Check for updates

AJAE

Open Access

Vol. 14, No. 1, pp. 62-72, March 2020
doi: https://doi.org/10.5572/ajae.2020.14.1.062
ISSN (Online) 2287-1160, ISSN (Print) 1976-6912

Technical Information

Missing Value Imputation for PM_{10} Concentration in Sabah using Nearest Neighbour Method (NNM) and Expectation-Maximization (EM) Algorithm

Muhammad Izzuddin Rumaling[1)], Fuei Pien Chee[1),*], Jedol Dayou[1)], Jackson Hian Wui Chang[2),3)], Steven Soon Kai Kong[3)], Justin Sentian[4)]

[1)]Faculty of Science and Natural Resources (FSNR), Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia
[2)]Preparatory Centre for Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia
[3)]Cloud and Aerosol Laboratory, Department of Atmospheric Science, National Central University, Taoyuan, Taiwan (ROC)
[4)]Climate Change Research Group (CCRG), FSNR, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia

***Corresponding author.**
Tel: +60-88-320-000
E-mail: fpchee06@gmail.com

Received: 28 November 2019
Revised: 28 January 2020
Accepted: 7 February 2020

ABSTRACT Missing data in large data analysis has affected further analysis conducted on dataset. To fill in missing data, Nearest Neighbour Method (NNM) and Expectation Maximization (EM) algorithm are the two most widely used methods. Thus, this research aims to compare both methods by imputing missing data of air quality in five monitoring stations (CA0030, CA0039, CA0042, CA0049, CA0050) in Sabah, Malaysia. PM_{10} (particulate matter with aerodynamic size below 10 microns) dataset in the range from 2003–2007 (Part A) and 2008–2012 (Part B) are used in this research. To make performance evaluation possible, missing data is introduced in the datasets at 5 different levels (5%, 10%, 15%, 25% and 40%). The missing data is imputed by using both NNM and EM algorithm. The performance of both data imputation methods is evaluated using performance indicators (RMSE, MAE, IOA, COD) and regression analysis. Based on performance indicators and regression analysis, NNM performs better compared to EM in imputing data for stations CA0039, CA0042 and CA0049. This may be due to air quality data missing at random (MAR). However, this is not the case for CA0050 and part B of CA0030. This may be due to fluctuation that could not be detected by NNM. Accuracy evaluation using Mean Absolute Percentage Error (MAPE) shows that NNM is more accurate imputation method for most of the cases.

KEY WORDS Particulate matter, Missing data, Nearest neighbour method, Expectation maximization algorithm, Performance indicators

1. INTRODUCTION

Air quality monitoring in Malaysia is continuously conducted by Department of Environment (DOE) and is done in stations around Malaysia (Dominick *et al.*, 2012). These stations collect PM_{10} concentration data at one-hour interval. However, due to maintenance, calibration of monitoring instruments and power outage, data collected by monitoring stations may suffer missingness. Missing data mechanism can be categorized into three different types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Nakai and Ke, 2011). Missingness is categorized as MNAR when it depends on the missing value itself. MNAR is known to be non-ignorable and missing data due to MNAR is not possible to be recovered (Graham, 2009). On the other hand, missingness due to MAR depends on the observed data. MAR is ignorable and missing data can be recovered because its missingness does not depend on missing data itself. MCAR is a special case of MAR, where missingness is independent of both missing data and observed data (Dong and Peng, 2013). A set of data containing missing data due to MCAR can be considered as complete dataset because the missingness does not introduce bias (Dong and Peng, 2013). Little's MCAR test can be used to determine whether the missingness is due to MCAR (Li, 2013). If the missingness is not MCAR instead, this test cannot be used to determine whether the missingness is due to MAR or MNAR (Dong and Peng, 2013). In terms of air quality data in Malaysia, missingness can be considered as MAR because the missingness is mainly caused by maintenance, calibration of monitoring instruments and power outage. It does not depend on whether the value of data is lower or higher than certain value. Missingness can affect further analysis that requires complete dataset such as Fourier analysis and principal component analysis.

Particulate matter (PM) is mixture of substances in the form of small particles suspended in the air. PM is one of the critical components of air pollution (Li *et al.*, 2017b). Due to its small size, PM can enter respiratory system, thus becoming one of major concerns in public health (Chang *et al.*, 2018). Because of this, scientific attraction has been attracted towards PM (Shahraiyni and Sodoudi, 2016). PM mainly comes from motor vehicles, dust from construction sites and landfills. It also comes from biomass burning and brought by haze, a typical challenge in Southeast Asia since 1980s (Shaadan *et al.*, 2015). PM_{10} (particulate matter with aerodynamic diameter less than 10 microns) is one of major concern because it possesses hazardous properties towards human health compared to other pollutants such as carbon monoxide and nitrogen dioxide (Kim *et al.*, 2015; Ny and Lee, 2010). This is because it can enter respiratory system while defending natural defences of human body (Chang *et al.*, 2018). PM_{10} can increase risk of asthma, aggravate bronchitis, respiratory syncytial virus (RSV) bronchiolitis and other lung diseases (Carugno *et al.*, 2018; Lelieveld *et al.*, 2015). This is especially true for children aged between 5–15 years (Cadelis *et al.*, 2014). Other than respiratory problems, cardiovascular disease and cancer can be developed due to PM_{10} in the air (Li *et al.*, 2017a).

Many agencies around the world such as European Union (EU) and World Health Organization (WHO) implemented guidelines and set limit on air pollution concentration levels (Abd. Rani *et al.*, 2018). In Malaysia, the guidelines are implemented by DOE. According to New Malaysia Ambient Air Quality Standard, PM_{10} concentration has its standard set to 50 μg/m^3 (1-year averaging time) on 2015 before it is gradually lowered to 40 μg/m^3 by 2020 (Department of Environment, n. d.). The implementation of this standard is important in order to ensure that air quality can be maintained at safe level. Therefore, there is a need to continuously monitor ambient air quality around Malaysia.

