International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-8)

Research Article | **August 2017**

A Pragmatic Study on Time Series Models for Big Data

B. Arputhamary
Research Scholar, Mother Teresa Women's University,
Kodaikanal, Tamil Nadu, India

Dr. L. Arockiam
Associate Professor, St. Joseph College,
Tiruchirappalli, Tamil Nadu, India

Abstract—

Recent years have witnessed the growth of Big Data, particularly Time Series data which initiates major research interest in Time Series analysis and forecasting future values. It finds interest in many applications such as business, stock market and exchange, weather forecasting, electricity demand, cost and usage of products and in any kind of place that has specific seasonal or trendy changes over time. The forecasting of Time Series data provides the organization with useful information that is necessary for making important decisions. In this paper, a detailed study is performed to find the total number of bike users with respect to the season and weather on Capital Bike Sharing System (CBS) dataset. The study covers the Auto Regressive Integrated Moving Average (ARIMA), Holt-Winters Additive and Multiplicative forecasting models to analyse the seasonal and trendy fluctuations of the given dataset to improve performance and accuracy.

Keywords— Big Data Analytics, Forecasting, Time Series, ARIMA, Holt-Winters

I. INTRODUCTION

Big Data Analytics (BDA) applications empower data scientists, predictive modellers, statisticians and other analytics professionals to analyze massive volumes of data[15]. These data are structured transaction data and also other forms of data such as semi-structured and un-structured that are often left idle by conventional Business Intelligence (BI) and other analytic programs [14]. The internet click stream data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records and machine data captured by sensor devices that are connected to the Internet Of Things (IOT) are some of the examples for Big Data. Time Seriesmodelling is a dynamic research area which has attracted attentions of researcher's community over last few decades. The main aim of Time Seriesmodelling is to carefully collect and rigorously study the past observations of a Time Series to develop an appropriate model which describes the inherent structure of the series. Time Series forecasting thus can be termed as the act of predicting the future by understanding the past [1]. Due to the indispensable importance of Time Series forecasting in numerous practical fields such as business, economics, finance, science, engineering and so on it is important to generate future values for the series that is to make forecasts.It is obvious that a successful Time Series forecasting depends on an appropriate model fitting. So, proper care should be taken to fit an adequate model. Llot of efforts have been taken by researchers over many years for the development of efficient models to improve the forecasting accuracy. As a result, various important Time Series forecasting models have been evolved.

II. AIM AND OBJECTIVES

Today data are generated in an unprecedented manner and most of them are Time Series data. In Time Series analysis, forecasting plays an important role in the area of statistics, econometrics, quantitative finance, seismology, geophysics, weather, demand and sales forecasting. It is mainly used for signal detection and estimation in the context of signal processing, control and communication engineering[13]. In the context of data mining, pattern recognition and machine learning Time Series analysis can be used for clustering, classification and query by content, irregularity detection as well as forecasting. The primary objective of this paper is, to perform a pragmatic study on the famous Time Series models and to identify the appropriate forecasting model for the Time Series data set provided by Kaggle2 collection from Capital Bikesharing System of ten years around Washington.

III. TIME SERIES MODELS AND FORECASTING

A. Time Series

A Time Series is a sequential set of data points, measured typically over successive times. It is mathematically defined as a set of vectors x(t), t = 0,1,2,... where t represents the time elapsed [1]. The variable x(t) is treated as a random variable. The measurements taken during an event in a Time Series are arranged in a proper chronological order. A Time Series containing records of a single variable is termed as univariate. But if records of more than one variable are considered, it is termed as multivariate. A Time Series can be continuous or discrete. In a continuous Time Series, observations are measured at every instance of time, whereas in a discrete Time Series observations are measured at discrete points of time such as hourly, daily, weekly, monthly or yearly time separations. For example temperature readings, flow of a river, concentration of a chemical process and so on can be recorded as a continuous Time Series. On the other hand, population of a particular city, production of a company, exchange rates between two different currencies may represent discrete Time Series[2].

B. Components of a Time Series

In general, a Time Series is affected by four main components, namely, Trend, Cyclical, Seasonal and Irregular. The general tendency of a Time Series is to increase, decrease or stagnate over a long period of time is termed as Secular Trend or simply Trend. And hence, trend is a long term movement in a Time Series. For example, series relating to population growth, number of houses in a city show upward trend, whereas downward trend can be observed in series relating to death rates, epidemics and so on.Cyclic variation takes account of regular cyclic variation at periods other than one year. Examples include business cycles over a period of five years and the daily measure in the biological behaviour of living creatures. Irregular fluctuations is often used to describe any variation that is left over after trend, seasonality and other systematic effects have been removed. As such, they may be completely random in which case they cannot be forecasted. Seasonal variations in a Time Series are fluctuations within a year during the season. The important factors causing seasonal variations are: climate and weather conditions, customs, traditional habits, etc. For example sales of ice-cream increase in summer, sales of woolen cloths increase in winter. Hence, seasonal variation is an important factor for businessmen, shopkeeper and producers for making proper future plans.

C. Time Series Analysis

The procedure of fitting a Time Series to a proper model is termed as Time Series Analysis [2]. It comprises of methods that attempt to understand the nature of the series and is often useful for future forecasting and simulation. In Time Series forecasting, past observations are collected and analyzed to develop a suitable mathematical model[3]. The future events are then predicted using this model. This approach is particularly useful when there is not much knowledge about the statistical pattern followed by the successive observations or when there is a lack of a satisfactory explanatory model. Valuable strategic decisions and precautionary measures are taken based on the forecasted results. Thus, a good forecast depends on fitting an adequate model to a Time Series. Over the past several decades many efforts have been made by researchers for the development and improvement of suitable Time Series forecasting models.

IV. FORECASTING MODELS

In recent years, Time Series prediction is done in many applications that deal with numerical data. The prediction can be done on the basis of three different time spaces as:

Short-Term period: The Short-Term period focus on a time frame or period of less than three months.

Mid-Term period: The Mid-Term period focuses on a time frame of three months to one year.

Long-Term period: The Long-Term considers a time period more than a year.

Based on the time period and the type of data, many techniques have been used for prediction. This research study provides a detail study on the various forecasting models that have been used for prediction. The prediction and forecasting were performed using traditional forecasting techniques that make use of mathematical formula. These basic techniques can be improved further using various advancements in different tools and automation methods. The traditional forecasting techniques which are considered for the empirical study are discussed below:

A. Auto regressive Integrated Moving Average (ARIMA) Model

The ARIMA model is a generalization of an Auto Regressive Moving Average (ARMA) model to include the case of non-stationary as well. In ARIMA model a non-stationary Time Series is made stationary by applying finite differencing of the data points[5]. The mathematical formulation of the ARIMA (p,d,q) model using lag polynomial is given below:

$$\phi(L)(1-L)^d y_t = \theta(L)\varepsilon_t, i.e.$$

$$\left[1-\sum_{i=1}^{p}\phi_i L^i\right](1-L)^d y_t = \left[1+\sum_{j=1}^{q}\theta_j L^j\right]\varepsilon_t \quad \text{……………………………………………………. (1)}$$

wherep, d and q arethe autoregressive, integrated, and moving average integers greater than or equal to zero.

