ST4064 Assignment

Maxim Chopivskyy, 118364841

Dataset

The dataset used for this project is a monthly time series dataset of Irish Cattle exports. This dataset is publicly available on the Central Statistics Office (CSO) website [data.cso.ie].
The time series shows the monthly number of cattle exported from Ireland in thousands. This is for a period of 192 months ranging over the years 2006-2021 inclusive.
The time series is split into two groups. The first group contains the first 168 months and will have the seasonal ARIMA (SARIMA) model fit to it. The second group contain the last 24 months and will be withheld until the end. Forecasts will be generated from the SARIMA model and be compared against the withheld data.

Exports of cattle and beef is among Ireland's most profitable businesses. Bovine meat is one of Ireland's top two commodity exports, being valued at $1.9 billion as of 2021 [https://commodity.com/data/ireland/]. The Covid-19 pandemic has impacted many sectors of life in 2020, affecting not only the economy, but also the livestock industry. Lockdown restrictions around the world had decreased the overall meat production in many countries during the pandemic. It would be interesting to use this time series dataset to observe how Irelands cattle exports was affected by global situations.

The time series exhibits *seasonal* behaviour as can be seen in Fig 1. We can also see that the variance increases around the year 2009. A slight upwards trend can be observed happening from 2018 onwards. The most exports happen in the months March, April, May and have the highest variance as seen in Fig 2. Spring has the best conditions of the year for Ireland to export cattle. The `decompose()` R function can easily give us a good view of the trend and seasonality of the time series, shown in Fig 3. Around 2010 and 2020 were Irelands most profitable years in exporting cattle.

