Time Series Analysis

Aleksandra Kalisz
Postgraduate Diploma in
Science in Data Analytics
National College of Ireland
Dublin, Ireland
21118876@student.ncirl.ie

***Abstract*— This report presents an analysis of two time series data sets representing the average temperature from January 1844 to December 2004. The preliminary assessment of the nature and components of the raw time series was conducted using visualizations. The report estimates and discusses suitable time series models for our data set. The optimum model for this series is discussed, and its adequacy for forecasting purposes is commented on. Finally, the data from the testing part and original data are compared.**

I. INTRODUCTION

The Climate Institute of the University of East Anglia has provided two time series datasets of average temperatures representing monthly and yearly data from 1844 to 2004. In this report, we aim to estimate and report on suitable time series models for both datasets. We will conduct a preliminary assessment of the raw time series and estimate appropriate models from three categories: exponential smoothing, ARIMA/SARIMA, and simple time series models. Diagnostic tests will be carried out to ensure the validity of our models. Finally, we will use the data up to and including 2003 as a training set to forecast the average temperatures for 2004 and evaluate our forecasts against the actual data for that year. Our aim is to identify the optimum model for forecasting these time series.

II. DATA

The data are provided in.csv format and include:

- Yearly temperatures – with 161 Rows and 1 column "Temperature"
- Monthly Temperatures- with 1932 rows and 1 column "Temperature"

Data was read and necessary transformations were made to prepare the data to place it into a Time Series

```
month_ts <- ts(month_temp.df, start=1844, frequency=12)
year_ts<- ts(year_temp.df,  start=1844,end = 2004)
```

Figure 1. Time Series

III. PRELIMINARY ASSESMENT

3.1 Dickey – Fuller Test

In the beginning of our assignment, we did Dickey – Fuller Test which shows if statistical properties of the data remain constant over time. As p- value in our monthly data is 0.01 that is strong evidence that monthly time series is stationery.

For the yearly time series, the p-value is 0.2 so, we fail to reject he null hypothesis of non-stationarity. Therefore, we cannot conclude that the yearly time series is stationary.

3.2 Plotting the Time Series

We plotted the time series to visually assess the trend, seasonality, and random irregularities in the data.

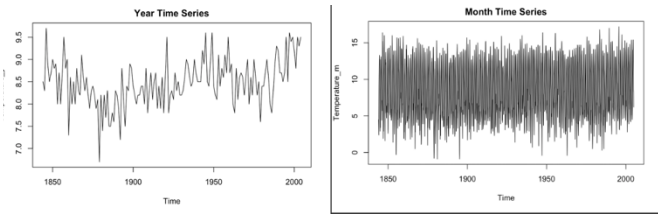

Figure 4. Plotted Time

3.3 Seasonal Plots

Created seasonal plots of the month series to visualize the seasonal pattern.

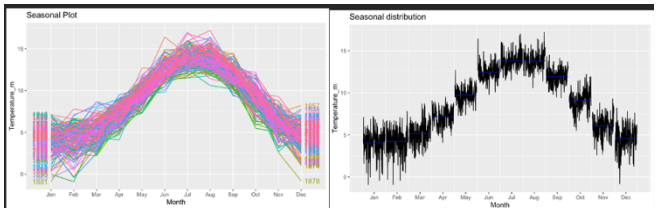

Figure 6. Seasonal Plots

From the plots we can see that in the yearly data the frequency of the data is one reading per year. The temperature average readings range from 6.7 to 9.7 degrees Celsius. It is not clear what the overall trend of the temperature readings is.

In Monthly data the temperature readings show a regular seasonal pattern with temperatures peaking in the summer and dropping in the winter. The data also exhibits year-to-year variability, with some years being warmer or cooler than others. Additionally, there is a slight upward trend in the temperature readings over time, but this trend is relatively small compared to the year-to-year changes.

IV. SMOOTHING TECHNIQUES

We also smoothed the time series using moving averages and plotted the simple moving average for the month. Smoothing time series data is a common technique used in data analysis to remove noise or fluctuations from a

dataset, making it easier to identify underlying trends and patterns.