1. The integer d controls the level of differencing. Generally d =1 is enough in most cases.
2. When d = 0, then it reduces to an ARMA (p,q) model.
3. An ARIMA (p,0,0) is nothing but the AR(p) model and ARIMA(0,0,q) is the MA(q) model.
4. ARIMA (0,1,0), i.e. yt= yt-1+ εt is a special one and known as the Random Walk model [4]. It is widely used for non-stationary data, like economic and stock price series.

B. Holt–Winters (HW) Model

The Holt-Winters model uses a modified form of exponential smoothing. It applies three exponential smoothing formulae such as level or mean, trend and seasonal component to the series. A combination of these three series is used to calculate the prediction output. The goal is to develop a high accuracy and low cost forecasting model that could integrate with the existing system. In 1957s, the researcher Charles Holt showed that the forecasting method most often used at the time, the method of exponentially moving average, could be used not only to smooth the level of a variable, but also to smooth the trend, seasonality and other components of aprediction. The new model created by Holt could control multiplicative and additive seasonality, additive and multiplicative trend and standard errors [11].

In addition, the new system was quick and easy to program but required minimal data storage. It uses simple initial conditions and robust parameters and allowed automatic adaptation. This model was then studied and tested in several Time Series by graduate student P.R. Winters, who found that the prediction formulae was surprisingly accurate. Winters

published results in 1960s and the new model was called Holt-Winters method. The formulas of Holt and Holt-Winters were quickly incorporated into marketable software systems for estimating and more than 50 years later, still have been used in researches[1][2].

In general, it can be said that the Holt-Winters method, sometimes called the method of Winters or seasonal exponential smoothing, is a sophisticated extension of the exponential smoothing methodology. It generalizes the methodology to deal with trend and seasonality. To do so, it considers α , β and γ as the three smoothing parameters. The exponential smoothing formulae applied to a series with a trend and constant seasonal component using the Holt-Winters Additive technique is as follows [6]:

$$a_t = \alpha(Y_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1})$$

…… (2)

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

…….. (3)

$$s_t = \gamma(Y_t - a_t) + (1-\gamma)s_{t-p}$$

…….. (4)

where α , β and γ are the smoothing parameters.

a_t is the smoothed level at time t.

b_t is the change in the trend at time t.

s_t is the seasonal smooth at time t.

p is the number of seasons per year.

The initial values required for Holt Winter's algorithm are

$$a_p = \frac{1}{p}(y_1 + y_2 + ... + y_p)$$

…… (5)

$$b_p = \frac{1}{p}\left[\frac{y_{p+1} - y_1}{p} + \frac{y_{p+2} - y_2}{p} + ... + \frac{y_{p+p} - y_p}{p}\right]$$

…… (6)

$$s_1 = Y_1 - a_p \; s_2 = Y_2 - a_p \; ..... , s_p = Y_p - a_p$$

……(7)

$$y_{T+\tau} = a_T + b_T + s_T$$

…… (8)

Multiplicative Model:

$$a_t = \alpha\frac{Y_t}{s_{t-p}} + (1-\alpha)(a_{t-1} + b_{t-1})$$

…… (9)

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

…… (10)

$$s_t = \gamma\frac{Y_t}{a_t} + (1-\gamma)s_{t-p}$$

…….. (11)

The initial values for multiplicative model are:

$$a_p = \frac{1}{p}(y_1 + y_2 + ... + y_p)$$

…… (12)

$$b_p = \frac{1}{p}\left[\frac{y_{p+1} - y_1}{p} + \frac{y_{p+2} - y_2}{p} + ... + \frac{y_{p+p} - y_p}{p}\right]$$

…….. (13)

$$s_1 = \frac{Y_1}{a_p} \; s_2 = \frac{Y_2}{a_p} \quad ,\ldots\ldots\ldots\ldots, \quad s_p = \frac{Y_p}{a_p}$$

……..(14)

The following section describes the experimentations of these forecasting models using CBS samples with R environment[12]-[14].

V. EXPERIMENTAL SUMMARY

The programming language R is used to build the Time Series models such as ARIMA, Holt-Winters Additive and Multiplicative with the aim of forecasting the bike counts for a certain set of season under various weather conditions. In this analysis, Microsoft Excel is used for initial analysis. The input file is downloaded as Comma Separated File(CSV) and uploaded into R-studio. The main part of the analysis is done through the use of R programming language [11].

A. Empirical Findings

The following steps explains the empirical analysis and findings of the CBS dataset.

Step 1: Loading the Data

The following Figure 4.2 describes the statistics of sample instances. Information regarding mean, median and the mode are discussed. The purpose of taking this summary is to get a feel for the data, to use in statistical tests and to indicate the error that are associated with results. Summary function in R is used to give an output of some of the commonly used description statistics. In description statistics, the details about 1st and 3rd quartile isgiven. The first quartile gives the middle number between the smallest number and the median of the dataset. The third quartile gives the middle value between the median and the highest value of the dataset. The mean and median are also described to discover outliers which may disturb forecasting accuracy when building the forecasting model.

```
> data <- read.csv("D:/1.JESUS NEVER FAILS/DataSet2/hour.csv")
> summary(data)
    instant              dteday          season           yr
 Min.   :    1   2011-01-01:   24   Min.   :1.000   Min.   :0.0000
 1st Qu.: 4346   2011-01-08:   24   1st Qu.:2.000   1st Qu.:0.0000
 Median : 8690   2011-01-09:   24   Median :3.000   Median :1.0000
 Mean   : 8690   2011-01-10:   24   Mean   :2.502   Mean   :0.5026
 3rd Qu.:13034   2011-01-13:   24   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :17379   2011-01-15:   24   Max.   :4.000   Max.   :1.0000
                 (Other)   :17235
      mnth             hr             holiday           weekday
 Min.   : 1.000   Min.   : 0.00   Min.   :0.00000   Min.   :0.000
 1st Qu.: 4.000   1st Qu.: 6.00   1st Qu.:0.00000   1st Qu.:1.000
 Median : 7.000   Median :12.00   Median :0.00000   Median :3.000
 Mean   : 6.538   Mean   :11.55   Mean   :0.02877   Mean   :3.004
 3rd Qu.:10.000   3rd Qu.:18.00   3rd Qu.:0.00000   3rd Qu.:5.000
 Max.   :12.000   Max.   :23.00   Max.   :1.00000   Max.   :6.000

   workingday      weathersit         temp           atemp
 Min.   :0.0000   Min.   :1.000   Min.   :0.020   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.340   1st Qu.:0.3333
 Median :1.0000   Median :1.000   Median :0.500   Median :0.4848
```

```
   workingday      weathersit         temp           atemp
 Min.   :0.0000   Min.   :1.000   Min.   :0.020   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.340   1st Qu.:0.3333
 Median :1.0000   Median :1.000   Median :0.500   Median :0.4848
 Mean   :0.6827   Mean   :1.425   Mean   :0.497   Mean   :0.4758
 3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:0.660   3rd Qu.:0.6212
 Max.   :1.0000   Max.   :4.000   Max.   :1.000   Max.   :1.0000

      hum          windspeed          casual         registered
 Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :  0.0
 1st Qu.:0.4800   1st Qu.:0.1045   1st Qu.:  4.00   1st Qu.: 34.0
 Median :0.6300   Median :0.1940   Median : 17.00   Median :115.0
 Mean   :0.6272   Mean   :0.1901   Mean   : 35.68   Mean   :153.8
 3rd Qu.:0.7800   3rd Qu.:0.2537   3rd Qu.: 48.00   3rd Qu.:220.0
 Max.   :1.0000   Max.   :0.8507   Max.   :367.00   Max.   :886.0

      cnt
 Min.   :  1.0
 1st Qu.: 40.0
 Median :142.0
 Mean   :189.5
 3rd Qu.:281.0
 Max.   :977.0
```

