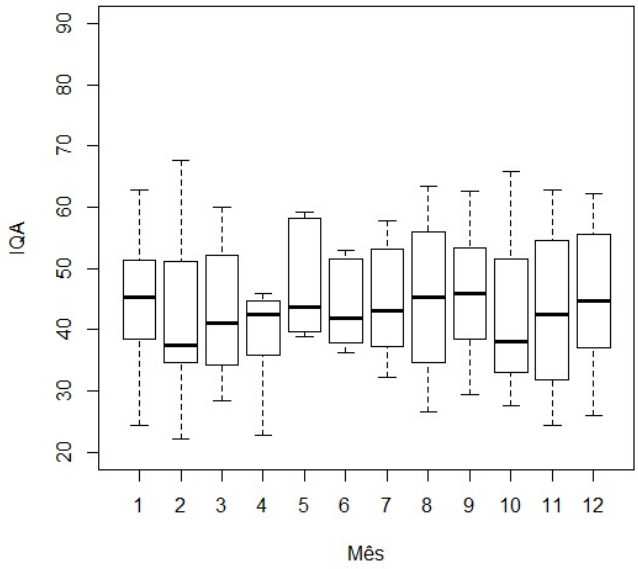


Figura 9: Boxplot das observações mensais para o Ponto 3.

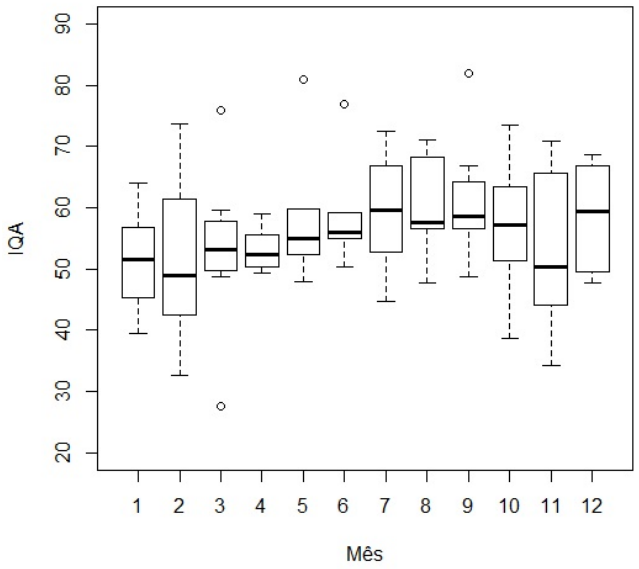


Figura 10: Boxplot das observações mensais para o Ponto 5.

Nota-se que aparentemente não há grande efeito de sazonalidade. A variabilidade dos dados possui alguma diferença dentro dos meses, mas não parece haver grande diferença entre os meses, devido ao período ao qual pertencem, no valor de tendência central. Além disso, é interessante notar que a variabilidade diferente não ocorre da mesma maneira entre os pontos.

Na Tabela 23 temos os resultados dos testes formais de t-estacionariedade e autocorrelação, levando-se em conta Lag 3.

Tabela 23: Testes formais de Estacionariedade

Teste	Ponto	Valor-p
PP test	Ponto 1	0.01
	Ponto 3	0.01
	Ponto 5	0.01
Ljung-Box	Ponto 1	< 0.01
	Ponto 3	< 0.01
	Ponto 5	< 0.01

Pode-se notar que, para Lag 3, o teste de Philipps-Perron rejeita a hipótese de que os pontos tenham raízes unitárias, assim dando indícios de haver t-estacionariedade. Por Ljung-Box nota-se que as autocorrelações, levando em conta Lag 3, não possuem dependência.

Para a matriz dos contrastes entre pontos há resultados similares, havendo evidências de t-estacionariedade.

5.2.1.2 Análise pelo Bootstrap em blocos

Agora passaremos à análise por meio do Bootstrap em blocos. O tamanho dos blocos utilizados foram determinados através dos métodos de seleção automática presentes na função "b.star", considerando-se os tamanhos para o Bootstrap em blocos os escolhidos para o Bootstrap circular e os valores esperados para os tamanhos dos blocos no Bootstrap estacionário os resultantes da seleção automática para esse método. Assim tem-se que os blocos tem tamanho 14 para todos os pontos no que se diz ao Bootstrap em Blocos e o valor esperado para o tamanho dos blocos no Bootstrap estacionário é de 10 para o Ponto 1, de 12 para o Ponto 3 e de 5 para o Ponto 5.

As Figuras 11 e 12 são apresentadas para melhor visualização da distribuição das estimativas das médias ao longo da realização das $B = 1000$ réplicas de Bootstrap para o Bootstrap em blocos móveis e nas Figuras 13 e 14 para o estacionário.

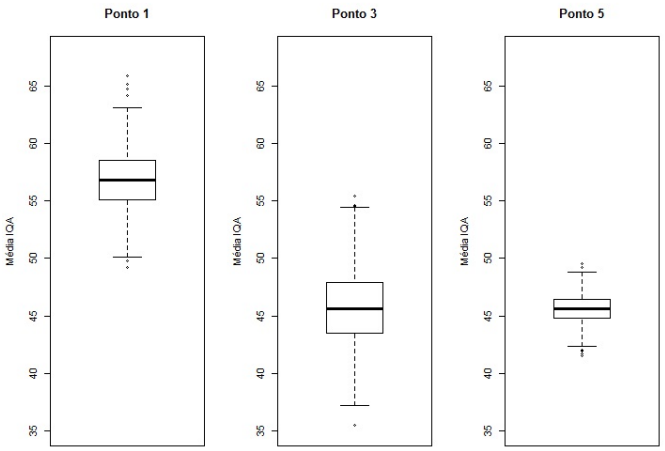


Figura 11: Boxplots das estimativas Bootstrap em blocos móveis.

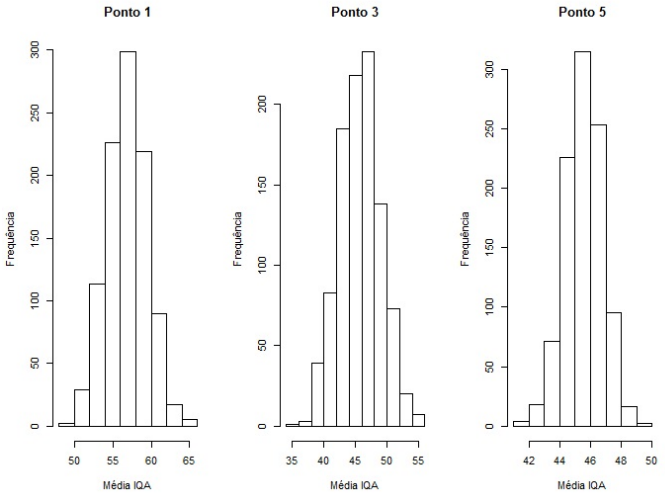


Figura 12: Histogramas das estimativas Bootstrap em blocos móveis.

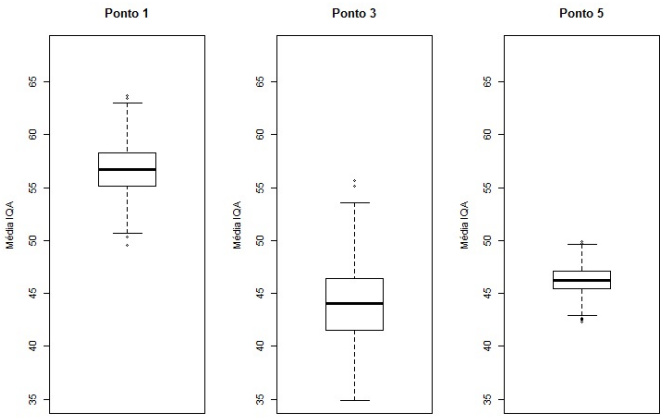


Figura 13: Boxplots das estimativas Bootstrap estacionário.

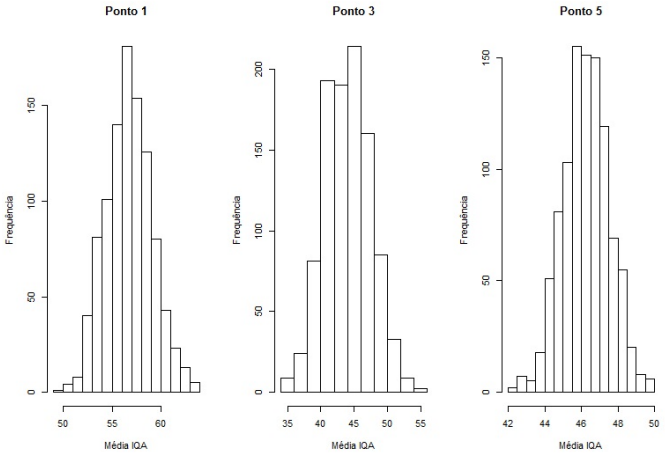


Figura 14: Histogramas das estimativas Bootstrap estacionário.

Nota-se que não há um número elevado de outliers e a distribuição é razoavelmente simétrica, o que é bom para a qualidade do Bootstrap. Os resultados do Bootstrap estacionário apresentam ainda menos valores de outliers.

Uma vez obtidos tais resultados passarremos à análise de possíveis diferenças entre os pontos no referente ao Índice de Qualidade da Água. Agora, os vetores utilizados para a realização do Bootstrap em blocos são os formados pelos contrastes entre os pontos, subtraindo o IQA de um ponto ao outro em observações para os mesmos meses.

Na Tabela 24 são apresentados os resultados para cada ponto de acordo com cada técnica utilizada. A letra “B” na frente de uma estatística indica o resultado da análise feita com o uso do Bootstrap em blocos móveis e a letra “E” pelo Bootstrap estacionário.

Tabela 24: Bootstrap para a Matriz Original por Ponto

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1	56.821	56.735	2.583	2.368	0.100	0.014
Ponto 3	45.729	44.021	3.242	3.442	1.678	-0.030
Ponto 5	45.571	46.259	1.194	1.270	-0.712	-0.025

Ainda é possível a obtenção dos intervalos de confiança individuais para cada ponto, com coeficiente de confiança de 95%, através dos quantis 0.025 e 0.975 de cada realização do Bootstrap. Tem-se assim os seguintes intervalos:

- Ponto 1 - Blocos: (51.794 ; 61.801) Estacionário: (52.433 ; 61.551)
- Ponto 3 - Blocos: (39.449 ; 52.042) Estacionário: (37.430 ; 50.733)
- Ponto 5 - Blocos: (43.224 ; 47.813) Estacionário: (43.761 ; 48.684)

Pode-se notar que os resultados são consideravelmente similares na maioria dos casos para ambas as técnicas de Bootstrap. O Ponto 3 é o ponto que apresenta maior diferença entre as estimativas obtidas. Para tal ponto, como o erro padrão apresenta pouca diferença e o vício utilizando o Bootstrap estacionário é bastante menor, o intervalo de confiança obtido por meio do Bootstrap estacionário parece mais fidedigno.

Quanto à classificação da água de acordo com o IQA-Cetesb temos que no Rio Jaguarí (Ponto 1) a água pode ser classificada como boa, com a média contida nessa classificação, porém tendo uma pequena distância da passagem para regular, como pode ser visto no intervalo de confiança.

A lagoa de estabilização (Ponto 3) tem o intervalo de confiança contendo tanto as classificações regular quanto boa da água, segundo o Bootstrap em blocos móveis com mais amplitude dentro da avaliação como regular do que como boa. Ao analisarmos o resultado obtido pelo Bootstrap estacionário, verifica-se que o intervalo de confiança contém apenas a classificação da água como regular, estando o limite inferior já próxima da passagem para ruim. Em ambos a média pontual encontra-se em regular.

Já o Rio Atibaia (Ponto 5) apresenta dentro de seu intervalo de confiança a classificação da água como sendo regular. O limite superior tem certa proximidade da classficação como boa, porém somente a classificação regular está contida.

Agora, quanto aos tamanhos dos blocos, para os três contrastes no Bootstrap em blocos foram utilizadas 13 observações por bloco. Já o comprimento esperado para o Bootstrap estacionário foi de 8, para os contrastes Ponto 1-Ponto 3, de 5 para Ponto 1-Ponto 5 e de 11 para Ponto 3 - Ponto 5.

A distribuição dos valores obtidos pelas médias à cada réplica do Bootstrap seguem o apresentado nas Figuras 15 e 16 para o Bootstrap em blocos móveis e nas Figuras 17 e 18 para o Bootstrap estacionário.

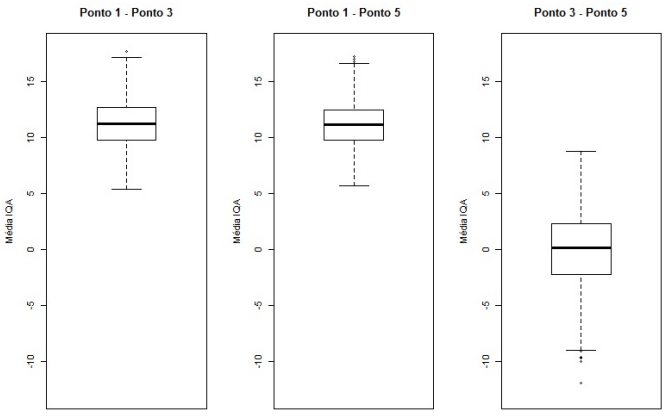


Figura 15: Boxplots das estimativas Bootstrap em blocos móveis - Contrastes.

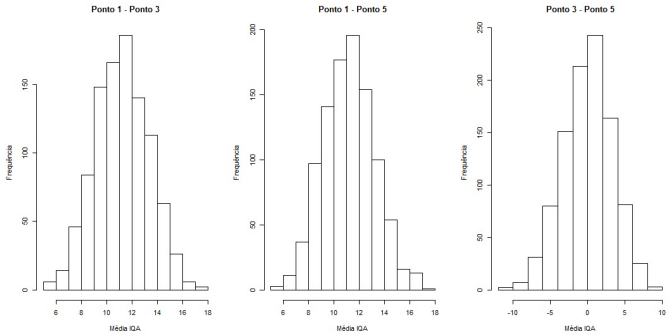


Figura 16: Histogramas das estimativas Bootstrap em blocos móveis - Contrastes.

