International Journal of Advancements in Electronics and Electrical Engineering – IJAEEE
Copyright © Institute of Research Engineers and Doctors
Volume 8 : Issue 1 - [ISSN : 2319-7498] - Publication Date: 27 Dec, 2019

Time Series Model for Predicting Ground-Level Ozone

Kunlawee Manwong and Pasapitch Chujai

Abstract— **This research is a comparison for forecasting methods for ground-level ozone using ARIMA (Auto-Regressive Integrated Moving Average) and GARCH (Generalized Autoregressive Conditionally Heteroskedastic) for the forecasting of four places in Thailand, from January 2013 to August 2018. The results obtained compare between the two models above mentioned in order to find the most accurate one, considering the lowest RMSE (Root Mean Square Error) and the lowest MAPE (Mean Absolute Percentage Error). According to the experiment, the most suitable method is GARCH which is good for the 1-2 hours early forecasting.**

Keywords—**Ground-Level Ozone, Time Series, ARIMA model, GARCH model**

I. Introduction

Air pollution is one of the problems that occur in many urban areas, especially the ones with heavy traffic or industrial production processes, causing air pollution which, in turn, affects citizens' health [1, 2]. There are many substances who affect air quality, one of them is ground-level ozone [3,4]. Ground-level ozone is corrosive, as such, it can cause eyes irritation, affect the respiratory system, causing nasal irritation, it can reduce lungs capacity and may cause respiratory problems [5]. Furthermore, people with respiratory problems may experience symptoms more severe than they would usually do.

The information technology has been integrated into weather condition data collection so that the data collection is up-to-date and continuous in serial format. If the statistical data or data of the occurred events are used to forecast the next data or events using mathematical principles, it will help to increase the accuracy of forecasting which is important in preparing and planning for future situations. Time series are a set of data that changes continuously in order of time and have equal time spacing. Time series analysis requires time series characteristics. In general, time series consists of four parts: trend, seasonal variations, cycle, and irregular movement [6].

Due to the above mentioned, the researcher has developed Thailand's ground-level ozone predictions using time series model as guidelines for creating ground-level ozone coping measures in the future by using *R* language to create forecasting models.

II. Background and Related Works

A. Ground-Level Ozone

Generally, ozone can be classified into two types. The first type is natural ozone [7]; it can be found in the atmosphere which is more than 40 kilometers above the ground. It protects the earth from ultraviolet radiation. This type of ozone is a non-toxic gas. The second type is ozone that can be found at an altitude of less than two kilometers above ground. This type of ozone is a toxic ozone found in the air we breathe which has a direct impact on humans.

Ground-level ozone [3] is considered dangerous ozone. Not only might it occur naturally with the light as its catalyst for reaction, it can also be caused by other factors such as excessive sunlight, smoke emission from engine combustion, such as passenger transportation vehicles, industrial production processes and man-made devices that produce ozone. Therefore, the surrounding ozone gas is not ozone found in the atmosphere. It is a toxic ozone that affects the human body.

B. Time Series

Time series [8, 9] is a set of quantitative data collected over consecutive time, such as the SET index at closing time each day, the income of a company in each year, weather forecast, and data of atmospheric gas, etc. Time series data means data sets or quantitative observations obtained from data collection by continuously arranging the time spent. Data sorting may be on a yearly base, quarterly base, monthly base, weekly base, daily base or hourly base, etc. The sorting characteristics depend on the purpose of using or analyzing such data or forecasting events.

C. The ARIMA Model

Auto-Regressive Integrated Moving Average or ARIMA [8] is a model for forecasting information or events. ARIMA consists of the combination of three techniques of time series including *AR* (Autoregressive), *I* (Integrated) and *MA* (Moving Average). Its purpose is to eliminate "Noise" from data in order to try to reduce error terms as much as possible and have the following stationary properties. The series is with constant

Kunlawee Manwong
Department of Electrical Technology Education, KMUTT,
THAILAND

Pasapitch Chujai
Department of Electrical Technology Education, KMUTT,
THAILAND

mean and variance in order not to change with time [10], so that the information is reliable which will increase the prediction efficiency in the next step.

