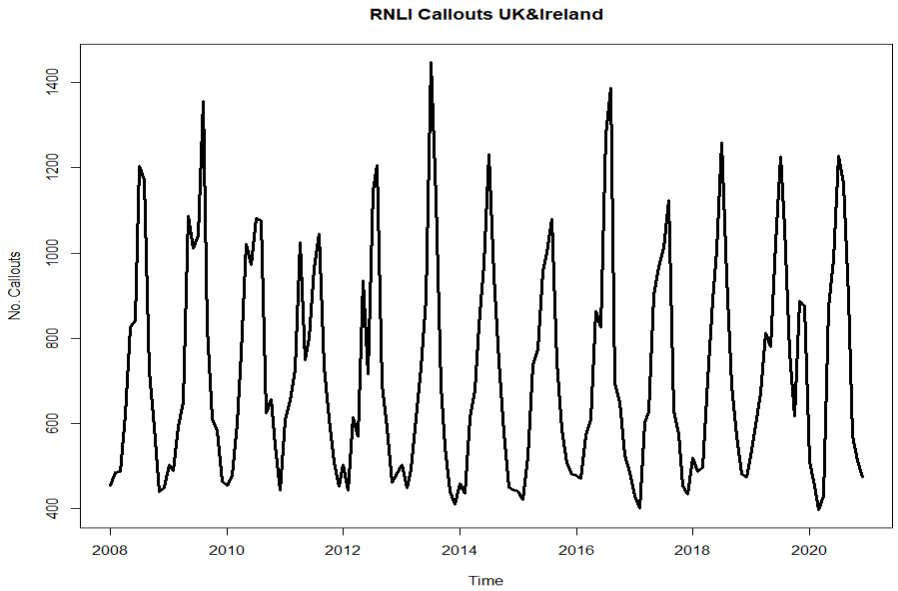
Time Series Analysis on RNLI Callouts In UK & Ireland

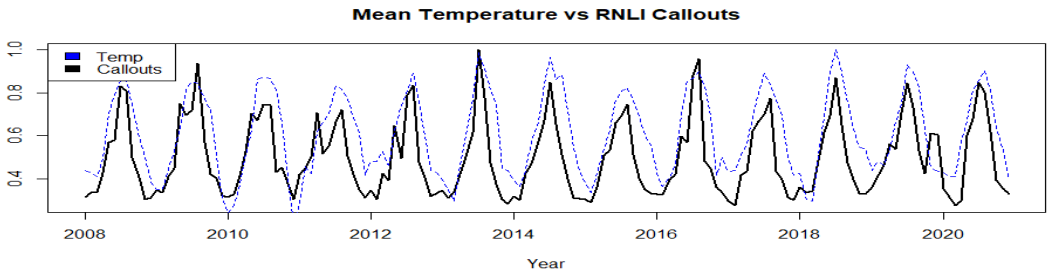
Student Number : 118359016

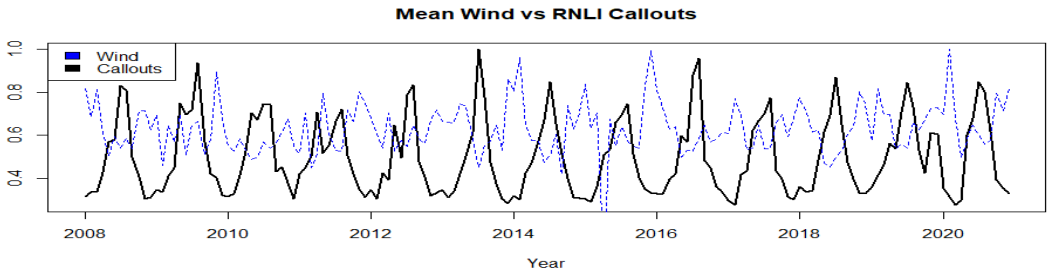
Dataset

The data used in this project was downloaded from the official RNLI website which has an open dataset available to the public. Link: https://data-rnli.opendata.arcgis.com/ . This analysis is aimed to give an insight into number of incidents where a lifeboat is required to assist people who may find themselves in possible life-threatening scenarios out at sea in the UK & Ireland and also how weather may impact the number of callouts. The dataset contains 112248 observations, where each observation is an individual callout for a lifeboat in Ireland or the UK in the period from 01/01/2008 to 31/12/2020. Each observation has 29 features, including a date and time of callout. For our time series analysis, we are most interested in the date and time feature as it allows us to sum daily data into monthly data, which is achieved using pandas. Feature selection reduced the dataset down to 2 features, Date(Month and Year) and the total number of callouts during each month. After plotting the initial monthly observations, a seasonal trend is quickly identified where the number of callouts is significantly higher each year in the mid to late summer months (June, July, August).



