MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

Application of Functional Time Series Model in Forecasting Monthly Diurnal Maximum API Curves: A Comparison between Multi-Step Ahead and Iterative One-Step Ahead Approach

Wan Najiha Wan Mat Din, Norshahida Shaadan*

Center of Statistical and Decision Science Studies, Faculty of Computer & Mathematical Sciences, 40450, UiTM Shah Alam, Selangor, Malaysia

Abstract In Malaysia, Air Pollution Index (API) is used to assess the status of background air quality. The computation of API involved six major air pollutants including PM10, PM2.5, O_3, CO, SO_2 and NOx. Due to the harmful effect of air pollution, forecasting API is important. Exposure and risk assessment play important role to measure the impact of air pollution towards humans health for controlling and mitigation purposes. Information based on continuous diurnal level of air pollution is more fruitful compared to daily average point data in the assessment process. This requires the needs for analyzing and predicting continuous diurnal level of API in the form of curves. To enable the analysis to be conducted, this paper introduces the application of Functional Time Series (FTS) model in forecasting monthly diurnal maximum API curves at two selected sites in Peninsular Malaysia; namely Shah Alam Selangor and Pasir Gudang Johor. Two FTS models were compared which include Multi-Step ahead and Iterative One-Step ahead approach. The results show that the Multi-Step ahead model has produced better performance giving the lowest error measures; FMSE, FRMSE and FMAPE compared to Iterative One-Step ahead. This study has shown that FTS model has the advantage because it enables the prediction of continuous API levels within a defined continuum time, which in this study was the interval time within 24 hours. Functional descriptive mean shows a bimodal pattern with a peak at 3.00 pm and the average levels are at a healthy level. Functional mean of API exhibits an increasing pattern after sunrise towards 10.00 am at both sites, which inform that, this is the time with a higher contribution of vehicles emission while the standard deviation differs in the pattern. The model is recommended as an alternative model to be used by the government and environmentalists in providing input for guiding pollution control and protecting public health at the early stage. Furthermore, as for the private sector and industries, this study might provide a predictive analytic tool for forecasting daily API curves instead of a single daily average API value.

Keywords: Functional Time Series, Forecasting, Air Pollution Index, Air Quality Prediction.

***For correspondence:**
shahida@tmsk.uitm.edu.my

Received: 26 Sept 2021
Accepted: 17 Feb 2022

Introduction

Air pollution has become an issue of concern worldwide due to its harmful impact on society and the environment. This becomes the reason for the urgency for air quality monitoring with the ultimate aim to assess and examine the air quality status of an environment. As for Malaysia's experience, Air Pollution Index (API) is used to describe air quality. The index is calculated involving six major air pollutants including particulate matter less than 10µm (PM10), particulate matter less than 2.5µm (PM2.5), Ozone (O_3), Carbon Monoxide (CO), Sulphur Dioxide (SO_2) and Nitrogen Dioxide (NO_2) [2]. Future predicted values of air pollution levels are also important as input to the environmentalist, industries or even to the many respective decision-makers in general. Thus, predicting the future level of API is important. A suitable prediction model and approach are required for more accurate prediction results and better information gathering.

Based on the literature review, many studies have been conducted to obtain information on air quality behavior and status in various fields of scope or angle using various methods and approaches. These include the study to investigate the temporal and spatial pattern of Malaysia air quality, the type and the possible emission sources [10], predictive modeling involving the application of advanced statistical analysis such as principal component analysis (PCA) and artificial neural network (ANN) to forecast the air pollutant index [1], seasonal ARIMA for forecasting API [8] and several more.

In the context of univariate prediction and forecasting models, based on the previous study done, the researcher had commonly used the classical time series technique such as ARIMA, SARIMA to predict air quality (API) or air pollutant values. However, a gap persists related to the modeling process whereby the data used to predict API were discrete observations which are based on points of hourly API. For example, if the aim is to forecast future daily API, consequently, the model used will produce a single forecasted daily point API values. These forecasted values are static, thus contains less information, which limits the pattern or dynamic prediction of API. It is more meaningful if ones can predict the continuous level of API within a period of 24 hours so that the whole pattern of API level can be observed or visualized. This requires the prediction for daily API curves. In conjunction to this matter, to overcome the limitation, this paper proposes an application of Functional Data Analysis (FDA) to predict API in the form of curves. Curves consist of a continuous level of API within a defined continuum. For example, a day curve consists of the continuous level within 24 hours continuum while a month curve consists of the continuous level within 168 hours continuum and so on. In this study, data on the monthly diurnal API curves is the focus and is considered.

FDA consists of a set of techniques and methods to analyze data. The data are in the form of curves or shapes or images. This data is also known as functional data because the physical representation of the data is defined by a set of mathematical functions. Functional data refers to a continuous observation that exists within a continuum of time or space [12]. Among the popular application of FDA includes the analysis for financial stock price curves [6] and monthly sea surface temperature in climatology [14]. When curves data are recorded according to a sequence series of time, the data set is known as functional time series [7]. In this study, FDA is applied to predict future monthly diurnal maximum API using Functional Time Series (FTS) modeling and methodology. The continuous nature of daily API within a continuum of 24 hours period has inspired this research to be conducted. It is believed that predicting the daily API curves would provide more information compared to predicting the average point of daily levels. The FTS method is adapted based on the application of the functional data approach in predicting demography curves in many research in the literature [4,5]. For example, FDA is applied to analyze annual age-specific mortality rates [3] involving series of 41 yearly mortality curves from 1975 to 2015. Another application is by Hyndman and Shang [4] to forecast Australian fertility rate curves using the curve-smoothing method without monotonic constraint [5].

Noticeably, in the literature, the application of FTS in air quality data set is still lacking. Realizing the importance of API and air quality forecasting, this study is conducted to investigate the capability and the advantage of FDA and FTS models in predicting monthly diurnal maximum API curves at two air quality monitoring stations; Shah Alam, Selangor and Pasir Gudang, Johor Malaysia. Two existing FTS models; the Multi-Step ahead and Iterative One-Step ahead are taken into consideration and their performance are compared.

