XXIII SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS

AVALIAÇÃO DE TÉCNICAS DE TRATAMENTO DE DADOS DE QUALIDADE DE ÁGUA CENSURADOS À ESQUERDA

Germano Alves de Sousa Pirajá Martins[1]; Davide Franco[2]; Cátia Regina Silva de Carvalho Pinto[3]

RESUMO – A presença de dados censurados na análise da qualidade da água dificulta a extração de informações sobre as variáveis. Métodos estatísticos para lidar com dados censurados foram desenvolvidos no campo de monitoramento ambiental. O objetivo deste artigo é avaliar algumas técnicas para lidar com dados abaixo dos limites de detecção. A remoção de valores censurados, a substituição por metade do limite de detecção e os métodos Kaplan-Meier (KM), Regressão de Ordem Estatística (ROS) e Estimador de Máxima Verossimilhança (EMV) foram utilizados para estimar média, mediana e desvio padrão, em dados amostrados aleatoriamente de uma distribuição log-normal, com limites simulados de detecção variando entre 20 e 60%. Os resultados mostraram que a remoção de valores censurados deve ser evitada, devido à grande discrepância entre os parâmetros estimados e os valores populacionais. Os métodos EMV e ROS apresentam um erro menor na estimativa dos parâmetros para as amostras testadas.

ABSTRACT– The presence of censored data in water quality analysis create difficulties to extract information about the variables. Statistical methods for dealing with censored data were developed in environmental monitoring field. The aim of this article is to evaluate some techniques for dealing with data below of limit detection. The deletion of censored values (REM), substitution by half of the detection limit (SUB), and the methods Kaplan-Meier (KM), Statistical Order Regression (ROS) and Maximum Likelihood Estimator (EMV) were used for estimating mean, median and standard deviation, in randomly sampled data from a log-normal distribution, with simulated limit ranging between 20 and 60%. The results showed that the deletion of censored values should be avoided, due to the large discrepancy between the estimated parameters and population values. The EMV and ROS methods present a smaller error in the estimation of the parameters for the samples tested.

Palavras-Chave – qualidade da água; dados censurados; limite de detecção.

1 Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA), Florianópolis, Santa Catarina, Brasil. Programa de Pós-Graduação em Engenharia Ambiental, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil. Telefone: +55(48)32123300. e-mail: germano.ibama@gmail.com.

2 Departamento de Engenharia Sanitária e Ambiental, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil. Telefone: +55(48)3721-9992. e-mail: d.franco.ocean@gmail.com.

3 Departamento de Engenharias de Mobilidade, Universidade Federal de Santa Catarina (UFSC), Joinville, Santa Catarina, Brasil. Telefone: +55(47)3204-7409. e-mail: catia.carvalho@ufsc.br

INTRODUÇÃO

A avaliação da qualidade de água em um rio demanda um esforço considerável no que diz respeito à obtenção de amostras, ao escopo dos elementos a serem avaliados, ao tempo de duração do monitoramento, entre outros. Quando bem sucedidos, esses programas geram uma quantidade considerável de dados para análise. A estatística é uma ferramenta essencial para que o dado gerado no programa de monitoramento de qualidade da água possa ser convertido em informação que sirva à gestão ambiental do corpo hídrico.

Ocorre que os dados de qualidade de água apresentam características próprias, que demandam ferramentas estatísticas adaptadas à essa realidade. De acordo com LiyaFu e You-GanWang (2012), os dados de qualidade da água em geral possuem: distribuição não-normal, presença de outliers, valores abaixo do limite de detecção, dados vazios e dependência serial.

Um dos aspectos relevantes da análise dos dados de qualidade da água é a gestão dos valores censurados.

As análises laboratoriais para a quantificação dos constituintes da água possuem limitações que impossibilitam a estimativa precisa de alguns elementos em determinadas quantidades. O limite de detecção (LD) pode variar de acordo com o laboratório, com o método utilizado para análise, com as limitações dos instrumentos, entre outros fatores. Os valores abaixo do limite de detecção na análise de qualidade da água são identificados como dados censurados à esquerda, e devem ser compreendidos como observações não quantificadas, mas que se têm conhecimento de que estão entre zero e o valor do limite. Nos laudos de análise, os valores abaixo do limite de detecção são apresentados em geral como "<LD", onde LD representa o limite de detecção do método.

