MSCI 718 - Mini-Project

GLOBAL AVERAGE LAND TEMPERATURE USING ARIMA

As many people believe climate change is a myth according to dodgy science, majority argue that it is the biggest threat in this world. The whole experiment was performed on the data collected by technicians using mercury thermometers from the 1940s, till data collection using electronic thermometers began in the 1980s.

OBJECTIVE:
The project is all about performing Exploratory Data Analysis on time series data of land temperatures collected globally between 1750s and 2015, performing time series decomposition, fitting a time series model and extracting forecast values of the data over the next 50 years using **ARIMA**.

HYPOTHESIS: Given the current trend, there will be an increase in the average land temperature observed over the next few years.

DATA SUMMARY: The dataset consists of **3192 observations** having **9 variables**. The first variable "**dt**" is a factor variable, which gives the date starting from 1750. The next variable is "**Land Average Temperature**", with numeric data type, denotes the global average land temperature in Celsius, while the third variable "**Land Average Temperature Uncertainty**" is also of the numeric data type, which shows the 95% confidence interval around the average. The remaining variables Land Max Temperature, Land Max Temperature Uncertainty, Land Min Temperature, Land Min Temperature Uncertainty, Land And Ocean Average Temperature and Land And Ocean Average Temperature Uncertainty are all of the type numeric data type, but all the observations under these variables are NULL values. LandAndOceanAverageTemperature is not a good parameter as the individual influence of land and ocean could not be checked. Therefore, we are focusing just on the **Land Average Temperature** dataset.

For this project we are working on the **Global Land average temperature** for forecasting the average temperatures for 50 years starting 2016. Since the data is collected over the time as a sequence of time gaps, **ARIMA**, an **Auto Regressive Integrated Moving Average** model has been used to perform time-series analysis and forecast the future results. This model will help us understand the trend in the data to find patterns and predict future results. These results can be extremely beneficial for the farming industry and scientists for research purposes.

In our model, we worked with **auto ARIMA** and manually set values of **p,d** and **q**. The reasons for the same have been discussed below.

DATA CLEANING:

- The data set was tidy with each row, column and cell having one value only.
- There were missing values for LandAverageTemperature. Since we are checking the trend of the past 25 years only, we encountered no missing value in that subset.

- The data was also checked for outliers. Since we are working on time series data, therefore, instead of removing them, the values were imputed. **tsclean() function** was used to replace the outliers such that the series is smoothed out.

ASSUMPTIONS:

- Data should be stationary
- Data should be univariate. This condition has been satisfied as we are working on one variable only.

Visualizing the data for the past 100 years:

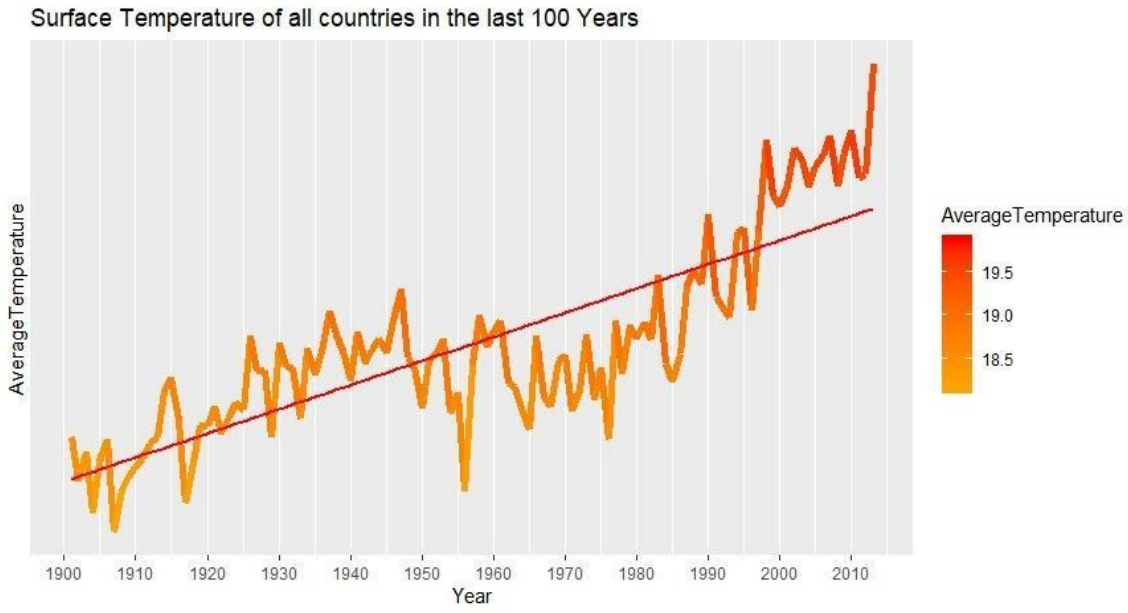

Fig: Surface Temperature of All Countries In The Last 100 Years

From the graph above, an increasing trend over the average temperature of the land can be inferred. A steep increase around 1995 can be observed. After that, the fluctuations are nominal with a sudden shoot around 2013. To understand how the land temperatures would change in the upcoming years, the results of the past 25 years were taken into consideration. Beyond that point, the temperatures have been pretty low compared to results after 1990.