Methodology

Data and study locations

To achieve the study objective, a secondary data set of hourly Air Pollutant Index (API) over the period from the year 2011 to 2018 is employed in the analysis. The data was obtained from the Air Quality Division of the Department of Environment (DOE), Malaysia for two air quality monitoring sites from the available stations; Shah Alam and Pasir Gudang. The reason for choosing these two sites is because of the location of the sites and the frequent record of high API in sometimes of the years. Shah Alam is often chosen to be a popular study location in several studies [1,11,15]. Shah Alam is located within the Klang Valley region, having active development of large-scale industrial and commercial activities, densely populated areas, and also high vehicular traffic. Due to these characteristics, Shah Alam is more susceptible to air pollution. Meanwhile, Pasir Gudang is a busy industrial site located in the southern part of Johor which is often reported to experience air quality deterioration episodes annually. Both sites were also reported to be impacted by transboundary haze.

Study framework and analysis process

The study is conducted according to a designed framework to accomplish the objectives involving data collection, data preparation which includes data cleaning, data arrangement, data conversion from discrete values into curves and data analysis.

Data preparation: Cleaning and arrangement

In this study, missing values are treated using Principal Components Analysis (PCA) imputation method through missMDA package in R software. The data are arranged into a matrix format. The rows are daily cases and the columns are the variables hours, i.e.: there will be 24 variables representing variable 1 is hour 1, variable 2 is hour 2 and so on until variable 24 is hour 24. The variable hour represents the recorded API of the hour. Since the model is developed for the prediction of monthly maximum API cures, the data set is then filtered and rearranged into a matrix of monthly by hourly maximum (i.e. within 24 hours period) API which consists of 96 rows (i.e. 12 months x 8 years (2011-2018)) and 24 columns. Table 1 illustrate the matrix of data presentation to be used for FTS modeling. For the model development, the data set is partitioned into training and testing data set. A training data set is used to establish the FTS model and testing data set is used for the validation phase. In this study, the training data set consists of monthly diurnal maximum data from January 2011 until June 2018 (i.e. 90 monthly/rows/ curves) while the testing data set involved monthly diurnal maximum data from July 2018 until December 2018 (i.e. 6 monthly/rows/curves).

Table 1. Matrix representation of data set for analysis.

	Maximum API record within continuum I=[1,24] hours					
Month	Hour 1	Hour 2	Hour 3	…	Hour 23	Hour 24
1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	…	$y_{1,23}$	$y_{1,24}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	…	$y_{2,23}$	$y_{2,24}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Data conversion and smoothing

FTS models aim to forecast the future monthly diurnal maximum API curves. To develop the FTS model, a series of monthly diurnal maximum API curves is needed. The development of FTS begins by converting the observed data recorded in Table 1 into curves. This conversion process involves basis expansion methodology. Assuming that there is a set of discrete observations for a month *t*, $y_t = [y_{t,1}, y_{t,2}, \dots, y_{t,24}]$ recorded for *n* number of months $t = [1,2,\dots,n]$ and time point $i = [1,2,\dots,24]$. Data conversion is defined as transforming the discrete data y into a continuous function $y_t(x_i)$ that can be computed at any time point *i* by the following equation:

$$y = y_t(x_i) + \varepsilon \tag{1}$$

Where the term $y_t(x_i)$ is assumed as a random smooth function and the term ε is random noise. The function $y_t(x_i)$ is established using the following equation:

$$y_t(x_i) = \sum_{k=1}^{K} C_k \varphi_k(x_i) \tag{2}$$

The term $\varphi_k(x_i)$ is the basis system consisting of k number of basis functions. Researcher may has several choices of basis function, among the Fourier, spline, polynomial, quadratic and constant that can be used as the basis. The determination on choosing the appropriate one is depending on the behavior and nature of the data. This can be determined just by conducting a simple exploratory analysis. Fourier is more suitable for curves that have seasonal fluctuation or periodic pattern meanwhile splines are more flexible for any kind of pattern [12]. In this study due to flexibility reason, B-spline basis was employed. The term C_k is the corresponding basis coefficient estimated using the least-square method. The smoothness of the curves depends on the size of basis function k. Too small k causes curves to be too smooth while the too large size of k impacts curves to be unsmooth or rough. The appropriate number of k can be determined using several statistics such as deviant value and AIC. In this study, the smallest AIC value is used to determine the appropriate number of basis function k [13].

Functional time series models

The development of FTS model to be used for predicting monthly diurnal maximum API curves in this study is adapted from previous research that has been conducted by Hyndman and Shang [4]. They have proposed the model for annual mortality curves forecasting. The procedure and methods involve the combination of Functional Principal Component Regression (FPCR) and Functional Principal Component Analysis (FPCA) methodology. Mathematically, lets $y_t(x_i)$ represent a function of API for the continuous hour variable x_i in month *t*. It is assumed that there is an underlying smooth function $f_t(x_i)$, which observes with error at discretized grid points of x.

