

Hybrid Seasonal ARIMA and Artificial Neural Network in Forecasting Southeast Asia City Air Pollutant Index

Nur Haizum Abd Rahman^{*1}, Muhammad Hisyam Lee², Suhartono³, and Mohd Talib Latif⁴

¹*Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.*

²*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.*

³*Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, 60111 Indonesia.*

⁴*School of Environmental and Natural Resource Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor, Malaysia.*

**Corresponding author: nurhaizum_ar@upm.edu.my*

The rise of air pollution has received much attention globally. As an early warning system for air quality control and management, it is important to provide precise future concentrations pollutant information. Using time series forecasting methods, the forecast of daily Air Pollutant Index (API) is presented here. The hybrid method between seasonal autoregressive integrated moving average (SARIMA) and artificial neural network (ANN) are chosen. To verify, the accuracies are measured using error magnitude approach. However, evaluation of forecasting API is also influenced by the health classification based on the threshold value assigned in air quality guidelines. Thus, forecast accuracies based on index value, namely as true predicted rate (TPR), false positive rate (FPR), false alarm rate (FAR) and successful index (SI) are also used for forecast validation. As shown in the results, the hybrid model performs better in both model's evaluations group used. Hence, the hybrid method must be considered in the forecasting area due to the capability to analyze real data consisting of both linear and nonlinear patterns. Besides, using the appropriate measurement in accordance to the purpose of forecasting is important to produce an accurate forecast. **Keywords:** air pollutant index (API), artificial neural network, forecasting evaluation, hybrid, SARIMA.

I. Introduction

Unpolluted air is considered to be a crucial necessity for human health and well-being. However, air pollution continues to pose a major threat to health globally. The presence of globalized development for both developed and developing countries has contributed to an increase in air pollution problems (Hassanzadeh et al., 2009). Air pollution has been found to affect human health, the environment and in the long run, tends to exacerbate risks to the earth, by increasing global warming and the

greenhouse effect (Heo and Kim, 2004, Kumar and Jain, 2010, Kurt and Oktay, 2010). In order to reduce the worst impacts of air pollution, air quality guidelines are designed.

Among the principal air quality guidelines (AQGs) were produced in 1987 by the World Health Organization (WHO), which are an organization that direct and coordinate authority on international health within the United Nations' system. The guideline values recommended by WHO are important in guiding to build the air quality standards set by each country in order to protect the pub-

lic health developing national management and environmental policies. As an example, from the standpoint of human health implications, Malaysia Ambient Air Quality Standard (MAAQS) used the WHO standard to measure the pollutants. One of the simplest developed air quality guidelines is measured based on an index. Thus, the ambient air quality measurement in Malaysia was described in terms of Air Pollutant Index (API).

The API scale and terms based on human health implications have been used for several years (Kumar and Goyal, 2011). The API in Malaysia was developed based on the API introduced by the United State Environmental Protection Agency (USEPA) and is determined by the calculation of sub-indexes of the five main pollutants, namely particulate matter (PM_{10}), ozone (O_3), carbon dioxide (CO_2), sulphur dioxide (SO_2) and nitrogen dioxide (NO_2). It can be used by government agencies to characterize the status of air quality since API providing an easy assessment and harmonization at the given location throughout the countries (Sansuddin et al., 2011). API also helps the general public to understand easily the air quality status for their own health precaution. Therefore, the API scales and terms are used in assessing and describing the air quality status on human health. This information with different ranges reflects as “Good (0-50), Moderate (51-100), Unhealthy (101-200), Very Unhealthy (201-300) and Hazardous (301 and above)” can be the benchmark of air quality management or data interpretation in decision-making processes (Afroz et al., 2003).

Warning systems for air quality are required in order to provide accurate information and act as advance notice whether the ambient pollution levels might exceed the air quality guidelines or limit values. Forecasting procedures involving time series analysis are one of the warning systems that provide reliable and effective in air pollution control measures (Kumar and Goyal, 2011, Sansuddin et al., 2011). Using time series, the historical data observations are analyzed to develop a model that describes the

relationship between time and variable and will be used to extrapolate the time series into future (Cryer, 1986). Thus, the sensitive groups in the population concerned, such as children, asthmatics and elderly people could get the benefit from the forecast.