DATA EXAMINATION:

- The data subset of 25 years was cleaned and imputed for outliers and null values.
- Log transformation was implemented to settle the development pattern.

Since the data contains the average temperatures for each month of the year, we observe a lot of fluctuations despite transformation. To have a better idea of the trend, the data is smoothed using the moving averages over a monthly basis.

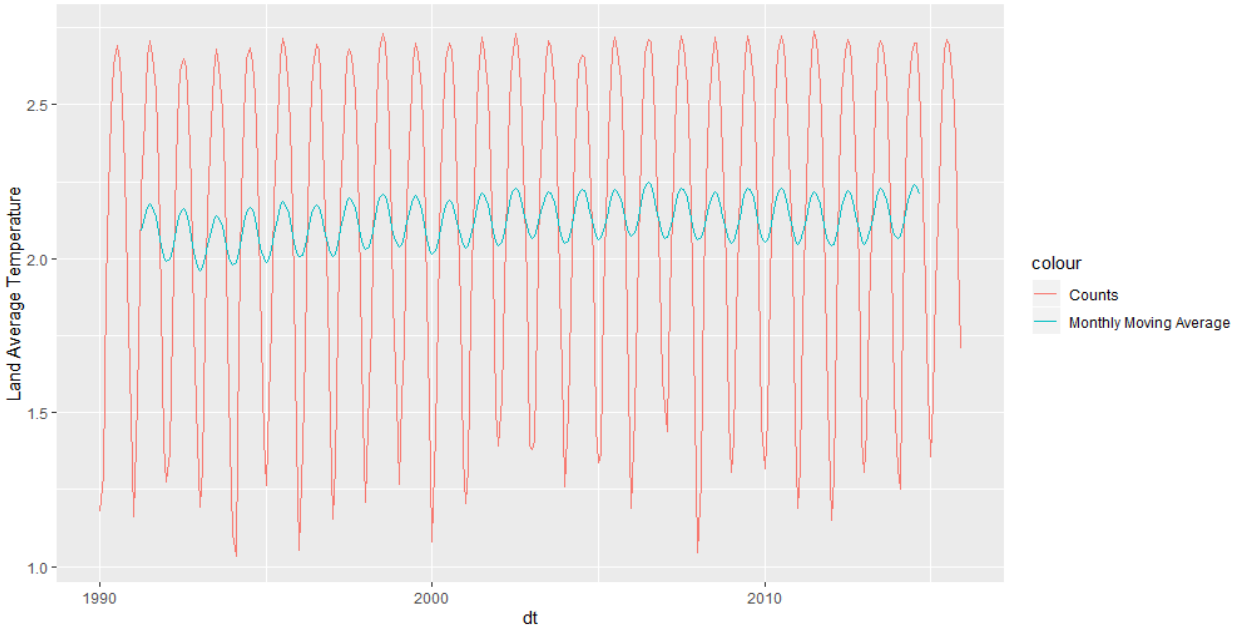

Fig: Moving Average to Average the Monthly Value

Once the data is visualized, it is decomposed to observe cycle, trends and seasons.

