Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Assessment of Time Series Model for Predicting Long-Interval Consecutive Missing Values in Air Quality Dataset

Daniel Bong Kim Boon[1], Norazian Mohamed Noor[1,2,*], Ahmad Zia Ul-Saufie[3], Faizal Ab Jalil[4], György Deák[5]

[1] Faculty of Civil Engineering & Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia
[2] Sustainable Environment Research Group (SERG), Centre of Excellence Geopolymer and Technology (CEGeoGTech), 02600 Arau, Perlis, Malaysia
[3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM), 40450 Shah Alam, Selangor, Malaysia
[4] Department of Environment, Kedah State, Kulim Branch, Lot 6, Jalan Hi-Tech 2/7, Kulim Hi-Tech Park, 09000 Kulim, Kedah, Malaysia
[5] National Institute for Research and Development in Environmental Protection INCDPM, Splaiul Inde-pendentei 294, 060031 Bucharest, Romania

ARTICLE INFO

Article history:
Received 24 August 2024
Received in revised form 20 November 2024
Accepted 7 April 2025
Available online 25 April 2025

Keywords:
Air pollutant dataset; missing data; imputation method; time series analysis; ARIMA

ABSTRACT

Air pollutant concentration in Malaysia is continuously monitored using the Continuous Ambient Air Quality Machine (CAAQM). During the observation phase by CAAQM, some air pollutant datasets were detected missing due to machine failure, maintenance, position changes and human error. Incomplete datasets especially with the longer gaps of consecutive missing observation may lead to several significant problems including loss of efficiency, difficulties in using some computational software and bias estimation due to differences of observed and predicted dataset. This study aim evaluates the performance of the time series method i.e. Auto Regression Integrated Moving Average (ARIMA) for filling long hours of missing data in an air pollution dataset. The dataset of PM_{10}, SO_2, NO_2, O_3, CO, wind speed, relative humidity and ambient temperature for Pegoh and Kota Kinabalu in 2018 were used for analysis. Monte Carlo Markov Chain (MCMC) and Expectation-Maximization (EM) were employed to compare with ARIMA's effectiveness in filling the simulated missing gaps in air quality dataset. Existing missing data in the raw data were pre-treated and then simulated into 5%, 10% and 15% of missing data ranging from 24-hour to 120-hour intervals. The performance of the imputation approach was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Prediction Accuracy (PA) and Index of Agreement (IA). Overall, the Expectation-Maximization technique was selected the most effective at filling in simulated long gaps of missing data of air pollutant dataset with the range of IA from 0.74 to 0.77. In contrast, the ARIMA approach performed poorly in this research with range of IA value of 0.44 to 0.48. This was because of it requires past time-series data to generalize a forecast or impute missing data, hence, the forecast becomes a straight line and performed poorly at predicting series with long hours of missing observation.

* *Corresponding author.*
E-mail address: norazian@unimap.edu.my

https://doi.org/10.37934/ard.127.1.1631

1. Introduction

Ambient air quality monitoring station is set to monitor ambient air quality; if there is a significant change in air quality level, the public should be alerted. In Malaysia, Department of Environment, Malaysia has setup 65 Continuous Air Quality Monitoring Stations (CAAQMS) and these stations require routine maintenance to make sure that the accuracy of the measured air pollutant concentration are reliable. However, maintenance procedures will render the station's air pollution data incomplete [1]. According to Guarnaccia *et al.,* [2], air pollution data might be missing due to excessive uncontrolled factors, such as device malfunction, maintenance or repair and calibration. As complete and continuous data are required for numerous mathematical analyses, such as time series, principal component analysis (PCA) and multivariate analysis, missing data might be troublesome [2-4].

Accurate prediction has always been hampered by missing values in the dataset. This may result in a misleading understanding of the air pollution scenario. Each problem may have a limited number of incomplete solutions; however, the missing details may vary [5]. Simple missing data or short gaps missing data can be easily treated using a simple imputation method. However, treating the data for long gaps in missing data is more complex. More missing data mean more critical data is absent from the data and such gaps are often the only sign of a massive change in the data series. Other possible consequences of long gaps in missing data include a decrease in the sample size and statistical power and a decrease in the precision and accuracy of parameter estimations [6]. Loss of accuracy results in incorrect conclusions or skewed judgments regarding outcomes and relationships of interest decreased precision of estimates resulting in decreased performance of confidence intervals and rising standard errors. Therefore, randomly assuming the long gaps of missing data appears risky.

The most common way to deal with missing data in a dataset is to delete those data points, also called listwise deletion. According to Little *et al.,* [7], eliminating missing values using the deletion approach might add considerable bias to the study. In addition to the deletion approach, single imputation fills in the exact value for each missing item with single imputation. Only one estimate is substituted for each missing item in single imputation [8]. Mean substitution is a common single imputation technique because of its simplicity. The mean imputation method estimates the missing value by substituting it with the mean value of each variable on the relevant missing variables [9]. Moshenberg *et al.*, [1] stated that the mean imputation understates the variance in the dataset and can change any other chemometric study based on the dataset. Furthermore, this strategy might result in bias and huge inaccuracies.

In other hand, Multiple Imputation (MI) imputed every missing observation in the dataset multiple to build a complete dataset. MI was developed in response to the limitations of single imputation approaches for dealing with nonresponse in surveys, where the inferences based on MI enable us to reflect the uncertainty in the missing data [7]. Though, MI is not limited to survey analysis and may be used in any scenario to impute missing data [10]. There are various MI methods available, but the Markov Chain Monte Carlo (MCMC) approach is the most used in statistical software.

Another well-known method is Expectation Maximize (EM); it is common practice in data analysis to utilize EM as a means of dealing with missing data. Indeed, EM overcomes some of the limitations of other techniques, such as mean substitution or regression substitution [11]. Numerous imputation methods can fill in missing data in air pollution datasets [12]. Moshenberg *et al.,* [1] stated that the missing gaps duration and the type of study must be considered while selecting the best imputation method. Therefore, this work attempt on applying time series methods to fill in the long gaps of simulated missing observations in an air quality dataset. The performance of the time series method was evaluated and compared with MI and EM using performance measure. Reliable imputation

method is important in ensuring high quality dataset, hence further analysis carried out using the dataset are not bias.