Statistical methods are usually used to analyze and forecast data with linear methods and have been applied widely to a lot of data series including air quality (Chelani and Devotta, 2006, Moustris et al., 2010, Slini et al., 2002, Wang and Wei-Zhen, 2006). The linear method introduced by Box-Jenkins in 1976 is among the widely important classical time series models (Suhartono, 2011, Zhang, 2003). In previous research done by Kurt and Oktay (2010), Wang and Wei-Zhen (2006) and Ibrahim et al. (2009), the performance of Box-Jenkins, or also referred to as autoregressive integrated moving average (ARIMA) model, gives satisfactory results in forecasting the daily pollutants such as CO , O_3 , NO and NO_2 . However, the major limitation in ARIMA is that the model only can capture the linear form of time series data and suffers from the assumptions of stationarity and linearity.

Artificial neural networks (ANNs) overcome the limitation of ARIMA and become the comparison of linear methods. ANNs are a branch of artificial intelligence that was firstly developed in the 1950s (Moustris et al., 2010). The ANNs aim in imitating the biological brain architecture. Progressive growth of ANN has been reported in literature, which is mainly interested in developing convenient hardware for fast data analysis and information processing (Khashei and Bijari, 2010). The significant advantage of the ANNs model is that no prior assumption of the model form is required in the model building process because the network model is highly determined by the characteristics of the data.

ANNs are nonlinear methods in forecasting with unspecified particular model form. The ANNs are flexible in modelling any type of relationship with high accuracy and are widely used forecasting models with a lot of applica-

tions (Khashei and Bijari, 2010, Palmer et al., 2006, Zhang et al., 1998, Zhang and Qi, 2005). Moreover, the ANNs suit the problems with huge amount of data (Moustris et al., 2010, Papanastasiou et al., 2007). In air quality literature, the ANNs have been used for forecasting in many years with a wide range of pollutants with impressive results, which have proven to be the better approach compared to the linear methods (Brunelli et al., 2007, Gardner and Dorling, 1998, 1999).

The classical method, ARIMA and the modern method ANNs both have their own strength to analyze their own pattern data, where ARIMA is suitable for linear pattern while ANNs are suitable for nonlinear pattern. Even though in the literature, modern method can outperform classical method, there are still mixed views in conclusions where some researchers found that classical is better than the modern method (Makridakis and Winkler, 1983, Zhang, 2003). Moreover, the final chosen methods might not necessarily be suitable to be used in future as such changes that can exist in the data pattern and sample variation could affect the analysis. To overcome these limitations, a combination of different methods, also known as a hybrid method, has been introduced.

One of the hybrid approaches in time series forecasting is using both ARIMA and ANN models where ARIMA is known as successful linear model while ANN usually use as nonlinear model. Previous research has proven that hybrid or combining several models often lead to improvement in the forecasting accuracy (Bates and Granger, 2001, Makridakis et al., 1982, Shen et al., 2011). Chelani and Devotta (2006) and Wang and Wei-Zhen (2006) are among the earliest studies of hybrid methodology in air quality applications. Chelani and Devotta (2006) used hybrid between ARIMA and nonlinear dynamical modelling in forecasting the NO_2 concentrations in Delhi. Wang and Wei-Zhen (2006) used multi-layer perceptron (MLP) trained with a particle swarm optimization algorithm (MLP-PSO) and a hybrid

Monte Carlo (HMC) method. This was applied for ground level O_3 forecasting in Hong Kong. Besides, a study by Díaz-Robles et al. (2008) used hybrid ARIMA and ANNs model to forecast PM_{10} in urban areas in Temuco, Chile.

At the end of the forecast, the appropriate model's validation is required in accordance to the purpose of forecasting so that the validation of forecast message is useful to the public, government and authorities. Mean absolute percentage error (MAPE), mean absolute deviation (MAD), mean square error (MSE) and root mean squared percentage error (RMSE) are among the measurements commonly used to assess forecast accuracy. Generally, this type of measurement evaluates how vast the difference is between the forecasted and the actual or observed data. However, a forecast system may have different objectives. In the field of air quality, it is important to maintain the air quality within the guidelines. Thus, forecast accuracy based on threshold value, namely as true predicted rate (TPR), false positive rate (FPR), false alarm rate (FAR) and successful index (SI) are taken into consideration for forecast validation.

The purpose of the present study is to develop models in forecasting API value. This is achieved by developing ARIMA models, ANNs and hybrid method in order to forecast the API value. Finally, based on the model's validation, the most accurate forecast will be determined.

