Monthly Average Temperature Forecasting using ARIMA Model

Introduction

Aim of this project is to construct a time series model which simply forecasts weather. Data used for this purpose is “Global Land Temperatures by Country” provided Berkeley Earth, a non-profit organization in the USA which focuses on environmental data analysis.

This data contains 577462 data points and four variables: 'Date', 'Average temperature', 'Average Temperature Uncertainty' and 'Country'. This has data from the year 1743 to 2013. Out of the 4 variables, the variable of interest to this analysis is 'Average temperature' and of course the ‘date’. ARIMA model is being used in this analysis to construct the forecasting model.

Literature Review

Weather forecasting is something that has touched all of our lives at one point or the other if not daily. Most of the time, it becomes a humorous topic for conversation because of the so-called inaccuracies in the predictions without actually realizing the advancements this sector has had.

MJC Hu and Halbert E. Root introduced ADALINE, an adaptive data processing system (ANN) (MJC Hu, 1964) for weather forecasting in 1964 and the sector has come a long way from there. Today, the forecasting systems get their data from tools like doppler radar, Satellites, Radiosondes and Automated surface-observing systems, process these data using systems like AWIPS (Advanced Weather Information Processing Systems) (Anon., 2017) and predict the weather with accuracies beyond our imagination. Foe example, the MET office, UK govt. predicts 92.5% of the next day temperature with 2-degree Celsius accuracy and they predict 92% of the next day wind speed within 5 knots accuracy (Anon., 2023).

ARIMA is one of the classic methods of time series forecasting models. ARIMA weather forecasting model gives reasonably comparable results to even that of ANFIS (Adaptive Network Based Fuzzy Inference System) (Tektaş, 2010). In this project this model has been used for learning purpose as well since it gives a better understanding about the process and the data compared to other methodologies in which the process employed is more or less a black-box.

Results and Discussion

First step of the analysis was to load and understand the data. In this step, it is understood that there are plenty of missing values in the dataset, which is rectified in the cleaning process.

Second step was cleaning of the data where the data was first filtered as per the requirement (1900-2013 and for countries USA, UK, Brazil, Kenya and India). After this step, it was understood that, in the data of interest, missing values are ignorable quantity (one missing for each country). In this same step, the data was split based on the countries of focus in this study.

In the next step, time series components for the data were analysed namely, Trend and Seasonality. There was no visible trend in the data for the countries USA, UK and India, while there was an upward trend for Brazil and Kenya. Seasonality in average monthly temperature was present in all five countries. More details on seasonality can be found in the notes section of the “ipynb” and appendix I.

A stationary data has its mean, variance and covariance constant with time. Data needs to be stationary to perform any form of time series analysis, else, the data needs to be transformed to stationary before proceeding further (Pandian, 2023). Stationarity can be checked via Dickey-Fuller test (Appendix II). According to the test, data for Brazil and Kenya are non-stationary and the rest is stationary. First step into transforming the data is differencing method (Appendix III). After the first order differencing, ADF test was conducted again, and the data turned out to be stationary.

In the next step, ACF and PACF plots were plot for all countries (for Brazil and Kenya, the new first order differenced data was used) to find out the input values p and q for model fitting (Appendix IV). PACF gives the p value and ACF gives the q value (TrainDataHub, 2021).

Using these input values, ARIMA model was fit into all five data sets. For this, the data was split into Train (1900-2010) and Test (2011-2013). Comparing the forecasted values to the actual values gave an RMSE value as shown in the table below. Since, the RMSE values have the same unit of measurement as the variable of interest (here, AverageTemperature), it is evident from the table below that the constructed model does a good job predicting the land temperature. As mentioned before, MET UK Gov predicts 92.5% of the next day temperature within 2-degree Celsius accuracy and here in this model, it is done with and RMSE of 1.36 degree Celsius.

Country	MSE	RMSE
USA	2.305	1.518
UK	1.853	1.361
Brazil	0.136	0.369
Kenya	0.213	0.461
India	5.776	2.403

Conclusion and Future Recommendations

As mentioned in the previous section, the result of this study, i.e., the ARIMA model for land temperature forecasting does an excellent job, even comparable to the industry standards. But this model is still pervious to the traits in the data hence, the variations in the RMSE values for different countries. One way to rectify this is to take the first order differencing for all the country's datasets. Other models like FB Prophet, XGBoost, ANFIS etc. can be tries to see if they have a better performance.