This may be strongly correlated to warmer weather and calmer weather conditions, represented in the form of lower wind speeds where more people may be out at sea. Viewing plots of mean temperatures and wind speeds against the number of callouts we can see that there is a correlation.

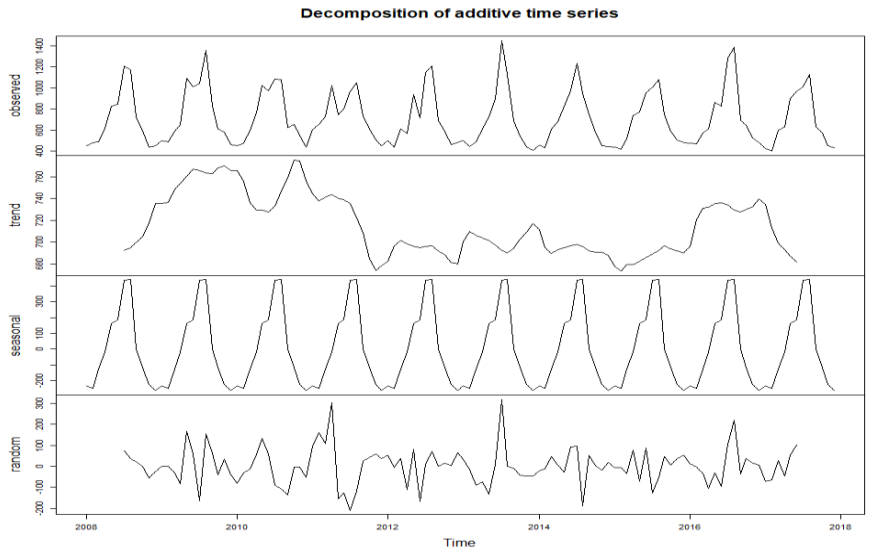




From the mean temperature plot, we can see as the mean temperature rises, the number of callouts also rises and in the winter months where the mean temperature falls, the number of callouts falls. However, the antithesis occurs in the mean wind speed plot where a negative relationship is apparent. As the mean wind speed falls, the number of callouts rises and vice-versa. This is important in the Further Discussion section of this report where these relationships can be used to explain the error between our predicted values and true values of our time series model.

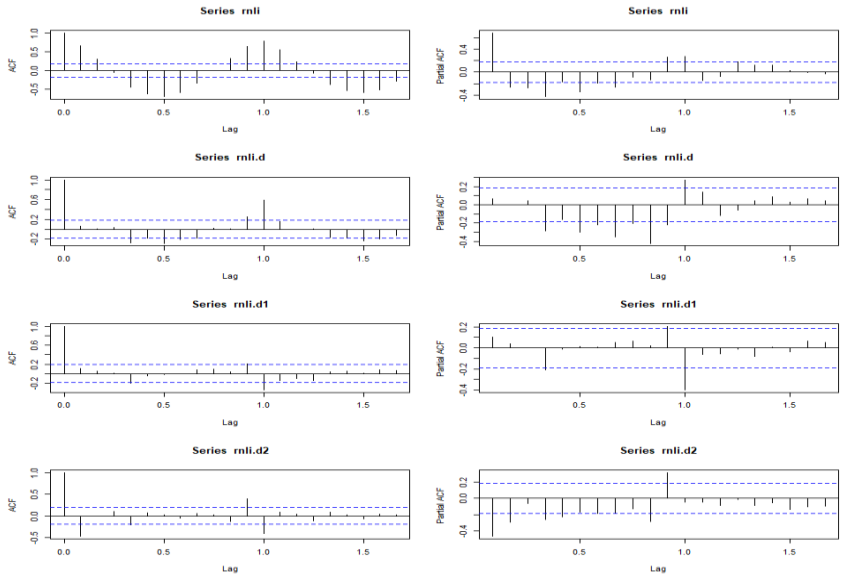
Reduction To Stationarity

A stationary time series is one whose properties do not depend on the time which the series is observed. Therefore, if a time series has a trend or seasonality aspect, it is not stationary.



Inspecting the diagram above, the trend feature of the time series appears to be random but there is evidence of a strong seasonal feature, concluding that the time series may be non-stationary. By carrying out the augmented Dickey-Fuller Test, which tests for the

presence of a unit root, we can conclude that the data is stationary as we get a p-value < 0.05. However, the data must be differenced. The data was differenced 3 different ways and the resulting auto correlation function (ACF) and partial auto correlation function (PACF) was compared for each.



