Malaysian Journal of Analytical Sciences, Vol 21 No 3 (2017): 552 - 559
DOI: https://doi.org/10.17576/mjas-2017-2103-05

MALAYSIAN JOURNAL OF ANALYTICAL SCIENCES
Published by The Malaysian Analytical Sciences Society
ISSN 1394 - 2506

MARKOV CHAIN MONTE CARLO METHOD FOR HANDLING MISSING DATA IN AIR QUALITY DATASETS

(Kaedah Rantai Markov Monte Carlo Untuk Mengurus Data Hilang Di Dalam Data Kualiti Udara)

Norhazlina Suhaimi[1], Nurul Adyani Ghazali[1]*, Muhammad Yazid Nasir [1], Muhammad Izwan Zariq Mokhtar [1], Nor Azam Ramli[2]

[1]*School of Ocean Engineering,*
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
[2]*School of Civil Engineering,*
Universiti Sains Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia

**Corresponding author: nurul.adyani@umt.edu.my*

Received: 27 October 2016; Accepted: 18 April 2017

Abstract

Missing data are a common problem in raw data especially in air quality datasets. Incomplete data due to machine or instruments failures, changes in the sitting air station monitors, calibration, routine maintenance and human error handling dataset. Multiple imputation of missing value technique was used to deal with selective air quality data by using Markov Chain Monte Carlo (MCMC). Expectation-maximization (EM) algorithm was used to compute the maximum likelihood estimate (MLE), assuming a multivariate normal distribution for the data. In this paper, the air quality monitoring stations selected namely Kemaman, Terengganu and Petaling Jaya, Selangor. The parameters selected were carbon oxide, ground level ozone, sulphur dioxide, nitrogen oxide, nitric oxide and nitrogen dioxide. A total of annual hourly data is 52,704 (8784 x 6 parameters) observations. Result shows that the coefficient of determination for all annual hourly monitoring data is consistently high (R^2 = 0.49 – 0.91) and small error (0.0001 – 0.3). Therefore, the multiple imputation technique by using MCMC method provides a good fit imputation and unbiased result of missing value to this data.

Keywords: missing data, air quality, multiple imputation, Markov Chain Monte Carlo

Abstrak

Data hilang ialah masalah biasa dalam data mentah terutamanya data kualiti udara. Data yang tidak lengkap disebabkan oleh kegagalan alat atau mesin, perubahan tempat stesen udara, kalibrasi, rutin penyelenggaraan dan kesilapan pekerja mengurus data. Teknik pelbagai imputasi digunakan untuk mengisi nilai hilang bagi data kualiti udara yang terpilih dengan menggunakan Markov Chain Monte Carlo. Jangkaan maksimum (EM) algoritma digunakan untuk menjangka kemungkinan maksimum dengan andaian taburan normal dari pelbagai pembolehubah. Dalam kajian ini, stesen kualiti udara yang dipilih ialah Kemaman, Terengganu dan Petaling Jaya, Selangor. Pembolehubah yang dipilih ialah karbon dioksida, ozon paras tanah, sulfur dioksida, nitrogen oksida, nitrik oksida dan nitrogen dioksida. Jumlah data jam tahunan ialah 52,704 (8784 x 6 pembolehubah) pemerhatian. Keputusan menunjukkan penentuan pekali untuk semua data jam tahunan ialah tinggi dan kesilapan kecil. Oleh sebab itu, teknik pelbagai imputasi dengan menggunakan kaedah MCMC menyediakan imputasi yang sangat bagus dan keputusan yang tiada keraguan bagi nilai hilang.

Kata kunci: data hilang, kualiti udara, pelbagai imputasi, rantai Markov Monte Carlo

Introduction

The air pollution in Malaysia has yet reached a critical level at the some areas [1]. However, even outside haze periods, pollution levels occurred despite tight regulations and this is exacerbated by the increase in the number of vehicle, distance travelled and growth in industrial production [1]. The haze phenomenon in Malaysia which contribute to the air pollutant reading including ozone concentration exceeds Malaysian standard [2] have already been observed in some urban [3] and industrial regions of Malaysia [4]. The air quality data was collected from the Air Quality Division of the Department of Environment, Ministry of Natural Resources and Environment of Malaysia. The National Continuous Air Quality Monitoring Network, manual air quality monitoring stations using high volume samplers were also established at the all air monitoring stations [1]. The main pollutants recorded at the Malaysian air quality monitoring stations are ground level ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and particulate matter of less than 10 microns in size (PM_{10}) [1].

