EnvironmentAsia
The international journal by the Thai Society of Higher Education Institutes on Environment

EnvironmentAsia 11(3) (2018) 251-262
DOI 10.14456/ea.2018.52
ISSN 1906-1714; ONLINE ISSN: 2586-8861

Application of Exponential Smoothing Holt Winter and ARIMA Models for Predicting Air Pollutant Concentrations

Arie Dipareza Syafei*, Nurul Ramadhan, Joni Hermana, Agus Slamet, Rachmat Boedisantoso and Abdu Fadli Assomadi

Department of Environmental Engineering, Faculty of Civil, Environmental, and Geo Engineering, Institut Teknologi Sepuluh Nopember (ITS), Indonesia

*Corresponding Author: dipareza@enviro.its.ac.id
Received: January 17, 2018; Accepted: June 11, 2018

Abstract

Two time series models, Holt Winter and Autoregressive Integrated Moving Average (ARIMA), were adapted to predict the concentrations of daily air pollutants in Surabaya, Indonesia. Two scenarios were developed to assess model performance in predicting PM_{10}, SO_2, CO, NO_2, and O_3 concentrations. In the first scenario, we used measured data, and, in the second scenario, we tested model performance when the data contained many missing values. We varied the percentage of missing values for three different sets of trained data and filled them with interpolations. It was found that the Holt Winter model was best at predicting CO, NO_2, and O_3 concentrations using measured data, whereas the ARIMA model was better at predicting PM_{10} and SO_2 concentrations. An assessment of model performance when there were missing values shows that the Holt Winter model was not affected by the number of missing values and missing data patterns in the prediction of CO and O_3 concentrations, although it was affected in the prediction of NO_2. On the other hand, the ARIMA model, which was used for the prediction of PM_{10} and SO_2 concentrations, was not affected by the amount of missing data and missing data patterns. The Holt Winter model is recommended for the prediction of CO concentrations based on the following model goodness of fit criteria for three different experimental runs with various amounts of missing data: the mean error, ME, (0.039; -0.878; -1106); root mean square error, RMSE, (0.315; 0.985; 1.175); coefficient of determination, R^2, (0.516; 0.612; 0.785); and correlation (0.719; 0.782; 0.886).

Keywords: ARIMA; Holt winter; Prediction; Air pollution

1. Introduction

Air pollution, especially in the world's big cities, including the Indonesian cities of Jakarta, Surabaya, Semarang, Bandung, and Medan, has caused a decrease in air quality and an increase in health disorders (Arifin and Sukoco, 2009). Based on the Minister of Environment's Regulation No. 12 of 2010 concerning the implementation of air pollution control in districts, the quality of ambient air has decreased due to the increase in air pollutant sources, leading to aggressive preventive and control actions. One step in air pollution management is the monitoring of ambient air quality status. A daily report on air quality can then be released so that citizens are aware of the level of and potential public health risks from air pollution (Afroz *et al.,* 2003).

Currently, there are three active monitoring stations in Surabaya, Indonesia recording particulate (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and ozone (O_3) levels (Sari, 2014). Accumulation of these pollutants, coming from a vast number of sources, causes negative impacts on public health. In order to avoid and reduce such impacts, citizens need to be aware of future air quality in advance. The accuracy of predictions helps end-users and stakeholders to develop appropriate plans to avoid being exposed to pollutants. Better predictions can increase profits, quality of life, and prevent deaths (Dahyot, 2014).

The predictions of air quality concentration can be done by using various statistical techniques, without consideration of physical and chemical processes, to analyze the existing historical data record (Kandya and Mohan, 2009), especially when it comes to short-term predictions. Time series analysis is a good model with which to monitor and forecast air quality conditions (Wei, 2006; Lee *et al.,* 2012). The data required in time series forecasting can be obtained from a wide variety of monitoring stations which record air quality in time sequence (Ip *et al.,* 2010).

The Autoregressive Integrated Moving Average (ARIMA) model is one of the statistical techniques used to predict urban air quality (Kandya and Mohan, 2009). This model has the capability of describing stationary data using the differencing method (Wei, 2006). ARIMA requires matched testing of each statistical modeling technique to obtain prediction values (Omane *et al.,* 2013) which consider past and erroneous values in the data (Makridakis *et al.,* 1999).