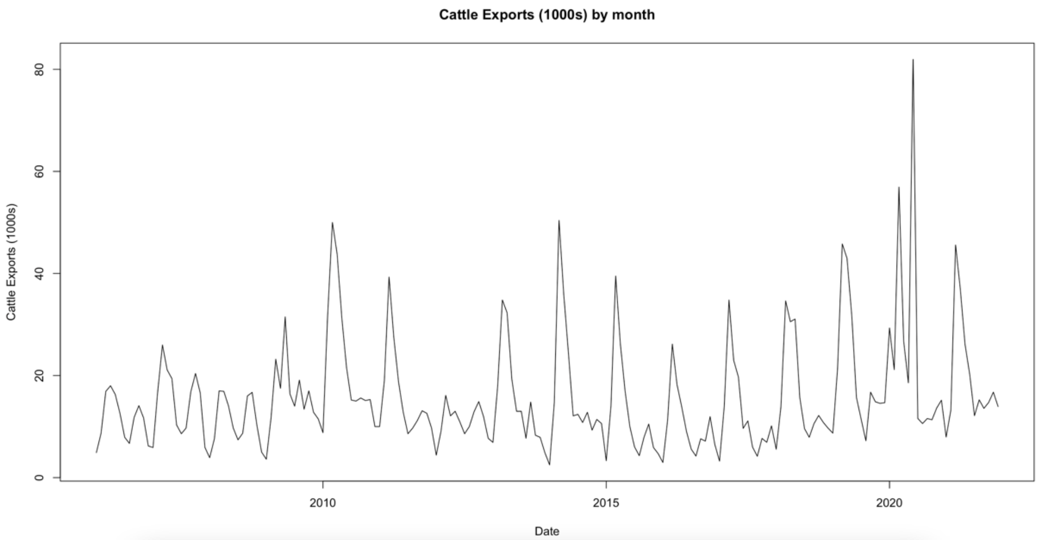

***Fig 1**: Irish Cattle Exports (1000s) time series, 2006-2021 inclusive*

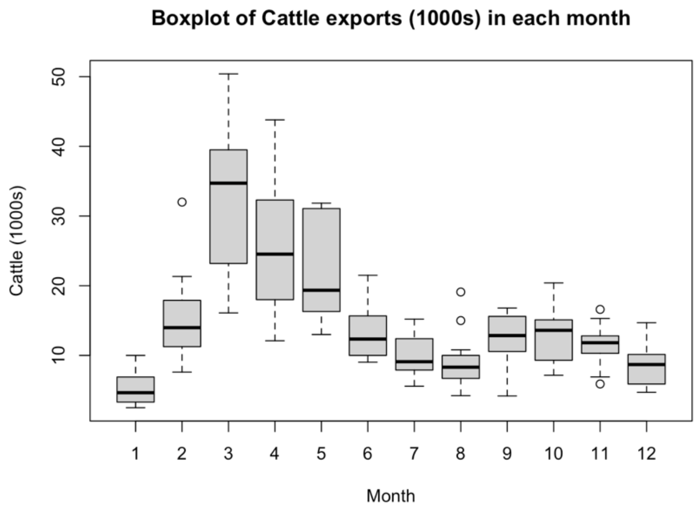

Fig 2: *boxplot of Irish Cattle Exports (1000s) by month*

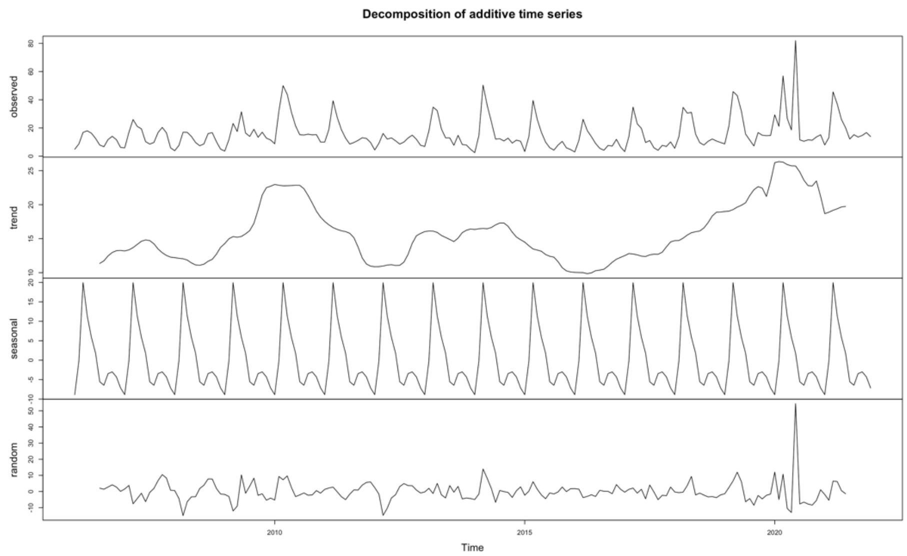

Fig 3: *Decomposition of Cattle Exports time series*

From our above analysis, it seems Irelands cattle exports were not badly affected in any way by the Covid-19 pandemic. In fact, the only noticeable dilemma is the large spike in June 2020 in Fig 1.

Reduction to Stationarity

As seen in Fig 1 variance appears to increase after 2009. We log transform the time series to make the variance constant. This is done with the `log()` function. Log transforms is effective at removing exponential variance in time series data.

To remove the slight trend, the log data is differenced at lag 1. This removes the dependence on time. Next, to remove the seasonality, the data is differenced again but at lag 12. These are done using the `diff()` function. The time series is stationary now, as shown in Fig 4, where X_i is log(cattle exported (1000s)) and Y_i is the stationary time series of cattle exported (1000s).

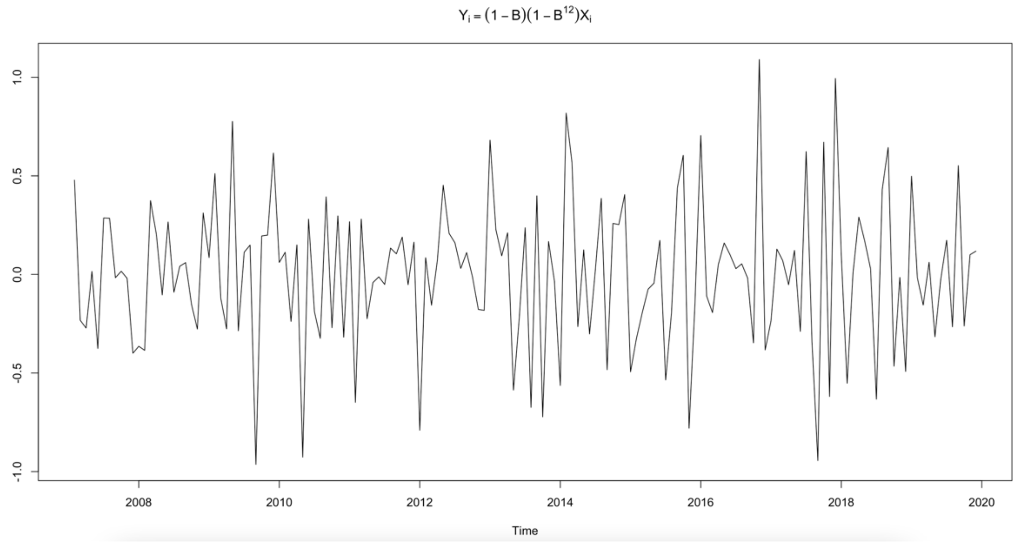

Fig 4: *stationary transformed Cattle exports (1000s) time series*

Model Fitting

The ACF and PACF of the stationary time series is shown in Fig 5. The seasonal ARIMA (SARIMA) model is composed of its non-seasonal and seasonal components.

Non-seasonal parameters: One could interpret it as the ACF tailing off, while the PACF cuts off after lag 1. This suggests an ARIMA(1, 0, 0) model. Another interpretation is that both the ACF and PACF tail off after lag 1. This suggests an ARIMA(1, 0, 1) model. It is unclear which to choose, so we will fit both models and compare them.

