Introduction

A dataset containing the number of people arriving to Ireland by air and sea travel was analysed using time series analysis. Several steps were taken in the analysis process and are outlined in the following sections. To begin, the data was retrieved and prepared for analysis. The data was differenced in order to reduce it to stationarity. ARIMA models were implemented and assessed for their performance. Forecasts with the chosen model were conducted and compared to a section of data that had been removed from the original dataset for this purpose. A discussion of the entire process concludes the document.

Dataset

The dataset was obtained from the Central Statistics Office (CSO) in Ireland via their website (https://www.cso.ie/en/index.html). The original dataset contained several monthly records from 2010 to 2021 for people arriving to and departing from Ireland via sea and air travel. For the purposes of this assignment, the dataset was reduced to the total number of arrivals to Ireland by air and sea travel (the analysed data has been submitted and the original dataset is available upon request). The dataset was split for model fitting and model testing. Values from January 2010 to April 2018 were used to fit the model and the model was then tested on values from May 2018 to January 2021. The dataset has important economic and societal value. For the tourism industry, it is important to know how many people are expected to arrive to the country each month so that preparations can be made in advance. For example, tourist attractions having the required number of staff in place for peak season, restaurants knowing how many ingredients they need to order each month in order to reduce waste and excessive costs, and hotels knowing how many rooms they can expect to fill. From a societal viewpoint, it is important to know how much traffic will be arriving to airports and ports around the country. This will affect things like traffic congestion and also the trade of imports and exports, with inadequate planning during peak arrival times potentially resulting in long delays.

The initial impressions of the time series are that it is showing an upward trend along with a clear seasonal effect (Figure 1). The seasonal effect can be seen further in the boxplots in Figure 2.

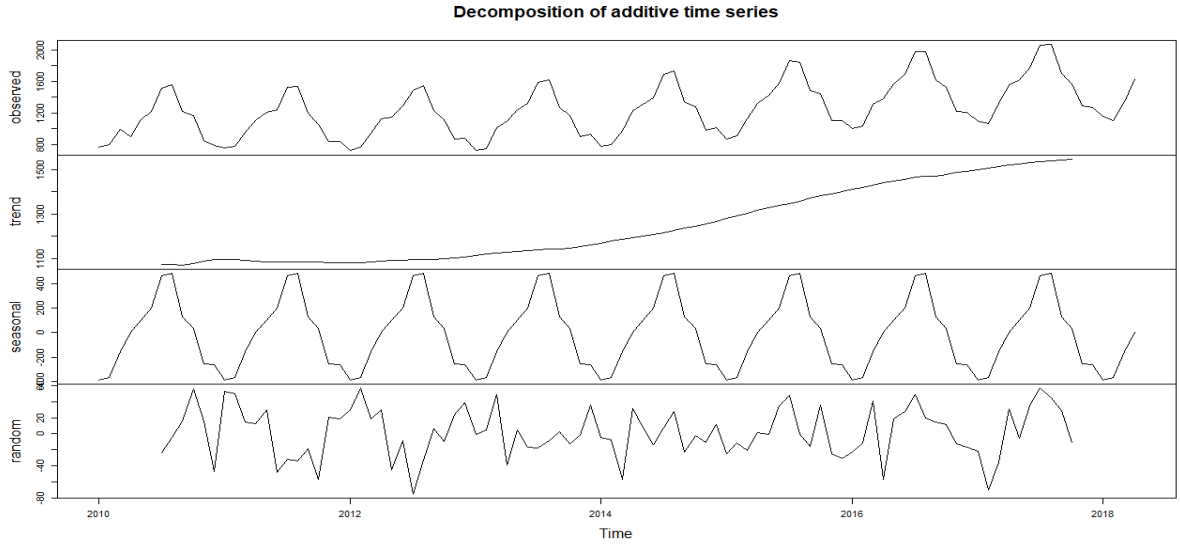

Figure 1: Number of people arriving to Ireland decomposed into its trend, seasonal and random effects

