MATH5802

Time Series and Spectral Analysis

Analysing historical forced migrations across the Atlantic Ocean

1. Introduction

Having given a collection of records documenting slave voyages across the Atlantic Ocean between 1654 and 1807, our primary goal is to uncover insights through careful analysis. First, we must address trends or seasonality in the data as they might overshadow the underlying variations. Trends represent long-term movement or patterns that show a general direction in which the data moves over time, while seasonality refers to a pattern of variation in the data that occurs at regular intervals.

When the trends and seasonality are removed, we can go on to identify the underlying process. We use tools like Correlograms and Yule-Walker Equations that help us analyse the autocorrelation or the correlation between data points at different time intervals. By examining the autocorrelation, we can get more insights into the underlying process and its characteristics. To further understand the underlying process, we use Periodograms; this helps us identify dominant frequencies in the data, representing the repeating patterns or cycles within the data.

Once we have the process that generated the data, we can utilize the residuals – the difference between the actual data and the fitted model, to make prediction values.

In summary, the process involves analysing the data, removing trends and seasonal effects, identifying the underlying process, understanding the dominant frequencies, and using the residuals to make predictions. This approach provides insights into the patterns and trends in the slave voyage data and helps us make informed predictions.

2. Data

In this section, we load the data, inspect it, and remove the overall trends and the seasonality from the data. First, we load our data – "*slavery.RData*" into R, which is stored in the variable *xf.* This *xf* variable has two parts – *xf$num* and *xf$year*; *xf$num* contains information on the yearly estimated number of enslaved people who embarked on the voyage, and *xf$year* contains information on which year it was. ***X is xf$num.***

When we plot the original dataset along with its trend, we get ***Figure 1***. This plot shows a gradual increase over time in the number of slaves who embarked on a voyage across the Atlantic Ocean. This data starts in 1654, in which about 3054 people embarked on a voyage across the Atlantic, and ends in 1807 when approximately 36000 slaves embarked on the voyage; this is about 12 times more than 1654— indicating an overall increasing linear trend.

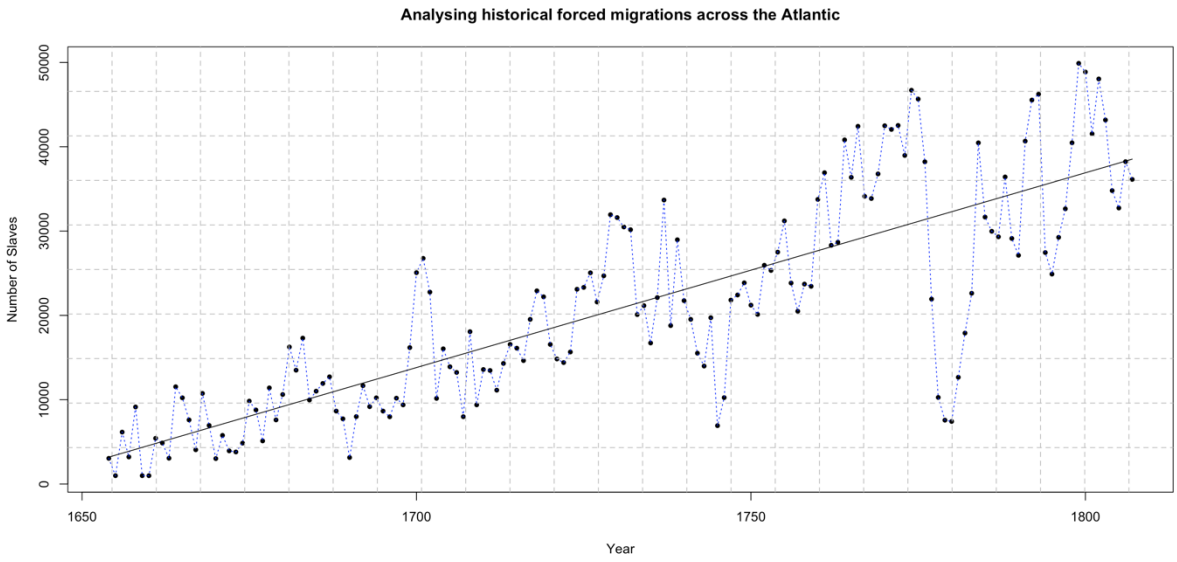

Figure 1: Plot of the time series.