D. *The GARCH Model*

Generalized Autoregressive Conditionally Heteroskedastic, or GARCH [11], is another model that has been developed from the ARCH (Autoregressive Conditionally Heteroskedastic) model in order to use time series analysis, especially focusing on the data variance. This model can be used in creating a variety of trading strategies, especially volatility estimation, including volatility clustering on time-based products.

E. *Akaike Information Criterion (AIC)*

Akaike Information Criterion or AIC [12] is the criterion determined by the estimation of tolerances with information of observations, using then the minimum value of data to adjust the forecasting values for more accuracy, AIC formula as (1).

$$AIC = -2(\text{log-likelihood}) + 2K \tag{1}$$

where *K* is the number of model parameters (the number of variables in the model plus the intercept).

Log-likelihood is a measure of model fit.

F. *Forecast Error*

Forecasting error methods [13, 14]. This research uses two methods of forecasting error, which are (2) and (3) as below:

1. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{actual})^2}{n}} \tag{2}$$

2. Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{y_{actual} - y_{pred}}{y_{actual}}\right|}{n} \times 100 \tag{3}$$

G. *Related Works*

J.W. Taylor [15] studied the comparison of the time series forecasting techniques to forecast incoming calls during daytime of a customer service center using two models, Box-Jenkins model and Holt’s exponential smoothing method. The results showed that Box-Jenkins’ time series forecasting provides better forecasting performance because of less root mean square error (RMSE).

L. CheeNian [16] studied ARIMA model and GARCH model for forecasting crude oil prices by using daily time series data of the WTI crude oil price between January 2, 1986, to September 30, 2009. He compared between ARIMA (1,2,1) and GARCH (1,1) models based on the RMSE. It was found that the GARCH (1,1) model is the most suitable model for this data set because it provided the least RMSE.

R. Selvaraj *et al*. [17] studied the model and forecasting total ozone and rainfall in Codicanal using ARIMA to assist in coping and managing agriculture by creating a prediction model of total ozone and rainfall with monthly data gathered during 12 years. It was found that ozone content can be predicted with very high accuracy because ozone has an RMSE of 8.182 and the RMSE of rainfall is 94.21.

M. Ohyver *et al*. [18] studied the model and forecasting for the price of medium quality rice detail, using the ARIMA model. They found that ARIMA (1,1,2) is a suitable model for this dataset, based on the RMSE price of medium quality rice which is 14.22.

M. Matyjaszek *et al*. [19] studied forecasting coking coal using ARIMA and Neural Networks models, which include the ROBUST model, the Generalized Regression Neural Networks (GRNNs) model, and the MLFN model. This consideration is based on the MAPE since it is considered a percentage deviation. The results showed that the most accurate prediction model was the ARIMA model, followed by the GRNNs and the least accurate was the MLFN model.

From the study research of various related models, it can be concluded that: ARIMA model is a popular model and effective in forecasting. In addition, GARCH is an interesting model. Therefore, the researcher is interested in comparing the accuracy of ground-level ozone forecasting with both ARIMA and GARCH models using data from four agencies, Map Ta Phut Health Promoting Hospital in Rayong, Hat Yai Municipality of Songkhla, four Regional Hydrology under Water Resource Office of Khon Kaen, and Learning Space of Phayao Provincial Administrative Organization for the preparation for ground-level ozone gas intensity in the future.

