Final Assignment
Statistics for Data Analysis

Peter Nolan PGDDA (x22154116@student.ncirl.ie)

I. Introduction

R was implemented using *RStudio 2022.12.0+353* on MacOS. The analysis was also done in parallel using a Jupyter notebook in the cloud hosted on the Google Colab service using an R core; this preserves the results of analysis without having to run it.

Google Colab in the cloud as back up and with GPUs to provide extra processor power.

A GitHub repository also holds the code and the analysis files, which can be downloaded or, for the Jupyter notebook, run in place, via Google Colab, at `https://github.com/dpnolan/taba`.

Data was taken from the course Moodle site for both parts.

II. Time Series Analysis

A. Descriptive Statistics

In this section, I estimate and analyse time series models based on monthly observations of departures from Irish airports. The time series data itself is the only input into this analysis, no explanation in other variables is available from the data set given and the models I use do not look for one, instead looking to capture and then to understand the other

Data was downloaded in the file named *Departure.csv*. The file has 153 observations of the numbers in every calendar month of passengers departing from all Irish airports, reported in thousands.

A simple graph of the departures against dates in figure 1 gives a visual summary behaviour that we should aim to understand and forecast. In later sections, I use different time series tools to test, model and forecast these numbers and attempt to evaluate how they can be used in practise.

I manually checked that there was one for every single month from and including January 2010 to and including September 2022, with all those months having one observation. No NA or otherwise unreadible results not suitable for numerical analysis were found in a visual inspection of this data.

To allow processing with the specialist R libraries for time series, I transformed the data into a *ts* object defined in the library *tseries*. Apart from the observations, the start and end dates were included in the definition, along with their frequency of 12 i.e. monthly.

Behaviour in the data was visible that we can use to understand and then to forecast the time series. Four aspects in particular, I believe, we have to make modelling choices about:

1) Seasonality: Seasonality is clearly visible and the same pattern consistently shows in all years, with the summer holiday months of June, July and August showing the highest passenger numbers in each year, with January and February consistently the lowest. For making decisions involving airport operations such as staffing or retail businesses, this behaviour is likely to be always relevant.

2) Trend and Cyclical Factors: While appearing level from 2010 to 2013, from 2013 to 2020, our chart of departures to time shows a pronounced growth trend, with minimum and maximum numbers each year increasing every year, reaching the peak value in the dataset of over two million in July 2019. The growth may be tied to economic growth and increased population in Ireland over the period. 2010 saw the bailout of Ireland's government and banking sector by the ECB, IMF and European Commission, with severe unemployment and loss of consumer confidence. After 2013, factors such as rising incomes across Europe, increasing population in Ireland and improved business conditions may explain the consistent rise over several years up until 2019. Operating on a longer-term than the seasonality within the year, this behaviour would be important for longer term decision making such as building terminals and buying planes.

3) Pandemic Shock: The COVID19 pandemic's impact is clear in the first quarter of 2020, as the disease spread from China to Europe. By April 2020, traffic at Heathrow airport in London was down by 97% and some 2.7 billion workers, 81% of the global workforce were under some form of lockdown [1]. The economic fallout from the pandemic were all apparent when the departures numbers had a rapid fall in Q1 of 2020 to their low point in this dataset in April 2020 of only 12,800.

Shocks like this to air travel are much rarer and hence much harder to plan for or model. One such was the grounding of air travel and slow recovery after the 9/11 terrorist attacks in 2001. Another was the outbreak of SARS, the first major coronavirus epidemic in 2004 and its spread through air travel [2]. However, the deaths and disruption have probably been unique since the global influenza pandemic of 1918. Pandemic insurance was offered by Munich Re but demand was very low and not part of regular operations planning. For our models then, this may not be something which can be incorporated.