Looking at ***Figure 1***, we can also see no seasonality in the data, so only the trends must be removed to get the residuals.
After removing the linear trends from the data, we get ***Figure 2***, which gives us the data residuals. Here, we indicate ***Y as the residuals of the time series.***

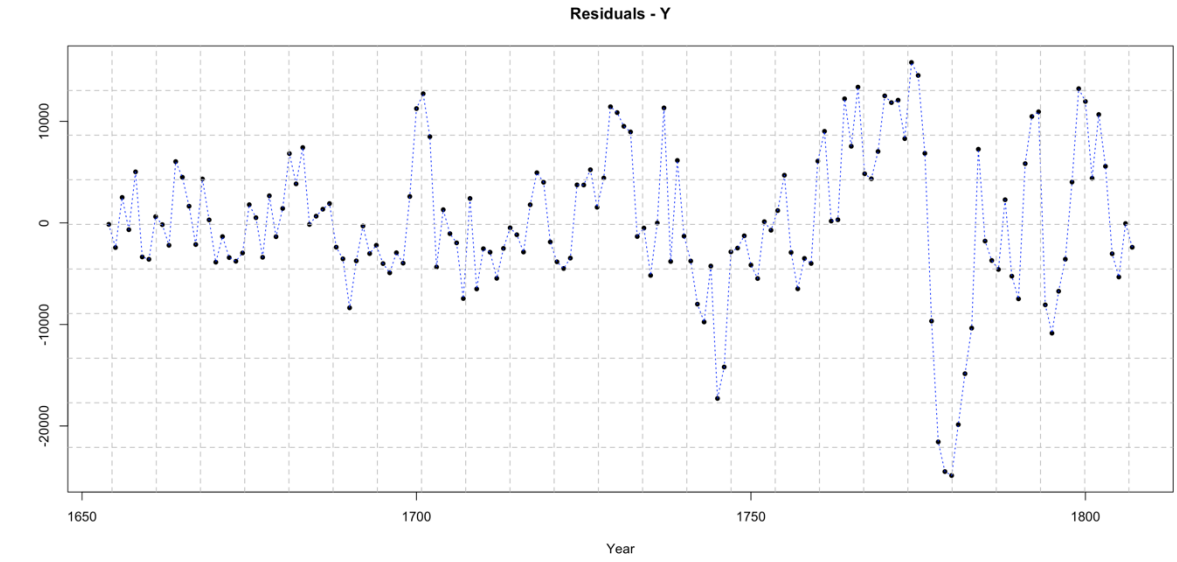

Figure 2: Plot of the Residuals, after removal of the trends from the time series - denoted as Y.

3. Process

Now to inspect the residuals - Y and determine whether its Moving Average Process (MA) or Autoregressive Process (AR), for this we need to plot the auto-correlation function plot (ACF) and the partial auto-correlation function plot (PACF).
The Autocorrelation Function (ACF) plot displays the correlation between a time series and its lagged values at different time lags. Analysing the ACF plot provides valuable insights

about the model identification – the decay or pattern in the ACF plot helps identity what process it could be, whether an MA or an AR process.