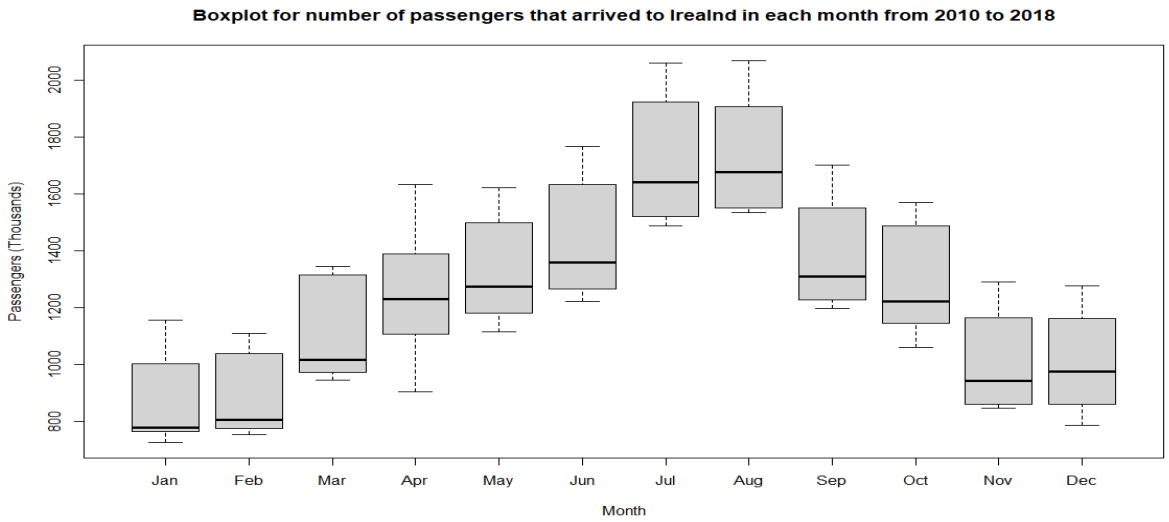

Figure 2: Monthly boxplots for the number of people that arrived to Ireland across the eight-year period 2010-2018

Reduction to Stationarity

A Dickey-Fuller Test on the data found that it was already stationary ($p < 0.01$). In order to remove the seasonal trend, the data was differenced at lag 12. Given that the Dickey-Fuller Test on the seasonally adjusted data found that the data was not stationary ($p = 0.54$) and the plot of the seasonally adjusted data did not look stationary, it was decided to difference the data at lag one month as well as difference it seasonally. The plot of this data looks stationary (Figure 4) and a Dickey-Fuller Test supports that it is stationary ($p < 0.01$).

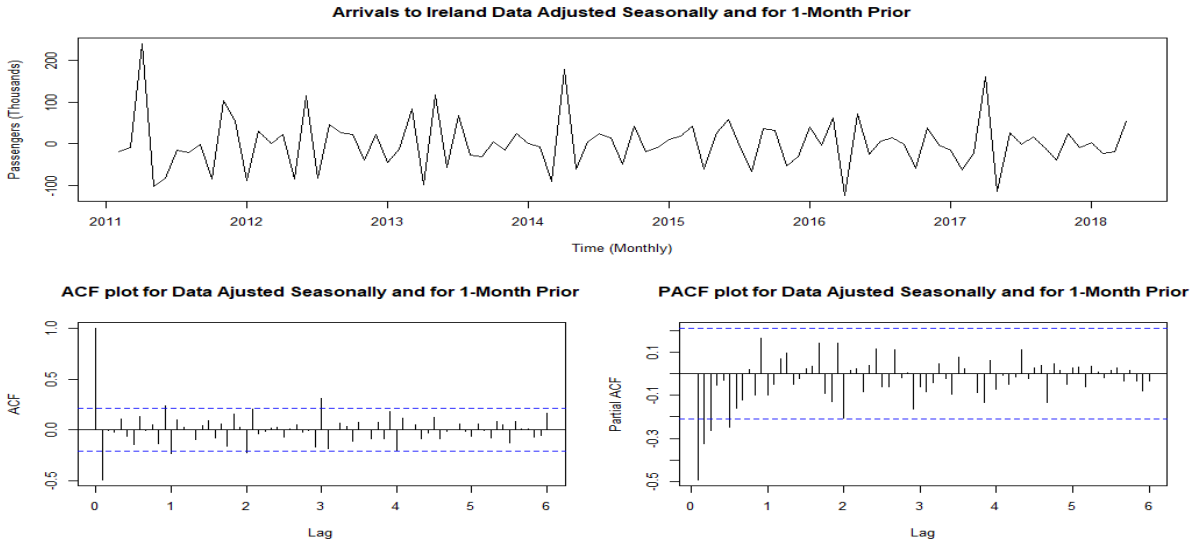

Figure 4: Seasonally and 1-month prior adjusted number of people arriving to Ireland each month along with the associated ACF and PACF

