Advances in Aeronautical Science and Engineering
ISSN: 1674-8190

Evaluating Statistical Techniques for Accurate Electricity Load Forecasting

Emma L. Reed, Lucas M. Grant & Oliver T. King
Department of Computer Science Engineering, University of Glasgow, Glasgow, Scotland

ABSTRACT

Load forecasting is the estimation of electrical power required to meet the long, medium or short term demand by the construction of models based on relative information, such as historic load data, climate etc. It helps the electric power generation and distribution systems to plan and optimize their operations. It includes energy purchase and generation, infrastructure development, load switching, and contract evaluation. Accurately forecasting the load enables the utility companies to gain profit. Different techniques have been developed for forecasting load, which is categorized as traditional or modern techniques. Time series and multiple regression models are the examples of traditional techniques, former considered historical load data and later take the relationship between the load and the factors affecting the load. These statistical methods are mathematically proven and are linear ones. Modern techniques include fuzzy logic, ANN, SVM etc are used to characterize the non- linear relationship between the load and various exogenous factors. This paper presents a study regarding ARIMA and regression for electricity load forecasting.

KEYWORDS: Electricity load forecasting, ARIMA, Regression

1. INTRODUCTION

The electricity demand in India is increasing day by day due to the growth of population and industrial development. The installed generation capacity is increased from 154.7 GW in 2007 to 346.62 GW in 2018 [1]. It makes India as the third largest electricity generating country in the world. In spite of this remarkable progress, India still faces some basic problems like unavailability of electricity in rural areas, transmission and distribution losses, load shedding. Hence India's grid needs support from modern technologies. The traditional grid is now advance to become smarter, by the use of digital technologies for automation and communication.Thesetechnologiesallowreal-timemonitoringofelectricityfromgeneration toconsumptionandanalyzinghistoricloaddatatopredictfutureenergyneeds.

Electricity load forecasting has now become an important research area owing to its increasing importance in the modern world. It is the prediction of upcoming demands of the load in utility companies (a company which supplies electricity) by examining historic loaddata,whichisdividedintothreecategoriessuchasshortterm,medium-term,andlong term load forecasting. The forecasts for different time horizons are important for operations within a utility company. The grid controllers predict the next minute load, grid operators predict the next day load and the grid designers forecast the load for next year.

Purposes of Load Forecasting

1. Proper planning and operation of power system.
2. Proper planning and operation of power system.
3. For providing thecapital.
4. Proper planning of distribution and transmission of electricity.
5. Helps electricity sales.
6. Helps in a grid formation
7. Helps manpower development.[2].

Categories of Load Forecasts

Type of load forecasts	Functions

1. Short term loadforecasting	• Predictingtheloadfromonehourtooneweek. • Estimateloadflowsandtomakedecisionsforpreventingoverloading. • It affects unitcommitment • Spinning reservescheduling • It improves networkreliability. • Itreducesequipmentfailures,blackoutandenergywastage.
2. Medium -term loadforecasting	• Itisfromoneweektooneyear. • Itisusedforoutageandgenerationandtransmissionmaintenanceplanning. • It helps budget assignment and fuelscheduling. • It helps load switchingoperation.
3. Long term loadforecasting	• Forecasting for more than ayear. • It helps to the schedule of the generatingunits. • Planningthefutureexpansionofthegeneratingcapacity. • Fordeterminingtheeconomicallocation,type,andsizeofthefuturepowerplants.

Inaccurate load forecasting is a very series issue which increases the cost of operation of a power system. The over estimation of the power demand can induce unbalancing in the network due to excess current flow. In a similar fashion, underestimation of the load cause insufficient power supply to meets the demand. This in turn will result in additional cost duetoproductionorpurchaseofextrapower.

