UMA ANÁLISE DAS ÁGUAS DOS RIOS JAGUARI E ATIBAIA COM USO DE IMPUTAÇÃO MÚLTIPLA LIVRE DE DISTRIBUIÇÃO E BOOTSTRAP EM BLOCOS

Eduardo Schneider Bueno de OLIVEIRA[1]
Antonio Carlos Simões PIÃO[2]
Dejanira de Franceschi de ANGELIS[3]

- RESUMO: Dados longitudinais tem grande relevância em análises para avaliar diversos fatores. Por vezes, há também o problema de ausência de dados coletados em algum momento durante o estudo, o que leva à necessidade de, ou desconsiderá-los, trabalhando assim com uma amostra de menor tamanho, ou utilizar técnicas adequadas para estimar tais dados ausentes. No presente estudo, com base em dados sobre a qualidade da água nos rios Jaguari e Atibaia, são apresentados resultados de imputação múltipla de dados para dados não normais, através da "Imputação Múltipla Livre de Distribuição" e, posteriormente, resultados para a análise dos dados através de técnicas de Bootstrap em blocos. Por fim, são apresentados os resultados práticos obtidos através da análise realizada, a fim de identificar resumidamente a situação de ambos os rios ao longo dos 15 anos de estudo.
- PALAVRAS-CHAVE: Imputação múltipla; bootstrap em blocos; índice de qualidade da água.

1 Introdução

Em diversas situações dados são coletados ao longo do tempo. Havendo, em geral, dependência entre dados longitudinais, a análise destes exige certos cuidados.

[1]Universidade Estadual Paulista "Júlio de Mesquita Filho" - UNESP, Instituto de Biociências, Departamento de Bioestatística, CEP: 18618-970, Botucatu, São Paulo, Brasil. E-mail: *eduardosb@ibb.unesp.br; eduardosbdeoliveira@hotmail.com*
[2]Universidade Estadual Paulista "Júlio de Mesquita Filho" - UNESP, Instituto de Geociências e Ciências Exatas, Departamento de Estatística, Matemática Aplicada e Computação, CEP: 13506-900, Rio Claro, São Paulo, Brasil. E-mail: *piao@rc.unesp.br*
[3]Universidade Estadual Paulista "Júlio de Mesquita Filho" - UNESP, Instituto de Biociências, Departamento de Bioquímica e Microbiologia, CEP: 13506-900, Rio Claro, São Paulo, Brasil. E-mail: *dangelis@rc.unesp.br*

Existem diversas técnicas, cada uma com suas limitações e suposições necessárias para a adequação da técnica de análise a ser utilizada. Uma das técnicas existentes é a de Bootstrap em blocos, cujos diferentes desdobramentos resultam em métodos de utilização da mesma com algumas variações.

Um possível problema na coleta longitudinal de dados é o da ausência de alguns valores. Em um estudo longitudinal ao longo de vários anos, com observações mensais, nem sempre em todos os meses há coleta de todas as variáveis de interesse. Para resolver este problema, pode-se optar pela exclusão dos meses nos quais há parcial ou completa ausência de dados. Outra abordagem possível é a da Imputação de Dados.

No presente artigo, visa-se avaliar o desempenho do método denominado "Imputação Múltipla Livre de Distribuição" (BERGAMO, 2007, BERGAMO *et al.*, 2008) para dados coletados mensalmente nos rios Jaguari e Atibaia, ao longo de quinze anos, os quais apresentam alguns problemas de ausência de dados ao longo do período, além de ter problemas de normalidade nas variáveis a serem imputadas.

Além disso, após realizada a imputação, pretende-se avaliar o desempenho de dois diferentes métodos de aplicação do Bootstrap em blocos, a saber, a do Bootstrap em blocos móveis (KÜNSCH, 1989) e a do Bootstrap Estacionário (POLITIS e ROMANO, 1994), avaliando seu desempenho e também obtendo resultados de suas aplicações.

Para a obtenção de resultados práticos, analisou-se o Índice de Qualidade da Água (IQA) (COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO, 2013) nos dois rios. A imputação foi realizada com base nos subíndices que compõem o IQA, levando-se em conta os dados para cada um dos rios analisados a cada imputação. Os meses para os quais não houve coleta de nenhuma variável que compõe o IQA não foram utilizados no estudo, uma vez que a imputação para os mesmos não teve bons resultados.

2 Objetivos

- Avaliar o desempenho da técnica de "Imputação Múltipla Livre de Distribuição" na imputação dos dados do estudo;
- Avaliar o desempenho do Bootstrap em blocos móveis e do Bootstrap estacionário na análise dos dados;
- Obter informações sobre a qualidade da água dos rios Jaguari e Atibaia ao longo dos últimos 15 anos.

3 Referencial teórico

3.1 O IQA

O IQA é utilizado desde 1975 pela Companhia Ambiental do Estado de São Paulo (CETESB), conforme pode ser consultado em COMPANHIA AMBIENTAL

DO ESTADO DE SÃO PAULO (2013). As informações obtidas por meio 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, que combinam diferentes medidas em um único valor, são medidas relevantes. Além disso, a facilidade de comunicação com o público é maior através de índices como o IQA.