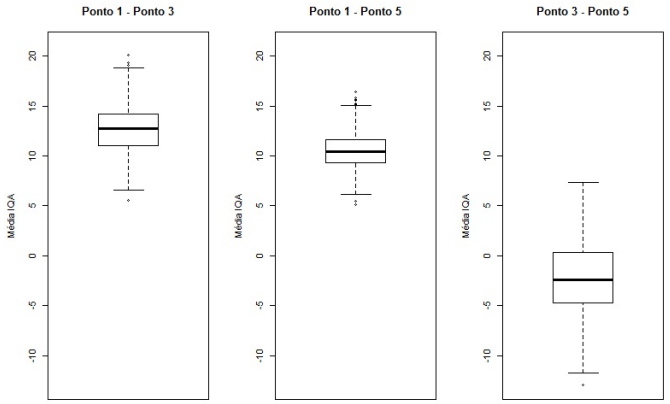


Figura 17: Boxplots das estimativas Bootstrap estacionário - Contrastes.

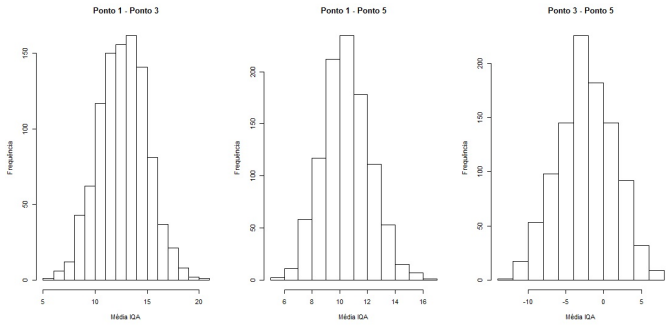


Figura 18: Histogramas das estimativas Bootstrap estacionário - Contrastes.

Nota-se que para os contrastes os dados comportam-se como o esperado. Há proximidade de simetria e há poucos valores de outliers levando em conta a média dos contrastes nas réplicas Bootstrap, tanto para o Bootstrap em blocos móveis quanto para o estacionário.

A Tabela 25 indica os resultados obtidos referentes as estimativas de algumas estatísticas para cada um dos contrastes realizados. As letras "B" e "E" tem o mesmo significado que na Tabela 24.

Tabela 25: Bootstrap para contrastes na Matriz Original

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1 - Ponto 3	11.200	12.663	2.111	2.300	-1.470	-0.007
Ponto 1 - Ponto 5	11.164	10.484	2.001	1.691	0.726	0.047
Ponto 3 - Ponto 5	-0.036	-2.320	3.273	3.715	2.196	-0.087

Quanto aos intervalos de confiança levando em conta os quantis 0.9% e 99.1% a fim de que a análise dos intervalos em conjunto atinja 95% de confiança temos:

- Ponto 1-Ponto 3 - Blocos: (6.227 ; 15.993) Estacionário: (7.207 ; 18.187)
- Ponto 1-Ponto 5 - Blocos: (6.637 ; 16.132) Estacionário: (6.907 ; 14.719)

- Ponto 3-Ponto 5 - Blocos: (-7.789 ; 7.243) Estacionário: (-10.886; 5.892)

Nota-se mais uma vez que existem pequenas diferenças quanto ao resultado das estimativas pontuais obtidas pelos dois métodos e até mesmo entre os intervalos de confiança, porém a conclusão quanto à diferença significativa do IQA entre os rios é a mesma tanto para o Bootstrap em blocos quanto para o Bootstrap estacionário.

Os resultados mostram que o Ponto 1 possui o IQA maior em relação tanto ao Ponto 3 quanto ao Ponto 5. Já os pontos 3 e 5 não tem diferença significativa quanto ao valor do IQA no período analisado.

Isso indica que a qualidade da água no Rio Atibaia está de fato inferior à do rio Jaguarí e inclusive não é significantemente melhor até mesmo que a água na lagoa de estabilização antes de ser devolvida para o rio.

5.2.2 Matriz Imputada

Agora passamos para a análise já com a matriz dos dados imputados, possuindo assim valores atribuídos ao IQA para maior número de meses ao longo dos anos de 2000 e 2014.

Há, após a imputação, 134 valores de IQA disponíveis para todos os pontos. Os meses nos quais não houve coleta de água não tiveram seus valores de IQA imputados, uma vez que a qualidade da imputação para esses meses era bastante baixa.

Para a análise dos dados o IQA calculado para cada ponto e estação será considerado com três casas decimais.

Como no caso anterior temos falhas na suposição da ANOVA que impossibilitam a realização da análise por meio de tal técnica. Agora as suposições de normalidade e homocedasticidade são válidas, para todas as 5 imputações, porém, a hipótese de independência dos resíduos continua sendo rejeitada, inviabilizando assim a utilização da técnica.

Com isso a utilização da Análise de Variância não é a adequada.

Como os dados apresentam dependência, não sendo i.i.d., até mesmo a utilização de métodos não paramétricos não é adequada, bem como na análise na matriz original.

Novamente será utilizado o Bootstrap a fim de comparar as médias para os pontos e períodos. O Bootstrap em blocos é novamente realizado, uma vez que leva em conta a dependência entre os dados.

5.2.2.1 Análise dos dados imputados

Primeiramente serão obtidos resultados para cada um dos grupos na matriz imputada, procedidos da análise dos contrastes que permitem a comparação da existência ou não de diferenças estatisticamente significantes entre os mesmos quanto ao IQA. Para isso, por tratar-se de um método de imputação múltipla, devem ser analisados os 5 conjuntos de dados resultantes, a fim de levar em conta a variabilidade existente na análise.

Os resultados para cada imputação foram bastante semelhantes, quanto ao IQA final, na Figura 19 têm-se a dispersão de IQA ao longo do tempo, considerando-se as ausências.

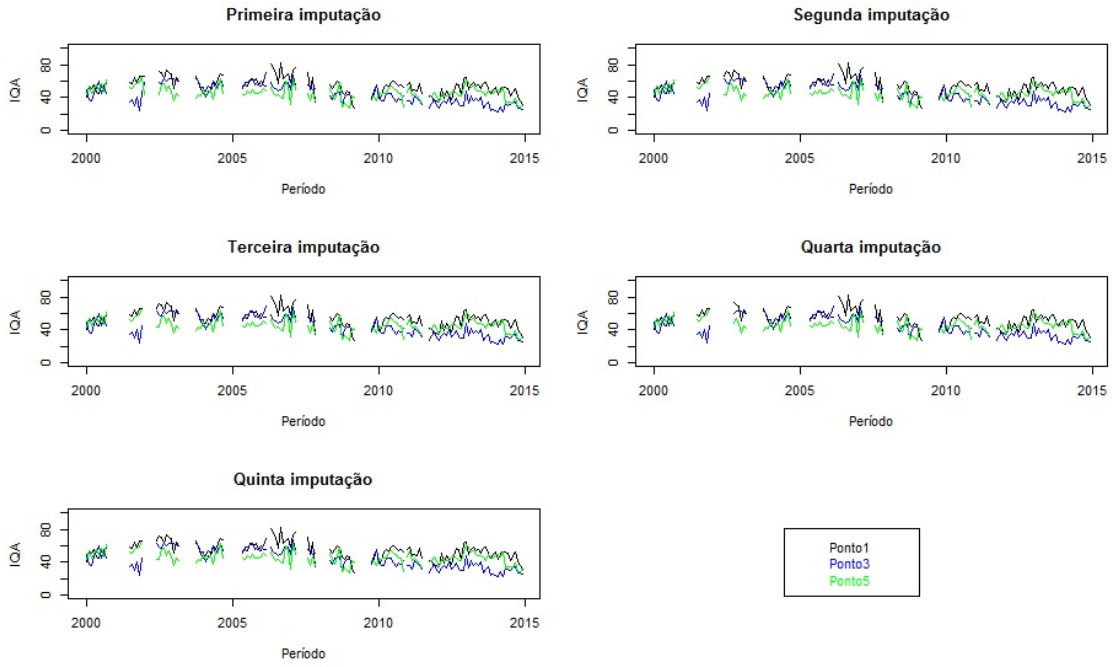


Figura 19: IQA para cada imputação por ponto.

Os resultados quanto à $t-$estacionariedade são parecidos com os anteriores à imputação, porém agora para Lag 4, como pode-se ver na Tabela 26.

Tabela 26: Testes formais de Estacionariedade

Teste	Ponto	Valor-p
PP test	Ponto 1	0.01
	Ponto 3	0.01
	Ponto 5	0.01
Ljung-Box	Ponto 1	< 0.01
	Ponto 3	< 0.01
	Ponto 5	< 0.01

Pode-se notar que, para Lag 4, o teste de Philipps-Perron rejeita a hipótese de que os pontos tenham raízes unitárias, assim dando indícios de haver t-estacionariedade. Por Ljung-Box nota-se as autocorrelações foram tidas como não dependentes, também levando em conta Lag 4. Para a matriz dos contrastes entre pontos também há resultados similares, havendo evidências de estacionariedade.

5.2.2.2 Análise pelo Bootstrap em Blocos

Para a realização do Bootstrap, o tamanho dos blocos selecionados foram de 19 para o Bootstrap em Blocos móveis e os tamanhos esperado para o Bootstrap estacionário de 12 para o Ponto 1, 17 para o Ponto 3 e 8 para o Ponto 5, considerando-se a primeira imputação. Para as demais imputações os valores foram os mesmos.

Pode-se notar que para todas as imputações não há grandes problemas de falta de simetria na distribuição empírica das médias Bootstrap obtidas, como pode ser visto nas Figuras $20 - 29$, tanto para o Bootstrap em blocos móveis quanto para o Bootstrap estacionário.

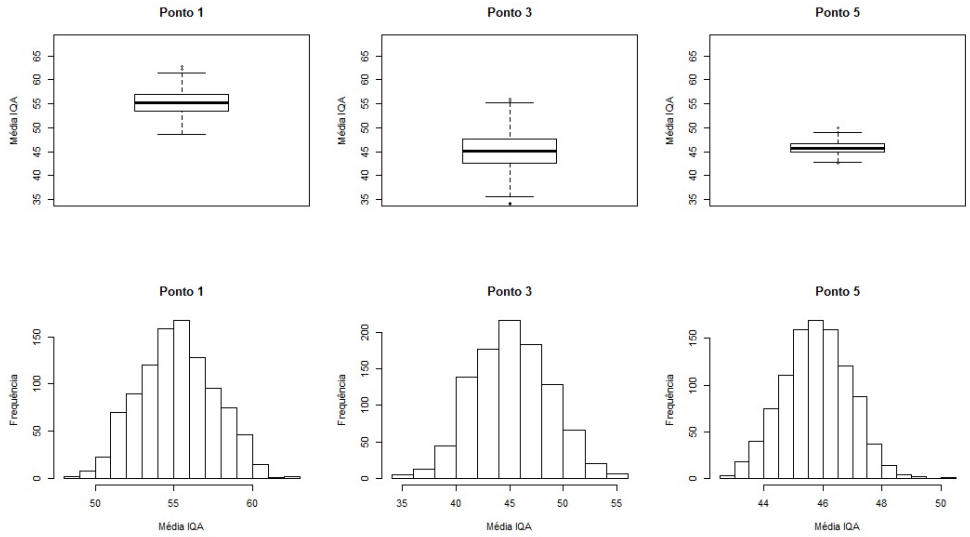


Figura 20: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a primeira imputação.

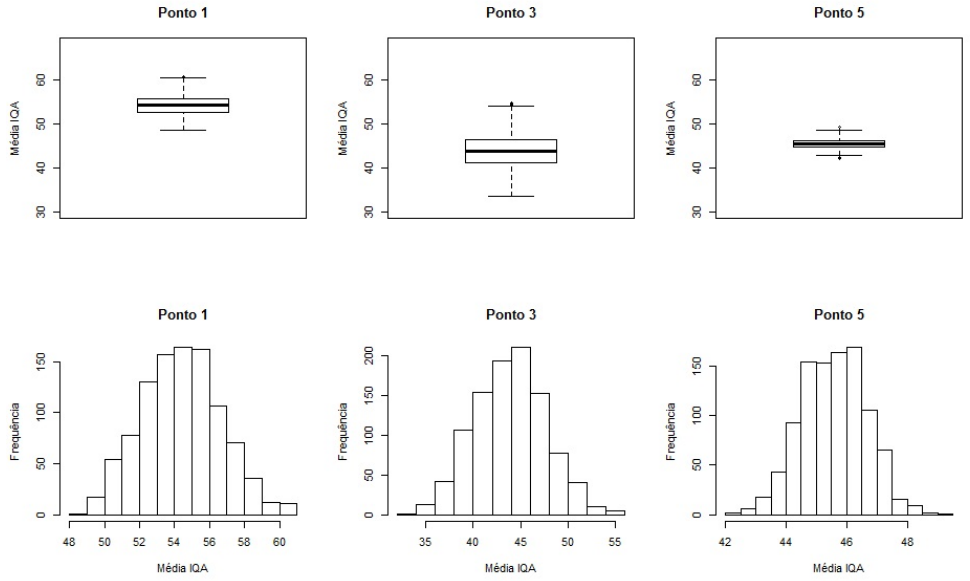


Figura 21: Histogramas e Boxplots das estimativas Bootstrap estacionário para a primeira imputação.

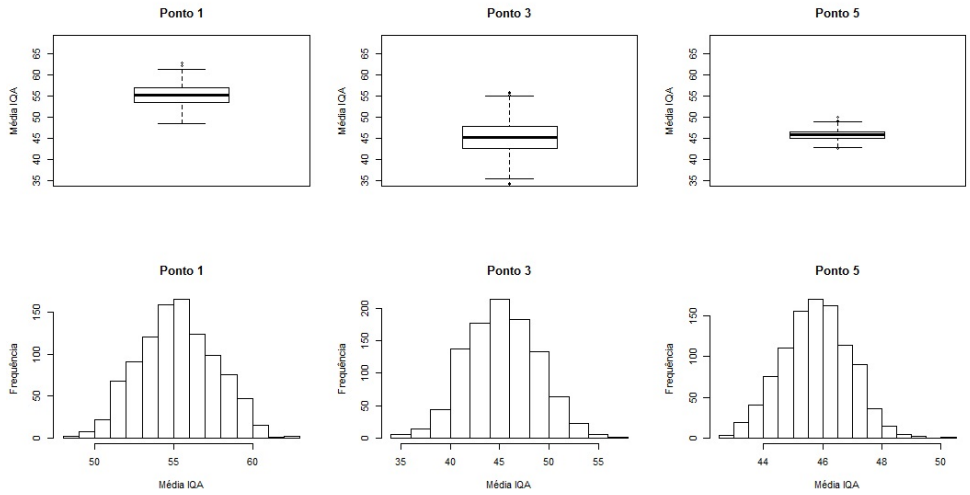


Figura 22: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a segunda imputação.

