MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Solar radiation forecast using hybrid SARIMA and ANN model: A case study at several locations in Peninsular Malaysia

Muhammad Zillullah Mukaram* and Fadhilah Yusof*

Department of Mathematical Sciences, University Teknologi Malaysia 81310 Skudai, Johor, Malaysia

* *Corresponding author: azil.utmskudai@gmail.com, fadhilahy@utm.my*

Article history
Received
Accepted

Abstract

Solar Energy have an enormous potential for generating renewable electricity. In the tropics solar energy are abundance all year long but suffer from uncertainty caused by rain and clouds. Accurate prediction of solar radiation can increase the affectivity and productivity of solar energy sources. Monthly average of solar radiation data are obtained from 3 stations in Malaysia. The data are modeled using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, artificial neural network (ANN) model and Hybrid ANN and SARIMA model. The SARIMA model is a reliable tool in forecasting seasonal data, on the other hand the ANN model have been proven to be a good model in forecasting non-linear data. By combining both model a more accurate model is obtained in this study. The model performance comparison is conducted by using mean absolute error (MAE), the mean absolute percentage error (MAPE) and root mean square error (RMSE). The result shows that the hybrid model is better in forecasting solar radiation data.

Keywords: SARIMA, ANN, Hybrid Mode, Solar Radiation, Forecasting.

INTRODUCTION

With the problem of climate change the world have begun shifting energy production from a carbon based energy sources such as petroleum and coal to alternative energy source such as solar, wind, wave and tides. This has led many countries to accelerate the process of shifting to green energy policy in order to minimize the effect of greenhouse gasses [1]. In the tropics where direct solar radiation is abundance all year long solar radiation has the biggest potential to be a sustainable energy source. However solar radiation has a very high variability in the tropics. In most cases the variation has caused weather (rainfall and clouds) obstructing direct sunlight and climate [2, 3]. This unpredictable nature of solar energy has led to an increase of cost of integrating renewable energy source to traditional energy source since energy market often require the prediction of hourly production of the following day [4]. Accurate solar radiation forecast have been proven to increase reliability and increase the financial value from revenue and reserve generation in a concentrated solar thermal plant [5]. Furthermore a better forecasting method in estimating the performance of solar energy source such as photovoltaic system can persuade financial backers to participate in its development [6].

Monthly average of daily solar radiation data from 3 stations in the peninsular Malaysia is obtained. The solar radiation data can be treated as a time series. The data can be fitted into a model that will be used to forecast solar radiations. Seasonal Autoregressive Integrated Moving Average (SARIMA) is very popular in modeling time series data. Another approach is to use Artificial Intelligence (AI) such as Artificial Neural Network (ANN). The use of AI have been recognized to have a better performance compare to traditional model in forecasting solar radiation [7]. The SARIMA is used when the data is assumed to be generated from a linear process while the ANN model excels in modeling data generated from a non-linear process [8]. Hybrid model is designed to capture both linear and non-linear process [9]. The hybrid model is proven to be a good model in forecasting multiple time series data such as annual energy cost budget [10], sunspot, Weekly BP/USD exchange rate [9], hourly solar radiation [11], goods subject to inspection and particulate matter [12], rainfall time series [13].

EXPERIMENTAL

Solar Radiation Data

Solar Radiation Data is obtained from 3 stations in Malaysia. Kluang, Hospital Jelebu and Batu Embun (Fig. 1) by Jabatan Meteorologi Malaysia. The data is in the form of daily data and the unit of measure is MJ/m^2. 7760 dailly solar radiation data is obtained from 1/1/1986 to 31/3/2007 in Batu Embun (3°58'N, 102°21'E), 2557 daily data from 1/1/2006 to 31/12/2012 in Hospital Jelebu (02°57'N, 102°04'E), 9740 dailly data from 1/1/1980 to 31/8/2006 in Kluang (02°01'N, 103°19'E). The data is then transformed to monthly data by using average into 168, 84, 320 monthly data points respectively.