2. Methodology

2.1 Air Pollutant Dataset

In this study, hourly measurement records of eight parameters of air quality in 2018 were acquired from the Department of Environment, Malaysia. There were two locations selected for this study i.e. Pegoh, Perak and Kota Kinabalu, Sabah. Table 1 describes the air pollutants parameters with the measurement unit.

Table 1
Air pollutants parameters and the measurement units

Parameters	Symbol	Unit
Particulate Matter	PM_{10}	$\mu g/m^3$
Sulphur Dioxide	SO_2	ppm
Nitrogen Dioxide	NO_2	ppm
Ozone	O_3	ppm
Carbon monoxide	CO	ppm
Windspeed	WS	m/s
Relative Humidity	RH	%
Ambient Temperature	T	°C

2.2 Data Pre-Treatment

In this study, the raw data was undergone pre-treatment first due to missing data in the dataset. This raw data must be treated first to obtain complete data before simulating the data into three different percentages of missingness. This raw data was treated using linear interpolation. In most air pollution data, linear interpolation is the most common imputation method to treat or impute the short gaps in missing data in the air pollution dataset [9]. Linear interpolation means estimating a missing value by connecting points in ascending order on a straight line. The linear interpolation function's Eq. (1) is [13]:

$$f(x) = f(x)_0 + \frac{f(x)_1 - f(x)_0}{x_1 - x_0}(x - x_0) \quad (1)$$

2.3 Simulation of Missing Data

The simulation of missing data mainly aims to investigate the efficiency of the time series method used in this study. The simulation data were divided into three groups of missing percentages: 5, 10 and 15%. The simulated dataset was used to compare the performances of the three individual imputation methods. The range of missing hours used in this study is (24 hours < L < 120 hours) which considered as long interval of missing observations in Malaysia air quality dataset [14]. The real goal is to assess time series method in imputing long-hour missing data gaps. Other methods such as Expectation Maximize and Multiple Imputation - Markov Chain Monte Carlo were used to compare the efficiency of time series method i.e. ARIMA. The simulations of the missing data are performed using the statistical software SPSS, E-views and Microsoft Excel for Windows.

2.4 Imputation Method

2.4.1 Time series method - ARIMA

Autoregressive Integrated Moving Average (ARIMA) is one of the most used time series forecasting model. Using the Box-Jenkins ARIMA model, the forecasting trend was generated by modelling the time series behaviour. ARIMA is based on the idea that the information in the past values of the time series can alone be used to predict the future values [15]. A four-step iterative procedure was utilized in this study's methodology. In the first stage, the historical data are utilized to identify an appropriate Box-Jenkins model provisionally. It is then followed by the estimation of the model's parameters. After that, a diagnostic check must be performed to ensure that the detected model is adequate before deciding on the finest one. If the model is insufficient, a new one should be found. It is then used to calculate the value of a time series forecast [16].

It is possible to anticipate the next few points in time using the AR(p) model because of the correlation between time series variables, which is why this model is named AR(p) [17].

$$Y_t = \epsilon_t + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \quad (2)$$

AR model can be replaced with the lower order MA(q) model in this case [15]:

$$Y_t = \epsilon_t + \beta_1 Y_{t-1} \quad (3)$$

It is essential to determine the best ARIMA order (p, d, q) and seasonal ARIMA order for air pollution prediction (P, D, Q, S). Using the grid search approach, different combinations of parameters can be tested iteratively. Firstly, the time series dataset was checked whether or not the time series is stationary. Time series line graph, scatter plot, autocorrelation function and partial autocorrelation function graphs are utilised to determine stationarity [18]. Typically, the unit root of Augmented Dickey-Fuller (ADF) is used to assess the variance, trend and seasonal variation and identify stationarity. The ARIMA model's autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) are determined using the autocorrelation order “p” and moving average order “q”. The autocorrelation and partial autocorrelation graphs are used to determine the number of autocorrelation coefficients and partial autocorrelation coefficients with a degree of statistical significance that is highly significant. The approximate sequence model can be chosen in this phase—tests for diagnosis and improvement [18]:

i. Selecting the model with the most significant terms (p-values 0.05) is necessary to ensure the optimal model selection.
ii. SigmaSQ: a volatility measure. The smallest option was chosen.
iii. Log Likelihood: Since we are trying to maximise the log-likelihood function, the largest number was chosen. (The largest value is the least negative)
iv. Model selection criteria: The design with the smallest Akaike, Schwarz and Hannan-Quinn was chosen.

If the model does not satisfy the step's requirements, reselect, generate several models and select the optimal model from all the test-fitted models.

2.4.2 Expectation-maximization (EM)

The EM approach replaces missing data with the value obtained from estimating the parameters of an incomplete data set by maximizing the probability of known data. EM requires the training dataset to be completed, e.g. all relevant interacting random variables are present. This method consists of two steps: prediction and estimation by iterative calculation [19]. Given data X and initial parameter $\ominus^t$. For completing the "Expectation" part, assume a hidden variable Y, where Y is distributed on current knowledge (X and $\ominus^t$ i.e. $p(Y|X,\ominus^t)$). The expectation of the joint likelihood under this distribution (Q function) was computed according to this [19]:

The E-Step: $Q(\theta^t, \theta^{t+1}) = E_{Y|X,\theta^t}[L(X,Y|\theta^{t+1})]$ (4)

The conditional expectation w.r.t. a random variable is a function on the range of Y, which is the desired parameter to be determined i.e. $\ominus^{t+1}$ was maximize until the iteration of E-Step and M-Step were converged.

