Borneo International Journal eISSN 2636-9826; Vol. 6 (3); 2023; 104-108
Published by *Majmuah Enterprise*
www.majmuah.com

Imputation of Carbon Monoxide (CO) Missing Data in Petaling Jaya Using Basic Statistical Methods

N. Z. A. HAMID*, N. S. A. O. ALI, R. A. TARMIZI, N. S. A. KARIM, N. W. M. JUNUS,
N. H. M. HUSIN, N. H. ADENAN & N. B. A. WAHID
Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900, Tanjong Malim, Perak, MALAYSIA
Email: nor.zila@fsmt.upsi.edu.my | Tel: +601548797416 |

Received: September 16, 2023
Accepted: September 22, 2023
Online Published: September 26, 2023

Abstract

Air pollution is the act of contaminating the cleanliness of the air, which deteriorates the air quality and severely affects human health and the environment. This study has examined the imputation of carbon monoxide (CO) missing data, conducted in the metropolitan area of Petaling Jaya, Malaysia using basic statistical methods. The purpose of this study is to determine the best method among the five basic statistical methods for the imputation of missing CO data in the study area. The five basic statistical methods used to impute the missing data are Linear Interpolation (*L*), Top Bottom Mean (*T*), Daily Mean (*D*), 12-Hour Mean (*M*12), and 6-Hour Mean (*M*6). Annual hourly data for CO gas pollutants from Petaling Jaya were taken in 2017, and the data used is complete and continuous. The percentages of missing data applied in this study are 10%, 20%, and 30%. The performance indices used to evaluate basic statistical methods are Correlation Coefficient (*C*), Mean Absolute Error (*M*), and Root Mean Square Error (*R*). Overall, the best basic statistical method is *L* ; it is hoped that this study can help the Malaysian Department of Environment (DOE) in imputing missing data for CO air pollutants in the future.

Keywords: Carbon Monoxide Time Series; Metropolitan Area; Missing Data Imputation; Basic Statistical Methods; Performance Index

1. Introduction

Air quality issues can be linked to the development process of a country, especially urbanization. According to Agustina et al. (2020), the increase in the number of vehicles every year has surged the concentration of pollutants in the air including carbon monoxide gas, CO, and nitrogen dioxide, NO_2. Air monitoring stations in the metropolitan areas of Klang Valley, are surrounded by busy main roads and are affected by vehicle smoke emissions from heavy traffic (Shuhaili, Ihsan, & Faris, 2013). According to Azmi et al, (2010) CO, NO_2, and Sulphur Dioxide (SO_2) concentrations were recorded high in Petaling Jaya because of motor vehicles and traffic activities.

Carbon monoxide (CO) is a poisonous, colorless, odorless and tasteless gas. Although it has no detectable odor, CO is often mixed with other gases that do have an odor. So, people can inhale CO right along with gases that they can smell and not even know that CO is present. CO is a common industrial hazard resulting from the incomplete burning of material containing carbon such as natural gas, gasoline, kerosene, oil, propane, coal, or wood. Forges, blast furnaces and coke ovens produce CO, but one of the most common sources of exposure in the workplace is the internal combustion engine. CO is harmful when breathed because it displaces oxygen in the blood and deprives the heart, brain and other vital organs of oxygen. Large amounts of CO can overcome people in minutes without warning, causing people to lose consciousness and suffocate. Besides tightness across the chest, initial symptoms of CO poisoning may include headache, fatigue, dizziness, drowsiness, or nausea. Sudden chest pain may occur in people with angina. During prolonged or high exposures, symptoms may worsen and include vomiting, confusion and collapse in addition to loss of consciousness and muscle weakness. Symptoms can vary widely from person to person. CO poisoning may occur sooner in those most susceptible: young children, the elderly, people with lung or heart disease, people at high altitudes, or those who already have elevated CO blood levels, such as smokers. Also, CO poisoning poses a special risk to foetuses (OSHA Malaysia, 2023). Therefore, it is essential to study the time series of CO pollutants.