Seasonal parameters: The seasonal ACF cuts off hard after lag s1 (s = 12). The seasonal PACF is tailing off. This suggests the seasonal component follows an ARIMA(0, 1, 1) model.

Combining the non-seasonal and seasonal components gives us two possible models: model 1, $SARIMA(1,0,0)x(0,1,1)_{12}$and model 2, $SARIMA(1,0,1)x(0,1,1)_{12}$. The diagnostic results of both models are shown in Table 1. The AIC and BIC of model 1 are significantly smaller than those of model 2. Model 1, $SARIMA(1,0,0)x(0,1,1)_{12}$ will be used to fit the Cattle exports time series.

The estimated model 1 is defined as:

$$Y_t = 0.691_{(.058)}Y_{t-1} + Y_{t-12} - 0.691_{(.058)}Y_{t-13} + e_t + 0.742_{(.073)}e_{t-12}$$

Where the values in parentheses are the standard errors of the corresponding coefficients.

$SARIMA(1,0,0)x(0,1,1)_{12}$	$SARIMA(1,0,1)x(0,1,1)_{12}$
AIC: **53.8** BIC: **64.948** Sigma^2: 0.0755	AIC: 54.71 BIC: 68.907 Sigma^2: **0.0751**

***Table 1**: Diagnostic results of two ARIMA models*

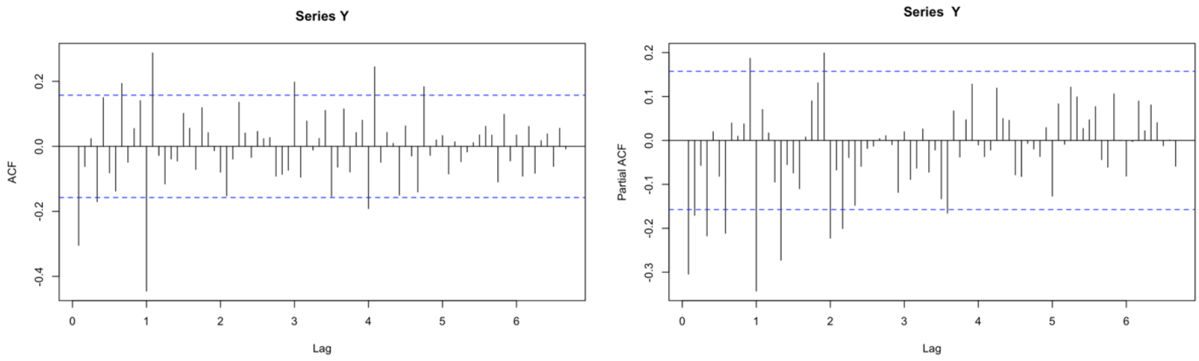

***Fig 5**: ACF and PACF of stationary time series Y*

Model Criticism

The residual and diagnostic plots are shown in Fig 6. The standardised residual plot on the right shows no obvious patterns with only one point exceeding 3 standard deviations. This is complemented by the cumulative plot. In the cumulative plot, the fitted line goes from (0,0) to (6,1) without crossing the dotted blue lines, meaning the residuals is a random process. This is good. The ACF of residuals cuts off at lag 1, thus the residuals are not autocorrelated. This is good. The p-values for Ljung-Box statistic are all above 0.05. No autocorrelation in any group of autocorrelations in the time series is different from 0, i.e. the model does not exhibit lack of fit.

The model appears to fit well.

