Regression Based Forecast of Electricity Demand of New Delhi

Aayush Goel*, Agam Goel**

*Electrical (Power) Engg., IIT Delhi
**Electrical (Power) Engg., IIT Delhi

Abstract- The forecast of electricity demand in India is of considerable interest since the electricity sector has been the prime focus of past as well as present Governments. This study presents three models of prediction of electricity demand of New Delhi, namely Multiple Regression, Trend Seasonality Model and ARIMA modelling. The significance of climatic and seasonal factors on electricity demand as well as comparison of the relative accuracy of the models have also been discussed

Index Terms- ARIMA, Electricity Forecast, Multiple Regression, Trend Seasonality

I. INTRODUCTION

Forecasting of electricity demand has an important function in short term load allocation and distribution as well as long term planning for future electricity generation facilities. Governments worldwide use energy demand forecasting as one of the most important policy tools. Accurate load forecasting can lead to better budget planning, overall reduction in cost and maintenance scheduling and fuel management. Without energy forecast, shortage in power or overcapacity may lead to large losses.

Demand of electricity is subject to a range of uncertainties such as weather conditions (temperature, humidity, precipitation etc.), population growth, technology, economy and irregularity in individual usage. It is also affected by some known calendar effects such as the time, day, year and holidays.

Load forecasting has been a subject of considerable research in the past. Ching-Lai Hor et al. [1] have used multiple regression as a means of investigating the effect of climatic factors on the electricity demand of England and Wales. Exponential smoothing has been investigated by Barakat et al. [2] and Infield and Hill [3] to forecast demand based on past trend in Saudi Arabia and Shetland, Scotland respectively. Adaptive load forecasting algorithms such as the Kalman Filter based regression used by Ojeda et al. [4] have also garnered interest in the past.

Apart from these, stochastic time series such as Autoregressive Integrated Moving Average Models (ARIMA) have also been efficient in predicting dependent variable, particularly when the independent variable data is missing or erroneous. Li Wei and Zhang zhen-gang [5] have successfully demonstrated this.

An accurate and robust demand forecast requires accurate and adequate data without which the results of the forecast are a bit unreliable. However, India does not have a robust system of collecting, maintaining and reporting data related to electricity usage in the public domain. It is hoped that with increasing stress on accurate electricity forecast, efficient data management systems will be devised to aid better quality forecasts.

The main contribution of this study is the development of robust statistical models to forecast electricity demand incorporating meteorological factors to improve the accuracy and bring out the relative contribution of the influence of these meteorological factors. Multiple regression, the Trend-Seasonality model and ARIMA models have been used to predict the electricity demand of New Delhi, India. In each case, the root mean square error has been used as a measure of the accuracy of the prediction.

The structure of this paper is as follows: Section II describes the data sources from which data for demand as well as the independent variables have been collected. Section III analyses the nature of electricity demand data of New Delhi. In section IV, we develop the statistical model to forecast electricity demand in New Delhi and relate electricity demand to climatic factors using multiple regressions. Section V gives a basic overview of the application of the Trend-Seasonality model on electricity demand. ARIMA modelling of the data and least error ARIMA formula are presented in Section VI. Finally, in Section VII, the main findings of the study are summarized and conclusions are drawn.

II. DATA SOURCES

To achieve an accurate electricity demand forecast, accurate data was required on electricity demand, climate and demography of Delhi and India. The required data was collected from the following sources.

1) Load generation balance report 2005-2013 by Central Electricity Authority, India for electricity demand data [6]
2) Tutiempo for Climate Data [7]
3) World Bank, India Data, 1961-2012 [8]

III. ANALYSIS OF ELECTRICITY DEMAND DATA

In this section the main characteristics of electricity demand data have been analyzed using standard electricity demand-time plots. The time frame can be a day, month or years depending upon the characteristic to be analyzed. Electricity demand data can be

considered to be a time-series data and all the mathematics for time-series data is applicable to electricity demand data.

A. Non-Stationary

Electricity demand data cannot be considered to be stationary because the mean of electricity demand data doesn't remain constant throughout the year.

B. Seasonality

There is seasonality in electricity demand data. Intuitively, the time-series has been divided into four seasons in the literature which are Winter, Summer, Pre-Monsoon and Post-Monsoon.