Originalmente, o IQA foi desenvolvido para águas americanas (BROWN *et al.*, 1970), sendo adaptado para águas brasileiras pela COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (2013), com algumas mudanças principalmente nos pesos atribuídos a cada subíndice. As curvas utilizadas para obter-se os subíndices componentes do IQA a partir dos valores brutos dos parâmetros equivalentes estão disponibilizadas em COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (2013), bem como os pesos de cada subíndice. Genericamente, a equação utilizada é a mesma, seguindo a Equação (1).

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

na qual IQA 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).

A classificação das águas, para o estado de São Paulo, é apresentada na Tabela 1.

Tabela 1 - 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	$19 < \text{IQA} \leq 36$
Péssima	$0 \leq \text{IQA} \leq 19$

Fonte: Companhia Ambiental do Estado de São Paulo (2013).

3.2 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. Tal situação ocorre no caso dos dados do presente estudo.

A imputação de dados é uma técnica que permite a substituição de dados faltantes por seus valores estimados, a partir das unidades observadas de um conjunto de dados. O objetivo do uso da imputação é estimar os valores ausentes, trazendo assim mais informações para a análise posterior dos dados.

Existe na literatura alguns padrões e mecanismos para os dados ausentes, definidos por Rubin (1976). Através desses padrões é possível se ter uma base de qual dos métodos existentes de imputação de dados é mais propício a ser utilizado 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 distribuição de probabilidades dos dados faltantes, visando desta maneira informar a causa da ausência, embora não ofereçam de fato uma explicação causal (ENDERS, 2006).

Entre as possíveis técnicas utilizadas para a imputação, existem as de Imputação Simples e as de Imputação Múltipla. Na imputação simples, cada valor para os dados ausentes é estimado uma única vez. Rubin (1976) faz uma crítica à utilização de tais métodos de imputação simples, uma vez que os mesmos apresentam aumento no viés das estimativas.

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).

3.2.1 Imputação múltipla livre de distribuição

Em diversos métodos de imputação múltipla, são feitas suposições sobre a distribuição de probabilidade dos dados. Para casos nos quais tais suposições não podem ser feitas, foi desenvolvido um método com estimativas precisas dos valores ausentes foi desenvolvido (BERGAMO, 2007; BERGAMO *et al.*, 2008).

Este método, baseado em ideias de decomposição em valores singulares (GOOD, 1969), as quais anteriormente já haviam sido utilizadas para a realização de imputação simples (KRZANOWSKI, 1988), realiza algumas alterações no método de imputação simples, permitindo a realização não mais de apenas uma, mas sim de M estimativas para o valor ausente, adicionando-se assim a presença de variabilidade nas imputações.

Segundo Rubin (1987) e Schafer e Graham (2002), um número M entre 3 e 5 imputações já é o suficiente para expressar a variabilidade existente entre as imputações.

3.3 Bootstrap em blocos

As técnicas de reamostragem consistem em, tomando uma amostra mestra de dados, realizar sucessivas reamostras com base nessa amostra original. O Bootstrap (EFRON, 1979), uma das técnicas de reamostragem existentes, é uma boa alternativa para a realização de análises que, em sua forma tradicional, exigem que diversas suposições de modelo sejam satisfeitas - o que muitas vezes não ocorre na prática. Por meio de realizações de reamostras, podem ser obtidas estatísticas de interesse, tanto pontuais quanto intervalares.

A ideia básica do Bootstrap consiste em, através dos dados disponíveis na amostra, reproduzir o mecanismo probabilístico que os gerou. Para casos com a distribuição de probabilidade desconhecida, é possível substituí-la por uma distribuição empírica que possa aproximá-la, utilizando assim o Bootstrap não paramétrico. No Bootstrap não paramétrico, a partir da amostra original, são realizadas B reamostras, com reposição, de mesmo tamanho da amostra original.

O número B de reamostras para a realização das análises estatísticas depende do intuito do estudo. Efron e 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.

Ainda há a possibilidade do cálculo de medidas de precisão do Bootstrap, bem como o viés e a variância entre as médias (${\sigma_B^2}^*$), conforme nas Equações (2) e (3).

$$\widehat{\text{Viés}}_B^* = \frac{\sum_{i=1}^{B} \theta^{*(i)}}{B} - \theta(\hat{F}_n), \tag{2}$$

$${\widehat{\sigma}_B^2}^* = \frac{1}{B-1} \sum_{i=1}^{B} (\theta^{*(i)} - \overline{\theta^*})^2. \tag{3}$$

na qual $\theta(\hat{F}_n)$ é a média estimada obtida por meio da amostra mestre original, $\theta^{*(i)}$ é a média da amostra gerada na realização do Bootstrap e $\overline{\theta^*}$ é a média das médias Bootstrap após B realizações.

Para casos nos quais há dependência entre os dados, mesmo o Bootstrap comum apresenta inadequações, uma vez que uma de suas únicas suposições é a de que os dados são independentes e identicamente distribuídos. Para casos assim, existe a alternativa do uso do Bootstrap em blocos, o qual tem grande valia em casos de análises de séries temporais e de cluster (POLITIS, 2003; CAMERON *et al.*, 2008). Se a dependência for neglicenciada, podem ocorrer erros (SINGH, 1981).

Inicialmente esse método foi proposto 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).