To enhance performance, seasonal and trend data patterns are incorporated into the model. The Exponential Smoothing Holt Winter model integrates these two aspects (Kalekar, 2004). The Holt Winter model has been applied to the prediction of electrical power requirements, which has daily data showing seasonal patterns and trends (Sudheer and Suseelatha, 2015). This model does not consider the stationarity of data, but, instead, uses repetitive steps and past weighting values to obtain new predictive values (Omane *et al.,* 2013). Daily data for air pollutant concentrations show seasonal patterns and trends (Kandlikar, 2007). The similarity in the seasonal patterns and trends of electricity requirements and air pollutant concentrations data indicates that the Holt Winter model can be used to predict the concentrations of air pollutants.

Both the Holt Winter and ARIMA models require past values to obtain future values, but cannot proceed if there is missing data (Hanzak, 2008; Terry *et al.,* 1986). Therefore, missing value imputation must be done to fill in these gaps through, e.g., interpolation. However, this process will affect model performance. The present study thus determines the effect of the amount of and patterns in missing data on model prediction accuracy.

2. Material and Method

2.1 Data Collection

Data used in this study are air quality data from the Environmental Agency in Surabaya taken from air quality monitoring stations located in Kebonsari, Surabaya (SUF-6). Air quality parameters are particulate (PM_{10}), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and ozone (O_3) concentrations. Observations on air quality were taken hourly each day for 15 days in 2014. Data were collected for 14 days in-sample (336 observations on each air quality parameter) and one day out-of-sample (24 observations) to determine the goodness of fit of the models. Total data used were 360 observations on each air quality parameter.

The analyses in this research were done by iterating the Holt Winter and ARIMA models until the best model parameters were found using the open-source platform R. Prediction of air pollutant concentrations was done one day ahead, and the prediction results were checked using observed data. The data on each parameter of air pollutant concentrations represents the dry season, January - February.

2.2 Scenario of Experiments

There were two experiments conducted, each with three test data sets. Both experiments were intended to test the consistency of the model under study in air pollutant concentrations prediction. The data used are shown in Tables 1 and 2.

Table 1. Test Data for Experiment I

Parameter	*In - Out Sample*		
	I	**II**	**III**
PM_{10}	4-18/2/2014	11-24/2/2014	24/1-7/2/2014
CO	4-18/1/2014	9-23/1/2014	7-21/2/2014
SO_2	2-16/1/2014	8-22/1/2014	6-20/2/2014
NO_2	4-18/1/2014	24/1-7/2/2014	29/1-12/2/2014
O_3	1-15/1/2014	7-21/1/2014	5-19/2/2014

Note: date format is dd/mm/yy; thus, 18/2/2014 refers to 18th February 2014. This format is used for the dates in all of the tables

Table 2. Test Data for Experiment II

% Missing	Data in-out sample				
Data	PM_{10}	CO	SO_2	NO_2	O_3
12.5					
37.5	24/1-7/2/2014	4-18/1/2014	2-16/1/2014	4-18/1/2014	4-18/1/2014
50					

Table 2 shows data collection dates for Experiment II, in which the data set has been manipulated by varying the amount of missing data. We randomly deleted 12.5%, 37.5%, and 50% of the data using with both regular and random patterns. Regular missing values means that we deleted several data in sequence, whereas random missing values means that we deleted data completely randomly, not in sequence. The manipulated data is then used to check consistency of the model when the missing data are filled in by interpolation.