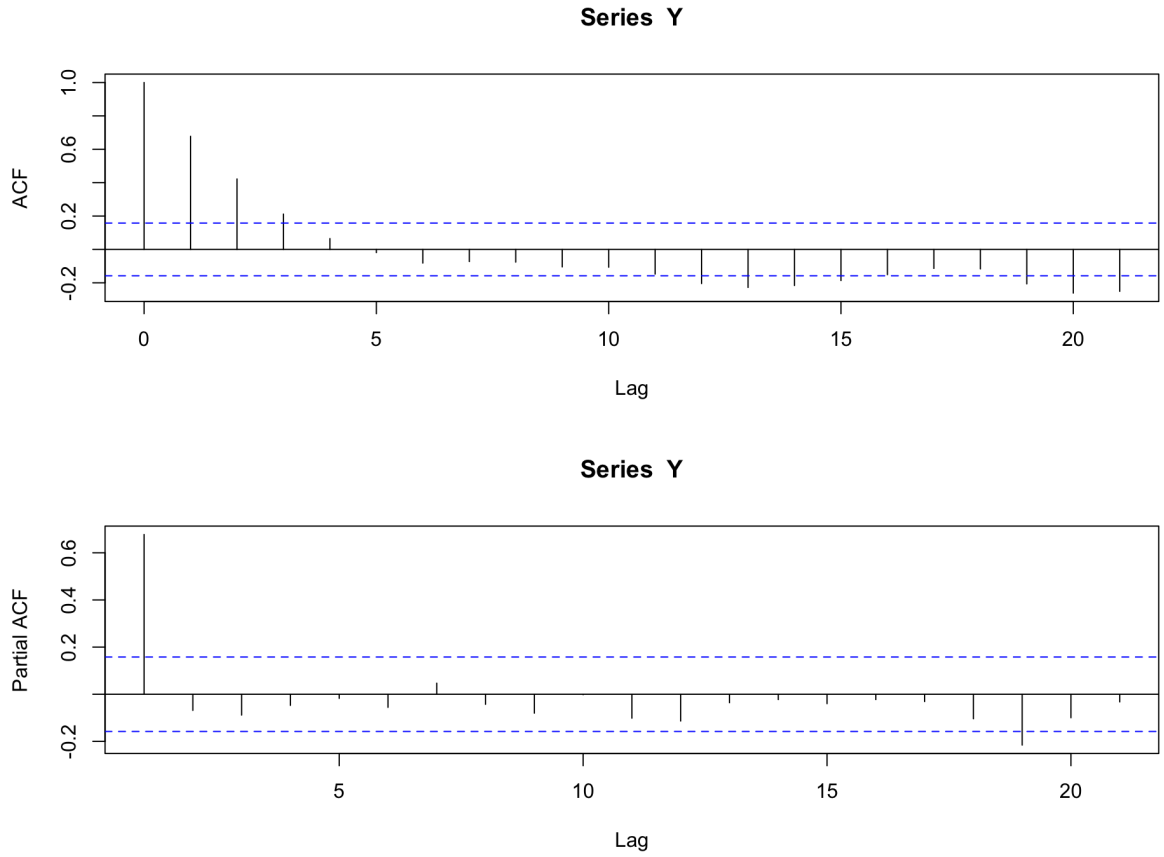

Figure 3: ACF and Partial ACF plot of Y.

From the ACF plot in **Figure 3**, we can see a decay at the start of the plot before lag 5, indicating that it is an AR process. The Partial ACF plot gives us the order of the process. The cut-off happens between 1 and 2 – as the difference is high. Therefore, we can say that it is an AR (1) process.

4. Yule – Walker Equations

The Yule–Walker equations are a set of equations used in time series analysis to estimate the autoregressive parameters of an AR process. In the previous part, we concluded it is an AR(1) process based on the ACF and PACF plot of the residuals – Y.

While using R to estimate the parameters of the AR (1) process, we can use an in-built function ar() and specify the method as "yule-walker."

```
> ar(Y,aic=FALSE,order.max=1,method = c("yule-walker"))

Call:
```

```
ar(x = Y, aic = FALSE, order.max = 1, method = c("yule-
walker"))

Coefficients:
     1
0.6777

Order selected 1 sigma^2 estimated as 28766227

> ar(Y,aic=FALSE,order.max=2,method = c("yule-walker"))

Call:
ar(x = Y, aic = FALSE, order.max = 2, method = c("yule-
walker"))

Coefficients:
      1        2
 0.7241  -0.0685

Order selected 2 sigma^2 estimated as 28820780
> ar(Y,aic=FALSE,order.max=3,method = c("yule-walker"))

Call:
ar(x = Y, aic = FALSE, order.max = 3, method = c("yule-
walker"))

Coefficients:
      1        2        3
 0.7180  -0.0045  -0.0884

Order selected 3 sigma^2 estimated as 28786148
```

Here we give three commands to get the parameters for each value of p = 1,2,3. P represents the order of the AR process.

For the first case, order is 1 so the co-efficient or α_1 is 0.6777.
For the second case, order is 2 so the coefficients are:

α_1= 0.7241 and α_2= - 0.0685

For the third case, the order is 3 so the coefficients are:

α_1= 0.7180, α_2= - 0.0045 and α_3= - 0.0884

These coefficient values give us the relationship between the current value of the time series and its lagged values. The positive coefficients indicate a positive correlation between the current value and the corresponding lagged value. The negative coefficients indicate that an

increase in the lagged value is associated with a decrease in the current value and the corresponding lag value.