III. Method

The objective of this research is to create and find the efficiency of the model for ground-level ozone forecasting in Thailand using time series model for a preparation to deal with the ground-level ozone gas intensity that may have serious consequences on human lives. The framework of the research is shown in Fig. 1 as follows:

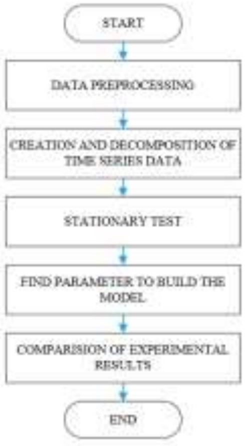

START
DATA PREPROCESSING
CREATION AND DECOMPOSITION OF TIME SERIES DATA
STATIONARY TEST
FIND PARAMETER TO BUILD THE MODEL
COMPARISION OF EXPERIMENTAL RESULTS
END

Figure 1. Conceptual Framework.

IV. Experimental Evaluation

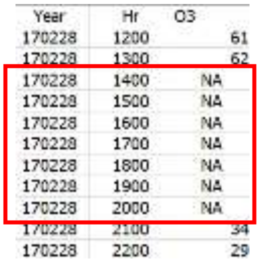
A. *Data Preprocessing*

The data in this research came from Bangkok's Air Quality and Noise Management Division and Bangkok's Pollution Control Department. The type of data is hourly data, selecting specifically the parameter O3 [20]. The data came from four sources, details shown in Table 1.

The raw materials could not be used for the forecasting model due to some data loss as seen in Fig. 2. It decreased the accuracy of the forecasting model. In consequence, it required to manage the raw data by changing to time series data. This research selected Linear Interpolation [21, 22] as a means to add data as seen in Fig. 3.

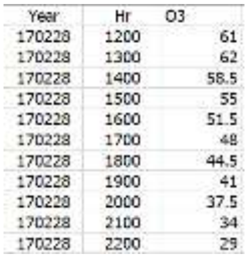
TABLE I. DETAILS OF DATA SET USED IN THIS RESEARCH

Dataset	Date	Number of Record	Missing Data	Missing Data (%)
Map Ta Phut Health Promoting Hospital in Rayong	1 January 2013 - 18 August 2018	49,656	7,366	14.83
Hat Yai Municipality of Songkhla	1 January 2013 – 18 August 2018	49,656	9,853	19.84
Regional Hydrology under Water Resource Office of Khon Kaen	1 January 2014 – 18 August 2018	40,896	8,663	21.18
Learning Space of Phayao Provincial Administrative Organization	1 January 2013 – 18 August 2018	49,656	4,179	8.41



Year	Hr	O3
170228	1200	61
170228	1300	62
170228	1400	NA
170228	1500	NA
170228	1600	NA
170228	1700	NA
170228	1800	NA
170228	1900	NA
170228	2000	NA
170228	2100	34
170228	2200	29

Figure 2. Samples of lost data of Learning Space of Phayao Provincial Administrative Organization.



Year	Hr	O3
170228	1200	61
170228	1300	62
170228	1400	58.5
170228	1500	55
170228	1600	51.5
170228	1700	48
170228	1800	44.5
170228	1900	41
170228	2000	37.5
170228	2100	34
170228	2200	29

Figure 3. Adding the lost data by Linear Interpolation.

B. *Creation and Decomposition of Time Series Data*

1) Creating Time Series Data

The data set is received in the form of a vector which must be turned into time series data with *ts()* function as the library of *R* program. This research created hourly time series data and the frequency of the time series is set as *8,766.*

2) Decomposition of Time Series Data

In this stage, we decomposed the time series data to see trend, seasonal variations, cycle, and irregular movement.

a) Map Ta Phut Health Promoting Hospital in Rayong.

The assessment of time series data decomposition of Map Ta Phut Health Promoting Hospital in Rayong seen in Fig. 4. The trend of the compositions in 2013 slightly increased but in 2014, it started to decrease slowly. After that, it contains the trait of multiplicative seasonal. As for the seasons, they were slightly changed. The pattern of the seasons seemed to begin and end like the time series data.

b) *Hat Yai Municipality of Songkhla.*

The assessment of time series data decomposition of Hat Yai Municipality of Songkhla shown in Fig. 5. The trend of the compositions in 2013 slowly increased but in 2015 it started to decrease. After that, it contains the trait of multiplicative seasonal. As for the seasons, they were slightly changed. With time passing, the pattern of the seasons seemed to begin and end like the time series data.

c) Four Regional Hydrology under Water Resource Office of Khon Kaen.