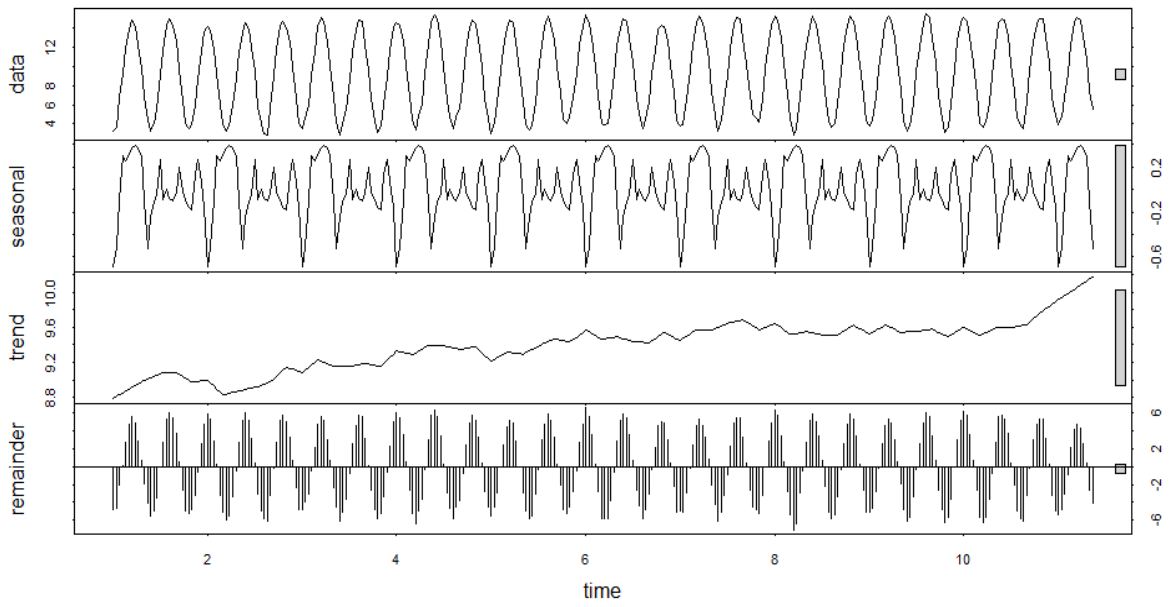

Fig: Seasonal Fluctuations and Trend

We can observe seasonal fluctuations in the data which are periodic over the years. The trend shows an increase in temperature over the years. Even the data is cyclic and follows the same patterns throughout. Since the data is seasonal, therefore the seasonal component is calculated and subtracted from the original data to deseasonalize it using **stl() function**.

STATIONARY DATA:

For data to be stationary, the mean, variance and autocovariance should be independent of time. To check for the same, an augmented Dickey-Fuller test is implemented which assumes the null hypothesis as non-stationary. Since the p-value=0.01, therefore null hypothesis can be rejected claiming that the data is stationary. But if the level of significance is 99%, then it would have to be further transformed using differencing to make the series stationary.

AUTOCORRELATION:

To further establish the fact that the data is stationary, ACF plots can be used for verification. ACF plot shows the correlation between time series and its lags. It is useful for determining the order of differencing and moving averages(q) model. Also, Partial ACF, displays the correlation between variable and the lags that have not been explained by previous lags. They help in determining the order of autoregression(p) model.

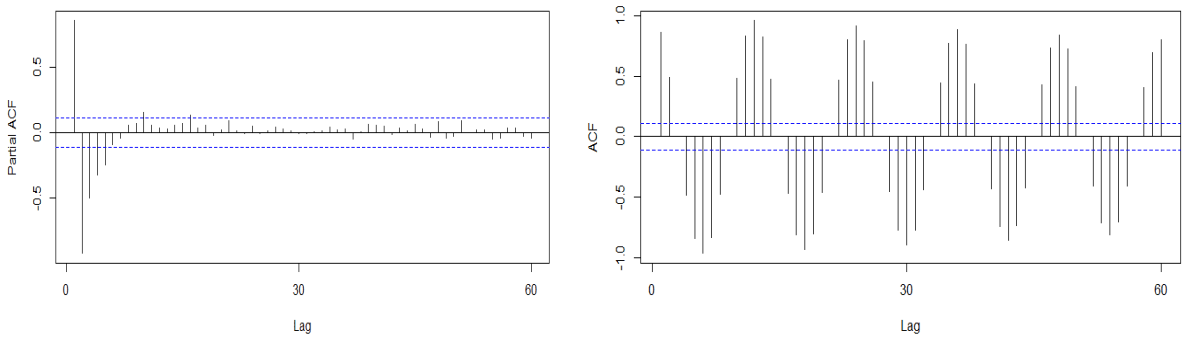

Fig: Partial Autocorrelation and Autocorrelation to Determine If A Series Is Stationary

These plots help determine the values of p, d, q by checking the lags on the plot. The blue lines seen on the plots are 95% significance boundaries. We can see above in PACF plot, there are 2 lags close to 0 that are falling outside the boundary.

FITTING THE MODEL:

The process of manually determining the values of p, d and q can be tiresome. To resolve that, auto.arima() can be used.

Auto ARIMA automatically chooses the values of p, d and q optimally that best fits the criteria. Also, the seasonal component can be set as FALSE. But when the auto.arima() model was evaluated, the residuals are plotted and showed patterns at regular lag intervals. This suggests that modifications need to be made to the values p, d and q.