II. Materials and Methods

A. Description of Sampling Sites

Johor, the southern state of the Peninsula of Malaysia located in Southeast Asia, was selected for this study due to the variety of locations it offered – urban and suburban. Moreover, these locations have often been affected by trans-boundary pollution from the neighbouring countries (particularly from biomass burning in Sumatra, Indonesia), which is usually the main factor behind hazardous occurrences. The dataset was obtained from the De-

partment of the Environment (DoE) Malaysia through Alam Sekitar Malaysia Sdn. Bhd., a private company which is responsible for monitoring and managing ambient air quality through a network of 51 stations. As shown in Figure 1, the station located in Johor Bahru city (N0129.815, E10343.617), the capital of Johor state, was established in December 1996 and this city is the second largest metropolitan area in Malaysia after the capital city, Kuala Lumpur (Rizzo and Glasson, 2012). Besides, it is home to a large number of the region's industries, residential, and commercial hotspots, and therefore, has congested roads.

The study used daily data set for seven years, which covered the period from 2005-2011. The data were divided into two data sets: (1) a training data set of 2005 –2010 (2191 observations) to identify the API model, and (2) a test data set for 2011 with a total of 365 observations to check the performance of the model. The data plot is shown in Figure 2.



Figure 1: Sampling station location

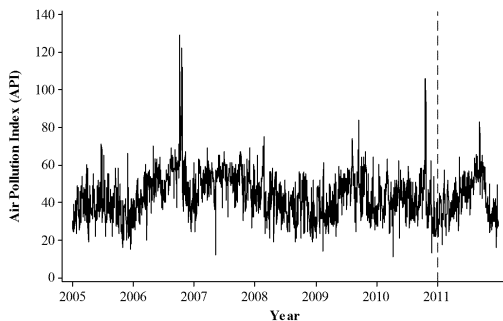


Figure 2: Air Pollution Index (API) recorded at the sampling stations from 2005 until 2011

B. Hybrid Seasonal ARIMA and Artificial Neural Network

The basic idea of hybrid method in forecasting is to make use the unique feature of the single methods to capture the different patterns in the data. Mainly, hybrid between the linear pattern and nonlinear pattern, such as the method used for each patterns, have limited capability depending on the characteristics of the data (Tseng et al., 2002, Zhang, 2003). The linear methods are only capable of detecting and modeling linear data pattern, meanwhile nonlinear methods for nonlinear data pattern. Therefore, with this capability, both linear and nonlinear time series models will be combined in order to achieve better forecasting performance (Taskaya-Temizel and Casey, 2005, Zhang, 2003).

A hybrid model comprising of a linear and a non-linear component has been employed in the experiments (Zhang, 2003):

$$Y_t = L_t + N_t \quad (1)$$

where, L_t is denoted as the linear component and N_t as the nonlinear component.

The concept of the hybrid model consists of two primary steps. Initially, data will be analyzed with the linear model. In this study, the linear model is estimated using SARIMA method. From the SARIMA model, the linear part of hybridization is obtained. Then, from the residual of linear model, the nonlinear model will be discovered. For non-linear patterns, the neural networks are used. In this way, the hybrid model is exploiting the strength of both components.

The Box-Jenkins or ARIMA is classified as a linear model that is capable of presenting both stationary and non-stationary time series. Most researchers have used this model to forecast univariate time series data. Three main steps that must be considered in building the model for forecasting are: a) tentative identification, b) parameter estimation, and c) diagnostic checking (Hanke and Wichern, 2005). ARIMA is flexible in represent-

ing different types of time series; autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA). In the case where seasonal components are included in the model, the ARIMA model is called as SARIMA model and abbreviated as SARIMA. Generally, the SARIMA model can be written as:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^DY_t = \theta_q(B)\Theta_Q(B)\epsilon_t \quad (2)$$

where

$$\begin{aligned} \phi_p(B) &= 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \\ \Phi_P(B) &= 1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_P B^P \\ \theta_q(B) &= 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q \\ \Theta_Q(B) &= 1 - \Theta_1 B - \Theta_2 B^2 - \dots - \Theta_Q B^Q \end{aligned}$$

where B is denoted as the backward shift operator, d and D are denoted as the non-seasonal and seasonal orders of difference respectively and ϵ_t is the error.

