International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Modelling and Forecasting the Unit Cost of Electricity Generated by Fossil Fuel Power Plants in Sri Lanka

W. P. M. C. N. Weerasinghe[a]*, D. D. M. Jayasundara[b]

[a,b]*Department of Statistics & Computer Science, University of Kelaniya, Kelaniya 11600, Sri Lanka*

[a]*Email: chayanweerasinghe44@gmail.com*

[b]*Email: jayasund@kln.ac.lk*

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. This study presents a statistical time series model for forecasting the Unit Cost (UC) of generation of electricity in fossil fuel power plants by using two approaches namely Auto Regressive Integrated Moving Average (ARIMA) and time series regression. This is conducted as a case study in a Diesel/Heavy Fuel Oil (HFO) power plant in Sri Lanka which consists of two sub stations. ARIMA (1,1,0) and ARIMA (2,1,2) were selected as the best models with the lowest Akaike Information Criterion (AIC) under the ARIMA model approach while two dynamic regression models with coefficient of determination (R^2) value 0.55 were selected under time series regression approach for Station 1 and Station 2 respectively. The regression model was identified as the best forecasting method for two stations with the minimum Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The forecasts of the future generation cost of electricity are extensively helpful for the national grid system for financial and capacity planning, fuel management and operational planning.

Keywords: AIC; ARIMA; coefficient of determination; dynamic regression mode; MAE; UC of generation of electricity.

---

* Corresponding author.

1. 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 biomass, coal and all other fuel oil sources, hydro power including small hydro 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 [8]. 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 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 base 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 area 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 [3,4,5,6,7,8]. 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, in order to maintain overall grid balance while minimizing costs, a system will normally 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 as demand changes. 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 such 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 of 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 lookup for the future generation cost of such plants and identify the key factors that effect for the unit generation cost of fossil fuel power plants in Sri Lanka. The study mainly aims at forecasting and finding the factors effecting to the UC of generation of electricity of fossil fuel power plants in Sri Lanka. Even though it is related to all fossil fuel plants in the 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 [3,4,5,6,7,8]. The selected power station 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. The electricity market has qualities which distinguish it from other commodities; it cannot be appreciably stored and requires constant balance between supply and demand and exhibits inelastic demand over short time periods. Therefore, in a deregulated market, forecasting the generation cost of power and power price are essential for market participant's survival [9]. The 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. These forecasts are used extensively by the industry for financial and capacity planning, fuel management, and operational planning. The production cost models account for the expected variation of load (i.e., demand for power) over time and the uncertainty in the utilization of the generating units resulting from their failures and repairs. The production cost of a power generating system over a given time interval is obtained by adding the amounts of energy produced by each unit in megawatt- hours (MWH), multiplied by its operating cost ($/MWH) [2]. There are two generally accepted methods for estimating power generation costs; 'model plant method' and the method using corporate financial statements. This study has used the corporate financial statements, though under some constraints, can provide useful information for computing thermal and nuclear power generation costs. This method was used in this study to estimate the thermal and nuclear power generation costs in Japan for past five years where there was wild fluctuations in primary energy prices [12]. 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. The production cost models consider only the basic factors which are effecting for the generation cost of electricity such as the expected variation of demand of power, availability of generating units and cost of fuel [2]. But time series modelling approach used throughout this study consider all possible factors which are affecting for the generation cost of electricity of the selected Diesel/HFO power plant together with the variations in time periods. Since the value of the all productions varies rapidly in a lower middle income country like Sri Lanka, it is not fair to consider only the basic factors that effects for the generation cost of electricity in a power plant as in production costing models. In that case, time series modelling approach seems a better way of forecasting the generation cost of electricity of a Diesel/HFO power plant.

2. Materials and Methods

Two main approaches to the research problem with their methodologies are discussed in here: univariate time series approach where only the historical data of UC of generation of electricity of the selected Diesel/Heavy Fuel Oil Power Plant is used to determine the future movement of UC and time series regression approach where other explanatory variables that describes the UC are used in developing the model for forecasting the future UC.

2.1. 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 and every variable used in the study). The data needed for the study were obtained from the selected power plant. In econometrics, a production company's total cost is composed with 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 are solely depend on the units of production. Fixed costs used in the study are personal expenses (LKR), maintenance cost (LKR), water treatment plant chemical cost (LKR) and variable costs used are Diesel cost (LKR), HFO cost (LKR), lube oil cost (LKR), Diesel price (LKR), HFO Price (LKR), lube oil price (LKR), water bill (LKR), 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 [3]. The data set was divided into two parts as 80% and 20% for the model building and model validation respectively. The statistical packages used for model building are R and E-views softwares.

2.2. Preliminary Analysis

Data cleaning is one of 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 [1], linear interpolation [10], forecasting backwards with Auto Regressive Integrated Moving Average (ARIMA) model [11] 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.

2.3. 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. When forecasting time series data, the aim 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 is useful because it incorporates information about other variables rather than only historical values of the variable to be forecast. Time series regression models can be considered under this explanatory models. This study have used two main approaches of time series forecasting methods; univariate time series approach and regression model approach.