Table 1: Average monthly electricity demand data of Delhi from 2005-2012

Month	Jan	Feb	Mar	Apr	May	Jun
Average	1764	1484	1620	1945	2351	2443

Table 2: Average monthly electricity demand data of Delhi from 2005-2012

Month	Jul	Aug	Sep	Oct	Nov	Dec
Average	2536	2418	2231	1909	1552	1634

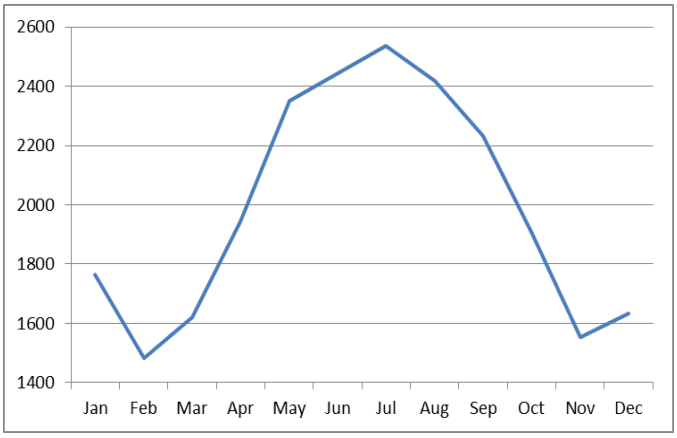

Figure 1: Plot of Average monthly electricity demand data of Delhi from 2005-2012

C. Trend

If the mean electricity demand data each year is analyzed over a period of many years, a tendency of the demand to increase is observed. Thus it can be said that there is a long-term trend in electricity demand data.

Year	2005	2006	2007	2008
Mean	1800.6	1860.8	1859.7	1896.6

Table 3: Average yearly electricity demand data of Delhi

Year	2009	2010	2011	2012
Mean	1983.4	2120.5	2218.7	2189.8

Table 4: Average yearly electricity demand data of Delhi

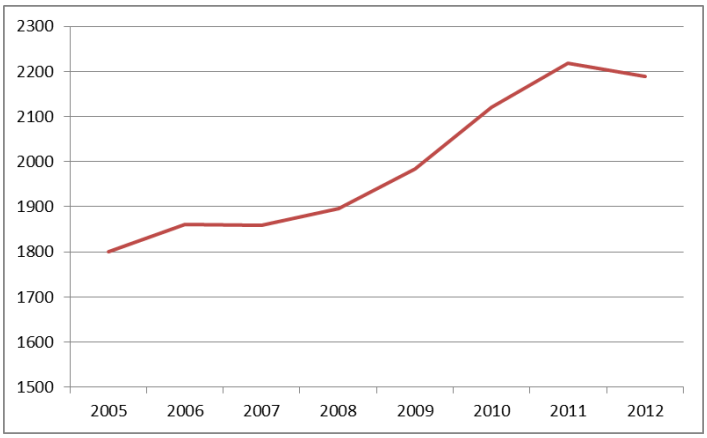

Figure 2: Plot of average yearly electricity demand data of Delhi

IV. MULTIPLE REGRESSION

In this section, we will forecast the electricity demand of New Delhi on a monthly basis through multiple regression. In multiple regressions, a dependent variable is predicted by two or more independent variables. Literature in the area indicates that temperature, humidity and precipitation are the major meteorological factors which tend to influence the electricity demand in a region. We will use past electricity demand and these factors to forecast electricity demand. The data was collected from January 2005 to March 2013.