Model Fitting

Non-Seasonal Component: In Figure 4, the ACF cuts off at the start while the PACF appears to decay. This suggests p=0 and q=1. Seasonal Component: In Figure 4, the PACF is in between the bands for all lags k where k = 1,2,3,… while the ACF is outside the bands at lags 1,2,3,4 and then cuts off. This suggests P=0 and Q=4. Therefore it was decided to begin the model fitting with the model S-ARIMA(0,1,1)x(0,1,4)[12]. This model had an AIC of 922.68 and reasonable diagnostics. After exploring

similar models, S-ARIMA(0,1,1)x(0,1,3)[12] was decided upon as it had a lower AIC and very little difference in the diagnostics which are shown in Figure 5. The following is the fitted model output:

```
arima(x = ts_modelsa, order = c(0, 1, 1), seasonal = list(order = c(0, 1,
3), period = 12))

Coefficients:

          ma1     sma1     sma2    sma3
      -0.7034  -0.1299  -0.3576  0.3349
s.e.   0.0917   0.1724   0.1775  0.1649
sigma^2 estimated as 1892:  log likelihood = -455.65,  aic = 921.3
```

Model Criticism

Inspection of the time plot of the standardised residuals in Figure 5 shows no obvious patterns. There are two values exceeding 2 standard deviations, which may indicate outliers. The p-values for the Ljung-Box statistic aren't close to being statistically significant, however, there is room for them to improve. The ACF and PACF of the standardised residuals show no severe departure from the model assumptions, with only one value falling outside the bands. The normal Q-Q plot of the residuals shows that the assumption of normality is reasonable, with only one case largely departing from the reference line. The residuals lie within the bands of the cumulative periodogram.