M-Step: $\ominus^{t+1} = argmax_{\ominus}[Q^t(\ominus) + logP(\ominus)]$ (5)

In this study, IBM SPSS Version 22 was used to perform EM following the procedures as listed below:

i. The mean, variance and covariance are estimated from the whole data on an individual basis.
ii. Maximum likelihood algorithms are used to estimate regression equations that tie each variable and construct the formula. The formula is used to estimate missing values.

2.4.3 Markov Chain Monte Carlo (MCMC)

Multiple imputation was performed using the MCMC method due to the assumption of multivariate normality. MCMC is a sequence of random variables whose distributions depend on the value of the previous variable [20]. MCMC is a simulation technique that can be used to determine and sample from the posterior function. A Markov Chain is a stochastic process that generate random variables, X_1, X_2, …, X_t where the distribution [20]:

$$P(X_t|X_1, X_2, \cdots, X_{t-1}) = (X_t|X_{t-1}) \quad (6)$$

i.e. the distribution of the next random variables depends only on the current random variable. X_i is typically highly correlated; thus, each sample is not an independent draw from the posterior. In this study, MCMC was computed using IBM SPSS software (Vers. 22).

2.5 Performance Indicators

The imputation methods were evaluated using four performance indicators namely Prediction Accuracy (PA), Index of Agreement (IA), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Table 2 shows the formulae of the performance indicators and their best fit.

Table 2
Performance indicators [6]

Performance Indicator	Formula	Best Fit
Prediction Accuracy (PA)	$PA = \sum_{i=1}^{N} \frac{[(P_i - \dot{P})(O_i - \bar{O})}{(N-1)\sigma_p \sigma_o}$	Close to 1
Index of Agreement (IA)	$IA = 1 - \left[\frac{\sum_{i=N}^{N}(P_i - O_i)^2}{\sum_{i=N}^{N}(\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}\right]$	Close to 1
Mean Absolute Error (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$	Close to 0
Root Mean Square Error (RMSE)	$RMSE = \sqrt{(\frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert)}$	Close to 0

Where,
N = Number of imputations
O_i = Observed data points
P_i = Imputed data points
$\dot{P}$ = Average of imputed data
$\bar{O}$ = Average of observed data
σ_o = Population standard deviation of the observed data
σ_p = Population standard deviation of the imputed data

3. Results

3.1 Air Pollutant Dataset

Tables 3 and 4 shows the percentages of missing data in air pollutant dataset in Pegoh and Kota Kinabalu respectively. From Tables 3 and 4, the majority of missing gaps was 1 hour for both Pegoh and Kota Kinabalu, constituting 23.78 and 18.45% of the overall percentages of missing observation for Pegoh and Kota Kinabalu, respectively. The longest gap for Pegoh was 24 hours and for Kota Kinabalu, was 31 hours.

Table 3
Percentages of missing data gaps for Pegoh, Perak

Length of Gap (Hour)	Percentage of missing data (%)								Total Percentage (%)
	PM_{10}	SO_2	NO_2	O_3	CO	WS	RH	AT	
1	22.2	68.7	78.0	77.3	73.1	9.6	2.2	2.9	23.8
2	6.9	18.9	2.2	4.4	4.4	1.9	2.2	1.4	3.0
3 – 9	38.9	35.9	14.5	12.9	16.8	12.5	10.9	16.4	15.9
10 – 17	9.0	10.4	3.2	5.5	5.7	30.8	34.8	32.9	26.5
18 – 24	13.2	-	-	-	-	41.4	45.7	30.0	30.8
Total									100

Table 4
Percentages of missing data gaps for Kota Kinabalu, Sabah

Length of Gap (Hours)	Percentage of missing data (%)								Total Percentage (%)
	PM_{10}	SO_2	NO_2	O_3	CO	WS	RH	AT	
1	22.6	7.2	78.1	95.5	54.6	11.8	26.1	4.5	18.5
2	19.6	61.1	10.0	1.4	6.5	2.6	-	-	8.3
3 – 12	50.0	20.4	11.9	2.9	19.3	63.2	73.9	11.0	31.0
13 – 22	11.8	3.6	-	-	8.4	-	-	22.6	21.3
23 – 31	10.5	5.1	-	-	11.9	-	-	61.9	20.9
Total									100

3.2 Simulated Missing Data

The percentage of simulated missing gaps for Pegoh and Kota Kinabalu's simulated are shown in Table 5. The gaps in the simulated missing data ranged from 24 hours to 120 hours and the percentage of missing data was simulated as 5, 10 and 15%. For 5% simulated missing data pattern in Pegoh, the most extensive distribution of missing gaps was roughly 35.97% of mean gaps for 72 to 96 hours, while the lowest distribution was 16.31% for 96 to 120 hours. For 10% simulated missing data patterns in Pegoh, the largest percentage of missing gaps was 36.88% for 72 to 96 hours of missing data and the lowest percentage was 13.23% for 24 to 48 hours of missing data. For 15% simulated missing data in Pegoh, the largest percentage of simulated missing data is 39.89% at 96 to 120 hours, while the lowest percentage is 15.23% for 24 to 48 hours. The distribution of gaps (5%, 10% and 15%) for the overall percentages of missing data for Pegoh and Kota Kinabalu was generally identical. The biggest distribution of missing gaps for Pegoh was around 31.94% of total percentages for the 72 to 96 hours gap and approximately 16.68% for the 24 to 48 hours gap. Kota Kinabalu had the highest and lowest percentages of simulated missing gaps of 33.67% and 15.82% for 96 to 120 hours, respectively.

