Imputation of missing values for air pollution data in Malaysia

Yong Zulina Zubairi*
Centre for Foundation Studies in Science, University of Malaya, Kuala Lumpur, Malaysia - yzulina@um.edu.my

Nuradhiathy Abd Razak
Institute of Graduate Studies, University of Malaya, Kuala Lumpur, Malaysia - tieramisu212@yahoo.com

Rossita M. Yunus
Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur, Malaysia - rossita@um.edu.my

Presence of missing values is unavoidable in most data collection. In the paper, we reviewed several missing values approaches. Simulation study is conducted to compare three methods of imputation namely mean substitution, hot deck, and expectation maximization (EM) imputation. The EM imputation is found to be superior especially when the percentage of missing values is high as it constantly gives the lowest RMSE values. Then, this method is applied for imputing PM10 concentrations data sets of two industrial areas in Malaysia, Petaling Jaya and Seberang Perai for year 2010. Then, two types of distributions namely the Weibull and lognormal are considered to describe the PM10 concentrations. The distribution that best fits the PM10 concentrations in Petaling Jaya is the Weibull distribution. Meanwhile, the results show that the lognormal distribution is the best to describe the PM10 concentrations in Seberang Perai.

Keywords: Missing value; expectation maximization; Mean imputation; Weibull