4) Recovery: Post-pandemic recovery seems to have seen different behaviour from the crisis conditions of the first few months of 2020, which saw the lowest value in the time series with 12,800 departures in April. The remainder of 2020 showed departures remaining well below previous values, but peaking in the summer months as before, but reaching a

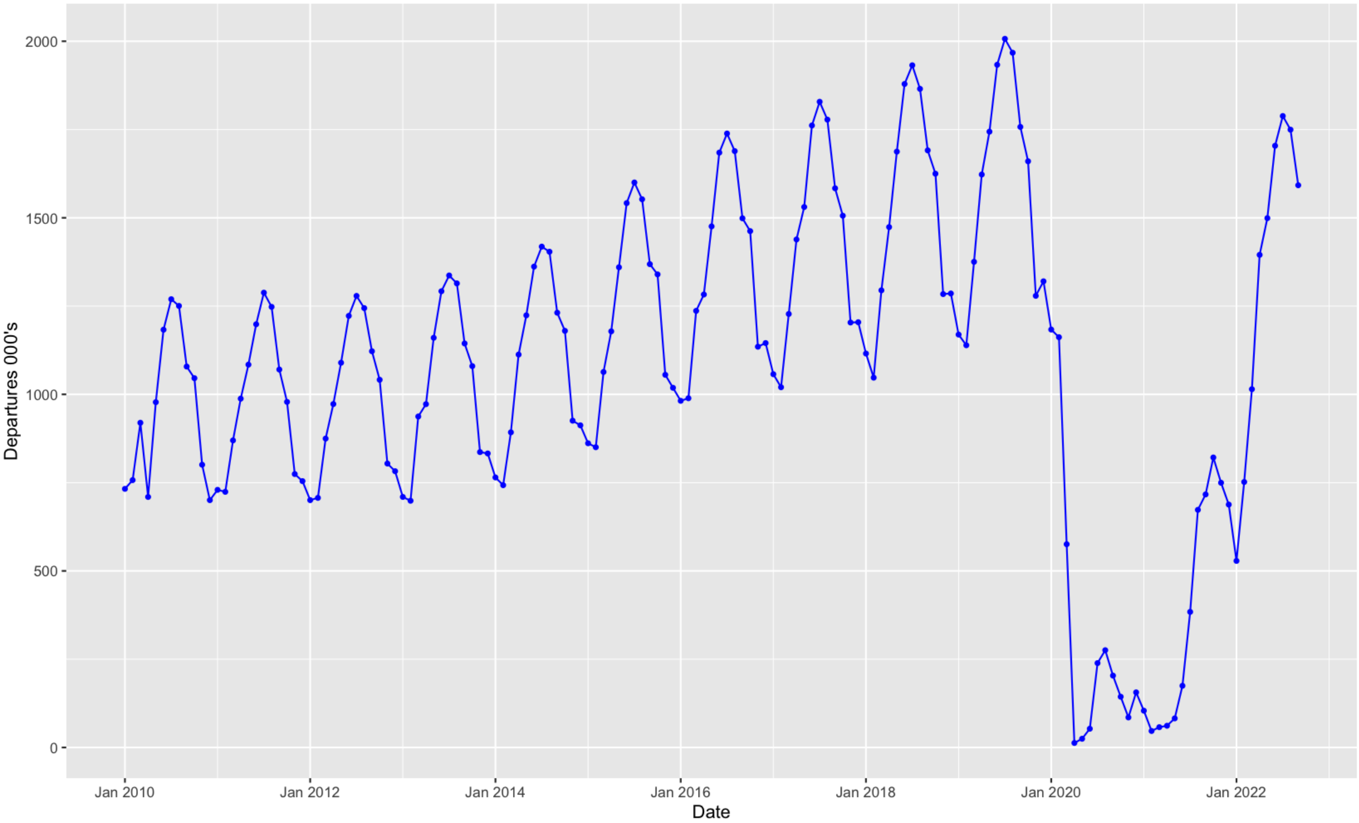

Fig. 1. Departures from Irish airports by month

maximum level for the year at only 275,500 in August.

Unemployment had risen substantially everywhere in the early stages of the pandemic, but emergency welfare payments soon began to provide financial support. From that low base, 2021 saw a partial recovery back to about half 2019 numbers by the autumn, peaking at 821,000 departures in October compared to 1,660,000 in October 2019. Starting from a lower point than before, 2022 shows an annual maximum in the summer months of some 1,788,200, close to the highest observations recorded in 2016 and 2017. This may be explainable by the economic rebound, as money saved during the pandemic is spent and as travel becomes possible again without onerous testing and quaratine.