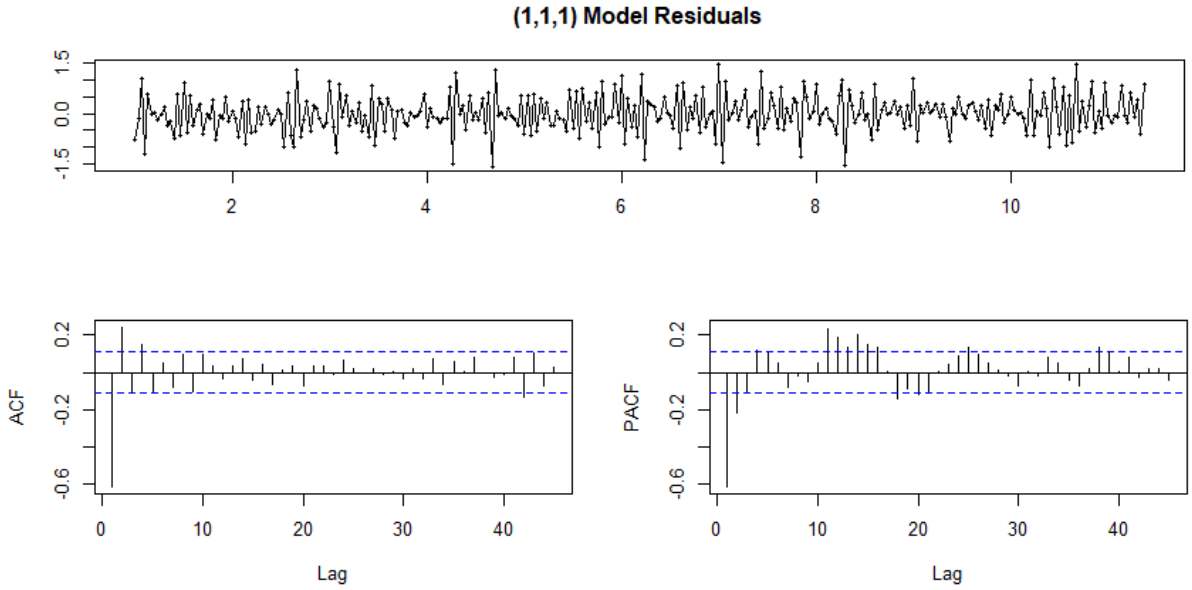

Fig: Autocorrelation Graphs

From the graphs, we observe a lag close to 0. All the points where residuals cut the 95% significance boundary are tested. After multiple tries, 8,1 and 14 best served the model. The main goal is to minimize the values of AIC and BIC for better results. In our case, a decrease in the value of AIC was observed.

When the residuals were plotted, no more patterns in the ACF and PACF were observed suggesting that this model is better than the previous one.

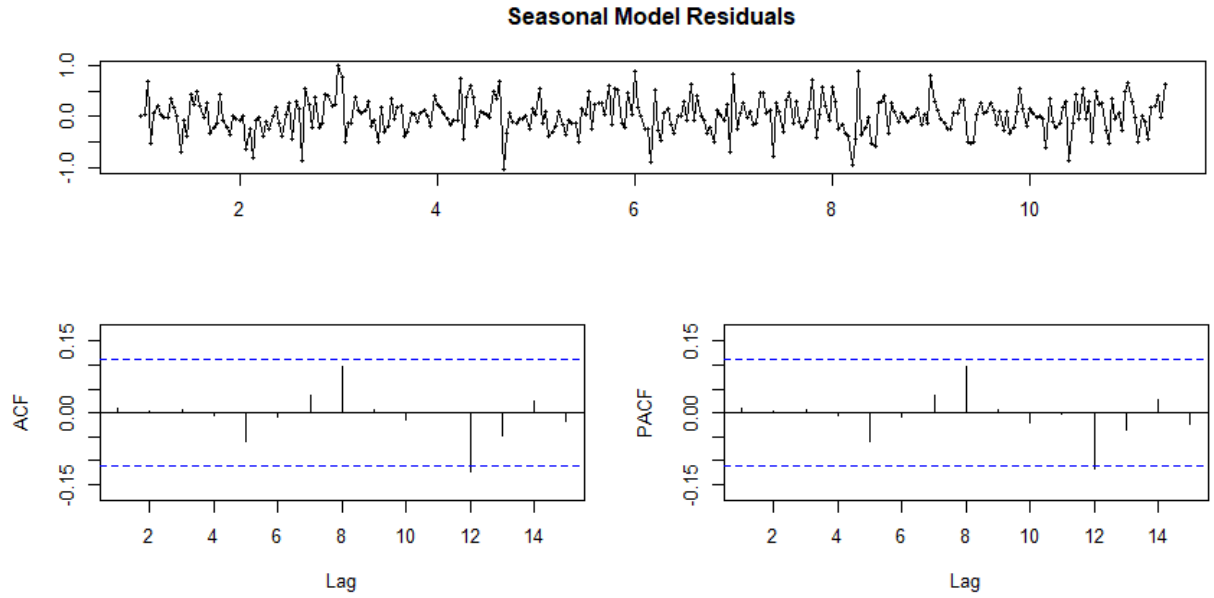

Fig: Seasonal Model Residuals

Data was also checked for the presence of seasonal components using auto.arima() function with seasonal components as true.