Fig. 1. Loading the data

The following Fig.2.gives the details about all seventeen variables in the dataset.

```
> str(data)
'data.frame':   17379 obs. of  19 variables:
 $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ dteday    : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 1 1 1 1 1 1$
 $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ yr        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ hr        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ weekday   : int  6 6 6 6 6 6 6 6 6 6 ...
 $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
 $ weathersit: int  1 1 1 1 1 2 1 1 1 1 ...
 $ temp      : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
 $ atemp     : num  0.288 0.273 0.273 0.288 0.288 ...
 $ hum       : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
 $ windspeed : num  0 0 0 0 0 0.0896 0 0 0 0 ...
 $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
 $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
 $ cnt       : int  16 40 32 13 1 1 2 3 8 14 ...
 $ cassqrt   : num  1.73 2.83 2.24 1.73 0 ...
 $ regsqrt   : num  3.61 5.66 5.2 3.16 1 ...
```

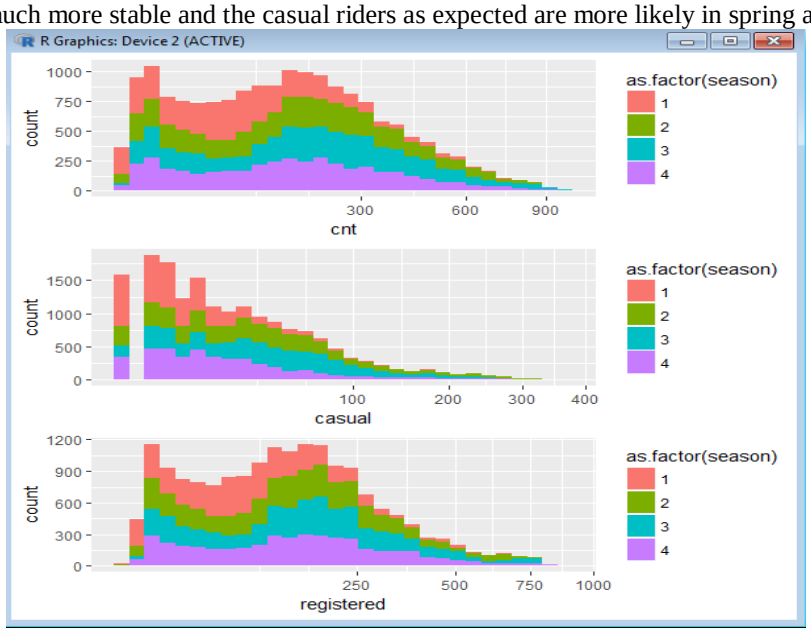
Fig. 2. Attributes of the data

Step 2: Exploring the Data

Data exploration is the process of summarizing the characteristics of the dataset which is very important in data analysis. It is commonly conducted using visual analytical tools but can also be done in more advanced statistical package such as R.

```
p1 <- qplot(x = cnt, data = data, fill = as.factor(season)) +     scale_x_sqrt()
p2 <- qplot(x = casual, data = data, fill = as.factor(season)) +    scale_x_sqrt()
p3 <-qplot(x = registered, data =data, fill = as.factor(season)) + scale_x_sqrt()
grid.arrange(p1, p2, p3)
```

The following Fig. 3 explores the total count of bikes of casual and registered users with respect to the season's spring, summer, fall and winter. The season values 1, 2, 3 and 4 represents the season's spring, summer, fall and winter respectively. The total count of casual users is greater than the registered users particularly in season 1 that is spring. Registered riders are much more stable and the casual riders as expected are more likely in spring and summer.



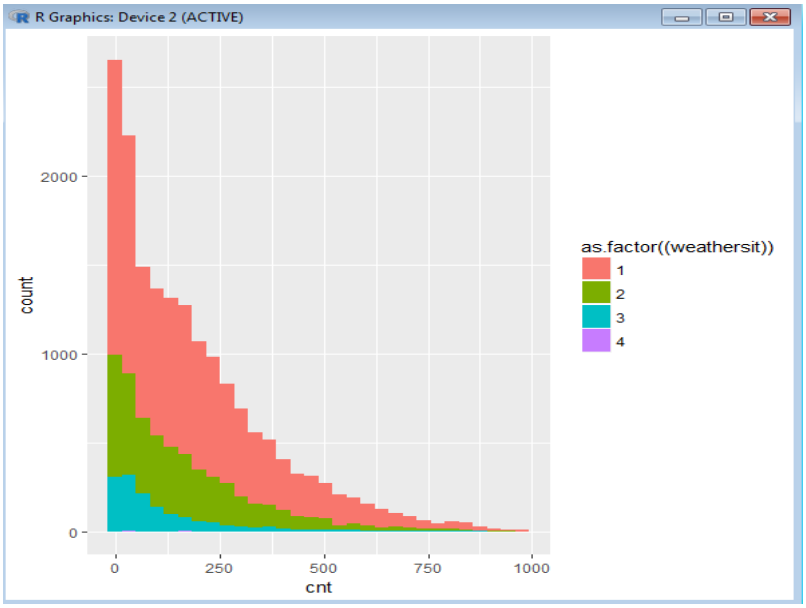


Fig. 3. Count of casual and registered usersFig. 4. Count with respect to weather

There are four categories of weather participation:
1 = clear, few clouds, partly cloudy
2 = mist & cloudy, mist & broken clouds, mist & few clouds, mist
3 = light snow, light rain & thunderstorm & scattered clouds, light rain & scattered clouds
4 = heavy rain & ice pellets & thunderstorm & mist, snow & fog

In Fig.4.total count of bikes are given with respect to the weather condition. And it is explored that more number of bikes are rented on Weather 1(Clear, few or partly cloudy) days. No bikes are rented on weather 4 (heavy rain, ice pellets, thunderstorm, mist, snow and fog) days. In the given dataset, weather and season are important attributes to determine the bike demand. The following Fig.5. depicts the bike counts with respect to temperature.