5. Correlograms of the Residuals and Squared Residuals

The Correlogram, also known as an autocorrelation function plot, is a graphical tool used in time series analysis to visualize the correlation between a time series and its lagged values. The primary purpose of a correlogram in Time Series is to check if the model is best fitted or whether the time series must be refined finer to remove any additional autocorrelation within the data.

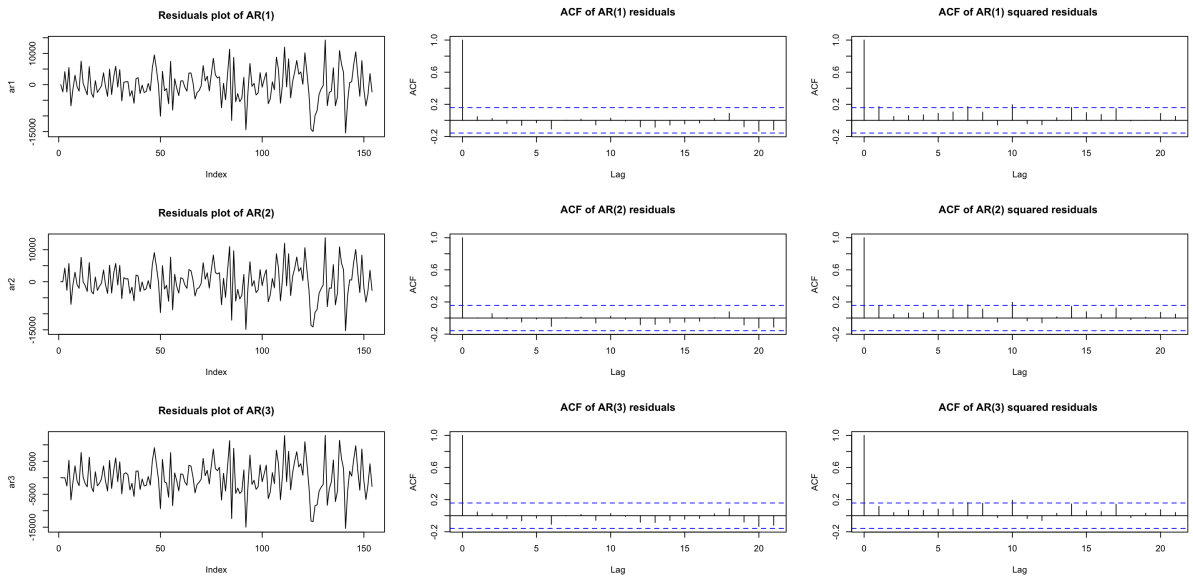

Figure 4: AR(p) residuals plot, Correlogram plots of the AR(p) residuals and squared residuals.

The ACF of the residuals from a well-fitting AR model should be white noise, meaning that the autocorrelations at all lags should be close to zero.

When we look at the ***Figure 4***, ACF of AR(1) residuals plot, ACF of AR(2) residuals plot and ACF of AR(3) residuals plot we do not see any significant autocorrelations at any lag, everything is below the blue dashed line, which indicates that it's within the 95% confidence interval which is [-1.96/√n, +1.96/√n]. Here n is the length of the time series which is 154. On calculating, the intervals for the blue dashed line is [-0.1579, +0.1579] . Rounding it off – it becomes [-0.158,+0.158].

As the ACF of AR(1), AR(2) and AR(3) residuals plots don't tell us much, we look into the squared residuals plot of the same.

When we look into the ACF of the AR(1) squared residuals plot, we see that there is a significant autocorrelation at lag 1, where the line is going above the blue dashed line, indicating that the AR(1) model may be underfitting the data.

Looking at the ACF of the AR(2) squared residuals, we can see that there is no significant correlation at lag one; at lag 7, it crosses the dashed blue line by a minimal margin; it can be considered a good fit for the model when compared to AR(1).

Finally, when we see the ACF of AR(3) squared residuals, we see that just like the ACF of AR(2) model, where the lag 7 crosses the dashed blue line by a very small margin, and therefore can again be considered as a good fit for the data.