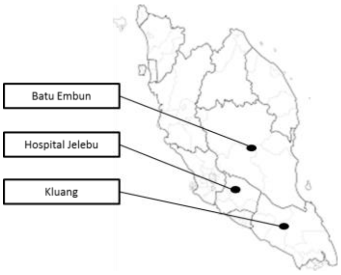

Fig. 1 The location of solar radiation data source

SARIMA model

The SARIMA model is a widely used time series model. It is an extension of the well-established ARMA model. The ARMA model is a combination of an autoregressive (AR) process and a moving average (MA) process. The AR part modeled current observation based on previous observation in the form of a linear regression, while the MA part model current data against previous value of process errors in the form of a linear regression [14]. In order to facilitate non-stationary seasonal data the ARMA model is extended to the SARIMA model. The SARIMA model utilize the well-known box-Jenkins methodology and able to extract useful statistical properties [15].

The SARIMA $(p,d,q)(P,D,Q)_s$ model has the following equation:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^S)^D Y_t = \theta_q(B)\Theta_Q(B^S)e^t \qquad (1)$$

where Y_t is the observed time series value at time t, e_t is the residual at time t, B is the backshift operator that converts Y_t(e.g $BY_t = Y_{t-1}$) , S is the seasonal period length, d is the number regular difference, D is the number of seasonal difference

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

is the non-seasonal autoregressive part of order p

$$\Phi_P(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^P$$

is the seasonal autoregressive part of order P

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

is the non-seasonal moving average part of order q

$$\Theta_Q(B) = 1 - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_Q B^Q$$

is the non-seasonal moving average part of order Q

In general there are 3 steps in modeling the SARIMA model: model Identification, validation and forecasting [16].

The data that is used in modeling must be a non-seasonal stationary data. This can be asses by using ACF and PACF that is given by the following equation:

$$ACF(k) = \rho_k = \frac{Cov(Y_t,Y_{t-k})}{Var(Y_t)} \qquad (2)$$

$$ACF(k) = \hat{\rho}_k = \frac{Cov(Y_t,Y_{t-k}|Y_{t-k},Y_{t-k-1})}{\sqrt{Var(Y_{t-1}|Y_{t-1},Y_{t-k-1})Var(Y_{t-k}|Y_{t-1},Y_{t-k-1})}} \qquad (3)$$

A seasonal data would have a spike of ACF and PACF at a certain lag. The data is then differenced by lag S where (S is the seasonal period) until it lost its seasonality. Another important characteristic of the modeled data is stationarity. To determine the stationarity of a time series a dickey-fuller test is used on the data. Dickey fuller test can determine the existence of a unit root in a time series. The existence of a unit root is evidence of a non-stationary time series [17]. The data is then differenced again with lag 1 until it is stationary. The number of seasonal differencing will determine the value of D while the number of non-seasonal differencing (differencing to obtain stationarity) will determine the value of d.

In determining the order of p, q, P and Q in the SARIMA model a few model is constructed with different orders. In estimating the parameter of the SARIMA model maximum likelihood method is used. As the order of the SARIMA model increases the model the error will be further reduced but this increase the complexity of the model and increase the possibility of over-fitting hence there might be information lost about the real underlying patter. To take this into consideration we will use Akaike information criterion (AIC). The model that has a lower value of AIC will be a more adequate model for the data [18]. The AIC is given in the following equation:

$$AIC = \ln\left(\frac{SSE}{n}\right) + \frac{2r}{n} \qquad (4)$$

The next step is to validate the model. It will determine the suitability of a given model. In this step the residual of the model is examined. The residual must be normally distributed with constant variance and zero mean besides there should be no autocorrelation between the residuals. This can be examined by looking at the residual ACF or Ljung-Box test [19]

ANN Model

The ANN model is designed based on the human neuron. A neural network has unique advantages over other model since it is designed to be able to learn from a given set of input and output. This enables the neural network model to be more flexible and data driven. Neural network model can capture a specific pattern given the examples for training [20].