This research focuses on evaluating performance of data imputation on air quality data from five monitoring stations around Sabah. To make performance evaluation possible, missingness is introduced to compare observed data with imputed data. Two methods of data imputation are studied in this research, namely Nearest Neighbour Method (NNM) and Expectation-Maximization (EM) algorithm. Many previous studies have employed nearest neighbour method and expectation-maximization algorithm to obtain complete dataset. However, not many of these studies emphasize on the efficiency of these two methods in data imputation. By comparing between both NNM and EM algorithm, further analysis that requires complete dataset can be made more accurate.

2. DATA AND METHODS

2.1 Study Area and Data

Five monitoring stations (CA0030, CA0039, CA0042, CA0049, CA0050) in Sabah are listed in Table 1. Respective cities of each monitoring station are located as shown in Fig. 1. Except for CA0049, other monitoring stations are located at low altitudes and are close to the sea. Furthermore, Labuan (CA0050) is situated on a small island located at western of Sabah. As shown in Fig. 2, PM_{10} concentration in Sabah differs between seasons and location (Kanniah *et al.*, 2016). Western coast of Sabah generally has higher PM_{10} concentration compared to other parts of Sabah all-year round. Also, PM_{10} concentration in Sabah is generally lower during inter-monsoon October.

Table 1. Location of monitoring stations in Sabah.

Station ID	Station name	Latitude	Longitude	Altitude (m)
CA0030	SM Putatan, Kota Kinabalu	5.9804° N	116.0735° E	13
CA0039	Pejabat JKR Tawau, Tawau	4.2447° N	117.8912° E	12
CA0042	Pejabat JKR Sandakan, Sandakan	5.8394° N	118.1172° E	10
CA0049	SMK Gunsanad, Keningau	5.3374° N	116.1567° E	288
CA0050	Taman Perumahan MPL, Labuan	5.3441° N	115.2404° E	13

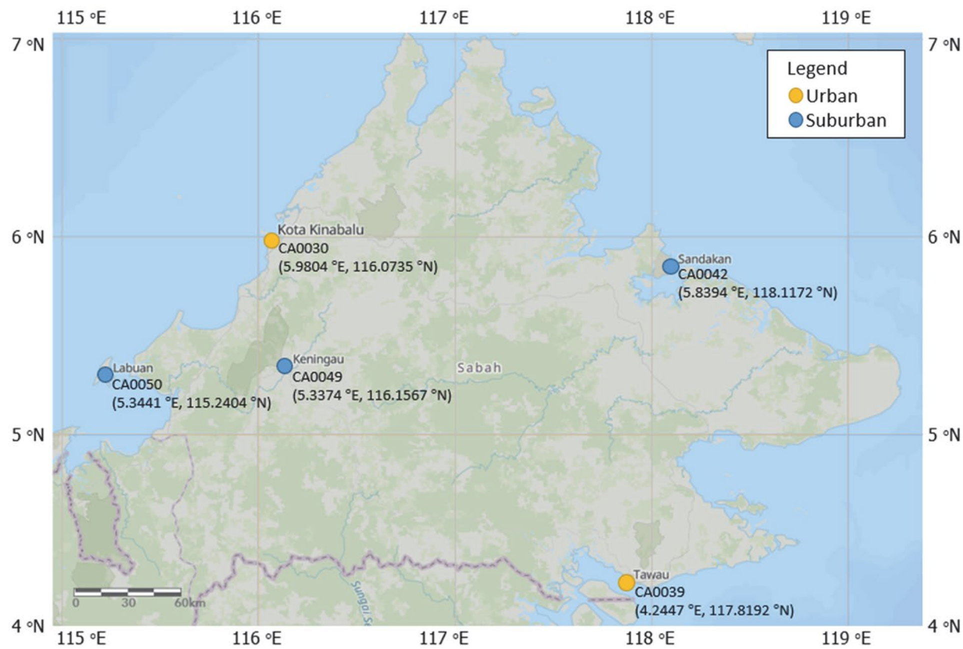

Fig. 1. Location of PM_{10} monitoring stations at urban and suburban areas (Kota Kinabalu, Tawau, Sandakan, Keningau, Labuan) in Sabah.

These monitoring stations, operated by DOE, continuously measures PM_{10} concentration data at 1-hour interval. PM_{10} concentration is measured using tapered element oscillating microbalance (TEOM), with temporal resolution of 1 h. As wind direction is angular quantity, wind speed and direction must be converted into x-component (east-west) and y-component (north-south) wind speed using equations (1) and (2). This prevents difficulty in analysis due to nature of angular quantity (Muhammad Izzuddin *et al.*, 2019; Kovač-Andrić *et al.*, 2009).

$$W_x = W_s \sin W_d \tag{1}$$

$$W_y = W_s \cos W_d \tag{2}$$

For the purpose of this research, 10-year hourly data from 2003 to 2012 are divided into two parts. The first part (Part A) ranges from 2003 to 2007, while the second part (Part B) ranges from 2008 to 2012. Due to climate change, trends of PM_{10} concentration data may differ from both parts. Thus, both parts may have difference in these data.