Table 5
Percentages of length of gap simulated missing data in Pegoh and Kota Kinabalu

Length of Gap (Hour)	Place	Percentage of Simulated Missing Data			Mean (%)	Total Percentage (%)
		5%	10%	15%		
24 ≤ L ≤ 48	P	22.78	13.23	14.04	15.23	16.68
	KK	18.84	16.46	12.16	14.95	15.82
48 < L ≤ 72	P	24.94	16.21	23.10	21.11	21.42
	KK	21.03	21.48	20.38	20.89	20.96
72 < L ≤ 96	P	35.97	36.88	22.96	29.77	31.94
	KK	30.13	30.99	27.53	29.23	29.55
96 < L ≤ 120	P	16.31	33.68	39.89	33.89	29.96
	KK	30.00	31.08	39.93	34.93	33.67
Total	P	100	100	100	100	100
	KK	100	100	100	100	100

Note: P = Pegoh, Perak and KK = Kota Kinabalu, Sabah

3.3 ARIMA Model

Everything outside the Partial Auto-Correlation Function (PACF) plot's perimeter or boundary indicates the order of the Auto Regression (AR) model. Usually, AR "p" has a fixed value throughout the data. In this study, all AR values are 1, as shown in Table 6. For the value of Integrated (I) “d”, it is always constant which value 1 throughout the series. Moving average (MA) is similar to choosing "p" for the AR model. In order to determine the correct "q" order for the MA model, all values outside the boundary must be analysed. Unlike the AR model, we may pick the order "q" for the MA (q) model from the ACF if this plot has a sharp cut-off after lag "q." The PACF plot decays more slowly, which is evidence of the MA process [15].

Analysing the Auto-Correlation Function (ACF) and PACF graphs to determine the proper "p" and "q" sequence for the ARIMA model can be tedious and challenging. An objective function that measures model performance on cross-validation is necessary to discover the best possible combination of p and q for a given situation [15].

Table 6
Value of "AR", "I" and "MA" based on ACF and PACF

Percentage of the Simulated Missing Data	Parameter	Pegoh			Kota Kinabalu		
		AR "p"	I "d"	MA "q"	AR "p"	I "d"	MA "q"
5%	PM_{10}	1	1	4	1	1	8
	SO_2	1	1	6	1	1	8
	NO_2	1	1	4	1	1	6
	O_3	1	1	4	1	1	4
	CO	1	1	7	1	1	5
	WS	1	1	9	1	1	12
	RH	1	1	10	1	1	9
	AT	1	1	3	1	1	3
10%	PM_{10}	1	1	4	1	1	7
	SO_2	1	1	7	1	1	8
	NO_2	1	1	3	1	1	8
	O_3	1	1	4	1	1	5
	CO	1	1	7	1	1	7
	WS	1	1	8	1	1	10
	RH	1	1	9	1	1	7
	AT	1	1	5	1	1	2
15%	PM_{10}	1	1	5	1	1	6
	SO_2	1	1	9	1	1	9
	NO_2	1	1	5	1	1	11
	O_3	1	1	5	1	1	4
	CO	1	1	5	1	1	4
	WS	1	1	7	1	1	7
	RH	1	1	10	1	1	6
	AT	1	1	10	1	1	4

3.4 Performances of the Imputation Methods

Table 7 illustrates the performance indicators of the three imputation methods (EM, MCMC and ARIMA) for 5, 10 and 15% simulated missing data in Pegoh and Kota Kinabalu. For 5% missing data, index of agreement (IA) calculated for EM reveals a greater performance compared to other methods with the values of IA ranging from 0.48 to 0.94. Meteorological data (Wind Speed, Relative Humidity and Ambient Temperature) has the highest value of IA compared to other parameters. IA value for ambient temperature is the highest compared to the other air pollutants. ARIMA performed better for filling the 5%-simulated missing dataset of SO_2 data.

EM was selected as the best way to fill in the 10%-simulated missing dataset for Pegoh and Kota Kinabalu (Table 5). It can be proved by the calculated values of Mean Absolute Error (MAE) that indicated lesser error value in the EM imputation approach ranging from 0 to 9.5 if compared to other methods. ARIMA shows the moderate performance for predicting 10%-simulated missing dataset with the range of MAE values of 0 to 13.1. MCMC performs slightly less accurate compared to ARIMA with the values of MAE ranging from 0 to 15.7. However, if comparing the values of IA for MCMC and ARIMA, MCMC outperformed with the range of 0.41 to 0.93 whereas for ARIMA, the values ranging from 0.29 to 0.79.

Meanwhile for 15%-simulated missing data, the Root Mean Square Error (RMSE) values for EM are the lowest compared to other imputation methods. Generally, higher error was calculated for higher percentage of simulated missing data for all imputation methods. EM gives the least error with the range of RMSE of 0 to 17.1. MCMC and ARIMA perform moderately with the values of RMSE

ranging from 0 to 21. This result was similar to Sukatis *et al.,* [21] that reported EM was the best method for filling both short and long gaps of missing dataset for air pollutant measurement records.

Figure 1 shows a ranking model for all imputation methods based on calculated Prediction Accuracy (PA) values for all parameters. From the figure, it was clearly seen EM method is the most effective imputation methodology for all simulated missing dataset i.e. 5, 10 and 15%. Though, In Pegoh, ARIMA ranks as the top method to impute SO_2 for 5%-simulate missing data. For 10% simulated missing data in Pegoh and Kota Kinabalu, the ranking model again shows the Time Series method - ARIMA ranked as the worst imputation method among the three imputation methods to impute long gaps in missing data in air pollution data. MCMC method was listed as the second most effective imputation method for air pollution data for Pegoh and Kota Kinabalu. According to research by Junninen *et al.,* [12], the MCMC approach fills in missing data by averaging or combining many simulated values. Applying Bayesian inference and repeating multiple phases, such as the imputation I-step and posterior P-step, were required to complete this process. These intricate techniques would be time-consuming but yield reasonable estimates of missing data. For 15% of simulated missing data for Pegoh and Kota Kinabalu, the rank of the best prediction method shows the same trend 5% and 10%-simulated missing dataset where EM outperformed other imputation methods. MCMC shows moderate performances for almost all air pollutant parameters whereas ARIMA performed least accurate in predicting simulated missing observation for all parameters except for SO_2 and wind speed.