2.4. 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).

$$y_t^{`} = c + \phi_1 y_{t-1}^{`} + \phi_2 y_{t-2}^{`} + \cdots + \phi_p y_{t-p}^{`} + \varepsilon + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (1)$$

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

2.5. Time Series Regression Model

There are lots of instances which require the investigation of relationships between two and more variables. Regression analysis is a typical method that is being used for this kind of problems. The aim of regression analysis is to estimate the dependencies between dependent variable and a set of independent variables. Under the time series regression, the time series of interest is forecast by assuming it has linear relationship with other times series. In here, both predictor and the response variables are time series. A common way to summarize how well a linear regression model fits the data is through the coefficient of determination or R^2. After selecting the regression variables and fitting a regression model, it is necessary to plot the residuals to check that the assumptions of the model have been satisfied. Existence of auto correlation and serial correlation was checked with Breusch Godfrey Serial Correlation LM Test where the test null hypothesis is presence of no serial correlation. Jarque – Bera Test and histogram of residuals were used to check the normality assumption of residuals in this study. The null hypothesis of the Jarque – Bera Test is that the residuals follow the normal distribution. One of the informal method to detect heteroscedasticity is drawing the error versus fitted values in a scatter plot. If it shows any systematic pattern, heteroscedasticity may be present in the problem. Breusch Pagan Godfrey test was used in this study to check the evidence of heteroscedasticity where the null hypothesis of the test is presence of no heteroscedasticity in the residuals. Multicollinearity which means the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of a regression model is a huge problem in regression models and it can be detected through the existence of high R^2 value but few significant t ratios, high pairwise correlation among regressors, auxiliary regressions and Variation Inflation Factors (VIF).

2.6. Forecasting Accuracy

The difference between actual and predicted values shows how well the model has performed. The main idea of forecasting techniques is to minimize this value since this should influence the performance and reliability of the model. The smaller the difference, the better the model is. Several criterias such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Weighted Mean Absolute

Percentage Error (WMAPE) 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).

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

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

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

3. Results

Results under the two approaches; univariate approach and time series regression are clearly explained in here.

3.1. Preliminary Analysis

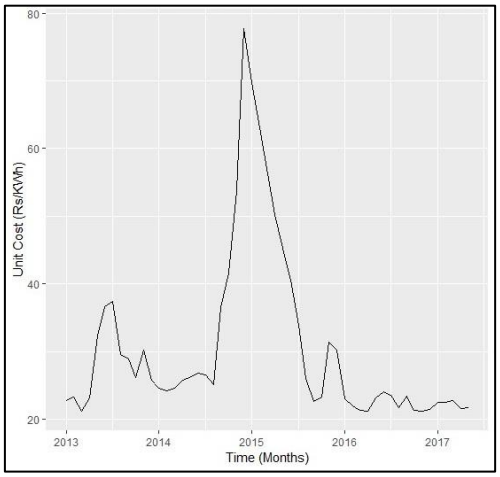


Figure 1: Time Series Plot of UC of generation of electricity in Station 1

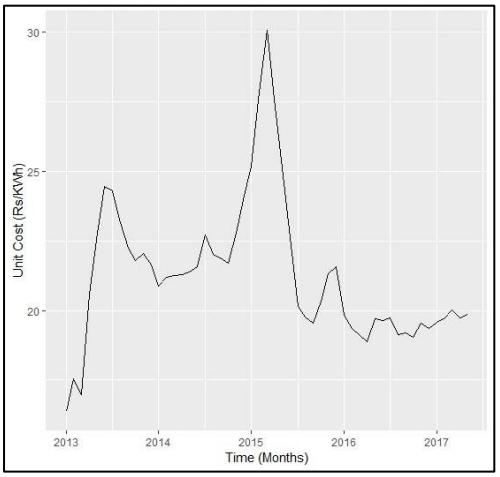


Figure 2: Time Series Plot of UC of generation of electricity in Station 2

The past UC values and the explanatory variables data 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% (5 missing values) 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. Linear interpolation method was identified as the best missing value imputation method with MAPE of 1.62429, 0.64024 and RMSE of 1.79484, 0.86869 in Station 1 and Station 2 respectively. 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 can be identified as the outliers in this study. In order to maintain the continuity of the time series and due to above mentioned reason for occuring the outliers, they were not removed in the study.

The time series plots of UC of Station 1 and Station 2 after imputing missing values are shown in Figure 1 and Figure 2. According to Figure 1 and Figure 2, it is clearly seen that the both series do not have a constant mean or constant variance so that series seem to be non – stationary. Stationary of the time series of UC were checked with statistical tests ADF, KPSS and PP where the tests concluded that the series of both stations were not stationary at 5% level of significance according to the results shown in Table 1.

Table 1: p values – Stationary Tests

Variable	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

3.2. Univariate Time Series Approach

Since the series is univariate and not stationary in both cases, an ARIMA model was selected among the tentative models identified through the investigation of ACF and PACF of the differenced series. ARIMA (1,1,0) model and ARIMA (2,1,2) were selected as the best models with the minimum AIC values (323.59 and 155.51) for Station 1 and Station 2 respectively for modelling the UC of generation of electricity. ARIMA (1,1,0) model and ARIMA (2,1,2) models were estimated as in Equation (4) and Equation (5) respectively.

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

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

Residuals of the fitted models for Station 1 and Station 2 were evaluated as shown in Figure 3 and Figure 4 respectively.

The Ljung Box test returns large p values which are equal to 0.8604 and 0.7909 for Station 1 and Station 2 respectively indicating that the successive residuals of the fitted models are independently distributed. According to Figure 3 and Figure 4, ACF plot of the residuals from both ARIMA (1,1,0) and ARIMA (2,1,2) model shows that all autocorrelations are within the threshold limits indicating the residuals are behaving random. Figure 5 shows the plot of UC values calculated from Equation (4) with the actual UC values of Station 1 while Figure 6 shows the plot of UC values calculated from Equation (5) with the actual UC values of Station 2. Hence the gaps between actual and fitted values is minimum, these models can be used to forecast the UC beyond year 2017.