The assessment of time series data decomposition of four Regional Hydrology under the Water Resource Office of Khon Kaen shown in Fig. 6. The trend of the compositions started in 2014 and slowly increased until 2017, it started to decrease rapidly until the end of time series. After that, it contains the trait of multiplicative seasonal. As for the seasons, they were slightly changed. The seasons seemed to be unchanged with time passing. The pattern of the seasons remained constant from the beginning to the end of the time series data.

d) Learning Space of Phayao Provincial Administrative Organization

The assessment of time series data decomposition of Learning Space of Phayao Provincial Administrative Organization shown in Fig. 7. The trend of the compositions has the characteristics of multiplicative seasonal, it started from 2013 until the end of time series. For the change of season, it does not change a lot. The pattern of the season remains constant from the beginning to the end of the time series data.

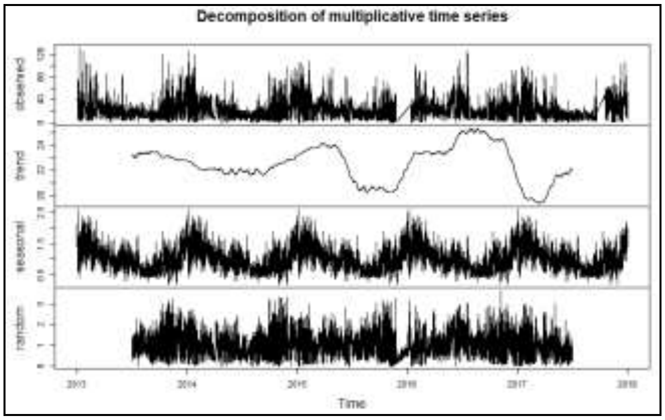


Figure 4. Decomposition of Grown Level Ozone Forecasting using Time Series Data of Map Ta Phut Health Promoting Hospital in Rayong.

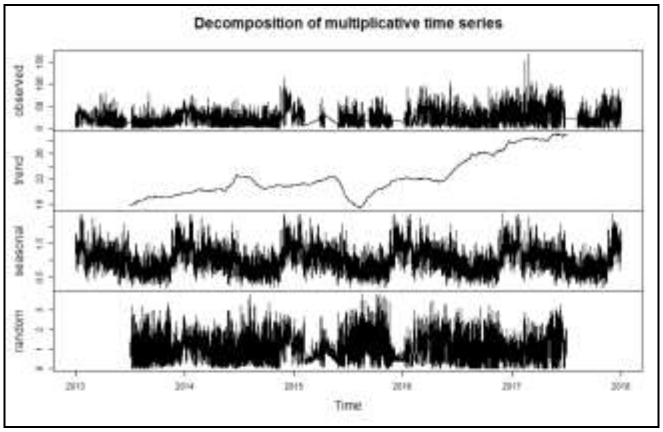


Figure 5. Decomposition of Grown Level Ozone Forecasting using Time Series Data of Hat Yai Municipality of Songkhla.