Table 7
The results of performance indicators for 5, 10 and 15% simulated missing data in Pegoh and Kota Kinabalu

%	Method	PI	PM_{10} (µg/m^3)		SO_2 (ppm)		NO_2 (ppm)		O_3 (ppm)		CO (ppm)		WS (m/s)		RH (%)		AT (°C)	
			Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK
5%	EM	PA	0.408	0.361	0.145	0.399	0.620	0.723	0.826	0.879	0.637	0.839	0.195	0.716	0.725	0.894	0.900	0.937
		IA	0.585	0.477	0.644	0.588	0.754	0.831	0.884	0.933	0.744	0.891	0.257	0.830	0.837	0.940	0.944	0.961
		MAE	10.412	8.944	0.000	0.000	0.003	0.002	0.008	0.003	0.128	0.146	0.545	0.391	7.761	3.712	1.144	0.935
		RMSE	13.982	15.888	0.000	0.000	0.004	0.003	0.010	0.005	0.163	0.180	0.685	0.522	9.268	4.898	1.438	1.221
	MCMC	PA	0.261	0.113	0.121	0.249	0.458	0.554	0.727	0.742	0.384	0.747	0.110	0.460	0.644	0.825	0.787	0.870
		IA	0.549	0.387	0.450	0.528	0.673	0.727	0.849	0.859	0.624	0.861	0.467	0.688	0.798	0.907	0.883	0.930
		MAE	13.991	11.970	0.000	0.000	0.004	0.003	0.009	0.005	0.176	0.171	0.696	0.594	9.375	4.562	1.712	1.323
		RMSE	18.069	19.496	0.001	0.001	0.005	0.004	0.013	0.007	0.231	0.223	0.885	0.830	11.635	6.362	2.132	1.681
	ARIMA	PA	-0.168	0.196	0.628	0.264	-0.044	-0.138	-0.019	0.059	0.091	0.527	-0.035	0.005	0.030	0.051	0.180	0.122
		IA	0.347	0.432	0.780	0.549	0.401	0.232	0.414	0.466	0.058	0.719	0.356	0.421	0.423	0.450	0.500	0.508
		MAE	15.829	13.127	0.000	0.000	0.004	0.005	0.019	0.010	5.123	0.227	0.667	0.770	13.308	10.712	2.842	3.365
		RMSE	21.163	19.389	0.000	0.000	0.006	0.007	0.025	0.013	5.377	0.285	0.833	0.967	16.568	13.055	3.634	4.137
10%	EM	PA	0.498	0.507	0.519	0.594	0.638	0.678	0.850	0.876	0.602	0.742	0.196	0.741	0.874	0.926	0.916	0.930
		IA	0.639	0.627	0.557	0.710	0.763	0.740	0.913	0.927	0.716	0.835	0.260	0.814	0.929	0.958	0.955	0.963
		MAE	9.489	6.848	0.000	0.000	0.003	0.003	0.007	0.004	0.124	0.162	0.501	0.416	4.928	3.242	1.053	0.930
		RMSE	12.699	11.838	0.001	0.000	0.004	0.004	0.009	0.005	0.182	0.209	0.637	0.617	6.409	4.397	1.283	1.162
	MCMC	PA	0.276	0.206	0.271	0.372	0.408	0.455	0.730	0.774	0.413	0.556	0.037	0.503	0.750	0.851	0.800	0.860
		IA	0.558	0.457	0.545	0.617	0.644	0.669	0.852	0.875	0.640	0.744	0.412	0.704	0.863	0.920	0.891	0.925
		MAE	13.353	15.701	0.000	0.000	0.004	0.004	0.009	0.006	0.177	0.229	0.716	0.705	7.320	4.917	1.662	1.338
		RMSE	17.432	20.716	0.001	0.001	0.005	0.005	0.012	0.007	0.244	0.287	0.914	0.919	9.408	6.284	2.083	1.693
	ARIMA	PA	0.038	0.206	0.216	0.151	0.045	0.218	0.063	0.096	0.092	0.663	-0.089	0.035	0.008	0.076	0.129	-0.036
		IA	0.400	0.466	0.480	0.496	0.437	0.521	0.467	0.454	0.411	0.791	0.290	0.402	0.402	0.481	0.449	0.416
		MAE	13.062	11.020	0.000	0.000	0.005	0.004	0.017	0.010	0.199	0.190	0.608	0.882	13.295	11.524	2.712	3.177
		RMSE	17.040	16.132	0.001	0.001	0.007	0.006	0.022	0.013	0.279	0.234	0.756	1.155	17.167	14.379	3.691	3.927
15%	EM	PA	0.464	0.427	0.506	0.502	0.684	0.676	0.828	0.846	0.516	0.772	0.184	0.662	0.803	0.888	0.835	0.895
		IA	0.571	0.430	0.628	0.650	0.786	0.782	0.897	0.916	0.655	0.844	0.264	0.763	0.882	0.939	0.906	0.941
		MAE	10.006	7.934	0.000	0.000	0.003	0.003	0.007	0.004	0.123	0.141	0.498	0.449	6.653	3.973	1.467	1.064
		RMSE	13.734	17.123	0.000	0.000	0.004	0.007	0.009	0.005	0.175	0.185	0.623	0.665	8.208	5.214	1.787	1.366
	MCMC	PA	0.225	0.161	0.182	0.371	0.462	0.479	0.686	0.718	0.246	0.620	0.063	0.463	0.669	0.807	0.706	0.807