Continuously sampling system is necessary to obtain a more reliable and accurate information about the air quality in atmosphere. However, real data are not usually complete that contain missing data. Missing data arise in almost real life datasets especially in air quality dataset. Incomplete data might be because of machine failures or break-down of measurement instruments, changes in the sitting air station monitors, calibration, routine maintenance and reading invalidation or human error handling dataset. Missing data are problematic in statistical analysis; the data will loss of information and efficiency [5], several problems related to data analysis, and bias due to systematic differences between observed and unobserved data [6]. According to Little and Rubin [5], rate of missing data less than 1% are considered trivial, 1 – 5% are manageable, 5 – 15% need powerful method to handle and more than 15% may harm the model's result.

There are various techniques of handling missing data such as interpolation techniques (linear, quadratic, cubic and nearest neighbor interpolation) [7], mean imputation techniques (mean-before after, mean before) [8], hot-deck imputation [9], regression imputation [10], k-nearest neighbor imputation [11]. Multiple Imputation (MI) provides a useful technique for dealing missing value in environmental research [12]. MI technique replaces each missing value with a plausible value that represents the uncertainty the right value to impute [13]. MI gives much better results to handle missing data [14]. The example for multiple imputation is Markov Chain Monte Carlo method [15]. Markov Chain Monte Carlo (MCMC) method was used for the MI procedure because of assuming multivariate normality [16]. Markov chain is a sequence of random variables in the distribution of each variable depends on the value of the previous variable. The expectation-maximization (EM) algorithm is a technique that estimates maximum likelihood for MCMC method [17]. Therefore, the aim of this study focuses on MI technique to replace missing value for air quality data.

Materials and Methods

This section explains the source of data and process of constructing the multiple imputation technique. We have designed a SAS version 9.0 codes for MI technique.

Study area

This study involving real secondary data that obtained from the Department of Environmental, Malaysia managed by a private company, Alam Sekitar Malaysia Sdn. Bhd (ASMA). There are two air monitoring stations were studied for this research; Kemaman, Terengganu (S1) and Petaling Jaya, Selangor (S2). Kemaman is developing Malaysian town located at the industrial Kertih Petrochemical Industrial Area in the North and the industrializing and urbanizing Gebeng Industrial Area in the South. The monitoring station was located at Sekolah Rendah Bukit Kuang with coordinate (4°14'21.9"N 103°11'31.8"E). Petaling Jaya, Selangor is located within Klang Valley region and covers an area of 97.2 km^2. It's surrounded by residential, commercial and industrial areas leads to a high volume of traffic. The monitoring station was located at Sekolah Rendah Sri Petaling with coordinate (3°08'16.7"N 101°36'29.7"E). Both stations were classified as industrial area although different geographical area. The locations of the continuous air quality monitoring stations for this research are shown in Table 1 and Figure 1.

Table 1. Air quality monitoring stations description

Air Monitoring Station	Location	Background	Coordinates	
			Latitude	Longitude
S1	Kemaman, Terengganu	Industrial	$4^0$14'21.9'' N	$103^0$11'31.8''E
S2	Petaling Jaya, Selangor	Industrial	$3^0$08'16.7''N	$101^0$36'29.7''E

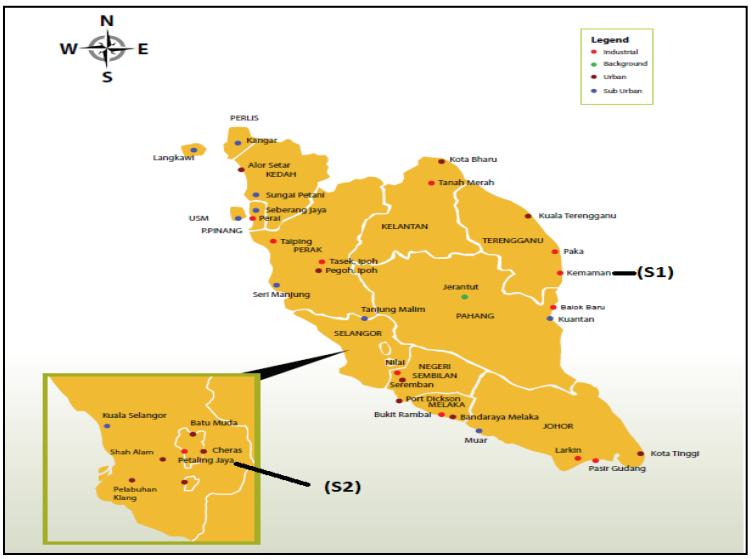

Figure 1. Location of Air Quality Monitoring Stations in Peninsular Malaysia, 2012

Data acquisition

Six air pollutants were selected to carry out replacing missing value. The overall air quality data used was collected from January to December for year 2012. The total number of observations is 52,704 records corresponding to the hourly value for each one of the following 6 parameters (8,784 observations x 6 parameters). The highest percentage missing data of parameter for S1 is NO (45.1%) and S2 is O_3 (20.4%). Table 2 show the parameters used in this study and percentage of missing data.

