Key Engineering Materials, vol.594-595, 2014, pages 930-934

Mean imputation techniques for filling the missing observations in air pollution dataset

Abstract

Almost all real life datasets consist missing values. These are usually due to machine failure, routine maintenance, changes in siting monitors and human error. The occurence of missing values requires special attention on analysing the data. Incomplete datasets can cause bias due to systematic differences between observed and unobserved data. Therefore, the need to find the best way in estimating missing values is very important so that the data analysed is ensured of high quality. In this research, three types of mean imputation techniques that are mean, mean above and mean above below methods were used to replace the missing values. Annual hourly monitoring data for PM_{10} were used to generate missing values. Four randomly simulated missing data were evaluated in order to test the efficiency of the methods used. They are 5%, 10%, 15%, 25% and 40%. Three types of performance indicators that are mean absolute error (*MAE*), root mean square error (*RMSE*) and coefficient of determination (R^2) were calculated to describe the goodness of fit for all the method. From all the method applied, it was found that mean above below method is the best method for estimating data for all percentages of simulated missing values.

Keywords; Air Pollution, Imputation, Performance Indicators, PM_{10}