2.2 Introduce Missingness to Data

In order to ensure that imputed data can be validated, a fraction of observed data must be replaced by missingness. Depending on complexity, missingness is introduced into data by percentage as conducted in previous research by Noor *et al.* (2014) as shown in Table 2. A sequence of zeros and ones (0 - do not replace observed data, 1 - replace observed data with missingness) is randomly generated using MATLAB 2018b and is used as a reference to introduce missingness to observed data. The actual percentage after introducing missingness may

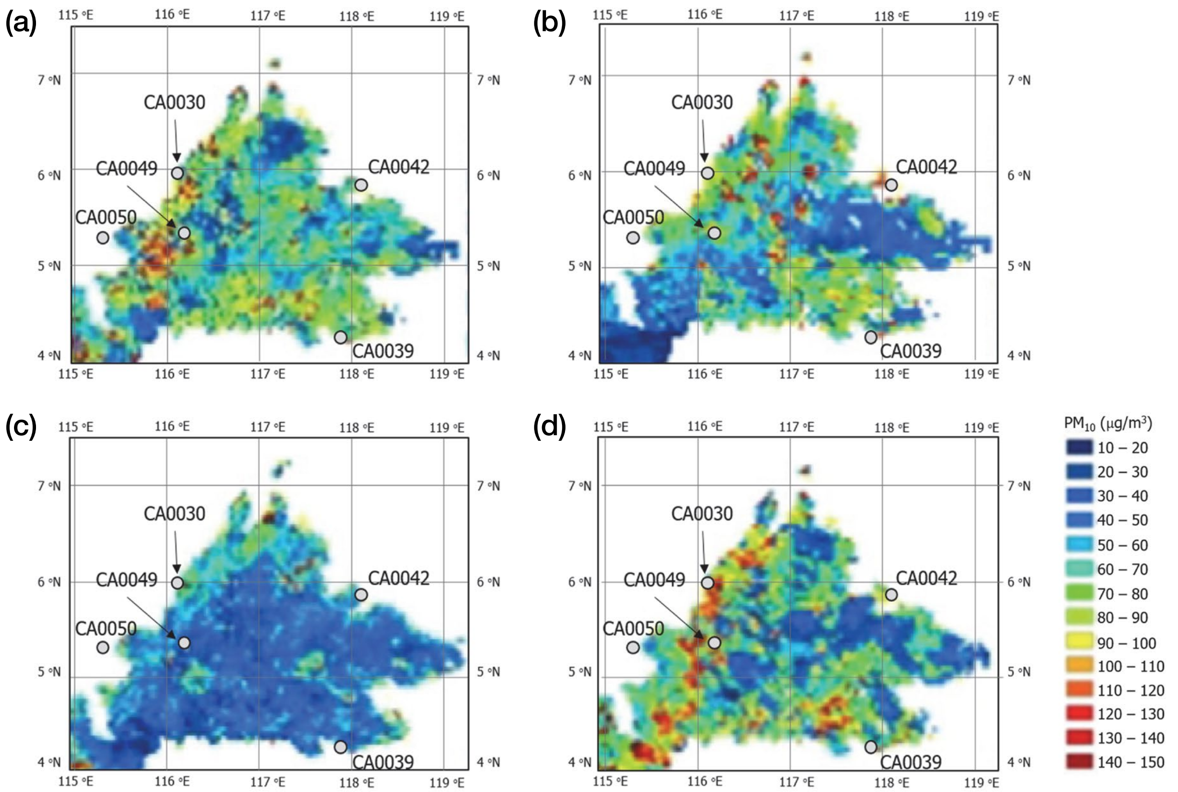

Fig. 2. Spatial distribution of estimated PM_{10} concentration in Sabah from 2007–2011 for (a) dry season (June–September), (b) wet season (November–March), (c) intermonsoon (April–May), and (d) intermonsoon (October) based on MODIS-AOD_{500} and meteorological variables (Kanniah *et al.*, 2016).

Table 2. Percentage of missingness as conducted by Noor *et al.* (2014).

Degree of complexity	Percentage of missingness (%)
Small	5
	10
Medium	15
	25
Large	40

deviate by up to 2% due to existing missingness in the data.

2.3 Data Imputation

A lot of data imputation method has been proposed for temporal dataset (Bai *et al.*, 2019). Due to simplicity, two of the most popular methods used in data imputation are NNM and EM. NNM is common in replacing missing air quality data (Li and Liu, 2014; Dominick *et al.*, 2012). For a stream of missing data bounded by observed data (x_1, y_1) in lower bound and (x_2, y_2) in upper bound, missing data is replaced with a value calculated using equations (3) and (4) (Abd Rani *et al.*, 2018; Zakaria and Noor, 2018; Siti Zawiyah *et al.*, 2010; Junninen *et al.*, 2004). NNM is performed by executing a code developed using MATLAB 2018b.

$$y=\begin{cases} y_1, \text{ when } x < x_1+\bar{x} \\ y_2, \text{ when } x \geq x_1+\bar{x} \end{cases} \tag{3}$$

$$\bar{x}=\frac{x_2-x_1}{2} \tag{4}$$

EM algorithm employs a set of iterative equations to estimate mean vector and covariance matrix of multivariate distribution from exponential family (Junger and de Leon, 2015). This method maximizes log likelihood to find parameters when there are missing values (Nakai and Ke, 2011). The simplicity and smooth operation of EM algorithm makes it unique among present multiple imputation methods. In addition, its faster operation compared to the alternatives makes EM algorithm one of the most popular imputation methods (Abd Rani *et al.*,

Table 3. Performance indicators for every station and missingness percentage for part A.