For non-linear patterns, the artificial neural networks (ANNs) are used. Many different ANN models have been proposed. The most influential ANNs used is multi-layer perceptron (MLP) (Moustris et al., 2010, Sarle, 1994). The model is characterized by three simple layers of processing units: the input layer, the hidden layer, and the output layer, which are connected to each other. Each layer consists of processing elements (PEs), called neurons. The input layer receives the external information whereby the input nodes correspond to the number of variables or the number of lagged observations are used to discover the underlying pattern in the time series (Zhang et al., 1998). The last or the highest layer is the output layer, where the problem solution is obtained. The input and output layers are separated by one or more intermediate layers called as the hidden layer. Most authors used only one or two hidden layers for forecasting purposes (Zhang et al., 1998). Hidden nodes are used to process the information received from the input nodes to perform nonlinear mapping between input

and output. Without hidden nodes in the hidden layer, the ANN is equivalent to a linear statistical forecasting model. The hidden layer is included in the neural network system to increase the flexibility of the model. Arrows from the input to the hidden layer, and hidden layer to the output layer indicate the strength of each connection and can be measured by a quantity called weight.

The MLP relationship between the output, y_t and the inputs, $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ has the following mathematical representation:

$$y_t = w_0 + \sum_{i=1}^n w_i y_i \quad (3)$$

$$y_t = w_0 + \sum_{j=1}^q w_j g(w_{0,j} + \sum_{i=1}^p w_{ij} y_{t-i} + \epsilon_t) \quad (4)$$

where w_j ($j = 1, 2, \dots, q$) and w_{ij} ($i = 1, 2, \dots, p; j = 1, 2, \dots, q$) are the model parameters that are often called as the connection weights; p is the number of input nodes and q is the number of hidden nodes. Commonly, in the hidden layer, the activation function used is the logistic function, which is $f(x) = 1/(1+exp^{-x})$, meanwhile the linear function, $f(x) = x$, is used at the output stage.

C. Performance Evaluations

From available historical data, the information obtained was used to predict future data by using forecasting methods. These forecasted values usually have different values from the actual observations. Hence, the forecast results should be evaluated in order to measure the accuracy. Thus, appropriate conclusion could be made. However, the evaluation of forecast accuracy is highly dependable on the choice of accuracy measurement used. It is important to use the appropriate measurement according to the purpose of forecasting. The measure is divided into two aspects that are the statistical test based on the magnitude error between the observed and the predicted data and the other forecast errors based on the boundary conditions.

1. Error Magnitude Measurements

It is important to check the forecasting performance to identify the best model with the smallest error. There are several statistical tests that can be considered to measure the model validation. Normally, the measurements of forecasting accuracy are based on the mean absolute percentage error (MAPE), the mean squared error (MSE), the root mean squared error (RMSE), and mean absolute deviation (MAD). The MAPE, MSE, RMSE and MAD can be written as follows:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| (100); y_t \neq 0 \quad (5)$$

$$MSE = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2 \quad (6)$$

$$RMSE = \sqrt{MSE} \quad (7)$$

$$MAD = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (8)$$

where y_t is an actual value, $\hat{y}_t$ is a predicted value, and n is the number of the predicted value. MAPE, MSE, RMSE and MAD can be computed directly from the actual and forecast values and are not involved with unknown parameters that need to be estimated. This statistical computations are widely applied in literature (Elliott et al., 2005). All the statistical computations measure the residual errors with the smallest value and are chosen as the best model to be used in forecasting.

2. Index Measurements

One of the sources of forecast evaluation is the incorrect forecast error based on the boundary conditions, which are also known as the exceedance measures. In this case, the binary system is used to determine and compare the occurrence of forecast and observed data, whether

the occurrence exceeds the threshold value assigned (van Aalst and de Leeuw, 1997). Definitions and values of statistical indexes related to the model's ability to forecast reliably the exceedances of the threshold value can be defined as A, the number of occurrences in the forecast and observed, B, the number of occurrences only in observed, C, the number of occurrences only in the forecast and D, the number of non-occurrence in the forecast and observed. From the count number of A, B, C and D, the statistical indexes can be defined as the true predicted rate (TPR), false predicted rate (FPR), false alarm rate (FAR) and success index (SI). The TPR, FPR, FAR and SI can be computed as follows:

$$TPR = \frac{A}{A + B} \quad (9)$$

$$FPR = \frac{C}{C + D} \quad (10)$$

$$FAR = \frac{C}{C + A} \quad (11)$$

$$SI = \frac{A + D}{A + B + C + D} \quad (12)$$

All these indices have values ranging from 0 to 1. TPR represents the fraction between correct occurrences over a total of occurrences in observed, the perfect score will be equal to 1 with the fraction of unexpected occurrences is given by $(1 - TPR)$ or $B/A + B$. FPR represents the occurrences only in the forecast over total non-occurrences in observed while the FAR represents the occurrences only in the forecast over total occurrences in the forecast, the perfect score will be equal to 0. Finally, SI represents the correct forecast of occurrences and non-occurrences over the total number of forecasts, with 1 as the perfect score.

