Forecasting of Electricity Demand

1 Introduction

Accurate and reliable forecasting of electricity demand plays a pivotal role in efficient energy management, grid stability, and optimal resource allocation. The increasing complexity of the energy landscape, driven by factors such as evolving consumption patterns, renewable energy integration, and the proliferation of electric vehicles, necessitates robust forecasting techniques to guide strategic decision-making. Time series analysis has emerged as a powerful tool for modeling and predicting electricity demand, leveraging historical data patterns to capture temporal dependencies and forecast future demand trends.

This research paper presents a comprehensive approach to forecasting electricity demand using time series analysis techniques. By exploring and adapting various methodologies, this study aims to enhance the accuracy, flexibility, and scalability of demand forecasting models. The proposed approach also tries to find out long-term trends and factors like seasonal variations in the data.

The significance of accurate electricity demand forecasting cannot be overstated. Utilities and power grid operators rely on precise predictions to optimize power generation, transmission, and distribution, thereby reducing costs and avoiding supply-demand imbalances. Additionally, policymakers and regulators can benefit from reliable forecasts to devise energy policies, set pricing mechanisms, and promote sustainable energy practices. Furthermore, industrial and commercial sectors can make informed decisions regarding production schedules, capacity planning, and energy procurement, resulting in improved operational efficiency and cost savings.

To achieve these goals, this research paper will explore various time series analysis techniques, such as Seasonal Naive, Autoregressive integrated moving average (ARIMA) and exponential smoothing methods (ETS). The performance of these models will be evaluated and compared using historical electricity demand data from different time periods.

The research findings presented in this paper aim to contribute to the existing body of knowledge in electricity demand forecasting, offering valuable insights and practical guidance for energy stakeholders and researchers alike. By improving the accuracy and reliability of these forecasts, it is anticipated that decision-makers can make more informed choices to meet future energy demands, minimize environmental impact, and ensure the sustainable development of the global energy sector.

Keywords: electricity demand forecasting, time series analysis, ARIMA, ETS and AIC

2 Data Description and Summary Statistics

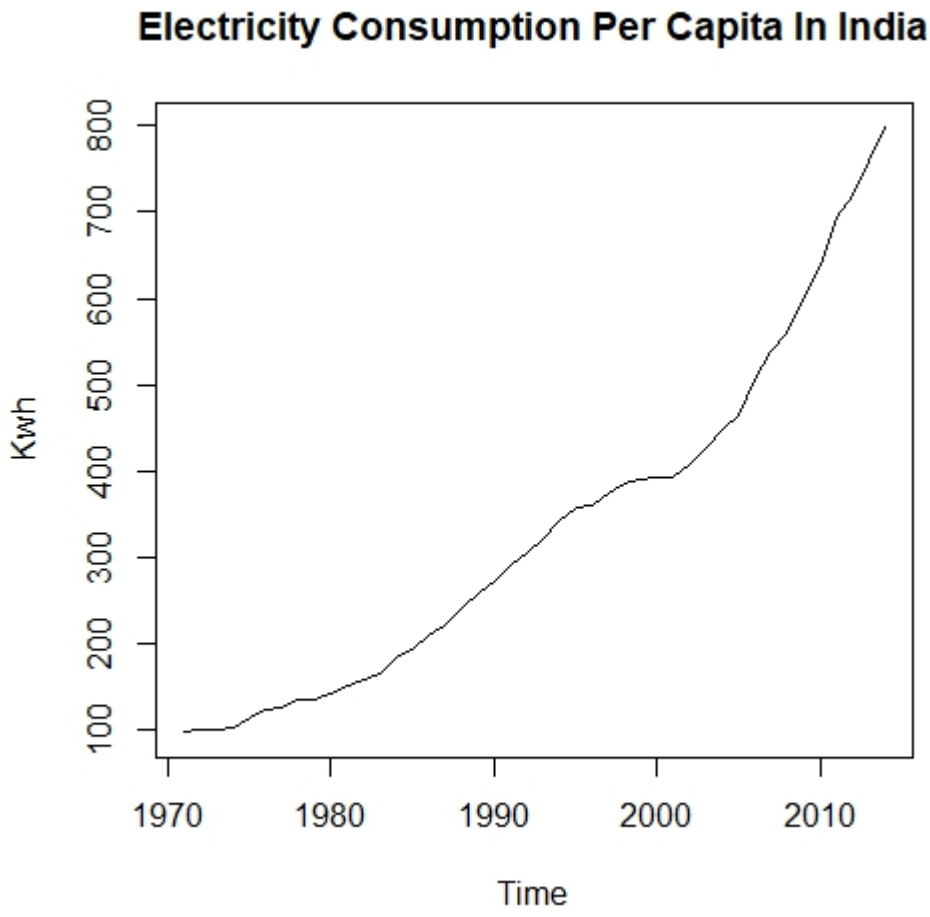
2.1 Data Source and Collection:

Drawing upon the Official World Bank Website, we diligently procure the Electric Power Consumption data for India. This invaluable dataset encompasses a vast array of annual Electric Power Consumption Per Unit Capita values, meticulously recorded from 1960 to 2014. The link to the data can be found here. The file in csv format contains data for several countries; however in this paper, we choose only the Indian data and try to analyse it and discover the hidden information that this data has to tell.