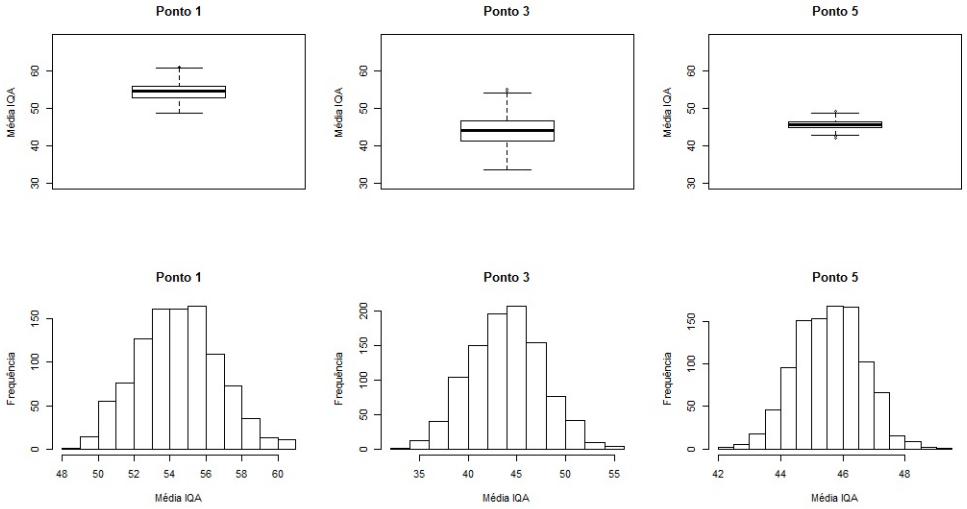


Figura 23: Histogramas e Boxplots das estimativas Bootstrap estacionário para a segunda imputação.

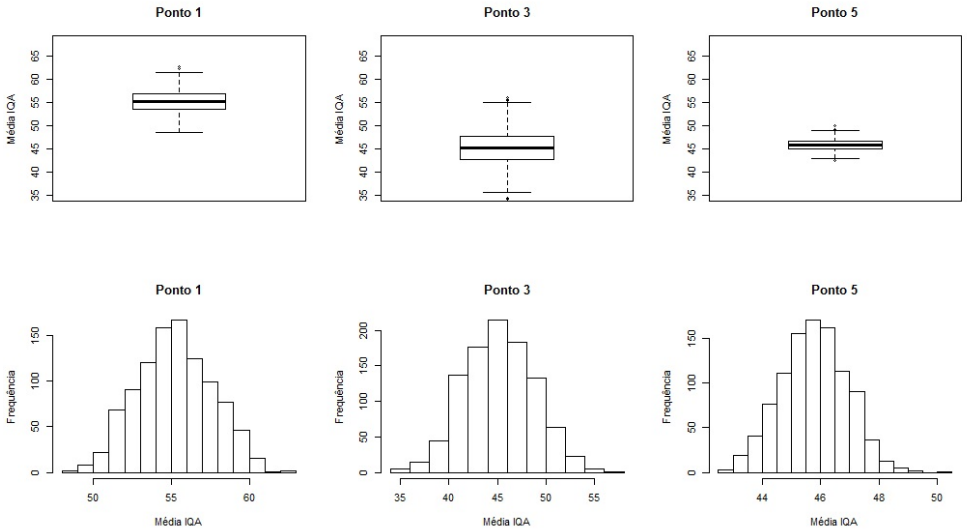


Figura 24: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a terceira imputação.

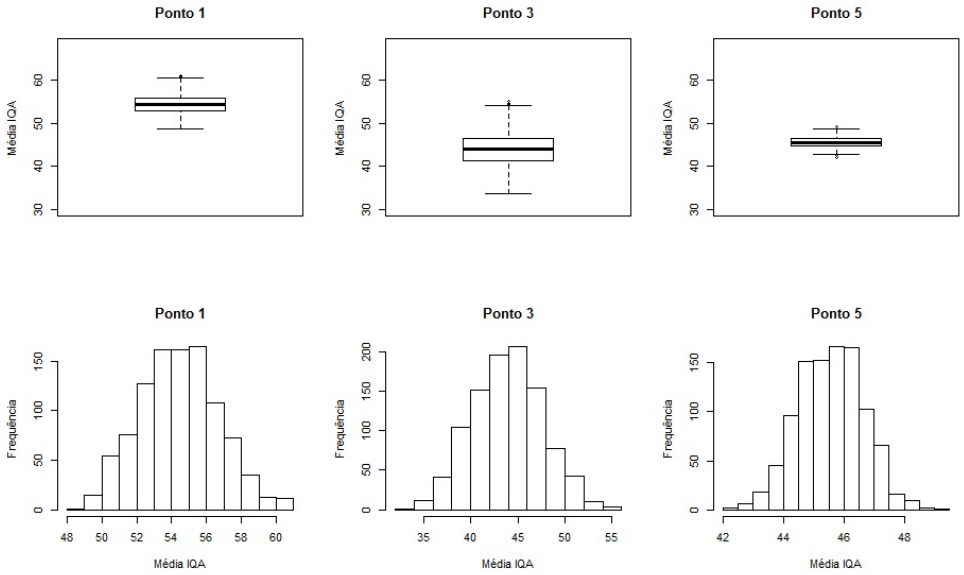


Figura 25: Histogramas e Boxplots das estimativas Bootstrap estacionário para a terceira imputação.

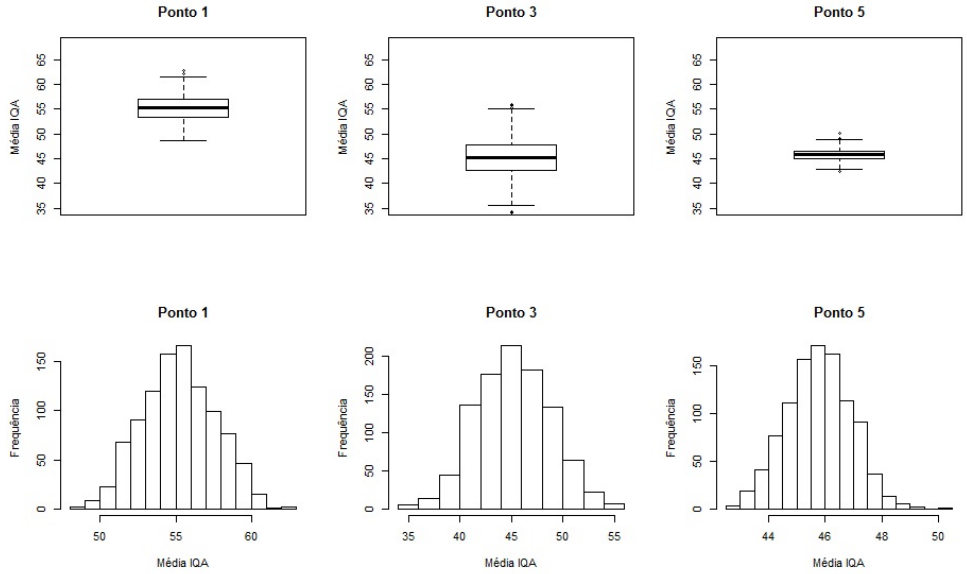


Figura 26: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quarta imputação.

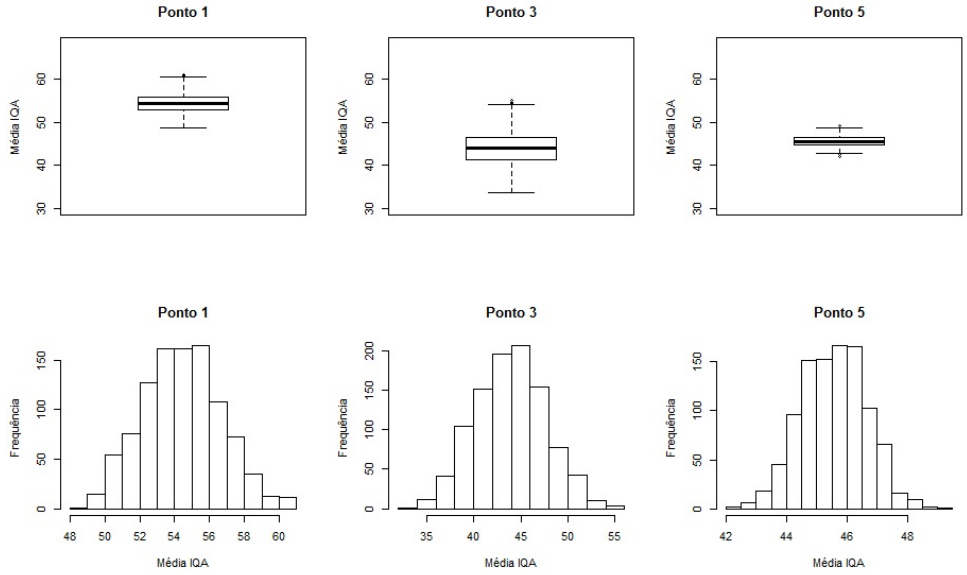


Figura 27: Histogramas e Boxplots das estimativas Bootstrap estacionário para a quarta imputação.

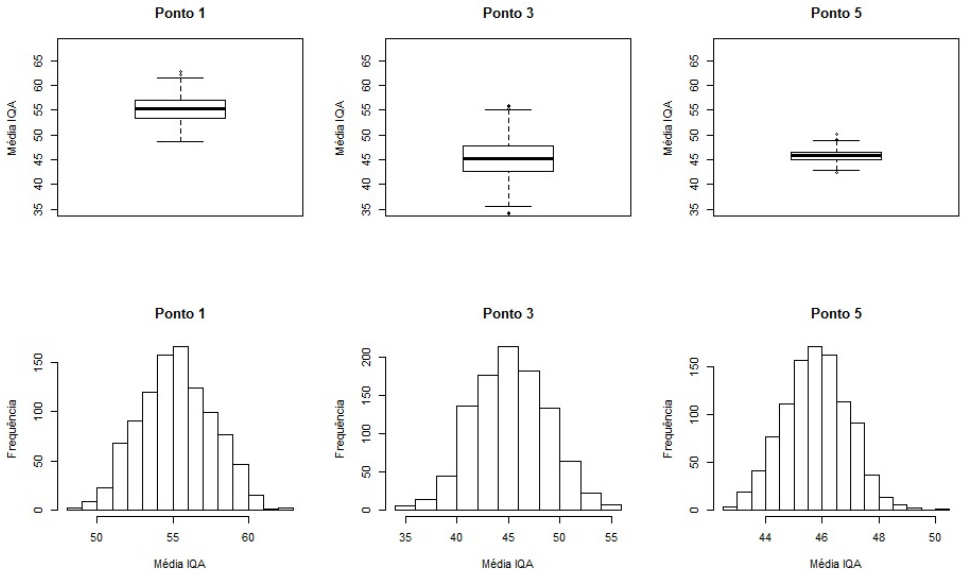


Figura 28: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quinta imputação.

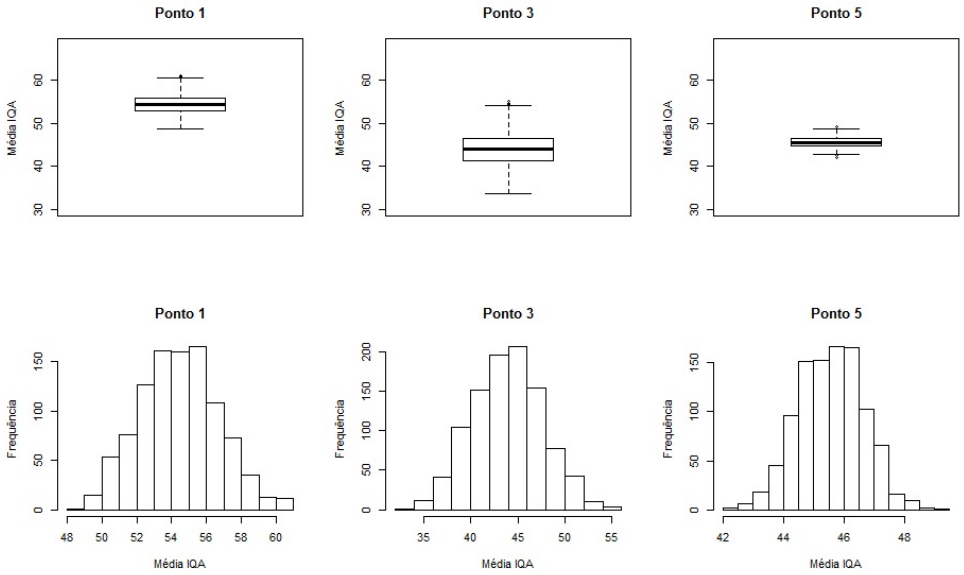


Figura 29: Histogramas e Boxplots das estimativas Bootstrap estacionário para a quinta imputação.

Nas Tabelas $27 - 31$ estão apresentados os resultados para cada ponto de acordo com cada técnica utilizada. As análises são realizadas considerando-se o resultado obtido para cada uma das 5 imputações, a fim de levar em conta a variabilidade da imputação múltipla. A letra "B" na frente de uma estatística indica o resultado da análise feita com o uso do Bootstrap em blocos móveis e a letra "E" pelo Bootstrap estacionário, para a primeira imputação.