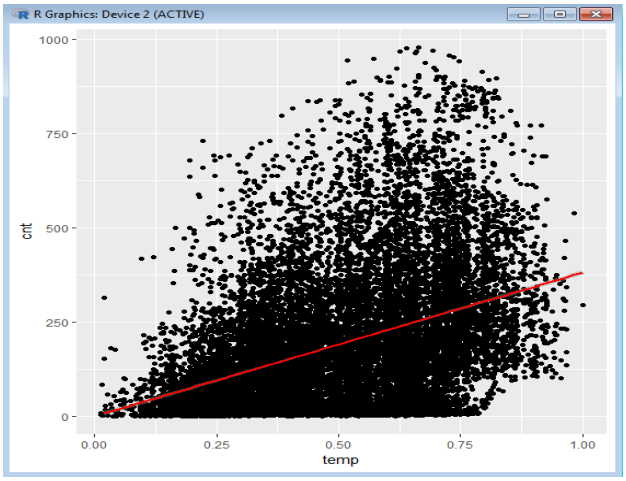


Fig.5. Bike counts with respect to temp

The temperature is normalized between 0 and 1 and maximum number of bikes are rented between 0.25 and 0.75. Therefore, the temperature between 0.25 and 0.75 are the preferable for renting the bikes.

The following figures Fig.6(a) and 6(b) represent the histograms of the attributes present in the dataset. Histograms are important to identify the outliers. If any outliers are identified, it must be considered carefully before giving the inputs. The following histograms depicts the attributes season, weather-sit, hum and holiday are not normally distributed. They are nominal data means categorical that is no computations can be performed on nominal data. And temp, hum and wind-speed are interval data. The cnt, casual and registered attributes are positively skewed data.

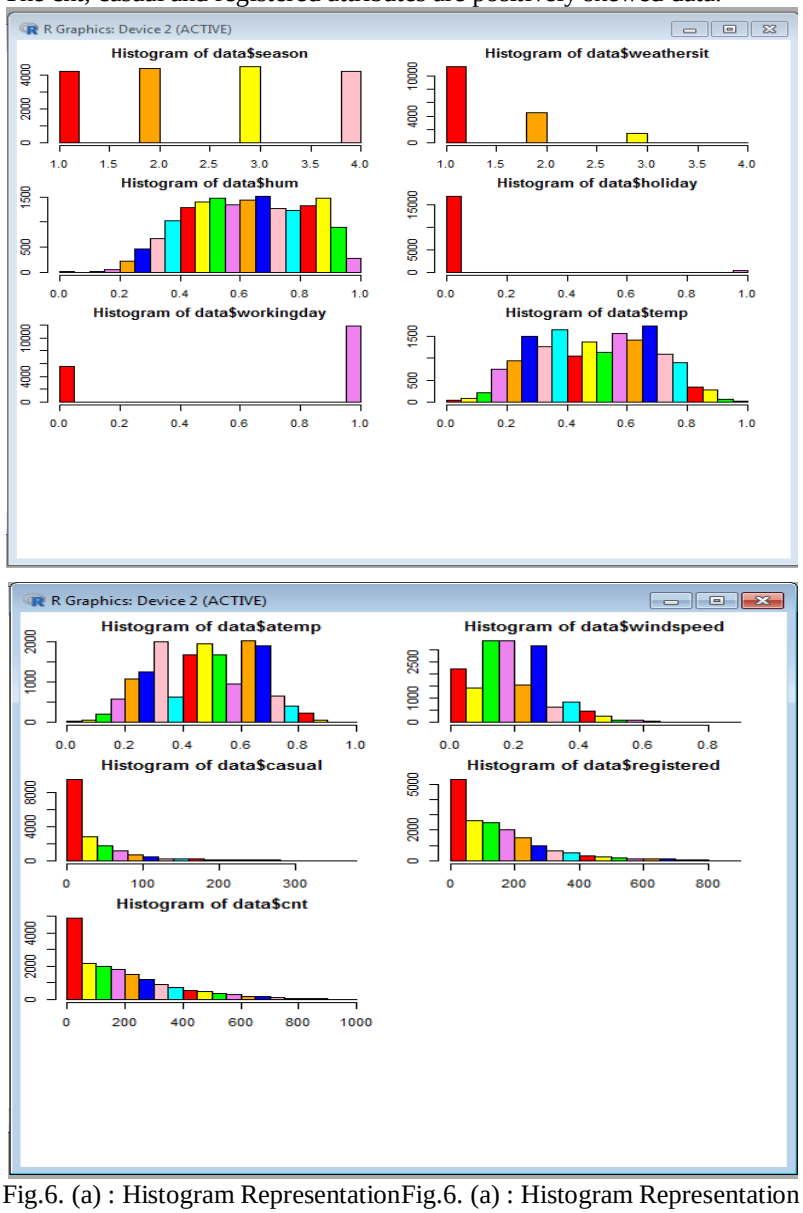


Fig.6. (a) : Histogram RepresentationFig.6. (a) : Histogram Representation

Few inferences can be drawn by looking at these histograms:

1. Season has four categories of almost equal distribution
2. Weather 1 has higher contribution i.e. mostly clear weather.
3. As expected, mostly working days and variable holiday is also showing a similar inference. Here a variable for weekday can be generated using holiday and working day. In case, if both have zero values, then it must be a working day.
4. Variables temp, atemp, humidity and wind speed looks naturally distributed.

Step 3: Hypothesis Testing

Hypothesis:Hour is a significant variable.

The followingFig.7.explores the trend of bike demand over hours.

