**Predicting Covid-19 UK Cases (Time Series Analysis)**

**Introduction**

The study aims to use the time series model, ARIMA(*p*, *d*, *q*), to produce the forecast of the daily number of cases of COVID-19 in the four UK nations for 29-31/03/2021 using RStudio. The four datasets we used to fit the model are provided by the UK government website (Daily cases as reported in <https://coronavirus.data.gov.uk/details/download>). Each dataset for each nation consisted of 434 daily reported cases for 03/01/2020-11/03/2021. Since there was a declining trend at the start of March, our forecasting will provide important guidelines for the government decision-maker.

**Methods**

The methods consisted of four parts. For **data exploration**, we first reversed and plotted the data, the mean-variance relationship and ACF graph to understand the patterns. We decided if we need to use data transformation to stabilize the variance. *(****Q1****)* For **model selection**, there are two ways. One is **selecting model orders ourselves**. There are three steps involved. **1.** To choose *d*, if the data are non-stationary, we differenced the data until the plot appears stationary. Since it is relatively subjective, we then used *ndiff* function in package *forecast* and two unit root tests (*ur.df* function in *urca* and *adf.test* function in *tseries*) to decide the optimal option. **2.** To choose *p* and *q*, we plotted the ACF and PACF of the differenced data and determined possible candidate models. **3.** Since AICc is most suitable for forecasting, we selected a better one with the lowest AICc. Another one is to **use automated algorithm**, *auto.arima* function in *forecast*. In this study, we considered both ways *(****Q1****). (****Q2****)* For **diagnostics**, we checked the residuals for the model by plotting the ACF, a portmanteau test to assess correlation and the Jarque-Bera test or QQ-plot to assess normality. If they do not look like white noise, we then tried a modified model and did analysis again *(****Q2****)*. For **forecasting**, once the residuals look like white noise, we calculated forecasts by using *forecast* function in *forecast*.

**Results**

For **data exploration**, since they are count data with increased variance for higher means, for all nations, we used log(x+1) transformation which considered the meaningful 0 cases. For **model selection**, for all nations, log data with first differencing are not stationary shown from the *adf.test*. Since the test detrends again the detrended data, the result is not valid. Thus, considering all other criteria, we chose d=1. For manually selected models, except for England and Scotland with order (4,1,3) and (2,1,3), the parameters for other nations are larger than 6, which seems overfitting. Since simpler models are more robust for forecasting, we firstly chose the models with fewer parameters. *(****Q2****)* For **diagnostics**, for all nations, the residuals are not correlated and most of the lines from residual ACF graphs did not exceed the boundaries; but the residuals are not normal. We then tested other models and all results seem identical*(****Q2****)(****Q1****)* Although AICc is slightly larger using *auto,arima*, simpler models are easier to interpret and overfitting can be a problem for forecasting.Thus, the best model for England, Wales and NI are ARIMA with order (1,1,2), (4,1,5), (4,1,3), respectively, from *auto.arima*. For Scotland, the residual ACF plot for order (2,1,3) performs better than (1,1,4) by *auto.arima*. We set *approximation=FALSE* in *auto.arima* and got ARIMA(2,1,3), which then seems to be the best one *(****Q1****)*. For **forecasting**, *(****Q3****)* the rounding forecasts for 29-31/03/2021 are 4554, 4543, 4532 (England); 170, 163, 166 (Wales); 179, 179, 181 (NI); 523, 549, 583 (Scotland) *(****Q3****)*.

**Conclusions**

*(****Q3****)* The observations for 29-31/03/2021 with forecasting differences in the bracket are 4112(442), 3384(1159), 3327(1205) (England); 125(45), 94(69), 60(106) (Wales); 65(114), 151(28), 123(58) (NI);

352(171), 411(138), 542(41) (Scotland) (as cited above). Since we used data before 11/03/2021, although there is a slight difference, our models seem useful *(****Q3****)*. But they did not pass normality tests and log transformation has issues for prediction intervals. We tried Box-Cox transformation but the performance was not improved. The data seems to have seasonal effects; thus further study could investigate on seasonal ARIMA or ETS model, and then check the forecasting performance.