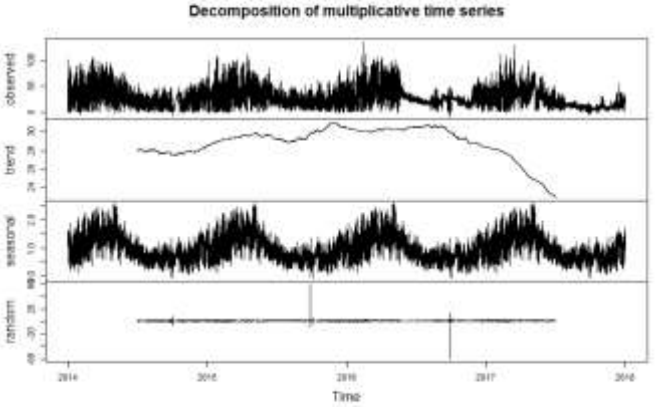


Figure 6. Decomposition of Grown Level Ozone Forecasting using Time Series Data of 4 Regional Hydrology under Water Resource Office of Khon Kaen.

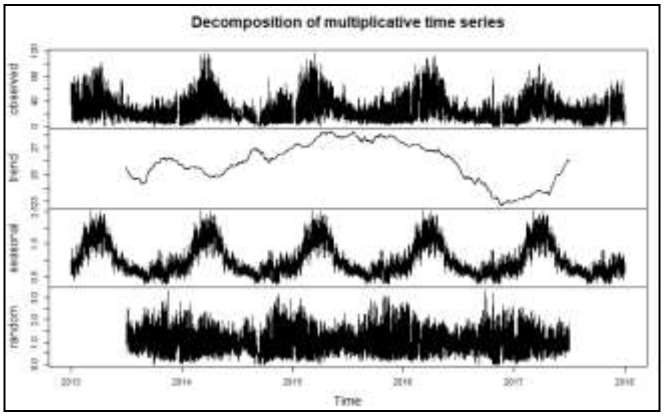


Figure 7. Decomposition of Grown Level Ozone Forecasting using Time Series Data of Learning Space of Phayao Provincial Administrative Organization.

C. *Stationary Test*

If the time series data has the mean or non-stationary, it needs to be processed to make it stationary before considering the model. The purpose of changing from a non-stationary time series to a stationary time series is to find the difference. The consideration of a stationary time series could be regarded from the Autocorrelation Function (ACF). If the time series is not stationary, the value will decrease slowly as seen in Fig. 8 which demonstrates the non-stationary time series; consequently, it requires to find the first difference value of the time series.

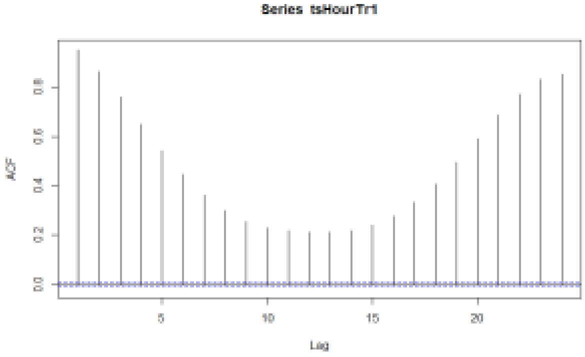


Figure 8. ACF graph sample of Learning Space of Phayao Provincial Administrative Organization

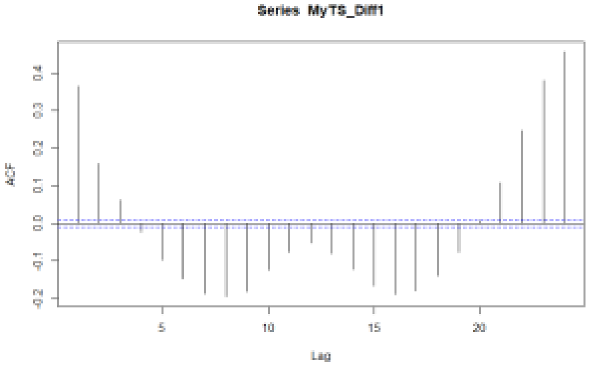


Figure 9. ACF graph sample of Learning Space of Phayao Provincial Administrative Organization demonstrating the first difference.