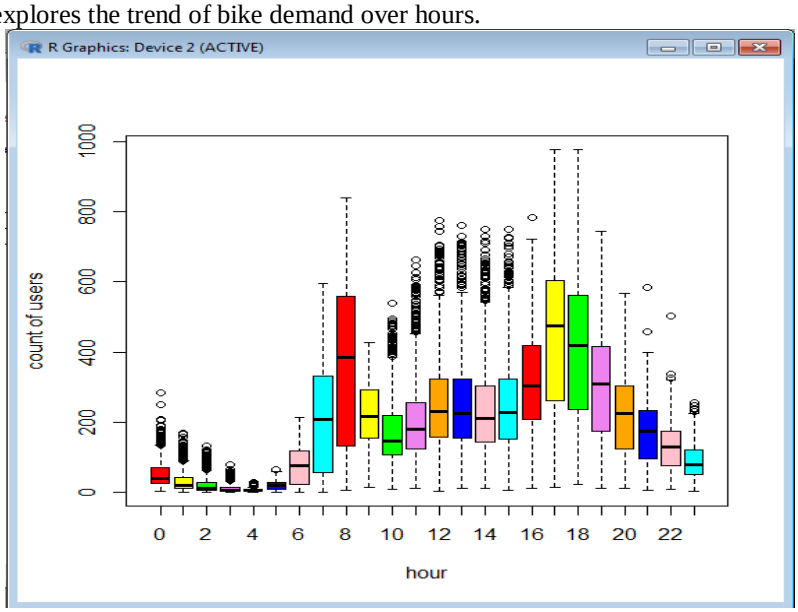


Fig.7.Trend analysis of Bike demand over hours

The bike demand can be categorized as follows:

- High : 7-9 and 17-19 hours
- Average : 10-16 hours
- Low : 0-6 and 20-24 hours

The following Fig.8.explores the trend in casual and registered users separately.

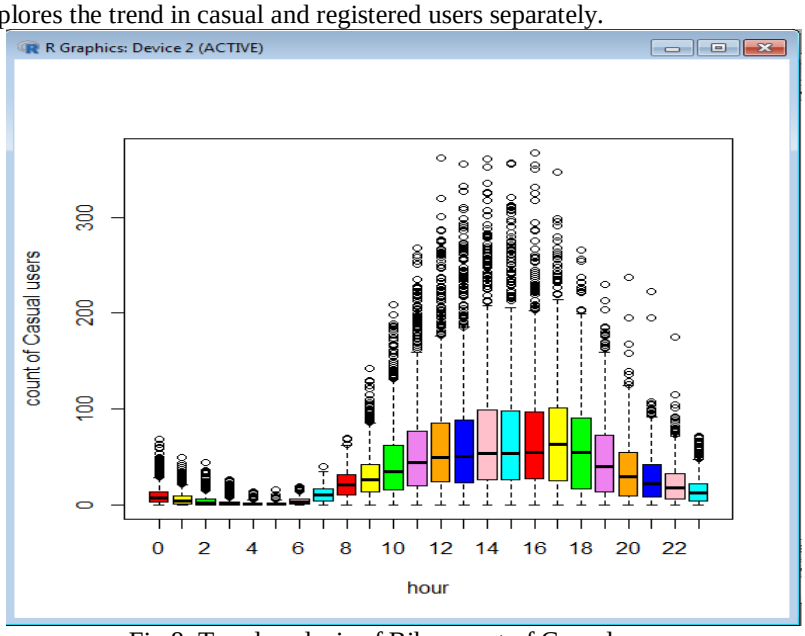


Fig.8. Trend analysis of Bike count of Casual users

The trend analysis of bike count of casual users are categorised as follows:

- High : 14-18 hours
- Average : 10-16 hours
- Low : 0-6 and 22-24 hours

In Fig.9.the trend analysis of bike count of registered users are given.

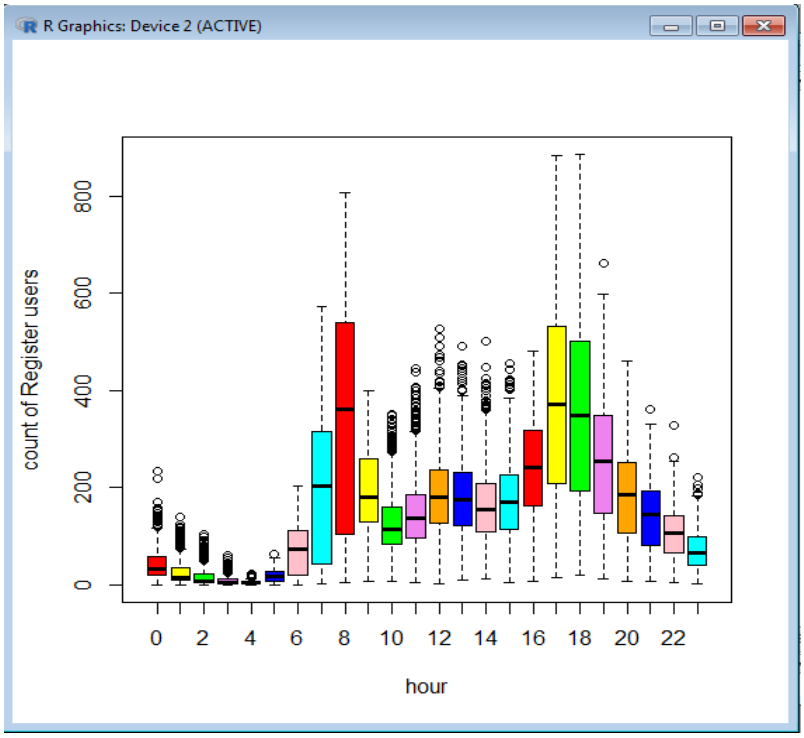


Fig.9.Trend analysis of Bike count of Registered Users

The trend analysis of bike count of registered users are categorised as follows:

- High : 6-8 and 16-18 hours
- Average : 10-16 hours
- Low : 0-6 and 22-24 hours

From the study, it is noted that registered users have similar trend as count whereas, casual users have different trends. Thus, hour is a significant variable and our hypothesis is true. And the following Fig.10.represents the logarithmic representation of bike count.