Um exemplo prático dessa situação foi encontrado em um programa de monitoramento do rio Uruguai, na divisa dos estados de Santa Catarina com o Rio Grande do Sul, onde a qualidade de água foi monitorada durante mais de 10 anos, em 67 campanhas, 14 pontos de monitoramento e 24 variáveis. Algumas dessas variáveis, com relevância para avaliação da qualidade da água, possuíam diferentes graus de censura. Os valores censurados (não-numéricos) impossibilitam a extração dos parâmetros descritivos da amostra, como médias e desvios padrão, assim os testes que utilizam o Teorema do Limite Central não podem ser aplicados (Clemens Reimann, 2015).

Os dados existentes no programa de monitoramento do rio Uriguai mostraram variáveis com níveis de censura que estavam entre 0 (zero) e 77,3% (Fenóis). Além disso, nas variáveis onde não constavam dados censurados (como turbidez, transparência, condutividade elétrica) foi possível notar distribuições amostrais que se aproximavam da curva log-normal. Esse tipo de distribuição é muito frequente em dados ambientais (Huston & Juarez-Colunga, 2009), se caracteriza pela

assimetria, com muita variabilidade, valores elevados, mas com maior parte dos dados concentrada nos valores menores (próximos a zero).

Há algumas alternativas para o manuseio de dados censurados. A eliminação do dado, de toda a variável ou do caso (amostra) em que ele se encontra são algumas delas. Outra abordagem é a substituição dos valores censurados por um valor constante igual ou menor do que o limite de detecção. Além dessas abordagens, existem métodos paramétricos e não-paramétricos que podem ser aplicados para obtenção da estatística descritiva de dados censurados.

Importante notar que as escolhas em relação ao manuseio do valor censurado têm consequências, como a alteração da distribuição quando se elimina o dado. Outra situação bastante frequente é a substituição dos valores censurados por um valor constante abaixo ou igual ao do limite de detecção (em geral uma fração). De acordo com Helsel (1990), a substituição dos valores é utilizada de maneira vasta, contudo carece de base teórica. Por outro lado, pesquisas com uso de dados simulados têm avaliado que as substituições podem ser realizadas para alguns testes estatísticos quando o nível de censura da variável é menor do que 25% (Antweiler, 2015; Stanimirova, 2013).

Para a extração de parâmetros de estatística descritiva em variáveis com dados censurados, podem ser ainda utilizadas as seguintes técnicas:

• Estimador da Máxima Verossimilhança (EMV): o método assume que os dados acima e abaixo do limite assumem a distribuição designada (normal ou lognormal) e os parâmetros são calculados para se adaptar da melhor forma a ela. O método tem sensibilidade a outliers e pode não funcionar adequadamente com pequenos conjuntos de dados, uma vez que neles não há informações suficientes para avaliar a aderência à distribuição de frequência escolhida (Helsel, 2011).

• Regressão de Ordem Estatística(ou Regression on Order Statistics – ROS) – Neste método os pontos censurados e não censurados são plotados no gráfico tipo Weibull, e um modelo linear é ajustado para estimar os valores com o método da máxima verossimilhança. A curva ajustada é utilizada para extrapolar os dados abaixo do limite de detecção (Helsel, 1990). Esses valores são considerados para o cálculo das estatísticas da distribuição, de forma coletiva. Importante ressaltar que o valor das amostras censuradas não devem ser utilizados individualmente em testes não paramétricos, uma vez que criam um ordenamento artificial das amostras (Christofaro & Leão, 2014). De acordo com Shumway (et. al., 2002), o ROS tem se mostrado uma das abordagens mais eficientes para estimativa de estatísticas em dados com múltiplos limites de detecção.

• Kaplan-Meier (KM) – É um método não-paramétrico, constituindo uma alternativa que não depende de uma determinada distribuição para ser utilizada para preenchimento dos dados

censurados. Apesar de poder se adaptar a dados com diversos limites de detecção, ainda não são vastamente utilizados com dados ambientais (Helsel, 2011). Originalmente o método foi desenvolvido em análises de sobrevivência, para uso com dados censurados à direita. Como adaptação para uso em dados ambientais (em geral com dados censurados à esquerda) é necessário inverter a ordem dos valores, o que permite criar uma função de distribuição empírica, e com isso estimar os parâmetros estatísticos da variável. O método é pouco sensível a outliers (frequente em dados ambientais) e funciona bem com amostras pequenas (Helsel, 2011).

