UNIVERSITI TEKNOLOGI MARA

LONG GAP IMPUTATION IN AIR QUALITY (PM10) DATA SET USING IMPROVISED EOF-BASED METHOD WITH ROUGHNESS PENALTY APPROACH

SHAMIHAH BINTI MUHAMMAD GHAZALI

Thesis submitted in fulfillment
of the requirements for the degree of
Master of Science
(Statistics)

Faculty of Computer and Mathematical Sciences

January 2022

ABSTRACT

Well-produced analysis results require good quality data. However, missing data is often a major problem in several scientific research, including air quality data set. Missing values lead to the problem of low accuracy prediction and bias of the analysis results. This situation shows the importance of imputation methods to replace the missing values with estimated values. Based on the literature search, investigation for an appropriate imputation method on Single-Site Temporal Time-Dependent (SSTTD) multivariate structure air quality dataset particularly with long gap sequence of missing values issue was found less discussed. Several empirical orthogonal functions (EOF) based imputation methods are proposed in this study to fill the gap. The EOF, sometimes named Principal Component Analysis (PCA) method, is a promising technique applied to solve for missing values. However, the existing EOF imputation method has a drawback because it uses data matrix centralization based on statistics mean for EOF computation. To be applied for the air quality dataset, the existing approach needs to be improvised because the air quality dataset often consists of extreme observations due to climatic variations and random processes. Therefore, the implementation of statistic median and trimmed mean seems better in the matrix centralization. In this study, several proposed EOF-based methods are introduced. The capability of the methods for estimating missing values for long gap problems focusing on air quality (PM_{10}) of the SSTTD multivariate data set in Malaysia is investigated. The performance of the existing EOF based method, the EOF mean centred approach (EOF-mean) and several proposed EOF based methods; the EOF based on median (EOF-median), EOF based on the trimmed mean (EOF-trimmean) and the newly applied Regularized Expectation Maximization Principal Component Analysis (R-EMPCA) are compared. The study was conducted using real PM_{10} data set from Klang and Shah Alam air quality monitoring stations. Performance assessment and evaluation of the methods was conducted by comparing the imputed values in the artificial missing data set with the true observed values in the reference (complete) data set. The artificial missing values data sets are created from an identified reference (complete) data set with respect to several patterns according to four different percentages (5, 10, 20 and 30) and long sequence (gap) size (12, 24, 168 and 720) of missing points (hours) at both study locations. Based on several performance indicators, including RMSE, MAE, Rsquare and AI, the results have shown that R-EMPCA has the most excellent performance with the highest accuracy in estimating the missing values, and the second best is EOF-trimmean. For further improvement, the estimation of the estimated values was improvised using B-spline Roughness Penalty (RP) Smoothing approach, which resulted in the proposed R-EMPCA-RP and EOF-trimmean-RP imputation methods. The application of the RP approach is proven fruitful.

ACKNOWLEDGEMENT

Firstly, I wish to thank Allah SWT, the almighty God who is the source of my life and strength of knowledge and also for giving me the opportunity to embark on my master and for completing this long and challenging journey successfully.

My gratitude and thanks go to my main supervisor Dr. Norshahida Shaadan and co-supervisor Dr. Zainura Idrus for giving support and guidance for completing this master study. Without their thoughtful encouragement and patient supervision, this thesis would never have been done. I am very grateful to them for their contribution to the direction and richness of this thesis.

My appreciation goes to the Department of Environment Malaysia (DOE) for providing the information and data. Special thanks to my friends for helping me with this project.

Finally, this thesis is dedicated to my dearest parents Muhammad Ghazali Ismail and Yuhana Ab Ghani, for their vision and determination to educate me. They have provided the utmost support during my master journey. May Allah bless both of them always. This piece of victory is dedicated to both of you. Alhamdulilah for this journey. May this success encourage my beloved siblings, as I know they are capable of having the best achievement in their future.

TABLE OF CONTENTS