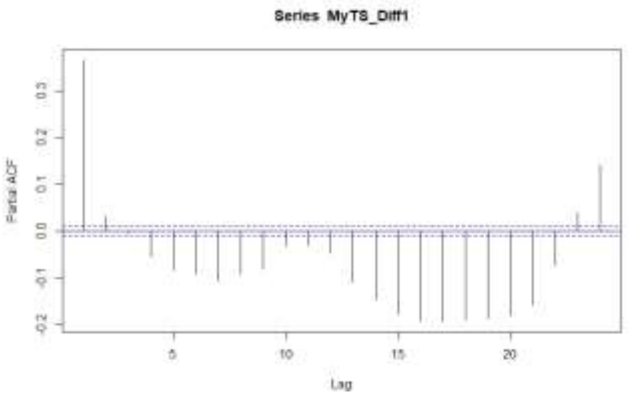


Figure 10. ACF graph sample of Learning Space of Phayao Provincial Administrative Organization

In Fig. 9, it is found that ACF reduces rapidly, meaning that the series is stationary.

D. *Finding Parameters to build the model*

Finding the *p*, *d* and *q* parameters in order to build the model can be determined by the ACF and Partial Autocorrelation Function (PACF). The ACF graph gives the value for the *q* parameter, the PACF graph gives the value for the *p* parameter, and the *d* parameter is determined by the difference.

Fig. 9, the cross line at the third Lag shows that the parameter *q* is equal to 3, and from Fig. 10, the cross line at the second Lag shows that the parameter *p* is equal to 2.

The received values would be applied to build the model by selecting the parameter with the least AIC as (1). For this research, the author studied the forecasting process of data or events by using the time series to predict the ground-level ozone. The author used two time-series models which are ARIMA model and GARCH model before comparing the two results to see which one provides a more accurate result and recorded it. Then, the author selected the model with better efficiency in forecasting to predict the ground-level ozone. Details are in Table 2.

E. *Comparison of experimental results*

According to the comparison of the forecasting results from the two mathematic methods which are RMSE [13] as it has the formula as (2) and MAPE [13] as it has the formula as (3). See the result in Table 3.

From Table 3, according to the comparison of forecasting by time-series data of the two models by considering the least RMSE, it is found that GARCH model provides RMSE lower than ARIMA model for all four data sets.

The GARCH (2,3) result of the data from Map Ta Phut Health Promoting Hospital in Rayong gives the lowest RMSE at the 2-hours early forecasting while RMSE remains stable at 5.35 and MAPE remains stable at 28.78%.

TABLE II. DETAILS OF DATA SET USED IN THIS RESEARCH

Dataset	ARIMA		GARCH	
	Model	***AIC***	***Model***	***AIC***
Map Ta Phut Health Promoting Hospital in Rayong	ARIMA(2,1,3)	271982.7	GARCH(2,3)	5.77
Hat Yai Municipality of Songkhla	ARIMA(2,1,3)	282682.7	GARCH(2,3)	5.95
Four Regional Hydrology under Water Resource Office of Khon Kaen	ARIMA(3,1,2)	221074.3	GARCH(3,2)	5.77
Learning Space of Phayao Provincial Administrative Organization	ARIMA(2,1,3)	262420.8	GARCH(2,3)	5.77

The GARCH (2,3) result of the data from Hat Yai Municipality of Songkhla gives the lowest RMSE at the 1-hour early forecasting while RMSE remains stable at 2.90 and MAPE remains stable at 3.97%. Meanwhile, the GARCH (2,3) result of the data from Regional Hydrology under Water Resource Office of Khon Kaen gives the lowest RMSE at the 1-hour early forecasting while RMSE remains stable at 1.81 and MAPE remains stable at 7.31%.

Next, the GARCH (2,3) result of the data from Learning Space of Phayao Provincial Administrative Organization gives the lowest RMSE at the 1-hour early forecasting while RMSE remains stable at 2.90 and MAPE remains stable at 3.97%.