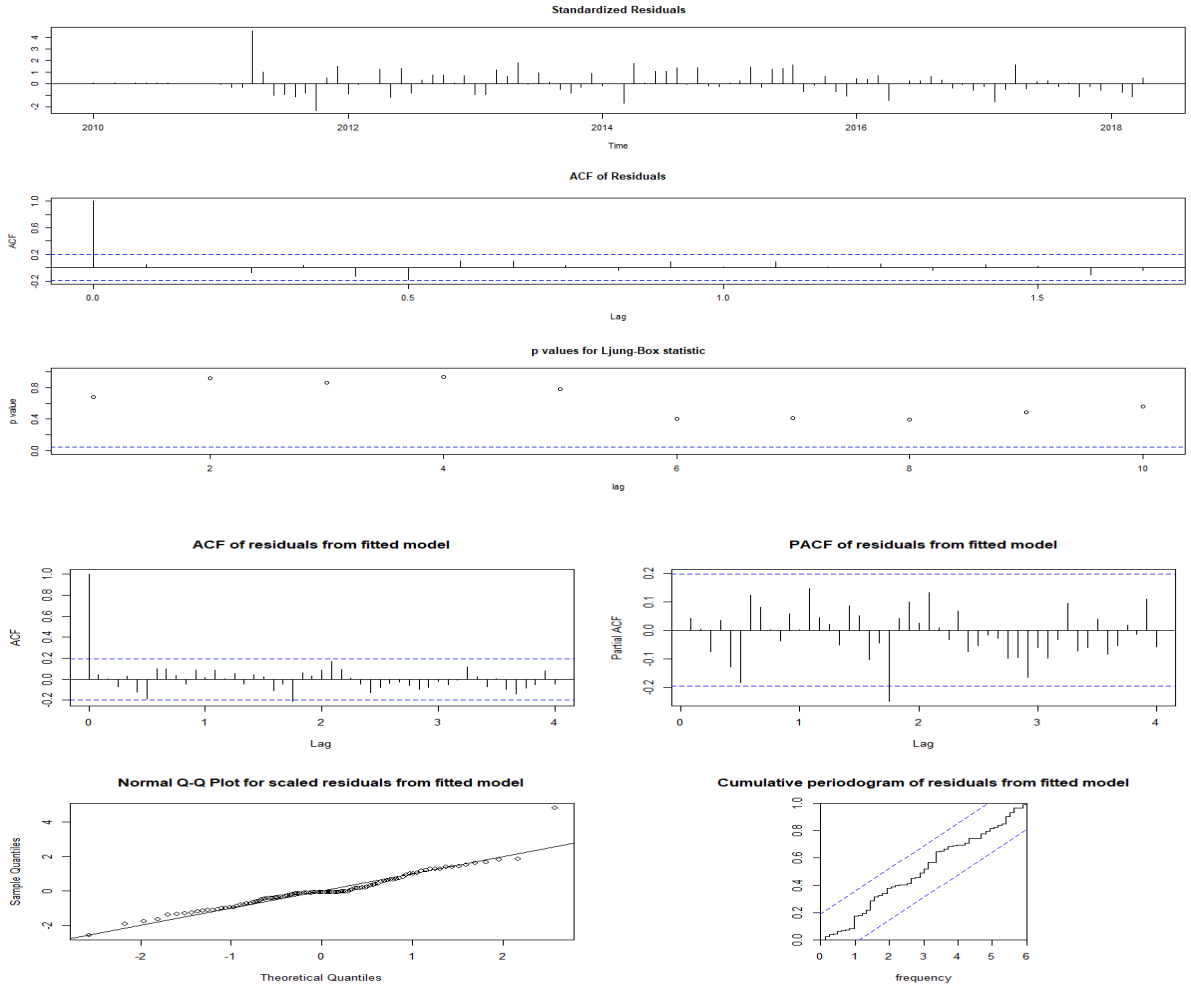

Figure 5: Diagnostics for the chosen model S-ARIMA(0,1,1)x(0,1,3)[12]

Forecasting

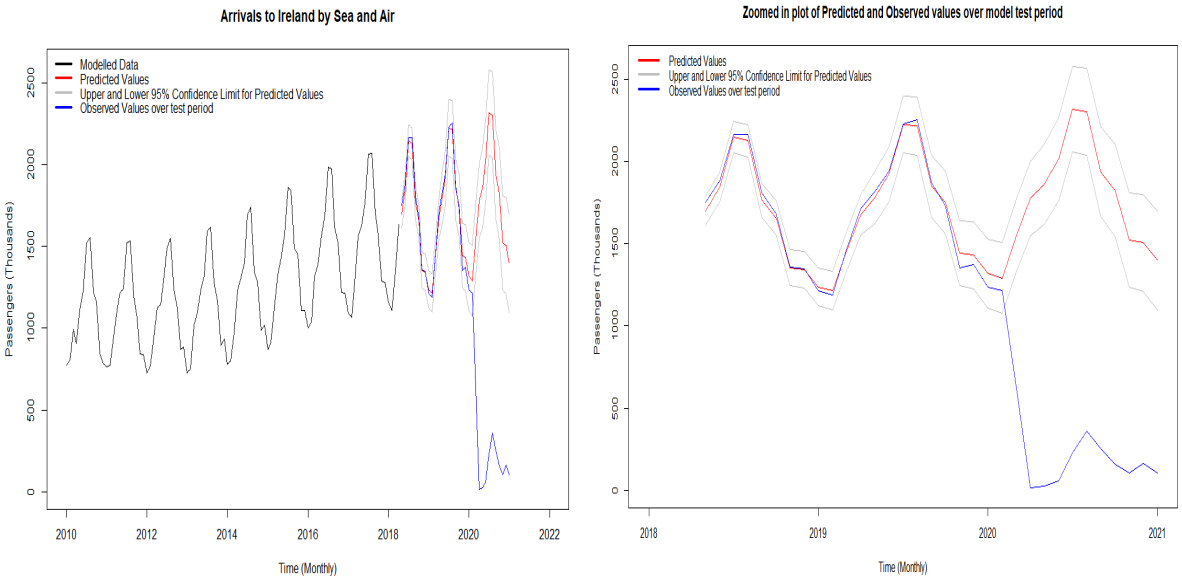

Figure 6: Model forecast and observed values from mid 2018 to 2021

The model performs reasonably well for 2018 and the first 8/9 months of 2019. In both years it followed the overall trend well. In both years it slightly underestimated the number of people arriving to Ireland during peak season (June and July) and in 2018 it sightly overestimated the off season in November and December. However, the observed values for this time period are still within the 95% confidence limits for the model predictions.

In the last 3 months of 2019, the effects of Covid-19 can be seen entering the observed values. By March 2020, when Ireland entered lockdown, the full effects of Covid-19 can be seen in the observed data. The model fails to predict such a shock and has very poor predictions for the remainder of 2020, with the observed values falling well outside the confidence limits for the model predictions.

Discussion

Compared to other fitted models, the chosen model of S-ARIMA(0,1,1)x(0,1,3)[12] has a lower AIC with good performance diagnostics. It may be possible to improve the diagnostics by trying a wider range of models, however, this introduces a risk of overfitting the model to the data used for fitting and predictive performance could become poorer for unseen data. The forecasts from the model were performing well until the shock to the system through Covid-19. This is an event the model (and the other models considered) would have never predicted given the data it was fitted with. It is unlikely that data extended further back into the past would have improved the performance of the model during such a shock event as shocks of this magnitude combined with accessibility of global travel are unlikely to have existed. This highlights the limitations of modelling a heavily seasonal system with little variation from year to year. However, in times of stability the model performs well, capturing the overall trend and the observed values fall within the 95% confidence limits for predictions.