Forecasting the Unit Cost of Electricity Generated by Fossil Fuel Power Plants Using ARIMA Technique: A Case Study in a Diesel/Heavy Fuel Oil Power Plant in Sri Lanka

WPMCN Weerasinghe[1#], DDM Jayasundara[1]

[1]*Department of Statistics & Computer Science, University of Kelaniya, Kelaniya 11600, Sri Lanka*
[#] WPMCN Weerasinghe; < chayanweerasinghe44@gmail.com >

Abstract— *The national grid system which is evolved to deliver electricity must be always kept in balance so that it must have a sufficient production to meet the demand of electricity while minimizing the generation cost of electricity. The forecasts made with the help of historical electricity generation cost data can support the national grid system in their decision-making activities. This study presents a statistical time series model for forecasting the Unit Cost (UC) of electricity generated by fossil fuel power plants using Auto Regressive Integrated Moving Average (ARIMA) technique. This is conducted as a case study in a Diesel/Heavy Fuel Oil (HFO) power plant in Sri Lanka which consists of two sub power stations. The model is developed and validated using 80% and 20% of monthly data that were obtained from the selected power plant from January 2013 to June 2018. ARIMA (1,1,0) and ARIMA (2,1,2) were selected as the best models with the lowest Akaike Information Criterion (AIC) for Station 1 and Station 2 respectively among many candidate models that were evaluated by the investigation of ACF and PACF of the series. The forecasting accuracy of above two models was measured with Mean Absolute Error (MAE) values (2.431 and 0.717) and Root Mean Square Error (RMSE) values (3.403 and 0.927). When comparing the UC of both stations, the forecasting values shows that UC of Station 1 are quite greater than Station 2 values and it is also relevant to past years cost data.*

Keywords— **Unit cost of electricity, time series model, ARIMA, AIC, MAE, RMSE**

I. INTRODUCTION

Energy exists in the different forms in nature but the most important form of energy is the electrical energy. All the facilities, devices, businesses, industries rely on electricity. At the same time, electricity is the most inconsistent of all types of energy, a source that must be consumed as far as it is produced because it is difficult to store the electricity. As well as electrical energy is superior to other forms of energy and a very convenient form of energy as it can be easily converted from one form to the desired form of energy. These factors together make electricity as the most significant and one of the most difficult production to understand economically.

Electricity in Sri Lanka is generated using three primary sources; thermal power which includes energy from coal and all other fuel oil sources, hydropower and other Non-Conventional Renewable Energy (NCRE) sources including solar power and wind power. Hydropower takes a share of nearly 25% of the total available grid capacity while 37% of power from coal and 34% from fuel in Sri Lanka (Utilities and Lanka, 2017). The remaining power was purchased from independent power producers including small power producers under standard power purchase agreements.

The generation cost of a unit of electricity is determined by a combination of the costs associated with the generation of the electricity and those associated with its delivery. The generation cost of electricity depends upon a large number of factors and it varies from one plant to the other. Once the plant begins to operate, the operational and maintenance costs are taken into account. Also, the costs include if there is any fuel required by the plant to produce electricity. The fuel cost is only applied to fossil fuel based power plants but not to renewable power plants. If there are any other specifications in the plants that required for the generation of electricity, the costs associated with those areas also taken into account. It is clear that the average Unit Cost (UC) of electricity generated by thermal sector (fuel and coal) incurs a high cost compared with renewable energy generation sources. As well as there is a fluctuation of UC of electricity generated by fossil fuel power plants among them (By, Commission and Lanka, 2011, 2012; Performance, 2014; Utilities, Of and Lanka, 2016).

Demand is an uncertain variable and as the network has no control over electricity demand, it must have a sufficient production at all times to meet the demand of the electricity. Some power generating plants can change the amount of electricity they produce quickly to meet any changes in the demand. But generally, the other plants that are cheapest to operate, cannot change output rapidly. Thus, a system will usually have a foundation of cheap base-load power plants that operate all the time together with a range of other, more expensive plants that are called into

service intermittently in order to maintain overall grid balance while minimizing costs.