To me, this points towards the effect of the pandemic largely fading away and the regular seaonality and cyclical behaviour returning instead.

5) Descriptive Statistics: Descriptive statistics were calculated. The median and mean departures per month were 1,134,900 and 1,088,600 travellers per month. Skewness was estimated for all observations at around -0.43, so leading to the right and kurtosis as 2.89, so with the central peak and tails close although not matching those of the normal distribution, as shown in the histogram in figure **??**.

The observations were grouped by calendar year and a box plot generated for each, as shown in figure **??**. This highlights the upward trend 2013–2019 and the discontinuity and possible non-stationarity between 2019, before the pandemic, the impact of the pandemic in 2020, and then the strong recovery in departures in 2021 and 2022.

A number of t-tests were performed on these sub-samples. One rejected the two-sided null-hypothesis of equality between the mean of departures for calendar 2013 and 2019 with p.value of 7.696756e-05. A second, also rejected equality of means for 2019 and 2020, the year of the pandemic crisis. A third rejected mean equality for the means for 2020 and 2022, the year that looks to be the second year of the recovery. All results reinforce the idea that we face three distinct periods in our time-series sample data.

B. *Exponential Smoothing*

Common solutions for forecasting are either exponential smoothing models that I cover in this section or ARIMA models that I discuss below in the next section II-C. The treatment here closely follows [3] chapter 7.

Given that our visualisation of the time series shows a what is likely to be a trend and seasonality, a sensible starting point is an exponential smoothing model that combines both these with a random element, namely the Holt-Winter (hereafter 'HW') model in its additive form i.e. from [3], section 7.3:

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)} \qquad (1)$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \quad (2)$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \quad (3)$$

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \quad (4)$$

where α is the smoothing parameter for the level, β for the trend and γ for the seasonal adjustment component. ℓ_t, the smoothed series at t, y_t the raw observed value of the series at t, h the number of periods ahead we are forecasting for, b_t the estimated trend component at t,

The level equation 2 shows a weighted average of a seasonally-adjusted observation in the first part and a non-seasonal component forecast as at t.

The trend equation 3 shows how the current estimate of the trend b_t is specified as a weighted average of the change in the smoothed series and the last value of the trend component.

The seasonal equation 4 shows an average weighted by γ between the current seasonal index component and the seasonal index of the same season m time periods ago.

A damping parameter ϕ was included in the model by default, to prevent unlimited growth in the trend component.

With ETS, solutions are estimated so as to minimise the sum of squared residuals of the model.

Using the ETS function from the *forecast* library, with the *model = 'ZZZ'* option allowing the software to automatically select the type, whether additive or multiplicative, for the error term, the trend and the seasonality components respectively, the system chose an additive model *departs_fit1* for all rather than multiplicative. This was in line with prior expectations, as the graphical representation of the time series shows no noticeable increase in variance in the years before the arrival of the pandemic.

The model *departs_fit1* scores an AIC corrected for small sample bias, AICc, of 2235.89 and Root Mean Squared Error ('RMSE') of 105.35. The smoothing parameters were $\alpha = 0.9999$, $\beta = 0.1413$ and $\gamma = 1 \times 10^{-4}$. A damping component was estimated with $\phi = 0.8$. The same parameter estimates and scores came from running the ETS function with the same inputs but with the model explicitly set as additive for all components i.e. 'AAA'.

A third model without a damping parameter was also run, *departs_fit3*. This generated worse scores than the earlier *departs_fit1* model, with AICc of 2245 compared to 2235 and RSME at 109.567 instead of 105.35.

Forecasts were generated for 8 months past the end of the dataset. The model *departs_fit1* with a damping parameter for the trend specified generated a range of forecasts of 746.3 to 2331.54 with a point forecast of 1538.9. The model *departs_fit3* with the same setup except for no dampening, shows a higher point forecast 1904.69 and a wider range of 941.48 to 2867.9.