Tabela 27: Bootstrap para a Matriz Imputada - Primeira imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1	55.194	54.403	2.415	2.266	0.849	0.057
Ponto 3	45.181	43.899	3.561	3.707	1.211	-0.070
Ponto 5	45.773	45.573	1.127	1.063	0.231	0.031

Tabela 28: Bootstrap para a Matriz Imputada - Segunda imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1	55.218	54.434	2.411	2.260	0.841	0.058
Ponto 3	45.207	43.933	3.576	3.718	1.204	-0.070
Ponto 5	45.764	45.563	1.132	1.068	0.232	0.031

Tabela 29: Bootstrap para a Matriz Imputada - Terceira imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1	55.216	54.433	2.410	2.259	0.841	0.058
Ponto 3	45.208	43.934	3.576	3.718	1.204	-0.070
Ponto 5	45.765	45.563	1.133	1.068	0.233	0.031

Tabela 30: Bootstrap para a Matriz Imputada - Quarta imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1	55.216	54.433	2.410	2.260	0.841	0.058
Ponto 3	45.208	43.934	3.576	3.718	1.204	-0.070
Ponto 5	45.765	45.563	1.133	1.068	0.232	0.031

Tabela 31: Bootstrap para a Matriz Imputada - Quinta imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1	55.217	54.434	2.410	2.260	0.841	0.058
Ponto 3	45.208	43.934	3.576	3.718	1.204	-0.070
Ponto 5	45.765	45.563	1.133	1.068	0.232	0.031

Ainda é possível a obtenção dos intervalos de confiança individuais para cada ponto, com 95% de confiança, através dos quantis 0.025 e 0.975 de cada realização do Bootstrap. Tem-se assim os seguintes intervalos:

Primeira imputação:

- Ponto 1 - Blocos: (50.813 ; 59.694) Estacionário: (50.233 ; 58.942)
- Ponto 3 - Blocos: (38.602 ; 52.119) Estacionário: (36.917 ; 50.820)
- Ponto 5 - Blocos: (43.586 ; 47.902) Estacionário: (43.505 ; 47.554)

Segunda imputação:

- Ponto 1 - Blocos: (50.852 ; 59.704) Estacionário: (50.272 ; 58.958)
- Ponto 3 - Blocos: (38.599 ; 52.157) Estacionário: (36.915 ; 50.879)
- Ponto 5 - Blocos: (43.563 ; 47.896) Estacionário: (43.490 ; 47.563)

Terceira imputação:

- Ponto 1 - Blocos: (50.851 ; 59.703) Estacionário: (50.272 ; 58.959)
- Ponto 3 - Blocos: (38.599 ; 52.157) Estacionário: (36.916 ; 50.879)
- Ponto 5 - Blocos: (43.565 ; 47.897) Estacionário: (43.491 ; 47.563)

Quarta imputação:

- Ponto 1 - Blocos: (50.851 ; 59.703) Estacionário: (50.272 ; 58.960)
- Ponto 3 - Blocos: (38.600 ; 52.157) Estacionário: (36.916 ; 50.879)
- Ponto 5 - Blocos: (43.564 ; 47.897) Estacionário: (43.491 ; 47.563)

Quinta imputação:

- Ponto 1 - Blocos: (50.851 ; 59.704) Estacionário: (50.273 ; 58.961)
- Ponto 3 - Blocos: (38.600 ; 52.157) Estacionário: (36.917 ; 50.879)
- Ponto 5 - Blocos: (43.564 ; 47.897) Estacionário: (43.491 ; 47.563)

Pode-se notar que os resultados são bastante similares na maioria dos casos para ambas as técnicas de Bootstrap. O Ponto 3 é novamente o ponto que apresenta maior diferença entre as estimativas obtidas. Como mais uma vez o erro padrão apresenta pouca diferença e o vício utilizando o Bootstrap estacionário é bastante menor, o intervalo de confiança obtido por meio do Bootstrap estacionário parece mais fidedigno.

Ainda é notável que há uma variação de fato bastante pequena nos resultados obtidos a cada imputação, tendo-se tanto estimativas pontuais quanto intervalos de confiança bastante similares a cada uma das 5 matrizes imputadas.

Quanto à classificação da água de acordo com o IQA-Cetesb temos que no Rio Jaguarí (Ponto 1) a água é classificada como boa, porém o limite inferior está bastante perto da passagem para regular.

A lagoa de estabilização (Ponto 3) tem o intervalo de confiança contendo tanto as classificações regular quanto boa da água, com muito mais amplitude dentro da avaliação como regular do que como boa. Agora a média para esse ponto encontra-se na qualidade regular de acordo com todos os métodos.

Já o Rio Atibaia (Ponto 5) apresenta dentro de seu intervalo de confiança a classificação da água como sendo regular.

Uma vez obtidos tais resultados, é interessante também analisar possíveis diferenças entre os pontos no referente ao Índice de Qualidade da Água, também considerando cada matriz imputada.

O tamanho dos blocos selecionados foram de 18 para o Bootstrap em Blocos móveis e os tamanhos esperado para o Bootstrap estacionário de 9 para o Ponto 1 - Ponto 3, 10 para o Ponto 1 - Ponto 5 e 15 para o Ponto 3 - Ponto 5, considerando-se a primeira imputação. Já para a segunda são de 18 para o Bootstrap em blocos móveis e os valores esperados são 9, 10 e 15 no Bootstrap estacionário, bem como na terceira, na quarta e na quinta matrizes imputadas.

Para os contrastes, observa-se ainda mais simetria na distribuição das médias e ausência de muitos valores de outliers, como pode ser visto nas Figuras $30 - 39$.

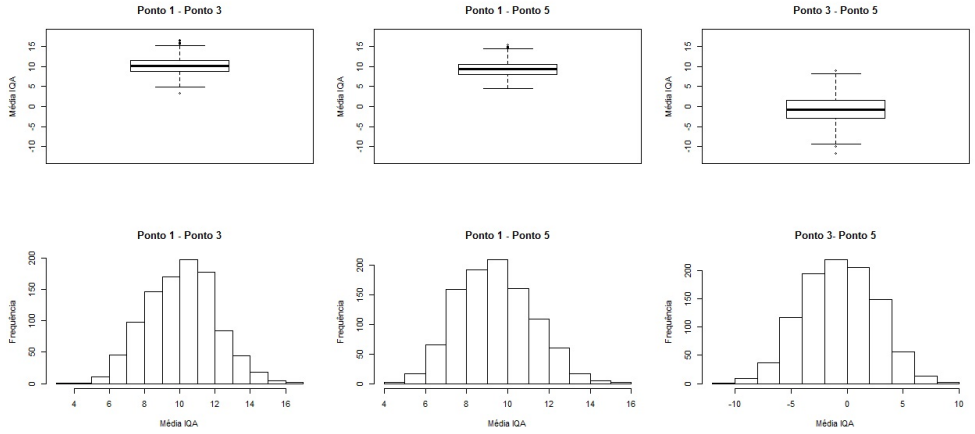


Figura 30: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a primeira imputação - Contrastes.

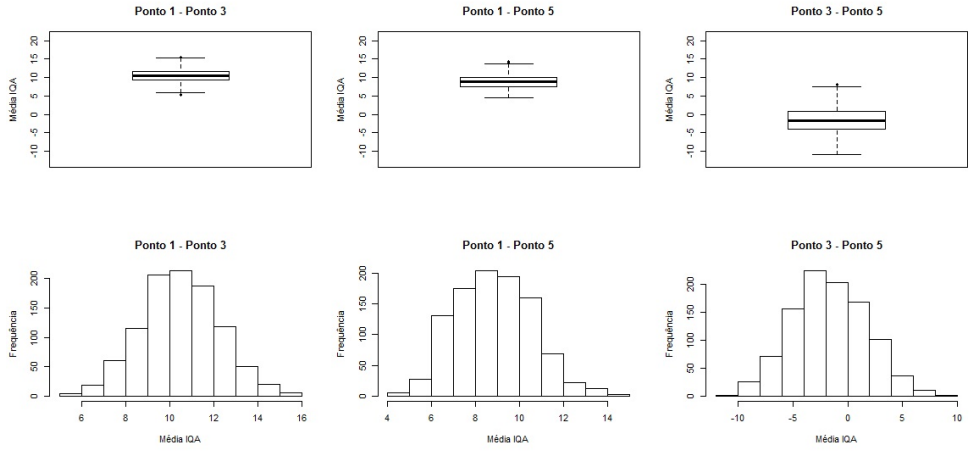


Figura 31: Histogramas e Boxplots das estimativas Bootstrap estacionário para a primeira imputação - Contrastes.

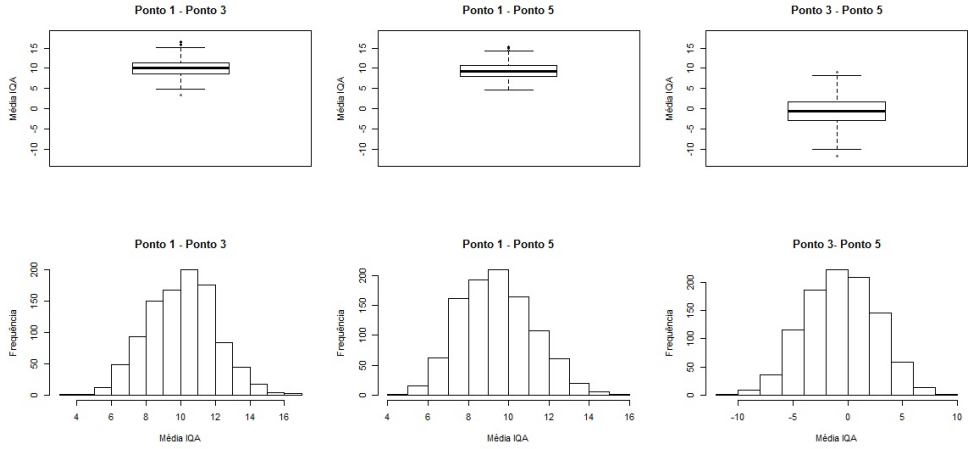


Figura 32: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a segunda imputação - Contrastes.

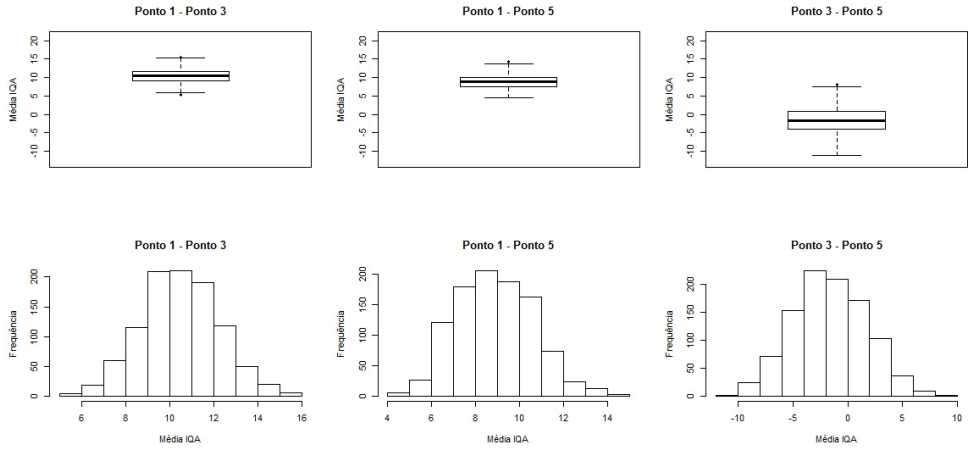


Figura 33: Histogramas e Boxplots das estimativas Bootstrap estacionário para a segunda imputação - Contrastes.

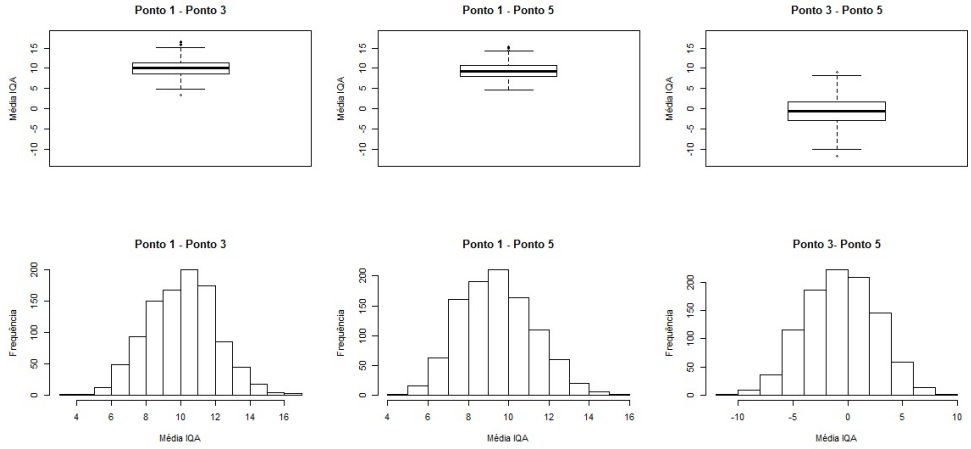


Figura 34: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a terceira imputação - Contrastes.

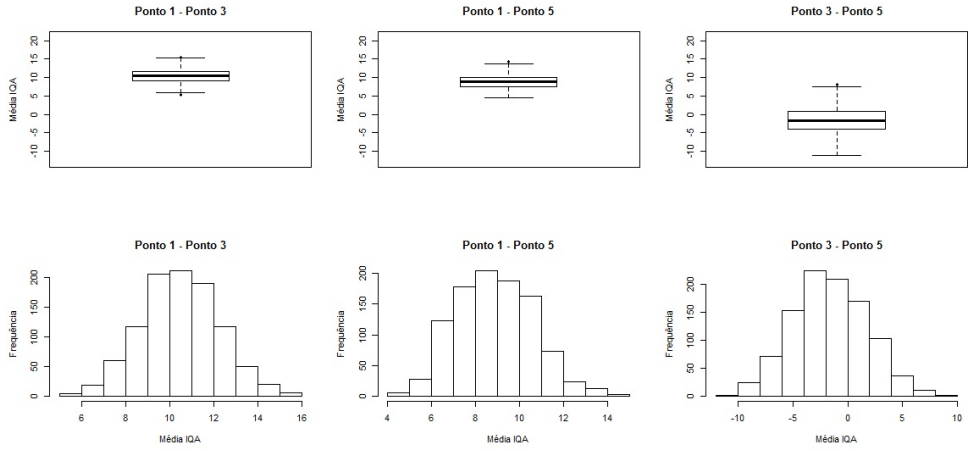


Figura 35: Histogramas e Boxplots das estimativas Bootstrap estacionário para a terceira imputação - Contrastes.