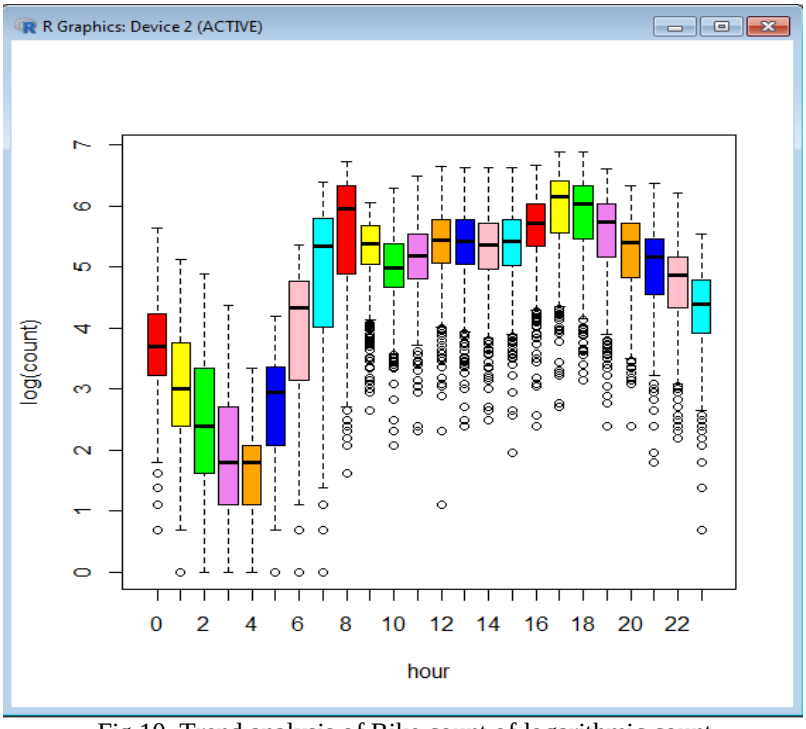


Fig.10. Trend analysis of Bike count of logarithmic count

Step 4: Finding Correlations

In Fig.11.the correlation matrix is given, showing how each individual attribute is related to all other attributes. The correlation matrix gives an insight about how attributes are related with each other. In correlation matrix values between

0 and 0.25 are considered as weak, 0.25 and 0.75 are moderate and values between 0.75 and 1.00 are strong. If the value is preceded by a negative symbol, the relationship is considered as indirect or negative, in that case, as one variable increases the other falls. That is, they are having negative relationship. If the value is positive, they have a direct relationship that is, if one increases the other will also increase. A value of 1 or -1 is estimated to have a perfect correlation.

```
> sub=data.frame(data$registered,data$casual,data$cnt,data$temp,data$hum,data$a$
> cor(sub)
                data.registered data.casual    data.cnt   data.temp    data.hum
data.registered      1.00000000  0.50661770  0.97215073  0.33536085 -0.27393312
data.casual          0.50661770  1.00000000  0.69456408  0.45961565 -0.34702809
data.cnt             0.97215073  0.69456408  1.00000000  0.40477228 -0.32291074
data.temp            0.33536085  0.45961565  0.40477228  1.00000000 -0.06988139
data.hum            -0.27393312 -0.34702809 -0.32291074 -0.06988139  1.00000000
data.atemp           0.33255864  0.45408007  0.40092930  0.98767214 -0.05191770
data.windspeed       0.08232085  0.09028678  0.09323378 -0.02312526 -0.29010490
                 data.atemp data.windspeed
data.registered  0.33255864     0.08232085
data.casual      0.45408007     0.09028678
data.cnt         0.40092930     0.09323378
data.temp        0.98767214    -0.02312526
data.hum        -0.05191770    -0.29010490
data.atemp       1.00000000    -0.06233604
data.windspeed  -0.06233604     1.00000000
```

Fig.11.Finding Correlation coefficients

Here are a few inferences drawn by looking at the above results:

1. Temp variable is positively correlated that is associated with dependent variables and the casual usersare more compared to registered users
2. Variable atemp is highly correlated with temp.
3. Wind-speed has lower correlation as compared to temp and humidity

Step 5: Model Fitting

In this research, Time Series models ARIMA, Holt-Winters Additive and Holt-Winters Multiplicative are taken into consideration. The reason for taking these models are, they are fitting well for the Time Series which has seasonal components. To analyse the forecast performance of these models, the measures RMSE, MSE and MAPE are considered. For each model, error rates are calculated and are tabulated in Table.1. Also in this paper, forecast diagram for all these models are given showing the actual and forecasted data points. The following Fig.12.shows forecasting using Holt-Winters (HW) Additive Model where x-axis represents date and y-axis represents bike counts on particular day.

In Fig.12, 2011 data are considered as historical or actual data which is represented in red color and the 2012 data are forecasted value which is in blue color. Fig.13. gives the forecasting of bikes demand using Holt-Winters multiplicative model. And Fig.14. depicts the forecasting of demands using ARIMA forecasting model.