Zainuri et al. (2015) stated that missing data is a pervasive problem in many fields, such as long-term research studies, experimental studies, and data obtained from questionnaire studies that are used for their respective needs. This issue also includes environmental studies like air quality data caused by various problems, such as machine malfunctions, human errors, and insufficient sampling. Incomplete data sets can cause results to be influenced by systematic differences between missing and non-missing data (Sukatis, Noor, Zakaria, Ul-Saufie, & Suwardi, 2019). Data with missing values can cause significant problems; for example, time series analysis requires continuous data to make

predictions (Noor, Yahaya, Ramli, & Al Bakri Abdullah, 2015; Saeipourdizaj, Sarbakhsh, & Gholampour, 2021; Sukatis et al., 2019; Zakaria & Noor, 2016). Therefore, the focus of this study is on the imputation of missing data.

Based on the complete data, a study on the missing data needs to be conducted to find the method of imputing the missing data. Sukatis et al. (2019) focuses on the use of various replacement methods and chooses the best method for the diversity of complete data that occurs in imputing the missing data for air quality datasets. The purpose of this study also similar with Noor et al. (2015), to use a variety of simple basic statistical methods since these methods are uncomplicated and do not require many calculations. The methods of missing data imputations employed are basic statistical methods, namely Linear Interpolation (*L*), Top Bottom Mean (*T*), Daily Mean (*D*), 12-Hour Mean (*M*12), and 6-Hour Mean (*M*6). To compare the performance of the missing data imputation method, performance indices such as Correlation Coefficient (*C*), Mean Absolute Error (*M*), and Root Mean Square Error (*R*) were employed.

2. Methodology

2.1 CO Time Series

The time series used in this study is Petaling Jaya's CO data in 2017, from April 16 at 2 pm until April 24 at 11 am. Since the data need to be continuous, therefore, this duration is selected. The entire duration of the time series is 190 hours and was recorded in units of parts per million (ppm).The selected study locations are metropolitan areas of Petaling Jaya.

2.2 Imputation of Missing Data

This study employs missing data imputation by using the data from the air monitoring station in Petaling Jaya. From the 190 hours, the data were taken out randomly by 10% (19 data), 20% (38 data), and 30% (57 data). Next, the missing data will be imputed by the five basic statistical methods. The imputations of missing data using basic statistical methods were compared and measured according to performance index calculations to determine the best among the five methods.

2.3 Basic Statistical Methods

In this study, five basic statistical methods were used to impute the missing data for three types of missing data percentages. The basic statistical methods employed in this study are Linear Interpolation (LI), Top Bottom Mean (TBM), Daily Mean (DM), 12-Hour Mean (M12), and 6-Hour Mean (M6).

Linear Interpolation *(L)*. Linear interpolation connects two data points with a straight line. Therefore, the missing value can be calculated directly using a linear equation. The equation is written as:

$$y^* = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x^* - x_1) \qquad (1)$$

where y^* is the missing data, x^* is the time point of the missing data, x_1 and y_1 are the coordinate of the starting point of the missing data, and x_2 and y_2 are the coordinate of the final point of the missing data.

Top Bottom Mean *(T)*. This method calculates the missing data using the mean calculation of the point above and the point below the missing data. The equation is as follows:

$$y^* = \frac{y_2 + y_1}{2} \qquad (2)$$

Daily Mean (*D*). The daily mean is the mean of the observation data calculated every 24 hours from hour 1 to 336. The missing data will be imputed with the 24-hour mean data. The equation is represented as below:

$$\bar{y} = \frac{\sum_{j=1}^{2} y_j}{2} \qquad (3)$$

Borneo International Journal eISSN 2636-9826; Vol. 6 (3); 2023; 104-108
Published by *Majmuah Enterprise*
www.majmuah.com

where $\bar{y}$ is the missing data and y_j is the data of the tide height at the j^{th} hour. For example, if the missing data is at the 50th hour, then $\bar{y}$ will be imputed with the average of the complete tidal height data from the 12 hours before and after the discarded data.