O tamanho da amostra, o percentual de censura, o tipo de distribuição dos dados exercem influência no tipo de técnica a ser utilizada para tratamento de dados com limite de detecção(Antweiler & Taylor, 2008; Christofaro & Leão, 2014; Helsel, 2011).

Apesar de diversos estudos a respeito terem sido realizados para avaliação das técnicas aqui mencionadas, há controvérsia a respeito da mais adequada a ser utilizada no tratamento dos dados censurados, em vista da grande variedade de cenários possíveis.

O presente trabalho tem o objetivo de avaliar as técnicas utilizadas para tratamento de dados censurados para extração dos parâmetros descritivos, com uso de dados simulados extraídos aleatoriamente de distribuição log-normal e com aplicação de níveis de censura variáveis. Para mensurar a frequência relativa dos resultados e reduzir os efeitos da aleatoriedade, as técnicas foram testadas via simulação de Monte Carlo (Meyer, 1983). A avaliação dos métodos foi realizada por comparação com os parâmetros da população de onde as amostras foram obtidas, com o uso da Raiz do Erro Quadrático Médio (RMSE) e de testes de hipótese para comparação de média, desvio padrão e mediana.

METODOLOGIA

Para avaliação do desempenho das diferentes técnicas para tratamento de dados censurados foram estabelecidos os critérios para sorteio de valores aleatórios extraídos de uma distribuição log-normal, com média 2 (dois) e desvio padrão 2 (dois) - que são valores tipicamente encontrados em variáveis de importância na análise de qualidade de água superficial, como DBO e Nitrogênio Total.

Inicialmente, foram realizados testes de hipótese para verificação da alteração da distribuição quando os dados censurados eram removidos do conjunto, com amostras de tamanho 30, 100 e 1000. Para isso, foi usado o método de Shapiro-Wilk, com os dados com transformação logarítmica. Os resultados do teste de adesão à distribuição demonstraram que, mesmo com graus de censura abaixo de 20%, as distribuições não podem ser mais classificadas como log-normal.

Para simulação dos valores de censura, foram definidos percentuais relativos à população que variam de 20 a 60% (com intervalos de 2%). Com base nesses quantis, foram obtidos os valores

dos limites de detecção, tendo como referência à distribuição log-normal da população ($\mu = 2$, $\sigma = 2$). As amostras aleatórias extraídas da população foram submetidas à censura com os valores de limite de detecção apresentados. A Figura 1 mostra a curva log-normal da população com a localização da moda , média, alguns percentuais da distribuição e seus respectivos valores (que serão usados como limite de detecção).

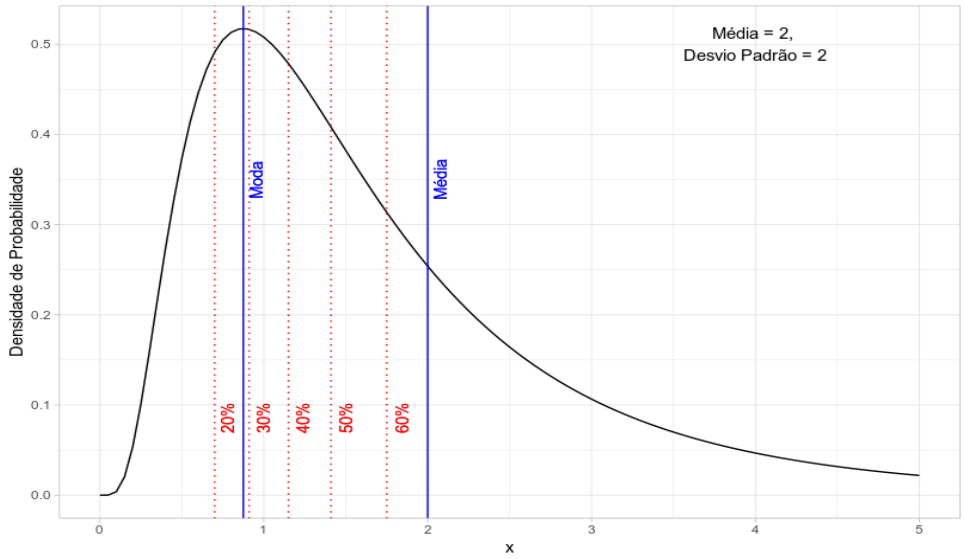

Figura 1: Curva teórica da densidade de probabilidade da população, com parâmetros moda, média.