Factors Influencing Load

There are several factors that affect the load consumption pattern. Some of them are:

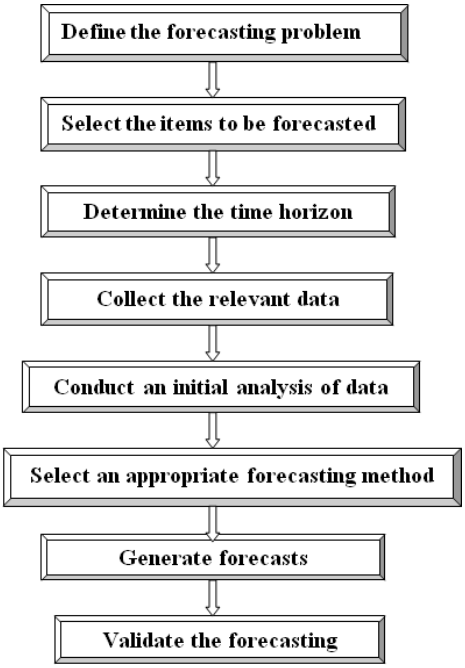
1. Weather: Temperate, humidity, cloud coverage, wind speed, rainetc.
2. Time and seasons: Length of day-light hours, sunrise and sunset time, alteration of seasons,

weekend, weekday, daytime, nighttime, holidays consumption pattern are different.

3. Economy: GDP rate, population, industrialization, per-capita income, type of customers like the commercial, industrial,agricultural.
4. Price of electricity: Time Of Use (TOU), Critical Peak Pricing (CPP), Real Time Pricing(RTP).
5. Random disturbance: Hartals, strikes, special functions, marriage functions, sudden industry startup or closingdown.
6. Other factors: Rural and urban area consumption is different, sales pattern of electronic equipment's, Television programs such as serials, sports etc affect theconsumption of electricity[3].

Steps For Forecasting

There are 8 basic steps for forecasting:



2 STATISTICAL LOAD FORECASTING MODELS

The series of data collected over a particular time interval is called as time series data. Time series analysis is a statistical technique which consists of some methods for analyzing time series data for extracting meaningful statistics and identifies the intrinsic structures in the time series. It is used to predict the upcoming values based on the understanding of the past observed data[4].

Types of Data:

There are three kinds of data:

1. Time series data: A variable takes a set of observations with fixed time intervals. They are usually collected at fixed intervals, such as daily, weekly, monthly, and annually, quarterly,etc.
2. Cross-sectional data: One or more variable's data collected at the same point in time.
3. Pooled data: A fusion of time series and cross-sectional data[5].

Terms and Concept used in Time series analysis:

- Dependence: It the association of two observations of a variable at preceding time points.
- Stationarity: Time series with statistical properties such as mean, variance, standard deviation, autocorrelation, etc which are constant over time is called a stationary time series. Some statistical prediction methods are approximately stationarized the time series by mathematical transformations

which are relatively easy to predict. That is its statistical properties will be the same in the future as in the past.

- Differencing: It is used to make the series stationary, to de-trend, and to control the auto-correlations. Some time series analyses do not require differencing and over- differenced series can produce in accurate estimates.
- Specification: Testing the linearity and non-linearity relationship of dependent vari- ables by use of many models such as ARIMA, regressionetc[6].

There are various time series forecasting methodologies are used for electricity load forecasting. Most important methods are discussed below.

3 ARIMA

Most popular statistical method used for forecasting is the ARIMA model. ARIMA stands for Autoregressive Integrated Moving Average. It combines AR model and MA model.

1. Autoregression model(AR)

The model that uses data from the same input variable at the previous time step is referred to as an

autoregression (regression to itself). *AR*(*p*) is an autoregressive model with *p* lags:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$$

whereμisac onstant andγ_p is the coefficient for the lagged variable in time t-1. AR(1) is expressedas:

$$y_t = \mu + \gamma y_{t-1} + \epsilon_t$$

2. Moving average model(MA)

The moving-average model describes the output variable relies on linearly on the present and various past values of a stochastic term (residuals from previous periods). *MA* (*q*) is a moving average model with *q*lags:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where $\theta_q$is the coefficient for the lagged error term int-q . *MA* (1) model is expressedas:

$$y_t = \mu + \epsilon_t + \theta \epsilon_{t-1}$$

3. Autoregressive Moving Average model(ARMA)

Autoregressive moving average (ARMA) models combine both *p* autoregressive terms and q moving average terms, also called *ARMA*(*p, q*).