Here we will try to find out various components present in our time-series data such as trend, seasonality and correlation.

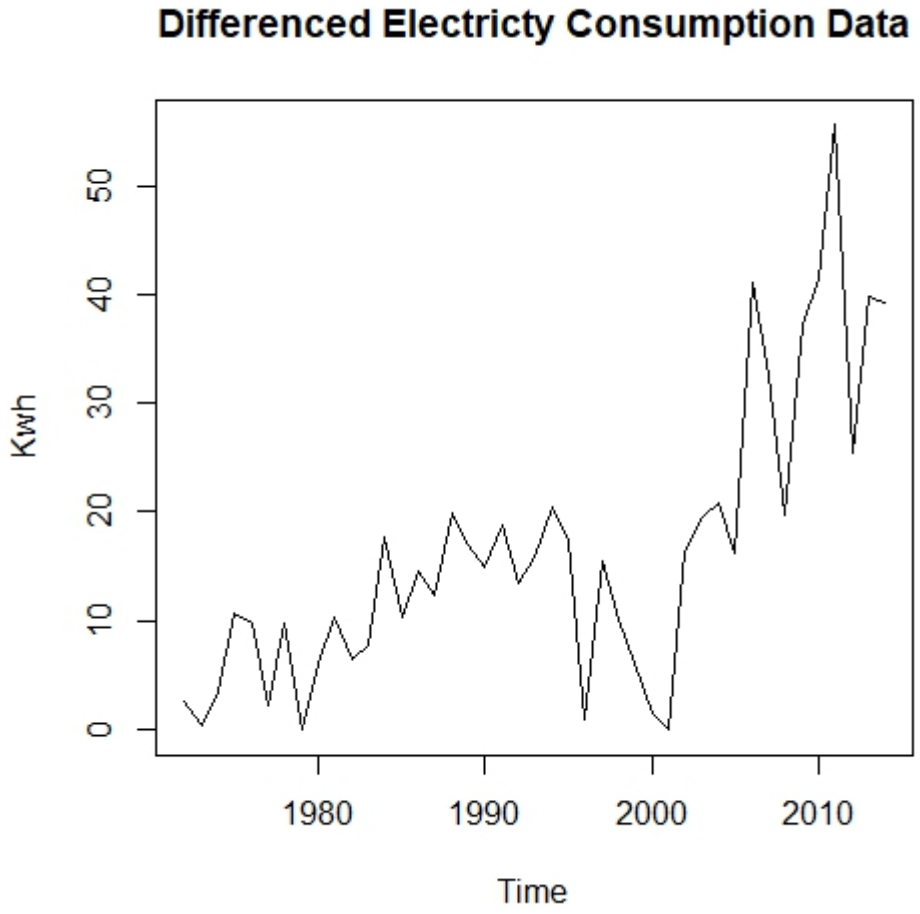
2.2 Data Analysis:

• **Trend Analysis:** The data contains a strong-trend component. We can easily observe the trend in the data from it's time series plot.



The strong trend component covers the many underlying features of the data. We need to remove the trend from the data to view those features. There are many methods to estimate and remove this trend such as Linear Regression, Differencing Technique etc. However, we have chosen to use the differencing technique of trend removal as differencing can help stabilise the mean of a time series by removing changes in the level of a time series.

We again make a time series plot of the differenced data. We refer to this as data at lag 1 as we have used the *first difference* operation on the data. As soon as the trend is removed, all underlying features of the data appears before us. We can easily identify that the smoothness of the actual time plot has suddenly converted into spikes.