In addition to variations in demand, some types of power plants have a variable output. They are renewable plants such as hydropower, wind and solar power plants. The output from renewable plants must be used when it is available, otherwise it is wasted. When the output from these types of plants changes, the network must have strategies for maintaining balance all the times to face any demand changes in electricity. In order for any of these aims to be achievable, the future generation cost of electricity must be predicted.

As there is a high average UC of generation of electricity among the plants that operate by Diesel/Heavy Fuel Oil (HFO) in the thermal sector in Sri Lanka, it can be considered as an important point to look up for the future generation cost of such plants. The study mainly aims at forecasting the UC of generation of electricity of fossil fuel power plants in Sri Lanka. Even though it is related to all fossil fuel power plants in Sri Lanka, due to the lack of access to the data needed for the study, this study is conducted as a case study in a prominent Diesel/HFO power plant in Western Province of Sri Lanka where the range of change of UC is very high through past few years (By, Commission and Lanka, 2011, 2012; Performance, 2014; Utilities, Of and Lanka, 2016).The selected power plant comprised of two sub power stations.

In the current literature, the reviews related to forecasting the UC of generation of electricity of fossil fuel power plants in Sri Lanka or any other country cannot be found which is done by using time series forecasting approaches. Currently there are only production costing models for forecasting the expected cost of producing electricity for a given power generation system. Production costing models are used in the electric power industry to forecast the expected amount of electricity produced by different power generation units and the expected cost of producing electricity for a given power generation system (Shih and Bloom, 2018). Time series modeling approach used throughout this study considers the time-based variation of the generation cost of electricity of the selected Diesel/HFO power plant. In that case, time series modeling approach seems a better way of forecasting the generation cost of electricity of a Diesel/HFO power plant.

II. METHODOLOGY AND EXPERIMENTAL DESIGN

The univariate time series modelling approach was used for the data set after evaluating its time series properties.

A. Data Collection

The data needed for the study were collected from the two sub stations separately in the selected power plant. The data set contains the monthly data from January 2013 to June 2018 (66 data points for each station). The data of UC of generation of electricity from the selected power plant were obtained. In econometrics, a production company's total cost is comprised of two types of costs as fixed costs and variable costs. Fixed costs do not change with the units of production of a company and usually not relevant to the output decisions while variable costs solely depend on the units of production.

Fixed costs associated with the UC values are personal expenses, maintenance cost, water treatment plant chemical cost and variable cost associated are Diesel cost, HFO cost, lube oil cost, Diesel price, HFO Price, lube oil price, water bill, plant factor (%), number of units generated from Diesel, and number of units generated from HFO. Plant Factor of a power plant is the ratio of the actual energy output of the power plant over a period of time to its potential output if it had operated at full nameplate capacity the entire time [5]. The data set was divided into two parts as 80% and 20% for the model building and model validation respectively. The statistical package used for model building is R software.

B. Preliminary Analysis

Data cleaning is one of the most common data pre-processing technique. It includes fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies. In this study, the data set is first explored to identify the outliers and the missing values. Four missing value imputation methods were used in this study namely mean imputation, linear interpolation, forecasting backward with Auto Regressive Integrated Moving Average (ARIMA) model and exponential smoothing. Outliers are simply the observations that are very different from the observations in a data set. The "tsoutliers" function in R software is designed to identify outliers and to suggest potential replacement values and it was used in this study to replace outliers.

A stationary time series can be identified as a time series whose properties specially mean and variance are constant over the time which the time series is observed. There are some statistical tests to identify whether a time series is stationary or not. The three tests Kwiatkowski-Phillips-Schmidt-Shin (KPSS), Augmented Dickey Fuller (ADF) and Phillips Perron (PP) were used in this study to check the stationary of the time series.

C. Time Series Forecasting Methods

A time series is a collection of observations made sequentially over time. There can be regular spaced time series that are observed at regular intervals of time such as hourly, daily, weekly, monthly, quarterly, annually or irregular spaced time series. The aim of forecasting time

series data is to estimate how the sequence of observations will continue into the future. Time series models used for forecasting include decomposition models, exponential smoothing models and ARIMA models. Predictor variables are also often useful in time series forecasting. That type of model is known as an explanatory model. An explanatory model incorporates information about other variables rather than only historical values of the variable to be forecast. Time series regression models can be considered under these explanatory models. This study has used one main approach of time series forecasting methods; univariate time series approach.