Para a realização do Bootstrap em blocos, parte-se da mesma ideia de reamostragem do Bootstrap, porém, agora baseando-se em reamostragens de blocos de dados de tamanho l, ao invés de utilizarem-se os dados individualmente. Um

algoritmo genérico para melhor entendimento da lógica de tal método é o seguinte (POLITIS e WHITE, 2004):

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 tem-se as estimativas para a média e a variância da média ao longo da realização das reamostragens Bootstrap e pode-se utilizá-las da mesma maneira que no Bootstrap usual para obter estimativas.

Entre as diferentes técnicas que se apropriam da ideia do Bootstrap em blocos estão a do Bootstrap em blocos móveis (KÜNSCH, 1989) e do Bootstrap estacionário (POLITIS e ROMANO, 1994).

A proposta de Künsch (1989) segue exatamente a formulação do Bootstrap em blocos, considerando-se sobreposição dos mesmos, com reamostragem com reposição. Para tal, o tamanho dos blocos é fixado, os blocos são montados de acordo com tal tamanho, havendo sobreposição, ou seja, o primeiro bloco parte do indivíduo 1, enquanto o segundo parte do indivíduo 2 e não do indivíduo seguinte ao término do primeiro bloco. Tais blocos são reamostrados, com reposição, seguindo a ideia do algoritmo exposto.

Já o Bootstrap estacionário apresenta algumas diferenças apenas quanto ao tamanho dos blocos. O tamanho de cada bloco, agora, não é considerado 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, tem-se que $E(L) = 1/p$ isso implica que uma vez estipulado o valor esperado para o tamanho (L) do bloco, 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 e Romano (1994).

Uma etapa importante é a da definição do tamanho e do número de blocos a serem utilizados. Na literatura existem diferentes sugestões referentes à escolha do número de blocos utilizados. Hall e Horowitz (1996) utilizam 5 blocos para uma amostra de tamanho 50 e 10 para uma de tamanho 100. Inoue e Shintani (2003) selecionam o tamanho dos blocos por meio de procedimentos automáticos, bem como Politis e White (2004) o fazem para o Bootstrap em blocos estacionário e o circular de blocos móveis. Em Patton, Politis e White (2009) são corrigidos alguns problemas de Politis e White (2004) e os métodos são aprimorados.

Os procedimentos automáticos propostos em Politis e White (2004) e Patton *et al.* (2009) baseiam-se na estrutura de dependência 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).

4 Metodologia

4.1 Os dados do estudo

Os dados utilizados no presente estudo são provenientes dos rios Jaguari (na coordenada 22°42'00.0"N 47°08'06.6"L) e Atibaia (22°44'25.3"N 47°071'35.2"L).

A análise foi realizada com base nos valores do Índice de Qualidade da Água (IQA). As variáveis que compõem o IQA foram coletadas ao longo dos meses entre 2000 e 2014 e as amostras foram analisadas quanto aos parâmetros químicos, físicos e biológicos segundo Standard Methods (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.

Verifica-se 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, ambos os rios 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 à ausências por não haver análise de determinada variável, ocorridas nos demais 134 meses, houveram situações diferentes nos pontos selecionados para o estudo, conforme apresentado na Tabela 2.

Tabela 2 - Proporção de Ausências para os Pontos

Variável	Proporção Rio Jaguari	Proporção Rio Atibaia
E. coli	0.00%	1.49%
pH	0.00%	1.49%
Nt	2.23%	3.73%
Pt	13.43%	11.19%
Turbidez	15.67%	17.16%
O.D.	2.23%	4.48%
DBO	0.00%	1.49%
S. Totais	2.23%	3.73%

4.2 Imputação de dados

No presente estudo, uma vez que os dados não apresentam normalidade, as imputações foram realizadas através da técnica de Imputação Múltipla Livre de Distribuição (BERGAMO, 2007; BERGAMO *et al.*, 2008), utilizando-se o *software* SAS (SAS, 2011), com base no código presente em Bergamo (2007). Foram imputados os valores para os meses nos quais houve coleta de água mas não houve a análise de alguma variável específica, o que impossibilitaria o cálculo do IQA, caso não fosse realizada a imputação ou utilizado algum método de cálculo que desconsidere aquela variável. Adotou-se a realização de 5 imputações.

A entrada de dados foi feita através de uma matriz do tipo:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} & y_{16} & y_{17} & y_{18} \\ y_{21} & y_{22} & y_{23} & y_{24} & y_{25} & y_{26} & y_{27} & y_{28} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & y_{n4} & y_{n5} & y_{n6} & y_{n7} & y_{n8} \end{bmatrix} \tag{4}$$

Com y_{ij} representando 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 rio, sendo as linhas os meses consecutivos.

Para a entrada de dados, seria possível tanto considerar os valores brutos de cada variável, imputar os valores ausentes, e com base neles calcular o subíndice q_i referente a tal valor, quanto calcular primeiramente os subíndices q_i das variáveis analisadas, com base nos valores brutos, e entrar com os mesmos na matriz de dados, tendo-se assim a imputação direta de subíndices para as variáveis ausentes, e não mais dos valores brutos das mesmas.