C. SARIMA Models

In this section, I work to fit a model of the SARIMA (Seasonal Autoregressive Integrated Moving Average) type to the time series data. While the exponential models fit trend and seasonality measurements, the ARIMA class focuses on identifying and estimating relationships between values of the time series at different times. The treatment here largely follows [3] chapter 8.

To begin, we can take first differences of our time series until it is stationary. The function *ndiffs* from R library *forecast* tests for stationary in the time series, whether its properties depend on the time of observation: In this case, the tests, using each of the three options available, namely a KPSS, Augmented Dickey-Fuller and Phillips-Perron all return a zero value, indicating that the time series is stationary and does not need differencing before applying an ARIMA model.

Doing the *nsdiffs* test, applying the same logic but at time lags matching the seasonality of every 12th value in the case of our monthly data, returns a value of 1, showing non-stationary between every 12th observation in our series. Together the two test results indicate that our data can be represented by a model of the form ARIMA(p,0,q)(P,1,Q)[12].

We may then continue on to plot the autocorrelation and partial autocorrelation functions of the time series, shown in figure **??** and **??**. The ACF plot shows the correlation of the time series with its own past values: Here, several lag intervals show statistically significant levels different from zero of correlation with a gradual reduction over time, which points to a model with one or more autoregression terms, showing that previous observations in our time series have an influence on our current observation.

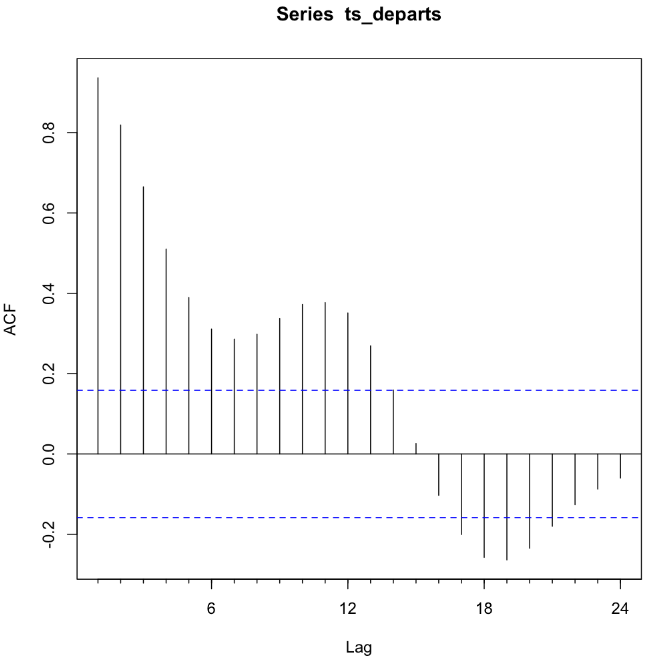

Fig. 2. ACF plot

The PACF graph shows the correlation between the time series and values at each time lag not explained by any lower order autocorrelations. The large and statistically significant partial autocorrelations at the lags of 1, 2, 3 and 5 are visible the PACF may indicate autoregression at those lags. The significant autocorrelations at lags of 12 and above may indicate AR or MA processes among the same months' values from different years also.