D. Univariate Time Series Approach: ARIMA Model

ARIMA model can be fitted to a univariate stationary time series. Non-seasonal ARIMA model can be obtained by combining the differencing with auto regression and a moving average model. The full model can be written as in Equation (1).

$$y`_t = c + \phi_1 y`_{t-1} + \phi_2 y`_{t-2} + \cdots\cdots + \phi_p y`_{t-p} + \varepsilon t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots\cdots + \theta_p\varepsilon_{t-p} \quad (1)$$

Where $y`_t$ is the differenced series. The "predictors" on the right-hand side include both lagged values of y_t and lagged errors. This is referred as ARIMA (p,d,q) model where p is the order of the autoregressive part, d is the degree of the differencing involved and q is the order of the moving average part.

ARIMA has four major steps as model building and identification, estimation, diagnostics and forecast. First tentative model parameters are identified through Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF), then coefficients of the most likely model are determined, next steps involve is to forecast, validate and check the model performance by observing the residuals through Ljung Box test and ACF plot of residuals.

E. Forecasting Accuracy

The difference between actual and forecasted values shows how well the model has performed. The main idea of forecasting techniques is to minimize the difference between actual and forecasted value since this should influence the performance and reliability of the model. The smaller the difference, the better the model is. Several criteria such as Root Mean Square Error **(**RMSE), Mean Absolute Error **(**MAE), Mean Absolute Percentage Error (MAPE) and Weighted MAPE can be used to compare different forecasting models. In this study, two different error metrics are considered for the evaluation of the forecasting models; MAE and RMSE.

RMSE is the square root of average of sum-squared errors and is given by the Equation (2) while MAE is given in Equation (3).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (2)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y_i}|}{y_i} \quad (3)$$

where $\hat{y_i}$, y_i , n represents the estimated value of y_i , actual value and number of observations respectively.

III. RESULTS

Results and the analysis of the performance of the concerned forecasting model ARIMA is presented and mainly the results and interpretations are separately given for two sub stations in the selected Diesel/HFO power plant.

A. Preliminary Analysis

The past UC values from January 2013 to May 2017 was evaluated under this part for both sub power stations separately while the data from June 2017 to June 2018 was used for model validation. There was 9·46% of missing values in the data set in both power stations. Four missing value imputation methods have been carried out in this study; mean imputation, linear interpolation method, back casting with ARIMA model and exponential smoothing. According to Table 1 and Table 2, linear interpolation method was identified as the best missing value imputation method with minimum MAPE and RMSE values.

In some time periods the number of units of electricity generated by the power plant can be very low due to many factors such as the demand is already fulfilled by another power plant, due to shut down of the plant for maintenance purposes. So that there can be high cost in that time periods which are defined to be outliers in this study. In order to maintain the continuity of the time series and due to above mentioned reason for occurring the outliers, they were not removed in the study.

Table 1. Missing value imputation – Station 1

Imputation method	MAPE	RMSE
Mean imputation	2.8393	6.9336
Linear interpolation	**1.6242**	**1.7948**
Back casting with ARIMA Model	19.1789	19.7874
Back casting with exponential smoothing	9.8738	22.5695

Imputation method	MAPE	RMSE
Mean imputation	1.1571	0.8889
Linear interpolation	**0.6402**	**0.8686**
Back casting with ARIMA Model	5.1907	1.8943
Back casting with exponential smoothing	5.7121	1.9603

Stationary of the time series of UC were checked with ADF, KPSS and PP test where the tests concluded that series of both stations were not stationary at 5% level of significance according to the results shown in Table 3.

Table 3. Stationary test results

Variable	p value		
	ADF	KPSS	PP
UC of generation of electricity (Station 1)	0.3817	0.0719	0.6256
UC of generation of electricity (Station 2)	0.3895	0.0510	0.3663

Since the series is univariate and not stationary in both cases, an ARIMA model was selected by evaluating the ACF and PACF of the series.