No presente estudo, cada casela da matriz de entrada de dados correspondeu ao subíndice q_i do rio em questão, em determinado mês, uma vez que, ao entrar-se com os dados brutos de cada variável, houve grandes diferenças na escala, que trouxeram problemas à imputação. Assim, são imputados diretamente os valores

dos subíndices q_i que compõem o IQA, bastando substituí-los na fórmula do índice, após a imputação, para assim calcular o valor de IQA para aquele mês. Com tais valores de IQA obtidos, é possível realizar análises quanto à qualidade da imputação com base no IQA e também outras análises estatísticas práticas, referentes à situação dos rios no período de estudo.

Para meses nos quais não há coleta de água, torna-se inviável a imputação por meio desse método, uma vez que, para todos os meses nos quais isso ocorre, os valores imputados seriam exatamente os mesmos, não levando em conta nenhuma informação específica daquele período na imputação.

4.2.1 Avaliação da qualidade de imputação

Antes de realizar a aplicação da técnica de imputação para os dados ausentes, foi realizado um estudo a fim de verificar a qualidade dos dados estimados pela imputação, em detrimento de outra maneira possível de se obter os resultados.

Com isso, baseando-se nos meses para os quais há os dados completos, em uma matriz na qual há os dados dos subíndices que compõem o IQA, foram retirados aleatoriamente algumas observações, seguindo-se o mesmo padrão de ausência existente na matriz a ser posteriormente imputada. Desta forma, pode-se inferir sobre as diferenças entre os valores obtidos pela imputação e os presentes na matriz original. Esse procedimento é semelhante ao realizado em outras análises da qualidade da imputação, bem como os presentes no próprio estudo de Bergamo (2007) e também em Arciniegas-Alarcón *et al.* (2014), entre outros.

Para a avaliação da qualidade de imputação, foram utilizados três estatísticas, a NRMSE (CHING *et al.*, 2010), o T_{acc} (BERGAMO, 2007) e o coeficiente de correlação de Spearman (SPEARMAN, 1904).

Tais estatísticas podem ser definidas como:

$$NRMSE = \frac{\sqrt{\text{média}(\mathbf{a}_{imp} - \mathbf{a}_{orig})^2}}{dp(\mathbf{a}_{orig})}, \tag{5}$$

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 de 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:

$$acc_l = \frac{\sum_{m=1}^{M}(\widehat{Y}_{l(m)} - VO_l)^2}{M-1}, \tag{6}$$

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:

$$T_{acc} = \frac{\sum_{l=1}^{na} acc_l}{na}. \tag{7}$$

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:

$$T_{acc} = V_E + VQM, \tag{8}$$

em que:

$$T_{acc}\ parte\ 1 = V_E = \frac{1}{na}\sum_{l=1}^{na}\frac{\sum_{m=1}^{M}(\widehat{Y}_{l(m)} - \overline{Y}_l)^2}{M-1} \tag{9}$$

e

$$T_{acc}\ parte\ 2 = VQM = \frac{1}{na}\sum_{l=1}^{na}\frac{M(\overline{Y}_l - VO_l)^2}{M-1}, \tag{10}$$

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 ambas as partes e, consequentemente, de T_{acc} melhor o ajuste.

E enfim, o coeficiente de correlação de Spearman:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{11}$$

na qual d_i=$x_i - y_i$ indica a diferença entre os postos dos pares x_i, y_i das variáveis em estudo. Quanto maior o valor da correlação, melhor a qualidade da imputação.

A imputação foi realizada baseando-se nos valores de q_i que compõem o IQA. Tais valores foram inseridos na matriz por ponto de coleta, ou seja, houve imputação de duas matrizes, uma para o rio Jaguari e outra para o rio Atibaia. As linhas da matriz utilizada para a imputação referem-se às diferentes datas de coleta para aquele rio e as colunas aos subíndices q_i.

A fim de realizar a comparação do resultado de IQA obtido com a imputação múltipla e o obtido por outro método, foi utilizado um método bastante intuitivo de cálculo do IQA quando da ausência de alguma variável que o compõe.

Para tanto, utilizou-se 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.

4.3 Bootstrap em blocos

No presente artigo foram realizadas análises por meio do Bootstrap em blocos móveis (KÜNSCH, 1989) e do Bootstrap estacionário (POLITIS e ROMANO, 1994). O desempenho de ambas as técnicas de aplicação da metodologia de Bootstrap em blocos foi comparado, bem como foram obtidos resultados, através dos mesmos, para análises práticas, referentes à situação dos rios no período estudado.

Quanto ao tamanho dos blocos, utilizou-se os valores obtidos por meio dos métodos de seleção automática. Para o Bootstrap em blocos móveis foram utilizados os tamanhos obtidos para o Bootstrap circular (POLITIS e ROMANO, 1992), uma vez que ambos apresentam grande semelhança metodológica.

Para a análise da qualidade da água, com base nesses dados, foi utilizado o valor do Índice de Qualidade da Água (IQA) em cada um dos rios, sendo os valores de IQA os componentes da amostra utilizada.

5 Resultados

5.1 Imputação de dados

Para a avaliação dos resultados de IQA imputados, obtidos imputando-se os valores dos subíndices q_i por meio do método de Bergamo (2007), foi realizada a comparação com os resultados de IQA imputados obtidos pela alteração dos pesos. Para tal, foram utilizadas as estatísticas de avaliação da qualidade da imputação trazidas na metodologia. Os resultados de imputação de IQA, e sua comparação com os originais, foram obtidos por simulações, retirando-se aleatoriamente dos dados coletados a proporção de ausência existente na amostra total. Com isso, é possível verificar o desempenho dos métodos de imputação.