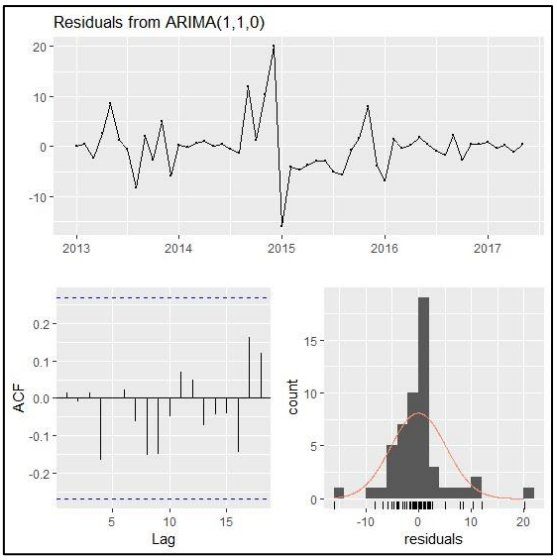


Figure 3: Model adequacy of ARIMA (1,1,0) Model

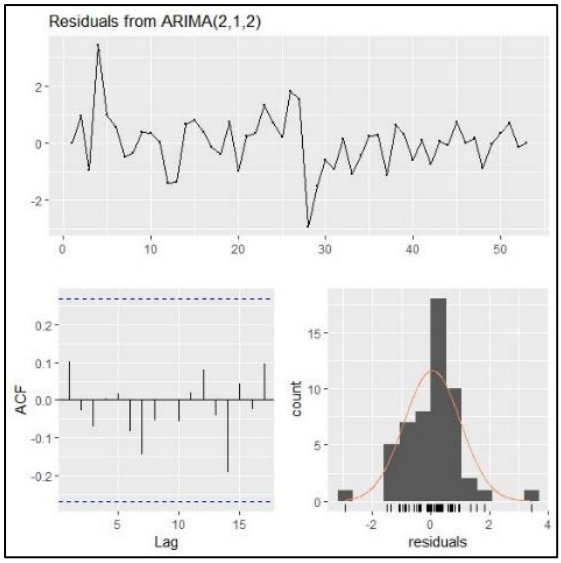


Figure 4: Model adequacy of ARIMA (2,1,2) Model

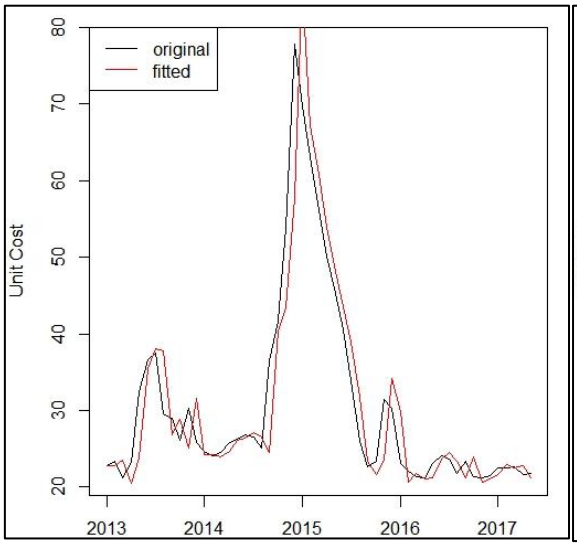


Figure 5: Actual and Fitted UC values – Station 1

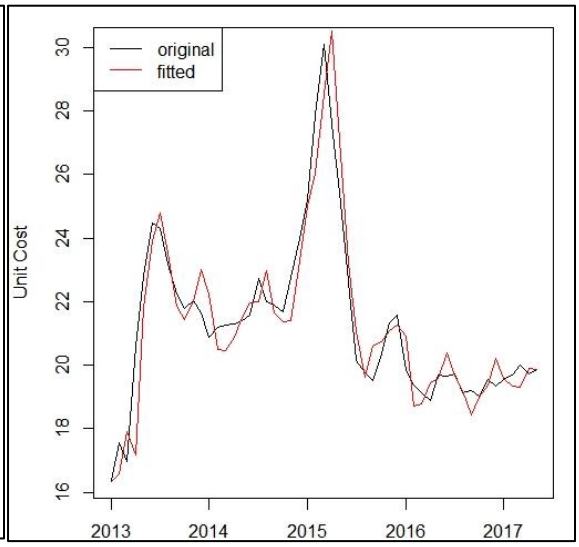


Figure 6: Actual and Fitted UC values – Station 2

Forecasting accuracy of the fitted two ARIMA models was measured with error metrics RMSE and MAE and the model validation is shown in Figure 7 and Figure 8.

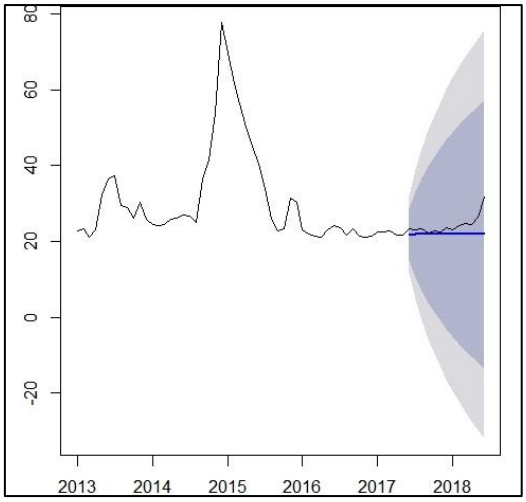


Figure 7: Model Validation – ARIMA (1,1,1)

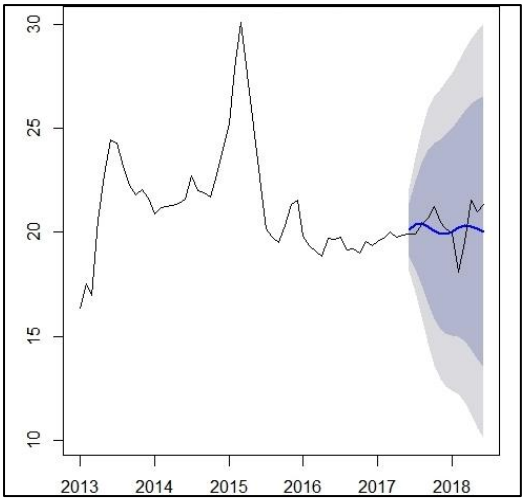


Figure 8: Model Validation – ARIMA (2,1,2)

3.3. Time Series Regression Approach

Two time series regression models were evaluated for modelling and forecasting the UC of generation of electricity in Station 1 and Station 2.