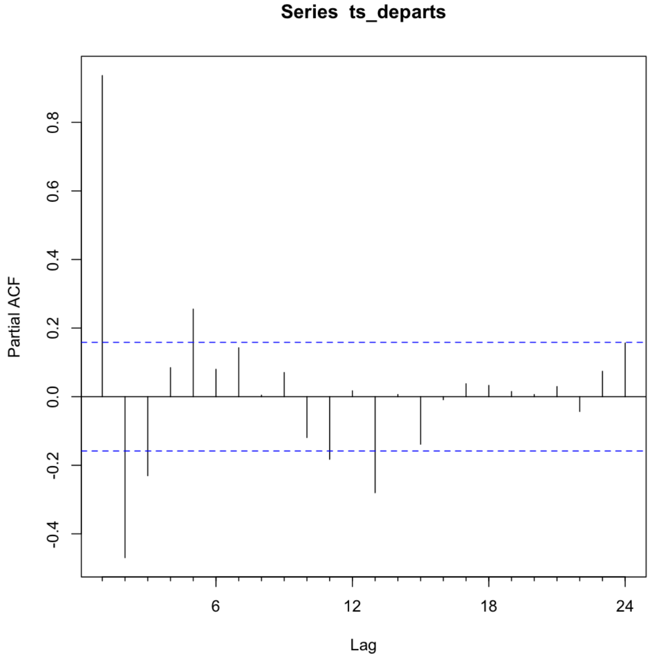

Fig. 3. PACF plot

Proceeding to test and select from a number of possible SARIMA models, we may use the *auto.arima* function, again from the *forecast* library, for an automated search. Given the large number of possible model specifications, searches for the best fit as determined by the AICc score, with the stepwise flag set to FALSE to specify the an exhaustive search and start.p and start.q both equal to 1 so as to search for the most models possible. This returns a model specification of ARIMA(2,0,0)(0,1,1)[12]

Looking at the residuals with the *checkresiduals* function, we can test if the errors left after applying our SARIMA model to the time series show any systematic structure not captured by the model. The plot is shown in figure 4 and the lack of any statistically-significant autocorrelations would support our hope that the residuals behave like white noise.

Running a Ljung-Box test, the null hypothesis of independence among the residuals is also upheld, with the p-value of 65.94%, again supporting the view that no autocorrelation and hence no systematic predictability remains within the residuals to capture in an updated version of our model.

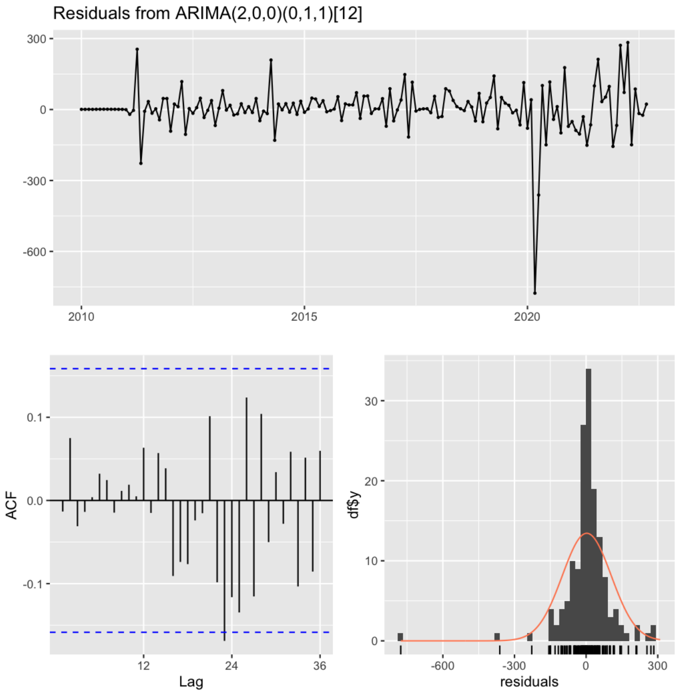

Fig. 4. Testing residuals from our SARIMA model

D. Simple Time Series Models

III. Applying a Forecasting Model

The percentile breakdown of Reddit polarity scores in table I confirms the prevalence of neutral sentiment in the data set:

TABLE I
Reddit Posts: Polarity Percentiles

Sentiment	**Textblob**	**VADER**
Positive	39.9%	42.4%
Negative	29.6%	15.7%
Neutral	28%	38.1%

IV. Conclusions and Future Work

Conclusions
Word Count = xxxx, excluding references

References

[1] A. Tooze, *Shutdown: How COVID Shook the World's Economy*. Penguin Allen Lane, 2021.

[2] K. T. Greenfeld, *China Syndrome: The True Story of the 21st Century's First Great Epidemic*. Harper Collins, 2006.

[3] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. OTexts, 2018. [Online]. Available: https://otexts.com/fpp2/