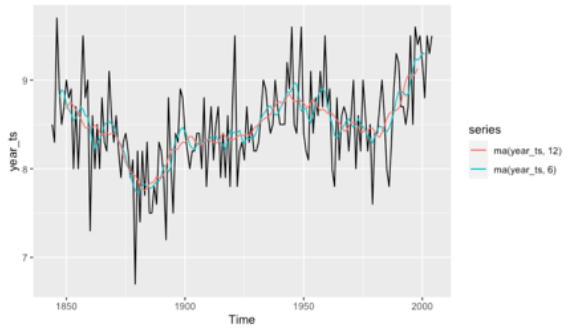

Figure 7. Moving Average

From the graph above we can see that 12 is the proper window size for year time series.

Seasonal Decomposition

This method allows for a more detailed analysis of the different components, such as identifying long-term trends, seasonal effects, or random fluctuations.

- **Multiplicative** decomposition assumes that the different components of a time series are multiplied together. This means that the size of seasonal changes in the data depends on the level of the time series. If the seasonal changes in the data increase or decrease as the time series increases or decreases, then a multiplicative model is more appropriate.

- **Additive** decomposition adds the components of a time series together, resulting in regular changes in the data that have the same size every time they occur that do not depend on the level of the time series. Therefore, an additive model is suitable when the seasonal variation has a consistent magnitude over time.

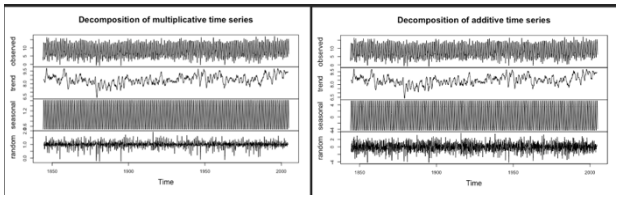

Figure 8. Multiplicative and Additive Decomposition

Our graphs above seem to be almost the same, only Random in Additive decomposition bars reaching a lower scale which would suggest that seasonal variation is smaller compering to scale of data. While it provides some indication which model may be more appropriate, it's important to note that this factor alone is not enough to make a definitive decision.

V. THE MODELS

Mean, Naïve and Seasonal Naïve Model

- The Mean model simply forecasts the future values of a time series as the average of the historical observations. It assumes that the time series is stationary and that any trend or seasonality has already been removed.

Our model suggests the forecast for "month_ts" is that the next four values will be 8.50. The forecast for "year_ts" is that the next four values will be 8.49. And RMSE is 0.52.

- The Naive model assumes that the best forecast for the next value in a time series is the current value, regardless of the historical pattern. It is useful when the time series is stable and there is no trend or seasonality.

Naïve's forecast for "month_ts" is that the next three values will be equal to the last observed value (i.e., no change). The forecast for "year_ts" is that the next three values will be equal to the last observed value plus a small constant (0.00625). And RMSE is 0.6.

- The Seasonal Naive (SNaive) model extends the Naive model to handle seasonality. It assumes that the best forecast for the next value in a time series is the value observed in the same season of the previous year. This model works well when there is a strong seasonality pattern in the data.

SNaive forecast for "month_ts" is that the next three values will be equal to the corresponding values from the same season in the previous year. And RMSE is 0.6.

A lower RMSE represents smaller differences between the forecasted and actual values, indicating better performance of the forecast model. In our case the lowest is in the Mean Model 0.52.

Holt's Model

Holt's model, also known as double exponential smoothing, is a time series forecasting method that uses a combination of a level and trend component to make predictions and is an extension of the simple exponential smoothing method.

The resulting model for year data had smoothing parameters alpha = 0.1544 and beta = 1e-04. The model had an RMSE of 0.4601316, indicating good performance. The forecasts were made using ETS with an RMSE of 0.4632.

Holt – Winters Method

Holt-Winters method is an extension of Holt's linear method and involves exponential smoothing with trends and seasonal components.

The smoothing parameters for month data in this additive method are alpha=0.0279, beta=0.0001, and gamma=0.0001. The RMSE for the training set is 1.20869.

The Ljung – Box

The Ljung-Box test was used to check for autocorrelation in the residuals of a time series model. The test on monthly data resulted in a large Q* value (18055) and a very small p-value (< 2.2e-16), indicating strong evidence of significant autocorrelation present in the residuals. This suggests that the model may not be capturing all the underlying patterns in the data. The test on yearly data resulted in a Q* value of 125.58 and a p-value less than 2.2e-16, indicating strong evidence against the null hypothesis of no autocorrelation in the residuals.

ARIMA

The time series modelling technique used to forecast future values based on past observations.
The ARIMA model was fitted with order (1,0,1) for both monthly and yearly time series data.