Table 2. Parameters used during period of study

Station	% Missing Data					
	CO	O_3	SO_2	NO_X	NO	NO_2
S1	6.9	5.3	34.4	8	45.1	10.1
S2	5.8	20.4	8.8	5.3	5.6	5.3

CO-Carbon Monoxide, O_3-Ozone, SO_2-Suphur Dioxide, NO_x-Nitrogen oxide, NO-Nitrogen Oxide, NO_2-Nitrogen Dioxide

Multiple imputation
The imputation has been performed by Monte Carlo simulation of MCMC method. The expectation-maximization (EM) [17] is a technique that finds maximum likelihood estimates for MCMC method to replace missing data. MCMC method is based on Bayesian inference with missing data by following several steps:

1. Imputation step. Estimate mean and covariance matrix, then simulates the missing values for each observation.
2. Posterior step (P-step). P-step simulates the mean vector and covariance matrix from the imputed step. This is the posterior distribution.

A current parameter estimate $\theta^{(t)}$ at t^{th} iteration, the I-step draws $Y_{mis}^{(t+1)}$ from $p(Y_{mis} \mid Y_{obs}, \theta^{(t)})$ and the P-step draws $\theta^{(t+1)}$ from $p(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$. Therefore, a Markov chain is created as shown below:

$$(Y_{mis}^{(1)}, \theta^{(1)}), (Y_{mis}^{(2)}, \theta^{(2)}) \tag{1}$$

which converges in distribution to $p(Y_{mis}, \theta \mid Y_{obs})$.

Performance indicators
There are four performance indicators were used to explain the efficiency of the imputation method used in this study. The observed data and theoretical data were compared to show the goodness of method for replacing missing values. Four performance indicators were used namely; coefficient of determination (R^2), prediction accuracy (PA), mean absolute error (MAE) and root mean squared error (RMSE).

Coefficient of determination
The coefficient of determination (R^2) explains how much the variability of the imputed data that related with observed data. The value of R^2 is between 0 and 1, with value closer to 1 implying a better fit model. The R^2 is computed using the following equation 2 [16]:

$$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)}{\sigma_p \sigma_o} \right]^2 \tag{2}$$

where N is the number of observations, O_i is observed data, P_i is the imputed data, $\overline{P}$ and $\overline{O}$ are the average of imputed data and observed data, and σ_P and σ_O are the standard deviation of the imputed data and observed data respectively.

Prediction accuracy
The Prediction accuracy (PA) takes on value range from 0 to 1 with closer to 1 describing a better fit model. The equation of PA is given as follows [18]:

$$PA = \sum_{i=1}^{N} \frac{\left[\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)\right]}{(N-1)\sigma_P \sigma_O} \tag{3}$$

where N is the number of observations, O_i is observed data, P_i is the imputed data, $\overline{P}$ and $\overline{O}$ are the average of imputed data and observed data, and σ_P and σ_O are the standard deviation of the imputed data and observed data respectively.

Mean absolute error
Mean absolute error (MAE) is the average of the difference between observed data and predicted data. The MAE is computed by Schafer [16]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - O_i\right| \tag{4}$$

where N is the number of observations, O_i is observed data, P_i is the imputed data, $\overline{P}$ and $\overline{O}$ are the average of imputed data and observed data, and σ_P and σ_O are the standard deviation of the imputed data and observed data respectively.

Root mean squared error
The root mean squared error (RMSE) is one of the common errors that measure for building of model and testing accuracy of the model. The small value explains the good performance of the model. This error is evaluated by the equation 5 [16]:

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}\left[P_i - O_i\right]^2\right)^{\frac{1}{2}} \tag{5}$$

where N is the number of observations, O_i is observed data, P_i is the imputed data, $\overline{P}$ and $\overline{O}$ are the average of imputed data and observed data, and σ_P and σ_O are the standard deviation of the imputed data and observed data respectively.