Station	Missing-ness (%)	Performance indicators							
		RMSE		MAE		IOA		COD	
		NNM	EM	NNM	EM	NNM	EM	NNM	EM
	5	18.130	**17.161**	11.765	**11.699**	**0.760**	0.649	0.495	**0.509**
	10	17.022	**16.864**	**11.279**	11.728	**0.782**	0.646	**0.536**	0.505
CA0030	15	16.887	**16.672**	**11.422**	11.715	**0.784**	0.651	**0.540**	0.503
	25	16.862	**16.205**	11.496	**11.488**	**0.782**	0.660	**0.538**	0.505
	40	17.362	**16.412**	11.745	**11.504**	**0.769**	0.660	**0.521**	0.506
	5	**21.701**	23.367	**13.371**	15.648	**0.787**	0.582	**0.522**	0.517
	10	**21.000**	23.001	**13.213**	15.517	**0.783**	0.566	**0.527**	0.484
CA0039	15	**21.591**	22.702	**13.640**	15.389	**0.770**	0.572	**0.504**	0.487
	25	**21.526**	22.648	**13.756**	15.376	**0.776**	0.569	**0.526**	0.497
	40	**22.654**	22.742	**14.220**	15.406	**0.750**	0.569	**0.488**	0.485
	5	**15.712**	17.525	**10.841**	11.761	**0.804**	0.511	**0.595**	0.370
	10	**15.393**	16.923	**10.609**	11.637	**0.801**	0.530	**0.586**	0.397
CA0042	15	**15.229**	16.543	**10.588**	11.574	**0.795**	0.544	**0.577**	0.407
	25	**14.930**	16.151	**10.506**	11.510	**0.798**	0.556	**0.585**	0.428
	40	**15.328**	16.313	**10.824**	11.656	**0.792**	0.553	**0.574**	0.425
	5	**13.560**	14.797	13.371	**10.451**	**0.841**	0.645	**0.630**	0.566
	10	**13.861**	15.218	13.213	**10.609**	**0.835**	0.626	**0.611**	0.534
CA0049	15	**13.707**	15.099	13.640	**10.639**	**0.834**	0.619	**0.613**	0.520
	25	**13.883**	15.305	13.756	**10.640**	**0.830**	0.610	**0.599**	0.514
	40	**14.302**	15.163	14.220	**10.597**	**0.819**	0.619	**0.586**	0.518
	5	14.937	**13.258**	10.666	**10.095**	**0.719**	0.676	**0.482**	0.473
	10	15.685	**13.314**	10.856	**10.093**	**0.702**	0.665	0.451	**0.467**
CA0050	15	15.552	**13.161**	10.818	**9.953**	**0.698**	0.664	0.453	**0.464**
	25	15.251	**13.266**	10.752	**9.903**	**0.711**	0.663	0.469	**0.471**
	40	15.239	**13.351**	10.843	**9.957**	**0.707**	0.661	**0.468**	0.467

Remark: Data ranges from year 2003 to 2007

2018).

Given a set of data consisting of observed data D^{obs} and missing data D^{mis}, EM algorithm starts by defining parameter θ as a random value. Then, E-step (expectation step) calculates the likelihood of each values of D^{mis} for every missingness. M-step (maximization step) uses computed values of D^{mis} to find better estimation of θ. Given the likelihood function L and expected value of log likelihood function $Q(\theta|\theta^{(t)})$, both E-step and M-step iterate until the value converges (Abd Rani *et al.*, 2018). Both E-step and M-step are executed using equations (5) and (6).

$$Q(\theta|\theta^{(t)}) = E[\log L(\theta; D^{obs}, D^{mis})] \quad (5)$$

$$\theta^{(t+1)} = \arg\max Q(\theta|\theta^{(t)}) \quad (6)$$

2.4 Performance Evaluation

The performance of data imputation is evaluated by using performance indicators. The performance indicators that have been used are root mean square error (RMSE), mean absolute error (MAE), index of agreement (IOA), and coefficient of determination (COD). The performance indicators are calculated by using equations (7) to (10) (Abd. Rani *et al.*, 2018; Nuryazmin *et al.*, 2015; Ul-Saufie *et al.*, 2013; Junninen *et al.*, 2004):

$$RMSE = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(P_i - O_i)^2} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (8)$$

Table 4. Performance indicators for every station and missingness percentage for part B.

Station	Missing-ness (%)	Performance index							
		RMSE		MAE		IOA		COD	
		NNM	EM	NNM	EM	NNM	EM	NNM	EM
CA0030	5	16.676	**14.553**	12.117	**10.864**	**0.734**	0.698	0.506	**0.521**
	10	16.299	**14.486**	12.003	**10.864**	**0.751**	0.703	0.526	**0.533**
	15	16.578	**14.595**	12.075	**10.933**	**0.743**	0.703	0.512	**0.528**
	25	16.971	**14.660**	12.247	**10.975**	**0.726**	0.695	0.495	**0.519**
	40	17.758	**14.988**	12.506	**11.055**	**0.706**	0.688	0.461	**0.508**
CA0039	5	**13.996**	18.965	**9.867**	15.145	**0.860**	0.666	**0.661**	0.535
	10	**14.383**	18.737	**9.977**	15.062	**0.846**	0.672	**0.629**	0.531
	15	**14.365**	18.913	**9.994**	15.164	**0.849**	0.668	**0.647**	0.528
	25	**14.602**	19.104	**10.187**	15.258	**0.852**	0.669	**0.654**	0.522
	40	**15.484**	18.984	**10.753**	15.246	**0.830**	0.673	**0.632**	0.530
CA0042	5	**10.615**	12.535	**7.450**	9.652	**0.845**	0.616	**0.640**	0.476
	10	**10.308**	12.533	**7.260**	9.675	**0.847**	0.606	**0.642**	0.466
	15	**10.430**	12.699	**7.287**	9.712	**0.851**	0.602	**0.647**	0.461
	25	**10.505**	12.602	**7.368**	9.729	**0.847**	0.602	**0.644**	0.469
	40	**10.722**	12.809	**7.526**	9.770	**0.835**	0.591	**0.632**	0.451
CA0049	5	18.255	**17.987**	**9.867**	12.430	**0.756**	0.529	**0.457**	0.400
	10	**16.754**	17.404	**9.977**	12.274	**0.780**	0.551	**0.492**	0.428
	15	**16.966**	18.046	**9.994**	12.457	**0.777**	0.530	**0.486**	0.399
	25	**16.771**	18.225	**10.187**	12.574	**0.780**	0.521	**0.495**	0.394
	40	**17.564**	18.143	**10.753**	12.629	**0.758**	0.523	**0.462**	0.396
CA0050	5	23.646	**16.463**	14.435	**11.360**	0.693	**0.795**	0.405	**0.609**
	10	23.071	**16.22**	14.070	**11.317**	0.701	**0.798**	0.427	**0.616**
	15	23.210	**16.007**	14.114	**11.272**	0.695	**0.809**	0.418	**0.633**
	25	23.163	**16.279**	14.173	**11.278**	0.696	**0.800**	0.414	**0.622**
	40	22.865	**16.203**	14.213	**11.319**	0.698	**0.800**	0.424	**0.625**