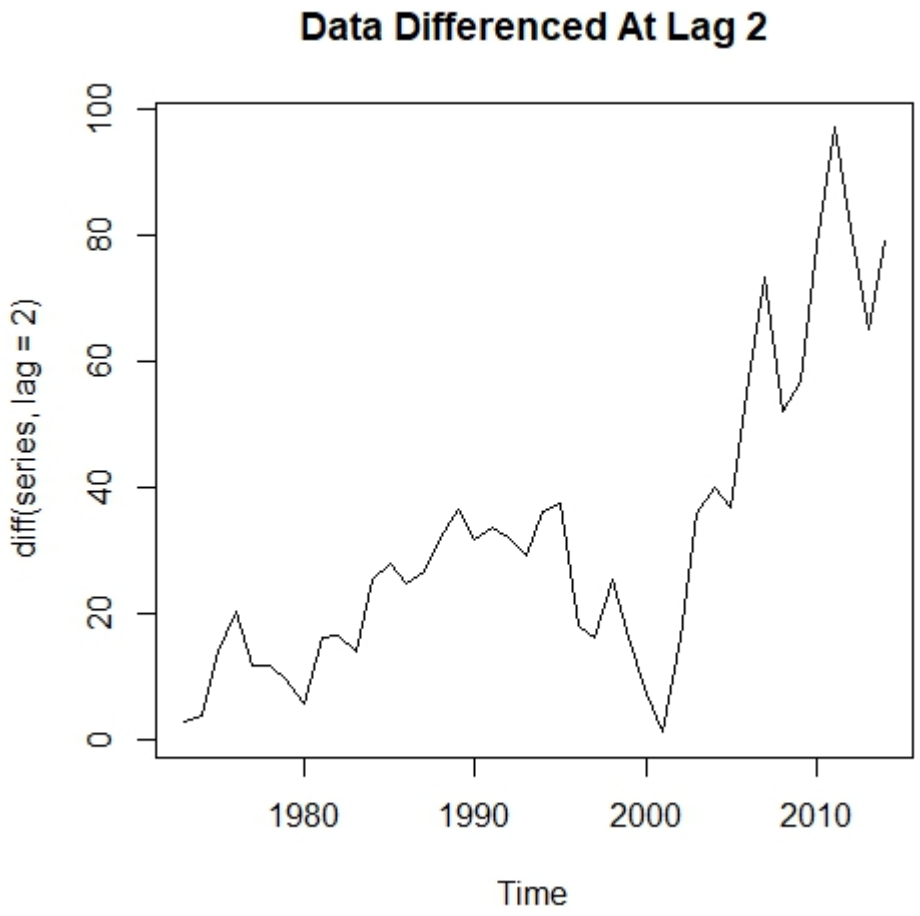


Is the trend really removed now? This question becomes important as the trend component might be dependent on time; and in such cases first difference won't be able to remove the trend completely. We have to difference the data again, i.e, differencing at lag 2 in such cases.

Let us consider the time-series with a time-dependent trend:-

$$x_t = \alpha + \mu t + x_{t-1} + \epsilon_t$$
$$\nabla x_t = \mu + \nabla x_{t-1} + \epsilon_t$$
$$\nabla^2 x_t = \nabla^2 x_{t-1} + \epsilon_t$$

However, we stop differencing further at lag=2 and don't continue differencing the data further. The reason for the same will be discussed later in this paper.



• **Seasonality Analysis:** After trend removal, our next task is to examine the data for any seasonality component, if present. The electricity consumption data should get influenced by the various seasons in a year and thus must have seasonality component. However, the data available to us is yearly and not a dat recorded on daily basis. So, it has not been possible to analyse the data for seasonality component.

•**Correlation Analysis:** The next step in our analysis is to examine the data for Autocorrelation(ACF) and Partial Autocorrelation Components(PACF). The Autocorrelation function of a time series, $\rho(s,t)$, at two different time instances s and t is defined as:-

$$\rho(s,t) = \frac{\gamma(s,t)}{\sigma_s \sigma_t}$$

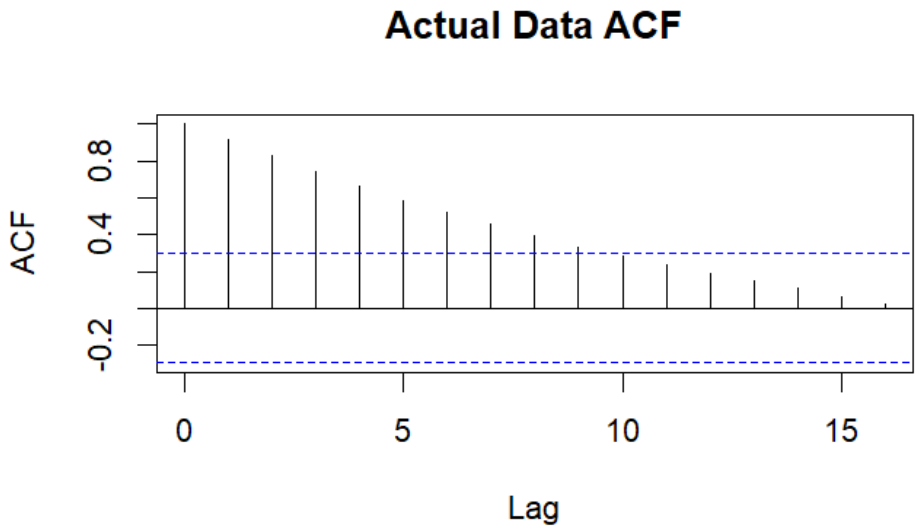
where $\gamma(s,t)$ refers to the autocovariance function of a time-series and σ_t refers to the standard error at time t. When s=t, the autocovariance function becomes variance.

