



IMPUTAÇÃO DE DADOS FALTANTES EM SÉRIES TEMPORIAS DE QUALIDADE DE ÁGUA: O CASO DO RIO DOCE

MELLO, Kyze Quadros¹;
NICOLETTE, Raquel da Fontoura²
Universidade Federal do Rio Grande

¹ Aluna PPGAmb. IMEF – FURG | kyzemello@hotmail.com ;

² Orientadora. Prof^ª. Doutora do PPGAmb IMEF - FURG | raquelnicolette@furg.br

A ausência de valores em uma série temporal dificulta o processamento dos dados e causa riscos na sua utilização, considerando que essa é uma das etapas mais importantes de um estudo com tal metodologia. O monitoramento ambiental consiste na coleta de dados ao longo de um período de tempo, seja ele provisório ou permanente, com a finalidade de avaliar o impacto das ações humanas ou naturais que acontecem em um determinado ambiente. Projetos subsidiados pelo Governo, como é o caso do Projeto Águas de Minas coordenado pelo Instituto Mineiro de Gestão das águas (IGAM), podem sofrer com interrupções no cronograma de coletas causadas por fatores diversos. É comum que o problema da existência dessas lacunas seja sanado por meio de técnicas simples, tais como a imputação dos valores de média ou mediana. Porém, as técnicas de imputação pouco sofisticadas podem não captar relações de dependência existentes entre os registros, como a autocorrelação e a autocovariância. Em virtude de tal dificuldade, diversos métodos estatísticos para lidar com dados em falta têm sido desenvolvidos, logo, considera-se relevante neste estudo efetuar a avaliação de dois métodos de Imputação Múltipla por Equações Encadeadas (MICE). Os métodos selecionados foram o “*pmm - Predictive Mean Matching*” (Correspondência preditiva da média) e o “*sample - Random sample from the available observed values*” (Amostra aleatória a partir dos valores observados), a fim de selecionar o mais apropriado para aplicar em um conjunto de dados de monitoramento da qualidade de águas superficiais. O objetivo deste trabalho é o de apresentar a comparação e a avaliação realizadas para dois métodos iterativos de imputação de dados ausentes em uma série temporal de registros de um monitoramento ambiental que verifica a qualidade da água do Rio Doce, nos períodos pré e pós “desastre de Mariana”, acidente ambiental decorrido pelo rompimento da barragem do Fundão/Mariana-MG. Utilizando a série histórica de observações do IGAM, o conjunto de dados selecionado compreende o período que inicia no ano de 2012 e vai até o final do ano de 2020. As observações registradas na estação de monitoramento RD035, situada dentro da calha do rio Doce, a jusante do município de Itatinga/MG, na região conhecida como Vale do aço, de relevante destaque para economia do Estado mineiro, recebeu o tratamento de imputação dos registros ausentes para as seguintes variáveis: Alumínio dissolvido e Oxigênio dissolvido, que correspondem a um total de trinta e três observações, e aproximadamente 30% de dados faltantes. Neste estudo, verificamos a eficiência dos métodos de imputação aplicados através da análise de sensibilidade por meio da extração do RMSE para conjunto de dados completados com o “*pmm*” e o “*sample*”. Foi validado aquele que apresentou o menor viés nos registros imputados, quando confrontados com os valores originais. De acordo com a raiz

¹ PPGAmbientometria IMEF - FURG – kyzemello@hotmail.com

² PPGAmbientometria IMEF - FURG – raquelnicolette@furg.br

do erro quadrático médio (RMSE), o método de imputação de dados faltantes que apresenta melhor acurácia é o “*pmm*”, tendo em vista que seu resultado é o $RMSE = 20,677$, o que demonstra que há uma boa medida de precisão do método, se comparado ao “*sample*” que teve o $RMSE = 583,677$. A imputação múltipla utilizada pareceu eficiente, pois os resultados conseguidos com os dados completados por imputações foram próximos dos obtidos com o conjunto completo no momento da análise de sensibilidade. No caso dos dados analisados da estação RD035, pode-se a partir dessa série fazer previsões do oxigênio dissolvido e da quantidade de alumínio presente no ambiente, em busca do entendimento das consequências dessas concentrações para a biota da região, após o rompimento de uma barragem. Concluindo, os métodos de imputação tornam-se ferramentas eficazes na análise de dados ambientais reais incompletos, os quais são bastante comuns, visando um melhor entendimento de processos associados a esses dados e viabilizando a realização previsões de estados futuros a partir dos mesmos.

Agradecimentos: Agradecemos a CAPES/PNPS pelo apoio financeiro.

Palavras-chave: Águas superficiais; Imputação múltipla de dados, Monitoramento ambiental, Séries temporais.