References

Anon., 2017. *6 tools our meteorologists use to forecast the weather.* [Online] Available at: https://www.noaa.gov/stories/6-tools-our-meteorologists-use-to-forecast-weather [Accessed November 2023].

Anon., 2023. *How accurate are our public forecasts?.* [Online] Available at: https://www.metoffice.gov.uk/about-us/what/accuracy-and-trust/how-accurate-are-our-public-forecasts#:~:text=The%20Met%20Office%27s%20four%20day,are%20correct%20within%205%20knots.
[Accessed November 2023].

MJC Hu, H. E. R., 1964. An Adaptive Data Processing System for Weather Forecasting. *Journal of Applied Meteorology and Climatology,* pp. 513 - 523.

Pandian, S., 2023. *Time Series Analysis and Forecasting | Data-Driven Insights.* [Online] Available at: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/#Methods_to_Check_Stationarity
[Accessed November 2023].

Tektaş, M., 2010. Weather Forecasting Using ANFIS and ARIMA MODELS.. *Environmental Research, Engineering and Management,* 51(1).

TrainDataHub, 2021. *How to Interpret ACF and PACF plots for Identifying AR, MA, ARMA, or ARIMA Models.* [Online]
Available at: https://medium.com/@ooemma83/how-to-interpret-acf-and-pacf-plots-for-identifying-ar-ma-arma-or-arima-models-498717e815b6#:~:text=Look%20for%20tail%20off%20pattern,and%20PACF%20%E2%86%92%20ARMA%20model
[Accessed November 2023].

Appendix

I

Country	Seasonality
USA & UK	Seasonality is similar in the USA and the UK average temperature with lowest temperatures during Nov-Feb period (Winter) and highest average temperatures during June-Sep period (Summer). For the rest of the countries, seasonality is a bit different.
Brazil	In Brazil, summer seems to be from December to March and winter is from May to September. This is because the country is in the southern hemisphere and the weather there would be more or less reverse of that of northern hemisphere.
Kenya	In Kenya, the lowest average temperature is between June and September and the highest is between February and April. This is consistent with the weather in Kenya where the winter runs from July to october and the long rainy season runs from April to June.
India	Seasonality in Indian temperature data is more or less similar to that of the USA and the UK, except for the small dip in temperature during July-September period. This is consistent with the monsoon season during this period in India.

II

ARIMA model can only be fit on a stationary data, i.e. a time series data which has no recurring time dependent features. So, the data needs to be checked for stationarity before proceeding further. This is done via Augmented dickey-Fuller test. This test has a null hypothesis : the time series data is non-stationary. So, if the p-value is less than the critical value (0.05) then we can reject the null hypothesis.