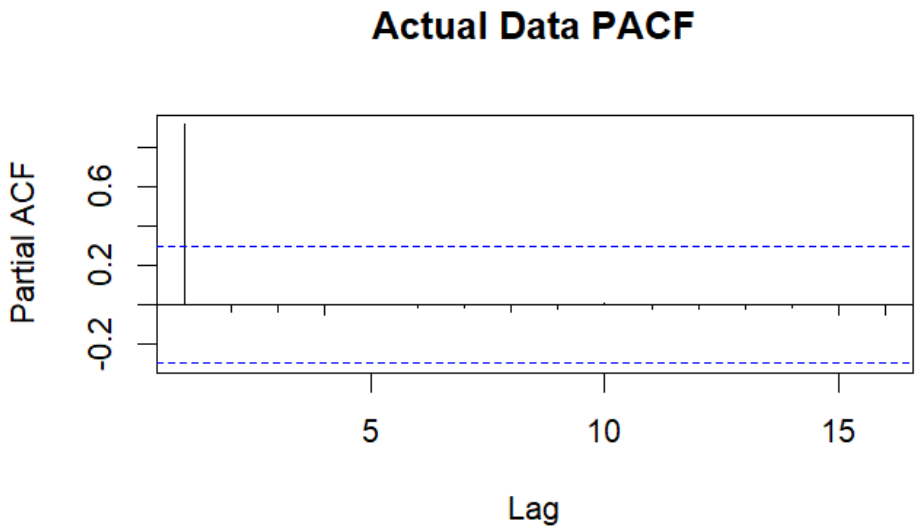
Now let us consider a time-series of the form:-

$$x_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + ... + \beta_{h-1} x_{t+1}$$

The Partial Autocorrelation function(PACF) is broadly defined as the correlation between x_{t+h} and x_t with the linear dependence of $\{x_{t+1}, x_{t+2}, ..., x_{t+h-1}\}$ removed. We can estimate these dependencies with the help of various methods such as Linear Regression etc.

The ACF and PACF plots of the data are very useful in determining the order of an ARIMA model during model fitting of the data. These will be discussed later in the paper.





In the next section, we discuss the various time-series models and how we fit them to our data and compare their accuracies.

3 Time Series Methods

3.1 ARIMA process:-

A time series y_t is said to follow an ARIMA process of order (p,d,q) if $\nabla^d y_t$ follows an ARMA (p,q) process. That is:-

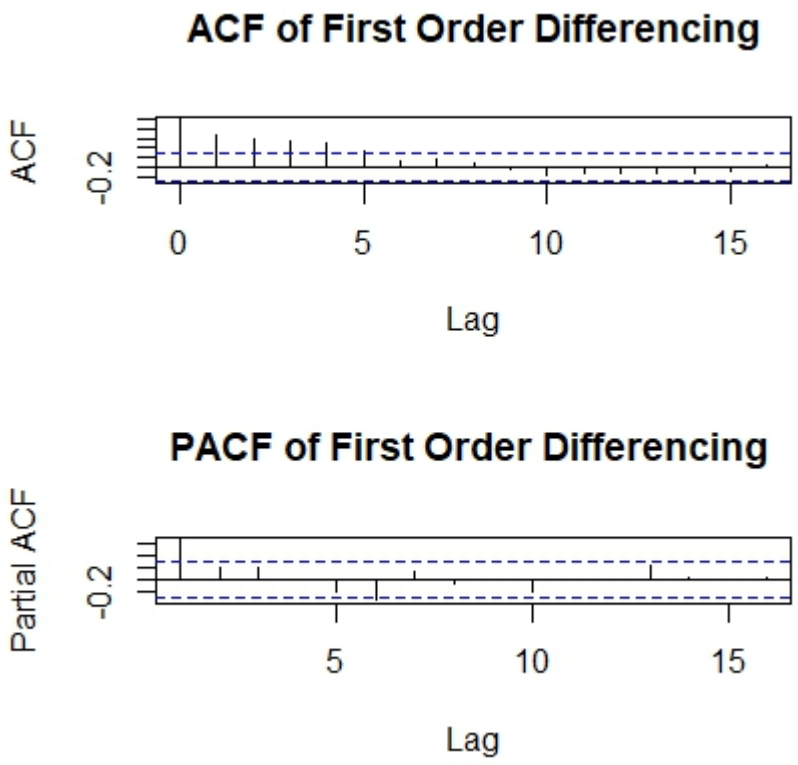
$$\begin{aligned} & y_t^* = ARMA(p,q) \\ where, \quad & y_t^* = \nabla^d y_t \end{aligned}$$