Para o T_{acc} há nos resultados a divisão em duas partes, sendo a primeira a referente à diferenças entre imputações e a segunda a referente aos erros a cada imputação em relação ao valor original. Para essa segunda parte é possível obter um valor mesmo para métodos de imputação simples, a fim de comparação, ao se considerar todas as imputações como iguais.

O método analisado, de Imputação Múltipla Livre de Distribuição consta nas tabelas como Bergamo e o da alteração dos pesos como Pesos. As Tabelas 3 e 4 apresentam os resultados para a qualidade da imputação.

Tabela 3 - Estatísticas para os diferentes métodos - IQA Rio Jaguari

Estatística	Bergamo	Pesos
NRMSE	0.190	0.192
SPEARMAN	0.983	0.977
T_{acc} 2 (VQM)	5.066	5.147

Nota-se que o método de Bergamo (2007), quanto ao IQA, apresenta resultados bastante interessantes. O método dos pesos não é de todo ruim, apresentando um

Tabela 4 - Estatísticas para os diferentes métodos - IQA Rio Atibaia

Estatística	Bergamo	Pesos
NRMSE	0.323	0.440
SPEARMAN	0.936	0.905
T_{acc} 2 (VQM)	4.333	40.953

bom valor para a correlação de Spearman, principalmente, porém, o método de Bergamo apresenta valores maiores de correlação e, além disso, tem menor NRMSE e parte 2 do T_{acc} (VQM), principalmente no que se refere ao rio Atibaia.

É interessante notar, também, que a variabilidade entre as imputações é bastante baixa. Como foram imputados os subíndices, é interessante checar tal variabilidade na imputação dos mesmos. Para o Jaguari, enquanto a parte 2 do T_{acc}, considerando-se os subíndices foi de 618.32, a parte 1 (da variabilidade interna das imputações), foi de apenas 1.27. Para o Atibaia a parte 1 do T_{acc} foi de apenas 0.83, com a parte 2 tendo o valor 368.38. Isso evidencia a baixa variabilidade existente entre as imputações realizadas, o que será visível quanto ao IQA quando forem apresentados os resultados do Bootstrap, que são bastante similares entre as imputações.

Vale ressaltar que, para o rio Jaguari, ao analisar os resultados de imputação dos subíndices em si, sem levar em conta o cálculo do IQA, a correlação de Spearman não foi tão alta (0.332), sendo que, caso o intuito do estudo fosse analisar os subíndices de maneira pormenorizada, poderia haver questionamentos quanto à qualidade de imputação para tal finalidade. No rio Atibaia mesmo para os subíndices há alta correlação (0.827). Para a análise do IQA o método mostra-se adequado para ambos os rios.

Ainda é interessante dizer que, caso sejam utilizadas as variáveis brutas, antes da aplicação na curva dos subíndices q_i, o método de Bergamo (2007) não apresenta bons resultados. Nesse caso existe grande diferença nas escalas, e isso acaba influenciando fortemente, de maneira negativa, na qualidade das estimativas obtidas.

5.2 Bootstrap em blocos

Passando para a análise da qualidade do ajuste pelo Bootstrap em blocos, foram analisadas estatísticas como o viés e o erro padrão obtido pelas $B = 1000$ reamostras para cada técnica. Antes da realização de tal análise foi verificada a t-estacionariedade dos dados, obtendo-se a rejeição de raiz unitária, a 5% de significância, pelo teste de Philipps-Perron (PHILLIPS e PERRON, 1988), para Lag=4.

O tamanho dos blocos utilizados foram determinados por meio dos métodos de seleção automática presentes na função "b.star", do *software R*, presente no pacote "np". O tamanho dos blocos para cada uma das imputações foi de 19, para ambos os rios, no Bootstrap em blocos. Os valores esperados para o tamanho dos blocos

no Bootstrap estacionário são de 12 para o rio Jaguari e de 8 para o rio Atibaia.

Nas Tabelas 5-9 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 5 - Bootstrap para a Matriz Imputada - Primeira imputação

Rio	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Jaguari	55.194	54.403	2.415	2.266	0.849	0.057
Atibaia	45.773	45.573	1.127	1.063	0.231	0.031

Tabela 6 - Bootstrap para a Matriz Imputada - Segunda imputação

Rio	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Jaguari	55.218	54.434	2.411	2.260	0.841	0.058
Atibaia	45.764	45.563	1.132	1.068	0.232	0.031

Tabela 7 - Bootstrap para a Matriz Imputada - Terceira imputação

Rio	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Jaguari	55.216	54.433	2.410	2.259	0.841	0.058
Atibaia	45.765	45.563	1.133	1.068	0.233	0.031

Tabela 8 - Bootstrap para a Matriz Imputada - Quarta imputação

Rio	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Jaguari 1	55.216	54.433	2.410	2.260	0.841	0.058
Atibaia	45.765	45.563	1.133	1.068	0.232	0.031

Tabela 9 - Bootstrap para a Matriz Imputada - Quinta imputação

Rio	MédiaB	MédiaE	Erro PadrãoB	Erro PadrãoE	VícioB	VícioE
Jaguari	55.217	54.434	2.410	2.260	0.841	0.058
Atibaia	45.765	45.563	1.133	1.068	0.232	0.031

Nota-se que os resultados obtidos pelo Bootstrap estacionário tem valores bastante inferiores de vício, principalmente no que diz respeito ao rio Jaguari. Descritivamente, observa-se que não há muita diferença no erro padrão para ambos os métodos. Os resultados para cada uma das imputações são bastante similares, com diferenças muito leves.