2.3 Model Performance Measurements

Model verification is used to determine the best model. The accuracy level refers to the goodness of fit of a prediction model. The verification method used is as follows:

1. Mean Error (ME). If there are observation values and predictions for n time periods, then there will be n false and standard statistical measures. The ME is defined in equation (1).

$$ME = \sum_{t=1}^{n} e_1/n \quad (1)$$

2. Root Mean Square Error (RMSE) is statistical parameter that shows the actual error size of the model. RMSE is used to compare prediction performance with the prediction model (Omane *et al.,* 2013). The formula for RMSE is shown in equation (2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_1} \quad (2)$$

3. Coefficient of Determination. R^2 represents the proportion of the variance (fluctuations) in the predicted variable that can be explained by the independent variables. It is used as a gauge of how well a model explains and predicts future outcomes (Berk, 2008). R^2 is thus used to check the goodness of fit of the model. R2 can be computed as seen in equation (3).

$$R^2 = \frac{SSE}{SST} = \frac{\Sigma(y_1 - \hat{y}_1)^2}{\Sigma(y_1 - \bar{y}_1)^2} \quad (3)$$

4. Correlation Coefficient. Correlation is a statistical method used to determine the strength or degree of the linear relationship between the predicted and independent variables (Bamston, 1992). Correlation is denoted by r, $-1 \leq r \leq 1$. If r = -1, this means a perfect negative (inverse) relationship; r = 0 means no relationship; and r = 1 means a perfect positive (direct) relationship (Muinah, 2011). The equation for r can be seen in equation (4).

$$r = \sqrt{R^2} \quad (4)$$

3. Results and Discussion

3.1 PM_{10} Modeling

The first step in PM_{10} concentration modeling is done by separating the data into two parts; namely, *in sample* (336 observations) and *out* of *sample* (24 observations). The data pattern in the PM_{10} air pollutant concentration will provide a description of the data and a visual stationarity and can be determined by a time series plot, as shown in Figure 1.

Based on the verification results in Table 3, the ARIMA model has better capability and consistency in PM_{10} air pollutant concentration prediction. Overall, statistical measures show consistent support for the ARIMA model. The forecasted values obtained from the ARIMA model align nicely with the observed data, as shown in Figure 2.

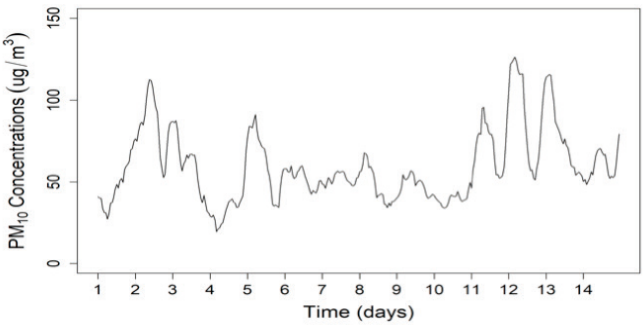

Figure 1. Time Series Plot of PM_{10} – time plot III (24/1-7/2/2014)

Table 3. Model Performance for PM_{10} Data

Model	Verification Results			
	ME	RMSE	R^2	Corr
HW I	34.858	56.335	0.527	-0.726
ARIMA I	-9.638	17.515	0.305	0.552
HW II	18.037	23.117	0.017	-0.132
ARIMA II	5.607	13.376	0.017	0.131
HW III	-125.72	143.893	0.677	-0.823
ARIMA III	2.443	5.123	0.823	0.907

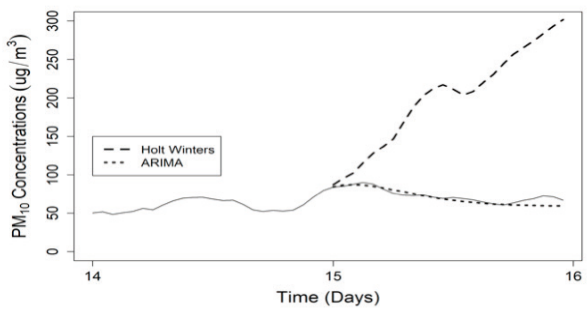

Figure 2. Graph of PM_{10} Prediction - time plot III (24/1-7/2/2014)

3.2 CO Modeling

A representation of the data pattern of CO concentration is shown in Figure 3 (7-21/2/2014). It exhibits a bimodal shape, representing two peak sessions every day.