A partir da aplicação dos limites de detecção na amostra, é possível aplicar as técnicas e extrair os parâmetros dos dados com censura simulada. Para este trabalho foram testados os métodos de remoção (REM), substituição pela metade do valor do limite de detecção (SUB), Kaplan-Meier (KM), Regressão de Ordem Estatística (ROS) e Estimador da Máxima Verossimilhança (EMV).

Para se reduzir os efeitos da aleatoriedade, as operações de sorteio dos dados, simulação de censura e aplicação das técnicas é repetida 100 vezes em cada um dos percentuais testados.

A comparação dos resultados obtidos por cada técnica para a extração dos valores dos parâmetros é realizada por meio da raiz do erro médio quadrado (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \quad (1)$$

Onde n é o tamanho da amostra, e y -$\hat{y}$ é a diferença entre o valor previsto e o verdadeiro.

Além do RMSE, foram aplicados teste estatísticos para avaliar se a média, desvio padrão e mediana podem ser consideradas iguais às estabelecidas na população de origem da amostra. Para a média, foi aplicado o teste *t* de Student. O desvio padrão foi comparado com o teste do Qui-Quadrado. Os valores obtidos para a mediana, foram comparados com o da população por meio do teste dos sinais de Wilcoxon.

As análises foram realizadas no programa R (R Core Team, 2019), utilizando o pacote NADA: Nondetects and Data Analysis for Environmental Data (Lee, 2017).

RESULTADOS

Após a aplicação da censura simulada nas 100 amostras, para cada valor de limite de detecção, os resultados de previsão dos valores de média, desvio padrão e mediana são extraídos com as técnicas REM, SUB, KM, ROS e EMV.

A Figura 02 mostra a distribuição de probabilidades em algumas faixas de censura (20, 40 e 60%) para a média em cada método testado. A linha vermelha indica o valor do parâmetro da população.

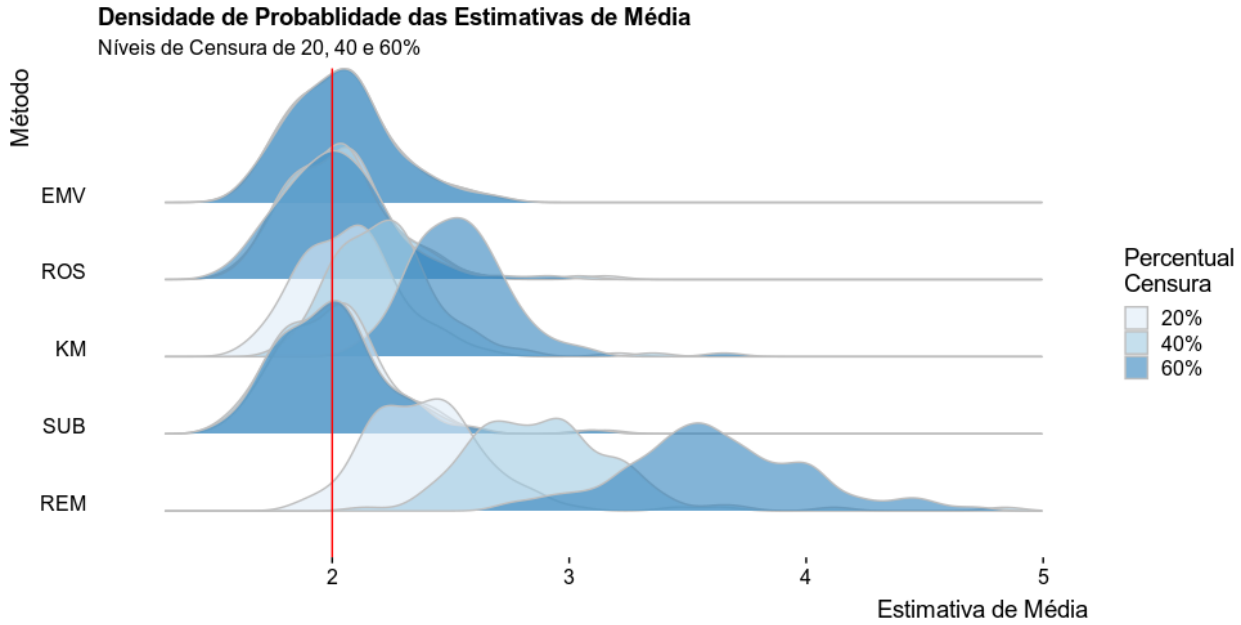

Figura 2: Gráfico da densidade de probabilidades dos resultados de estimativa de média, nos diferentes métodos de tratamento de censura, nas faixas de 20, 40 e 60%.