B. Univariate time series approach – Station 1

Among the candidate models in Table 4 which was selected based on the cut off values of ACF and PACF, ARIMA (1,1,0) model was selected as the best model with the minimum Akaike Information Criterion (AIC) value 323·59 for forecasting the UC of generation of electricity in Station 1.

Table 4. Candidate ARIMA models in Station 1

ARIMA model	AIC value
ARIMA (1, 1, 1)	352.56
ARIMA (0, 1, 1)	323.90
ARIMA (1, 1, 0)	**323.59**

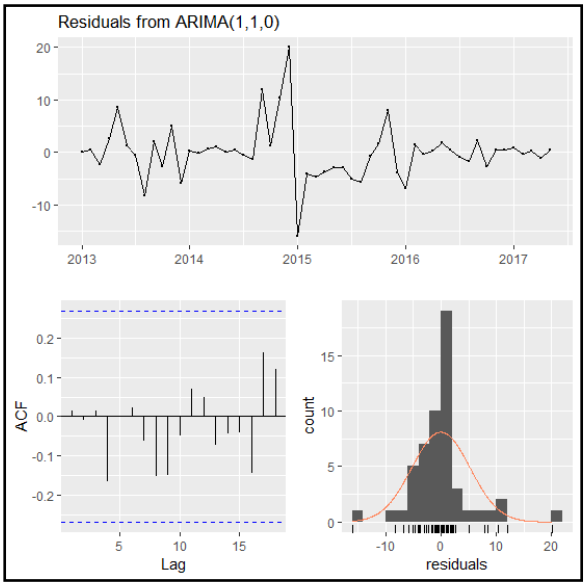

Figure 1. Model adequacy of ARIMA (1, 1, 0) model

ARIMA (1,1,0) model was estimated as in Equation (4).

$$X_t = X_{t-1} + 0.3329\,(X_{t-1} - X_{t-2}) \quad (4)$$

p value of the Ljung Box test is 0·8604 which is greater than 0·05 significance level and it suggests that the residuals are independently distributed. As in Figure 1, the ACF plot of the residuals from the ARIMA (1,1,0) model also shows that all autocorrelations are within the threshold limits indicating that the residuals are independently distributed.

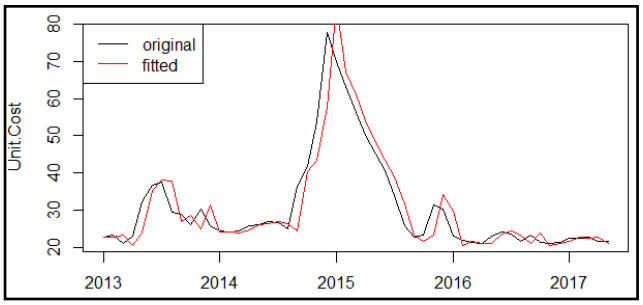

Figure 2. Actual and fitted values of UC of Station 1

Figure 2 shows the plot of UC values calculated from Equation (4) with the actual UC values of Station 1. Hence the gaps between actual and fitted values are minimum, this model can be used to forecast the UC beyond the year 2017.

Forecasting accuracy with the test data is measured with RMSE and MAE with values 3·403 and 2·431 under the model validation as shown in Figure 3.

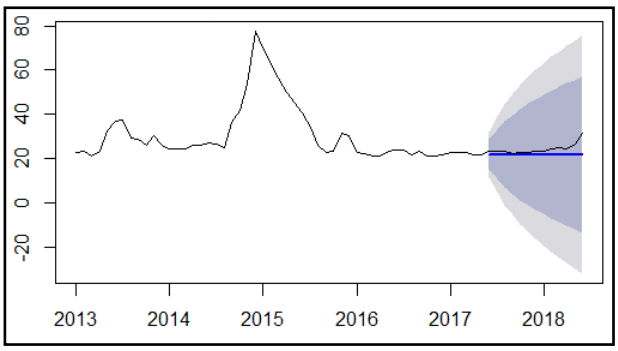

Figure 3. Forecasts from ARIMA (1, 1, 1) model

Table 5 represents some forecasted values of the UC compared with the actual values in Station 1.