Table 4 shows the results of the statistical measures for both models when predicting the concentration of the pollutant gas CO. The Holt Winter model is superior to the ARIMA model in this case. Figure 4 shows that the Holt Winter prediction pattern is similar to the observed data, and the correlation coefficients are fairly strong. However, the R^2 ranges from 0.5 to 0.8, which means that the performance relies on data training and is, therefore, likely to fluctuate in the near future.

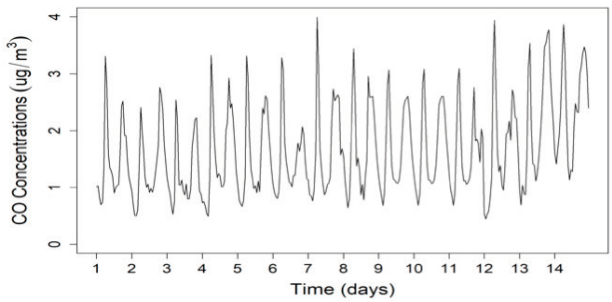

Figure 3. Time Series Plot of CO – time plot III (7-21/2/2014)

Table 4. Model Performance for CO Data

Model	Verification Results			
	ME	RMSE	R^2	Corr
HW I	0.04	0.315	0.513	0.716
ARIMA I	-0.432	0.415	0.357	0.517
HW II	-0.878	0.985	0.612	0.782
ARIMA II	-0.014	0.595	0.305	0.552
HW III	-1.106	1.175	0.785	0.886
ARIMA III	0.288	0.885	0.046	0.215

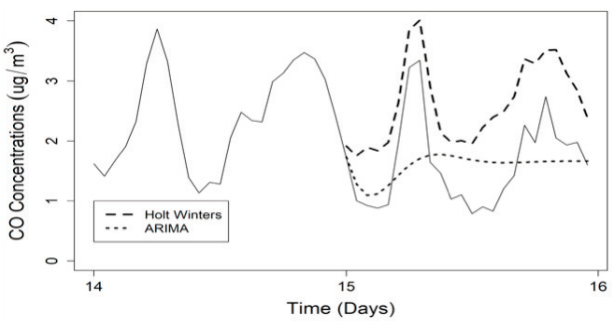

Figure 4. Graph of CO Prediction – time plot III (7-21/2/2014)

3.3 SO_2 Modeling

A representation of the data pattern for SO_2 concentration can be shown by the time series plot in Figure 5. The pattern shows several spikes, or high concentrations, due to manufacturing activities in several time slots.

Based on the verification results in Table 5, the ARIMA model has better capability and consistency in SO_2 pollutant concentration prediction. However, we note that, for this gas, the performance is not as good as those of the models used to predict CO and PM_{10}. CO concentration is very low in ambient air, and its fluctuations may occur over a very long period of time. This may cause the concentration data to not follow any clear pattern. Figure 6 shows that both models fail to produce accurate predicted concentrations.

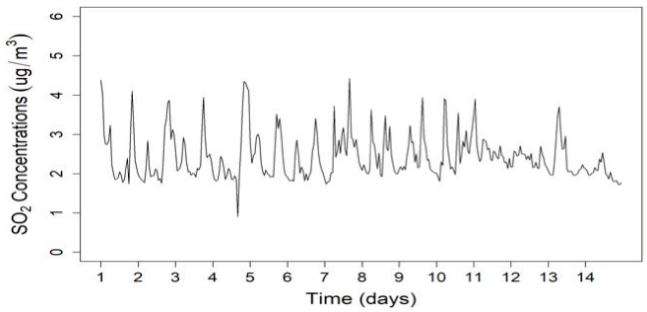

Figure 5. Time Series Plot of SO_2 – time plot II (8-22/1/2014)

Table 5. Model Performance for SO_2 Data

Model	Verification Results			
	ME	RMSE	R^2	Corr
HW I	1.99	2.431	0.023	-0.152
ARIMA I	-0.362	1.198	0.035	0.187
HW II	1.132	1.193	0.050	0.223
ARIMA II	-0.131	0.256	0.261	0.511
HW III	0.412	1.117	0.002	-0.043
ARIMA III	0.376	1.091	0.029	0.171

