# SIIC107

## EVALUATION OF SINGLE MISSING VALUE IMPUTATION APPROACHES FOR INCOMPLETE AIR POLLUTION DATA IN MALAYSIA

Wan Suhailah Wan Mohamed Fauzi[1], Zuraira Libasin[2] and Ahmad Zia ul-Saufie[3]
*[1]Faculty of Chemical Engineering, Universiti Teknologi MARA Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang Malaysia*
*[2]Faculty of Computer and Mathematical Science, Universiti Teknologi MARA Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang Malaysia*
*[3]Faculty of Computer and Mathematical Science, Universiti Teknologi MARA Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang Malaysia*

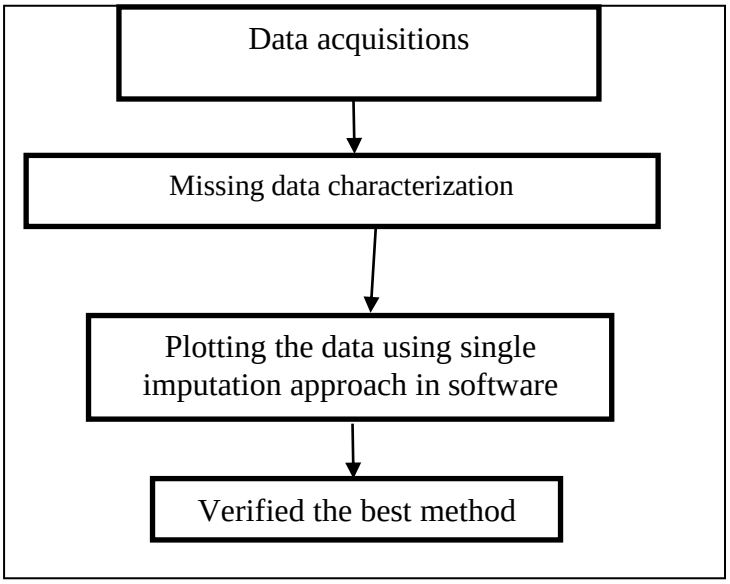
**Corresponding author: [2]zuraira946@ uitm.edu.my*

***Abstract:***
This research is mainly focused on environmental scope, which is air pollution. İt is about evaluation of single missing value imputation approaches for incomplete air pollution data in Malaysia. Single missing value imputation means the replacement of blank space in monitoring dataset from chosen DOE monitoring station with calculated value from the best method for long gap hours. The variable that mainly being monitor is $PM_{10}$. This variable is the primary source of air pollution release from industrial and transporation of everyday activities. Single imputation method focused in this research is mean imputation method. Furthermore, this methd will be tested on the dataset from Tanjung Malim monitoring station by fitting with many performance indicators such as MAE, RSME, $R^2$, PA and IA. The result will be compared with previous study whether it is the best used for long gap hour data. Four stages need to be followed in order to complete this research. The steps are data acquisitions, characteristic analyzing of missing value, single imputation approach and lastly, verification of approach and suggestion of the best method. The four existing imputation method for missing data implemented in this research are series mean method, mean of nearby points, linear trend and linear interpolation. The finding from this research shows that interpolation method is the best method to be applied for particulate matter missing data replacement with least mean absolute error and the better in performance accuracy.
***Keywords*:**
Imputation; Air pollution; Performance indicator; Interpolation; Missing data

***Objectives:***

- To determine the characteristic of missing value in air pollution data.
- To evaluate the best single imputation methods to the long gaps missing data using four different methods.

***Methodology:***




***Results:***

| METHOD | MISSING DATA % | MAE | RMSE | IA | PA | $R^2$ |
|---|---|---|---|---|---|---|
| LINEAR INTERPOLATION | 2.5 | 0.043233 | 1.659556 | 0.975495 | 0.962742 | 0.918349 |
| SERIES MEAN | | 0.593523 | 17.193990 | 0.336215 | 0.000000 | 0.000000 |
| MEAN NEARBY TWO POINTS | | 0.095029 | 5.423340 | 0.803035 | 0.699729 | 0.485118 |
| MEAN NEARBY ONE POINT | | 0.055462 | 2.018434 | 0.962313 | 0.942526 | 0.880187 |
| MEAN NEARBY THREE POINTS | | 0.096861 | 5.448565 | 0.798063 | 0.691625 | 0.473946 |
| LINEAR TREND | | 0.898617 | 25.231054 | 0.261694 | -0.515198 | 0.262989 |
| | | | | | | |
| LINEAR INTERPOLATION | 5 | 0.038920 | 2.159368 | 0.991016 | 0.982460 | 0.960815 |
| SERIES MEAN | | 0.256419 | 11.939487 | 0.285478 | 0.000000 | 0.000000 |
| MEAN NEARBY TWO POINTS | | 0.054667 | 2.914869 | 0.983292 | 0.967771 | 0.932299 |
| MEAN NEARBY ONE POINT | | 0.055381 | 2.945380 | 0.982898 | 0.967092 | 0.930991 |
| MEAN NEARBY THREE POINTS | | 0.055480 | 2.940319 | 0.982823 | 0.967325 | 0.931441 |
| LINEAR TREND | | 0.337788 | 15.867781 | 0.431738 | -0.679798 | 0.460012 |
| | | | | | | |
| LINEAR INTERPOLATION | 10 | 0.040230 | 0.871611 | 0.998997 | 0.998313 | 0.996402 |
| SERIES MEAN | | 0.313707 | 5.806496 | 0.950983 | 0.909237 | 0.826523 |
| MEAN NEARBY TWO | | 0.067985 | 1.567689 | 0.996726 | 0.993977 | 0.987764 |

| | | | | | | |
|---|---|---|---|---|---|---|
| POINTS | | | | | | |
| MEAN NEARBY ONE POINT | | 0.063324 | 1.504180 | 0.996996 | 0.994413 | 0.988631 |
| MEAN NEARBY THREE POINTS | | 0.071544 | 1.626157 | 0.996466 | 0.993557 | 0.986931 |
| LINEAR TREND | | 0.377668 | 6.434816 | 0.948203 | 0.905342 | 0.819457 |

## *Conclusion:*

This study concludes that the efficiency of Linear Interpolation method is be used in predicting the missing values closed to actual data for the case of particulate matter (PM10) variables for the long gaps regardless how many percentages of missing data complexity. Simulation results for this research demonstrate that Linear Interpolation method produces the lowest performance error and most accurate compared to others method. It is to be believed that when dealing with another data set for $PM_{10}$, the result produced will still be the same, which consists of lowest MAE and RSME. It is also noticeable that the IA, PA and $R^2$ values really approaches to digit one, which is the best-fit conditions for performance accuracy. However, further research needs to be done since the limitation of this research is the stated methods of imputations already implemented in SPSS and are used vastly by previous researchers. All in all, it can be said that if further experiment needs to be conducted, Linear Interpolation method still the best among five available method in SPSS., based on the experiment results.