Multiple imputation efficiency
The relative efficiency is computed by using m imputation and λ as shown in equation 6

$$RE = (1 + \frac{\lambda}{m})^{-1} \tag{6}$$

Results and Discussion

Table 3 gives the summary of parameters for air quality data. The mean values for both stations and for all parameters are higher than median which indicates that the parameters distributions are skewed to the right. The variance value shows the variability of parameter concentration. The station 1 explained the distribution for CO (sk = 1.36), SO_2 (sk = 6.947), NO_x (sk = 2.14), NO (sk = 4.15), and NO_2 (sk = 2.08) are highly skewed to the right. Meanwhile, the distribution for O_3 (sk = 0.922) is slightly skewed to the right. For the station 2, the distribution for O_3 (sk = 1.36), SO_2 (sk = 5.23), and NO (sk = 1.55) are highly skewed to the right. Meanwhile, the distribution for CO (sk = 0.93), NO_x (sk = 0.99), and NO_2 (sk = 0.94) are slightly skewed to the right. Hence, all parameters are skewed to the right because of nonnegative values. The data was analyzed assuming missing at random because the air quality data was missing being due to the monitoring site down [19].

MI technique was performed by using SAS version 9 with the imputation of m = 5 values for each missing record hence creating five complete datasets. Table 4 shows the relative efficiency of missing data for both stations. The result shows that all parameters are high efficiency very close to 1. This indicates that imputations with five replications are needed for simulation of this data.

Table 5 shows that the parameters give the close fit between the actual data and imputed data because of high value for R^2 and PA implying closer to 1. Meanwhile, the errors of the model for both stations are small and closer to 0 for parameters. According to the performance of R^2, the highest R^2 for both stations is CO parameter in S2 which is 0.906. Meanwhile, the highest performance of PA is NO in S2 which is 0.9513. The smallest RMSE is NO for S1

which is 0.0005 and the smallest MAE is also for NO in S1 which is 0.0002. Although R^2 value is quite low for SO_2 (0.497) in S1 but the performance of PA is quite high and also given small error. Therefore, this assures us that the MI technique by using MCMC method produced consistent result and suitable for the replace missing value for air quality datasets. It is supported with [20] reported that MI is suitable method to solve the high ratio of missing values for example 43.5% missing data. Another study from [21] shows that MCMC method was more effective than simple mean for the monthly rainfall data in the northern region of Thailand.

Table 3. Details of the two original datasets

Station	US	CO	O_3	SO_2	NO_x	NO	NO_2
S1	M	0.35	0.021	0.0017	0.005	0.002	0.004
	Med	0.34	0.018	0.001	0.004	0.001	0.003
	Var	0.018	0.0002	0.00001	0.00002	0.000003	0.00001
	Sk	1.36	0.922	6.947	2.14	4.15	2.08
S2	M	1.17	0.015	0.0034	0.05	0.03	0.03
	Med	1.10	0.009	0.003	0.055	0.023	0.028
	Var	0.31	0.0002	0.00001	0.001	0.001	0.0001
	Sk	0.93	1.36	5.23	0.99	1.55	0.94

**S-Station, US-Univariate Statistic, M-Mean, Med-Median, Var-Variance, Sk-Skewness*

Table 4. Relative efficiency (RE)

Station	CO	O_3	SO_2	NO_X	NO	NO_2
S1	0.99	0.99	0.95	0.97	0.97	0.98
S2	0.99	0.98	0.99	0.99	0.99	0.99

Table 5. The performance indicators for both stations

Station	Parameter	Performance indicator			
		R^2	PA	RMSE	MAE
S1	CO	0.8911	0.9439	0.0985	0.0242
	O_3	0.8865	0.9416	0.0061	0.0011
	SO_2	0.4970	0.7050	0.0016	0.0007
	NO_x	0.8627	0.9288	0.0017	0.0004
	NO	0.7606	0.8721	0.0005	0.0002
	NO_2	0.8822	0.9392	0.0013	0.0003
S2	CO	0.9060	0.9518	0.3039	0.0652
	O_3	0.7695	0.8771	0.0070	0.0024
	SO_2	0.8576	0.9260	0.0012	0.0003
	NO_x	0.9013	0.9494	0.0149	0.0031
	NO	0.9050	0.9513	0.0084	0.0015
	NO_2	0.8859	0.9413	0.0074	0.0016

Conclusion

This paper discussed the use of multiple imputation technique to replace missing values in air quality dataset. The MCMC technique was used. The hourly air quality data was used to test the performance of the imputation technique. The relative efficiency was calculated to examine the accuracy technique for replacing missing values. The result shows that the relative efficiency is high and more than 0.9. Based on the result of performance indicators shows that the MCMC method give the good linear relationship because of R^2 and PA are high approaches to 1 for both stations and followed with small errors for RMSE and MAE. Therefore, it can be concluded that MCMC method is suitable method for replacing missing air quality data. MCMC method is also reliable to replace missing value either low or high percentage of missing data. Missing values are always arises, but a proper imputation can help remedy the analysis as much as possible.