In this study, the researcher observes $\{x_i, y_t(x_i)\}$ for $t = 1,2,\dots,n$ months and $i = 1,2,\dots,24$ point hours of maximum diurnal API readings from which a smooth function $f_t(x_i)$ is obtained given by the following equation,

$$y_t(x_i) = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i} \tag{3}$$

where $\varepsilon_{t,i}$ is an independent and identically distributed (iid) standard random variable, $\sigma_t(x_i)$ allows the amount of noise to vary with x_i, and $\{x_1, x_2, \dots, x_{24}\}$ is a set of discrete data points. To forecast or predict the next month diurnal maximum API curve when $t = n + 1$, the following function defined by equation (4) below is used.

$$y_{n+1}(x_i) = f_{n+1}(x_i) + \sigma_t(x_i)\varepsilon_{n+1,i} \tag{4}$$

Looking into the component of the term $y_{n+1}(x_i)$, a stochastic process indicated by f represented by function $f_{n+1}(x_i)$ can be composed into two-component functions, the mean function and the sum of the products of orthogonal functional principal components and uncorrelated principal component scores. The term f can be written as below.

$$f = \mu + \sum_{k=1}^{\infty} \beta_k \phi_k \tag{5}$$

The term μ refers to the unobservable population mean function, $\beta_k$is the K^{th}principal component scores, and $\phi_k$is the K^{th}population functional principal component. To obtain the function for the term, that is $f_t(x_i)$, for a given month *t,* the following model equation is defined.

$$f_t(x_i) = \mu(x_i) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x_i) + \varepsilon_t(x_i) \tag{6}$$

The term $\phi_k(x_i)$ is the k^{th} principal component function. The term $\{\beta_{1,k}, \beta_{2,k}, \dots, \beta_{n,k}\}$ are the corresponding scores while the term $\varepsilon_t(x_i)$ $\varepsilon_t(x_i)$ are the independent and identically distributed random functions with zero mean. It is required that $K < n$. It is important to note that towards the forecasting of a future curve, referring to Hyndman and Ullah [5], it is recommended that each univariate time series $\{\beta_{t,k}\}, k = 1,2,\ldots,K$ can be forecasted separately using the univariate time series model. This is due to the uncorrelated properties of the principal component scores. Therefore, the future curve $y_{n+1}(x_i)$ can be obtained by first estimating $f_{n+1}(x_i)$ using the multiplication of the forecasted principal component scores with the principal components.

The mean function $\mu(x_i)$ is estimating using weighted average as shown below.

$$\hat{\mu}(x_i) = \textstyle\sum_{t=1}^{n} w_t \hat{f}(x_i) \tag{7}$$

where $\hat{f}_t(x_i)$ is the smoothed curve estimated from y_t, and $w_t = K(1-K)^{n-t}$ is a geometrically decreasing weight with $0 < K < 1$. If one requests for a robust estimator, then L median of the estimated smoothed curves can be used [5]. The mean or median-adjusted functional data are represented as

$$f_t^*(x_i) = \hat{f}(x_i) - \hat{\mu}(x_i) \tag{8}$$

The estimation of the smoothed function $\{f_t(x_i)\}$ is conducted following the research of Hyndman and Ullah [5] using a nonparametric smoothing approach where the K^{th}principal component and their scores is obtained using a functional principal component analysis (FPCA) by decomposing them from the monthly series of diurnal maximum API curves.

Multi-Step Ahead Versus Iterative One-Step Ahead Forecast

There will be two models of FTS namely the Multi-Step ahead and Iterative One-Step ahead forecast are applied in this study. With a given forecast horizon *h*, the *h*-step ahead forecasts can be obtained using $y_{n+h}(x)$ as given in equation (9).

$$\hat{y}_{n+h|n}(x) = E[y_{n+h}(x)|f(x), B] = \bar{f}(x) + \textstyle\sum_{k=1}^{K} \hat{\beta}_{n+h|n,k}\hat{\Phi}_k(x) \tag{9}$$

where $\hat{\beta}_{n+h|n,k}$ denotes the h-step-ahead forecasts of $\hat{\beta}_{n+h,k}$ using a univariate time series. The forecasted curve of Equation (9) can be obtained by conditioning on the set of smoothed functions $f(x) = [f_1(x), f_2(x), \ldots, f_n(x)]^T$and the fixed functional principal components $B = [\hat{\Phi}_1(x), \hat{\Phi}_2(x), \ldots, \hat{\Phi}_K(x)]^T$.

Meanwhile, using a one-period ahead model, iterated forward for the number of periods desired, the Iterative One-Step ahead time series forecasts can be obtained. For example, as a comparison for the above forecast, the two (2) iterative one-step-ahead forecasts can be obtained as illustrated below:

Iteration 1:

$$\hat{y}_{366} = \hat{y}_{365+1|365}(x) = E[y_{365+1}(x)|f(x), B] = \bar{f}(x) \sum_{k=1}^{K} \hat{\beta}_{365+1|365,k}\hat{\Phi}_k(x)$$

Iteration 2:

$$\hat{y}_{367} = \hat{y}_{366+1|366}(x) = E[y_{366+1}(x)|f(x), B] = \bar{f}(x) \sum_{k=1}^{K} \hat{\beta}_{366+1|366,k}\hat{\Phi}_k(x)$$

Performance indicator

Several performance indicators are considered to describe the accuracy measures of the two applied models, the Multi-Step ahead and Iterative One-Step ahead forecast by investigating residuals of a prediction model in FDA. The residuals are also known as functional errors which can be computed based on the L^{p} norm. The respective formulae of the performance indicator are as follows.

The Functional Mean Average Error (FMAE), also called the Mean Integrated Average Error (MIAE), as given below:

$$FMAE = N^{-1}\sum_{t=1}^{T} ||Y_t - \hat{Y}_t||_{L_1} = N^{-1}\sum_{t=1}^{T} \int |Y_t - \hat{Y}_t(u)\,|\,du \tag{10}$$

The functional Mean Square Error (FMSE), also called the Mean Integrated Square Error (MISE), is defined as

$$FMSE = \sqrt{N^{-1}\left\|Y_t - \hat{Y}_t\right\|_{L2}^{2}} = \sqrt{N^{-1}\sum_{t=1}^{T}\int (Y_t - \hat{Y}_t(u))^2 du} \tag{11}$$

Meanwhile, functional MAPE is defined as the norm of the residual normalized by the norm of the real functional observation such that

$$FMAPE = N^{-1}\sum_{t=1}^{T}\left(||Y_t - \hat{Y}_t||_{L_2}/||Y_t||_{L_2}\right) = N^{-1}\sum_{t=1}^{T}\left(\int |Y_t(u) - \hat{Y}_t(u)\,|du \,/\, \int |Y_t(u)\,|du\right) \tag{12}$$