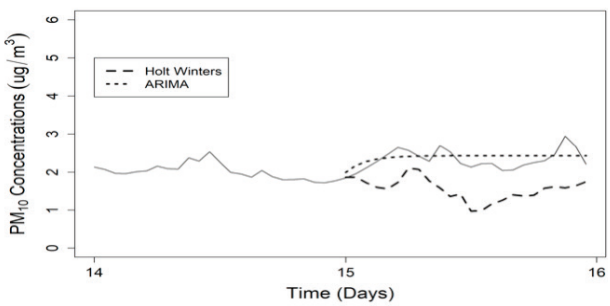

Figure 6. Graph of SO_2 Prediction - time plot II (8-22/1/2014)

3.4 NO_2 Modeling

NO_2 concentration is shown in Figure 7. NO_2 is emitted largely through a combustion process, mainly from vehicles, with a smaller portion coming from industry. The pattern in NO_2 concentrations shows a bimodal shape, which occurs in the morning and afternoon each day.

The performance of the Holt Winter model in forecasting the next day's pollutant levels is superior to that of the ARIMA model, as indicated by all statistical measures. The correlations are relatively strong, and the forecasted values match up well with the observed data (Figure 8).

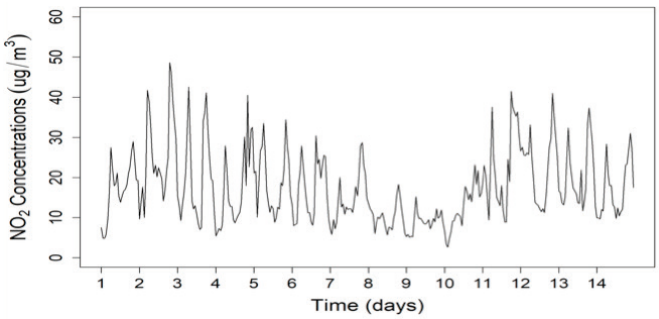

Figure 7. Time Series Plot of NO_2 – time plot II (24/1-7/2/2014)

Table 6. Model Performance for NO_2 Data

Model	Verification Results			
	ME	RMSE	R_2	Corr
HW	-9.954	12.721	0.427	0.654
ARIMA	-15.766	16.126	0.072	0.269
HW	-9.805	11.270	0.516	0.718
ARIMA	-0.172	6.041	0.133	0.365
HW	4.693	8.192	0.644	0.802
ARIMA	0.281	10.332	0.123	0.350

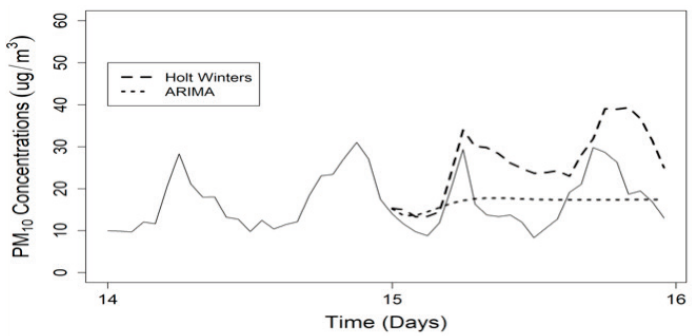

Figure 8. Graph of NO_2 Prediction – time plot II (24/1-7/2/2014)

3.5 O_3 Modeling

Ozone (O_3) is formed in the tropospheric layer by photochemical reactions involving the presence of NO, NO_2, and UV light. Therefore, O_3 is a secondary pollutant. High concentrations of O_3 may lead to skin irritation. The ozone concentration pattern follows a one-peak-a-day pattern, peaking each day at noon (high point in UV radiation), as seen in Figure 9. However, on days when the cloud cover is high, lower O_3 peak concentrations are found.

The Holt Winters model once again outperforms the ARIMA model when used to forecast O_3 concentrations for all three sets of data. The results follow the pattern of the observed data (Figure 6). On the other hand, the smoother results from the ARIMA model fail to accurately reflect the observed data. Statistical measures from all three sets (Table 7) consistently show the superiority of the Holt Winters model over the ARIMA model. The flexibility of Holt Winters is apparent when predicting pollutants with a clear seasonal pattern.