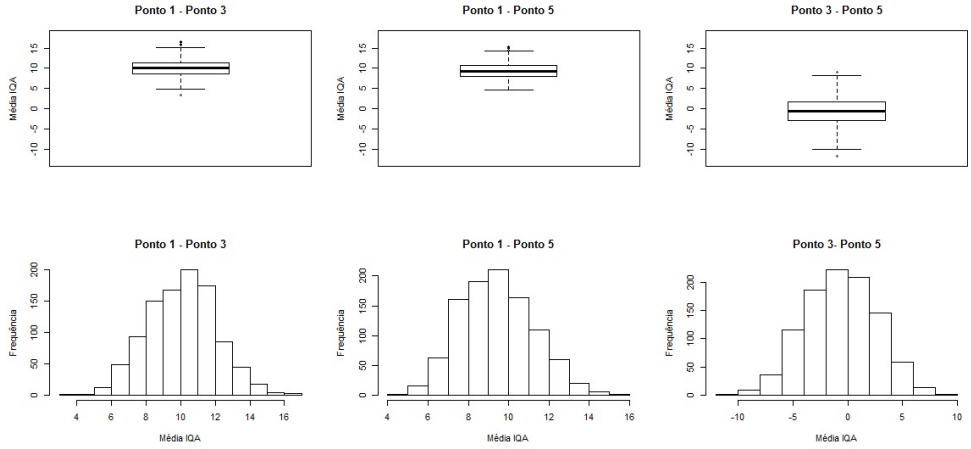


Figura 36: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quarta imputação - Contrastes.

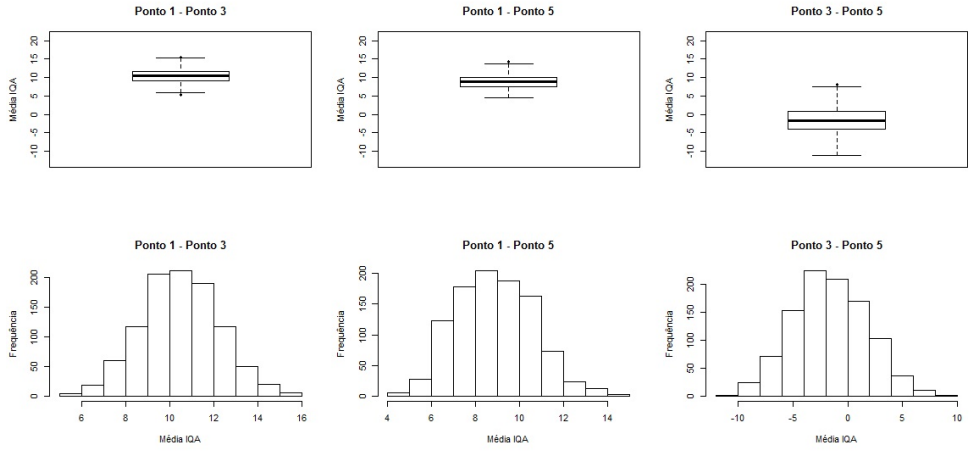


Figura 37: Histogramas e Boxplots das estimativas Bootstrap estacionário para a quarta imputação - Contrastes.

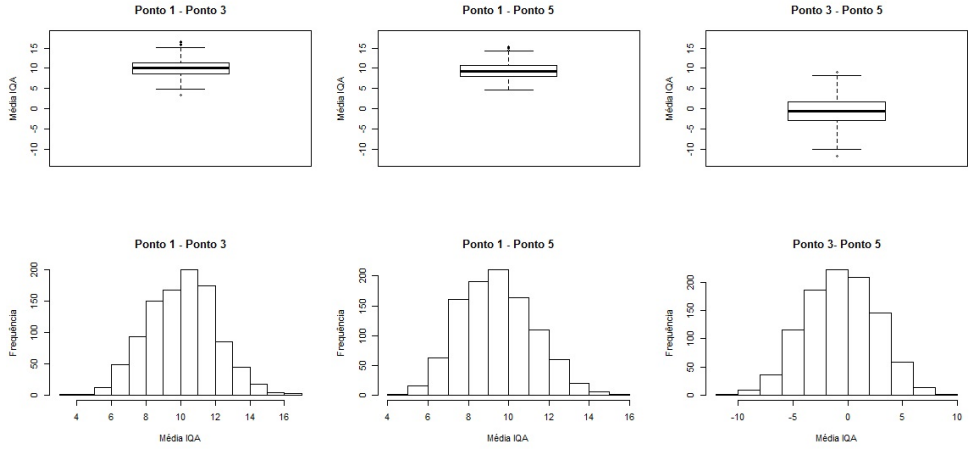


Figura 38: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis para a quinta imputação - Contrastes.

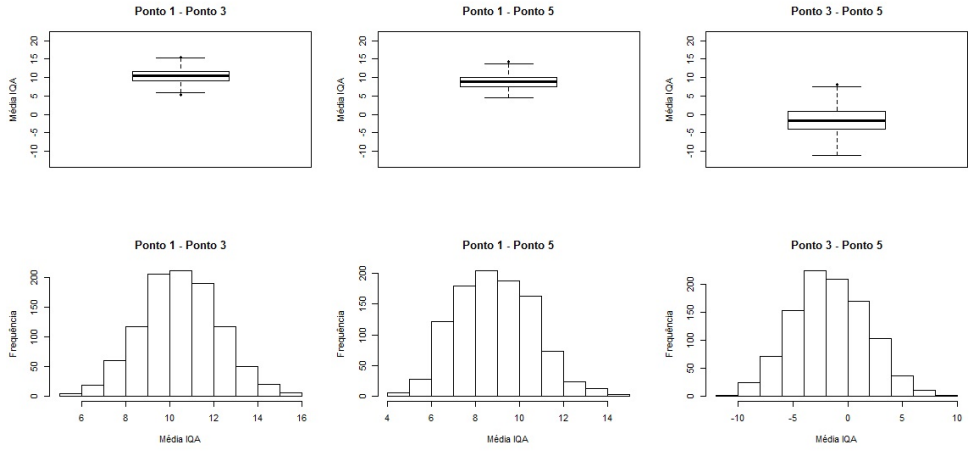


Figura 39: Histogramas e Boxplots das estimativas Bootstrap estacionário para a quinta imputação - Contrastes.

As Tabelas $32 - 36$ indicam os resultados obtidos referentes as estimativas de algumas estatísticas para cada um dos contrastes realizados. As letras "B" e "E" tem o mesmo significado que na Tabela 24.

Tabela 32: Bootstrap para contrastes na Matriz Imputada - Primeira Imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1 - Ponto 3	10.070	10.473	1.970	1.763	-0.302	0.098
Ponto 1 - Ponto 5	9.372	8.844	1.799	1.738	0.569	0.040
Ponto 3 - Ponto 5	-0.701	-1.672	3.214	3.353	0.871	-0.100

Tabela 33: Bootstrap para contrastes na Matriz Imputada - Segunda Imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1 - Ponto 3	10.069	10.468	1.977	1.768	-0.304	0.095
Ponto 1 - Ponto 5	9.404	8.844	1.797	1.737	0.559	0.039
Ponto 3 - Ponto 5	-0.665	-1.627	3.224	3.361	0.864	-0.099

Tabela 34: Bootstrap para contrastes na Matriz Imputada - Terceira Imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1 - Ponto 3	10.066	10.466	1.977	1.768	-0.305	0.095
Ponto 1 - Ponto 5	9.401	8.882	1.796	1.736	0.559	0.039
Ponto 3 - Ponto 5	-0.665	-1.628	3.224	3.361	0.864	-0.099

Tabela 35: Bootstrap para contrastes na Matriz Imputada - Quarta Imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1 - Ponto 3	10.067	10.466	1.977	1.768	-0.305	0.095
Ponto 1 - Ponto 5	9.402	8.882	1.796	1.736	0.559	0.039
Ponto 3 - Ponto 5	-0.665	-1.628	3.223	3.361	0.864	-0.099

Tabela 36: Bootstrap para contrastes na Matriz Imputada - Quinta Imputação

Ponto	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Ponto 1 - Ponto 3	10.067	10.466	1.977	1.768	-0.305	0.095
Ponto 1 - Ponto 5	9.402	8.883	1.796	1.736	0.559	0.039
Ponto 3 - Ponto 5	-0.665	-1.627	3.223	3.360	0.863	-0.099

Quanto aos intervalos de confiança levando em conta os quantis 0.9% e 99.1% a fim de que a análise dos intervalos em conjunto atinja 95% de confiança temos:

Primeira imputação:

- Ponto 1-Ponto 3 - Blocos: (5.840 ; 14.714) Estacionário: (6.343 ; 14.730)
- Ponto 3-Ponto 5 - Blocos: (5.546 ; 13.622) Estacionário: (5.126 ; 13.098)
- Ponto 3-Ponto 5 - Blocos: (-8.112 ; 6.532) Estacionário: (-8.664 ; 6.213)

Segunda imputação:

- Ponto 1-Ponto 3 - Blocos: (5.826 ; 14.783) Estacionário: (6.356 ; 14.789)
- Ponto 3-Ponto 5 - Blocos: (5.585 ; 13.655) Estacionário: (5.170 ; 13.137)
- Ponto 3-Ponto 5 - Blocos: (-8.132 ; 6.589) Estacionário: (-8.687 ; 6.227)

Terceira imputação:

- Ponto 1-Ponto 3 - Blocos: (5.822 ; 14.781) Estacionário: (6.352 ; 14.781)
- Ponto 3-Ponto 5 - Blocos: (5.577 ; 13.647) Estacionário: (5.170 ; 13.137)
- Ponto 3-Ponto 5 - Blocos: (-8.133 ; 6.687) Estacionário: (-8.686 ; 6.273)

Quarta imputação:

- Ponto 1-Ponto 3 - Blocos: (5.823 ; 14.781) Estacionário: (6.353 ; 14.781)
- Ponto 3-Ponto 5 - Blocos: (5.577 ; 13.648) Estacionário: (5.170 ; 13.139)
- Ponto 3-Ponto 5 - Blocos: (-8.132 ; 6.687) Estacionário: (-8.686 ; 6.273)

Quinta imputação:

- Ponto 1-Ponto 3 - Blocos: (5.823 ; 14.781) Estacionário: (6.354 ; 14.781)
- Ponto 3-Ponto 5 - Blocos: (5.578 ; 13.648) Estacionário: (5.171 ; 13.140)
- Ponto 3-Ponto 5 - Blocos: (-8.132 ; 6.687) Estacionário: (-8.685 ; 6.273)

Mais uma vez os resultados não tem grande variação entre as matrizes imputadas. Para os contrastes nas matrizes imputadas o Bootstrap estacionário novamente apresenta menor vício, tendo erro padrão bastante parecido com o Bootstrap em blocos móveis.

Quanto aos resultados obtidos, têm-se que, com 95% de confiança, o Ponto 1 apresenta IQA estatisticamente maior que os demais pontos, considerando-se todo o período estudado. Além disso, os Pontos 3 e 5 não apresentam diferença significativa no IQA. A qualidade da água no Rio Atibaia mostra-se assim como preocupante.

5.2.2.3 Análise por quinquênios

Agora, a fim de verificar se houveram melhorias com o tempo, serão apresentados resultados para cada um dos quinquênios no estudo ($2000-2004; 2005-2009; 2010-2014$). O intuito é comparar, dentro de cada ponto, se com a passagem de cada quinquênio há diferença significativa na qualidade da água. Como os dados são desbalanceados levando-se em conta os quinquênios, uma vez que nos anos mais recentes há maior número de coletas, não é possível realizar o contraste entre os períodos dentro de cada ponto.

Considerando-se os dados por quinquênio, há 37 dados para os primeiros cinco anos do estudo, 40 para o segundo quinquênio e 57 para o último.

Ao serem analisados os dados a fim de realizar a análise pelo Bootstrap em blocos, percebeu-se que, apesar de para os dados ao longo dos 15 anos como um todo haver um comportamento adequado, ao dividir o banco de dados em quinquênios passam a existir alguns problemas para a aplicação da técnica.

Para o Ponto 1, em todas as cinco matrizes imputadas, há evidências de não estacionariedade para o primeiro quinquênio (pelo teste de Phillips-Perron). Para o Ponto 3 há evidências de não estacionariedade para o primeiro e também para o segundo quinquênio. Apenas o Ponto 5 apresenta indícios de estacionariedade para todos os quinquênios, considerando-se eles um a um.

Além disso, verifica-se que, ao realizar a análise pelo Bootstrap em blocos, tanto pelos Blocos móveis quanto pelo Estacionário, há grande assimetria das médias obtidas ao longo das B=1000 médias obtidas ao longo da realização das reamostras em vários pontos em determinados quinquênios, além de existir o problema de caudas pesadas em diversas realizações. Como em todas as imputações os resultados são semelhantes, serão apresentados os gráficos apenas para uma das imputações, como pode ser visto nas Figuras $40 - 42$ para o Bootstrap em blocos móveis e nas Figuras $43 - 45$ para o Bootstrap estacionário.