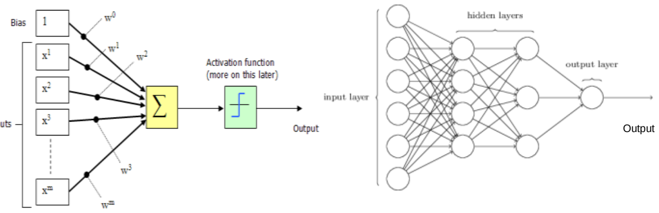

Fig. 2 A single artificial neuron (left) and a multilayer perceptron (right)

A single neuron in a neural network consists of individual inputs, weights, summation, an activation functions (Fig. 2, left). A multilayer perceptron on the other hand added a hidden layer between the input and output (Fig. 2 right).

In a time series model a neural network input consist of previous value of observations while the output is the observation at time t. The following equation illustrate the relationship between the input and output in a neural network.

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-i}\right) \quad (5)$$

where y_t is the output, α_0 is the hidden layer bias, α_j is the hidden layer weight, $g(x)$ is the activation function, β_{0j} input bias, β_{ij} is the input weight and y_{t-i} is the input .

The data of a neural network model must be differenced until there are no seasonality. The data are then transformed to a value of $-1 \leq y_t \leq 1$ [21]. A neural network usually starts with a random weight. Then the weight is changed by using training point. A training point is an input that has a known output. A training point input is presented to the perceptron. After the input is processed, the output of this process is compared with the real output (the training point output). We will do this by subtracting the process output with the real output to obtain an error. This error is then used to correct the weight.

Hybrid Model

In a SARIMA model data are assumed to be generated from a linear process. On the other hand, the ANN model assume the data are generated from a non-linear process. However it is challenging to determine whether a particular real world data is purely linear or not, more often it is mostly a mix between a linear process and a non-linear process [8]. To capture the both process a hybrid model of SARIMA-ANN is proposed by Zhang [9]. In this model the data is assumed to originate from a combination of linear and nonlinear process by a summation given by the following equation:

$$Y_t = L_t + N_t \ (6)$$

where Y_t is the observed time series data L_t is the linear component of the time series and N_t is the nonlinear component.

The hybrid model consist of two step [9]. First the data are modeled using the SARIMA model. The residual of the SARIMA model is assume to be the non-linear part that have not been captured by the SARIMA model, Hence the residual is then used to model the ANN part of the hybrid model. The forecasted value of the hybrid model will be the sum of the forecasted value of the SARIMA model and the ANN model.

Performance metrics

The performance of each model is assessed utilizing the standard error metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These metrics are commonly used in evaluating the performance of time series models. The metrics are as followings:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{N}} \qquad (7)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \qquad (8)$$

$$MAPE = \frac{1}{N}\left(\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|\right) \times 100 \qquad (9)$$

where N is the number of data, Y_i is the solar radiation data and $\widehat{Y_\iota}$, is the forecasted time series data. Model with a smaller value of the error metric is a more suitable model in predicting the solar radiation data.

Model building and validation

First the data are divided into two parts, with 12 data point at each end. These are used in calculating the performance matrix, while the rest of the data will be used in building the models. R is used as the programing langue for this research.

Next the data from each station is examined for seasonality and non-stationarity using ACF and Dickey Fuller test. From Fig. 3 it can be seen that there are yearly seasonality on all the data since there is a significance autocorrelation at a lag that is a multiple of 12. Hence the data need to be differenced by a lag of 12.