Prediction interval

For predictive inference purposes, a functional prediction interval is obtained. A prediction interval is an estimate of an interval for which a future curve will fall within a certain probability. The 95% prediction interval for $y_{n+h}(x)$ is given by

$$\hat{y}_{n+h|n}(x) \pm z_{0.025}\sqrt{\eta_{n+h}(x)} \tag{13}$$

where $z_{0.025}$ is the $\left(1-\frac{0.05}{2}\right)$ standard normal quantile. To interpret, a 95% confidence interval consists of a range of values with 95% confidence that it contains the true mean of the population curve. The term $\eta_{n+h}(x)$ is the forecast variance. The term can be approximated by the sum of component variances such that,

$$\eta_{n+h}(x) = Var[y_{n+h}(x)|f(x),\beta] = \hat{\sigma}_{\mu}^{2}(x) + \sum_{k=1}^{K} u_{n+h,k}\phi_k^2(x) + v(x) + \sigma_{n+h}^{2}\,(x) \tag{13}$$

where $u_{n+h,k} = Var(\beta_{n+h,k}|\beta_{1,k},\beta_{2,k},\dots,\beta_{n,k})$ and it can be obtained from the time series model while the model error variance $v(x)$ is estimated by averaging $\{\hat{\varepsilon}_1^2(x),\hat{\varepsilon}_2^2(x),\dots,\hat{\varepsilon}_n^2(x)\}$ while the term $\hat{\sigma}_{\mu}^{2}(x)$ is computed from the nonparametric smoothing methods used in Hyndman and Ullah [5].

In this study, the validation on the capability of the chosen FTS model is assessed using the 95% prediction interval. This is done by comparing how does the forecasted and the actual curve lies within the estimated functional confidence interval. The model is said to be capable, if both actual and forecasted curves lies within the interval while exhibiting not much different in the pattern.

Results and discussion

The following section discusses the results of analysis including the physical view of monthly maximum functional time series as the results of the data conversion process, functional descriptive analysis and then followed by the results of the forecasting models and the performance.

Functional descriptions of monthly diurnal maximum API data

The monthly maximum diurnal curves for Shah Alam and Pasir Gudang are shown in Figure 1. The majority of the curves represent the normal level, which is seen to have a peak at around hour 3.00 pm for most of the months. There are two significant groups of curves across the 24 hours. A group of curves with consistent behavior was observed at the lower level of maximum API and more a sparser curve at the higher API level indicated anomalous behavior that records the peak at around 10.00 am. High API exceeds 100 indicates unhealthy air quality. The deterioration of air quality is believed due to local and transboundary pollution [10].The existence of several extreme API curves which magnitude levels much higher than the rest are clearly shown Figure 1 for both Shah Alam and Pasir Gudang. In Shah Alam, the five identified outlying curves; arranged in descending order, were found occured on October 2015, June 2013, September 2015, July 2011 and March 2014. In Shah Pasir Gudang there were three identified extreme curves. The highest was recorded on June 2013, followed by September 2015 and the third highest was on October 2015.

As shown in Figure 2, the functional mean of monthly maximum diurnal API at both locations fluctuate and exhibits an increasing pattern after sunrise reaching the peak at 10.00 am at both sites, which inform that this is the time with a higher contribution of vehicles emission. The API is then gradually decreased after the hour and becomes constant after 8.00 pm. Day to day basis, the air quality in Shah Alam and Pasir Gudang is under healthy conditions as the mean level is within a good to a moderate status category below 100. As shown by the standard deviation curves monthly maximum diurnal API in Pasir Gudang is more volatile compared to Shah Alam with the highest variation at 10.00 am at both sites. The lowest variation occurs at 8.00 pm in Shah Alam and 3.00 pm in Pasir Gudang.

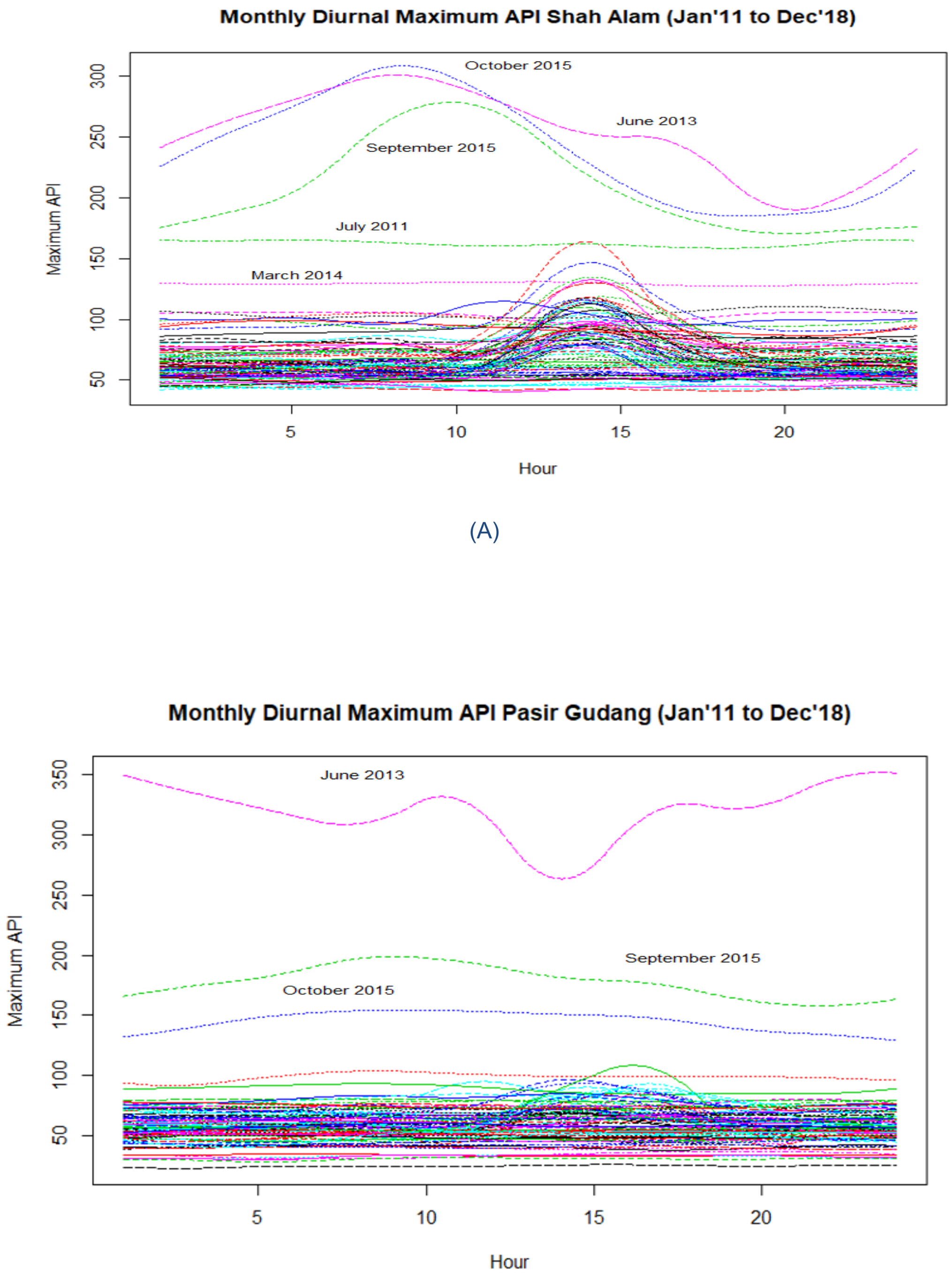
Multi-step ahead and iterative one-step ahead forecasting models: A comparison analysis