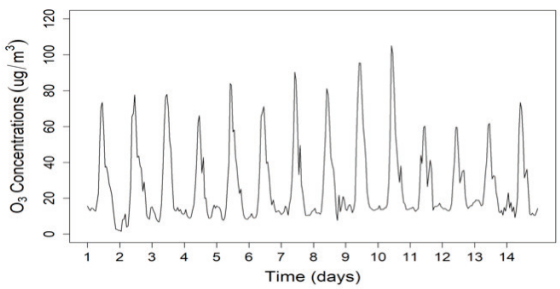

Figure 9. Time Series Plot of O_3 – time plot I (1-15/1/2014)

Table 7. Model Performance for O_3 Data

Model	Verification Results			
	ME	RMSE	R^2	Corr
HW I	-6.403	15.747	0.616	0.785
ARIMA I	-0.156	10.482	0.007	0.084
HW II	-4.486	10.911	0.465	0.682
ARIMA II	6.04	8.776	0.282	0.531
HW III	-20.237	24.488	0.165	0.406
ARIMA III	-15.575	18.683	0.019	0.139

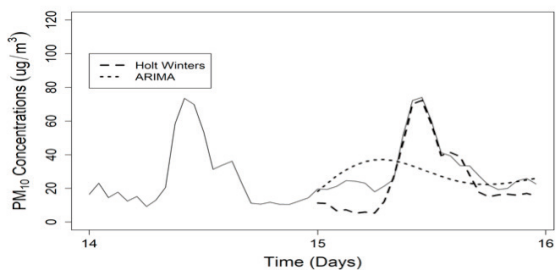

Figure 10. Graph of O_3 Prediction – time plot I (1-15/1/2014)

3.6 *Model Consistency When Considering the Amount of Missing Data*

Missing data causes prediction to become difficult in time series models because continuity should be maintained in order to obtain coefficients for the models. A wide range of statistical methods, especially interpolation, are available to fill in the missing data (Honaker et al., 2010). Missing data are filled with new values, and, once the trained data is complete, prediction simulations using the Holt Winter and ARIMA models can be performed.

The generation of missing data, at 12.5%, 37.5%, and 50% of data, shows that Holt Winter is able to consistently produce better predictions when missing values are filled in with interpolation, as shown in Table 8. Whether the missing pattern is random or regular does not affect significantly model performance (Table 9), except in the case of NO_2, for which the Holt Winters model performance was better when the missing values were in sequence. The Holt Winter model failed to produce an accurate forecast when the missing values occurred randomly.

The results also show that there is consistency between the scenarios developed in this study. The performance of both models with missing values which were interpolated was the same as their performance with no missing values. This result indicates that missing values will not cause a deterioration in model performance. However, there is a slight effect on prediction from how missing values are generated.

The ARIMA model is not affected by the distribution pattern in the air of a pollutant's concentration when it is used to predict PM_{10} and SO_2 concentrations. The Holt Winter model is not affected by missing value patterns when predicting CO and O_3 concentrations; however, its performance was slightly affected when predicting NO_2 concentrations.

Table 8. The Best Model Based on the Percentage of Missing Data

Missing Data	PM_{10}	CO	SO_2	NO_2	O_3
12.5%					
37.5%	ARIMA	Holt Winter	ARIMA	Holt Winter	Holt Winter
50%					

Table 9. Distribution Pattern of the Best Data

Missing Data	PM_{10}	CO	SO_2	NO_2	O_3
12.5%	Regular	Random	Random	Regular	Regular
37.5%	Random	Random	Regular	Regular	Regular
50%	Random	Regular	Regular	Regular	Random

4. Conclusion

The use of a time series model is important for air quality management, especially when used to forecast short-term concentrations of pollutants. However, selecting an appropriate model is essential. In this study, we assess successfully the performance of ARIMA and Holt Winter models in predicting pollutants. It is observed that, for pollutants with seasonal variations in concentration, such as CO, NO_2 and O_3, the Holt Winter model outperforms the ARIMA model. ARIMA was better at predicting PM_{10} and SO_2 concentrations, both of which had little seasonal variation. This result is not affected by the number of missing values filled in by interpolated data contained in the data training. The present study shows the promising use of the Holt Winter model when forecasting air pollutant concentrations compared with ARIMA.