	IA	0.518	0.368	0.503	0.619	0.681	0.683	0.826	0.845	0.531	0.786	0.428	0.683	0.814	0.897	0.838	0.896
	MAE	14.083	11.071	0.000	0.000	0.004	0.003	0.010	0.006	0.182	0.195	0.678	0.615	8.541	5.234	1.956	1.442
	RMSE	18.258	21.019	0.001	0.001	0.005	0.005	0.014	0.008	0.248	0.254	0.856	0.903	11.116	7.071	2.482	1.913
ARIMA	PA	0.163	0.022	0.436	0.413	0.008	0.040	0.112	0.072	0.045	0.678	0.016	0.072	0.126	0.008	0.191	0.212
	IA	0.495	0.249	0.663	0.621	0.346	0.392	0.463	0.461	0.412	0.808	0.400	0.428	0.471	0.438	0.504	0.543
	MAE	15.366	9.899	0.000	0.000	0.004	0.004	0.016	0.010	0.191	0.171	0.646	0.910	13.399	11.354	2.899	2.831
	RMSE	20.236	20.161	0.000	0.000	0.006	0.005	0.020	0.013	0.246	0.215	0.826	1.179	16.537	14.919	3.608	3.433

Fig. 1. The ranking of all imputation methods for 5, 10 and 15% simulated missing data in Pegoh and Kota Kinabalu based on Prediction Accuracy. 1 = PM_{10}; 2 = SO_2; 3 = NO_2; 4 = O_3; 5 = CO; 6 = Wind Speed; 7 = Relative humidity; 8 = Atmospheric temperature

4. Discussions

Table 8 displays the average values of all performance indicators for all simulated missing dataset in Pegoh and Kota Kinabalu. The best imputation approach for filling in long gaps of missing data in air pollution was the EM method, followed by the MCMC method and then the ARIMA method.

Across all percentages of simulated missing data, the EM method was the best imputation technique for filling in a long gap of missing values in the air pollution dataset. This is demonstrated when all the performance indicators for each percentage of simulated missing data concurred that this imputation method was the most effective. Even for datasets with long-missing hour gaps, this technique's performance was deemed exceptional. This conclusion is consistent with what Abd Razak *et al.,* [22] observed: the EM approach performed exceptionally well despite the large percentages of missing values. Using the EM approach, both Pegoh and Kota Kinabalu can demonstrate good performance despite a long interval and a significant proportion of missing data simulations.

The MCMC approach was the second-best imputation method for computing simulated missing data. This technique performed exceptionally well, even as the proportion of simulated missing data increased with more missing data in the dataset. Junninen *et al.,* [12] have indicated that MCMC is the ideal approach for imputation due to its complicated procedure that may reflect the uncertainty

associated with missing data. In this investigation, however, EM outperformed MCMC due to the linear relationship between missing data and accessible data in air pollution [1,12,23,24]. This is because the primary assumption of the EM approach is that the missing data has a linear relationship with the available data, such as time-series data [10]. Air pollution data is one example of a time series.

The ARIMA method performed least accurate for all percentages of the simulated missing dataset. It requires just the past time-series data to generalize a forecast or impute missing data. The classic model identification methods for finding the proper model from the class of alternative models are typically challenging to comprehend and computationally costly. Firstly, this method is also subjective and the predictor's skill and experience might influence the model's accuracy. Second, the underlying theoretical model and structural links are not as different as they are in particular straightforward for imputation methods, such as EM and MCMC. Moreover, ARIMA models, like other imputation methods, are fundamental "retrospective," which let the past predict the future [2]. Therefore, in the long run, the forecast becomes a straight line and is not very good at predicting series with turning points.

Table 8
The average of all performance indicators of all simulated missing data

Method	PI	5%	10%	15%	Average
EM	PA	0.638	0.693	0.656	0.662
	IA	0.756	0.769	0.741	0.755
	MAE	2.133	1.732	2.020	1.962
	RMSE	3.017	2.466	3.069	2.851
MCMC	PA	0.503	0.516	0.479	0.500
	IA	0.699	0.707	0.682	0.696
	MAE	2.787	2.884	2.751	2.807
	RMSE	3.848	3.751	4.010	3.870
ARIMA	PA	0.109	0.119	0.163	0.131
	IA	0.441	0.460	0.481	0.461
	MAE	4.126	3.544	3.606	3.759
	RMSE	5.341	4.676	5.088	5.035

Figure 2 presents the scatter plot of the observed and predicted data using the EM method for all parameters to predict missing observations of 15%-simulated missing data in Pegoh. The closer the value of R^2 to 1, the more significant the correlation between the expected and observed data [25]. Overall, R^2 values for all Pegoh observations were quite near 1. This suggested that the expected and actual values in Pegoh were virtually identical. This study determined that the EM approach is better than other methods for air pollution datasets with long-missing hours. This indicates that the EM method is a superior method and the most stable among other methods.