Antes de prosseguir para a análise prática dos dados, é interessante verificar a simetria nos dados para cada rio, a fim de ter noção sobre a plausibilidade da construção usual do intervalo de confiança, por meio dos quantis. Também é interessante verificar se há muitos *outliers* nas médias das reamostras calculadas, a fim de verificar algum possível problema nos dados para a utilização dessa técnica. Uma vez que os resultados são bastante parecidos para cada imputação, a fim de tornar as análises mais concisas, serão apresentados os gráficos apenas para a terceira imputação, nas Figuras 1 e 2.

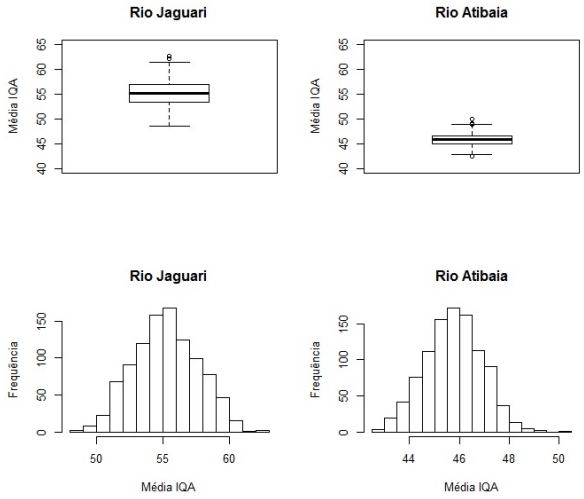

Figura 1 - Boxplots e Histogramas do Bootstrap em blocos móveis para a terceira imputação.

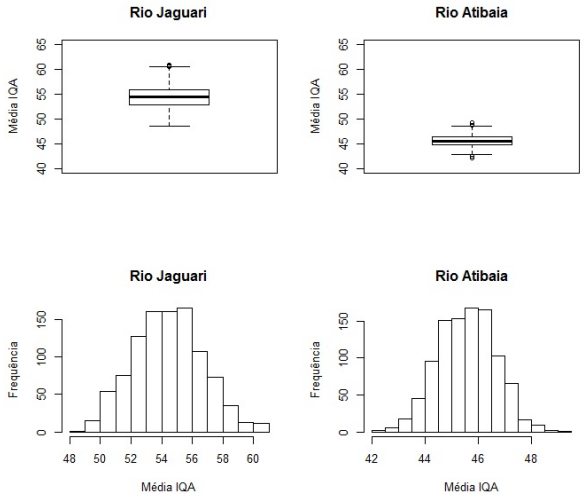

Figura 2 - Boxplots e Histogramas do Bootstrap estacionário para a terceira imputação.

Nota-se que não há muitos *outliers* entre as $B = 1000$ médias obtidas, bem como não há evidências de problemas mais sérios de assimetria, para nenhum dos dois métodos. Assim, foram prosseguidas as análises para verificar a situação dos rios no período estudado. A seguir, também foram utilizadas as técnicas de Bootstrap em blocos para o vetor de contrastes entre os rios, a fim de verificar se o intervalo de confiança entre os mesmos não possui o valor 0, a 5% de significância, o que indicaria diferença significativa entre os IQAs desses rios.

Os resultados quanto às estatísticas para comparação dos métodos Bootstrap são bastante semelhantes aos obtidos pelos vetores dos rios de maneira separada, sendo novamente o Bootstrap estacionário o que obtêm os menores vícios e com valores de erro padrão bastante similares. O tamanho dos blocos para o vetor de contrastes foi de 18 para o Bootstrap em blocos móveis e o tamanho esperado para o Bootstrap estacionário foi de 10.

Ainda é interessante salientar que ao dividir o banco de dados em períodos de tempo menores, como, por exemplo, cinco anos, os resultados do teste de Philipps-Perron indicam não haver t-estacionariedade e, além disso, a assimetria das médias Bootstrap passa a ser grande. A fim de exemplificar a existência de maior assimetria, as Figuras 3 e 4 apresentam os resultados para a terceira imputação, levando em conta apenas o período de 2005 a 2009.

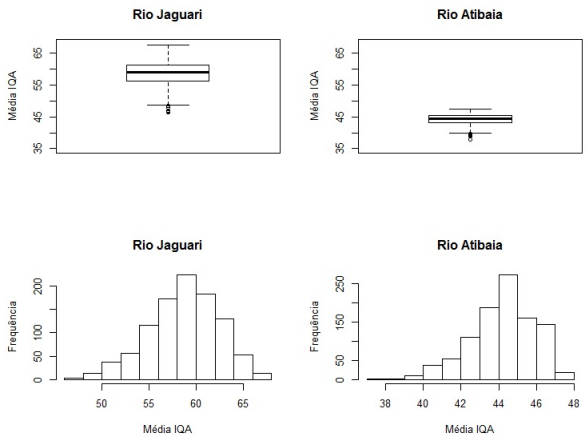

Figura 3 - Boxplots e Histogramas do Bootstrap em blocos móveis para a terceira imputação - Período 2005 a 2009.