The following are the results of the Multi-Step ahead forecast model and Iterative One-Step ahead forecast model for monthly diurnal maximum API that has been developed based on historical data for the two monitoring stations; Shah Alam and Pasir Gudang from January 2011 until June 2018. The obtained model is then used to predict the next following month's curves of July 2018 until December 2018. The performance of these two models is assessed based on their capability to forecast.

Figure 3 shows the forecasted monthly maximum API of Shah Alam and Pasir Gudang from July 2018 to December 2018 highlighted in the overlapped blue color curves, while the data used for estimation are greyed out. The forecasted curves of the Multi-Step ahead forecast method are forecasted into six similar patterns of API monthly maximum diurnal curves for the month July until December 2018. Figure 3 (A) also shows that the blue curve in Shah Alam depicts that the maximum API achieved the highest peak level of about 80 at hour 3:00 p.m. The curve is below 100 level throughout the 24 hours of the forecasted months indicating that Shah Alam’s air quality is predicted to be in a healthy state for the second half of the year 2018. Meanwhile, in Figure 3 (B), the blue curve indicates that the future maximum API in Pasir Gudang is predicted to have a peak level of about 53 at 3.00 pm. then remained at the lower level during the night. Thus, these forecasting results give information to the public that Pasir Gudang’s air quality status for the second half of the year 2018 is in a good to moderate status as the API is below 100.

Figure 4 shows the forecasted month of maximum API of Shah Alam and Pasir Gudang respectively using Iterative One-Step ahead forecast. The forecasted curves, from month July 2018 to December 2018 produce a slightly different pattern of curves highlighted by rainbow-colored curves, while the data used for estimation are the grey-colored curves. This method produces some variations of forecasted curves.

The results suggest that, in forecasting the monthly diurnal maximum API data set, the two (2) models of Multi-Step ahead and Iterative One-Step ahead models have recognized that Malaysia's air pollution has varied for each location. Overall the predicted levels across the 24 hour period for the month is within the API range between 51-100 which indicates a moderate level of health concern Over exposure can bring potentially mild impact to extremely sensitive groups such as those who have asthma and the elderlies. Both location is predicted to have a common peak time around hour 3.00 pm.



Monthly Diurnal Maximum API Shah Alam (Jan'11 to Dec'18)

(A)

Monthly Diurnal Maximum API Pasir Gudang (Jan'11 to Dec'18)

(B)

Figure 1. Monthly Diurnal Maximum API Data Curves of (A) Shah Alam and (B) Pasir Gudang.