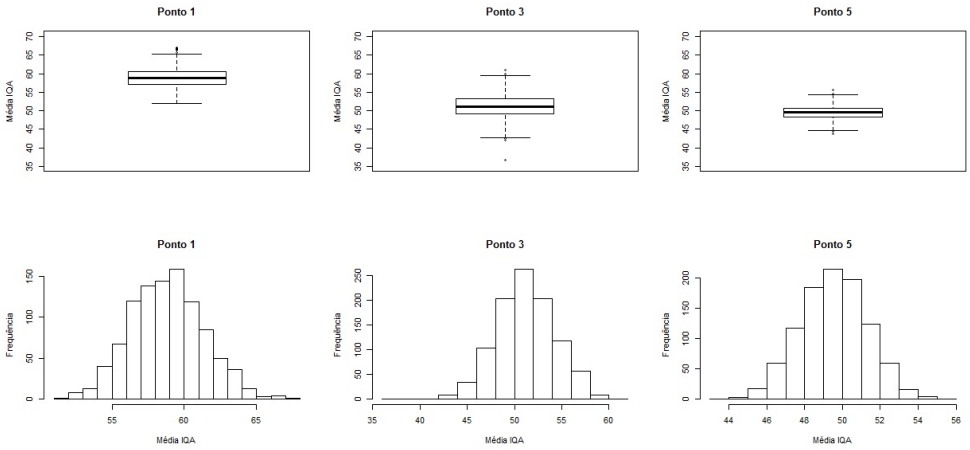


Figura 40: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis - Primeiro quinquênio.

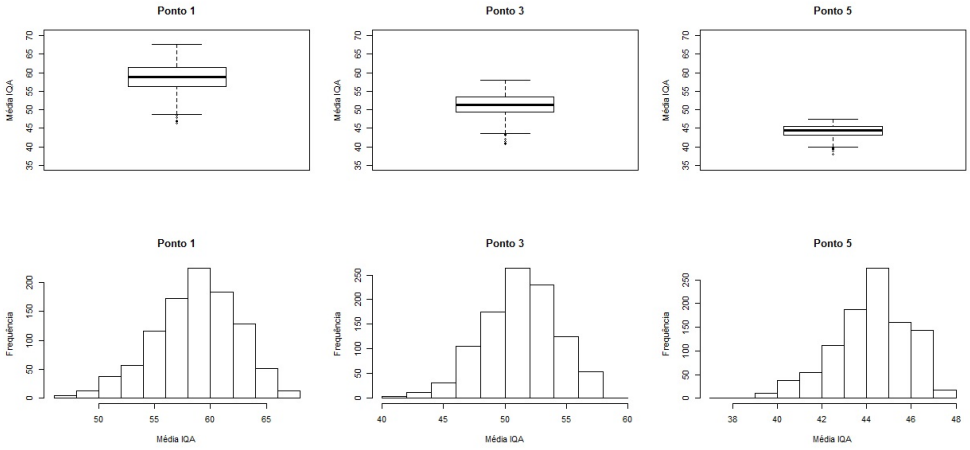


Figura 41: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis - Segundo quinquênio.

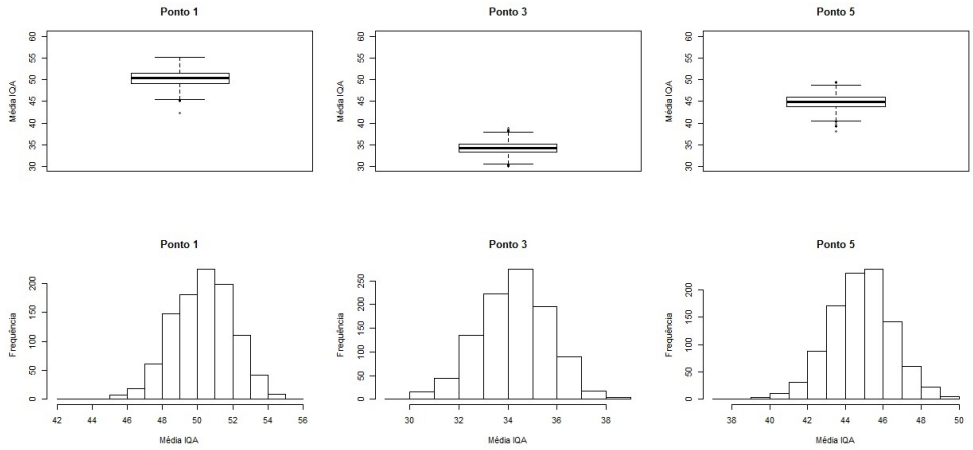


Figura 42: Histogramas e Boxplots das estimativas Bootstrap em blocos móveis - Terceiro quinquênio.

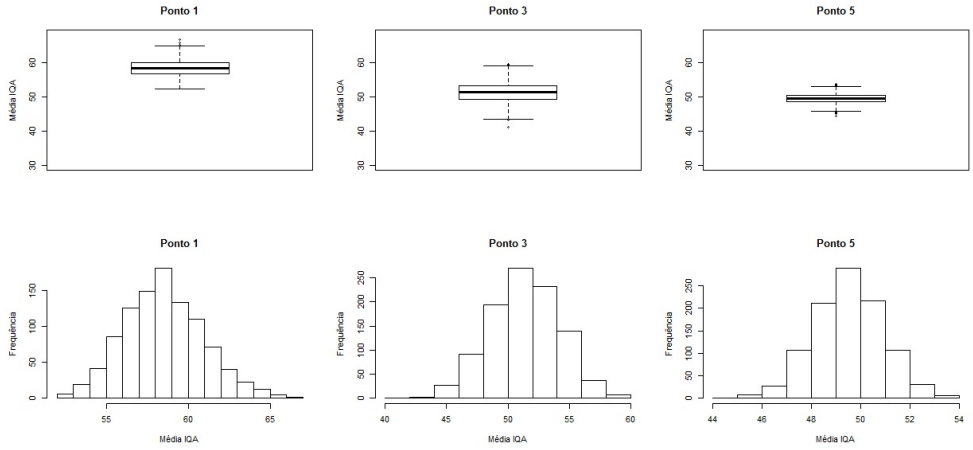


Figura 43: Histogramas e Boxplots das estimativas Bootstrap estacionário - Primeiro quinquênio

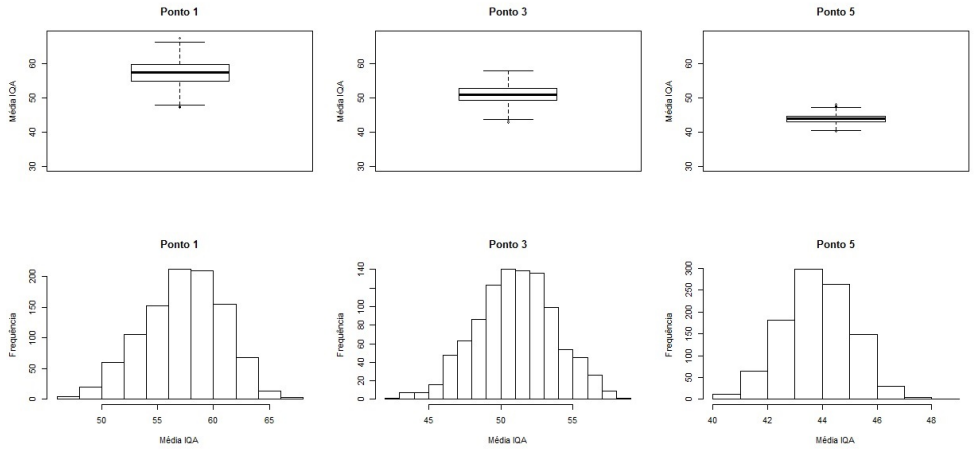


Figura 44: Histogramas e Boxplots das estimativas Bootstrap estacionário - Segundo quinquênio

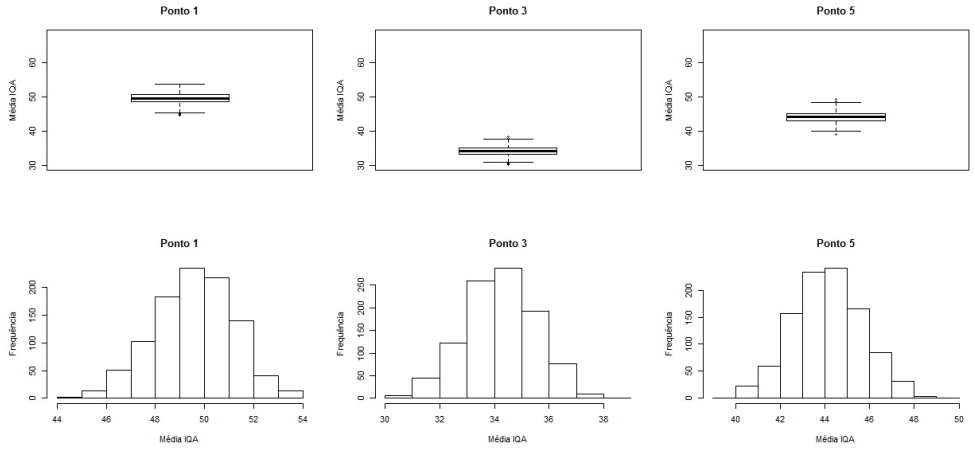


Figura 45: Histogramas e Boxplots das estimativas Bootstrap estacionário - Terceiro quinquênio

Com tudo isso, a análise por meio do Bootstrap em blocos perde em sua significância, não sendo muito adequada a sua utilização. Sendo assim, primeiramente será realizada uma análise descritiva dos dados, sem a realização de reamostra, a fim de verificar como os valores de IQA se comportam a cada quinquênio, para cada uma das cinco imputações realizadas.

Os problemas para a realização de técnicas como a ANOVA persis-

tem, também sendo essa inadequada considerando-se todos os dados do estudo. Agora existe normalidade, porém, persiste o problema de falta de independência dos resíduos.

Analisando os valores de IQA ao longo dos quinquênios, temos os seguintes resultados, apresentados nas Figuras 46 – 50.