12-Hour Mean (M). The missing data were calculated with the 12-hour average values. The equation is as follows:

$$\bar{y} = \frac{\sum_{j=1}^{1} y_j}{1} \quad (4)$$

6-Hour Mean (M). Similar to the method in the previous section, the 6-hour average is the average of observation data calculated every 6 hours. The equation is:

$$\bar{y} = \frac{\sum_{j=1}^{6} y_j}{6} \quad (5)$$

2.4 Performance Index

To determine the best of the five basic statistical methods, the performance index is used in this study. The imputed missing data and the original data for the missing data were compared to find the best method for imputing the missing data (Mohamed Noor et al., 2015). The three performance index methods used in this study are Correlation Coefficient (CC), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

Correlation Coefficient (C). This indicator describes the variability in the calculated data and how much it relates to the observed data. It takes a value between 0 and 1, with a value closer to 1 implying a better fit. The equation can be expressed as:

$$C = \frac{\sum_{i=1}^{N}(P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(P_i - \bar{P})^2} \cdot \sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}} \quad (6)$$

where N is the total of the imputed data, P_i is the imputed data, O_i is the original data, $\bar{P}$ is the mean of the imputed data, and $\bar{O}$ is the mean of the original data.

Mean Absolute Error (M). The M is the average difference between the calculated and observed data and is given by:

$$M = \frac{1}{N} \sum_{i=1}^{N} |P_i - O_i| \quad (7)$$

Root Mean Square Error (R). R is one of the most common methods for evaluating numerical predictions. The value is calculated with the equation:

$$R = \left(\frac{1}{N} \sum_{i=1}^{N} [P_i - O_i]^2\right)^{\frac{1}{2}} \quad (8)$$

A smaller R value indicates a better performance of the method.

3. Study Findings and Discussion

Overall, the results obtained for the missing data of 10%, 20%, and 30% of the imputed missing data are shown in **Table 1**. Linear Interpolation (L) is the best method for imputing missing data for gas pollutants CO compared to other basic statistical methods. Other methods demonstrated similar results of the correlation coefficient to the value of L , but the L method result showed the closest to the correlation value of 1.

For the 10% forecasting results, L has shown the best performance for imputing missing data compared to other methods. Meanwhile, for 20%, the most effective method is the $M6$, but for 30%, the best method is also L . Based on these three results, L is the best method, and this is supported by the study of Mohamed Noor et al. (2015) who stated that the less the average hours used, the better the results of imputing the missing data.

Table 1. Performance index results for five basic statistical methods for 10%, 20%, and 30% missing data

Method	10% Missing Data			20% Missing Data			30% Missing Data		
	C	M	R	C	M	R	C	M	R
LI	**0.993**	**0.355**	**0.942**	0.771	0.456	0.972	**0.866**	**0.480**	1.123
TBM	**0.993**	0.791	1.539	0.833	**0.437**	**0.877**	0.550	0.496	1.124
DM	0.858	1.538	2.812	0.679	0.679	1.213	0.343	0.648	1.108
M12	0.875	1.554	2.803	0.648	0.705	1.194	0.356	0.663	1.155
M6	0.983	1.086	2.091	**0.837**	0.543	0.894	0.484	0.564	**1.080**

Since the majority of the most effective performance index results are L ; for example, 10% and 30%, hence, each result needs to focus on the *CC*. As shown through 10%, the highest *CC* values were found in LI and TBM, which is 0.993. Aided by the lowest value of M and R in the L results, L shows the best result for the 10% percentage. A similar case occurred with the 30% percentage result. As for the percentage of 20%, the highest *CC* is noted at P6, which is 0.837, and the *CC* value showed a high positive C value in line with the findings of Alias et al. (2021), Jusoh et al. (2021), Shahizam et al. (2021), Ruslan et al. (2020) and Ruslan and Hamid (2019) who also predicted CO pollutants, but with different method. Therefore, the method used shows that, the imputation of CO pollutant missing data in Petaling Jaya can be forecasted using the L method.