Table 5. Actual and predicted values of UC of Station 1

Time	Actual value	Predicted value
June 2017	23.33	21.81326
July 2017	22.13	21.83099
August 2017	23.35	21.83689
September 2017	22.10	21.83885
October 2017	22.76	21.83950
November 2017	22.62	21.83972
December 2017	22.53	21.83979
January 2018	22.14	21.83982

C. Univariate time series approach – Station 2

The same procedure above was followed in finding a suitable forecasting model for Station 2. ARIMA (2,1,2) model returned the minimum AIC value (155·51) among the candidate models evaluated in Table 6 and it was selected as the best model for forecasting the UC of generation of electricity of the Station 2.

Table 6. Candidate ARIMA models in Station 2

ARIMA model	AIC value
ARIMA (1, 1, 1)	159.22
ARIMA (1, 1, 2)	159.49
ARIMA (2, 1, 1)	161.27
ARIMA (2, 1, 2)	**155.51**

The best model can be expressed as in Equation (5).

$$X_t = X_{t-1} + 1.2607\,(X_{t-1}-X_{t-2}) - 0.889\,(X_{t-2}-X_{t-3}) - 0.9708\varepsilon_{t-1} + 0.8275\,\varepsilon_{t-2} \quad (5)$$

Ljung Box test returns a large p value of 0·7909 indicating that the residuals are independent. According to Figure 4, it is obvious there is no significant spike of ACF of residual series of ARIMA (2,1,2) model. It also confirms that the residuals of this selected ARIMA model are independently distributed.

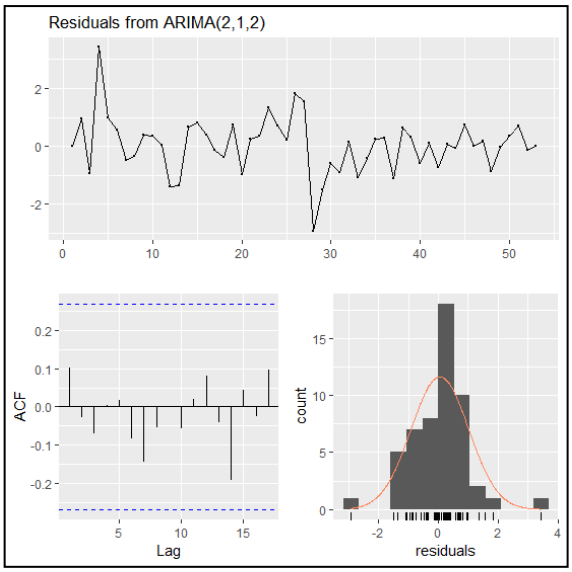

Figure 4. Model adequacy of ARIMA (2, 1, 2) model

Figure 5 shows the plot of fitted UC values against the actual UC values to demonstrate the correlation of accuracy.

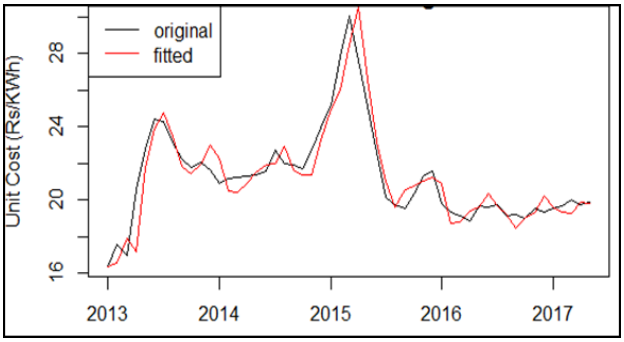

Figure 5. Actual and fitted values of UC of Station 2

According to Figure 5, it is clear that the performance of the ARIMA (2,1,2) model selected is quite impressive as there are some instances of closely related to actual and fitted values.