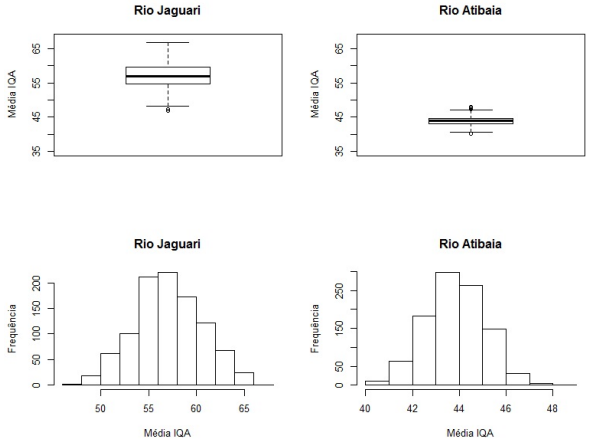

Figura 4 - Boxplots e Histogramas do Bootstrap estacionário para a terceira imputação - Período 2005 a 2009.

Tal situação dá indícios de que a utilização do Bootstrap em blocos para períodos de tempo menores, ao menos no caso desse estudo, não é a mais adequada. Vale ainda ressaltar que dividindo os períodos dessa maneira há alguns que tem poucas observações, e com isso, o tamanho dos blocos fica mais próximo ao tamanho de toda a amostra.

5.3 Análises práticas

Primeiramente, antes da realização das análises para todo o período, é interessante ao menos observar como os dados se portam ao longo dos anos. A Figura 5 dá um panorama de tal comportamento, levando em conta cada uma das imputações.

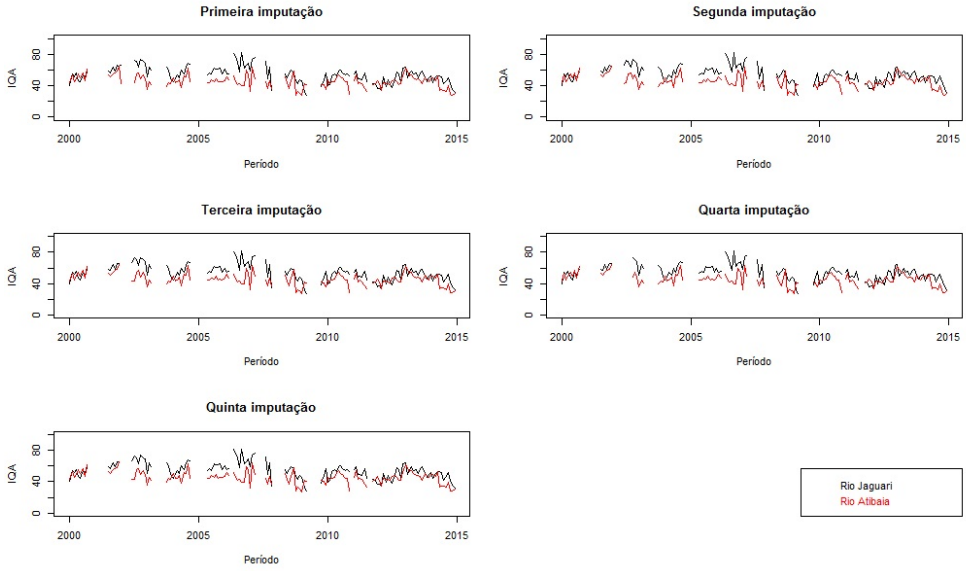

Figura 5 - IQA para cada imputação por ponto.

Os espaços em branco referem-se àqueles meses nos quais não houve coleta de água. Nota-se, descritivamente, que o IQA, em ambos os rios, teve alguma melhoria há alguns anos atrás, porém, voltou a cair nos últimos anos, tendo atualmente resultados não superiores aos do início do estudo. Não há também tendência a comportamentos muito típicos levando-se em conta, por exemplo, os espaços de tempo de 12 em 12 meses.

Realizando a análise dos dados, com todo o período sendo levado em conta, obtêm-se os seguintes intervalos quantílicos de confiança (quantil 0,025 e quantil 0,975):

- Rio Jaguari - Blocos móveis: (50,851 ; 59,703) Estacionário: (50,272 ; 58,959)
- Rio Atibaia - Blocos móveis: (43,564 ; 47,897) Estacionário: (43,491 ; 47,563).

Pode-se notar que, para o Rio Jaguari, o intervalo de confiança apresenta algumas leves diferenças entre as técnicas, porém, no que diz respeito à categoria de acordo com a tabela da CETESB, ambas as técnicas contêm, em sua maior parte, valores que classificam a qualidade da água como boa. Vale ressaltar que o limite inferior dos intervalos encontra-se já um pouco abaixo da passagem para a categoria cuja classificação é regular.