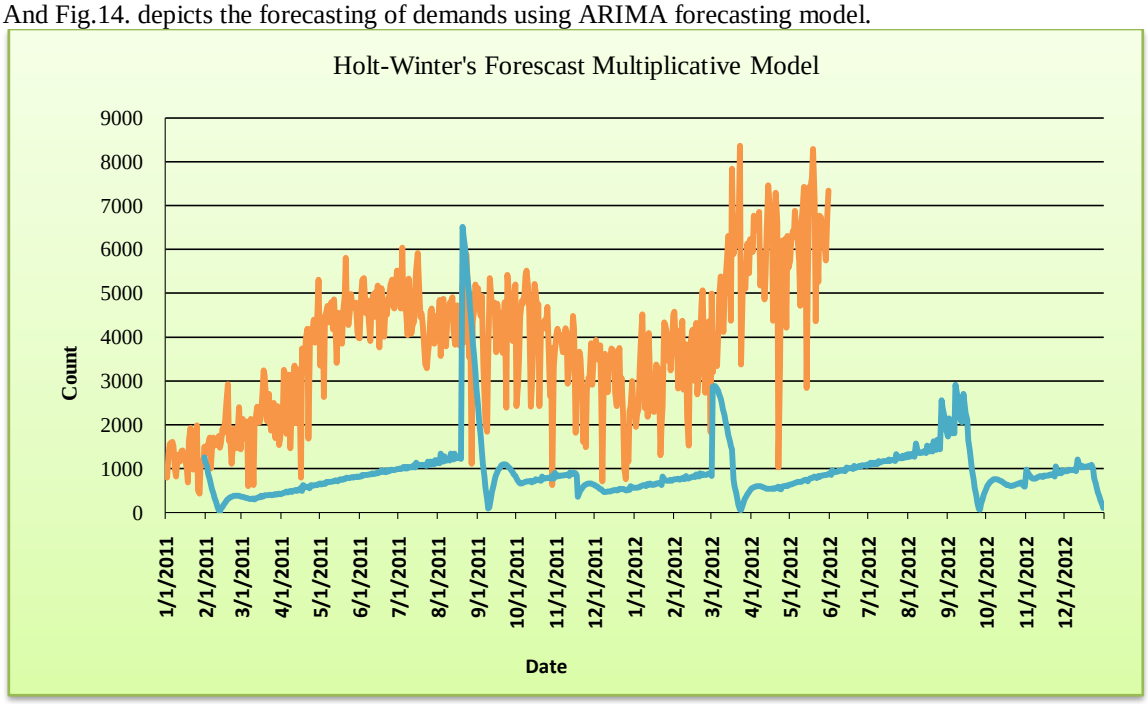


Fig.12. Holt-Winters(HW) Additive Model

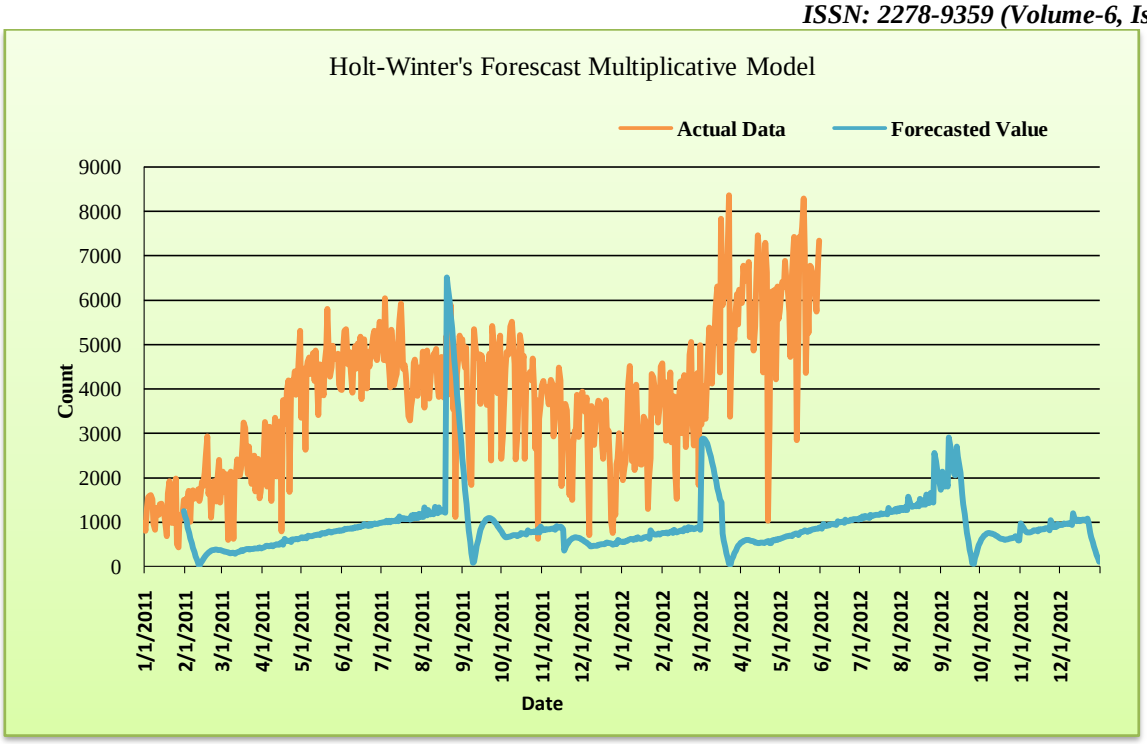


Fig.13. Holt-Winters(HW) Multiplicative Model

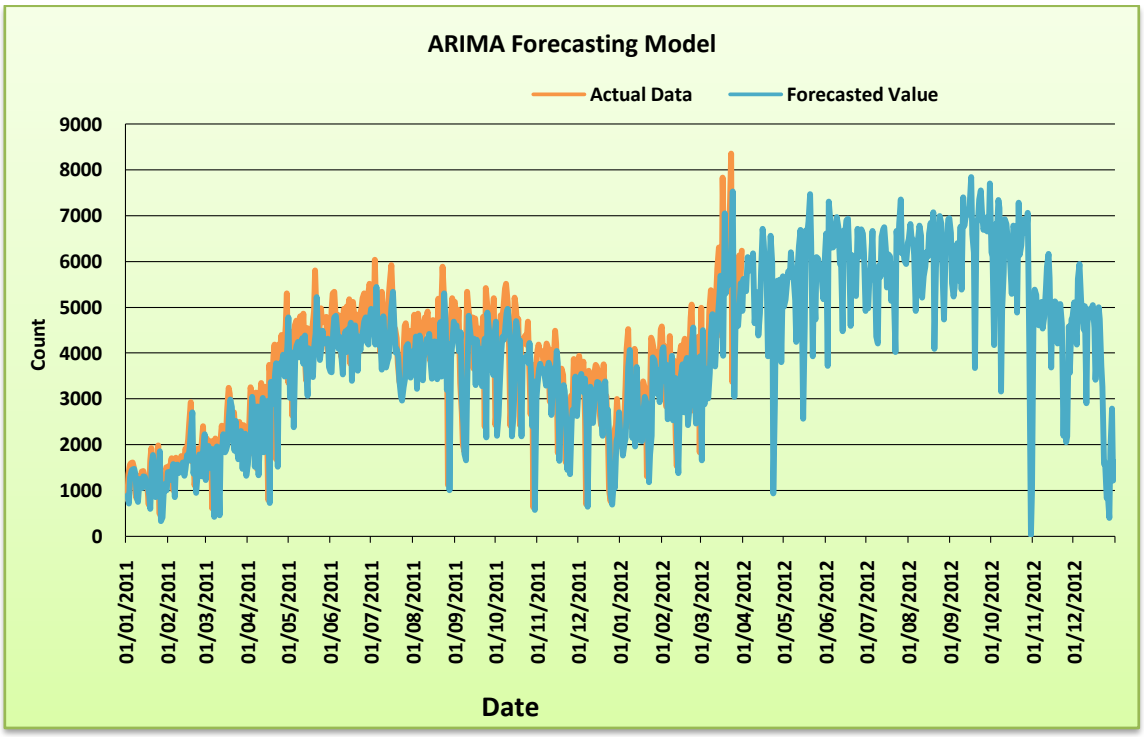


Fig.14. ARIMA forecasting

Table.1. Findings on Empirical Analysis

Fit Statistic	RMSE	MAE	MAPE
Holt-Winter's Forecast (Additive)	809.0434	238.208	27.879
Holt-Winter's Forecast (Multiplicative)	3253.252	629.308	92.366
ARIMA Model	1123.567	463.89	63.325

The following Fig.15.gives the evaluation of models where x-axis represents the evaluation metrics and the y-axis gives the error rate. From the discussion, it is observed that the Holt-Winters Additive model is suited well for the CBS dataset.

