Missing Value Imputation for PM10 Concentration in Sabah using Nearest Neighbour Method(NNM) and Expectation-Maximization (EM) Algorithm

ABSTRACT

Missing data in large data analysis has affected further analysis conducted on dataset. To fill in missing data, Nearest Neighbour Method(NNM) and Expectation Maximization(EM) algorithm are the two most widely used methods. Thus, this research aims to compare both methods by imputing missing data of air quality in five monitoring stations (CA0030, CA0039, CA0042, CA0049, CA0050) in Sabah, Malaysia. PM10 (particulate matter with aerodynamic size below 10 microns) dataset in the range from 2003-2007(Part A) and 2008-2012 (Part B) are used in this research. To make performance evaluation possible, missing data is introduced in the datasets at 5 different levels(5%, 10%, 15%, 25% and 40%). The missing data is imputed by using both NNM and EM algorithm. The performance of both data imputation methods is evaluated using performance indicators(RMSE, MAE, IOA, COD) and regression analysis. Based on performance indicators and regression analysis, NNM performs better compared to EM in imputing data for stations CA0039, CA0042 and CA0049. This may be due to air quality data missing at random (MAR). However, this is not the case for CA0050 and part B of CA0030. This may be due to fluctuation that could not be detected by NNM. Accuracy evaluation using Mean Absolute Percentage Error (MAPE) shows that NNM is more accurate imputation method for most of the cases.