Para a média, é possível observar que a remoção dos valores censurados (REM) altera de forma significativa a estimativa do parâmetro, em especial a medida que os percentuais de dados abaixo do limite de detecção se elevam. Um comportamento semelhante é observado na técnica de Kaplan-Meier, onde valores de média da amostra afastam-se do valor do parâmetro para a população, quando os graus de censura se elevam. As técnicas de substituição pela metade do limite de detecção (SUB), ROS e EMV apresentam comportamentos em torno do valor real da população.

No caso da estimativa do desvio padrão, as curvas de densidade dos métodos mostram uma distribuição que não segue a normal, apresentando dados com assimetria a direita. Os resultados mostram que os métodos SUB, ROS e KM tem comportamentos semelhantes, com pouca variação em relação ao nível de censura. A REM apresenta o pior resultado, com elevação dos erros a

medida que a censura aumenta. O método EMV apresenta a menor diferença entre a estimativa e o desvio padrão, contudo tem sua precisão afetada pelo aumento da censura nos dados.

Para a previsão de mediana, o comportamento da estimativa para o conjunto de dados com os valores censurados removidos (REM) se assemelha ao da média. A medida que os valores de censura se elevam, a precisão e acurácia da estimativa se deteriora. Com valores de mediana sendo estimados para maior. A substituição dos dados pela metade do valor apresenta dois comportamentos distintos, enquanto a censura é menor que 50%, os valores de previsão convergem para a mediana da população. Quando esse grau de censura é alcançado, os valores de mediana estimados se afastam do valor real, para menor. O método KM não consegue estimar a mediana, quando o grau de censura ultrapassa 50%. Os métodos ROS e EMV apresentam boa acurácia nos testes, contudo perdem precisão com níveis mais elevados de censura.

A Figura 3, mostra o Erro Quadrático Médio (RMSE) de previsão dos parâmetros média, desvio padrão e censura para cada uma das técnicas.

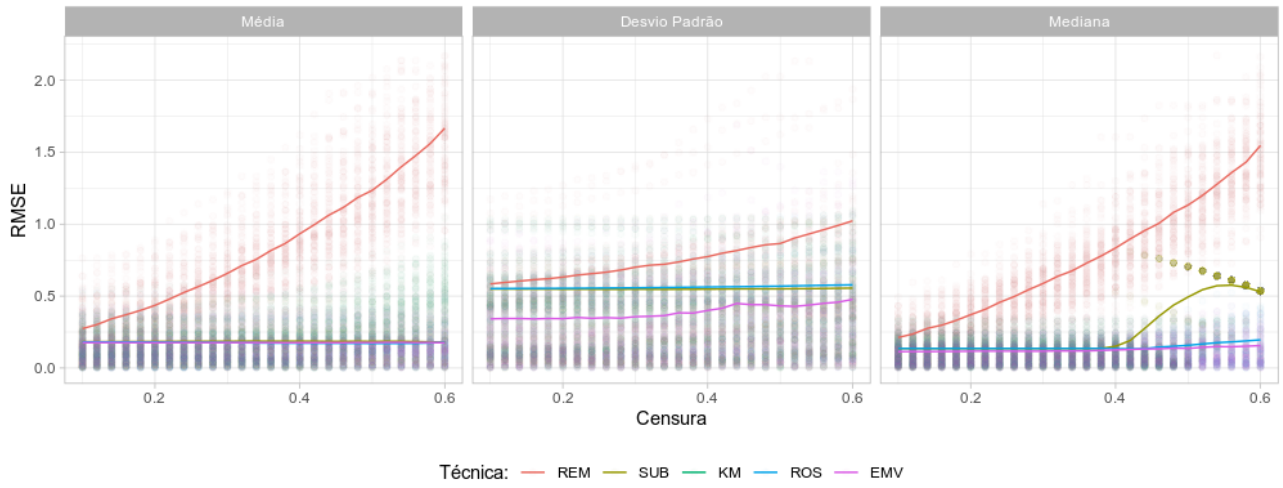

Figura 3: Gráficos do Erro Quadrático Médio dos métodos de tratamento de dados censurados, de acordo com o grau de censura, para previsão de Média, Mediana e Desvio Padrão.