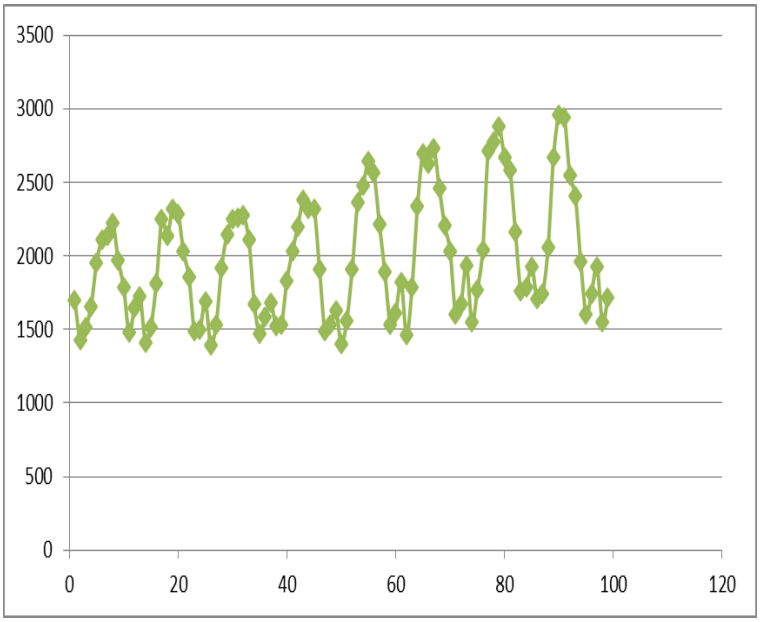

Figure 3: Electricity Demand in Delhi

Clearly, as we can see from Figure 1,2 and 3, the data has a time trend, seasonality and irregularity.

Temperature, humidity and precipitation are taken as independent variables and electricity demand is the dependent variable. Since, the data has a time trend; time will also be taken as a dependent variable. Also, the data collected is monthly data and hence 11 dummy variables will be taken to account for monthly variation in electricity demand. The final model is

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 \ldots\ldots + B_{14}X_{14} + B_{15}X_{15}$$

Where

Y = Monthly electricity demand in Delhi (MU)
X_1 = Temperature (ºC)
X_2 = Mean humidity percentage
X_3 = Precipitation (mm)
X_4 = Time trend
X_5 to X_{15} = Dummy variables for January to November

The B's are regression coefficients. A regression coefficient in multiple regressions is the slope of the linear relationship between the dependent variable and the part of a dependent variable that is independent of all other independent variables

Regression performed in Microsoft Excel gives the coefficients in as specified in Table 5.

	Coefficients	T Statistics
B_0	1016.70	3.39
B_1	37.66	2.93
B_2	-3.13	-1.24
B_3	-0.27	-1.01
B_4	5.22	12.60
B_5	233.48	4.10
B_6	-232.30	-3.50
B_7	-364.37	-3.03
B_8	-305.18	-1.65
B_9	0.31	0.00
B_{10}	101.22	0.42
B_{11}	368.94	1.57
B_{12}	303.31	1.30
B_{13}	143.70	0.69
B_{14}	-147.10	-1.02
B_{15}	-281.15	-3.23

Table 5: Regression coefficients – Multiple regression model

We need to analyze the goodness of fit of the model. The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. The regression has an adjusted goodness of fit of 0.926 indicating an extremely good fit model.

The T Statistics indicates the confidence level with which we can say that a particular dependent variable indeed predicts the independent variable and that the corresponding regression coefficient is non zero. We can say that temperature regression coefficient is non zero with confidence level 99.5%, mean humidity percentage and precipitation regression coefficients are non-zero with confidence level 70%. This indicates strong correlation between electricity demand and climatic variables such as temperature, mean humidity percentage and precipitation. However we can neglect mean humidity and precipitation if their data is not available to us as 70% is not deemed to be a very good confidence interval.

We plot the error term i.e. the difference between Actual electricity demand and electricity demand predicted according to the model. The Root Mean Square Error obtained is 102.54. The plot of the error terms is shown in Figure 4