For the monthly data, the model estimated coefficients of ar1=0.7057, ma1=0.3324, and an intercept of 8.5004. The model's sigma^2 was estimated as 4.645, and the AIC was 8459.06. The training set error measures showed an RMSE of 2.155171 and an MAE of 1.747039.

For the yearly data, the model estimated coefficients of ar1=0.9681, ma1=-0.8043, and an intercept of 8.6102. The model's sigma^2 was estimated as 0.2103, and the AIC was 214.63. The training set error measures showed an RMSE of 0.4585462 and an MAE of 0.3518872. Overall, the model fits the data well, with low error measures and AIC values.

Ljung Test on Arima

The Box-Ljung test results show that the ARIMA model fit to monthly data has small (1.686e_05) value and significant autocorrelation in the residuals, while the ARIMA model fit to yearly data does not. This suggests that the ARIMA model fit to yearly data is likely capturing the underlying patterns in the data well, while the model fit to monthly data displays the test statistic is only 0.14189 with 1 degree of freedom, and the p-value is relatively large (0.7064) may not be fully capturing all the patterns.

Forecasting Using Arima

The ARIMA (1,0,1) model with non-zero mean was used to make forecasts for monthly and yearly data. For the monthly data, the model didn't perform very well, and caution should be used when using its forecasts. For the yearly data, the model performed well, and its forecasts can be used with reasonable accuracy. It's important to keep an eye on the model's performance and make sure it stays reliable.

Automatic Arima

For the yearly data, the Auto ARIMA model was fit with a coefficient for one moving average term. The model had relatively small errors and low values for RMSE (0.46), indicating that it is likely capturing the underlying patterns well. The forecasts for the yearly data can be used for forecasting purposes with reasonable accuracy.

SARIMA

Seasonal Arima – for the monthly data, the ARIMA model with order (2,1,2) was found to be a good fit, and the model's coefficients suggest that past values and errors of the series can help predict future values. The training error measures indicate that the model has an acceptable level of accuracy.

To evaluate our time series forecasting methods, we typically look at the root mean squared error (RMSE) metric. However, it is important to note that RMSE should not be the only metric used to evaluate models. Other metrics such as mean absolute error (MAE), mean absolute percentage error (MAPE) are very important.

Overall results showed that Auto ARIMA performed the best for yearly temperature data with an RMSE of 0.46. This indicates that the Auto ARIMA model is highly accurate and can be relied upon to forecast temperature data for future years.

For monthly temperature data, the SARIMA method performed the best with RMSE 1.1 as well as Holt-Winters method with RMSE 1.2 is a good option for forecasting monthly temperature data. This implies that the Holt – Winters model can be used as an alternative to the SARIMA method for monthly temperature data forecasting.

VI. SPLITTING DATA

We are splitting data to evaluate the performance of a model on new, unseen data. If we use all the data to build and evaluate our model, we risk overfitting.

```
# split monthly temperature data into training and test sets
train_month_ts <- window(month_ts, end=c(2003,12))
test_month_ts <- window(month_ts, start=c(2004,1), end=c(2004,12))

# split yearly temperature data into training and test sets
train_year_ts <- window(year_ts, end=2003)
test_year_ts <- window(year_ts, start=2004, end=2004)
```

VII. FORECASTING

Figure 12. Yearly Temperatures

Monthly Forecast Using SARIMA

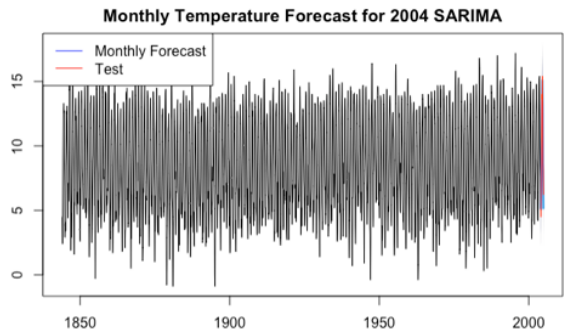

Figure 9. Monthly Forecast Using SARIMA

Above plot shows the blue line in 2004 represents the forecasted monthly temperatures, while the red line represents the actual test data. The closer the blue line is to the red line, the more accurate the forecasted values are.