In some situation, these measures have the disadvantage of being infinite or undefined if the denominator is equal to 0. This infinite number reflects that no boundaries or limits

are impossible to measure or calculate. However, based on equation 9 to 12 presented previously, the values are based on the binary system where number of data presented the A , B , C and D with the equation based on the ratio between numerator and denominator. Therefore, the infinite has limit value, which is bounded between 0 and 1.

III. Results and Discussion

In this study, the classical method, seasonal autoregressive integrated moving average (SARIMA), the modern method, artificial neural networks (ANNs) and the advance method, hybrid method have been used in forecasting daily air pollution index (API). The methodology for analyzing the data consists of the following steps which are:

Step 1: Modelling of seasonal ARIMA by using Box-Jenkins procedure.

Step 2: Modelling of ANNs with three types of transformation data (normalize, $[-1,1]$, $[0,1]$) and with three types of input, i.e.

- a. The inputs based on the order of the best seasonal ARIMA model at the first step
- b. The inputs based on seasonal lag
- c. The inputs are lag 1 and seasonal lags with ± 1

Step 3: Modelling of hybrid model based on the best model from the first step and taking the residual from the first step and repeat the second step to perform hybrid forecast

Step 4: Compare the forecast accuracy between seasonal ARIMA, ANNs and hybrid model to find the best forecasting model.

The Box-Jenkins model was implemented via MINITAB version 16 and SAS software. The analysis begins with the time series plot of the data used to identify the existence of a trend or seasonal pattern. As shown in Figure 2, the API data contains a seasonal pattern with higher values during certain months, especially during the southwest monsoon between May to September. The main causes of high API data recorded within this period is because Malaysia was affected by trans-boundary

sources of biomass burning around Southeast Asia, especially from Sumatra, Indonesia (Samsuddin et al., 2011). Before ascertaining the tentative model, identification of the differencing process is necessary for both non-seasonal and seasonal parts to obtain the stationary data set before model development can be undertaken. As the data contain seasonal pattern, the seasonal ARIMA (SARIMA) was used.

By taking the difference for non-seasonal, and with for seasonal, the data had become stationary series. Then, the tentative model based on the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) from stationary series was determined to find the best combination order of SARIMA. Based on the Ljung-Box statistics, the adequacy of the tentatively identified model was identified. The result shown in Table 1 with the possible model were SARIMA(3, 1, 3)(0, 1, 1)³⁶⁵ and SARIMA(3, 1, 3)(2, 1, 0)³⁶⁵ with the best model based on error magnitude was SARIMA(3, 1, 3)(0, 1, 1)³⁶⁵. The model can be written as:

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)(1 - B)(1 - B)^{365} Y_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)(1 - \Theta_1 B^{365}) \epsilon_t \quad (13)$$

The ANNs was suggested as an alternative method of classical method in time series forecasting. ANN has been used widely, such as in sales, productions, and load forecasting areas (Julian and Chris, Perez and Salini, 2008, Zhang et al., 1998). For this study, the computations were performed using Matlab software. Initially, before applying the ANN, the data pre-processing was carried out. It included:

- a. Normalizing the original data into the standard normal distribution
- b. Making range of the original data into $[-1,1]$
- c. Making range of the original data into $[0,1]$

Next, in the selection of input nodes, the selected input, p was found through the ARIMA model (Julian and Chris, Zhang and Qi, 2005). The possibilities lags are 1, 2, 3, 4, 365, 366, 367, 368 and 369. As for the hidden layer, the single hidden layer is sufficient for any complex

nonlinear function (Zhang et al., 1998). With all three data pre-processing, input nodes based on ARIMA and hidden layers with up to five hidden nodes, the ANN model was built. The result is shown in Table 1.

Based on Table 1, the result shows that SARIMA obtained the best performance in forecasting API in the selected stations compared to ANNs. As mentioned in the introduction, many results in the air quality literature find ANNs to have the best performance. However, our results show contrary findings to previous studies. With this mixed conclusion, the analysis is extended to forecast the API using hybrid methodology. The basic principles and modeling process of the ARIMA and ANN models were used. Using the residual from the best model SARIMA model, the ANN model was built. Based on equation 4, the hybrid forecast was obtained and the results are shown in Table 1.