As y_t^* follows an ARMA model, so we can represent it in the form of:-

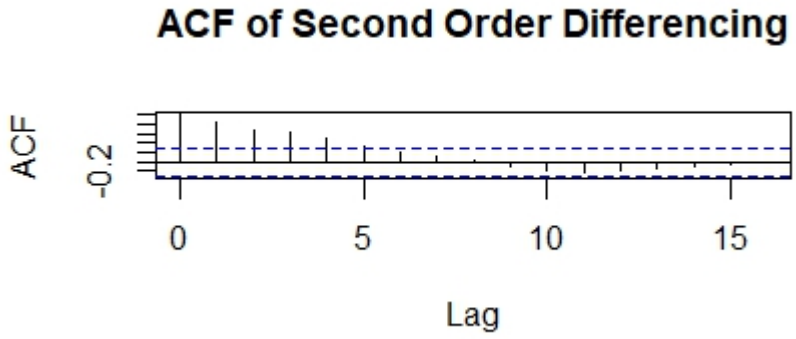
$$y_t^* = \alpha + \phi_1 y_{t-1}^* + ... + \phi_p y_{t-p}^* + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

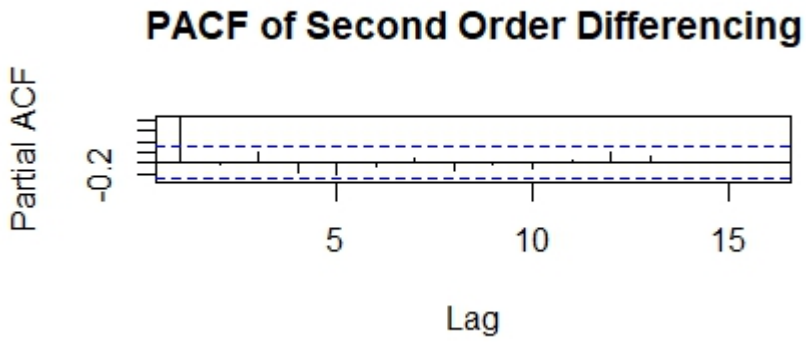
The above equation can also be represented as:-

$$\Phi(L) y_t^* = \alpha + \Theta(L) \epsilon_t$$



To estimate the orders p,d and q, we need to look at the ACF and PACF plots of the time-series data. We have already plotted ACF and PACF plot of the Actual Data. We can clearly see the **tail-off** pattern in the ACF plot. However, as a strong trend component is present in the data, we have to difference and plot the ACF and PACF of the differenced data. So, we plot the ACF and PACF functions after first-order differencing and second-order differencing.





Once we have identified the orders of the ARIMA model, it's now time fo estimation of the paramerers. The perfect ARIMA model that fits our data the best is ARIMA(0,2,1). The model can be expressed in the form:-

$$\nabla^2 y_t = \alpha + \epsilon_t + \theta_1 \epsilon_{t-1}$$

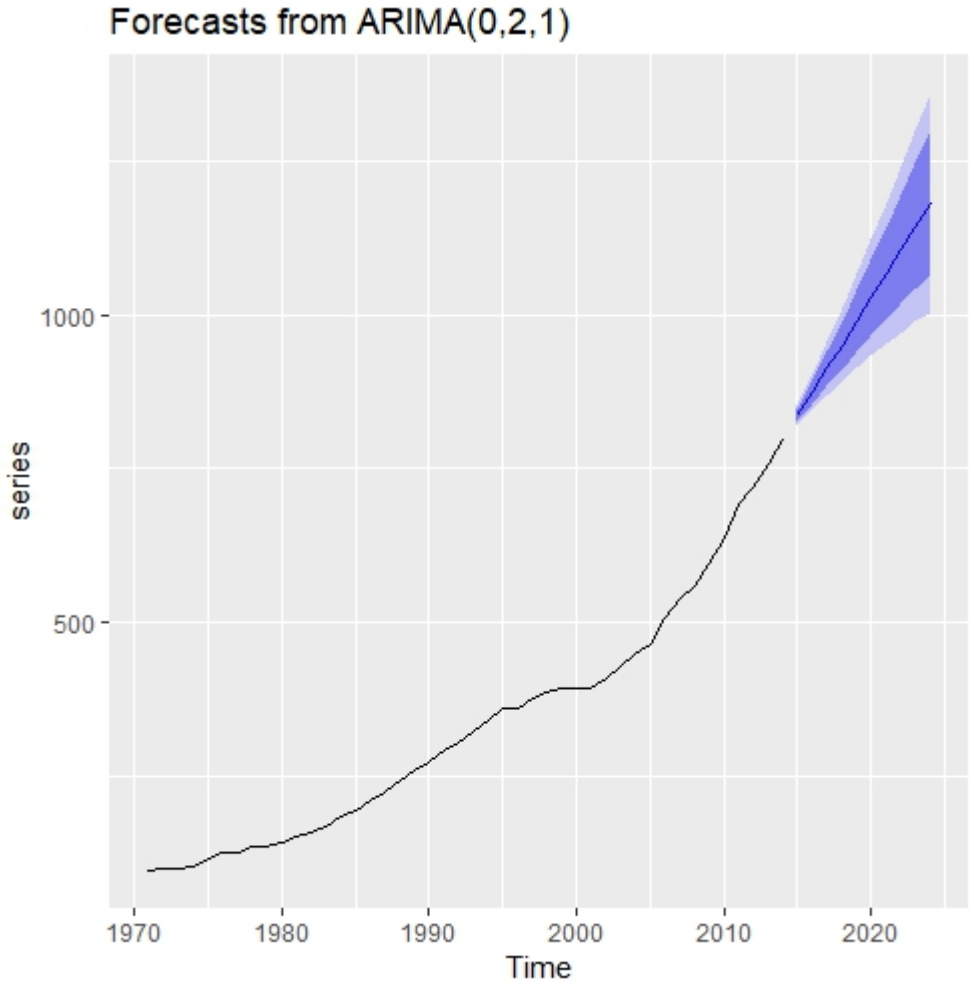
All of these model order determination and parameter estimation is automatically done in R using the *auto.arima()* function which uses a variation of the **Hyndman-Khandakar** algorithm.