Monthly Forecast Using Holt Winters Method

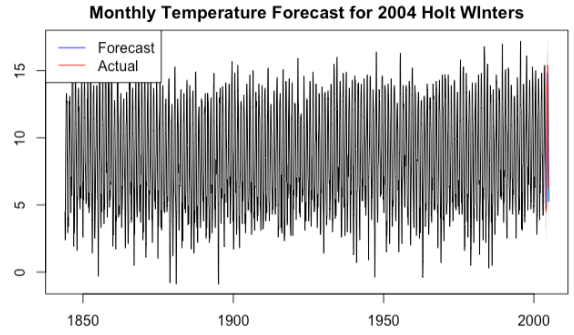

Figure 10. Monthly Forecast using Holt - Winters

Holt-Winters method shows very similar results as SARIMA method.

Yearly Temperatures Forecast Using Seasonal Arima

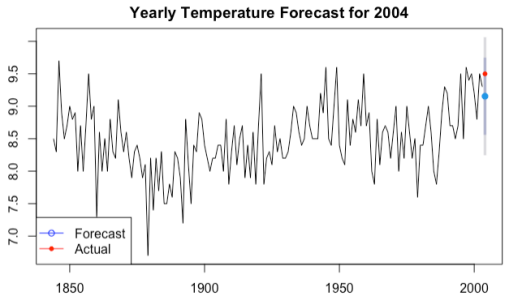

Figure 11. Splitting Data

I the graph the blue line represents the forecasted values, while the red dot at 9.5 represents the actual temperature values for the test set and its mean at 9.2. The plot shows that the model was able to capture the overall pattern and trend of the actual data well.

Compering Predicted and Actual Data

To compare predicted and actual data we created a plot, for monthly data*:*

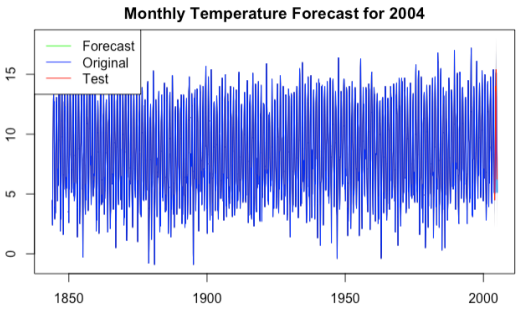

Figure 13. Monthly Temperature Compared

Which indicates that temperatures forecasted for 2004 were between 5 and 15 and original temperatures were 4 – 15 which makes the SARIMA method good forecasting method for seasonal data.

To compare predicted and actual values for yearly data we created plot:

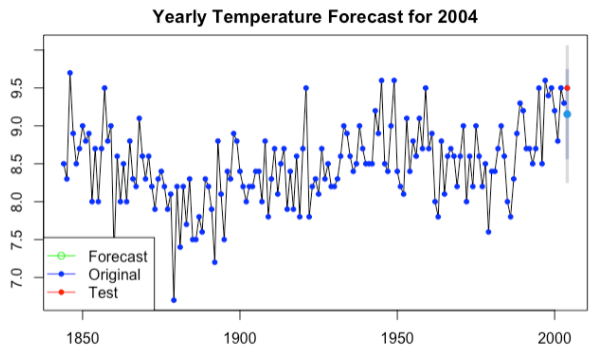

Figure 14. Compared Yearly Temperature

Above plot shows forecasted temperatures between 8.6 and 9.8 where actual temperatures were on average 9.4. This indicates that Auto Arima is a good method of forecasting yearly time series.

VIII. Conclusion

In conclusion, our analysis of two time series datasets on average temperatures provided insights into the underlying patterns and allowed for future forecasting. We used various time series techniques such as ARIMA, SARIMA, Holt-Winters, and Seasonal ARIMA to model and forecast the data, and evaluated their performance using error measures such as RMSE and MAE. Our findings showed that the Sarima and Auto Arima models had the lowest RMSE and performed well in forecasting future values for both monthly and yearly data. The comparison of predicted and actual data demonstrated the effectiveness of the various time series forecasting methods in capturing the overall pattern and trend of the data. Overall, this analysis provided valuable insights into the temperature trends over time and the effectiveness of various time series forecasting methods in predicting temperature values.

Table of Figures

DECLERATION

I declare that the work I have submitted for this project is my own work and has been completed by me alone. I have acknowledged all material and sources that have been used in its preparation, whether they be books, publications, lecture notes, or any other kind of document. I have not copied or otherwise plagiarised any part of the work submitted for this project from other students and/or persons.

Date: 09/05/2023

Signature: Aleksandra Kilisz