The forecast of API in the testing period was done using three different approaches, namely Box-Jenkins model, ANNs and hybrid models. Based on the results obtained in Table 1, the best forecasting accuracy within the approaches was chosen to be evaluated for comparison purposes. According to the MAPE, MAD, MSE, and RMSE, the conventional ARIMA model outperformed the ANNs model. These two methods showed inconsistent results in performing for forecasting API. In contrast, the advance method, hybrid model gave the most accurate forecast and had been consistent in all four magnitude error measurements as shown in Table 1.

Within the area of forecasting air quality, the error measurement to define the quality of forecast model to be able correctly forecast concentrations above the threshold should be taken into consideration. Normally, the concentration of fine dust, PM_{10} is the highest compared to other pollutants and so the PM_{10} determine the API levels in selected station. According to WHO and MAAQS air quality guidelines, the limit value assigned is $50 \mu g/m^3$. Thus, the API limit value of 50 was selected in order to

verify the forecast quality of developed models. To evaluate the performance of the exceedance forecasting, the frequency of the forecast is obtained as mentioned in section 2.3.2.

The out sample data chosen containing 365 data with 92 data exceeded and the remaining 273 did not exceed the assigned limit value. According to the values in Table 2, the ANNs are able to capture all the data that do not exceed the limit value. However, ANNs perform worst in capturing the exceedances in the observed data as shown in A. For SARIMA, the frequency in determining non-exceedances is higher than for hybrid model, but in determining the exceedance observed value, the hybrid model performs better. More importantly, forecast obtained is used to build a warning system that is able to give prior notice to the public especially among the sensitive groups. Thus, the frequency of B, the occurrence in observed, but not predicted is needed to be aware as SARIMA and ANN recorded higher number compared to hybrid. In other situation, the forecast value predicts above the threshold value, although the observed does not. This can be called as a false alarm. Denoted by C, the hybrid shows that 19 days of false alarm occurred.

For validation of the model, the performance statistics using the frequency information are defined. The hybrid model with the TPR values of 0.5 indicates that the model predicted exceeds successfully compared to SARIMA and ANN. In particular, the FPR with the best model equal to 0 is shown by the ANN following SARIMA and hybrid model. Earlier, in section 2.4.2, this measurement could not be defined. In this study, this situation occurs in finding FAR for ANN where the values of denominator are equal to zero. However, as mentioned previously, the value can be defined. Thus, in Table 2, the FAR infinite result of ANN model is defined by 0 since the frequency of nominator is equal to 0. Hence, ANN is the best model followed by hybrid and SARIMA.

In addition, Table 2 shows the SI measurement where the hybrid model gives the best

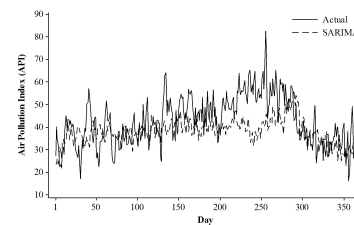
Table 1: Forecast accuracy comparison based on error magnitude for SARIMA, ANN and hybrid models

Models	MAPE	MAE	MSE	RMSE
SARIMA				
(3, 1, 3)(0, 1, 1) ³⁶⁵	19.39	8.36	110.60	10.52
(3, 1, 3)(2, 1, 0) ³⁶⁵	23.22	10.47	175.83	13.26
ANN				
All lags	22.17	8.70	111.76	10.57
Seasonal Lags 365,730	23.87	9.02	118.71	10.90
Lag 1 with Seasonal Lags ± 1	22.46	9.03	122.31	11.06
Hybrid				
All lags	13.91	5.41	46.39	6.81
Seasonal Lags 365,730	19.41	8.34	109.75	10.48
Lag 1 with Seasonal Lags ± 1	214.65	6.03	55.72	7.46

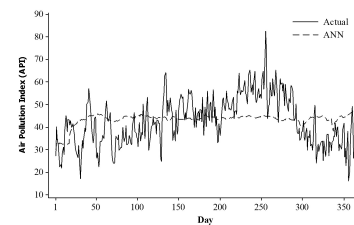
result in identifying the capability of the prediction to capture the observed either exceeding or not exceeding the agreed threshold value. In order to support that the models can predict accurately the exceeds based on the assigned limit, the values of TPR and SI must be high while FPR and FAR must be low. Thus, the hybrid model is better in TPR and SI measurements, while ANN is better in FPR and FAR measurements. Concerning both groups of forecast accuracy measurements in order to determine the best forecasting method, the hybrid model becomes the best developed model to forecast API values. Figure 3 shows the observed and forecasted value of the API using SARIMA, ANN and hybrid models.