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

4. Autoregressive Integrated Moving Average Model ARIMA model

ARMA model only used for stationary time series. Most of the real world problems have non-stationary behavior, which is modeled by the use of ARIMA model. It is described as *ARIMA*(*p, d, q*). This represents the order of the autoregressive components (*p*), the number of differencing operators (*d*), and the highest order of the moving average term (*q*). Make a stationary time series from a non-stationary series performing differencing on data pointsuntilit's mean and variance remains stationary[7].

When a variable $y_t$is not stationary, Differenced variable:

$\delta y_t = y_t - y_{t-1}$ for first order differences.

An *ARIMA*(*p*, 0, 0) is the *AR*(*p*) model and *ARIMA*(0, 0, *q*) is the *MA*(*q*) model. *ARIMA*(2, 1, 1) means

that, a second-order autoregressive model with a first-order moving average component whose series has been differenced once to introduce stationarity.

The full model can be in the form:

$$y_t' = \mu + \gamma_1 y_{t-1}' + \ldots\ldots + \gamma_p y_{t-p}' + \theta_1 \epsilon_{t-1} + \ldots\ldots \theta_q \epsilon_{t-q} + \epsilon_t$$

Steps of ARIMA modeling

- Plot the time series data.
- Logarithmic transformation to make data stationary on variance.
- Differencing on data to make data stationary on mean (remove trends).
- Plot ACF and PACF for identifying potential AR terms and MA terms.
- Identification of best fit ARIMA model.
- Forecasting by use of best fit model.
- Plot ACF and PACF for residuals of ARIMA model to ensure no more information is present.

BasicstepsofARIMAmodelingasshowninFigure2.

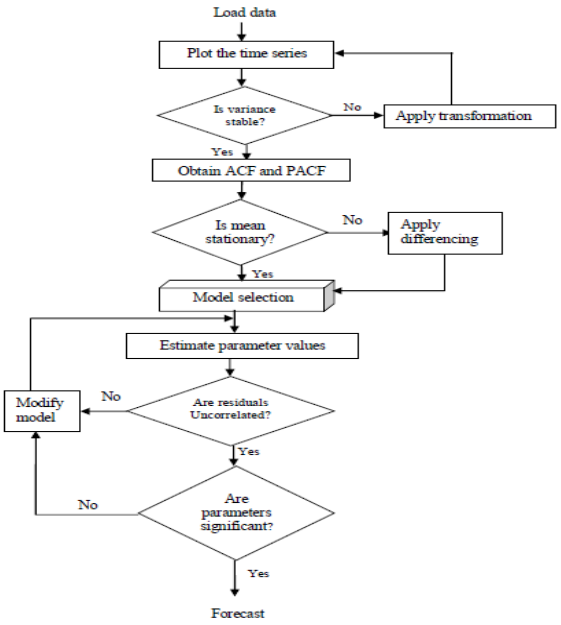


Figure 2: Steps regarding the ARIMA

ARIMA Based Methods

Fadhilah Abd. Razak *et.al* [8] proposed a method for forecasting the monthly maximum demand for electricity for a utility company by identifying a suitable time series model using ARIMA.

In this paper authors use monthly maximum demand of load data of 52 months from September 2000 to December 2004. As the first step, authors have plotted the time series data, as shown in figure 3.