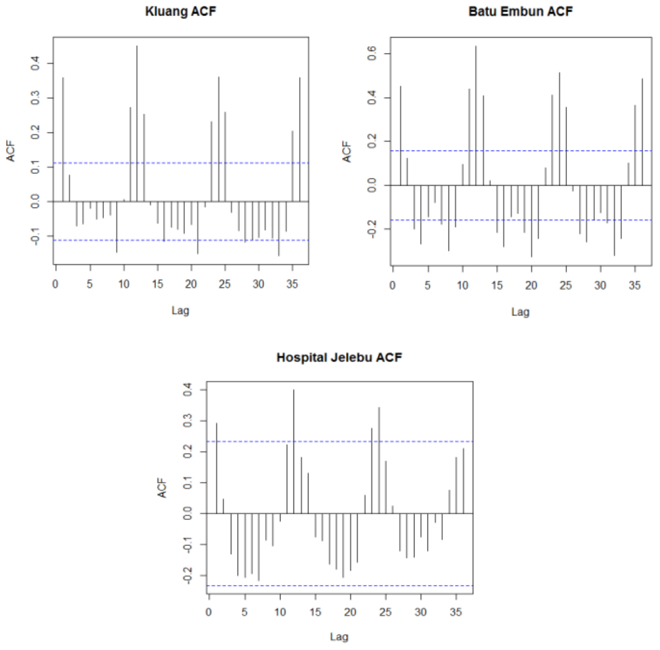
Kluang ACF
Batu Embun ACF
Hospital Jelebu ACF

Fig. 3 ACF plot from each station

After differencing we can see that there is no more significance seasonal correlation from the ACF plot (Fig. 4). The data is also stationary since there only a significance correlation. To validate this augmented dickey-fuller is used. The dickey fuller test yields a p-value that is lower than 0.1 this means that the differenced time series data is also stationary.

The next step is to determine the suitable SARIMA model. A few model is selected to be compared and the one with the smallest AIC is the best to model the data. A total of 64 models are tested with $0 \leq p \leq 3$, $0 \leq q \leq 3$, $0 \leq P \leq 1$ and $0 \leq Q \leq 1$. The results are as follows:

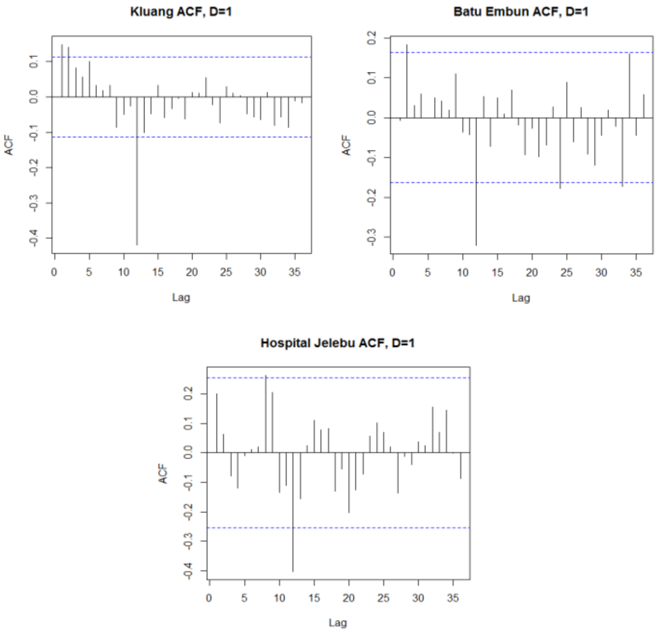
Kluang ACF, D=1
Batu Embun ACF, D=1
Hospital Jelebu ACF, D=1

Fig. 4 ACF of differenced data sets

Table 1 AIC for Kluang

$SARIMA(p,d,q) \times (P,D,Q)_S$	AIC
$(0,0,3) \times (0,1,0)_{12}$	1144.9726
$(2,0,2) \times (0,1,0)_{12}$	1144.391
$(2,0,3) \times (1,1,0)_{12}$	1144.048
$(3,0,3) \times (0,1,0)_{12}$	1144.048

Table 2 AIC for Batu Embun

$SARIMA(p,d,q) \times (P,D,Q)_S$	AIC
$(2,0,3) \times (0,1,0)_{12}$	603.1501
$(2,0,3) \times (1,1,0)_{12}$	592.4306
$(2,0,3) \times (1,1,1)_{12}$	592.4306
$(3,0,3) \times (0,1,0)_{12}$	592.4306
$(3,0,3) \times (0,1,1)_{12}$	592.8313