Table 2: Performance statistics for the validation of SARIMA, ANN and hybrid models

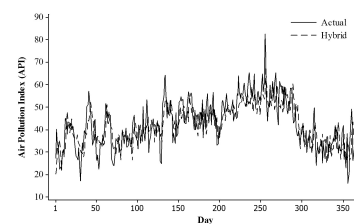
	SARIMA	ANN	Hybrid
Frequency			
A	1	0	46
B	91	92	46
C	7	0	19
D	266	273	254
Measurements			
TPR	0.01	0	0.50
FPR	0.03	0	0.07
FAR	0.86	0	0.29
SI	0.73	0.75	0.82



(a)



(b)



(c)

Figure 3: The forecasted air pollutant index (API) in 2011 (a) SARIMA, (b) ANN, and (c) Hybrid

IV. Conclusion

A detailed study of daily air pollution index (API) in urban areas, Johor Bahru located in Southeast Asia is reported in this paper. Based on the data recorded, the selected station shows seasonal data pattern with high API in the southwest monsoon (May to September) period due to biomass burning. Concerning the air pollution affecting the human health, forecasting using time series reported here is hoped to give an advance notice to the public and the authorities. Therefore, the seasonal autoregressive integrated moving average (SARIMA), artificial neural networks (ANNs) and hybrid method of SARIMA and ANNs were used to forecast the daily API.

Regarding the objective of the study to forecast the API with highest accuracy, various accuracy measurements have been tested. This includes the measurements based on the magnitude of forecasted and observed and the model's validation based on the ability to forecast the exceedances of the assigned limit value. Comparing the three models using magnitude error measurements (MAPE, MAE, MSE, RMSE), the hybrid model showed better skills in forecasting API compared to SARIMA model and ANNs model. If comparing the three models based on model's validation using exceedances indices, the hybrid model gives the best result in forecasting the capability to forecast these exceedances with a limit value of 50. Therefore, we can conclude that the hybrid method can be used for future forecasting of air pollutants since it is able to outperform single methods. For example, predicting the real data usually contains both linear and nonlinear pattern. In addition, in research on air quality concerning the guidelines in order to control the pollution, the chosen best forecast must consider the indices based on exceedance threshold value together with the statistical performance using the magnitude error.

Acknowledgements

This study was supported by the Universiti Putra Malaysia, Malaysia, under Putra Grant (IPM) 9587700.

References

- [1] Afroz, R., Hassan, M. N., and Ibrahim, N. A. (2003). Review of air pollution and health impacts in malaysia. *Environmental Research*, 92(2):71 – 77.
- [2] Bates, J. M. and Granger, C. W. J. (2001). Essays in econometrics. chapter The Combination of Forecasts, pages 391–410. Cambridge University Press, New York, NY, USA.
- [3] Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., and Vitabile, S. (2007). Two-days ahead prediction of daily maximum concentrations of so₂, o₃, pm₁₀, no₂, co in the urban area of palermo, italy. *Atmospheric Environment*, 41(14):2967 – 2995.
- [4] Chelani, A. B. and Devotta, S. (2006). Air quality forecasting using a hybrid autoregressive and nonlinear model. *Atmospheric Environment*, 40(10):1774 – 1780.
- [5] Cryer, J. D. (1986). *Time Series Analysis*. Prindle, Weber and Schmidt, United States, 1st edition.
- [6] Díaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., and Moncada-Herrera, J. A. (2008). A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of temuco, chile. *Atmospheric Environment*, 42(35):8331 – 8340.
- [7] Elliott, G., Timmermann, A., and Komunjer, I. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, 72(4):1107–1125.