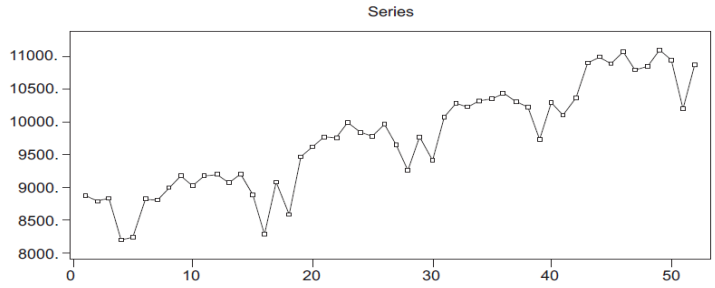


Figure 3: The maximum demand from September 2000 to December 2004.

It shows an upward linear trend and its variance is almost stable. So de-trend the time series. Each year shows a seasonal pattern with a few troughs occurring between November to February. This accounts to the various holidays, which conveys that the series is not stationary and hence needs to be transformed. The transformation is made by differencing atlag12and1forobtaininganapproximatestationaryseries.

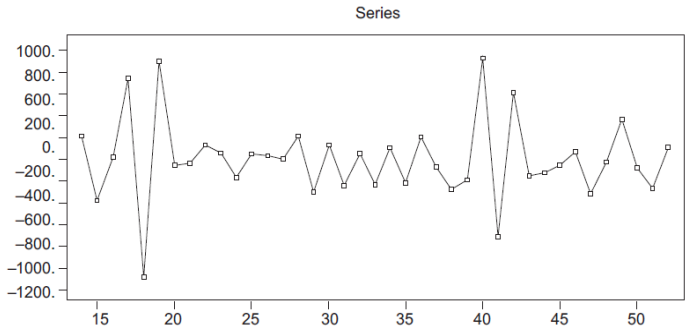


Figure 4: The time series of the residuals after differencing at lag 1 and 12

Figure4showsthedifferenced series, which is appropriatet of ita zero-meanARIMAmodel.

The next step is to determine the AR or MA terms and finding these terms by looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the differenced series. From

this plot identify the numbers of AR and MA terms that are needed. The ACF will represent a pure MA (q) model and the PACF will represent a pure AR (p) model.

Once a model is obtained, evaluate whether the model is fit well on the data or not. It is done by examining the errors from a time series forecast model. The forecast errors ideally are white noise. If it is white noise, it means that all of the signal information in the time series has been utilized by the model in order to make predictions. If it is not white noise, then further improvements are needed on the forecast model.

Many validations tests are run on the proposed models. The residuals of the selected model must pass all the tests. If it passes all tests then it is chosen for forecasting the future values. For example, by looking at the autocorrelation of the residuals and try to find any relationships exists between them. If there is no relationship, then it predicted well otherwise it is not good enough. Thus residualanalys is estimated model quality.

If many models pass the validation tests, the most suitable model can be obtained by looking at the forecasting accuracy criteria. This is done by calculating Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Relative Percentage Error(MAPE).

$$APE = \frac{actual(i) - forecast(i)}{actual(i)} * 100\%$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{actual(i) - forecast(i)}{actual(i)} * 100\%$$

4 REGRESSION

Regression is a modeling method that isused to find the relationship between a dependent variable y and one or more independent variables x_1, x_2...x_k. The main aim is to detect a function that tells the relationship between these variables and to forecast the dependent variable by using the independent variable. In the case of load forecasting, the load is described in terms of many exogenous variables like temperature, humidity, wind speed, cloud coverage etc. These variables greatly affect the consumption of the load pattern. The load model using this method isexpressedintheformas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k + \epsilon$$

where y is the load, $x_i$is the affecting factors, $\beta_i$is regression parameters with respect to x i.e, the coefficients measure the marginal effects of the predictor variables and ε is an error term. It reflects the difference between the observed and fitted linear relationship [9]. The n-tuples of observations are also assumedas:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n \end{aligned}$$

These n equations again can be written as:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$$x = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

or written as:

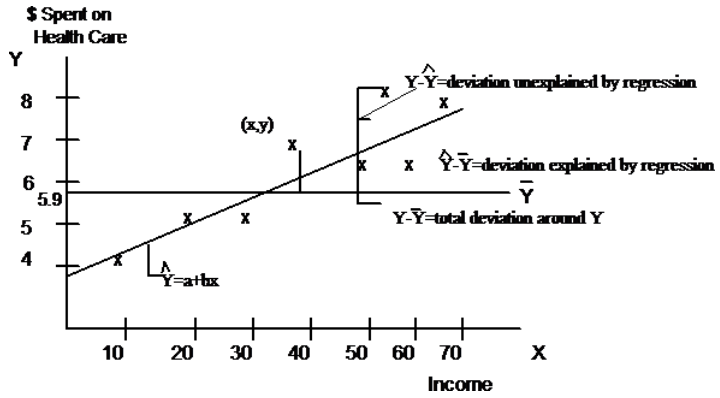
$$y = X\beta + \epsilon$$

If intercept term is present, take first column of X to be (1,1,...,1) [10]. The least square method is used to

obtain the parameters $\beta_i$which determine the best-fit line for the given data in order to minimize the sum of the squares of the vertical deviations from each data point to the line.

SSE, SSR, SST, R^2, Standard error(s):

A least squares regression selects the line with the lowest total sum of squared prediction errors. This value is called the Sum of Squares of Error, or SSE. The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean. The Total Sum of Squares (SST) is equal to SSR + SSE. This is shown in graph 1.



Mathematically,

$$SSR = \sum(\hat{Y} - \bar{Y})^2$$

$$SSE = \sum(Y - \hat{Y})^2$$

$$SST = \sum(Y - \bar{Y})^2$$

The proportion of total variation (SST) that is explained by theregression (SSR) is known as the coefficient of determination and is often referred to as R^2. The standard error of the regression(s)andR^2aretwokeygoodness-of-fitmeasuresforregressionanalysis.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

The range of R^2 will be in between 0 and 1 and its value is higher, it results in a more precise regression model. The Standard Error of a regression is a measure of its variability which is similar to standard deviation. If all independent variable is exactly recognized, then the standard error will be small. It is calculated by taking the square root of the average prediction error. The equation is representedas:

$$s = \sqrt{\frac{SSE}{n-k-1}}$$

where n is the number of observations in the sample and k is the total number of variables in the model[11].

Regression BasedMethods

Tanel Kivifold *et.al* [12] analyzing the dependency between load and temperature andone day ahead consumption of electricity forecasted using regression analysis on time series.

First make a mathematical model that describes the load. Generally, the load can be modeled as:

$l(t) = l_e(t, C, b) + \theta(t)$

where $l(t)$ stands for actual load, $l_e(t, C)$ for mathematical load expectation, $\theta(t)$ for stochastic- tic component, t for time, C for temperature, and b for wind.

Mathematical load expectation tells regular variation in the load, for example, overall growth, seasonal, intra-week, intra-day periodicity. Stochastic component describes random changes occurring in the load pattern that cannot be estimated. The proportion of these this component can be reduced by considering more exogenous variable. However, the influence of the stochastic component cannot be completely eliminated. If we use the temperature, humidity, wind speed etc as input in load forecasting which arise a

difficulty about accuracy of available historical data. It will affect the forecasted load value precision.

Here consider three main factors that influence the load: (a) Day (weekend, working dayetc.) (b) Time (c) Temperature. The measurement period was 1 year (8784 hours). After removing various erroneous measurements, 6497 hours of data (74%) remained. The temperature and load have a linear relationship, which depends on a particular day and the time ofday.

Figure 5 describe a negative relationship between the load and temperature which means the rise in temperature leads to a fall in consumption. There is also a positive relationship means that the increase in temperature leads to an increase in consumption. The dependency between temperature and load is based on earth geographic locations. A positive correlation observed in regions with warmer climatic conditions and negative correlation observed in colder places.