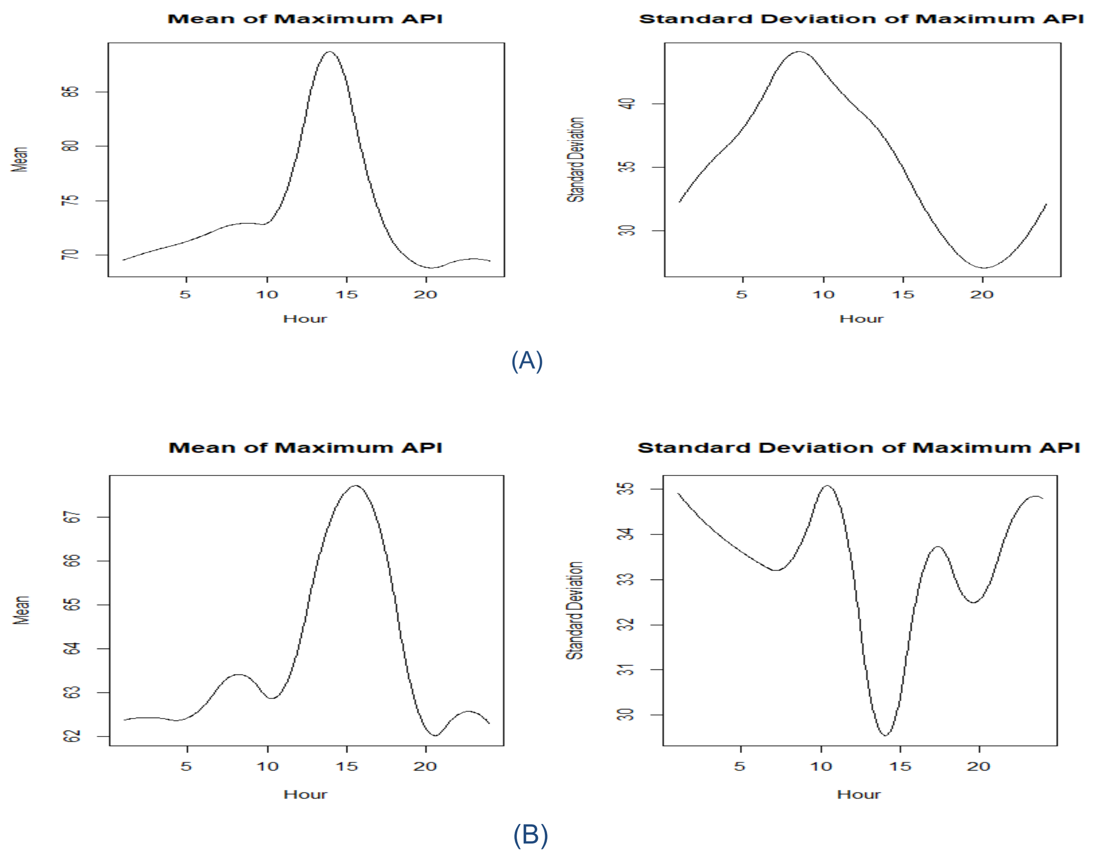
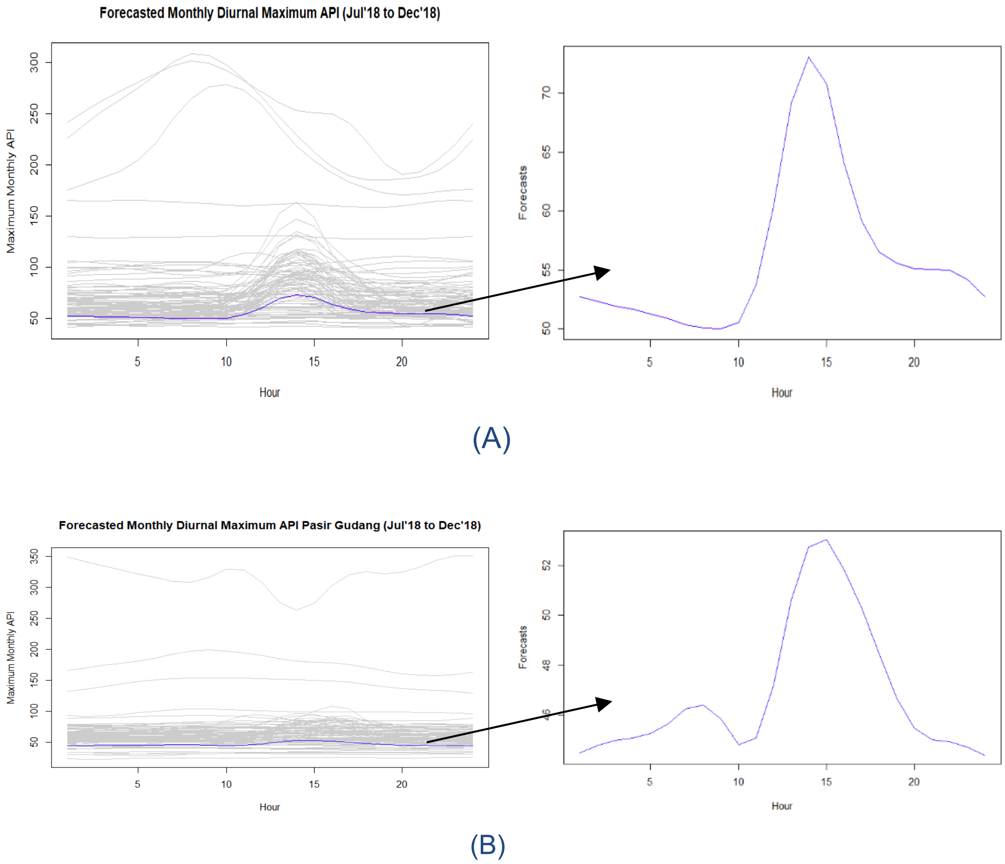


Figure 2. (A) Functional mean and functional standard deviation of monthly maximum diurnal API Shah Alam, (B) Functional mean and functional standard deviation of monthly maximum diurnal API Pasir Gudang.

Table 2 and Table 3 summarize the results on the performance of the two models approach when applied to the testing data set for the two locations.

The two models are tested by comparing the observed API readings and the forecasted readings during July – December 2018. Based on the results, it is concluded that the Multi-Step ahead model is better in predicting the maximum API data because the model produces lower errors of FMSE, FRMSE and FMAPE; 91.17, 9.55 and 11.77 respectively for Shah Alam and 46.78, 6.84 and 6.62 respectively for Pasir Gudang. Even though the results indicate that the Multi-Step ahead is a better model compared to the Iterative One-Step ahead, however, the error measures are not much different. Their similar performance is due to the modeling procedures for both methods do not differ. Both are using the same number of ten (10) basis in producing smooth curves. The only difference between those two (2); Multi-Step ahead forecast and the Iterative One-Step ahead forecast is the Multi-Step ahead times series forecasts are created using a horizon-specific forecasted model while Iterative One-Step ahead forecasts are created using a one-period ahead model, iterated forward for the number of period desires. A previous study by McElroy [9] also stated the same that if the model is fixed, and the same parameters are used for the direct (Multi-Step ahead forecast) and iterative forecasting formulas, then the forecasts are virtually indistinguishable. Thus, further research for exploring this matter is suggested.