Table 2: Notations used in Regression Equations to denote variables

Variable	Station 1	Station 2
UC of generation of electricity	Y_1	Y_2
Diesel Cost	X_1	X_1
Diesel Price	X_2	Z_2
HFO Cost	X_3	Z_3
HFO Price	X_4	Z_4
Y_1Lube Oil Cost	X_5	Z_5
Lube Oil Price	X_6	Z_6
Maintenance Cost	X_7	Z_7
Personal Expenses	X_8	Z_8
Plant Factor	X_9	Z_9
Units generated from Diesel	X_{10}	Z_{10}
Units generated from HFO	X_{11}	Z_{11}
Water Bill	X_{12}	Z_{12}
Water treatment plant chemical cost	X_{13}	Z_{13}

UC of generation of electricity is the dependent variable while personal expenses, maintenance cost, Diesel Cost, HFO Cost, lube oil cost, Diesel price, HFO price, lube oil price, water bill, plant factor, number of units generated from Diesel, number of units generated from HFO were selected as the independent variables. Table 2 shows the notations used to represent each dependent and independent variables in writing the time series regression

models. Initially a scatterplot matrix is drawn for each station with independent and dependent variables. According to the scatter plot matrix in Figure 9 and Figure 10, it can be seen that the relationship between dependent and independent variables is not linear in case of both sub stations. A transformation was used in later models to gain the linearity.

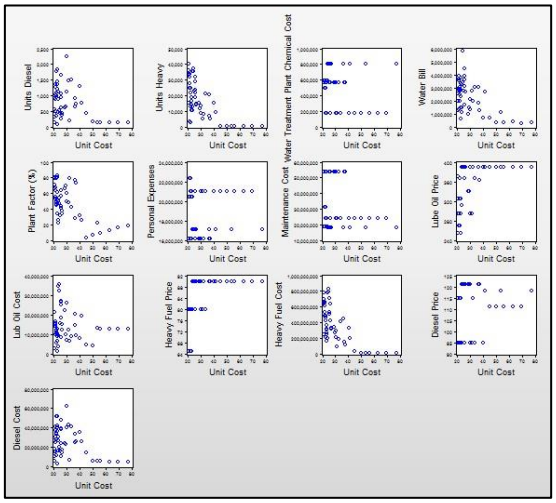


Figure 9: Scatter plot matrix of variables – Station 1

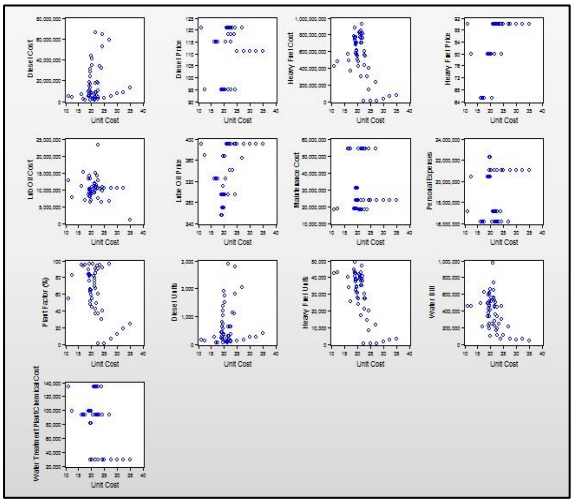


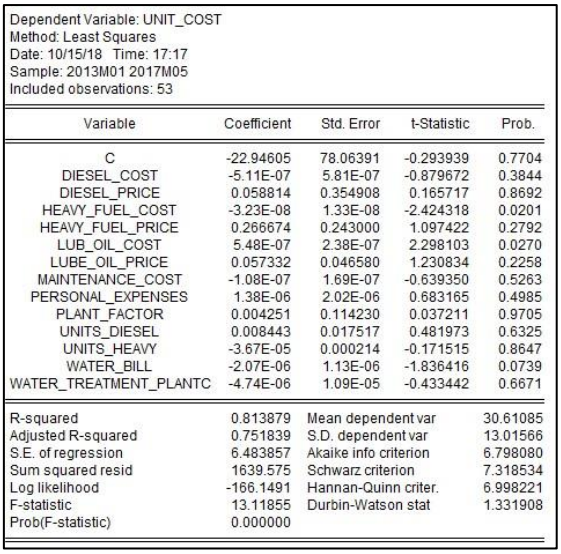
Figure 10: Scatter plot matrix of variables – Station 2

Initially, a time series regression model with all level variables was fitted to both stations but both models give the evidence of multi collinearity. The initial models fitted to Station 1 and Station 2 are given in Equation (6) and Equation (7) respectively.

$$Y_{1,t} = c_0 + c_1 X_{1,t} + c_2 X_{2,t} + c_3 X_{3,t} + c_4 X_{4,t} + c_5 X_{5,t} + c_6 X_{6,t} + c_7 X_{7,t} + c_8 X_{8,t} + c_9 X_{9,t} + c_{10} X_{10,t} + c_{11} X_{11,t} + c_{12} X_{12,t} + c_{13} X_{13,t} \quad (6)$$

$$Y_{1,t} = c_0 + c_1 Z_{1,t} + c_2 Z_{2,t} + c_3 Z_{3,t} + c_4 Z_{4,t} + c_5 Z_{5,t} + c_6 Z_{6,t} + c_7 Z_{7,t} + c_8 Z_{8,t} + c_9 Z_{9,t} + c_{10} Z_{10,t} + c_{11} Z_{11,t} + c_{12} Z_{12,t} + c_{13} Z_{13,t} \quad (7)$$