The parameter values that we get after estimation is:-

$$\theta_1 = -0.536$$

So our overall model becomes: $$\nabla^2 y_t = \alpha - 0.536\epsilon_{t-1} + \epsilon_t$$

where $\epsilon_t \tilde{i} id N(, 73.21)$ Once our model is ready, we can now use it to forecast the future values of electricity consumption. As we have data available only upto 2014, we use the model to forecast the electricity consumption upto 2020 and a little beyond it.



The above plot shows the forecasted electricity consumption per unit capita a little beyond the year 2020. We have mentioned the different **confidence intervals** of our forecast. The deep blue shaded region denotes the 80% confidence interval, whereas the light blue region denotes the 95% confidence interval.

The strong increasing trend is clearly visible in the plot and the forecast suggests that this electricity consumption is going to increase in the future. The model has a Mean Absolute Percentage error (MAPE) of 1.957 and a corrected AIC of 303.14.

3.2 Exponential Smoothing Methods:

Now we move on to our next model which is known as the Exponential Smoothing Methods. There are three types of methods that come under Exponential Smoothing:-

1. Simple Exponential Smoothing(SES)
2. Double Exponential Smoothing(DES)
3. Triple Exponential Smoothing(TES)

- **Simple Exponential Smoothing:-**

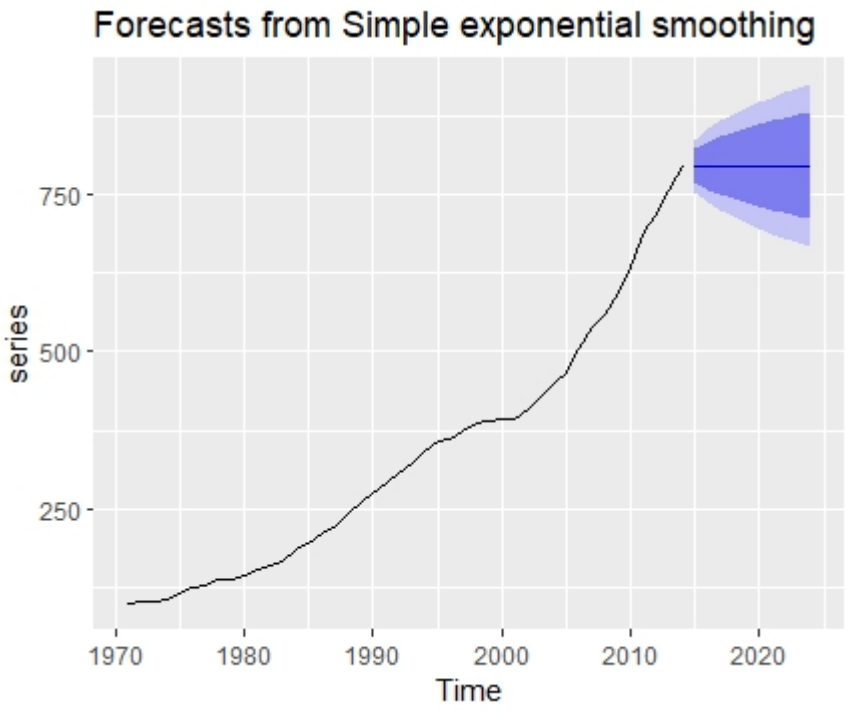
  The **level updating equation** of a Simple Exponential Smoothing process at time t can be written as:-

$$\begin{aligned} L_t &= \alpha y_t + (1-\alpha)L_{t-1} \\ &= \alpha y_t + (1-\alpha)\{\alpha y_{t-1} + (1-\alpha)L_{t-2}\} \\ &. \\ &. \\ &. \\ L_t &= \sum_{i=0}^{t-1} \alpha(1-\alpha)^i y_{t-1} + (1-\alpha)^t L_0 \end{aligned}$$

  And the **Forecasting Equation** at h time intervals ahead of time for the same is:-

$$F_{t+h} = L_t$$

  We have fitted the Simple Exponential Smoothing method in our data and forecasted the results using the Forecast equation:-



  We can clearly see that the SES method provides us with a forecast as a line of zero-slope. For our model, the y-intercept of the forecast line is 750. We have also marked the 80% and 95% confidence intervals in deep blue and light blue shades respectively.