Acknowledgement

The authors wish to thank the Department of Environmental Engineering, Faculty of Civil, Environmental, and Geo Engineering, Institut Teknologi Sepuluh Nopember (ITS) for providing facilities during this research. We would also like to expand our thank to the Directorate of Research and Community Services, Directorate General of Research and Development, Ministry of Research, Technology and Higher Education of the Republic of Indonesia for providing access and funding for the publication of this manuscript under the umbrella of air quality monitoring station research project.

References

Afroz R, Hasan MN, Ibrahim NA. Review of Air Pollution And Health Impacts In Malaysia. Environmental Research 2003; 92(2): 71–77.

Arifin Z, Sukoco. Pengendalian Polusi Kendaraan. Alfabeta, Bandung. 2009.

Barnston AG. Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. Climate Analysis Center 1992; 7: 669-709.

Berk R. Statistical Learning from a Regression Perspective. New York: Springer; 2008.

Dahyot R. Time Series and Applied Forecasting. School of Computer Science and Statistics, Trinity Collage Dublin, Ireland. 2014.

Hanzak T. Improved Holt Method for Irregular Time Series. Proceeding of Mathematics and Computer Science 2008; 8: 62-67.

Honaker J, King G. What to Do about Missing Values in Time-Series Cross-Section Data. American Journal of Political Science 2010; 54(2): 561-581.

Ip V, Yang Y, Wong. Forecasting Daily Ambient Air Pollution Based On Least Squares Support Vector Machines. International Conference on Information and Automation 2010; 571-575.

Kalekar, Prajaka S. Time series Forecasting using Holt-Winter Exponential Smoothing [Thesis]. Mumbai: Kanwal Rekhi School of Information Technology. 2004.

Kandlikar M. Air Pollution at a Hotspot Location in Delhi: Detecting Trends, Seasonal Cycles and Oscillations. Atmospheric Environment 2007; 41(28): 5934–5947.

Kandya A, Mohan M. Forecasting the Urban Air Quality Using Various Statistical Techniques. The Seventh International Conference on Urban Climate, Yokohama, Japan. 2009.

Lee M, Abd R, Suhartono, Latif MT, Nor ME, Kamisan NA, Bazilah. Seasonal ARIMA for Forecasting Air Pollution Index: A Case Study. American Journal of Applied Sciences 2012; 9(4): 570-578.

Omane A, Oduro T, Oduro D. Determining the Better Approach for Short-Term Forecasting of Ghana's Inflation: Seasonal-ARIMA vs. Holt-Winter. International Journal of Business, Humanities and Technology 2013; 3(1): 69-79.

Makridakis S, Wheelwright S, McGee, Victor E. Metode dan Aplikasi Peramalan. 1st ed. Jakata: Errangga; 1999.

Muinah. Analisis Pengaruh Tingkat Pendapatan Dan Tingkat Pendidikan Masyarakat Terhadap Permintaan Produk Asuransi Jiwa Bersama Bumiputera 1912 Kantor Wilayah Medan [Thesis]. Medan: Universitas Sumatera Utara. 2011.

Sari N. Penentuan Korelasi Perubahan Tekanan Udara dan Curah Hujan Terhadap Lapisan Inversi dan Hubungannya Dengan Kualitas Udara Ambien Kota Surabaya [Thesis]. Surabaya: Institut Teknologi Sepuluh Nopember. 2014.

Sudheer, Suseelatha. Short term load forecasting using wavelet transform combined with Holt–Winters and weighted nearest neighbor models. International Journal of Electrical Power and Energy Systems 2015; 64: 340–346.

Terry, Leet, Kumar. Time Series Analysis in Acid Rain Modeling: Evaluation Of Filling Missing Values By Linear Interpolation. Atmospheric Environtment 1986; 20(10): 1941-1945.

Wei W. Time Series Analysis: Univariate and Multivariate Methods Second Edition. USA: Pearson Education; 2006.