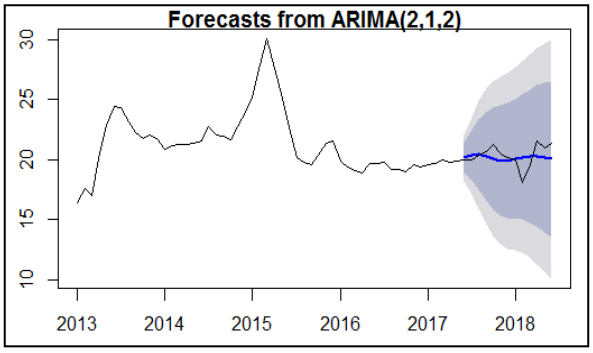

Figure 6. Model validation of ARIMA (2, 1, 2) model

Forecasting accuracy of the ARIMA (2,1,2) was measured with RMSE and MAE values of 0·927 and 0·717 respectively and it can be shown graphically under model validation as shown in Figure 6.

Some forecasted values of the UC in Station 2 compared with the actual values are given in Table 7.

Table 7. Actual and predicted values of UC of Station 2

Time	Actual value	Predicted value
June 2017	19.92	20.12149
July 2017	19.93	20.35327
August 2017	20.39	20.41301
September 2017	20.68	20.28228
October 2017	21.22	20.06436
November 2017	20.46	19.90583
December 2017	20.15	19.89970
January 2018	19.88	20.03290

IV. DISCUSSION AND CONCLUSION

The past UC values from the selected Diesel/HFO power plant from January 2013 to May 2017 was evaluated in this study. The data points were keenly analysed. There were found to be 5 missing values in each station and the missing values were imputed using linear interpolation method which was identified as the best missing value imputation method according to Table 1 and Table 2. The number of units of electricity generated by the power plant can be very low due to many reasonable factors and hence there can be high cost in that time periods. Those points are identified as the outliers in the study. Due to above mentioned reason for occurring the outliers, they were not removed in the study.

After investigating time series univariate approach separately for Station 1 and Station 2, ARIMA (1,1,0) and ARIMA (2,1,2) were the best models that were evidently selected for forecasting the UC of generation of electricity of the selected power plant. According to Box –Jenkins, when the differencing order is greater than 0, constant should not be included in the ARIMA model except for the series showing significant trend. As the series of UC does not show any growth or a deterministic trend, constant term is not included in fitting the above ARIMA models given in Equation (4) and Equation (5).

The study also statistically tested and validated that the successive residuals in the fitted univariate time series models were independent and the residuals seems to be normally distributed with mean zero and constant variance. Both Station 1 and Station 2 shows a minimum values for error metrics MAE (2·431 and 0·717) and RMSE (3·403 and 0·927) concluding that two ARIMA models have a strong potential for forecasting the UC of generation of electricity of the selected power plant and can compete favorably with existing techniques for prediction of UC.

One month ahead forecast of UC value of generation of electricity of Station 1 is influenced by past two months UC values while one month ahead forecast of UC value of generation of electricity of Station 2 is influenced by past three months UC values as well as by past two months error terms. These models could guide the national grid system to make profitable power generating plan by considering the time-based variations. The two forecasting models were estimated by using monthly recorded data available at the selected power station. Further accurate forecasts can be obtained if there were weekly or daily data records.

This forecasting method can be generalized to other fossil fuel power plants with necessary alterations.

ACKNOWLEDGEMENT

Authors wish to thank Department of Statistics & Computer Science , the Diesel/HFO power plant selected for the case study, Mr. W A Sirisena and Mr. S B M S S Gunarathne for their enormous contribution.

REFERENCES

By, P., Commission, P. U. and Lanka, S. (2011) 'Generation Performance in Sri Lanka'.

By, P., Commission, P. U. and Lanka, S. (2012) 'Generation Performance in Sri Lanka Prepared By : Public Utilities Commission of Sri Lanka'.

Performance, G. (2014) 'Generation Performance Generation Performance 2014 (First Half) in Sri Lanka in Sri Lanka', 2014.

Shih, F. and Bloom, J. A. (2018) 'Asymptotic Mean and Variance of Electric Power Generation System Production Costs via Recursive Computation of the Fundamental Matrix of a Markov Chain Author (s): Fen-Ru Shih , Mainak Mazumdar and Jeremy A . Bloom Published by : INFORMS Stable URL : ht', 47(5), pp. 703–712.

Utilities, P. and Lanka, S. (2017) 'Annual Report 2017 (Draft)', 2017.

Utilities, P., Of, C. and Lanka, S. R. I. (2016) 'Generation Report', pp. 1–35.