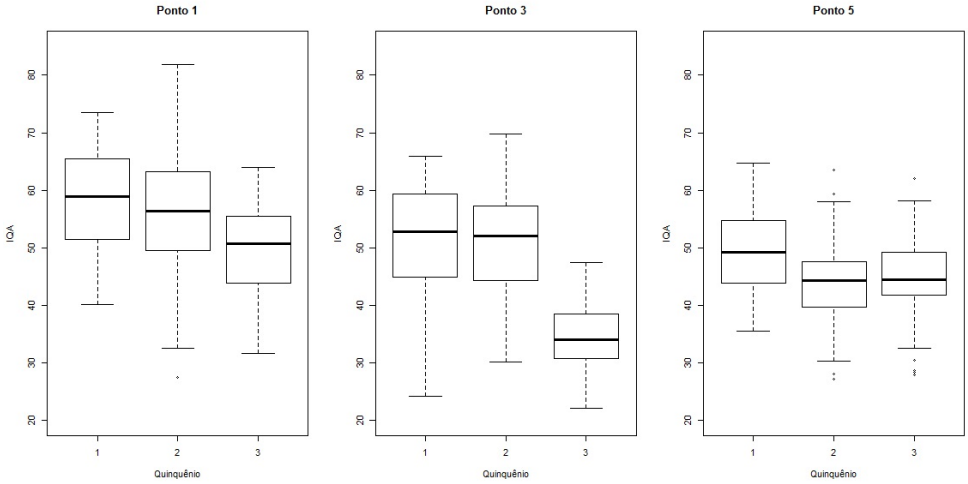


Figura 46: Boxplots dos IQAs por ponto para os quinquênios - Primeira Imputação

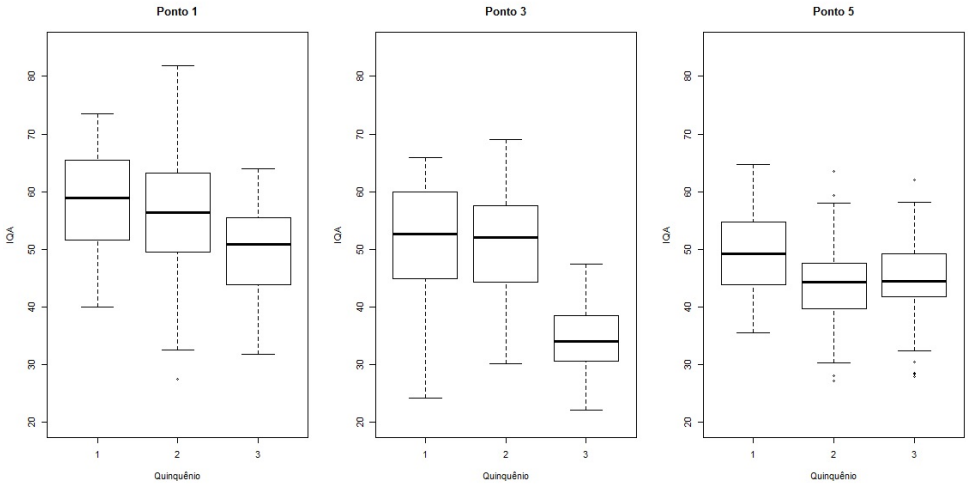


Figura 47: Boxplots dos IQAs por ponto para os quinquênios - Segunda Imputação

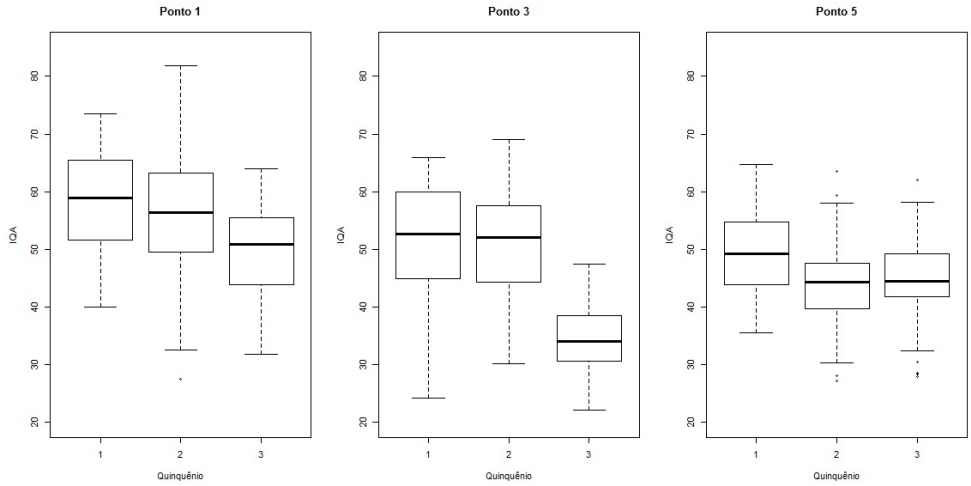


Figura 48: Boxplots dos IQAs por ponto para os quinquênios - Terceira Imputação

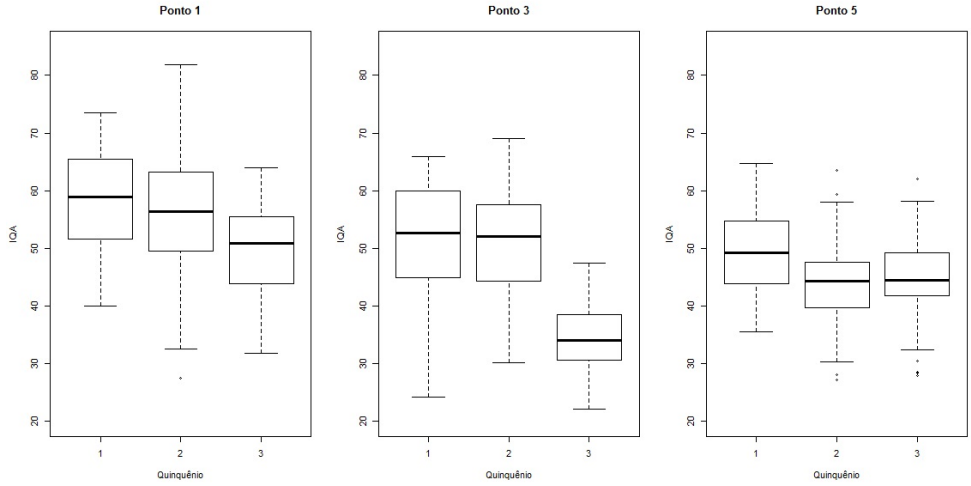


Figura 49: Boxplots dos IQAs por ponto para os quinquênios - Quarta Imputação

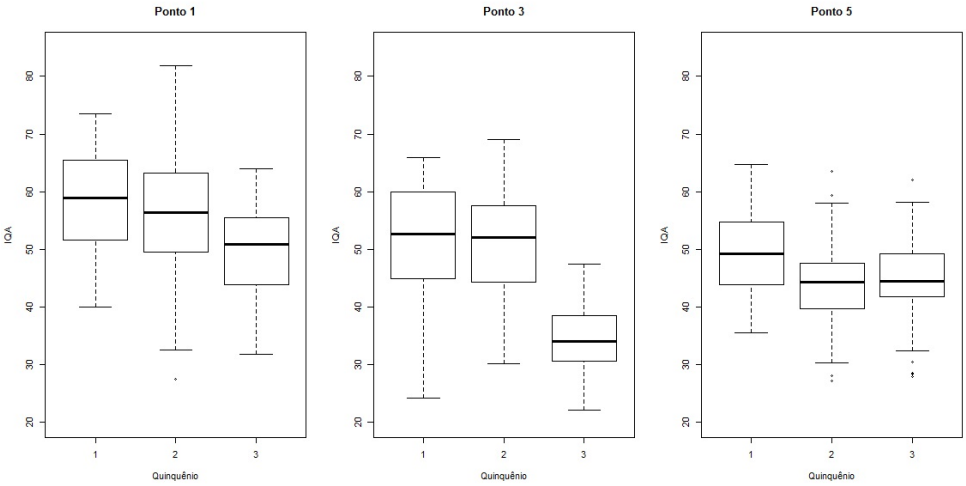


Figura 50: Boxplots dos IQAs por ponto para os quinquênios - Quinta Imputação

Para todas as imputações os resultados são bastante semelhantes. Há alguns indícios, pela análise descritiva, de que principalmente para o Ponto 3 (lagoa de estabilização) os valores de IQA sejam diferentes entre os quinquênios. O valor no último quinquênio apresenta os piores resultados, o que é um tanto quanto preocupante, uma vez que, com os avanços tecnológicos, espera-se que o tratamento dos resíduos melhore cada vez mais a qualidade da água a ser devolvida ao rio.

Apesar disso, o Ponto 5, referente ao rio Atibaia após a devolução dos resíduos, não parece estar sendo tão afetado por essa redução no IQA no Ponto 3, apresentando até mesmo uma leve melhora no último quinquênio. Isso pode ocorrer devido ao tratamento de outros resíduos despejados no rio, como despejos de esgoto urbano, possivelmente ter melhorado.

Já para o Ponto 1, nota-se uma leve redução na qualidade das águas do rio Jaguarí para o último quinquênio.

A fim de ser possível a realização de um teste formal, foi utilizada uma mudança na variável resposta, a fim de garantir a independência dos dados, assim permitindo o uso da ANOVA. Para tal foram utilizadas médias de um bloco de meses, a fim de quebrar a dependência existente entre os mesmos. Com isso há perca de grande parte da variabilidade dos dados, sendo esse um grande ônus de utilizar-

se os dados dessa maneira, porém, as suposições de normalidade, independência e homocedasticidade são todas elas verificadas.

Para que haja a menor perca possível na variabilidade dos dados, foram testadas todos os possíveis tamanhos de cada bloco que compõe as médias. Assim, iniciando-se com as médias de dois meses consecutivos, foram verificadas as suposições da ANOVA, a fim de, com o menor número de meses compondo possível a compor cada bloco, realizar a análise.

Para o Ponto 1 a suposição de independência dos resíduos (pelo teste de Durbin-Watson a 5% de significância) é verificada somente a partir do uso de médias de 7 meses consecutivos. Com isso, é possível a utilização de 5 valores para o primeiro quinquênio, 5 para o segundo e 8 para o terceiro. A normalidade (teste de Shapiro-Wilk) e a homocedasticidade (teste de Breusch-Pagan) também são verificadas. O mesmo vale para o Ponto 3.

Para o Ponto 5 a suposição de independência é verificada a partir do uso de médias de 6 meses, a 5% de significância. Com isso, é possível utilizar 6 valores para o primeiro quinquênio, 6 para o segundo e 9 para o terceiro.

Realizando a ANOVA com os dados de tais médias como variáveis respostas as suposições são verificadas e obtêm-se os seguintes resultados, apresentados nas Tabelas $37 - 39$.

Tabela 37: ANOVA entre quinquênios para o IQA no Ponto 1 - Primeira Imputação

Fonte de Variação	G.L.	Soma de Quadrados	Quadrado Médio	Valor F	Valor-p
Quinquênios	2	378.56	189.279	4.919	0.0228
Resíduos	15	577.19	38.479		

Tabela 38: ANOVA entre quinquênios para o IQA no Ponto 3 - Primeira Imputação

Fonte de Variação	G.L.	Soma de Quadrados	Quadrado Médio	Valor F	Valor-p
Quinquênios	2	1316.04	658.02	14.888	0.0003
Resíduos	15	662.96	44.20		

Tabela 39: ANOVA entre quinquênios para o IQA no Ponto 5 - Primeira Imputação

Fonte de Variação	G.L.	Soma de Quadrados	Quadrado Médio	Valor F	Valor-p
Quinquênios	2	104.22	52.108	3.1081	0.0693
Resíduos	18	301.77	16.765		

Os resultados finais para todas as imputações são os mesmos, havendo apenas mudanças bastante pequenas nas somas de quadrados, as quais não afetam a conclusão final, bem como ocorrera na análise anterior, referente à comparação entre os pontos ao longo de todo o período, pelo Bootstrap em blocos. Por esse motivo, não serão apresentadas as demais tabelas da ANOVA, para as outras imputações.

Nota-se que, a 5% de significância, há a rejeição de igualdade de médias de IQA ao longo dos quinquênios para o Ponto 1 e para o Ponto 3. Sendo assim, é interessante a verificação de qual dos quinquênios apresenta tal diferença, para o que pode ser utilizado o teste de Tukey.

Através do teste de Tukey, nota-se que a diferença significativa, a 5% de significância, para o Ponto 1 ocorre entre o terceiro e o segundo quinquênio. A média dos IQAs para o terceiro quinquênio é menor que a para o segundo. Entre o terceiro quinquênio e o primeiro existe diferença significativa apenas a 10% de significância.

Já para o Ponto 3, também a 5% de significância, há diferença significativa entre o terceiro quinquênio e os dois outros, confirmando o que era visível

através das Figuras $46 - 50$. No terceiro quinquênio o IQA é inferior aos dos demais quinquênios.

6 DISCUSSÃO

Quanto à imputação de dados, nota-se que o método proposto por Bergamo apresenta bons resultados se comparado ao método simples da média em algumas das variáveis, na primeira maneira de entrada de dados, por variáveis, porém, em outras ele é bastante ineficiente. Os dados apresentam-se de fato como de difícil imputação, uma vez que para muitas das variáveis não há normalidade e também devido ao padrão de ausência ser de tal maneira que em alguns tempos há ausências para todas as variáveis, dificultando assim o uso de algo observado especificamente naquele tempo para a previsão das ausências e ocasionando repetições de valores nas linhas que isso ocorre.

Devido ao fato de as imputações para casos nos quais há ausência da variável, para os três pontos, ao mesmo tempo utilizar os mesmos valores (ou seja, as demais linhas sem ausência), o método de Bergamo acaba resultando em valores repetidos para tais casos, bem como o da média o faz. A maior diferença ocorre naquelas datas para as quais são observados dados em duas variáveis e apenas uma é ausente, porém, esse tipo de ausência representa uma baixa quantidade dentre as ausências existentes na matriz de dados.

Já se levarmos em conta a entrada de dados por pontos, o método de Bergamo mostra-se eficiente para a avaliação final do IQA. Ainda persiste o problema de só ser possível realizar uma imputação consideravelmente boa pra casos nos quais seja observado ao menos algum dos subíndices do IQA, sendo inviável imputar dados para quando há ausência na coleta de dados. Além disso, caso o intuito fosse a avaliação dos subíndices do IQA, a imputação não apresentaria resultados tão bons.

A parte da análise estatística da diferença da qualidade das águas entre

os pontos selecionados apresentou dificuldades quanto à utilização dos métodos mais usuais. A metodologia de Bootstrap em blocos, principalmente com o Bootstrap estacionário, mostrou-se uma técnica bastante útil nesse caso, bem como o é em muitos outros. Através dele é possível realizar análises tanto para casos nos quais suposições usuais de distribuição de probabilidade não são atingidas quanto para problemas com relação à dependência entre os dados, muito comum em casos de análises de séries temporais.

Vale ressaltar que, mesmo que corrija-se parcialmente o problema de dependência, o método ainda pressupõe que os blocos sejam independentes entre si, o que não é completamente verdadeiro para nenhum caso. Porém, mesmo assim é uma das mais eficientes opções existentes para análises de dados dependentes. Além disso, quando há ausências de dados, como no caso desse estudo, alguns blocos conterão valores que não provém de meses sequenciais dentro dos mesmos. Essas nuances fazem parte do que ainda pode ser melhor desenvolvido na técnica, porém, mesmo com elas, ela possui grande valor na melhoria da análise dos dados com relação a outros métodos.

Ainda é interessante ressaltar que, ao dividir os dados em quinquênios, o que diminuiu bastante o número de observações a serem analisadas, levando-se em conta cada quinquênio de maneira separada, as técnicas de Bootstrap em blocos apresentaram problemas em sua aplicação.

Quanto aos resultados obtidos na etapa da análise estatística, é notável a diferença na qualidade das águas dos rios em estudo. O Rio Atibaia apresenta índices de IQA bastante preocupantes e inferiores aos do Rio Jaguarí, estando os dois em regiões próximas, nas quais o efeito natural é bastante semelhante. Há fortes evidências de que os despejos, tanto domésticos quanto industriais, estão sendo bastante nocivos ao rio e a qualidade do tratamento da água utilizada em tais processos deve ser continuamente estudada e melhorada.

Outro resultado interessante e preocupante é o referente à diferença no IQA entre os quinquênios analisados. Para 2010 até 2014 os resultados na lagoa de

estabilização apresentaram uma queda significativa no valor do índice. Além disso o Rio Jaguarí também teve queda de IQA no último período analisado. Apesar disso, no último quiquênio, mesmo havendo queda na qualidade da água na lagoa de estabilização, não houve o mesmo padrão para o rio Atibaia.

7 CONCLUSÃO

A análise estatística é de suma importância na avaliação das mais diversas realidades, assim sendo uma ferramenta utilíssima para a tomada de decisões com base em conclusões obtidas por meio de dados disponíveis para tal análise.

Mesmo quando os métodos usuais não funcionam adequadamente, existe uma gama de possibilidades para que seja possível realizar inferência sobre uma amostra. Tais métodos alternativos tem sido cada vez mais desenvolvidos, o que é possibilitado, principalmente, pelo avanço computacional. Tais avanços tem grande importância, uma vez que, ao lidar com diferentes tipos de dados, por vezes surgem problemas não solucionáveis por métodos mais simples, usualmente utilizados.

Nesse estudo, foi de grande valia o uso de técnicas que exigem implementação computacional, possibilitando assim que sejam obtidos resultados importantes e mais confiáveis que aqueles que seriam obtidos através de métodos usuais, os quais apresentam falhas nas suposições necessárias para que seu funcionamento adequado seja garantido.

O método de imputação múltipla aplicado mostrou-se útil para a obtenção de estimativas do IQA, para meses nos quais alguma das variáveis que o compõem não constava no banco de dados original. Com isso, pôde-se obter resultados tanto referentes ao desempenho da imputação e do Bootstrap em blocos, quanto que dizem respeito à situação na qual se encontram os pontos analisados, no que tange à qualidade da água.

O Bootstrap em blocos mostrou-se útil na análise dos dados ao longo dos 15 anos de estudo, sendo que o Bootstrap estacionário obteve resultados melhores

que o Bootstrap em blocos móveis, principalmente em relação ao viés. Porém, para períodos menores de tempo, com menos observações, a metodologia do Bootstrap em blocos apresentou problemas em sua aplicação.

Na parte prática, referente à qualidade da água em si, pode-se notar que de fato é necessário o esforço conjunto advindo tanto da parte de empresas e governo quanto da população em geral, uma vez que a qualidade da água, principalmente no rio Atibaia, apresentou resultados preocupantes.

Quanto aos resultados obtidos, podemos ainda concluir que é necessária a contínua vigilância e estudo sobre temas relacionados ao ecossistema em geral, inclusa em tais temas a qualidade das águas dos sistemas de água doce, pois esse é um bem não somente importante, mas sim imprescindível ao ser humano para sua subsistência.