Table 3 AIC for Hospital Jelebu

$SARIMA(p,d,q) \times (P,D,Q)_S$	AIC
$(2,0,2) \times (0,1,0)_{12}$	260.6008
$(2,0,3) \times (0,1,0)_{12}$	255.1970
$(2,0,3) \times (1,1,0)_{12}$	256.2620
$(3,0,3) \times (0,1,1)_{12}$	257.9624

Tables 1 to 3 contain the model with the smallest AIC from all 64 model that have been tested From Table 1 we chose SARIMA $(2,0,3) \times (1,1,0)_{12}$ as the model for Kluang since it has the lowest AIC. Although SARIMA $(3,0,3) \times (0,1,0)_{12}$ have the same value of AIC, the model is rejected since it has no seasonal component. The same also applied to data from Batu Embun and Hospital Jelebu. Hence all of the data will be modeled by SARIMA $(2,0,3) \times (1,1,0)_{12}$.

For the neural network we choose a network of 13 inputs and 7 hidden neurons. This networks have been automatically chosen by the R package. The final model is the hybrid model which use the same SARIMA $(2,0,3) \times (1,1,0)_{12}$ the neural network on the other hand only have 2 inputs and 1 hidden neuron. This is due to the simplicity of the residual data. The forecast of each model is then compared with the true solar radiation value. In Fig. 5 the black line represent the true solar radiation data, red is the SARIMA model, blue is the ANN model and green is the Hybrid model.

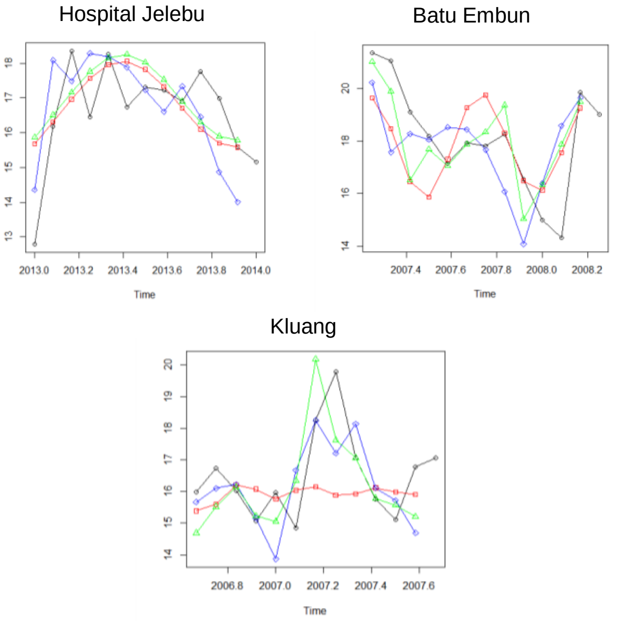

Fig. 5 Comparison of forecasted value monthy solar radiation data: x-axis time, y-axis solar radiation amount, black line true data, red line SARIMA model, blue line ANN model, green line Hybrid model

From Fig. 5 almost all of the models can modestly predict the value of solar radiation data. To accurately compare each model the performance metrics are used. The result can be seen in tables 4 to 6

Table 4 Performance Metrics for Kluang

Method	RMSE	MAE	MAPE
SARIMA	1.4936	1.1311	0.0657
ANN	1.3200	0.9900	0.0593
Hybrid	1.1982	0.9431	0.0557

Table 5 Performance Metrics for Batu Embun

Method	RMSE	MAE	MAPE
SARIMA	1.8202	1.4815	0.0833
ANN	1.9831	1.5086	0.0879
Hybrid	1.4908	1.0921	0.0648

Table 6 Performance Metrics for Hospital Jelebu

Method	RMSE	MAE	MAPE
SARIMA	1.2286	0.9099	0.0575
ANN	1.3180	1.1297	0.0699
Hybrid	1.2581	0.9404	0.0599

By examining Table 4 to 6 it can be seen that the hybrid model have the least RMSE, MAE and MAPE by a considerable difference in both Batu Embun and Kluang. The hybrid model also perform fairly well although the SARIMA model is slightly outperform the hybrid model.