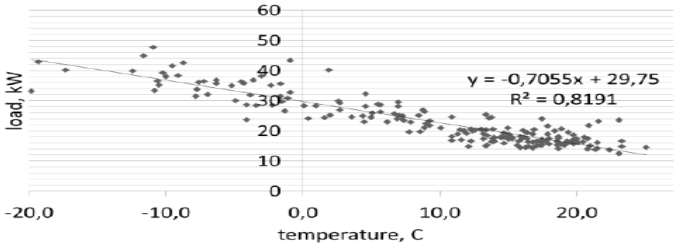


Figure 5: Correlation between load and temperature during working days at 1 pm.

In this paper, 192 separate functions are created for 24 hours per day, 7 days of weeks and public holidays. These functions characterize the dependency of the load with temperature.
The general shape of each function can be provided as:
$l_{Elspot}(t, d) = a(t, d).C + b(t, d)$

where $l_{Elspot}(t, d)$ stands for hourly consumption depending on the time of day and a par- ticular day consumption(unit kWh); a(t,d) and b(t,d) for parameters, which depend on the time of day and day of; and C for temperature (unit in degree Celsius). The forecasted results are shown in Figure6.

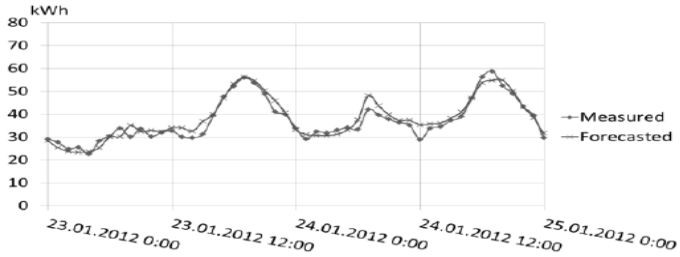


Figure 6: One day-ahead forecast.

N. Amral *(et al.)* [13] proposed three STLF models using multiple linear regression. Data used for forecasting are the rainy season and dry season load consumption pattern.

The hourly load is modeled as:

1. Intercept component: During different intervals of time of the day, it is assumed to be as aconstant.
2. Time of observation: Consumption patterns are different during each hour of a day.
3. Temperature sensitive component: It is the function of the difference of temperature at time t and the average temperature in time intervals.

The relationship between the temperature sensitive component, time of observation and the load fluctuation are in the form of polynomial lterm. Aregression equation is a polynomial regression equation if the power of independent variable is more than 1. In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points. There are three models are proposed during the intervals, 1 to 6 am, 7am to 17pm and 18 to 24pm.

For the interval 1 to 6 am, the hourly load can be modeled as follows:

$$y_i(t) = b_0 + b_1(T_i(t) - T_i(t-1)) + b_2(T_i(t-1) - T_i(t-2)) + b_3 t + b_4 t^2$$

where:

- $y_i(t)$:predictedloadathourtintheintervaliftheday.
- b_0: Intercept component (regression constantcoefficient).
- $T_i(t)$: Temperature at time t in the interval i of theday.
- t: Time ofobservation.

For interval 7 am to 17 pm, the load curve is modeled as follows:

$$\begin{aligned} y_i(t) = b_0 + b_1(T_i(t) + \\ b_2(T_i(t-1) - T_{iav} + \\ b_3(T_i(t) - T_{iav})^2 + \\ b_4(T_i(t) - T_{iav})^3 + \\ b_5 t + b_6(T_{ava} - T_{avb}) + \\ b_7(T_i(t) - T_i(t-1) + \\ b_8(T_i(t-1) - T_i(t-2)) + \\ b_9(T_i(t-2) - T_i(t-3)) \end{aligned}$$

where:

- $b_1..b_9$: Regression parameters of temperature sensitive component and time of observations.
- T_{iav}: Temperature average in the interval i of theday.
- T_{ava}:Averagetemperatureofprevious24hourstothetimet.
- T_{avb}: T_{ava}lagged 3hours.

For interval 18 to 24 pm, the load curve is modeled as follows:

$y_i(t)=b_0+b_1T_i(t)+b_2t+b_3t^2+b_4t^3$

This is different from the other two models due to the sharp increase of load when sky brightness decrease as during night consumers turn on the lights.