Quanto maior o RMSE mais afastada está a previsão do parâmetro do valor real da população de onde foi extraída a amostra.

O teste t de Student foi aplicado para avaliar as médias obtidas das amostras censuradas em cada método. A hipótese nula, portanto, era de que a amostra possuía a média igual a da população. A probabilidade (ou valor-p) foi calculado para cada amostra e sumarizado com a média dos valores na Figura 3. As probabilidades e intervalos de confiança dos valores de *p* nos gráficos apresentados nas Figuras 3, 4 e 5 foram elaborados com o método Loess(Peng & Matsui, 2016).

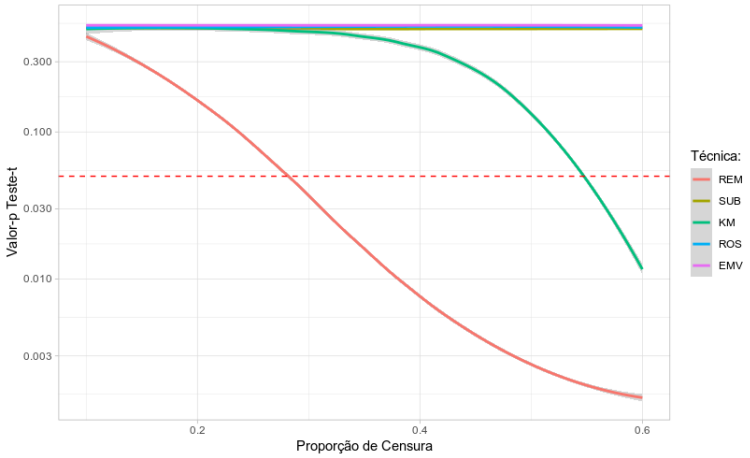

Figura 4: Gráfico da probabilidade estatística do teste-t de Student para as médias.

Para a média, é possível notar que as técnicas KM e REM perdem eficiência na previsão, fazendo com que a hipótese nula seja rejeitada a partir de 28% e 55%, respectivamente, considerando o nível de significância α de 5% (linha tracejada em vermelho).

Para a comparação dos valores do desvio padrão foi aplicado o teste de Qui-quadrado, onde a hipótese nula era de que os desvios eram iguais.

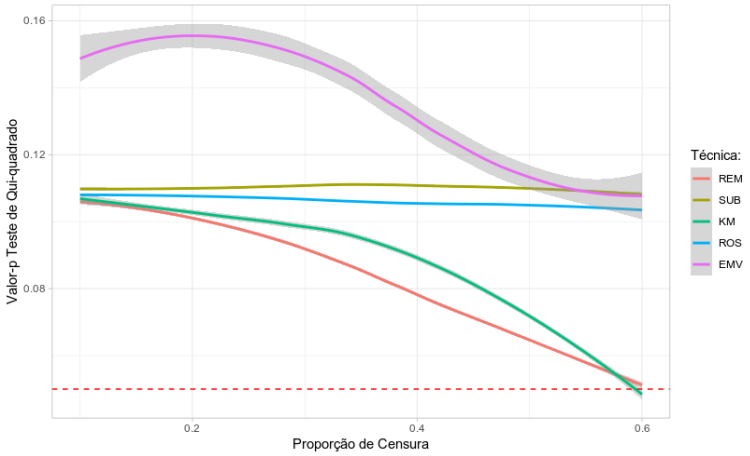

Figura 5: Gráfico da probabilidade estatística do teste Qui-Quadrado.

No caso do desvio padrão, a hipótese de que os valores são iguais só é rejeitada quando os percentuais elevados de censura na amostra são aplicados para KM e REM.

Para avaliação dos valores de mediana previstos em cada método, foi aplicado em cada intervalo de censura o teste de Wilcoxon, que é baseado no *ranking* dos valores. Neste caso, a hipótese nula é de que as medianas das amostras censuradas e da população são iguais.