Remark: Data ranges from year 2008 to 2012

$$IOA = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \tag{9}$$

$$COD = R^2 = \left(\frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{n \cdot s_p \cdot s_o}\right)^2 \tag{10}$$

where n is total number of data, P_i is predicted value of ith data, O_i is observed value of ith data, $\bar{P}$ is mean predicted value, $\bar{O}$ is mean observed value, s_p is standard deviation of predicted values, and s_o is standard deviation of observed values.

2.5 Mean Absolute Percentage Error (MAPE)

Mean absolute percentage error (MAPE) is a measure that evaluates accuracy of a prediction model (Khair *et al.*, 2017). MAPE indicates error in predicting the value of missing data when comparing to real value. MAPE is calculated using equation (11) as follows (Khair *et al.*, 2017).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - P_i|}{O_i} \times 100\% \tag{11}$$

3. RESULT AND DISCUSSION

3.1 Performance Indicators

PM_{10} concentration datasets for five monitoring stations in Sabah are analysed. RMSE, MAE, IOA, and COD are calculated for every percentage of missingness

Table 5. Coefficient of correlation for dataset in Part A.

Missingness (%)	Station									
	CA0030		CA0039		CA0042		CA0049		CA0050	
	NNM	EM	NNM	EM	NNM	EM	NNM	EM	NNM	EM
5	**0.593**	0.569	**0.633**	0.575	**0.653**	0.406	**0.714**	0.604	**0.524**	0.363
10	**0.624**	0.547	**0.626**	0.535	**0.648**	0.429	**0.707**	0.580	0.504	**0.507**
15	**0.626**	0.552	**0.606**	0.539	**0.639**	0.438	**0.704**	0.561	0.497	**0.504**
25	**0.622**	0.555	**0.613**	0.541	**0.644**	0.456	**0.698**	0.553	0.513	**0.514**
40	**0.602**	0.560	**0.575**	0.537	**0.634**	0.449	**0.680**	0.558	0.506	**0.512**

Remark: Data ranges from year 2003 to 2007

Table 6. Coefficient of correlation for dataset in Part B.

Missingness (%)	Station									
	CA0030		CA0039		CA0042		CA0049		CA0050	
	NNM	EM	NNM	EM	NNM	EM	NNM	EM	NNM	EM
5	0.544	**0.577**	**0.746**	0.558	**0.721**	0.493	**0.594**	0.386	0.498	**0.711**
10	0.570	**0.587**	**0.723**	0.566	**0.725**	0.487	**0.628**	0.412	0.506	**0.716**
15	0.560	**0.586**	**0.728**	0.560	**0.732**	0.485	**0.622**	0.386	0.498	**0.729**
25	0.533	**0.571**	**0.732**	0.565	**0.724**	0.490	**0.625**	0.369	0.500	**0.723**
40	0.506	**0.564**	**0.695**	0.571	**0.704**	0.473	**0.594**	0.363	0.501	**0.723**

Remark: Data ranges from year 2008 to 2012

and station for both part A and B. Tables 3 and 4 reveals performance indicators for NNM and EM at 5 missingness levels and 5 different stations for part A and part B respectively. The desirable attributes between these methods are highlighted in bold. In terms of missingness level, there is no definite relationship between performance of data imputation and missingness level. This is because both NNM and EM impute missing data based on available data. As long as available data is sufficient, missing data can still be effectively imputed.

Most of the data show that nearest neighbour method is better imputation method. This may be due to the nature of missingness in relation to ability of EM algorithm to impute data. EM algorithm works best for missing data caused by MCAR (Nakai and Ke, 2011; Graham, 2009). However, air quality data collected in monitoring stations are not caused by MCAR as the cause of missingness is known. This may attribute to lower performance of EM algorithm compared to NNM.

However, this is not the case for CA0050, where most of the performance indicators for that station show that EM algorithm is a better imputation method. This may be due to the fact that Labuan is surrounded by sea. One study has shown that air humidity is affected by bodies of water due to high heat capacity and strong evaporation (Zhu and Zeng, 2018). Furthermore, cold-wet air that surrounds a water body enhances air flow away from bodies of water by changing the local air circulation (Zhu and Zeng, 2018). The local air circulation highly affects humidity in Labuan. Another study suggests that different levels of humidity affects PM_{10} concentration differently (Lou *et al.*, 2017). PM_{10} concentration increases with humidity up to 60%. Beyond that point, gravity deposition occurs and PM_{10} concentration begins to drop (Lou *et al.*, 2017). PM_{10} concentration as monitored by CA0050 may fluctuate due to continually changing of humidity level, traffic congestion and active industrial activity. This fluctuation is not accounted by NNM, leading to indication that EM algorithm is better imputation method for data collected by CA0050.