According to outputs of the models in Figure 11 and Figure 12, initial regression model with level variables give the R^2 values of 0.81 and 0.91 for Station 1 and Station 2 respectively but they give only few significant independent variables. It was found that there is a high correlation between the variables; Diesel cost and units generated from Diesel (0.9821 for Station 1 and 0.9891 for Station 2), HFO cost and units generated from HFO (0.9031 for Station 1 and 0.9476 for Station 2) in both stations. This was also confirmed with VIF values and auto regressions results in Table 3.

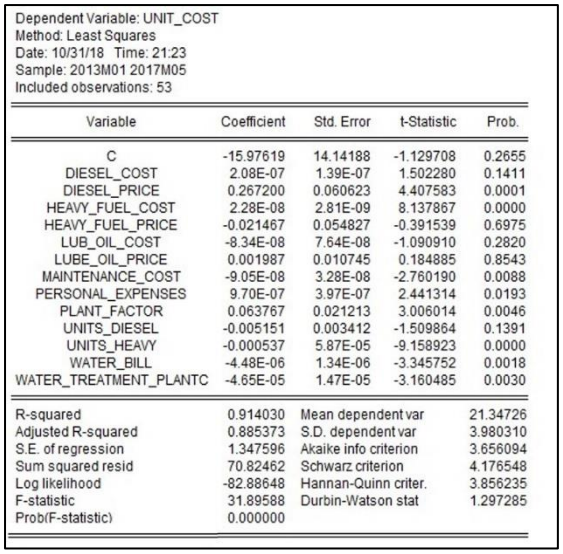


Dependent Variable: UNIT_COST
Method: Least Squares
Date: 10/15/18 Time: 17:17
Sample: 2013M01 2017M05
Included observations: 53

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-22.94605	78.06391	-0.293939	0.7704
DIESEL_COST	-5.11E-07	5.81E-07	-0.879672	0.3844
DIESEL_PRICE	0.058814	0.354908	0.165717	0.8692
HEAVY_FUEL_COST	-3.23E-08	1.33E-08	-2.424318	0.0201
HEAVY_FUEL_PRICE	0.266674	0.243000	1.097422	0.2792
LUB_OIL_COST	5.48E-07	2.38E-07	2.298103	0.0270
LUBE_OIL_PRICE	0.057332	0.046580	1.230834	0.2258
MAINTENANCE_COST	-1.08E-07	1.69E-07	-0.639350	0.5263
PERSONAL_EXPENSES	1.38E-06	2.02E-06	0.683165	0.4985
PLANT_FACTOR	0.004251	0.114230	0.037211	0.9705
UNITS_DIESEL	0.008443	0.017517	0.481973	0.6325
UNITS_HEAVY	-3.67E-05	0.000214	-0.171515	0.8647
WATER_BILL	-2.07E-06	1.13E-06	-1.836416	0.0739
WATER_TREATMENT_PLANTC	-4.74E-06	1.09E-05	-0.433442	0.6671

R-squared	0.813879	Mean dependent var	30.61085
Adjusted R-squared	0.751839	S.D. dependent var	13.01566
S.E. of regression	6.483857	Akaike info criterion	6.798080
Sum squared resid	1639.575	Schwarz criterion	7.318534
Log likelihood	-166.1491	Hannan-Quinn criter.	6.998221
F-statistic	13.11855	Durbin-Watson stat	1.331908
Prob(F-statistic)	0.000000		

Figure 11: Output of the model in Equation (6)



Dependent Variable: UNIT_COST
Method: Least Squares
Date: 10/31/18 Time: 21:23
Sample: 2013M01 2017M05
Included observations: 53

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-15.97619	14.14188	-1.129708	0.2655
DIESEL_COST	2.08E-07	1.39E-07	1.502280	0.1411
DIESEL_PRICE	0.267200	0.060623	4.407583	0.0001
HEAVY_FUEL_COST	2.28E-08	2.81E-09	8.137867	0.0000
HEAVY_FUEL_PRICE	-0.021467	0.054827	-0.391539	0.6975
LUB_OIL_COST	-8.34E-08	7.64E-08	-1.090910	0.2820
LUBE_OIL_PRICE	0.001987	0.010745	0.184885	0.8543
MAINTENANCE_COST	-9.05E-08	3.28E-08	-2.760190	0.0088
PERSONAL_EXPENSES	9.70E-07	3.97E-07	2.441314	0.0193
PLANT_FACTOR	0.063767	0.021213	3.006014	0.0046
UNITS_DIESEL	-0.005151	0.003412	-1.509864	0.1391
UNITS_HEAVY	-0.000537	5.87E-05	-9.158923	0.0000
WATER_BILL	-4.48E-06	1.34E-06	-3.345752	0.0018
WATER_TREATMENT_PLANTC	-4.65E-05	1.47E-05	-3.160485	0.0030

R-squared	0.914030	Mean dependent var	21.34726
Adjusted R-squared	0.885373	S.D. dependent var	3.980310
S.E. of regression	1.347596	Akaike info criterion	3.656094
Sum squared resid	70.82462	Schwarz criterion	4.176548
Log likelihood	-82.88648	Hannan-Quinn criter.	3.856235
F-statistic	31.89588	Durbin-Watson stat	1.297285
Prob(F-statistic)	0.000000		

Figure 12: Output of the model in Equation (7)

Hence a second model was tested in both stations with some modifications in the initial model. In the second model, a logarithmic transformation of the dependent variable was taken to overcome non-linearity. Dependent variable and the independent variables were non- stationary in their level series. To gain the normality of the residuals and to overcome the spurious regression, first difference of all non-stationary series was taken. Second type of models fitted to both stations are shown in Equation (8) and Equation (9) respectively.