- [8] Gardner, M. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627 – 2636.
- [9] Gardner, M. and Dorling, S. (1999). Neural network modelling and prediction of hourly nox and no2 concentrations in urban air in london. *Atmospheric Environment*, 33(5):709 – 719.
- [10] Hanke, J. E. and Wichern, D. W. (2005). *Business Forecasting*. Pearson, New Jersey, 5th edition.
- [11] Hassanzadeh, S., Hosseinibalam, F., and Alizadeh, R. (2009). Statistical models and time series forecasting of sulfur dioxide: a case study tehran. *Environmental Monitoring and Assessment*, 155(1):149–155.
- [12] Heo, J.-S. and Kim, D.-S. (2004). A new method of ozone forecasting using fuzzy expert and neural network systems. *Science of The Total Environment*, 325(1):221 – 237.
- [13] Ibrahim, M. Z., Zailan, R., Ismail, M., and Lola, M. S. (2009). Forecasting and time series analysis of air pollutants in several area of malaysia. *American Journal of Environmental Sciences*, 5(5):625–632.
- [14] Julian, F. and Chris, C. Time series forecasting with neural networks: a comparative study using the air line data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(2):231–250.
- [15] Khashei, M. and Bijari, M. (2010). An artificial neural network (p,d,q) model for timeseries forecasting. *Expert Systems with Applications*, 37(1):479 – 489.
- [16] Kumar, A. and Goyal, P. (2011). Forecasting of daily air quality index in delhi. *Science of The Total Environment*, 409(24):5517 – 5523.
- [17] Kumar, U. and Jain, V. K. (2010). Arima forecasting of ambient air pollutants (o3, no, no2 and co). *Stochastic Environmental Research and Risk Assessment*, 24(5):751–760.
- [18] Kurt, A. and Oktay, A. B. (2010). Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks. *Expert Systems with Applications*, 37(12):7986 – 7992.
- [19] Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, H., Parzen, E., and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1:111 – 153.
- [20] Makridakis, S. and Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, 29(9):987–996.
- [21] Moustris, K. P., Ziomas, I. C., and Paliatsos, A. G. (2010). 3-day-ahead forecasting of regional pollution index for the pollutants no2, co, so2, and o3 using artificial neural networks in athens, greece. *Water, Air, & Soil Pollution*, 209(1):29–43.
- [22] Palmer, A., Montaña, J. J., and Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management*, 27(5):781 – 790.
- [23] Papanastasiou, D. K., Melas, D., and Kioutsioukis, I. (2007). Development and assessment of neural network and multiple regression models in order to predict pm10 levels in a medium-sized mediterranean city. *Water, Air, and Soil Pollution*, 182(1):325–334.
- [24] Perez, P. and Salini, G. (2008). Pm2.5 forecasting in a large city: Comparison of three methods. *Atmospheric Environment*, 42(35):8219 – 8224.

- [25] Rizzo, A. and Glasson, J. (2012). Iskandar malaysia. *Cities*, 29(6):417 – 427.
- [26] Sansuddin, N., Ramli, N. A., Yahaya, A. S., Yusof, N. F. F. M., Ghazali, N. A., and Madhoun, W. A. A. (2011). Statistical analysis of pm10 concentrations at different locations in malaysia. *Environmental Monitoring and Assessment*, 180(1):573–588.
- [27] Sarle, W. S. (1994). Neural networks and statistical models. In *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, volume 1994, pages 1–13. SAS Institute Inc.
- [28] Shen, S., Li, G., and Song, H. (2011). Combination forecasts of international tourism demand. *Annals of Tourism Research*, 38(1):72 – 89.
- [29] Slini, T., Karatzas, K., and Moussiopoulos, N. (2002). Statistical analysis of environmental data as the basis of forecasting: an air quality application. *Science of The Total Environment*, 288(3):227 – 237.
- [30] Suhartono, S. (2011). Time series forecasting by using seasonal autoregressive integrated moving average: Subset, multiplicative or additive model. 7:20–27.
- [31] Taskaya-Temizel, T. and Casey, M. C. (2005). A comparative study of autoregressive neural network hybrids. *Neural Networks*, 18(5):781 – 789. IJCNN 2005.
- [32] Tseng, F.-M., Yu, H.-C., and Tzeng, G.-H. (2002). Combining neural network model with seasonal time series arima model. *Technological Forecasting and Social Change*, 69(1):71 – 87.
- [33] van Aalst, R. and de Leeuw, F. (1997). National ozone forecasting systems and international data exchange in northwest europe. Technical report, European Environment Agency.
- [34] Wang, X.-K. and Wei-Zhen, L. (2006). Seasonal variation of air pollution index: Hong kong case study. 63:1261–72.
- [35] Zhang, G. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159 – 175.
- [36] Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting*, 14(1):35 – 62.
- [37] Zhang, G. and Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, 160(2):501 – 514. Decision Support Systems in the Internet Age.