As for PM_{10} concentration read by CA0030, several performance indicators show that EM algorithm is better imputation method especially for part B of the data. This may be due to fluctuation of PM_{10} concentration in Kota

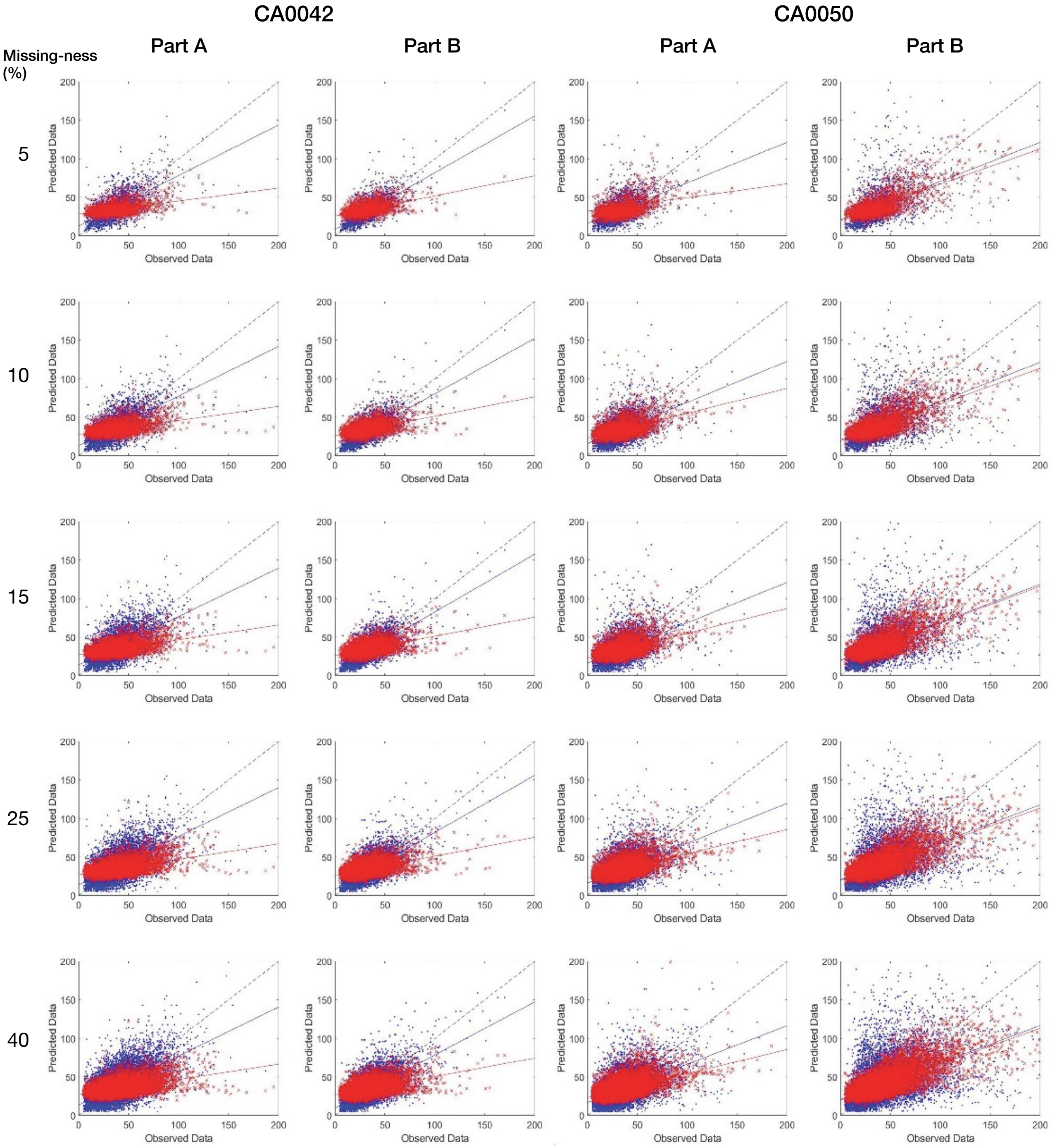

Fig. 3. Scatter plot for imputation of data from CA0042 and CA0050 for part A and B at various missingness percentage. Blue indicates NNM, red indicates EM, while dashed line represents the point where predicted data equals observed data.

Kinabalu especially between year 2008 and 2012. One study shows that PM_{10} concentration from 16^{th} to 18^{th} January 2012 spiked at 7.00 a.m. and fluctuates at the other time (Chang *et al.*, 2018). When this portion of data is missing, NNM may not be able to restore the missingness as well as EM algorithm.

3.2 Regression Analysis on Imputed Data

The performance of data imputation is further evaluated by calculating correlation of coefficient R on predict-

Table 7. Mean absolute percentage error (MAPE) of stations in Sabah for various missingness level.

Set	Missing-ness (%)	Station									
		Kota Kinabalu		Tawau		Sandakan		Keningau		Labuan	
		NNM	EM	NNM	EM	NNM	EM	NNM	EM	NNM	EM
A	5	**35.771**	38.619	**25.091**	27.545	**33.003**	36.986	**26.217**	31.333	**37.707**	39.461
	10	**34.874**	39.585	**24.925**	27.759	**32.558**	37.109	**25.916**	31.282	**38.180**	38.842
	15	**35.260**	39.367	**26.160**	27.763	**32.853**	37.231	**25.717**	31.038	**38.274**	38.285
	25	**35.835**	38.428	**26.496**	27.745	**32.794**	37.773	**25.935**	31.108	**37.627**	37.652
	40	**36.051**	37.932	**27.407**	27.668	**33.795**	37.859	**26.467**	30.888	**37.645**	37.812
B	5	**42.717**	44.935	**28.481**	59.879	**25.373**	40.454	**27.214**	42.656	40.381	**37.333**
	10	**43.452**	46.053	**28.743**	59.450	**25.204**	41.661	**26.583**	43.593	39.701	**37.232**
	15	**43.741**	45.783	**29.204**	59.890	**25.409**	41.824	**26.510**	43.857	39.440	**36.847**
	25	**44.431**	46.577	**30.007**	60.157	**25.918**	42.443	**26.655**	43.965	39.875	**36.972**
	40	**45.364**	47.055	**32.239**	60.652	**26.601**	42.348	**27.922**	44.606	40.546	**37.309**

ed data against observed data. The most ideal case of imputed data occurs when predicted data equals observed data (R=1). Tables 5 and 6 reveals coefficient of correlation of data in part A and B respectively, for all five missingness percentages and five stations.