5 ANALYSIS

Models	Analysis
ARIMA	• It only requires the prior data of a time series. • It is very effective and precise method, but the model requires more sequence data. • It constantly standardized to meet the model requirements. • It constantly tested in p, q value.
Regression	• It produces results faster because of the direct mathematical computations. • Good results can be obtained with small data sets. • Implementation is time-consuming and very costly. • Weather changes cannot be easily integrated in linear regression models.

6 CONCLUSION

In the modern power system, electricity load forecasting is an important area for its promising management and operation. It helps the operator of a power system for making decision on maintenance, unit commitment, allocation of fuel and scheduling spinning re- serve. Limiting the operating costs and reliable power system operations greatly depends on prediction accuracy. The selection of load forecasting methods depends on factors such as time horizon of forecasts, availability, and usefulness of historical data, the degree of accuracy desirable etc. Statistical methods have transparency and solve the problem of forecastingbyusingaknownequation.Butitrequiresalargeamountofaccuratedata.

ARIMA and regression are two techniques with different approaches. The former uses the past values of a variable and previous error terms for forecasting while the later models the variable based on the values of other variables. ARIMA requires long time series data with constant structure and it doesn't have an automatic updating feature. The whole procedure must be repeated when new data available. The ARIMA is best suited for short term forecasts with high-frequencydata.

Theoutputfromtheregressiongivesusameasureofhowstrongtherelationshipbetween the variables used. The functional relationship that is established between any two or more variables on the basis of limited data may not hold good if more and more data are taken intoconsideration.

REFERENCES

[1] [1] Wikipedia contributors, Electricity sector in India — Wikipedia, the free encyclopedia, [Online; accessed 7-March-2019](2019).
URL https://en.wikipedia.org/w/index.php?title=Electricity_sector_in_India&oldid= 883580990
[2] Loadforecasting-purpose,classificationandprocedure,studyelectrical.com.
[3] M. U. Fahad, N. Arbab, Factor affecting short term load forecasting, Journal of Clean Energy Technologies 2 (4) (2014) 305–309.
[4] Wikipedia contributors. (2019, March 13). Time series. In Wikipedia, The Free Encyclopedia. Retrieved 08:22, March 18, 2019, from https://en.wikipedia.org/w/index.php?title=Time_series&oldid=887619148
[5] Wikibooks, Econometric theory/data — wikibooks, the free textbook project, [Online; accessed 9-March-2019](2018). URL https://en.wikibooks.org/w/index.php?title=Econometric_Theory/Data&oldid=3442506
[6] Statisticssolutions, https://www.statisticssolutions.com/time-series-analysis/ ((accessed March09)
[7] YouTube https://www.youtube.com/playlist?list= PLXxXXnEc7-nyKhOAZwtnrOliFslUs-9WC.
[8] F. A. Razak, M. Shitan, A. H. Hashim, I. Z. Abidin, "Load forecasting using time series models," Jurnal Kejuruteraan 21 (2009) 53–62.
[9] B.-J. Chen, M.-W. Chang, et al., "Load forecasting using support vector machines: A study on eunite competition 2001", IEEE transactions on power systems 19 (4) (2004) 1821–1830.
[10] Multiple linear regression model-iit kanpur,http://home.iitk.ac.in/~shalab/regression/Chapter3-Regression-MultipleLinearRegressionModel.pdf.
[11] The coefficient of determination, r-squared, https://newonlinecourses.science.psu.edu/stat501/node/255 (2018(accessed March 09,2019).
[12] T. Kivipold, J. Valtin, "Regression analysis of time series for forecasting the electricity consumption of small consumers in case of an hourly pricing system", Adv Autom Control Model Simul 34 (2) (2013) 127–132.
[13] N.Amral,C.Ozveren,D.King,"Shorttermloadforecastingusingmultiplelinearregression",in:UniversitiesPowerEngineeringConference,2007.UPEC2007.42ndInternational,IEEE,2007,pp.1192–1198.