Last, the GARCH (2,3) result of the data from Regional Hydrology under Water Resource Office of Khon Kaen gives the lowest RMSE at the 1-hour early forecasting while RMSE remains stable at 1.65 and MAPE remains stable at 9.41%.

TABLE III. COMPARING THE MODEL OF FORECASTING BY RMSE AND MAPE.

Dataset	Model	Hour	RMSE	MAPE
Map Ta Phut Health Promoting Hospital in Rayong	ARIMA(2,1,3)	1	6.91	40.65
		2	7.42	39.59
	GARCH(2,3)	1	5.35	31.47
		2	5.35	28.78
Hat Yai Municipality of Songkhla	ARIMA(2,1,3)	1	6.26	8.58
		2	5.92	8.38
	GARCH(2,3)	1	2.90	3.97
		2	3.80	5.32
Four Regional Hydrology under Water Resource Office of Khon Kaen	ARIMA(3,1,2)	1	2.89	12.04
		2	2.53	10.82
	GARCH(3,2)	1	2.32	9.67
		2	1.81	7.31
Learning Space of Phayao Provincial Administrative Organization	ARIMA(2,1,3)	1	3.29	19.35
		2	4.03	22.59
	GARCH(2,3)	1	1.68	9.88
		2	1.65	9.41

V. Conclusion

The objective of this research was to create and find the efficiency of forecasting models for ground-level ozone in Thailand by the time series model, using three following data sets gathered between January 2013 and December 2017 from the relating units which are the air quality record of Map Ta Phut Health Promoting Hospital in Rayong, the air quality record of Hat Yai Municipality of Songkhla, the air quality record of Learning Space of Phayao Provincial Administrative Organization, and one data set for the air quality record of four Regional Hydrology under Water Resource Office of Khon Kaen from January 2014 to December 2017. Two forecasting models, the ARIMA model and the GARCH model, were tested from data gathered from January 2018 until August 2018 in order to compare the accuracy of the results.

Comparing RMSE and MAPE of ground-level ozone for each agency, the values obtained by using both ARIMA and GARCH Models, show that the results obtained with GARCH model have the lowest RMSE and MAPE values.

As conclusion, when forecasting the time series data who share a similarity, GARCH method should be selected for forecasting because the RMSE is less or it might be studied with other methods additionally in order to compare the patterns and to find the least RMSE.

Acknowledgment

This research was supported by the Electrical Technology Education Department, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, Thailand.

References

[1] S. Dougall, "Workplace health and safety (WHS) implications for farmers hosting unconventional gas (UG) exploration & production," Policy and Practice in Health and Safety, 2019, vol.17, no. 2, pp. 156-172.

[2] G. McCarron, "Air Pollution and human health hazards: a compilation of air toxins acknowledged by the gas industry in Queensland's Darling Downs," International Journal of Environmental Studies, 2018, vol. 75, no. 1, pp. 171-185

[3] W. S. Beckett, "Ozone, air pollution, and respiratory health," The Yale journal of biology and medicine, 1991, vol. 64, no. 2, pp.167-75.

[4] C. Walter, E. S. Futschik, and L. Irving, "Traffic pollution near childcare centres in Melbourne," Australian and New Zealand Journal of Public Health, 2019, vol. 43, no. 5, pp. 410-412.

[5] J. A. Bernstein, N. Alexis, C. Barnes, I. L. Bernstein, A. Nel, D. Peden, and P. B. Williams, "Health effects of air pollution," Journal of Allergy and Clinical Immunology, 2004, vol. 114, no. 5, pp. 1116 – 1123.

[6] H. C. Godwin, "Modified Trend and Seasonal Time Series Analysis for Operations: ACase Study of Soft Drink Production", International Journal of Engineering Research in Africa, 2012, vol. 7, pp 63-72.