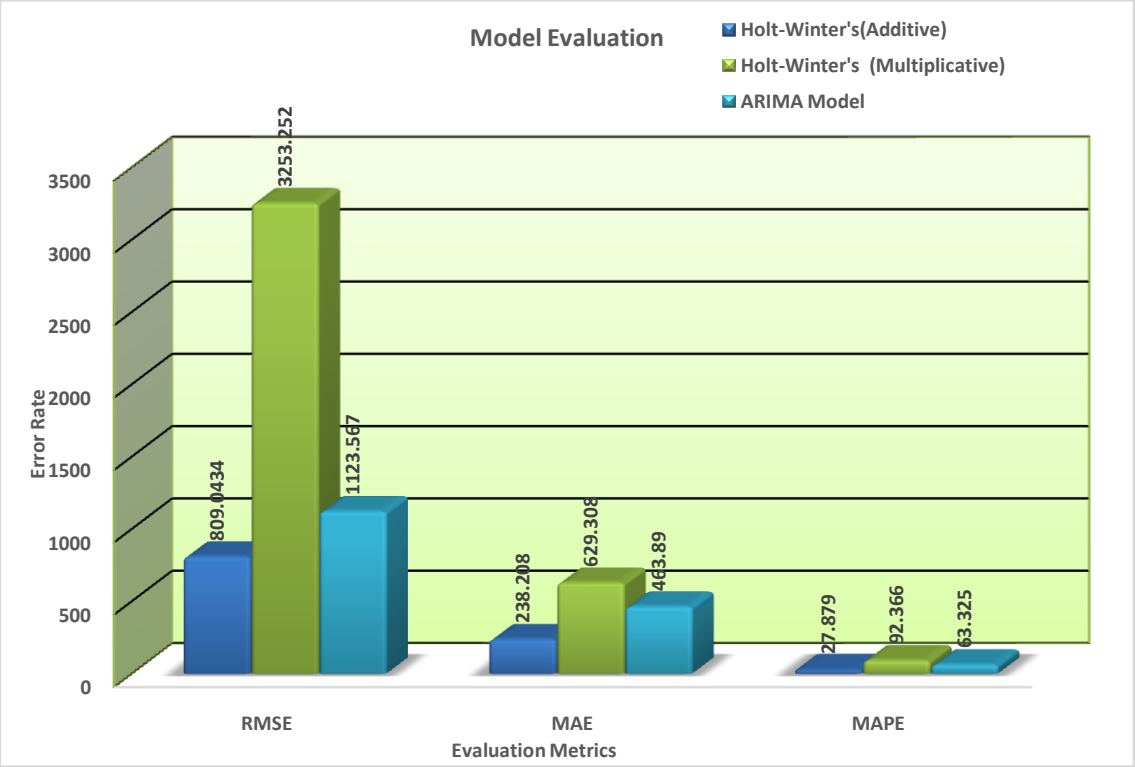


Fig.15. Model Fitting

VI. CONCLUSION

In this paper, empirical study on CBS data set is performed. The results show that Holt- Winters Additive model performs better for the above dataset which has more seasonal fluctuations. For Big Data, it is better to parallelize Holt-Winters (HW) model. The Holt-Winters (HW) model is iterative in nature, it is difficult to parallelize at task level. But it can be achieved at the data level by partitioning the data. In future, a parallel approach for Time Series models can be proposed to handle the Big Data efficiently and effectively.

REFERENCES

[1] R. K. Agrawal, "An Introductory Study on Time SeriesModeling and Forecasting",LAP Lambert Academic Publishing, Germany, pp 1-67, 2013.
[2] Hipel K.W. and McLeod A.I., "Time Series Modelling of Water Resources and Environmental Systems", 2005.
[3] G.P. Zhang, "A neural network ensemble method with jittered training data for Time Series forecasting", Information Sciences, pp 5329-5346, Volume 177, Issue 23, 2007.
[4] Rob J. Hyndman and YeasminKhandakar, "Automatic Time Series Forecasting: The forecast Package for R", Journal of Statistical Software (JSS),pp 1-22, Volume 27, Issue 03, 2014.
[5] S.Hamm, "How Big Data can Boost Weather Forecasting", 2013.
[6] LiljanaFerbarTratar, "Improved Holt Winters Method:A Case of Overnight stays of Tourists in Republic of Slovenia", Economic and Business Review , 2013.
[5] Aditi Jain, ManjuKaushik, "Performance Optimization in Big Data Predictive Analytics",International Journal of Advanced Research in Computer Science and Software Engineering(IJARCSSE), ISSN: 2277 128X , pp 126-129, Volume 04, Issue 08, 2014.
[6] Ekaterina Gonina, AnithaKannan, John Shafer, MihaiBudiu, "Parallelizing large-scale data processing applications with data skew: a case study in product offer matching", International Workshop on MapReduce and its Applications,2011.
[7] Min Chen, Shiwen Mao, Yunhao Liu, "Big Data: A SurveyMobile Networks and Applications, The Journal of Special Issues on Mobility of Systems, ISSN: 1383-469X (Print) 1572-8153 (Online),", pp 171-209, Volume 19, Issue 02,Springer, 2014.
[8] Lei Li, FarzadNoorian, Duncan J.M. Moss, Philip H.W. Leong, "Rolling Window Time Series Prediction Using MapReduce", 15th International Conference on Information Reuse and Integration (IEEE IRI 2014),ISBN: 978-1-4799-5879-5,pp 1-4, 2006.
[9] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", International Conference on Communication, Information and Computing Technology(ICCICT),ISBN : 978-1-4577-2078-9, Oct 19-20, 2012.
[10] Dilpreet Singh and Chandan K Reddy, " A survey on platforms of Big Data Analytics", Journal of Big Data, Springer Open Access, Volume 02, Issue 08,2014.
[11] RashmiRanjanDhall and B.V.A.N.S.S. Prabhakar Rao, " Shrinking the Uncertainty In Online Sales Precdiction With Time Series Analysis", Journal on Soft Computing(ICTACT), pp 869-874,Volume 05, Issue 01, 2014

[12] B. Arputhamary, L.Arockiam, R.ThamaraiSelvi, "Analysis of Prediction Techniques in Time Series for Big Data Using R", International Conference on Engineering Technology and Science(ICETS'15),ISSN 0973-4562, pp 6712-6715, Volume 10, Issue 09, 2015.
[13] B. Arputhamary, L.Arockiam, "Parallel Prediction Model for Big Data using MapReduce Programming Model", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Issue 82, 2015.
[14] B. Arputhamary, L. Arockiam, "Improved Time Series Based Algorithm for Big Data using MapReduce Programming Model", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Issue 85, 2015.
[15] Kanagalakshmi R, "Big Data: Performance Analysis of Vendor and Value Creation through Big Data Analytics", International Journal of Engineering Sciences and Research Technology (IJESRT), ISSN: 2277-9655, Volume 03, Issue 12, 2014, pp 429-434.