Table 3: VIF values and Auto Regression Results in Station 1 and Station 2

Variable	Station 1		Station 2	
	VIF	R^2 value	VIF	R^2 value
Diesel Cost	92.05	0.99	156.81	0.99
Diesel Price	21.99	0.95	14.85	0.93
HFO Cost	12.13	0.92	14.50	0.93
HFO Price	3.86	0.74	4.54	0.78
Y_1 Lube Oil Cost	4.22	0.76	1.52	0.84
Lube Oil Price	6.57	0.85	809	0.88
Maintenance Cost	7.20	0.87	6.29	0.84
Personal Expenses	25.86	0.96	23.14	0.96
Plant Factor	7.30	0.86	9.29	0.89
Units generated from Diesel	106.01	0.89	159.96	0.99
Units generated from HFO	7.58	0.87	18.34	0.94
Water Bill	2.32	0.59	2.10	0.52
Water treatment plant chemical cost	7.13	0.86	8.29	0.88

$$d\left(log(Y_{1,t})\right) = c_0 + c_1\, d(X_{2,t}) + c_2 d(X_{4,t}) + c_3 d(X_{5,t})\; c_4 d(X_{6,t}) + c_5 d(X_{7,t}) + c_6 d(X_{8,t}) +$$

$$c_7 d(X_{9,t}) + c_8 d(X_{10,t}) + c_9 d(X_{11,t}) + c_{10} d(X_{12,t}) + c_{11} d(X_{13,t}) \quad (8)$$

$$d\left(log(Y_{2,t})\right) = c_0 + c_1 d(Z_{2,t}) + c_2 d(Z_{4,t}) + c_3 d(Z_{5,t}) + c_4 d(Z_{6,t}) + c_5 d(Z_{7,t}) + c_6 d(Z_{8,t}) + c_7 d(Z_{9,t}) + c_8 d(Z_{10,t}) + c_9 d(Z_{11,t}) + c_{10} d(Z_{12,t}) + c_{11} d(Z_{13,t}) \quad (9)$$

By considering the auto regressions, correlations of the regressors and the VIF values in Table 3, two variables in the regression models with higher values for above criterias were dropped. Among the variables Diesel cost and the units generated from Diesel, the variable 'Diesel Cost' was dropped because it might be a combination Diesel price and units generated from Diesel. In the same way, the variable 'HFO Cost' was dropped among the variables HFO cost and units generated from HFO. It was done in order to overcome the multi collinearity. Hence the results of the second models in both stations were not satisfying the model diagnostics; normality of the residuals, a third model was tested by adding some extra modifications in second model. The second difference and logarithm was gained because there were some variables which are integrated at order 2 and lag values of the dependent variable were added to the model as explanatory variables to overcome the serial correlation. Both new models give a considerable R^2 value of 0.55 in the final fitted regression models and the final time series regression models for Station 1 and Station 2 are shown in Equation (10) and Equation (11) respectively.

$$d\left(d\left(log(Y_{1,t})\right)\right) = c_0 + c_1 d\left(d\left(log(X_{2,t})\right)\right) + c_2 d\left(d\left(log(X_{4,t})\right)\right) + c_3 d\left(d\left(log(X_{5,t})\right)\right) + c_4 d\left(d\left(log(X_{6,t})\right)\right) + c_5 d\left(d\left(log(X_{7,t})\right)\right) + c_6 d\left(d\left(log(X_{8,t})\right)\right) + c_7 d\left(d\left(log(X_{9,t})\right)\right) + c_8 d\left(d\left(log(X_{10,t})\right)\right) + c_9 d\left(d\left(log(X_{11,t})\right)\right) + c_{10} d\left(d\left(log(X_{12,t})\right)\right) + c_{11} d\left(d\left(log(X_{13,t})\right)\right) + c_{12} d\left(log(Y_{1,t-1})\right) + c_{13} d\left(log(Y_{1,t-2})\right) + c_{14} d(log(Y_{1,t-3})) \quad (10)$$

$$d\left(d\left(log(Y_{2,t})\right)\right) = c_0 + c_1 d\left(d(Z_{2,t})\right) + c_2 d\left(d(Z_{4,t})\right) + c_3 d\left(d(Z_{5,t})\right) + c_4 d\left(d(Z_{6,t})\right) + c_5 d\left(d(Z_{7,t})\right) + c_6 d(d(Z_{8,t})) + c_7 d(d(Z_{9,t})) + c_8 d(d(Z_{10,t})) + c_9 d(d(Z_{11,t})) + c_{10} d(d(Z_{12,t})) + c_{11} d(d(Z_{13,t})) \quad (11)$$

Final models in both stations follow all the model assumptions; normality of the residuals, absence of multi collinearity, serial correlation and heteroscedasticity. p value associated with the Jarque-Bera test statistic is 0.088 for Station 1 and 0.2161 for Station 2 which are both greater than 0.05 and they indicate that the residuals of the fitted models are normally distributed at 5% level of significance. Further, p value of the Breusch –Godfrey Serial correlation LM test is 0.2707 for Station 1 and 0.5177 for Station 2 which are also greater than 0.05 and they give the evidence of absence of serial correlation at 5% level of significance. The most important assumption in regression which is known as homoscedasticity is also met by these two final regression models. It is also proven with the Breusch Pagan Godfrey test for heteroscedasticity with a p value 0.8389 for station 1 and 0.9995 which are greater than 0.05 significance level.