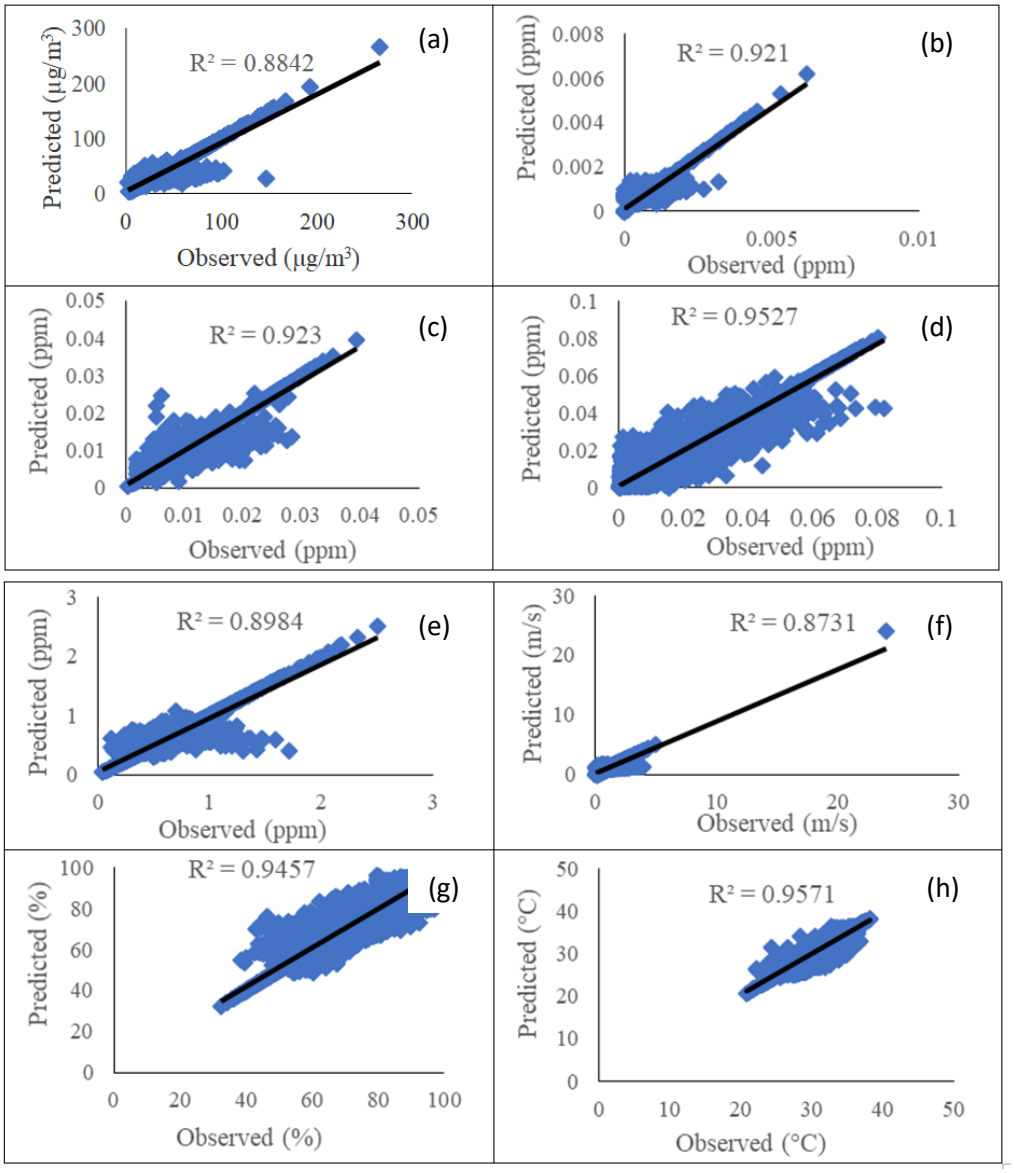

Fig. 2. Scatter plot of observed vs predicted data (using EM) in Pegoh for 15% simulated missing data; (a) PM_{10} (b) SO_2 (c) NO_2 (d) O_3 (e) CO (f) wind speed (g) relative humidity (h) ambient temperature

4. Conclusions

In this study, Auto-Regression Integrated Moving Average (ARIMA), Expectation-Maximization (EM) and Markov Chain Monte Carlo were used as imputation methods (MCMC) for three percentages of simulated missing dataset i.e. 5, 10 and 15%. Four performance indicators, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Prediction Accuracy (PA) and Index of Agreement (IA), were used to describe how well each of these imputation methods fit the simulated missing dataset. The main purpose of this is to evaluate the performances of the time series method i.e. ARIMA to fill in the long gap of missing observations in air pollution dataset. Most of the time, the EM method was chosen as the best method to fill in the long gaps of simulated missing data in the air quality monitoring dataset. Compared to EM and MCMC, ARIMA was selected as the least performed method in estimating the long gap missing observations in air pollutant dataset. This study has the potential to be improved in the future by:

i. Model identification: at this stage, particularly for nonstationary models, the data must undergo differencing to become stationary and look for opportunities to expand the sample size (use at least four years of data).
ii. Estimation is the process of choosing models that are as simple as possible. Here, fully utilize the Correlogram of the different data. Do not overlook the ACF and PACF in the model. Consider any lag of the ACF and PACF that falls outside of the 95% confidence bound to determine the AR and MA values in the model.
iii. Diagnostics: To check to see if the models are still beneficial. The best fit model is most likely the one with the lowest Akaike Information Criterion (AIC) and the highest Log-Likelihood values. Also, the residuals correlation chart must be stable.
iv. If forecasts cannot be compared to actual data, there is a problem with the modelling procedure. To make a non-stationary series into stationary data, it must be differentiated. For the ARIMA model to determine the order, a correlogram of the difference must be used. The predictions should then be consistent with the observed data in some manner.
v. Before initiating the simulation and imputation process, maybe consider replacing or removing the outliers and extreme outliers with suitable values to improve the performance of the EM and MCMC methods. This is necessary because any abnormality present in the dataset will impact the performance of the EM and MCMC methods.

Acknowledgement
The authors would like to thank Department of Environment Malaysia for the air pollutants dataset.