[7] R. SamuelSelvaraj, "Modeling and Predicting Total Ozone Column and Rainfall in Kodaikanal, Tamilnadu By ARIMA Process", IJECS, vol. 2, no. 8, 2013, pp. 2512-2526

[8] G. Li, W. Yan and Z. Wu, "Discovering shapelets with key points in time series classification,Expert Systems with Applications", 2019, vol. 132, pp. 76-86.

[9] W. W. S. Wei. "Time Series Analysis : Univariate and Multivariate methods," New York USA: Addison – Wesley Publishing company, 1990.

[10] P. Ramos, N. Santos and R. Rebelo, "Performance of state space and ARIMA models for consumer retail sales forecasting", Robotics and Computer-Integrated Manufacturing, 2015, vol. 34, pp. 151-163.

[11] W. Enders, "Applied Econometric Time Series, 2nd Edition," In: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken. 2004.

[12] A. Zanini and A. Woodbury, "Contaminant source reconstruction by empirical Bayes and Akaike's Bayesian Information Criterion", Journal of Contaminant Hydrology, 2016, vol. 185–186, pp. 74-86.

[13] T. Chai, and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature," Geoscientific Model Development. 2014, vol. 7, no. 3, pp. 1247-1250.

[14] M. Bratu, "The Reduction of Uncertainty in Making Decisions by Evaluating the Macroeconomic Forecasts Performance in Romania," Journal of scientific and industrial research, 2012, vol. 25, no. 2, pp. 239-262.

[15] J. W. Taylor, "A comparison of univariate time series methods for forecasting intraday arrivals at a call center," Management Science, 2008, vol. 54, pp. 253-265.

[16] L. CheeNian, "Application of ARIMA and GARCH models in forecasting crude oil prices," Master thesis. University Technology Malaysia. Falculty of Science, 2009.

[17] R. SamuelSelvaraj, C.P. Sachithananthem and K. Thamizharasan "Modeling and Predicting Total Ozone Column and Rainfall in Kodaikanal, Tamilnadu By ARIMA Process," Department of Physics, Presidency College, Chennai, 2013.

[18] M. Ohyver and H. Pudjihastuti, "Arima Model for Forecasting the Price of Medium Quality Rice to Anticipate Price Fluctuations", Procedia Computer Science, 2018, vol. 135, pp. 707-711.

[19] M. Matyjaszek, P. Fernández, A. Krzemień, K. Wodarski and G. Valverde, "Forecasting coking coal prices by means of ARIMA models and neural networks, considering the transgenic time series theory", Resources Policy, 2019, vol. 61, pp. 283-292.

[20] Notification of Pollution Control Department, "Historical data", Retrieve January 17,2019, http://air4thai.pcd.go.th/webV2/history,January 1, 2013.

[21] N. M. Noor, Norazian, M. M. AI B. Abdullah, A. S. Yahaya, and N. A. Ramli, "Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set", Materials Science Forum, 2014, vol. 803, pp. 278-281.

[22] M. M. AI B. Abdullah, L. Jamaludin, A. Abdullah, R.A. Razak and K. Hussin, "Filling Missing Data Using Intepolation Methods: Study on the Effect of Fitting Distribution". Key Engineering Materials, 2014, vol. 594., pp. 889-895.

About Author (s):

Knlawee Manwong is a developer at JCG Coporation Co.,Ltd. He received his bachelor's degree in computer engineer at the Electrical Technology Education Department, FIET, KMUTT, Thailand in 2018. His current research include Time Series, Machine Learning, and RPA.

Dr. Pasapitch Chujai is a lecturer at the Electrical Technology Education Department, FIET, KMUTT, Thailand. She received her bachelor's degree in Computer Science from Ramkhamhaeng University, Thailand in 2000, master's degree in Computer and Information Technology from KMUTT, Thailand in 2004 and a doctoral degree in Computer Engineering from Suranaree University of Technology, Thailand in 2015. Her current research include Ontology, Recommendation System, Time Series, Machine Learning, and Imbalanced Data Classification.