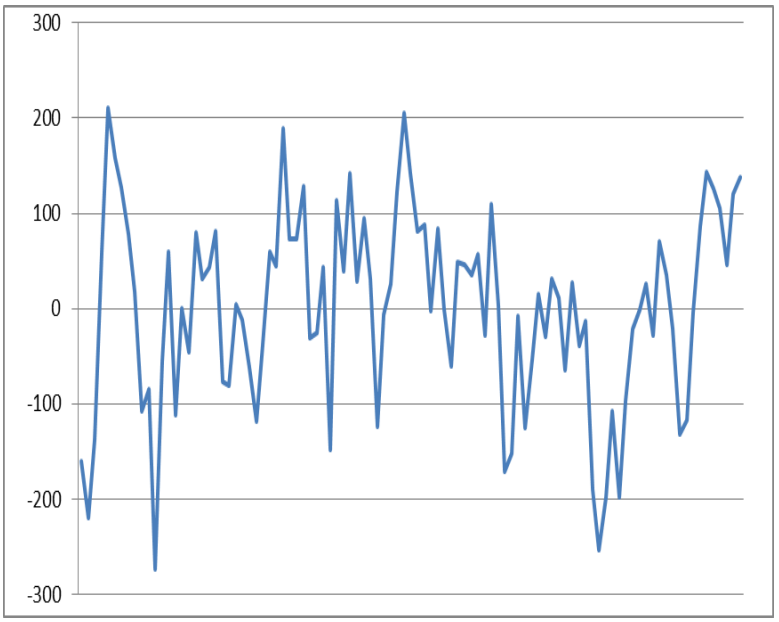

Figure 4: Errors – Multiple regression model

In order to improve the fit of the model and reduce error, we use another model where the dependent variable is logarithm of electricity demand. Literature suggests that due to the nature of electricity demand data, taking logarithm of electricity demand will lead to a better statistical result.

The coefficients from the multiple regressions for the logarithmic model are specified in Table 6.

	Coefficients	T Statistics
B_0	7.12	54.90
B_1	0.01	2.64
B_2	0.00	-0.96
B_3	0.00	-1.18
B_4	0.00	13.78
B_5	0.12	4.94
B_6	-0.12	-4.30
B_7	-0.14	-2.68
B_8	-0.06	-0.75
B_9	0.09	0.90
B_{10}	0.13	1.24

B_{11}	0.23	2.30
B_{12}	0.21	2.11
B_{13}	0.14	1.54
B_{14}	-0.01	-0.12
B_{15}	-0.13	-3.41

Table 6: Regression coefficients – Multiple regression logarithmic model

The regression has an adjusted goodness of fit of 0.943 indicating an improved model as compared to the previous case. We can say that temperature regression coefficient is non zero with confidence level 99%, mean humidity percentage regression coefficient is non zero with 65% confidence level and precipitation regression coefficients is non-zero with confidence level 70%. This indicates strong correlation between electricity demand and climatic variables such as temperature, mean humidity percentage and precipitation.

We plot the error term i.e. the difference between Actual electricity demand and electricity demand predicted according to the model. The plot of the error terms from the new model is shown in Figure 5. The errors have clearly reduced as compared to the previous case. The new Root Mean Square Error is 87.36.

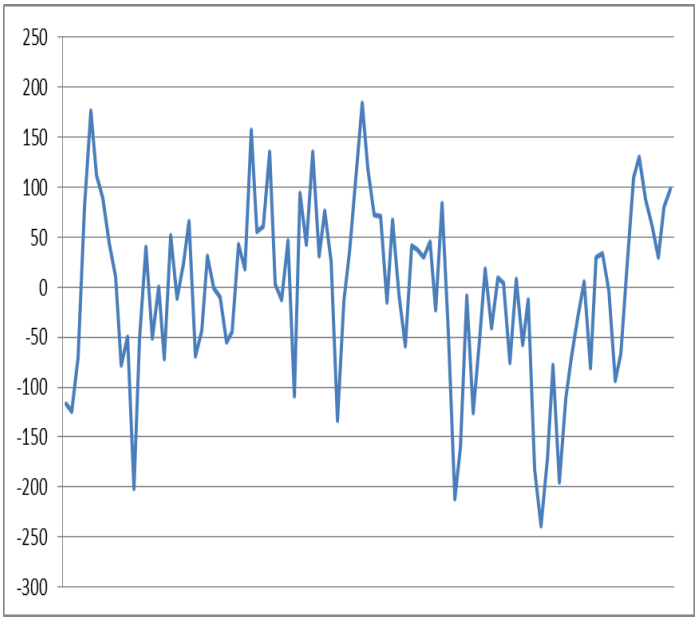

Figure 5: Errors – Multiple regression logarithmic model

V. TREND-SEASONALITY MODEL

The advantages of predicting a dependent variable on regression of itself has distinct advantages as compared to regressing on multiple independent variables.

Regressing on auto regressions implies that all of the factors that affect the dependent variable y_t are included in auto regressions of y_t i.e. y_{t-1}, y_{t-2} etc. according to Wooldridge. Therefore, one doesn't have need of all the independent factors, and this has a lot of advantages on the data side, particularly with respect to its collection and storage, as we only need to collect and store the dependent variable and we can lag it according to our need. This is particularly apt for India, as here data for independent variables is either unavailable for public or extremely costly.