Quanto ao Rio Atibaia, mais uma vez ambos os métodos tem algumas diferenças numéricas mas resultados similares quanto à categoria no qual tais valores se encaixam. Dessa vez a classificação da água está na categoria regular, tendo alguma distância ainda razoavelmente grande de chegar próxima à categoria ruim.

Para o contraste obtêm-se o seguinte intervalo de confiança, também a 5% de significância:

- Jaguari-Atibaia - Blocos móveis: (6,251 ; 13,017)
- Estacionário: (5,783 ; 12,363)

Esses intervalos evidenciam que de fato existe diferença significativa entre a qualidade da água dos rios Jaguari e Atibaia, uma vez que o 0 não está contido. O rio Atibaia, o qual recebe maior influência de despejos de resíduos da atividade antrópica, apresenta IQA consideravelmente menor que o obtido para o rio Jaguari.

Conclusões

O método da "Imputação Múltipla Livre de Distribuição" mostrou-se adequado para realizar a estimativa dos subíndices ausentes, tendo-se como intuito final a análise do IQA. Para a análise pormenorizada dos valores de subíndices, levando em conta o rio Jaguari, principalmente, tal método não seria completamente adequado.

A Imputação Múltipla Livre de Distribuição, inicialmente proposta para casos de estudo de genótipos em diferentes ambientes, funciona bem para o caso em que utilizam-se os subíndices q_i na matriz de imputação, os quais possuem escalas similares. Porém, vale ressaltar que, ao ser testado tal método de imputação com as variáveis brutas de cada parâmetro que compõe o IQA (antes de utilizar-se os mesmos para obter os subíndices q_i), cada uma em sua escala original, as quais apresentam grandes diferenças entre si, os resultados não foram bons.

O Bootstrap estacionário obteve melhores resultados se comparado ao Bootstrap em blocos móveis, principalmente no tocante ao viés. Quanto à conclusão da análise obtida por ambos os métodos, elas foram bastante similares.

Quanto às limitações do Bootstrap em blocos, vale ressaltar que ele tem que ser realizado com os dados como se fossem sequenciais. O fato de haverem meses sem observações não impede o uso dessa técnica, que permite que em seus blocos sejam usados meses saltados, como se fossem sequenciais, para a finalidade de obtenção de estimativas pontuais e intervalares da média. Porém, a consideração de tais saltos

poderia ser interessante para o desenvolvimento de uma técnica mais aprimorada para circunstâncias como essa.

Além disso, para utilizar os métodos de Bootstrap em blocos em dados mensais como os desse estudo, é necessário haver um número consideravelmente grande de anos. Dependendo do intuito do pesquisador, pode ser interessante a análise em menores períodos de tempo, à qual o Bootstrap em blocos pode apresentar algumas dificuldades.

Já no referente à conclusão prática, a qualidade da água no rio Atibaia para os últimos 15 anos é consideravelmente preocupante. O rio Jaguari encontra-se em situação um pouco melhor, mas também tem um resultado aquém daquele que seria o ideal.

Agradecimentos

À CAPES, pelo financiamento do projeto de Mestrado, à FUNDUNESP pela coleta da água que gerou as informações utilizadas, desde o ano 2000, e à UNESP-Rio Claro pela realização da análise física, química e biológica dos dados e por tê-los cedido.

OLIVEIRA, E. S. B.; PIAO, A. C. S.; ANGELIS, D. F. An analysis of the water of rivers Jaguari and Atibaia using distribution-free multiple imputation and blocks bootstrap. *Rev. Bras. Biom.,* Lavras, v.34, n.4, p.543-563, 2016.

- *ABSTRACT: Longitudinal data have a huge relevance to evaluate several factors. Sometimes, there is also a problem of non collected data at some moment during the study, which leads to the necessity of, or desconsidering them, or using suitable techniques to estimate missing data. In this study, results based on data about the water quality of rivers Jaguari and Atibaia, results of multiple imputation for non normal data, through the "Distribution-free multiple imputation" and posteriorly results for data analysis using blocks Bootstrap techniques are presented. Lastly, practical results obtained through the analysis are shown, aiming to briefly identify the situation of both rivers along the 15 years of study.*

- *KEYWORDS: Multiple imputation; blocks bootstrap; water quality index.*

Referências

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.

BERGAMO, G. C. *Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação*. Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 2007. 236p.

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.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. Apêndice C: Índices de Qualidade Das águas. In: Companhia Ambiental do Estado de São Paulo, *Relatório de Qualidade das Águas Superficiais no Estado de São Paulo*, 2013.

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.

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.; GREENBERG, A. E.; EATON, A. D. *Standard Methods for the examination of water and wastewater*. 20th.ed., Washington: American Public Health Association (APHA), 1998.

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 Introduction to the Bootstrap*. London: 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.

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

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.

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

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.

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

KÜNSCH, H. R. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, v.17, p.1217-1241, 1989.

BROWN, R. M.; MCCLELLAND, N. I.; DEININGER, R. A.; TOZER, R. A. Water Quality Index - Do We Dare?. *Water Sewage Works*, v.117, p.339-343, 1970.

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.

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

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

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, p.53–70, 2004.

R CORE TEAM (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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

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

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.

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

SPEARMAN, C. E. The proof and measurement of association between two things. *American Journal of Psychology* , v.15, p.72-101, 1904.

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.

Recebido em 28.12.2015.

Aprovado após revisão em 18.04.2016.