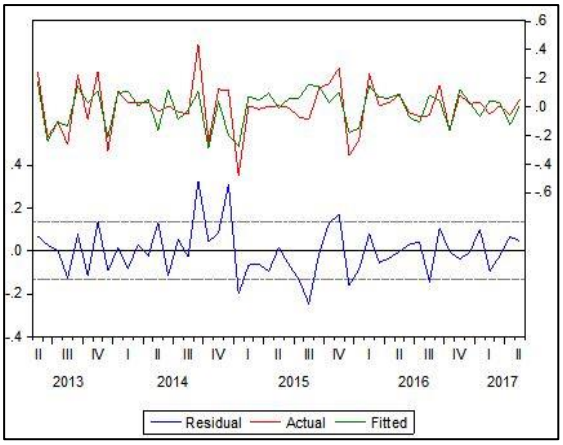


Figure 13: Actual and Fitted UC values–Station 1

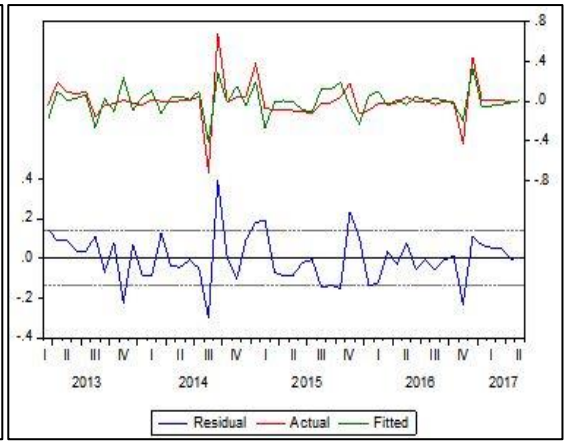


Figure 14: Actual and Fitted UC values – Station 2

Figure 13 and Figure 14 shows the plots where the actual and fitted values are drawn. Hence the actual and fitted lines have small gaps with each other, these final fitted models can be used in modelling the UC of generation of electricity in the two stations of the power station. Hence the final regression models in both power stations provide a considerable R^2 value and satisfy all model assumptions, these final fitted models can be used to forecast the UC beyond the year 2017 as shown in Figure 15 and Figure 16.

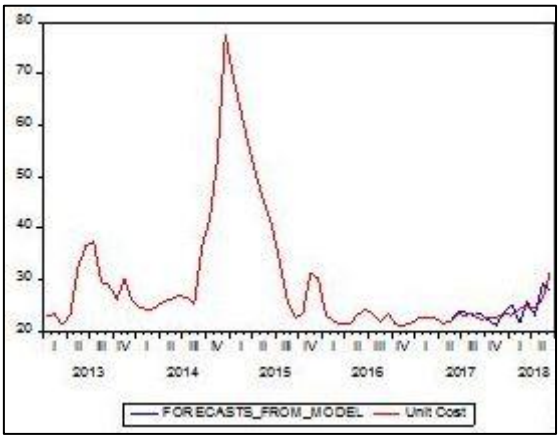


Figure 15: Model Validation–Final fitted regression model of Station 1

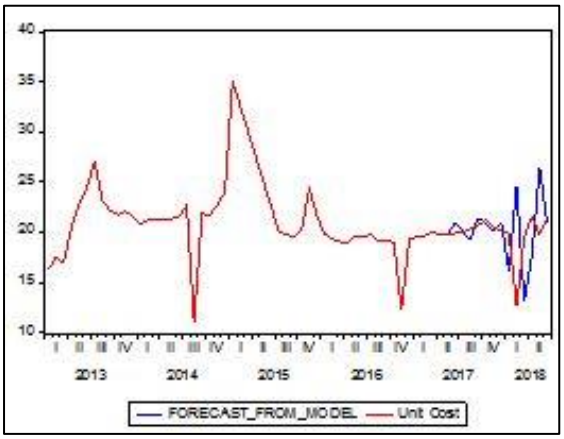


Figure 16: Model Validation–Final fitted regression model of Station 2

Forecasting accuracy of the 20 % of the test data is measured with RMSE and MAE with values 0.095196 and 0.079537 for Station 1 and 0.243976 and 0.154469 for Station 2 respectively.

3.4. Identified Significant Factors that Describe the UC of Generation of electricity of the Selected Power Plant

The significant factors identified from the time series regression models which describe the UC of generation of electricity of the two sub stations in the selected Diesel Power Station can be identified as follows.

Station 1: Number of units generated from Diesel and its lag 1 and lag 2, lag 1 and lag 2 of the UC

Station 2: Number of units generated from HFO and its lag 1 and lag 2, Lube Oil Price and its lag 1 and lag 2, Plant Factor and its lag 1 and lag 2, HFO Price and its lag 1 and lag 2, lag 1 value of the UC

The basic simplified time series regression model equations for Station 1 and Station 2 with their significant factors can be written as in Equation (12) and Equation (13) respectively.

$$log(Y_{1,t}) = 0.001824 + 0.064707\, log(X_{10,t}) - 0.129414 \log(X_{10,t-1}) + 0.064707\, log(X_{10,t-2}) + 1.437042 \log(Y_{1,t-1}) - 0.437042\, log(Y_{1,t-2}) \quad (12)$$

$$log(Y_{2,t}) = -0.006030 + 0.010421\, log(Z_{4,t}) - 0.129414\, log(Z_{4,t-1}) + 0.010421\, log(Z_{4,t-2}) - 0.001559\, log(Z_{6,t}) + 0.003118\, log(Z_{6,t-1}) - 0.001559\, log(Z_{6,t-2}) + 0.005433\, log(Z_{9,t}) - 0.010866\, log(Z_{9,t-1}) + 0.005433\, log(Z_{9,t-2}) - 0.000012\, log(Z_{11,t}) + 0.000024\, log(Z_{11,t-1}) - 0.0000121\, log(Z_{11,t-2}) + log(Y_{2,t-1}) \quad (13)$$

3.5. Model Comparisons

In order to identify the best model from ARIMA and the time series regression models, the forecasting accuracy of the both models were compared with error metrics MAE and RMSE for two stations separately. The model comparisons for Station 1 and Station 2 are shown in Table 4 and Table 5 respectively.