The trend seasonality model assumes that the dependent variable can be modeled as the product of a Trend T (which varies linearly with time), Cyclicity C (which is periodic on a short duration), Seasonality S (which is periodic on a long duration) and an Irregular component I (error). Electricity is a variable which does not vary on a short duration; therefore we can generally remove C without affecting the model to a great extent.

The methodology adopted for this method is as follows -

1) Assume trend to be constant for the period of the seasonality. In our case, the period of seasonality is 12. Calculate a moving average over y_t to y_{t-12}. This is the fixed trend component for that value.
$$MA(12) = {\sum_{i=1}^{12} y_{t-i}}/{12}$$
2) Divide the values by the Moving Average for that value. This means that we are left with the seasonal and the irregular component.
3) To get seasonal component, consider each month separately. Find all seasonal values for that particular month and average to remove irregular component. We get a fixed seasonal component S1 for a month. Repeat for every month.
4) Divide each value by its seasonal component to get the trend component.
5) Since trend is linear, hence we can easily regress this on time to get a linear relation w.r.t. time.
6) Forecast the trend from the linear relationship and multiply by the average seasonal value for the particular month to get prediction for the particular time.

Employing the method for New Delhi electricity demand, a plot showing the seasonal and trend components of the New Delhi and solely the linear trend component after deseasonalizing can be obtained as seen in Fig 6.

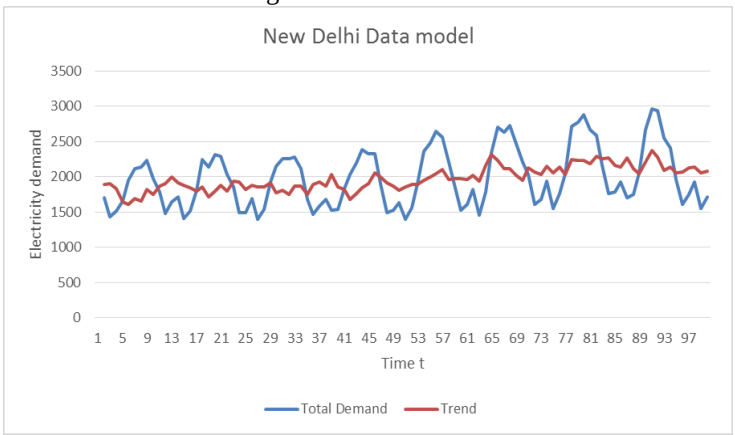

Figure 6: Seasonal and trend components of Demand

A plot of the error terms for this model is shown in Fig 7. The Root Mean Square Error obtained is 104.