4. Conclusions

In conclusion, basic statistical methods can be used to impute missing CO data in Petaling Jaya. Moreover, there is a difference between the Performance Index of the five basic statistical methods for imputing missing data. However, it can be concluded through this study that the best method among the five basic statistical methods for the imputation of missing CO data is Linear Interpolation (L). Thus, this missing data imputation method can be proposed to predict the time series data of other pollutants.

Acknowledgments

Our utmost appreciation to Malaysian Department of Environment (DOE) for sharing their data for the purpose of this study. This research is funded by FRGS grant with Project Code 2019-0005-102-02.

References

Agustina, D. P., Annisa, N., Riduan, R., & Prasetia, H. (2020). Konsentrasi Karbon Monoksida dan Nitrogen Dioksida pada Ruas Jalan Kuin Utara dan Kuin Selatan Kota Banjarmasin. *Urnal Tugas Akhir Mahasiswa*, *3*(1), 37–48. https://doi.org/10.20527/jernih.v3i1.480

Alias, S. N., Hamid, N. Z. A., Saleh, S. H. M., & Bidin, B. (2021). Predicting Carbon Monoxide Time Series Between Different Settlements Area in Malaysia Through Chaotic Approach. *Journal of Science and Mathematics Letter*, *9*, 45–54.

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Qual Atmos Health*, *3*, 53–64. https://doi.org/10.1007/s11869-009-0051-1

Jusoh, K. C., Hamid, N. Z. A., & Side, S. (2021). Forecasting Through the Chaotic Approach on Carbon Monoxide Time Series in Industrial Area. *Journal of Science and Mathematics Letters*, *9*, 55–62.

Malaysia, O. (2023). *Carbon monoxide poisoning : undetected*. Retrieved from https://www.osha.gov/sites/default/files/publications/carbonmonoxide-factsheet.pdf

Noor, N. M., Yahaya, A. S., Ramli, N. A., & Al Bakri Abdullah, M. M. (2015). Filling the Missing Data of Air Pollutant Concentration Using Single Imputation Methods. *Applied Mechanics and Materials*, *754–755*(March), 923–932. https://doi.org/10.4028/www.scientific.net/amm.754-755.923

Ruslan, A. B., & Hamid, N. Z. A. (2019). Application of Improved Chaotic Method in Determining Number of K-Nearest Neighbor for CO Data Series. *International Journal of Engineering and Advanced Technology*, *8*(3), 10–14. https://doi.org/10.35940/ijeat.F1003.0986S319

Ruslan, A. B., Hamid, N. Z. A., & Jusoh, K. C. (2020). Nonlinear Prediction of Carbon Monoxide Time Series in Highly Populated Area in Sabah. *Journal of Quality Measurement and Analysis*, *16*(2), 157–170.

Saeipourdizaj, P., Sarbakhsh, P., & Gholampour, A. (2021). Application of imputation methods for missing values of pm10 and o3 data: Interpolation, moving average and k-nearest neighbor methods. *Environmental Health Engineering and Management*, *8*(3), 215–226. https://doi.org/10.34172/EHEM.2021.25

Shahizam, S. I., Hamid, N. Z. A., & Side, S. (2021). Forecasting Carbon Monoxide (CO) Pollutant in Putrajaya using the Local Mean Approximation Method. *Journal of Science and Mathematics Letters*, *9*, 63–71.

Shuhaili, A. F. A., Ihsan, S. I., & Faris, W. F. (2013). Air Pollution Study of Vehicles Emission In High Volume Traffic : Selangor , Malaysia As A Case Study. *WSEAS Transactions on Systems*, *12*(2), 67–84.

Sukatis, F. F., Noor, N. M., Zakaria, N. A., Ul-Saufie, A. Z., & Suwardi, A. (2019). Estimation of missing values in air pollution dataset by using various imputation methods. *International Journal of Conservation Science*, *10*(4), 791–804.

Zainuri, N. A., Jemain, A. A., & Muda, N. (2015). A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana*, *44*(3), 449–456. https://doi.org/10.17576/jsm-2015-4403-17

Zakaria, N. A., & Noor, N. M. (2016). Imputation Methods for Filling Missing Data in Urban Air Pollution Data for Malaysia. *Urbanism*, *9*(2), 159–166.