Table 4: Model Comparison – Station 1

Model	MAE	RMSE
ARIMA (1,1,0)	2.430926	3.403056
Time Series Regression Model in Equation 12	0.079537	0.095196

Table 5: Model Comparison – Station 2

Model	MAE	RMSE
ARIMA (2,1,2)	0.717347	0.927269
Time Series Regression Model in Equation 13	0.154469	0.243976

According to Table 4 and Table 5, it is clear that time series regression models are the best models for forecasting the UC of generation of electricity in both sub stations in the selected Diesel Power Station because they have minimum MAE and RMSE.

4. Conclusion

Forecasting the UC of generation of electricity of a fossil fuel power plant with time series modelling approach

was carried out through this study while investigating the factors that effect for the UC of generation of electricity of a fossil fuel power plant. The Study was carried out as a case study in a leading Diesel/HFO power plant in Western Province of Sri Lanka which has two sub stations. Two main approaches to solve the research problem was used in this study; univariate time series approach where only the history data of UC of generation of electricity by the selected Diesel/HFO power plant is used to determine the future movement of UC and regression approach where other explanatory variables that describes the UC are used in developing the model. After investigating time series approaches separately for Station 1 and Station 2, the time series regression model was evidently selected as the best approach in forecasting the UC of generation of electricity of the selected Diesel/HFO power plant. ARIMA (1,1,0) and ARIMA (2,1,2) which was selected as the best models under the univariate approach also do a quite good job in forecasting UC but the best results can be achieved by the time series regression models proposed for each sub-station in the selected Diesel/HFO power plant. Two dynamic regression models were separately identified for Station 1 and Station 2 with R2 value 0.55 concluding that 55% of the variation of the UC of generation of electricity in the selected Diesel/HFO power plant is explained by the identified independent variables. Hence identified regression 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. Among the significant factors identified, past UC values were a significant factor in both sub stations. Number of units generated from Diesel has become significant in the Station 1 only the number of units generated from HFO is significant in Station 2. Plant factor, lube oil price and HFO price are some more significant factors identified for Station 2. HFO is used for prime mover engine while Diesel oil is used for engine start up and shut down. The Lube Oil system in Station 1 and Station 2 differ in their process of cooling the lubricants. Hence mainly it can be concluded that there is an effect from the process of cooling the lubricants and engine start up and shut down process for the UC of generation of electricity of the selected Diesel/HFO power plant. This study is useful for the power station as well as Ceylon Electricity Board (CEB). CEB usually dispatch from the stations with lower costs. If a power station can forecast the UC with a model like this, the results of the research will go a long way to help in selecting the suitable power stations for the dispatch in time periods properly without wasting our energy resources. As well as, in a time of unit cost fluctuation, CEB can get an idea to use the free energy sources like wind, solar and hydro since renewable energy resources do not require purchased fuel and the operating costs and generation cost over time are highly predictable, as opposed to fossil fuel based generation costs. Further, the identified significant factors can be given more consideration in electricity generation process in order to reduce the generation cost of electricity. There might be some other factors which are affecting to the generation of electricity in a Fossil Power Station than the factors that were used in this study. Those factors can be also considered in modelling further and the researchers who are interested in this field can do the studies without limiting this to a one station. This forecasting method can be generalized to other fossil fuel power plants with necessary alterations.

Acknowledgements

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

[1]. C. Yozgatligil, S. Aslan, C. Iyigun and I. Batmaz, "Comparison of missing value imputation methods in time series: The case of Turkish meteorological data." Theoretical and Applied Climatology, 112(1–2), pp. 143–167,2013. https://doi.org/10.1007/s00704-012-0723-x

[2]. Fen-Ru Shih, Mainak Mazumdar and Jeremy A. Bloom, "Asymptotic mean and variance of electric power generation system production costs via recursive computation of the fundamental matrix of a markov chain." Operations Research,47, pp. 703-712,1999.

[3]. Generation Performance in Sri Lanka (2011).

[4]. Generation Performance in Sri Lanka (2012).

[5]. Generation Performance in Sri Lanka (2013).

[6]. Generation Performance in Sri Lanka (2014).

[7]. Generation Performance in Sri Lanka (2016).

[8]. Generation Performance in Sri Lanka (2017).

[9]. M.Davison, C.L. Anderson, B. Marcus and K. Anderson, "Development of a hybrid model for electrical power spot prices." IEEE Transactions on Power Systems, 17(2), pp. 257–264. https://doi.org/10.1109/TPWRS.2002.1007890

[10]. S.Moritz and T. Bartz-Beielstein, "imputeTS: Time Series Missing Value Imputation version in T." The R Journal, pp. 1–12,2017.

[11]. T.A. Moahmed, N.E. Gayar and A.F. Atiya, "Forward and backward forecasting ensembles for the estimation of time series missing data." IAPR Workshop on Artificial Neural Networks in Pattern Recognition, 8774, pp.93–104, 2014. https://doi.org/10.1007/978-3-642-12159-3

[12]. Y. Matsuo, Y. Nagatomi and T. Murakami, "Thermal and Nuclear Power Generation Cost Estimates Using Corporate Financial Statements", pp. 1–20,2011