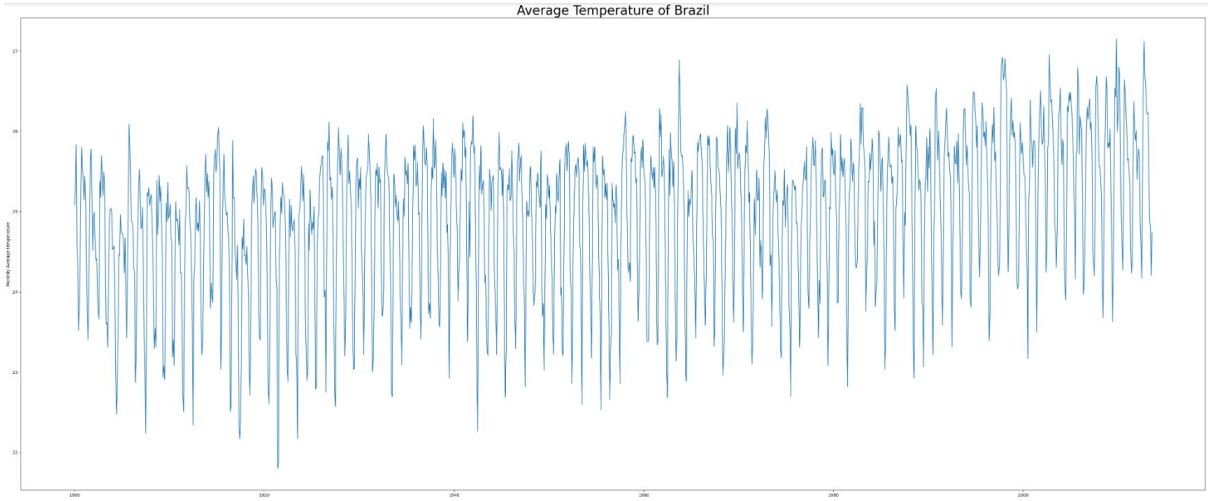
```
[ ]  data_list = [USA_df,UK_df, Brazil_df, Kenya_df, India_df]
     title_list = ['USA','UK','Brazil','Kenya','India']
     for (i,j) in zip(data_list,title_list):
       print('ADF Test for %s' %j)
       dftest = adfuller(i['AverageTemperature'].dropna())
       df_output = pd.Series(dftest[0:4], index = ['Test Statistic', 'p-value', '#Lags Used', 'Number of Observations Used'])
       for key, value in dftest[4].items():
         df_output['Critical Value (%s)' %key] = value
       print(df_output)
```

```
ADF Test for USA
Test Statistic                  -3.532340
p-value                          0.007194
#Lags Used                      24.000000
Number of Observations Used    1340.000000
Critical Value (1%)              -3.435239
Critical Value (5%)              -2.863699
Critical Value (10%)             -2.567920
```

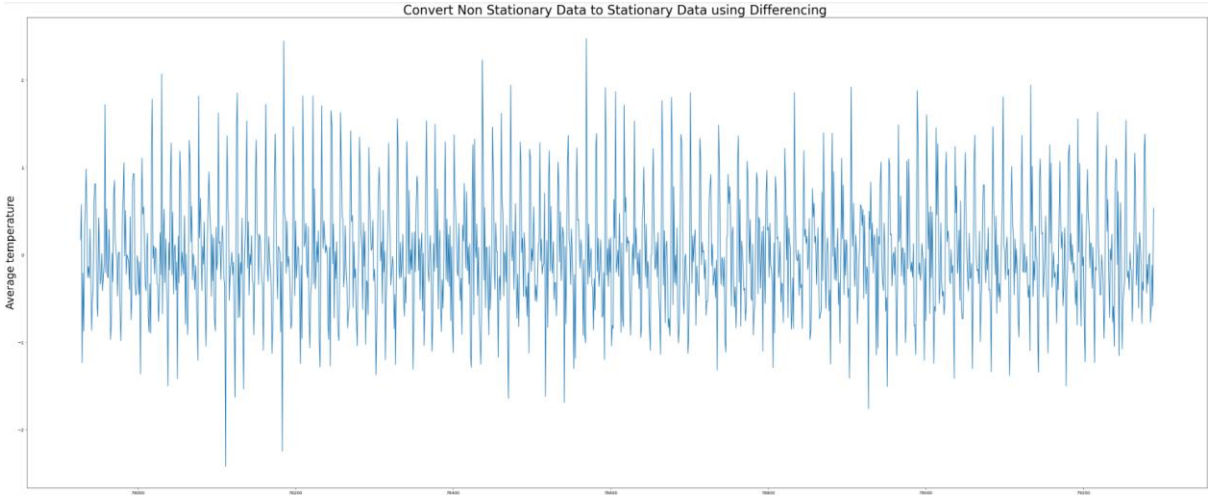
since critical p value < 0.05 for the countries USA, UK and india, we can reject the null hypothesis in these cases i.e. the data is stationary for these countries. But, for Brazil and Kenya, p-value is greater than 0.05, so the data is non-stationary and they will need to be converted to stationary

III

Brazil data before differencing



Brazil data after differencing



IV

RMSE values for first observed p and q values

```
values for USA
2.305274301063426
1.5183129786257594
values for UK
2.009345226239858
1.4175137481660833
values for Brazil
0.3026236890757734
0.5501124331223332
values for Kenya
0.4829137391840084
0.6949199516376029
values for India
8.05120719084023
2.8374649232792692
```