REFERÊNCIAS BIBLIOGRÁFICAS

APHA. **Standard Methods of Examination of Water and Wastewater**. 20. ed. Washington DC, USA: American Public Health Association, 2005.

ARCINIEGAS-ALARCÓN, S. Imputação de dados em experimentos com interação genótipo por ambiente: Uma aplicação a dados de algodão. Piracicaba, 2008. Dissertação (mestre em ciências com ênfase em estatística e experimentação agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C. T. S. Imputação de dados em experimentos com interação genótipoxambiente. **Asociación Interciencia Caracas**, v.36, n.6, p.444–449, 2011.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W.; DIAS, C. T. S. Deterministic Imputation in Multienvironment Trials. **ISRN Agronomy**, 2013.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W.; DIAS, C. T. S. Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison. **Communications in Biometry and Crop Science**, v.9, n.2, p.54–70, 2014.

AYUYEV, V. V.; JUPIN, J.; HARRIS, P. W.; OBRADOVIC, Z. Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data. **Data warehousing and knowledge discovery. Lecture notes in computer science**, v.5691, p.366–377, 2009.

BERGAMO, G. C. Imputação multipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação. Piracicaba, 2007. Tese (doutorado em estatística e experimentação agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

BERGAMO, G. C.; DIAS, C. T. S.; KRZANOWSKI, W. Distribution-free multiple imputation in an interaction matrix though singular value decomposition. **Scientia Agricola**, v.65, n.4, p.422–427, 2008.

BHARGAVA, A. On the Theory of Testing for Unit Roots in Observed Time Series. **The Review of Economic Studies**, v.53, n.3, p.369–384, 1986.

BRENNING, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package 'sperrorest', 2012.

BROWN, R. M.; MCCLELLAND, N. I.; DEININGER, R. A.; TOZER, R. A. Water Quality Index – Do We Dare? **Water Sewage Works**, p.339–343, 1970.

CAMERON, A. C.; GELBACH, J. B.; MILLER, D. L. Bootstrap-based improvements for inference with clustered errors. **Review of Economics and Statistics**, v.90, p.414–427, 2008.

CANTY, A.; RIPLEY, B. D. **boot: Bootstrap R (S-Plus) Functions**, 2015. R package version 1.3-15.

CARLSTEIN, E. The Use of Subseries Methods for Estimating the Variance of a General Statistic from a Stationary Time Series. **The Annals of Statistics**, v.14, p.1171–1179, 1986.

CBH-PCJ; COBRAPE. Plano de Bacias Hidrográficas dos Rios Piracicaba, Capivari e Jundiaí para o Quadriênio 2008-2011,. Rel. téc., 2008. Pág 9.

CETESB. Relatório de Qualidade das Águas Superficiais no Estado de São Paulo. Apêndice C: Índices de Qualidade Das águas. Rel. téc., CETESB, 2013.

CHING, W.; LI, L.; TSING, N.; TAI, C.; NG, T. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. **International Journal of Data Mining and Bioinformatics**, v.4, p.331–347, 2010.

CLESCERI, L. S.; GRENBERG, A. E.; EATON, A. D. **Standard Methods of the Examination of Water and Wastewater**. 20. ed. American Public Health Association (APHA), American Water Works Association (AWWA), Water Environmental Federation (WEF),, 1998. Cap. 9, Seção 9215.

COMITÊS-PCJ. Relatório da Situação dos Recursos Hídricos 2014. Rel. téc., 2014.

COMITÊS-PCJ. Plano de Bacias, 2015a. Disponível em: http://www.agenciapcj.org.br/novo/instrumentos-de-gestao/plano-de-bacias , consulta em 01 dez 2015.

COMITÊS-PCJ. Relatório de Situação, 2015b. Disponível em: http://www.agenciapcj.org.br/novo/instrumentos-de-gestao/relatorios-de-situacoes, consulta em 01 dez 2015.

COMITÊS-PCJ. Retrospectiva cronológica consórcio PCJ– de 1989 até os dias atuais, 2015c. Disponível em: http://agua.org.br/historico/ , consulta em 01 dez 2015.

CONNOVER, W. J. **Practical Nonparametric Statistics**. New York: Wiley, 1980. 493p.

CPRH Índice e Indicadores de Qualidade da Água – Revisão da Literatura. In: , 2013a. **PNMA II - índices e indicadores**; resumos. , 2013a.

CPRH Seleção de Índices e Indicadores de Qualidade da Água Aplicação dos Índices Selecionados. In: , 2013b. **PNMA II - índices e indicadores**; resumos. , 2013b.

DAEE; IG; IPT; CPRM. **Mapa de águas subterrâneas do Estado de São Paulo : escala 1:1.000.000 : nota explicativa**, 2005.

DAVISON, A. C.; HALL, P. On Studentizing and BloBlock Methods for Implementing the Bootstrap with Dependent Data. **Australian Journal of Statistics**, v.35, p.215–224, 1993.

DE LIVERA, A.; HYNDMAN, R.; SNYDER, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. **Journal of the American Statistical Association**, v.106, n.496, p.1513–1527, 2011.

DERISIO, J. C. **Introdução ao controle da poluição ambiental**. 3. ed. São Paulo: Signus Editora, 2007.

EFRON, B. Bootstrap methods: another look at the jackknife. **Annals of Statistics**, v.7, n.1, p.1–26, 1979.

EFRON, B.; TIBSHIRANI, R. J. **An Introdution to the Bootstrap**. Chapman & Hall., 1993.

ENDERS, C. K. A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research. **Psychosomatic Medicine**, v.68, p.427–436, 2006.

ESTADO DE SÃO PAULO. Decreto Nº 10.755, de 22 de Novembro de 1977, 1977. Decreto da Assembléia Legislativa do Estado de São Paulo.

ESTADO DE SÃO PAULO. Lei Nº 7663 de 30 de Dezembro de 1991 de São Paulo, 1991.

ESTADO DE SÃO PAULO. Constituição do estado de São Paulo, 2014. Título VI, Capítulo IV, Seção II, Artigo 215.

GAUCH JR, H. G. **Statistical analysis of regional yelds trials: AMMI analysis of factorial designs**. Elsevier, 1992.

GOOD, I. Some applications of the singular value decomposition of a matrix. **Technometrics**, v.11, p.823–831, 1969.

GRUNITZKI, R.; FERRARI, J. C.; SILVA, A. C.; ZAMBÃO, P. H.; NECKEL, E. V. P. Ferramenta WEB para Determinação do Indice de Qualidade de Agua a Partir da Reestruturação das Equações que descrevem as curvas dos Indicadores de Qualidade, 2013.

HALL, P.; HOROWITZ, J. L. Bootstrap critical values for tests based on generalized-method-of-moments estimators. **Econometrica**, v.64, n.4, p.891–916, 1996.

HAYFIELD, T.; RACINE, J. S. Nonparametric Econometrics: The np Package. **Journal of Statistical Software**, v.27, n.5, 2008.

HYNDMAN, R. J. **forecast: Forecasting functions for time series and linear models**. R package version 6.1, 2015. URL: http://github.com/robjhyndman/forecast.

INOUE, A.; SHINTANI, M. Bootstraping GMM estimators for Time Series. **Journal of Econometrics**, v.133, n.2, p.531–555, 2003.

JAECKEL, L. The infinitesimal jackknife. **Memorandum**, 1972.

JELICIC, H.; PHELPS, E.; LERNER, R. M. Use of Missing Data Methods in Longitudinal Studies: The Persistence of Bad Practices in Developmental Psychology. **Developmental Psychology**, v.45, n.4, p.1195–1199, 2009.

JUNGER, W. L. Imputação de dados faltando em séries temporais multivariadas via algoritmo EM, 2002. Monografia apresentada ao departamento de estatística - Universidade do Estado do Rio de Janeiro.

KÜNSCH, H. The jackknife and the bootstrap for general stationary observations. **Ann.Statist**, v.17, p.1217–1241, 1989.

KRZANOWSKI, W. J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. **Biometrical Letters**, v.25, n.1-2, p.31–39, 1988.

LANDWEHR, J. M.; DEININGER, R. A. A Comparison of Several Water Quality Indices. **J. Water Poll. Control Fed.**, v.48, n.5, p.954 – 958, 1976.

LEISCH, F.; TIBSHIRANI, R. **bootstrap: Functions for the Book "An Introduction to the Bootstrap"**, 2015. R package version 2015.2.

LEMOS, D. S.; LEMOS, T. R. M. S. Aspectos jurıdicos da sustentabilidade da agua. **Âmbito Jurıdico**, , n.Nº 71, 2009.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. Wiley-Interscience, 2002. ISBN 0-471-18386-5.

LIU, R. Y.; SINGH, K. Moving Blocks Jackknife and Bootstrap Capture Weak Dependence. In: WILEY, J. (Ed.). **Exploring the Limits of Bootstrap**. New York: John Wiley, 1992. p.225-248.

LOBO, R. S.; KALIL, R. C. E. Um estudo de caso do impacto da imputação de dados na mineração de dados. Universidade Federal do Pará, Belém., 2006. Trabalho de conclusão de curso (bacharel em ciência da computação) - .

MARITZ, J. S. **Distribution-Free Statistical Methods**. Chapman & Hall, 1981. ISBN: 0-412-15940-6.

MEDINA, F.; GALVÁN, M. Estudios Estadısticos y prospectivos. Imputacion de datos: Teorıa y Pratica. **Division de Estadıstica y Proyecciones Economicas**, 2007.

MOLENBERGHS, G.; FITZMAURICE, G.; KENWARD, M. G.; TSIATIS, A.; VERBEKE, G. **Handbook of Missing Data Methodology**. Chapman and Hall/-CRC, 2014.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 3. ed. New York: Wiley, 2001.

NELSON, C. R.; PLOSSER, C. I. Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications. **Journal of Monetary Economics**, , n.10, p.139–162, 1982.

NORDMAN, D. J.; LAHIRI, S. N. Optimal Block Size for Variance Estimation by a Spatial Block Bootstrap Method. **The Indian Journal of Statistics Special Issue on Statistics in Biology and Health Sciences**, v.69, p.468–493, 2007.

NUNES, L. N. Métodos de imputação de dados aplicados na area da saude, 2007. Tese de doutorado - Universidade Federal do Rio Grande do Sul.

PATTON, A.; POLITIS, D. N.; WHITE, H. CORRECTION TO "Automatic block-length selection for the dependent bootstrap"by D. Politis and H. White. **Econometric Reviews**, v.28, n.4, p.372–375, 2009.

PENNY, K. I.; JOLLIFFE, I. T. Multivariate outlier detection applied to multiply imputed laboratory data. **Statistics in Medicine**, v.18, p.1879–1895, 1999.

PERRY, P. Cross-validation for unsupervised learning, 2009. Tese (doutorado em estatística) - Stanford University.

PHILLIPS, P. C. B.; PERRON, P. Testing for a Unit Root in Time Series Regression. **Biometrika**, v.75, n.2, p.335–346, 1988.

POLITIS, D.; ROMANO, J. P. A Circular Block-Resampling Procedure for Stationary Data.. In: **Exploring the Limits of Bootstrap**. New York: John Wiley, 1992. p.263-270.

POLITIS, D. N. The Impact of Bootstrap Methods on Time Series Analysis. **Statist. Sci**, v.2, p.219–230, 2003.

POLITIS, D. N.; ROMANO, J. P. The Stationary Bootstrap. **Journal of the American Statistical Association**, v.89, n.2, p.1303–1313, 1994.

POLITIS, D. N.; WHITE, H. Automatic Block-Length Selection for the Dependent Bootstrap. **Econometric Reviews**, v.23, n.1, p.53–70, 2004.

PROJETO ANHUMAS. Recuperação ambiental, participação e poder público: uma experiência em Campinas. Rel. téc., 2006.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2015.

RENNO, C. D. Jacknife, Bootstrap e Afins, 2011. INPE.

RINO, C. A. F.; SAGGIORO, N. J.; HERCULIANI, L. A. Avaliação da qualidade das águas do rio bauru – Determinação do IQA, 2001.

RUBIN, D. B. Inference and Missing Data. **Biometrika**, v.63, n.3, p.581–592, 1976.

RUBIN, D. B. Multiple imputation in sample surveys: a phenomenological Bayesian approach to nonresponse. **Proceedings of the Survey Research Methods Section**, p.20–28, 1978.

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

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

SAS INSTITUTE INC. **SAS/STAT® 9.3 User's guide**. Cary, NC: SAS Institute, 2011.

SCHAFER, J. L.; GRAHAM, J. W. Missing Data: Our View of the State of the Art. **Psychological Methods**, v.7, n.2, p.147–177, 2002.

SCHEFFÉ, H. **The Analysis of Variance**. Wiley-Interscience, 1999.

SIDAK, Z. K. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. **Journal of the American Statistical Association**, v.62, n.318, p.626–633, 1967.

SILVA, D. N. O método bootstrap e aplicações á regressão multipla. Campinas, 1995. Dissertação (mestre em estatística) - UNICAMP.

SILVA, M. J. C. Imputação múltipla: uma comparação e eficiência em experimentos multiambientais. Piracicaba, 2012. Dissertação (mestre em ciências com ênfase em estatística e experimentação agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

SINGH, K. On the asymptotical accuracy of Efron's bootstrap. **Ann. Statist**, v.9, p.1187–1195, 1981.

SYNOWIECKI, R. Consistency and application of moving block bootstrap for non-stationary time series with periodic and almost periodic structure. **Bernoulli**, v.13, n.4, p.1151–1178, 2007.

TRAPLETTI, A.; HORNKI, K. **tseries: Time Series Analysis and Computational Finance**, 2015. R package version 0.10-34.

TUKEY, J. Bias and confidence in not-quite large samples. In: Tukey (1958), 1958.

VIEIRA, R. S. **Crescimento econômico no estado de São Paulo : uma analise espacial**. São Paulo: Cultura Acadêmica, 2009.

VOGEL, R. M.; SHALCROSS, A. L. The moving blocks bootstrap versus parametric time series models. **Water Resources Research**, v.32, n.6, p.1875–1882, 1996.