CONCLUSION

In conclusion the Hybrid model is the best model to forecast solar radiation data when compared to ANN only and SARIMA only model. This means that daily average solar radiation data arise form a combination of linear and non-linear process. Other hybrid method s such as SVM-ANN, SARIMA-GARCH [13] etc. should be tested to see whether other hybrid model also can capture the underlying process of the solar radiation data.

ACKNOWLEDGEMENT

This work was financially supported by the Universiti Teknologi Malaysia under the GUP#14H00 Research university grant

REFERENCES

[1] Grantham, Adrian, Gel, Yulia R. and Boland, John. (2016). Nonparametric short-term probabilistic forecasting for solar radiation. *Solar Energy*. 133, 465–475

[2] Jiang, He, Dong, Yao and Xiao, Ling. (2017). A multi-stage intelligent approach based on an ensemble of two-way interaction model for forecasting the global horizontal radiation of India. *Energy Conversion and Management.* 137, 142–154.

[3] Monjoly, Stephanie, Andre, Maina, Calif, Rudy and Soubdhan, Ted. (2017). Hourly forecasting of global solar radiation based on multi scale decomposition methods: A hybrid approach. *Energy*. 119, 288-298.

[4] Jiménez-Pérez, Pedro F., Mora-López, Llanos. (2016). Modeling and forecasting hourly global solar radiation using clustering and classification techniques. *Solar Energy*. 135, 682–691.

[5] Law, Edward W., Kay, Merlinde and Taylor, Robert A. (2016). Evaluating the benefits of using short-term direct normal irradiance forecasts to operate a concentrated solar thermal plant. *Solar Energy*. 140, 93–108.

[6] Huang, Jing, Korolkiewicz, Małgorzata, Agrawal , Manju and Boland, John . (2013). Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. *Solar Energy*. 87, 136–149.

[7] Jha, Sunil Kr., Bilalovic, Jasmin, Jha, Anju, Patel, Nilesh, and Zhang, Han. (2017). Renewable energy: Present research and future scope of Artificial Intelligence. *Renewable and Sustainable Energy Reviews*. 77, 297–317.

[8] Khashei, M. and Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl. *Soft Computing*. 11,2664–2675.

[9] Zhang, G.P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 50, 159–175.

[10] Jeong, Kwangbok, Koo, Choongwan and Hong , Taehoon. (2003). An estimation model for determining the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network). *Energy*. 2014. 71, 71-79.

[11] Wu, Ji and Chan, Chee Keong. (2011). Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Solar *Energy*. 85, 808–817.

[12] Dı´az-Robles, Luis A., Ortega, Juan C., Fu, Joshua S., Reed, Gregory D., Chowc, Judith C., Watson, John G. and Moncada-Herrera, Juan A. (2008). A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*. 42,8331–8340.

[13] Kanel, Ibrahim Lawal, Yusof, Fadhilah. (2013) Assessment of Risk of Rainfall Events with a Hybrid of ARFIMA-GARCH. *Modern Applied Science.* 7, 78-89.

[14] Fang, Tingting and Lahdelma, Risto. (2016). Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system. *Applied Energy*. 179, 544–552.

[15] Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control. San Francisco: Holden Day.

[16] Bas, María del Carmen, Ortiz, Josefina, Ballesteros, Luisa and Martorell, Sebastian. (2017). Evaluation of a multiple linear regression model and

SARIMA model in forecasting 7Be air concentrations. Chemosphere. 177, 326-333.

[17] Dickey, D.A. and Fuller, W.A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*. 49, 1057–1072.

[18] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*. 19,716–723.

[19] Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*. 65,279–303.

[20] Haykin, Simon O. (2009). Neural Networks and Learning Machines. Ontario: Pearson.

[21] Armstrong, J. Scott. (2002). Principles of Forecasting: A Handbook for Researchers and Practitioners New York: Kluwer Academic Publishers.