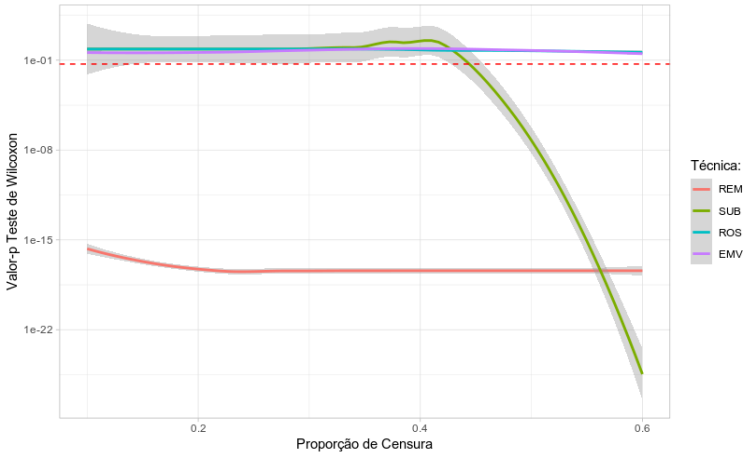

Figura 6: Gráfico da probabilidade estatística do teste de sinais de Wilcoxon.

Conforme Figura 6, mesmo em graus pequenos de censura, a remoção altera a avaliação da mediana. Como mencionado previamente, o método KM não consegue realizar a estimativa da mediana quando os valores de censura ultrapassam 50% na amostra. No caso da SUB, os valores da probabilidade caem a partir dos valores de censura de 40%.

CONSIDERAÇÕES FINAIS

Dados censurados são uma realidade na área ambiental que precisam ser tratados de forma adequada para que não interfiram na extração de informações a respeito do meio estudado.

Esse trabalho teve o objetivo de avaliar a forma como os tratamentos dos dados censurados se comportam quando se tem uma amostra derivada de uma população com distribuição log-normal. Em vista desse cenário, é necessária cautela na utilização dos resultados aqui apresentados, já que dados reais nem sempre seguem uma distribuição específica na natureza.

Além da limitação do teste em relação à distribuição log-normal, algumas outras variáveis devem ser consideradas na avaliação dos métodos para tratamento de dados censurados. Assim, seria importante realizar testes com variações de tamanho de amostras, múltiplos limites de detecção e simulação com dados reais censurados artificialmente para avaliar o desempenho das técnicas.

Os resultados obtidos mostram que os métodos EMV e ROS apresentam melhor desempenho que os demais na estimativa de média, desvio padrão e mediana. Por outro lado, fica claro que a remoção dos valores censurados deve ser evitada, pois altera de forma considerável os parâmetros de descrição dos dados.

REFERÊNCIAS

Antweiler, R. C. (2015). Evaluation of Statistical Treatments of Left-Censored Environmental Data Using Coincident Uncensored Data Sets. II. Group Comparisons. *Environmental Science and Technology*, *49*(22), 13439–13446. https://doi.org/10.1021/acs.est.5b02385

Christofaro, C., & Leão, M. M. D. D. (2014). Tratamento de dados censurados em estudos ambientais. *Quimica Nova*, *37*(1), 104–110. https://doi.org/10.1590/S0100-40422014000100019

Helsel, D. R. (1990). Less than obvious: Statistical treatment of data below the detection limit. *Environmental Science and Technology*, *24*(12), 1766–1774. https://doi.org/10.1021/es00082a001

Helsel, D. R. (2011). Statistics for Censored Environmental Data Using Minitab® and R: Second Edition. In *Statistics for Censored Environmental Data Using Minitab® and R: Second Edition*. https://doi.org/10.1002/9781118162729

Huston, C., & Juarez-Colunga, E. (2009). *Guidelines for computing summary statistics for data-sets containing non-detects*. 177. Retrieved from http://bvcentre.ca/files/research_reports/08-03GuidanceDocument.pdf

Lee, L. (2017). *NADA: Nondetects and Data Analysis for Environmental Data*. Retrieved from https://cran.r-project.org/package=NADA

Meyer, P. L. (1983). *Probabilidade - Aplicações à Estatística* (2º ed.). LCT.

Peng, R. D., & Matsui, E. (2016). The Art of Data Science: A guide for anyone who works with Data. *Journal of Chemical Information and Modeling*, *53*, 160. https://doi.org/10.1017/CBO9781107415324.004

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.r-project.org/

Shumway, R. H., Azari, R. S., & Kayhanian, M. (2002). Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits. *Environmental Science & Technology*, *36*(15), 3345–3353. https://doi.org/10.1021/es0111129

Stanimirova, I. (2013). Practical approaches to principal component analysis for simultaneously dealing with missing and censored elements in chemical data. *Analytica Chimica Acta*, *796*, 27–37. https://doi.org/10.1016/j.aca.2013.08.026