References

[1] Moshenberg, Shai, Uri Lerner and Barak Fishbain. "Spectral methods for imputation of missing air quality data." *Environmental Systems Research* 4 (2015): 1-13. https://doi.org/10.1186/s40068-015-0052-z
[2] Guarnaccia, Claudio, Julia Griselda Ceron Breton, Rosa Maria Ceron Breton, Carmine Tepedino, Joseph Quartieri and Nikos E. Mastorakis. "ARIMA models application to air pollution data in Monterrey, Mexico." In *AIP Conference Proceedings*, vol. 1982, no. 1. AIP Publishing, 2018. https://doi.org/10.1063/1.5045447
[3] Azizan, F. L., S. Sathasivam, M. Velavan, N. R. Azri and N. I. R. A. Manaf. "Prediction of drug concentration in human bloodstream using Adams-Bashforth-Moulton method." *J Adv Res Appl Sci Eng Technol* 29, no. 2 (2023): 53-71. https://doi.org/10.37934/araset.29.2.5371
[4] Mohamad, Muhammad Arif and Muhammad Aliif Ahmad. "Handwritten Character Recognition using Enhanced Artificial Neural Network." *Journal of Advanced Research in Computing and Applications* 36, no. 1 (2024): 1-9. https://doi.org/10.37934/arca.36.1.19
[5] Libasin, Zuraira, Wan Suhailah Wan Mohamed Fauzi, Nur Azimah Idris and Noor Azizah Mazeni. "Evaluation of Single Missing Value Imputation Techniques for Incomplete Air Particulates Matter (PM 10) Data in Malaysia." *Pertanika Journal of Science & Technology* 29, no. 4 (2021). https://doi.org/10.47836/pjst.29.4.46
[6] Kellermann, Anh Pham. "Missing Data in Complex Sample Surveys: Impact of Deletion and Imputation Treatments on Point and Interval Parameter Estimates." PhD diss., University of South Florida, 2018.
[7] Little, Roderick JA and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019. https://doi.org/10.1002/9781119482260
[8] Norazian, Mohamed Noor, Yahaya Ahmad Shukri, Ramli Nor Azam and Abdullah Mohd Mustafa Al Bakri. "Estimation of missing values in air pollution data using single imputation techniques." *ScienceAsia* 34, no. 3 (2008): 341-345. https://doi.org/10.2306/scienceasia1513-1874.2008.34.341
[9] Noor, Norazian Mohamed, Mohd Mustafa Al Bakri Abdullah, Ahmad Shukri Yahaya and Nor Azam Ramli. "Comparison of linear interpolation method and mean method to replace the missing values in environmental data set." In *Materials science forum*, vol. 803, pp. 278-281. Trans Tech Publications Ltd, 2015. https://doi.org/10.4028/www.scientific.net/MSF.803.278
[10] Rubin, Donald B. "Inference and missing data." *Biometrika* 63, no. 3 (1976): 581-592. https://doi.org/10.1093/biomet/63.3.581
[11] Dong, Yiran and Chao-Ying Joanne Peng. "Principled missing data methods for researchers." *SpringerPlus* 2 (2013): 1-17. https://doi.org/10.1186/2193-1801-2-222

[12] Junninen, Heikki, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen and Mikko Kolehmainen. "Methods for imputation of missing values in air quality data sets." *Atmospheric environment* 38, no. 18 (2004): 2895-2907. https://doi.org/10.1016/j.atmosenv.2004.02.026
[13] Chapra, Steven C. and Raymond P. Canale. *Numerical methods for engineers*. Vol. 1221. New York: Mcgraw-hill, 2011.
[14] Zakaria, Nur Afiqah. "Imputatation Methods for Filling the Long Interval of Missing Observations in Air Pollution Data and Meteorological Dataset." PhD diss., School of Environmental Engineering, 2018.
[15] Guarnaccia, Claudio, Julia Griselda Ceron Breton, Rosa Maria Ceron Breton, Carmine Tepedino, Joseph Quartieri and Nikos E. Mastorakis. "ARIMA models application to air pollution data in Monterrey, Mexico." In *AIP Conference Proceedings*, vol. 1982, no. 1. AIP Publishing, 2018. https://doi.org/10.1063/1.5045447
[16] Ye, Ziyuan. "Air pollutants prediction in shenzhen based on arima and prophet method." In *E3S Web of Conferences*, vol. 136, p. 05001. EDP Sciences, 2019. https://doi.org/10.1051/e3sconf/201913605001
[17] Lee, Muhammad Hisyam, Nur Haizum Abd Rahman, Mohd Talib Latif, Maria Elena Nor and Nur Arina Bazilah Kamisan. "Seasonal ARIMA for forecasting air pollution index: A case study." *American Journal of Applied Sciences* 9, no. 4 (2012): 570-578. https://doi.org/10.3844/ajassp.2012.570.578
[18] Ahn, Hyun, Kyunghee Sun and K. Pio Kim. "Comparison of missing data imputation methods in time series forecasting." *Computers, Materials & Continua* 70, no. 1 (2022): 767-779. https://doi.org/10.32604/cmc.2022.019369
[19] Ghapor, Adilah Abdul, Yong Zulina Zubairi and A. H. M. R. Imon. "Missing value estimation methods for data in linear functional relationship model." *Sains Malaysiana* 46, no. 2 (2017): 317-326. https://doi.org/10.17576/jsm-2017-4602-17
[20] Suhaimi, Norhazlina, Nurul Adyani Ghazali, Muhammad Yazid Nasir, Muhammad Izwan Zariq Mokhtar and Nor Azam Ramli. "Markov Chain Monte Carlo method for handling missing data in air quality datasets." *Malaysian journal of analytical sciences* 21, no. 3 (2017): 552-559. https://doi.org/10.17576/mjas-2017-2103-05
[21] Sukatis, Fahren Fazzer, Norazian Mohamed Noor, Nur Afiqah Zakaria, Ahmad Zia Ul-Saufie and Suwardi Annas. "Estimation of missing values in air pollution dataset by using various imputation methods." *International Journal of Conservation Science* 10, no. 4 (2019): 791-804.
[22] Abd Razak, Nuradhiathy, Yong Zulina Zubairi and Rossita M. Yunus. "Imputing missing values in modelling the PM10 concentrations." *Sains Malaysiana* 43, no. 10 (2014): 1599-1607.
[23] Junger, W. L. and A. Ponce De Leon. "Imputation of missing data in time series for air pollutants." *Atmospheric Environment* 102 (2015): 96-104. https://doi.org/10.1016/j.atmosenv.2014.11.049
[24] Aburashed, Laila, AL Amoush, Marah and Alrefai, Wardeh. "SQL Injection Attack Detection using Machine Learning Algorithms." *Semarak International Journal of Machine Learning*, 2(1) (2024): 1–12. https://doi.org/10.37934/sijml.2.1.112
[25] Zakaria, Nur Afiqah and Norazian Mohamed Noor. "Imputation methods for filling missing data in urban air pollution data formalaysia." *Urbanism. Arhitectura. Constructii* 9, no. 2 (2018): 159.