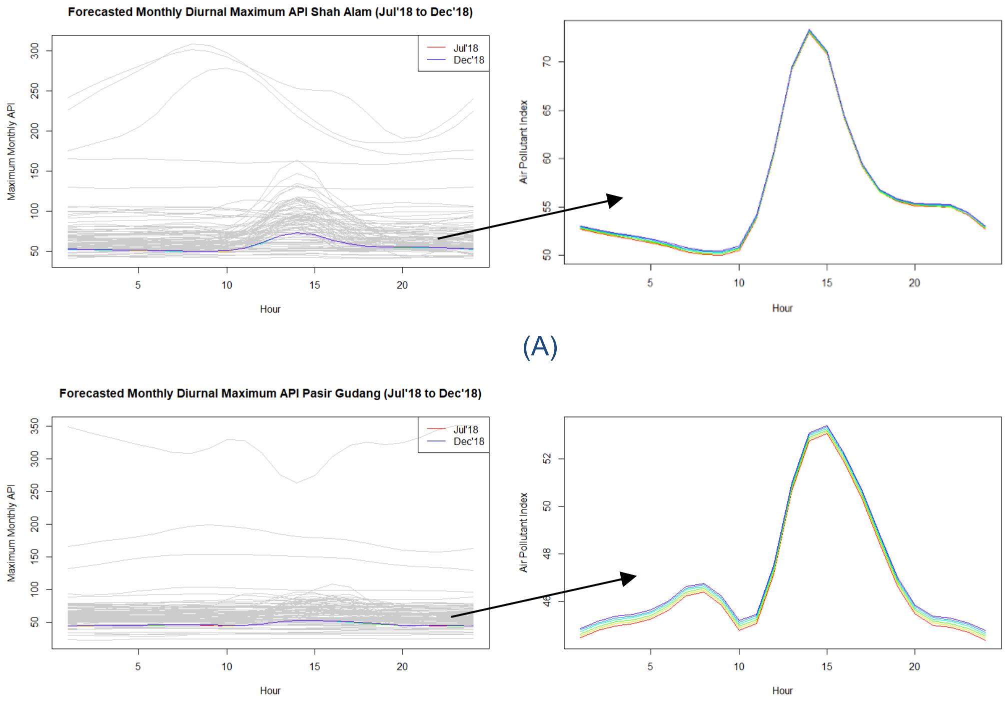
(A)

(B)

Figure 3. (A) Multi-Step ahead forecasted monthly diurnal maximum API curves for Shah Alam (July 18 to December 18), (B) Multi-Step ahead forecasted monthly diurnal maximum API curves for Pasir Gudang (Jul'18 to Dec'18).

Estimation using prediction confidence interval

Figure 5 illustrates the 95% functional confidence interval of the forecasted maximum API curves for Multi-Step ahead and Iterative One-Step ahead forecast for Shah Alam and Pasir Gudang air monitoring stations for month August 2018 and November 2018. The forecasted Multi-Step ahead, the Iterative One-Step ahead and the actual curve is indicated by blue, green and dark colored curves respectively. For each predicted month, both model produces the forecasted curves to be within the upper and lower confidence limits curve (i.e. the red colored lines). Both models produces about the same pattern of forecasted curves as shown by the overlapping blue and green lines. The pattern of the forecasted curves is about similar with the actual API curve at both sites except a slight difference is observed on the peak of predicted API at around 3.00 pm in Shah Alam. This situation can be also justified as the existing model is developed based on the concept of average as shown by equation (9). Supported by the findings in Figure 2, on the average the peak of API occurs at around 3.00 pm thus this provide the tendency for having a slightly higher API at that particular hour.



(A)

(B)

Figure 4. (A) Iterative One-Step ahead forecasted monthly diurnal maximum API curves for Shah Alam (Jul’18 to Dec’18), (B) Iterative One-Step ahead forecasted monthly diurnal maximum API curves for Pasir Gudang (Jul’18 to Dec’18).

Table 2. Comparison Model Performance Shah Alam.

Errors	Multi-Step Ahead Forecast	Iterative One-Step Ahead Forecast
FMAE	91.1719	92.7854
FRMSE	9.5484	9.6325
FMAPE	11.7742	12.1009

Table 3. Comparison Model Performance Pasir Gudang.

Errors	Multi-Step Ahead Forecast	Iterative One-Step Ahead Forecast
FMAE	46.7837	48.5807
FRMSE	6.8399	6.9700
FMAPE	6.6192	6.3674

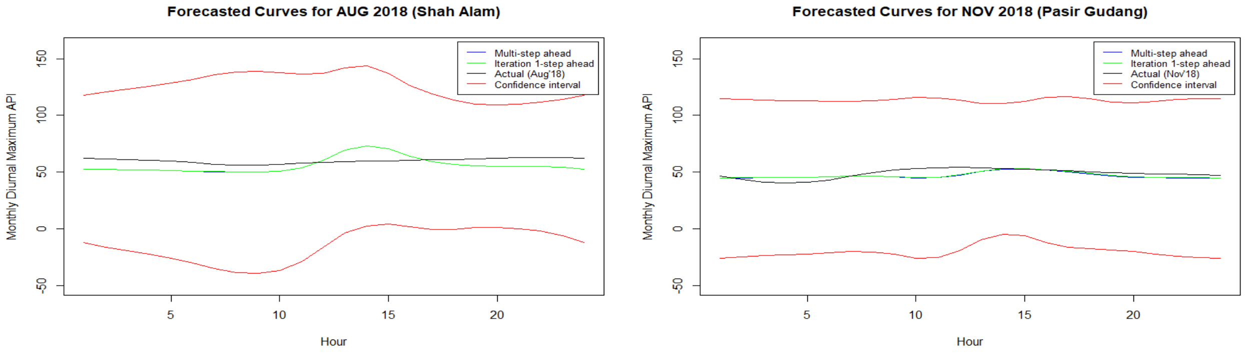


Figure 5. Comparison between Forecasted Curves and Actual Data for November 2018 (Pasir Gudang) using Curves Confidence Interval.