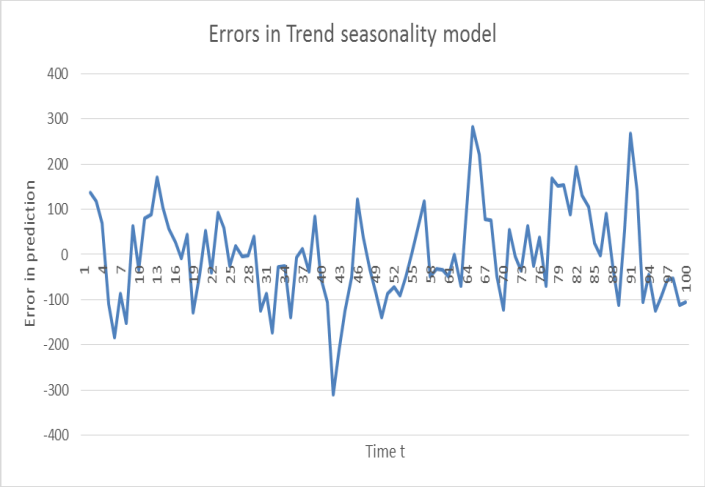

Figure 7: Errors – Trend Seasonality Model

Finally, a, plot of the actual and forecasted load demand is shown in Fig 8

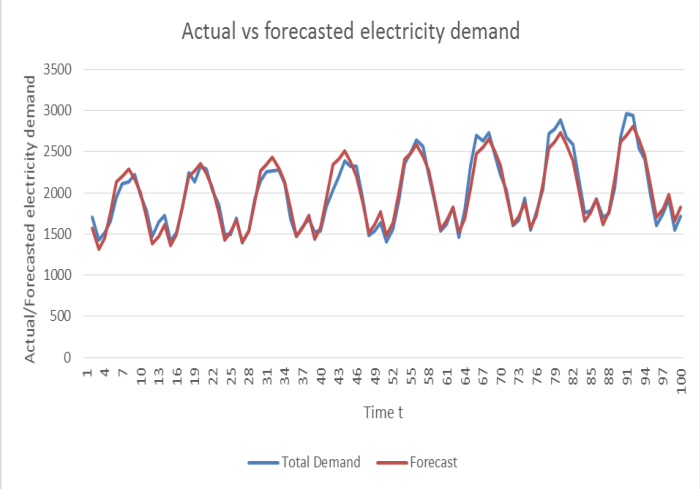

Figure 8: Actual vs Forecasted Demand – Trend Seasonality Model

VI. ARIMA MODELLING

The ARIMA model uses the fact that electricity demand is a stochastic time series.

This modelling regresses the dependent variable y_t on – p lags of the dependent variable (Autoregressive) and q lags of the error term (Moving Average).Sometimes instead of dependent variable y_t, $L^d y_t$ can be used as the dependent variable. Here L is the one step lag operator i.e.

$$Ly_t = y_{t-1}$$

The general equation of ARIMA model is as follows -.

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)(1-L)^d X_t = \left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t.$$

Where ε_t is white noise error.

A. *Dickey-Fuller Test*

The ARMA model (AR and MA) can only be applied is the dependent variable has no trend i.e. it is stationary w.r.t. time. Since our data has a trend variable it has to be differenced to remove the trend. This is the differencing which appears in the ARIMA model I and the order of differencing is decided by the Dickey-Fuller tests. Dickey-Fuller test identifies the unit root in the equation. If the equation has a unit root, the variable is non-stationary. [9]

The significance results when regression of lagged values of y_t are performed to determine the presence of a unit root are tabulated in Table 7.

Dependent variable	Coefficient of Var_{t-1}	t-stat for Var_{t-1}	P-Value for Var_{t-1}
Ly_t (Simple)	-0.008	-0.6364	0.526
Ly_t (With drift)	-0.203	-3.2954	0.0013
Ly_t (With time trend)	-0.218	-3.3459	0.0011
L^2y_t	0.663	6.8410	7.61E-10

Table 7: Dickey-Fuller Test Significance Results

As can be seen the t-statistic for the simple case is very low and therefore in this case the null hypothesis is rejected. Hence the simple variable cannot be used in the ARIMA model.

We can use the 'With drift' and 'With time trend' variable, however their t-stats are just on the verge of the 95% confidence interval level. Hence it would be better to take Ly_t as the dependent variable as its t-stat has the largest possible value amongst all.

Hence the order of differencing d = 1 for the ARIMA model as specified by the Dickey-Fuller tests.

A. *Box-Jenkins Method*

The Box Jenkins method is a tool that can be used to predict the degree of p and q for an ARIMA model by comparing the Autocorrelation function (ACF) and the Partial Autocorrelation function (PACF) of the data. [10]

The autocorrelation function at lag k is the correlation of the data with k lags of itself. In our case since the data is essentially Ly_t, hence the ACF(k) for our model is –

$$ACF(k) = {Cov(Ly_t, Ly_{t-k})}\Big/{\sqrt{Var(Ly_t)}\sqrt{Var(Ly_{t-k})}}$$

The Partial Autocorrelation function at lag k is the correlation of data with k lags of itself discounting for all effects of the k-1

lags. For Ly_t, the PACF(k) can be found out be regressing Ly_t on k lags of itself, and finding the coefficient of Ly_{t-k}.

Box and Jenkins have specified a wide variety of the characteristic ACF and PACF should obey in order to be an AR(p), MA(q), ARMA(p,q) or even a seasonal ARIMA model.

Therefore, the ACF and PACF of the data should be analyzed first which are shown in Fig 9 and 10 respectively. All of the plots have been conceived using Microsoft Excel.