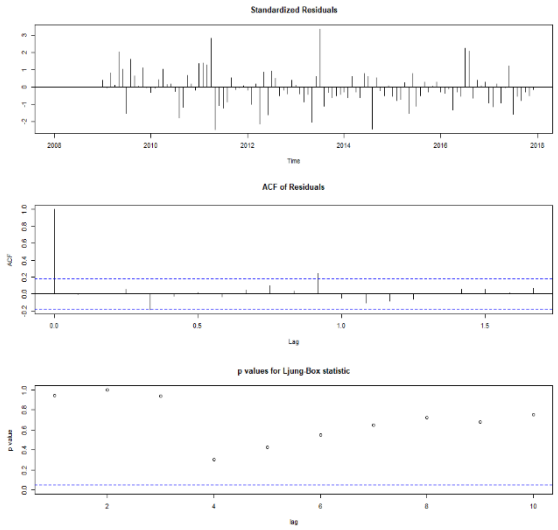
The first row contains the ACF & PACF of data that was not differenced (1). The second row contains data that was differenced at lag = 1 (2), next row has data differenced at the seasonal component (lag=12) (3), and the last row contains data that was differenced at lag =1 and then at lag=12(4). (3) was the preferred result, where the data was differenced just at the seasonal component lag=12, because it fit the criteria of both the ACF & PACF decaying to zero. Also, there is no sinusoidal shape present unlike what we see in the ACF in (1) and (2). Furthermore, we see many lags exceeding the significance bounds in (2) and (4) in both the ACF and PACF and so just using seasonal differencing (3) is our best result.

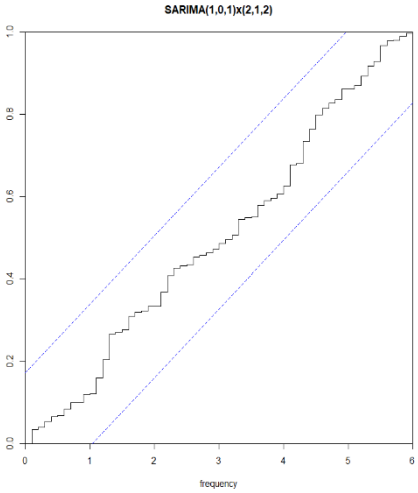
Model Fitting

The criteria for choosing a model was the preferred model would be the model that minimizes the AIC but also taking into the consideration the complexity of the model i.e. lag p in AR(p) or MA(p). Inspecting the chosen ACF and PACF plots (3), there are spikes at the seasonal lags which may suggest a seasonal AR(2) term. In the non-seasonal lags, there are two spikes outside the significant bounds, suggesting a possible AR(2) term. By looking at the pattern in ACF, we can identify significant spikes at seasonal lag suggesting an MA(2) and 1 significant spike at non-seasonal and so possibly suggesting a MA(1) term. These terms result in a final ARIMA model (2,0,1)(2,1,2) with an AIC of 1321.26. However, by trying to minimise the AIC, an ARIMA model (1,0,1)(2,1,2) produces a lower AIC of 1318.27 and is a simpler model and so that model was chosen.

Model Criticism

From the diagnostic plots for the residuals of the model, there are a few significant spikes that exceed the significance bounds in the ACF. The model passes the Ljung-Box Portmanteu Test for all lags (lag 4 is the min p-value and is =0.3055) and so doesn't have a lack of fit. This is also shown below in the diagnostic plot by all p-values being greater than the bound.