Predicted values were generated for both types of results - with seasonal components and without seasonal components.

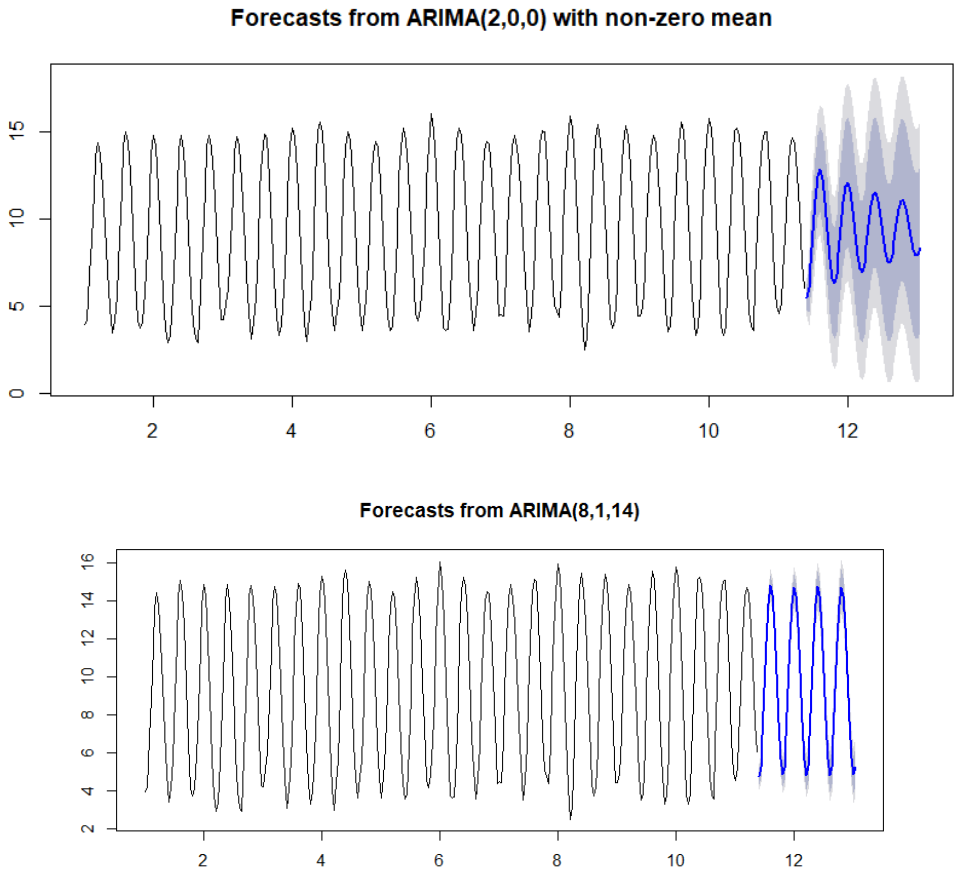

Fig: Prediction by ARIMA

CONCLUSION:

The objective of the study was to predict the average temperature of the land for the next 50 years. Taking seasonal components into account, results were generated for both the cases. This represents seasonality and trend in the temperature data. From the graph, it can be said that the temperature in the upcoming years will reciprocate the path of previous years. Also, when plotted with the past data, we see an increasing trend which satisfies our hypothesis.

FUTURE WORK:

This forecast has been done with only one variable. For future work, other features can be taken into consideration for better and more accurate results.

REFERENCES:

[1] Andy Field, Jeremy Miles, and Zoë Field (2012). *Discovering Statistics Using R. SAGE Publications.*
[2] Lecture Slides - L26 - Time Series 1 & L27 - Time Series 2
[3] Harvey, A. C. (1993). *Time Series Models*. 2nd Edition. Harvester Wheatsheaf. Sections 3.3 and 4.4.
[4] https://otexts.com/fpp2/arima-r.html
[5] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/arima.html
[6] https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/
[7]Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer, New York. Sections 3.3 and 8.3