Similar to performance indicators, coefficient of correlation shows that NNM is better imputation method for monitoring stations in Tawau, Sandakan and Keningau. As for CA0030, NNM is better imputation method for Part A, but not in Part B. Dataset recorded by CA0050 strongly suggests that EM algorithm is better imputation method.

Fig. 3 reveals scatter plot of data imputation for both CA0042 and CA0050. CA0042 and CA0050 are selected to be presented in the Fig. 3 because CA0042 is located at high altitude while CA0050 is located in a small island. The predicted-observed regression is shown for both stations due to different geographical condition in contrast to the other three stations. Coefficient of correlation for CA0042 shows relatively large difference between two methods compared to other stations. As shown in Fig. 3, all scatter plots for CA0042 shows that line representing NNM is closer to dashed line compared to line that represents EM algorithm. This shows that NNM has greater tendency to predict missing data closer to observed data compared to EM algorithm. This might be caused by missingness mechanism, in which data is Missing at Random. EM algorithm may not be able to impute MAR data as well as MCAR data (Nakai and Ke, 2011; Graham, 2009).

Meanwhile, CA0050 shows that EM algorithm gives better coefficient of correlation in contrast to other stations. Despite that, Fig. 3 reveals that NNM has either greater tendency (Part A) or approximately similar to EM algorithm (Part B) to predict missing data. This is because the lines representing NNM and EM are plotted at best fit. However, the scatter plot shows that imputed data by NNM for CA0050 are more dispersed away from line of best fit compared to that of CA0042, which might contribute to lower R value of NNM compared to EM algorithm. Although best fit line for NNM is closer to dashed line, the dispersion of scatter plot shows that EM algorithm is better imputation method compared to NNM.

3.3 Mean Absolute Percentage Error (MAPE)

Performance of data imputation is further evaluated using MAPE. Data imputation is most accurate when MAPE approaches zero. Table 7 reveals accuracy of data imputation using NNM and EM for all stations and various level of missingness. According to Table 7, it is shown that NNM is generally more accurate data imputation method compared to EM (except for CA0050 in set B). This is reflected by lower values for NNM for most of the cases. This may be due to its ability to predict missing data closer to actual data compared to EM.

4. CONCLUSION

Generally, it has been shown that NNM is better imputation method for data from all the monitoring stations in Sabah except CA0050. NNM works most efficient for CA0049 in Part A (RMSE < 14.302, MAE < 10.640, IA > 0.819 and COD > 0.586) and CA0042 in Part B (RMSE < 10.722, MAE < 7.526, IA > 0.835 and COD > 0.632). This may be due to missing data type of MAR. However, strong fluctuation which may be present in data from CA0050 and part B from CA0030 may cause NNM to impute data not as well as EM algorithm. This may be further confirmed by regression analysis for CA0050 (R > 0.711 for part B). Evaluation of accuracy using MAPE reveals that NNM is more accurate imputation method for most cases (except for set B in CA 0050). This shows that NNM can be used as data imputation for missing data found in dataset observed by stations in Sabah. Accurate data imputation is important for future research because this enables further analysis on air quality data to become more reliable.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Malaysia Sabah for supporting this research by providing grant (SBK0324-2018, SGI0054-2018 and GUG0378-2018) and Department of Environment Malaysia for providing meteorological and pollutant data for research purpose.

REFERENCES

Abd. Rani, N.L., Azid, A., Khalit, S.I., Juahir, H. (2018) Prediction Model of Missing Data: A Case Study of PM_{10} across Malaysia Region. Journal of Fundamental and Applied Science, 10(1S), 182-203, https://doi.org/10.4314/jfas.v10i1s.1.

Bai, K., Li, K., Guo, J., Yang, Y., Chang, N.B. (2019) Filling the gaps of in-situ hourly $PM_{2.5}$ concentration data with the aid of empirical orthogonal function constrained by diurnal cycles. Atmospheric Measurement Techniques, 1-29, https://doi.org/10.5194/amt-2019-317.

Cadelis, G., Tourres, R., Molinie, J. (2014) Short-Term Effects of the Particulate Pollutants Contained in Saharan Dust on the Visits of Children to the Emergency Department due to Asthmatic Conditions in Guadeloupe (French Archipelago of the Caribbean). PLOS ONE, 9(3), 1-11, https://doi.org/10.1371/journal.pone.0091136.

Carugno, M., Dentali, F., Mathieu, G., Fontanella, A., Mariani, J., Bordini, L., Milani, G.P., Consonni, D., Bonzini, M., Bollati, V., Pesatori, A.C. (2018) PM_{10} exposure is associated with increased hospitalizations for respiratory syncytial virus bronchiolitis among infants in Lombardy, Italy. Environmental Research, 166, 452-457, https://doi.org/10.1016/j.envres.2018.06.016.

Chang, H.W.J., Chee, F.P., Kong, S.K.S., Sentian, J. (2018) Variability of the PM_{10} concentration in the urban atmosphere of Sabah and its responses to diurnal and weekly changes of CO, NO_2, SO_2 and Ozone. Asian Journal of Atmospheric Environment, 12(2), 109-126, https://doi.org/10.5572/ajae.2018.12.2.109.

Department of Environment. (n. d.) New Malaysia Ambient Air Quality Standard. Available at http://www.doe.gov.my/portalv1/wp-content/uploads/2013/01/Air-Quality-Standard-BI.pdf.

Dominick, D., Juahir, H., Latif, M.T., Zain, S.M., Aris, A.Z. (2012) Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmospheric Environment, 60, 172-181, https://doi.org/10.1016/j.atmosenv.2012.06.021.

Dong, Y., Peng, C.Y.J. (2013) Principled missing data methods for researchers. SpringerPlus, 2(222), 1-17. https://doi.org/10.1186/2193-1801-2-222.