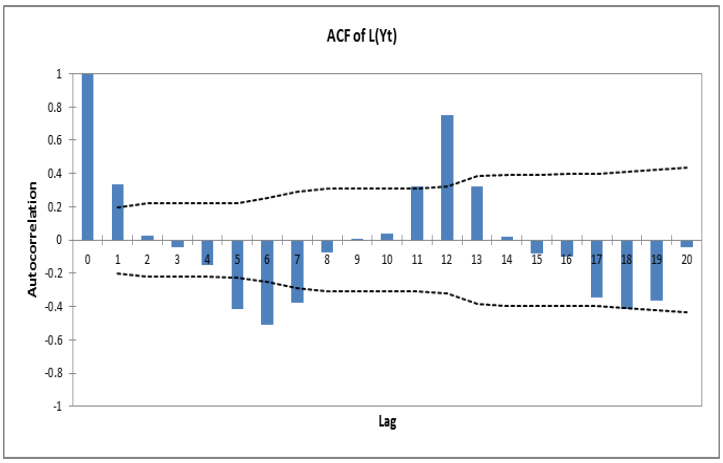

Figure 9: ACF of Ly_t

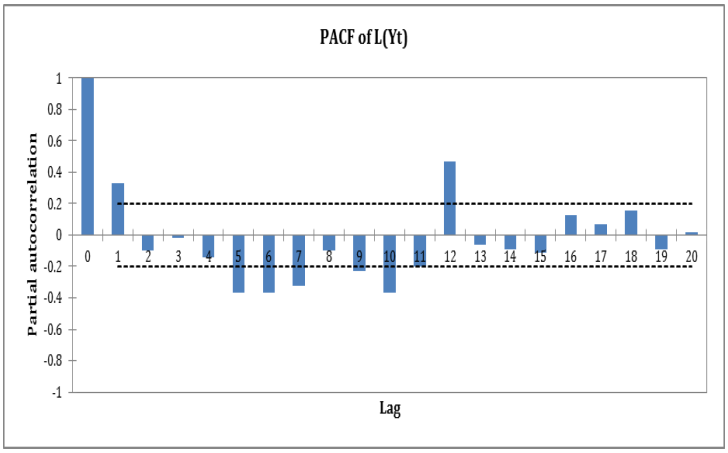

Figure 10: PACF of Ly_t

There is a certain periodicity with period 12 in the ACF plot and spikes are significant at 0 and 12, therefore according to Box Jenkins a seasonal ARIMA plot with period 12 must be considered.

Hence $Ly_t$is differenced by 12, i.e. we consider $Ly_t - Ly_{t-12}$ as the new variable to which ACF and PACF are applied. The plots for the new ACF and PACF are shown in Fig 11 and Fig 12 respectively.

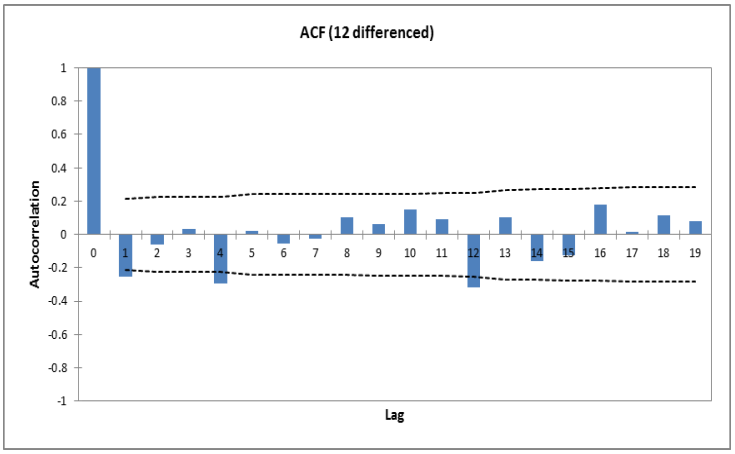

Figure 11: ACF of Ly_t differenced by 12

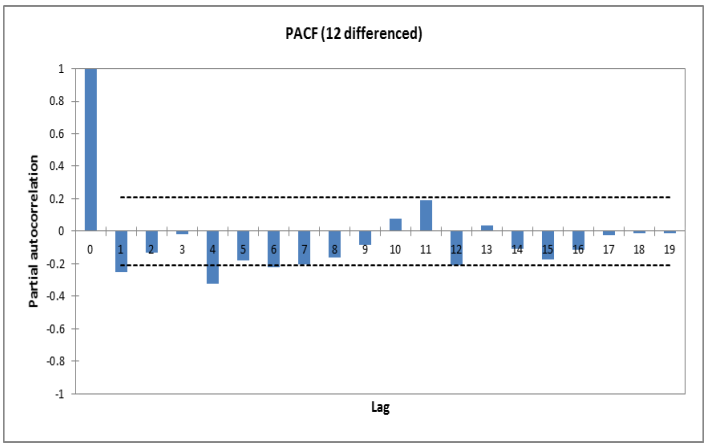

Figure 12: PACF of Ly_t differenced by 12

We observe that from the ACF, significant spikes are located at lag 1, 4 and 12 indicating MA(1), MA(4) and MA(12) terms and there are significant spikes at lag 1, 4 and 12 in the PACF indicating AR(1), AR(4) and AR(12) terms. We will not consider > AR(4) spikes in this as the spikes start to die out after lag 4, indicating that they are not to be considered.