  We have chosen $\alpha = 0.999$ for the above model as it was producing the least error and AIC. The corrected AIC of this model is 436.884. So, according to this model, the future electricity consumption per unit capita will remain constant over the years.

- **Double Exponential Smoothing (Holt's Method):-**

  Before we continue discussing model-fitting, we first discuss why did we choose Double Exponential Smoothing and not Triple Exponential Smoothing.

  The Double Exponential Smoothing method is suitable for time series with trend but no seasonality. And the Triple Exponential Smoothing method is suitable for time series data having both trend and seasonality components.

  The data available with us should have a seasonality component in it. However, the data is a time series with the yearly consumption provided; and not the daily/monthly. So, it was impossible to analyse the data for any seasonality component as we discussed earlier in this paper.

  Now we move on with fitting the Double Exponential Smoothing Model to this data. Similar to the SES method, the DES also has a level updating and a forecasting equation. In addition to those, the DES method also has a **trend updating equation** to include the trend component within the forecast.

  The level updating equation of DES is defined as:-

  $$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad where \quad 0 \leq \alpha \leq 1$$

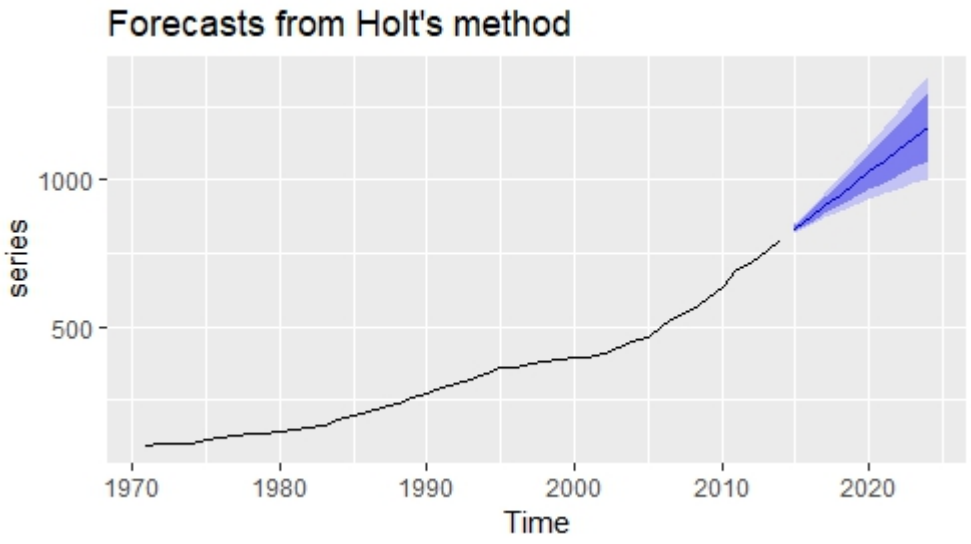
  The trend updating equation of DES is defined as:-

  $$T_t = (1-\beta)T_t + \beta(L_t - L_{t-1}) \qquad where \quad 0 \leq \beta \leq 1$$

  And the forecast equation of DES is defined as:-

  $$F_{t+h} = L_t + hT_t \qquad for \quad h = 0, 1, 2, ...$$

  The Double Exponential Smoothing Model after we fit the electricty consumption per unit capita data in to it looks like this:-



  The deep blue and the light blue shade again represents the 80% and 95% confidence interval as previous. The smoothing constants α and β was chosen such that it minimizes the AIC and the MAPE. The corrected AIC for this model is 363.436 and the MAPE is 1.888.

4 Data Analysis Results And Best Model Selection

Till now, we have analysed the data and discussed various features of the data such as trend, seasonality and correlation. We have also fitted three different time series models to the electricity consumption per capita data. But how do we decide which model is going to forecast the electricity consumption most accurately?

We can calculate the error in our forecast only when we have the actual data available with us. But if we wait for the actual data, we can never choose the best model and thus the whole purpose of forecasting will not be fulfilled.

However, statistics allows us to have an **estimate of the error** in forecasting. There are many methods to do it such as calculation of training error, Akaike Information Criteria(AIC), Bayesian Information Criteria(BIC), k fold cross-validation etc.

If we choose training error as the criteria for model selection, then probably training error is not a good estimate of the test error. In fact if we continue introducing more complex models, we will observe that the training error will continue decreasing. However, in this way, there is a fair possibility that we overfit our model. And it has been found that an overfitted model, though it performs greatly on the training error, it will perform very poorly when forecasting the data. And so, the test error will generally increase after a certain point of model complexity.

Also we can choose cross-validation neither as our criteria for model selection; because applying k fold cross-validation on a time series data becomes excessively complex due to the fundamental nature of calculation of k fold cross-validation.