Now, looking back at the ACF of AR(2) residuals and ACF of AR(3) residuals plot, we can decide which is the best fit for the data. AR(3) seems the best fit because of the decay, with the first correlation being the largest and the subsequent correlations decreasing in magnitude. ***Z is AR(3) residuals.***

6. Periodograms

A periodogram is a tool that estimates the spectral density of a signal in the frequency domain. It shows the power of each frequency in a signal. The higher the power at a frequency, the more likely it is to contain a component.
In Time series, we can use a periodogram to identify the dominant frequencies in the data. The main goal of this analysis is to identify the critical frequencies or periods in the observed time series.

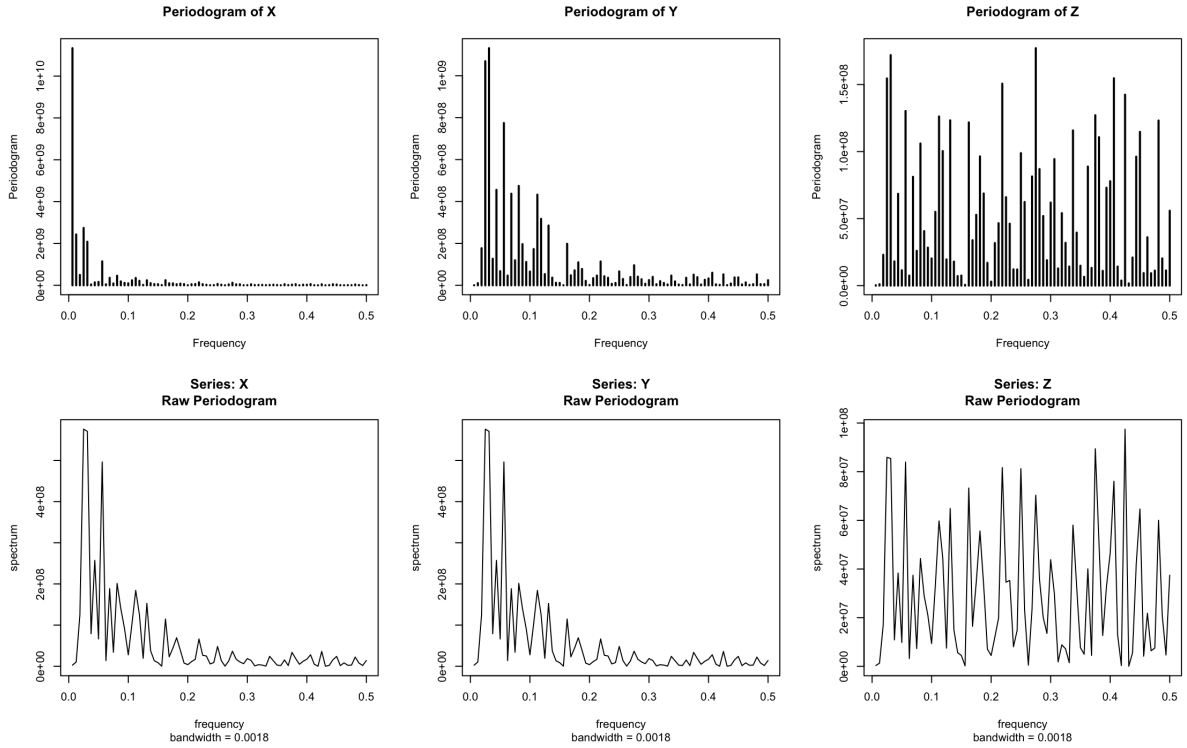

Figure 5: Periodogram and Raw Periodograms of X,Y and Z.

Plotting the periodograms of X (which is *xf$nums*), Y (which is the residuals after removing the trend and seasonality) and Z (which is the model with the best fit).

For the periodograms in the first row of ***Figure 5*** – a function called periodogram from the TSA library was used to plot them, and for the periodograms in the second row – the function spec.pgram() was used to plot them.
All three – Series X, Y, and Z have significant power in the low frequencies. X peaks were very close to 0.02 Hz, Y at 0.05 Hz, and Z at 0.4 Hz, indicating the time periods to be 50 seconds, 20 seconds, and 2.5 seconds, respectively.