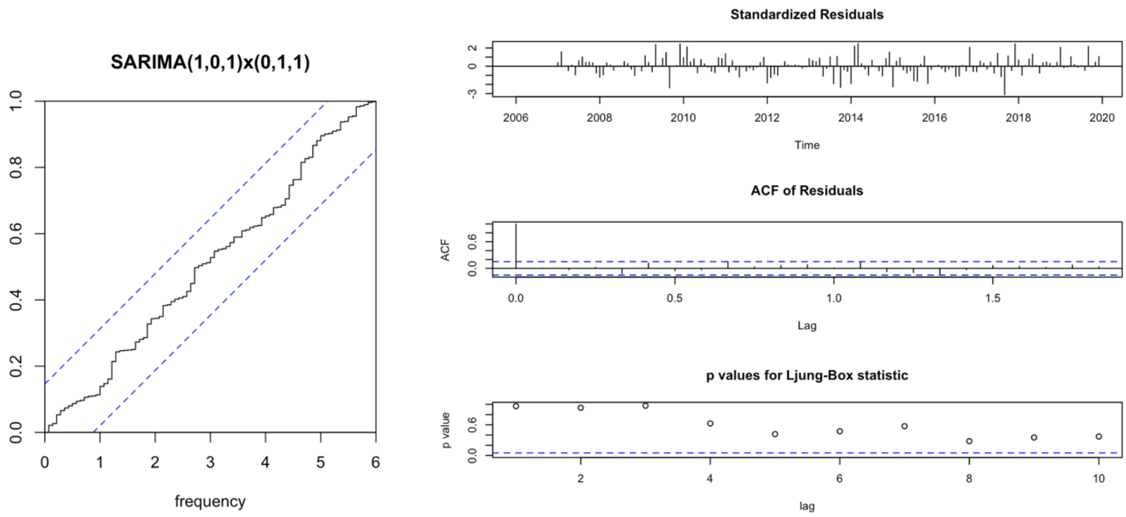

***Fig 6**: cumulative periodogram (left) and diagnostic plots (right) of fitted SARIMA model*

Forecasting

The plot of withheld data compared to the forecast values are shown in Fig 7. The actual forecast values are not shown but the 95% forecast interval is shown. The model predicts the true values have 95% chance being in the interval. We can see the model does generally well at predicting the true values, except for June 2020. As mentioned in the Dataset section, this is an unexpected outlier in the data which is hard for the model to predict.

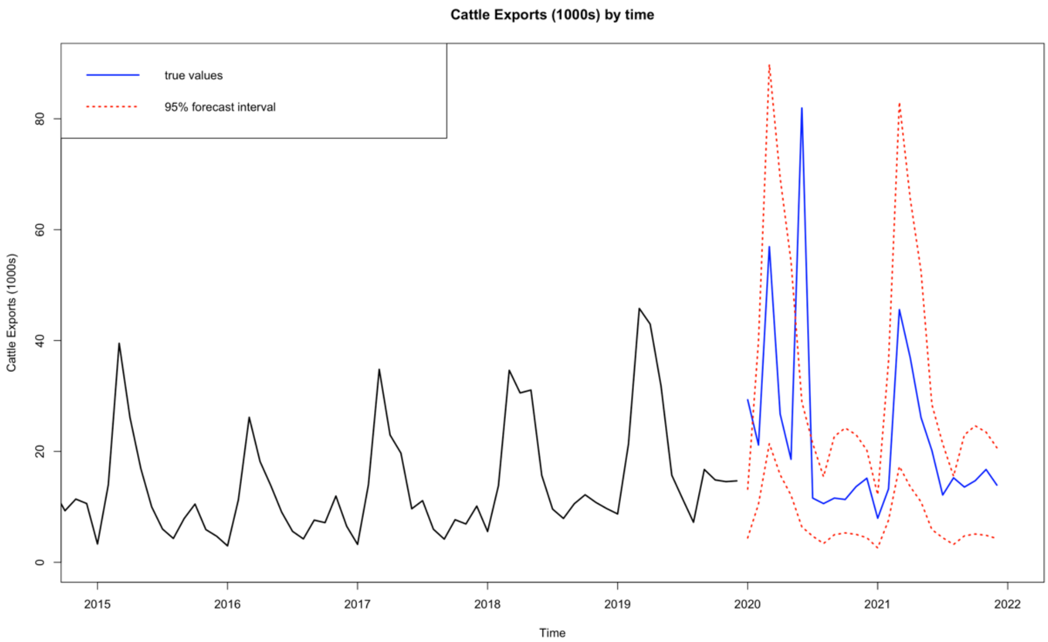

Fig 7: *time series plot with true values compared against 95% forecast interval, 2015-2022*

Discussion

The purpose of this project was to use the ARIMA model in R to forecast unseen data in Irelands Cattle exports. We also discovered interesting preliminary observations of the time series dataset. The upwards trend of cattle exports after 2018 was noticed. We also noticed which months of the year were busier than others, i.e. the seasonality.

Next, we outlined how to reduce the dataset to Stationarity, using a combination of log transform and differencing. The ACF and PACF of the stationary time series was used to identify the SARIMA model parameters and fit the model to the log time series. The final model chosen was $SARIMA(1,0,0)x(0,1,1)_{12}$. The model was confirmed to fit the data well, using the cumulative periodogram, studentised residual plot and other diagnostic plots. These plots showed the residuals are not correlated with each other and are not too large. Finally, a forecast of the withdrawn data was produced using the fitted ARIMA model. The forecast was generally accurate for most of the data but did not account for the large spike in June 2020.

An advantage of our SARIMA model is that it recognizes the seasonal patterns of the data, not just the upward trend. The SARIMA model has high accuracy and is simple to create. A disadvantage of the model is that it cannot account for outside effects on the time series model. For example, June 2020 had a huge spike which our forecast did not predict. This may have happened because of the Covid-19 pandemic. Unfortunately, there is not much that can be done to amend this weakness. Real life time series data requires knowledge of world events to predict the time series more well. ARIMA models cannot account for this.