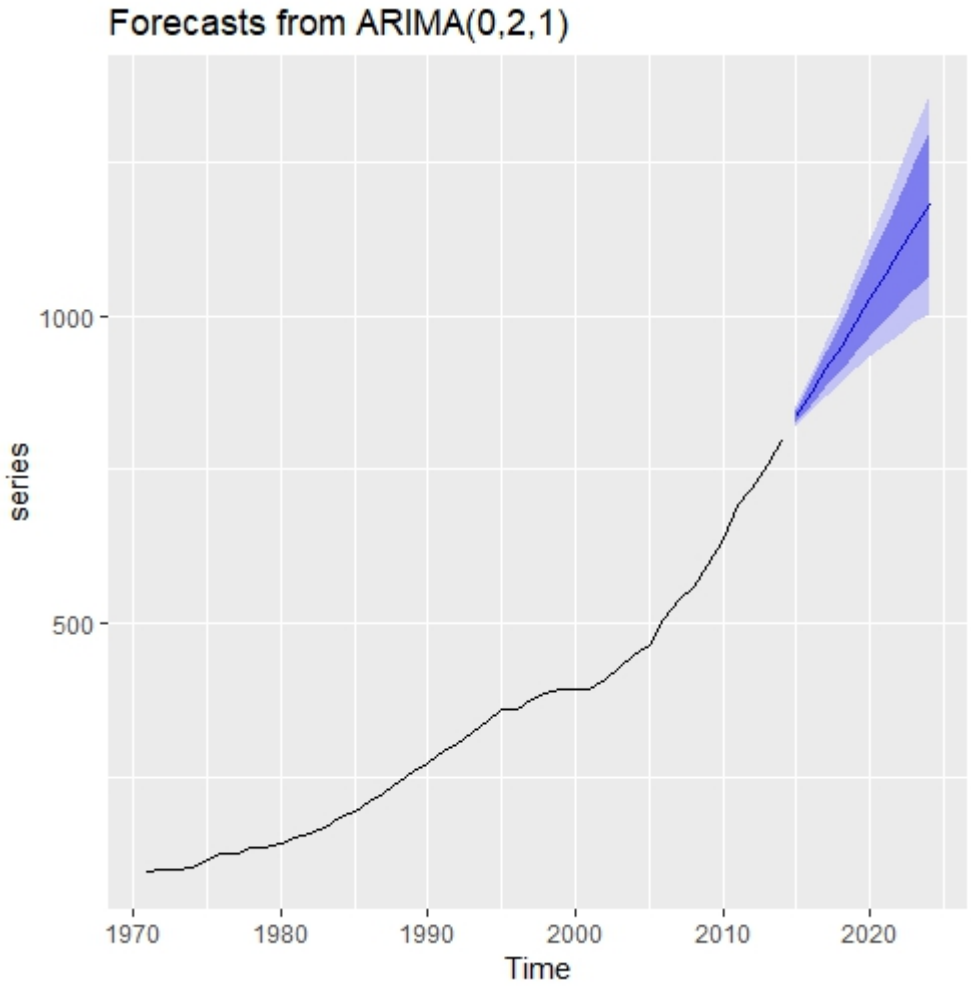
So **we choose Akaike Information Criteria(AIC) as our criteria for model selection**. AIC is just an adjustment made to the training error, which then can be used to estimate the test error. However, in this paper, we will be using the bias-corrected form of AIC as it is more accurate in estimation of the test error.

As AIC is an estimate of our test error, so naturally **the best model will have the least test error or least AIC**. Now when we go to compare our models based on their corrected AIC values, we will obeserve that Simple Exponential Smoothing has an AIC of 436.884, Holt's method has an AIC of 363.436. And finally, our ARIMA model has an AIC of 303.14.

Based on the above three AIC values, we come to the conclusion that **our ARIMA(0,2,1) model is the best model among the three and thus is able to forecast the electricity consumption per unit capita in India most accurately**. However, we must remember that all the models discussed above, including the ARIMA(0,2,1) model are to be used only for short-term forecasting purposes. They can't be used for long-term forecasting as we won't be obtaining satisfactory forecasts in those situations.

5 Conclusion

We now come to the conclusion that ARIMA(0,2,1) model is able to forecast the data the most accurately. But before moving forward, we must remember that the data available to us was only upto 2014. So, data only upto 2014 was considered for making these forecasts.



When we have a closer look at our forecast, we observe that the electricity consumption in India per unit capita is likely to maintain it's increasing trend in the future and is most likely to cross 1000KWh by the year 2020. And by 2025, electricity consumption is likely to reach a little lesser than 1250KWh by the year 2025.

6 R Codes

We have used R for our calculation purposes including model-fitting, making forecasts, making time-series plot of the data and also calculation of AIC values. R is a popular open-source statistical software which enables us to efficiently use the computing power of modern-day computers to compute the complex calculations involved in statistics and other fields.

The links to the R codes used to fit models and make forecasts in this paper can be found here.