Graham, J.W. (2009) Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology, 60, 549-576, https://doi.org/10.1146/annurev.psych.58.110405.085530.

Junger, W.L., de Leon, A.P. (2015) Imputation of missing data in time series for air pollutants. Atmospheric Environment, 102, 96-103, https://doi.org/10.1016/j.atmosenv.2014.11.049.

Junninen, J., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M. (2004) Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38: 2895-2907, https://doi.org/10.1016/j.atmosenv.2004.02.026.

Kanniah, K.D., Kaskaoutis, D.G., Lim, H.S., Latif, M.T., Kamarul Zaman, N.A.F., Liew, J. (2016) Overview of atmospheric aerosol studies in Malaysia: Known and unknown. Atmospheric Research, 182, 302-318, https://doi.org/10.1016/j.atmosres.2016.08.002.

Khair, U., Fahmi, H., Al Hakim, S., Rahim, R. (2017) Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error. Journal of Physics, 930(1), 1-6, https://doi.org/10.1088/1742-6596/930/1/012002.

Kim, K.H., Kabir, E., Kabir, S. (2015) A review on the human health impact of airborne particulate matter. Environment International, 74, 136-143, https://doi.org/10.1016/j.envint.2014.10.005.

Kovač-Andrić, E., Brana, J., Gvozdić, V. (2009) Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. Ecological Informatics, 4(2), 117-122, https://doi.org/10.1016/j.ecoinf.2009.01.002.

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A. (2015) The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525(7569), 367-371, https://doi.org/10.1038/nature15371.

Li, C. (2013) Little's test of missing completely at random. The Stata Journal, 13(4), 795-809, https://doi.org/10.1177/1536867X1301300407.

Li, L., Liu, D.J. (2014) Study on an Air Quality Evaluation Model for Beijing City Under Haze-Fog Pollution Based on

New Ambient Air Quality Standards. International Joutnal of Environment Research and Public Health, 11, 8909-8923, https://doi.org/10.3390/ijerph110908909.

Li, L., Wu, A.H., Cheng, I., Chen, J.C., Wu, J. (2017a) Spatio-temporal estimation of historical $PM_{2.5}$ concentrations using PM_{10}, meteorological variables, and spatial effect. Atmospheric Environment. 166, 182-191, https://doi.org/10.1016/j.atmosenv.2017.07.023.

Li, X., Chen, X., Yuan, X., Zeng, G., León, T., Liang, J., Chen, G., Yuan, X. (2017b) Characteristics of Particulate Pollution ($PM_{2.5}$ and PM_{10}) and Their Spacescale-Dependent Relationships with Meteorological Elements in China. Sustainability, 9(12), 2330-2443, https://doi.org/10.3390/su9122330.

Lou, C., Liu, H., Li, Y., Peng, Y., Wang, J., Dai, L. (2017) Relationships of relative humidity with $PM_{2.5}$ and PM_{10} in the Yangtze River Delta, China. Environmental Monitoring Assessment, 189(11), 1-16, https://doi.org/10.1007/s10661-017-6281-z.

Muhammad Izzuddin, R., Chee, F.P., Dayou, J., Chang, H.W.J., Soon, K.K.S., Sentian, J. (2019) Temporal Assessment on Variation of PM_{10} Concentration in Kota Kinabalu using Principal Component Analysis and Fourier Analysis. Current World Environment, 14(3), 400-410, https://doi.org/10.12944/CWE.14.3.08.

Nakai, M., Ke, W. (2011) Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. International Journal of Mathematical Analysis, 5(1), 1-13.

Noor, H.M., Nasrudin, N., Foo, J. (2014) Determinants of Customer Satisfaction of Service Quality: City bus service in Kota Kinabalu, Malaysia. Procedia - Social and Behavioral Sciences, 153, 595-605, https://doi.org/10.1016/j.sbspro.2014.10.092.

Nuryazmin, A.Z., Abdul Aziz, J., Nora, M. (2015) A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. Sains Malaysiana, 44(3), 449-456.

Ny, M.T., Lee, B.K. (2010) Size Distribution and Source Identification of Airborne Particulate Matter and Metallic Elements in a Typical Industrial City. Asian Journal of Atmospheric Environment, 4(1), 9-19, https://doi.org/10.4209/aaqr.2010.10.0090.

Shaadan, N., Jemain, A.A., Latif, M.T., Mohd. Deni, S. (2015) Anomaly detection and assessment of PM_{10} functional data at several locations in the Klang Valley, Malaysia. Atmospheric Pollution Research, 6, 365-375, https://doi.org/10.5094/APR.2015.040.

Shahraiyni, H.T., Sodoudi, S. (2016) Statistical Modeling Approaches for PM_{10} Prediction in Urban Areas; A Review of 21st-Century Studies. Atmosphere, 7, 1-24, https://doi.org/10.3390/atmos7020015.

Siti Zawiyah, A., Mohd Talib, L., Aida Shafawati, I., Liew, J., Abdul Aziz, J. (2010) Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. Air Quality, Atmosphere and Health, 3, 53-64, https://doi.org/10.1007/s11869-009-0051-1.

Ul-Saufie, A.Z., Yahaya, A.S., Ramli, N.A., Rosaida, N., Abdul Hamid, H. (2013) Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmospheric Environment, 73, 621-630, https://doi.org/10.1016/j.atmosenv.2013.05.017.

Zakaria, N.A., Noor, N.M. (2018) Imputation Methods for Filling Missing, Data in Urban Air Pollution Data for Malaysia. Urbanism, 9(2), 159-166.

Zhu, C., Zeng, Y. (2018) Effects of urban lake wetlands on the spatial and temporal distribution of air PM_{10} and $PM_{2.5}$ in the spring in Wuhan. Urban Forestry and Urban Greening, 31, 142-156. https://doi.org/10.1016/j.ufug.2018.02.008.