Acknowledgement

This study was funded by Ministry of Higher Education Malaysia under Research Acculturation Collaborative Effort (RACE) Grant Scheme Phase 2/2013. The author would like to thank the Ministry of Education Malaysia for its financial support to carry out this study under the MyBrain15 program. The author also would like to thank the Universiti Malaysia Terengganu for its financial support to carry out this study and Department of Environment (DoE) Malaysia for their permission to utilize air quality data for this study.

References

1. Department of Environment Malaysia (2012). Malaysia Environmental Quality Report 2012. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.
2. Awang, M., Jaafar, A. B., Abdullah, A. M., Ismail, M., Hassan, M. N., Abdullah, R., Johan, S. and Noor, H. (2000). Air quality in Malaysia: Impacts, management issues and future challenges. *Respirology,* 5: 183 – 196.
3. Nichol, J. (1998). Smoke haze in South East Asia: A predictable recurrence. *Atmospheric Environment,* 32: 2715 – 2716.
4. Yusoff, N. F. F., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A. and AlMadhoun, W. A. (2010). Monsoonal differences and probability distribution of PM_{10} concentration. *Environmental Monitoring and Assessment,* 163: 1 – 4.
5. Little, R. J. and Rubin, D. B. (1987). *Statistical* analysis with missing data*,* New York: John Wiley and Sons.
6. Noor, M. N., Yahaya, A. S., Ramli, N. A. and Mustafa Al Bakri, A. M. (2014). Mean imputation techniques for filling the missing observations in air pollution dataset. *Trans Tech Publications, Switzerland, Key Engineering Materials*: pp. 902 – 908.
7. Batista, G. E. A. P. A. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence,* 17: 519 – 533.
8. Li, D., Deogun, J., Spaulding, W. and Shuart, B. (2004). Towards missing data imputation: A study of fuzzy k-means clustering method. In S Tsumoto, R.Slowinski, J.Komorrowski, & J. W. Grzmala-Busse (Eds), Lecture notes in computer science: Rough sets and current trends in computing. Sweden:Springer-Verlag: pp. 573 – 579.
9. Noor, M. N., Yahaya, A. S, Ramli, N. A. and Abdullah, M. A. A. B. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia,* 34: 341 – 345.
10. Jerez, J. M., Molina, I., GarcÍa-Laencina, P. J., Alba, E., Ribelles, N., Martín, M. and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in real breast cancer problem. *Artificial Intelligence in Medicine,* 50: 105 – 115.
11. Silva-Ramírez E-L., Pino-Mejías R., López-Coello M. and Cubiles-de-la-Vega, M-D. (2011). Missing value imputation on missing completely at random data using multilayer perceptron. *Neural Networks,* 24: 121 – 129.
12. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality datasets. *Atmospheric Environment,* 38: 2895 – 2907.
13. Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys*,* New York: John Wiley & Sons, Inc.
14. Greenland, S. and Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology,* 142: 1255 – 1264.

15. Lu, W-Z. and Wang, D. (2008). Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. *Science of the Total environment,* 395: 109 – 116.
16. Schafer, J. L. (1997). Analysis of incomplete multivariate data. Chapman and Hall, New York.
17. Little, R. A. and Rubin, B. B. (2002). Statistical analysis with missing data 2nd edition. Wiley, New York: pp. 4 - 22.
18. Chen, J. L., Islam, S. and Biswas, P. (1998). Nonlinear dynamics of hourly ozone concentrations: Nonparametric short term prediction. *Journal of Atmospheric Environment,* 32: 1839 – 1848.
19. Plaia, A. and Bondi, A. L. (2006.) Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment,* 40: 7316 – 7330.
20. Gómez-Carracedo, M. P., Andrade, J. M., López-Mahlía, P., Muniategui, S. & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems,* 134: 23 – 33.
21. Ingsrisawang, L. and Potawee, D. (2012). Multiple imputation for missing data in repeated measurements using MCMC and copulas. *Proceedings of the International Multiconference of Engineers and Computers Scientists*. 14-16 March 2012, Hong Kong.