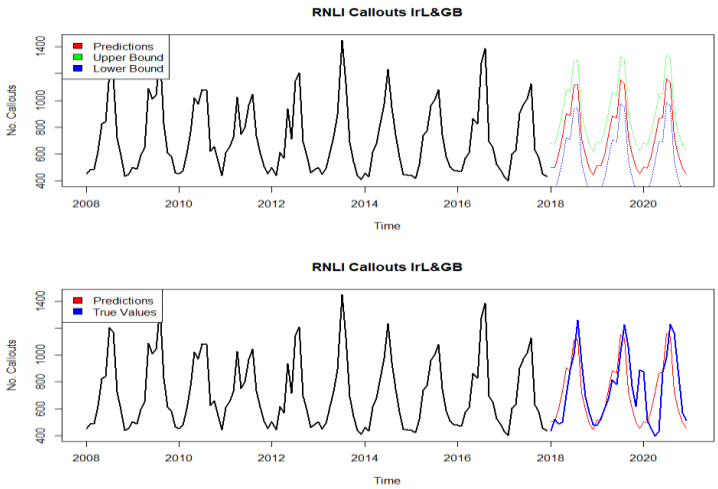


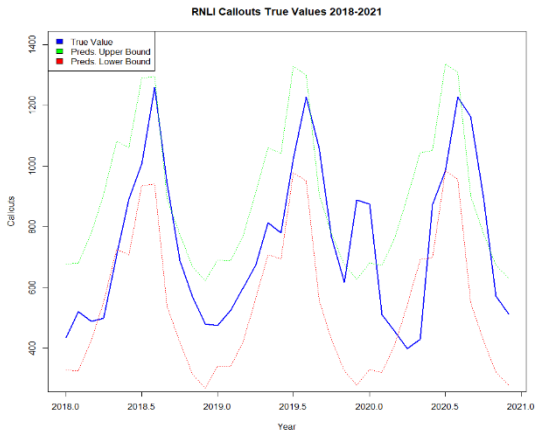


The cumulative periodogram starts at w=0 and stays within the bounds to (0.5,1). It closely replicates a random series and as the periodogram doesn't exceed the bounds at any point, it shows not to have an excess in high or low frequencies.

Forecasting

Before fitting the model, 36 observations were excluded from the dataset and instead were to be used as a metric to compare our predictions to. The seasonal ARIMA (1,0,1) x (2,1,2) was fitted and the following predictions for the 36 observations were made with upper and lower bounds shown and a comparison to the true values.

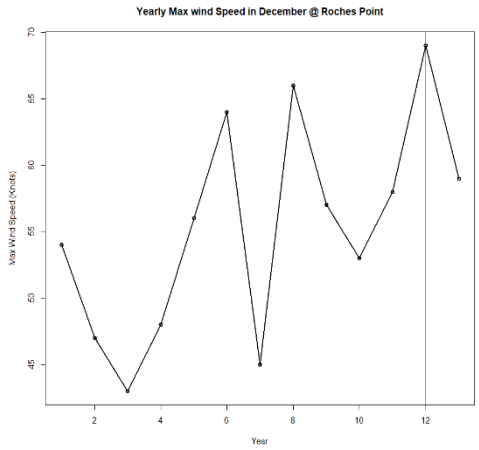


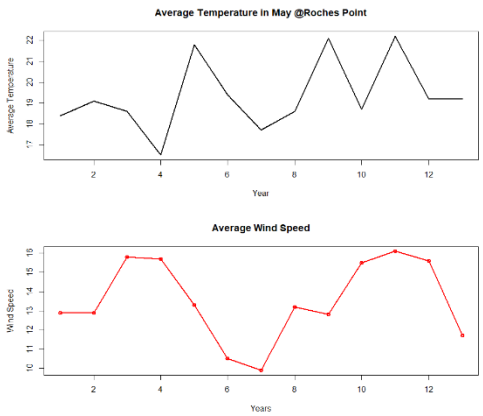


From second plot where we compare our predicted values to the true values for the next 36 observations, we can see a good fit with a strong correlation of 0.69 between them. The RMSE between the set of predictions and the set of true values is 32.57 and the max absolute difference is 439 which may need to be investigated. When we look at the plot of the true values in the period 2018-2021, we see that almost all values lie with the significant bounds of 2 standard errors (95% CI).

Discussion

The fitted ARIMA model (1,0,1)x(2,1,2) showed strong performance when compared to the true values of the data, as shown by the true value lying within the confidence intervals and the RMSE value. Also, the model is relatively simplistic with log p values for AR(p) and MA(p) while still maintaining high accuracy. However, two large absolute differences were found between the true values and prediction values. These occurred in May 2020 and December 2019. In December 2019, there were significantly more callouts than the number predicted. Investigating into this further and using meteorological data from the time period from Roches Point in Cork, we can see that at that period of time, there was a significant rise in the max gust speed and mean wind speed in comparison to previous Decembers, suggesting a storm. This would result in a more dangerous sea and resulting in more callouts as seen in the true data values.





Investigating May 2020, we see that the mean temperature was higher than previous years and therefore a larger population may have been at sea and resulting in more callouts from the graph in the Dataset section of the report. These two outlier values shows that the model performs weak against data that is out of the ordinary, which is expected but model can be improved on by considering more factors when modelling. Also, more models could be analysed to see which minimises the AIC but may be more complex. However, the fit may be better and from this we could test the performance of that model against ours.