Hence our final ARIMA model becomes –

$$\mathbf{ARIMA(0,1,0)(1,0,1)^{12}(4,0,4)^{12}(12,0,12)^{12}}$$

(The formula has been written in typical seasonal ARIMA representation.)

The Root Mean Square Error from this model is 99.61 which is less than the RMSE calculated from the Trend Seasonality Model.

VII. CONCLUSION

The Root Mean Square Error, used as a measure of accuracy, is tabulated for all the models presented.

Method	RMSE
Multiple Regression	102.54

Logarithmic Multiple Regression	87.36
Trend-Seasonality Model	104
ARIMA model	99.61

Table 7: Root Mean Square Error for various models

It is evident that Logarithmic Multiple Regression would be the best model to apply in the particular case of predicting Electricity Demand for New Delhi as it gives the lease RMSE as well as provides the significance of climatic factors on Electricity demand.

However, if data for independent variables is absent, the ARIMA models should be used as it gives a lesser RMSE than the Trend-Seasonality Model and also provides a definite formula which can be put into statistical software or Microsoft Excel which has been used to demonstrate all the methods presented here.

ACKNOWLEDGMENT

The authors would like to thank Prof. Abhijeet Abhyankar of the Electrical (Power) Engineering Dept., IIT Delhi for providing the authors the opportunity to pursue research in this area as well as for his valuable guidance.

REFERENCES

[1] Hor, Ching-Lai, Simon J. Watson, and Shanti Majithia. "Analyzing the impact of weather variables on monthly electricity demand." *Power Systems, IEEE Transactions on* 20.4 (2005): 2078-2085

[2] Barakat, E.H.; Al-Qassim, J. M.; Al-Rashed, S.A., "New model for peak demand forecasting applied to highly complex load characteristics of a fast developing area," *Generation, Transmission and Distribution, IEE Proceedings C* , vol.139, no.2, pp.136,140, Mar 1992

[3] Infield, D.G.; Hill, D.C., "Optimal smoothing for trend removal in short term electricity demand forecasting," *Power Systems, IEEE Transactions on* , vol.13, no.3, pp.1115,1120, Aug 1998

[4] Ojeda, L.L.; Kibangou, A.Y.; de Wit, C.C., "Adaptive Kalman filtering for multi-step ahead traffic flow prediction," *American Control Conference (ACC), 2013* , vol., no., pp.4724,4729, 17-19 June 2013

[5] Li Wei; Zhang zhen-gang, "Based on Time Sequence of ARIMA Model in the Application of Short-Term Electricity Load Forecasting," *Research Challenges in Computer Science, 2009. ICRCCS '09. International Conference on* , vol., no., pp.11,14, 28-29 Dec. 2009

[6] Load generation Balance Report, Central Electricity Authority, Government of India

[7] Climate Data, New Delhi http://www.tutiempo.net/en/Climate

[8] India Data, World Bank http://data.worldbank.org/country/india

[9] Halim, S.; Bisono, I. N.; Melissa; Thia, C., "Automatic seasonal auto regressive moving average models and unit root test detection," *Industrial Engineering and Engineering Management, 2007 IEEE International Conference on* , vol., no., pp.1129,1133, 2-4 Dec. 2007

[10] Paretkar, P.S.; Mili, L.; Centeno, V.; Kaiyan Jin; Miller, C., "Short-term forecasting of power flows over major transmission interties: Using Box and Jenkins ARIMA methodology," *Power and Energy Society General Meeting, 2010 IEEE* , vol., no., pp.1,8, 25-29 July 2010

AUTHORS

First Author – Aayush Goel, B.Tech. in Electrical (Power) Engineering, IIT Delhi, axgoel8@gmail.com

Second Author – Agam Goel, B.Tech. in Electrical (Power) Engineering, IIT Delhi, agamgoel18@gmail.com

.