Based on the figure, it can be 95% confident that the forecasted maximum diurnal API is between the upper and lower line curves. This result also helps to confidently infer the future levels of maximum API at any time within 24 hours for that particular future month.

Conclusions

This study highlights the application of Functional Data methods to analyze monthly diurnal API maximum with the aims to visualize, evaluate and predict the monthly day-to-day (diurnal) API level. The application of the methods is inspired by the nature of the continuous nature of air pollutants concentration in the air. In specific, this study was carried out to investigate the performance between Multi-Step ahead and Iterative One-Step ahead functional time series models in predicting monthly diurnal maximum API curves. It is hoped that the model can be used as an alternative predictive tool. The model enables to visualization of the future pattern of monthly diurnal maximum API curves at two industrial sites in Malaysia which include Shah Alam and Pasir Gudang.

Based on the results of the functional descriptive analysis in the preliminary investigation, the results have shown that the monthly day-to-day maximum API level within a day, on average is considered under healthy level (below 100). The variation in Pasir Gudang is more volatile in comparison to Shah Alam which is found high within the day at both locations with the peak occurred between 9 to 10 am. The results have provided the insight that air pollution emission from the transportations activities can be the reason since this interval hour is identified as busy hours of vehicles movement at the two locations.

Between the two functional time series models discussed, the Multi-Step ahead forecast model has produced a better performance as shown by the lowest FRMSE and FMAPE; 9.55 and 11.77 respectively compared to the Iterative One-Step ahead forecast model. The capability of the multi-step ahead model is validated by comparing the forecasted future month diurnal curves with the actual observed data. The pattern of the behavior is quite similar. The forecasted API level also shown can be estimated using confidence interval for functional data that enable the predicted API level to be determined continuously over 24 hours.

In general, the application of Functional Times Series (FTS) analysis in this study gives more advantage over the conventional time series model. FTS model provides the ability to visualize, describe, evaluate and predict the continuous variation of monthly diurnal air pollution over a continuum of 24 hours. On the other hand, in the conventional time series model, the method only produces a single point average diurnal API. The information is static since the data only inform the average future API. In the context of an application, this study has shown that the FTS model enables the assessment for the future health effect of air pollution across 24 hours period. Thus, these results provide greater information efficiency to the responsible bodies and the researcher on risk and exposure of air pollution.

Further investigation on FTS model and its application can also be done. For future research, the researcher may identify the effect of different curve smoothing techniques and the sample size on the model performance. An alternative approach, other than the functional mean, the researcher can also considers functional median or functional mode as the average measure in the concept and theory of FTS model development.

Data availability

The data that has been used in this study was provided by the Department of Environment (DOE) Malaysia.

Conflicts of interest

There is no conflict of interest regarding the publication of this paper.

Funding statement

This research is not funded or granted by any grant.

Acknowledgments

The authors would like to thanks the Department of Environment Malaysia for providing the data for the study analysis.

References

[1] Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., Aziz, N. A. A., Azaman, F., Latif, M. T., Zainuddin, S. F. M., Osman, M. R., & Yamin, M. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, and Soil Pollution* 2014, 225(8).

[2] Department of Environment. A Guide to Air Pollutant Index (API) in Malaysia. Kuala Lumpur, Malaysia. *Department of Environment* 2000. Retrieved 12th June 2020. Online: https://issuu.com/ universititeknologimalaysia/docs/ a_guide_to_pollutant_index__api__in

[3] Gao, Y., Shang, H. L., & Yang, Y. High-dimensional functional time series forecasting: An application to age-specific mortality rates. *Journal of Multivariate Analysis* 2019, 170: 232-243.

[4] Hyndman, R. J., & Shang, H. L. Forecasting functional time series. *Journal of the Korean Statistical Society 2009*, 38(3): 199–211.

[5] Hyndman, R. J., & Ullah. Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics and Data Analysis*, 2007, 51(10): 4942–4956.

[6] Kokoszka, P., Miao, H., & Zhang, X. Functional dynamic factor model for Intraday price curves. *Journal of Financial Econometrics* 2012, 13(2): 456–477.

[7] Kokoszka, P., & Reimherr, M. *Introduction to Functional Data Analysis*. Boca Raton 2017: Chapman and Hall, CRC Press.

[8] Lee, M. H., Rahman, N. H. A., Suhartono, Latif, M. T., Nor, M. E., & Kamisan, N. A. B. Seasonal ARIMA for forecasting air pollution index: A case study. *American Journal of Applied Sciences* 2012, 9(4): 570–57

[9] McElroy, T. When are direct multi-step and iterative forecasts identical. *Journal of Forecasting* 2015, 34(4): 315–336.

[10] Mutalib, S. N. S. A., Juahir, H., Azid, A., Mohd Sharif, S., Latif, M. T., Aris, A. Z., & Dominick, D. Spatial and temporal air quality pattern recognition using environmetrics techniques: A case study in Malaysia. *Environmental Sciences: Processes and Impacts* 2013, 15(9): 1717–1728.

[11] Rahman, N. H. A., Lee, M. H., & Latif, M. T. Forecasting of air pollution index with artificial neural network.

Jurnal Teknologi (Sciences and Engineering) 2013, 63(2): 59–64.

[12] Ramsay, J. O., & Silverman, B. W. Functional data analysis. 2nd Edition. New York 2006: Springer.

:[13] Shaadan, N., Deni, S. M., & Jemain, A. Application of functional data analysis for the treatment of missing air quality data. *Sains Malaysiana* 2015, 44(10): 1531–1540.

[14] Shang, H. L., & Hyndman, R. J. Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation* 2011, 81(7):1310–1324.

[15] Siew, L. Y., Chin, L. Y., & Wee, P. M. J. ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. *Malaysian Journal of Analytical Sciences*, 2008 12(1): 257-263.