These periodograms also suggest that X may be driving series Y and Z. This is because the peak frequency of X is lower than that of Y and Z.

7. Re-Fitting with ARIMA

Here for the chosen AR(3) model we refit it with arima() command in R.

```
> ar_arima=arima(Y,order=c(3,0,0))
> print(ar_arima)
Call:
arima(x = Y, order = c(3, 0, 0))

Coefficients:
         ar1      ar2      ar3  intercept
      0.7141  -0.0055  -0.0857   -18.9415
s.e.  0.0800   0.0990   0.0800  1122.7828

sigma^2 estimated as 28027823:  log likelihood = -1539.29,
aic = 3086.57
```

8. Forecasting

Forecasting in time series is the process of using historical data to predict the future values of a time series. The choice of forecasting method depends on the specific characteristics of the time series data. If the time series is stationary, then an AR or MA model might be sufficient; if the time series is non-stationary, then the ARIMA model may be necessary.
Forecasting time series data can be complex, as there is no guarantee that the forecasts will be accurate. However, forecasting is a valuable tool for making informed decisions about the future.

R has an inbuilt function predict() that would help us predict future values based on the best-fit model. In my opinion, AR(3) seems to be the best fit.

```
> predict(ar_arima1,n.ahead=5)
$pred
Time Series:
Start = 155
End = 159
```

```
Frequency = 1
[1] 37651.84 36599.25 35612.81 34688.34 33821.96

$se
Time Series:
Start = 155
End = 159
Frequency = 1
[1] 4216.056 5778.127 6862.813 7690.303 8349.704
```

Visualizing the prediction along with the original time series data, we get.

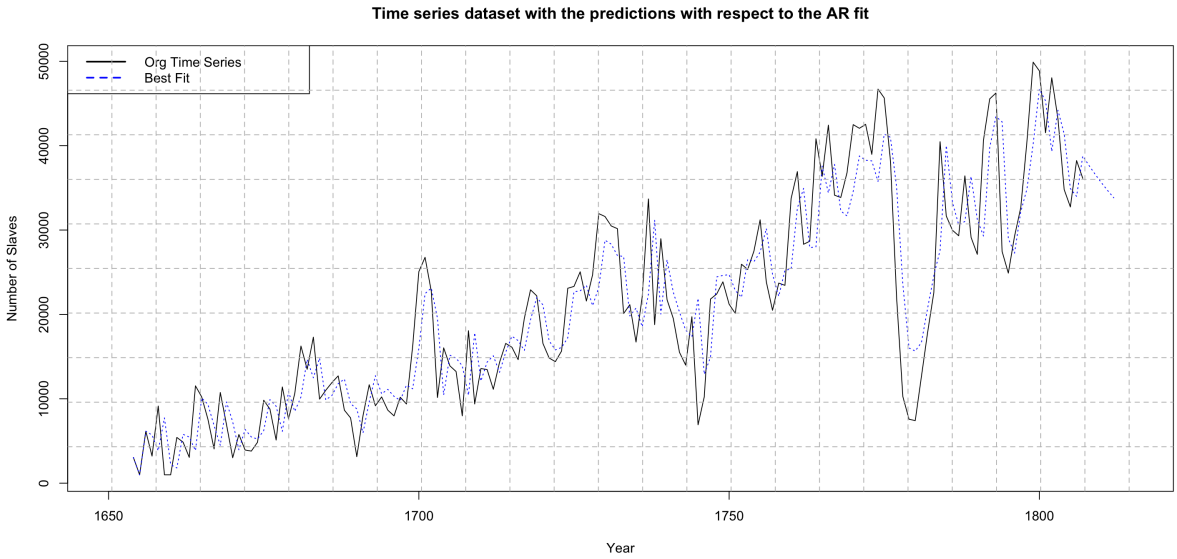

Figure 6: Original Time series, with the predicted values for the fit model

The $pred shows us the predicted values for the fit model, and $se gives us the standard error. The values we see for the prediction seem accurate. As for the original data, there is a drop after 1807, so it continues to drop by almost 1000 yearly.

9. Conclusion

The insights obtained through this analysis can help us better understand the patterns and trends in the slave voyage data and make informed predictions. The methodology presented here can also be applied to other historical data sets to uncover meaningful insights and trends.