MONASH University

Office Use Only

Semester One 2020
Examination Period

Faculty of Business & Economics

UNIT CODES: **ETC3550/ETC5550**

TITLE OF PAPER: **Applied Forecasting**

EXAM DURATION: **3.5 hours**

This is an open book exam. You may consult the online textbook and the materials used during the unit. You may not communicate with any other person during the exam.

The exam contains *five* sections. *All* sections must be answered. The exam is worth 100 marks in total.

SECTION A

Write about a quarter of a page each on any **four** of the following topics. (Clearly state if you agree or disagree with each statement. No marks will be given without any justification.)

1. The disadvantage of using a test set for choosing a forecasting model is that it uses only a small proportion of the data.
2. The best forecasting models adapt rapidly to changes in the trend and seasonal patterns.
3. With STL decompositions and ETS models, we always need to transform our data before estimating the components.
4. The mean of a stationary AR(3) process

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t,$$

   where $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, is equal to c.
   (You can write out your answer elsewhere and upload it as an image if you prefer)
5. ARIMA models are better than ETS models because there are more possible models available.
6. Regression models are not very useful for forecasting because you have to forecast all the predictors as well.

Total: 20 marks

— END OF SECTION A —

SECTION B

Figures 1, 2 and 3 relate to the number of students (in thousands) arriving in Australia from China over the period January 2005 – February 2020.

1. Using Figures 1, 2 and 3, describe the student arrivals from China. Carefully comment on the interesting features of all three plots.

4 marks

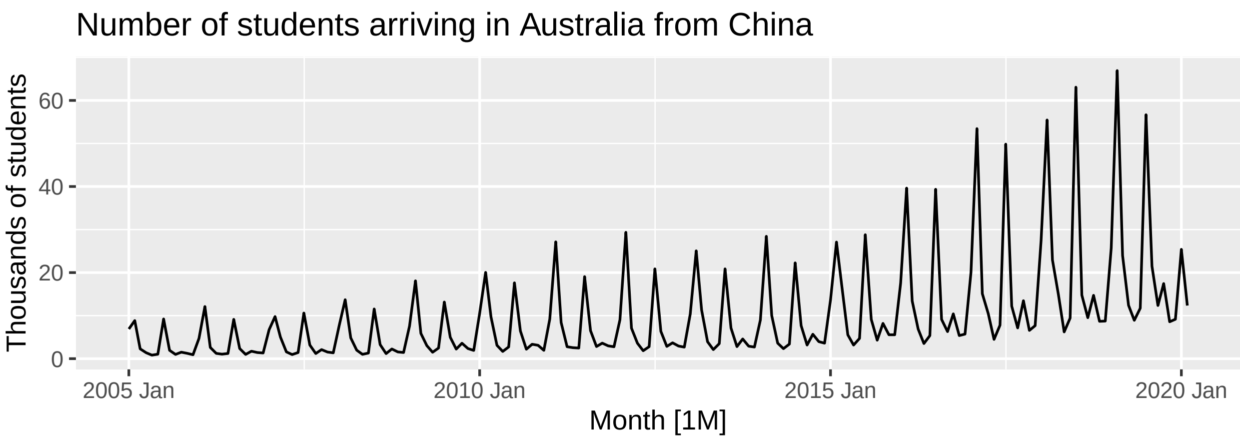

Figure 1:

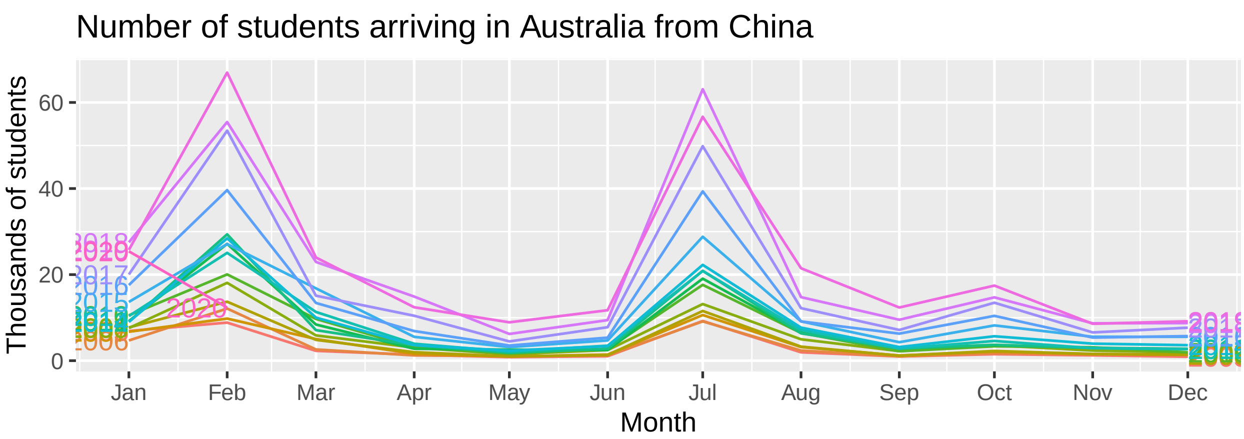

Figure 2:

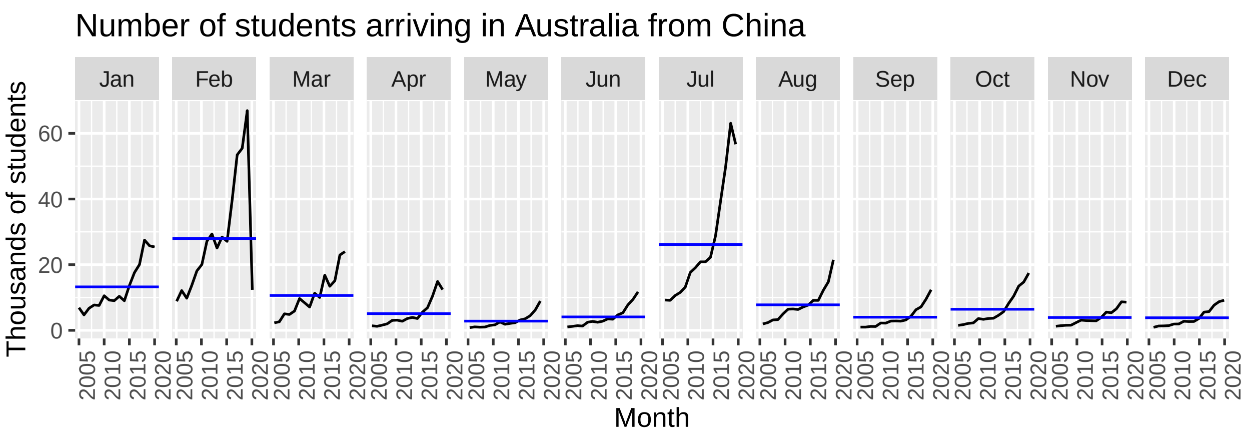

Figure 3:

2. Using the code below, describe what is plotted in all panels of Figures 4 and 5. Comment on the default settings for `window` in `trend()` and `season()`, and the effect of `robust=TRUE`. Which settings would you consider appropriate in this case?

6 marks

```
ch_edu_arrivals %>%
  model(STL(log(Count))) %>% components() %>%
  autoplot() + ggtitle("STL decomposition")
```

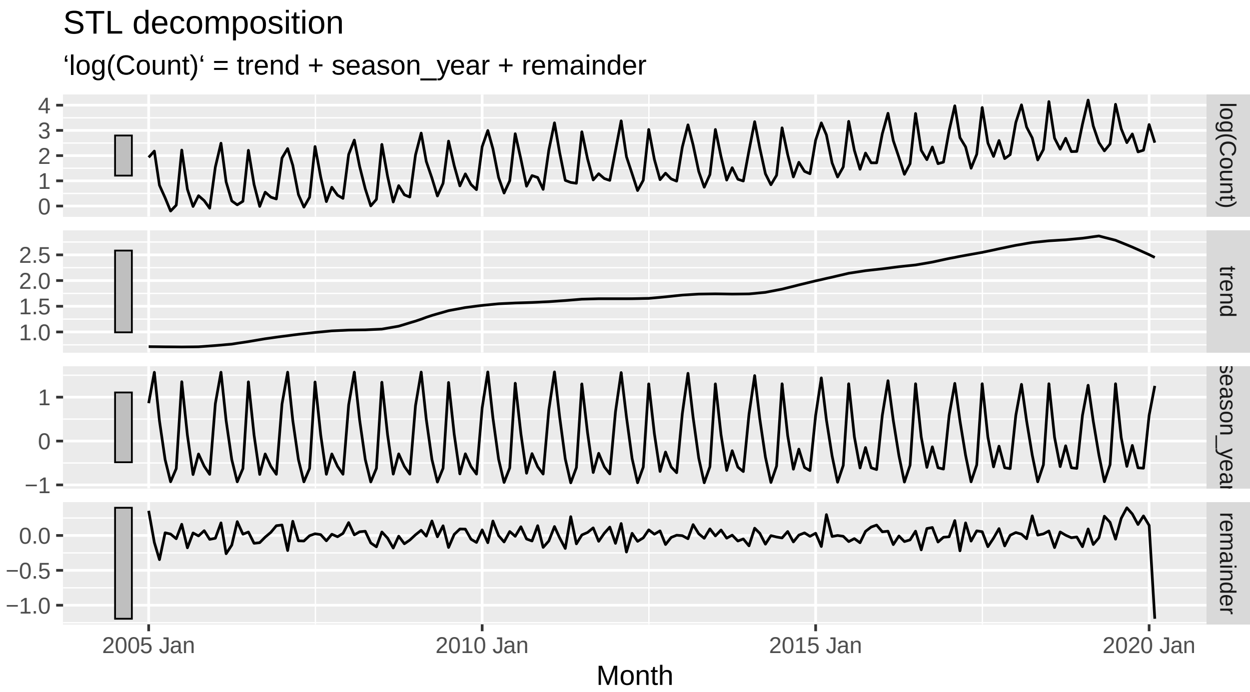

Figure 4:

```
ch_edu_arrivals %>%
  model(STL(log(Count), robust = TRUE)) %>% components() %>%
  autoplot() + ggtitle("Robust STL decomposition")
```

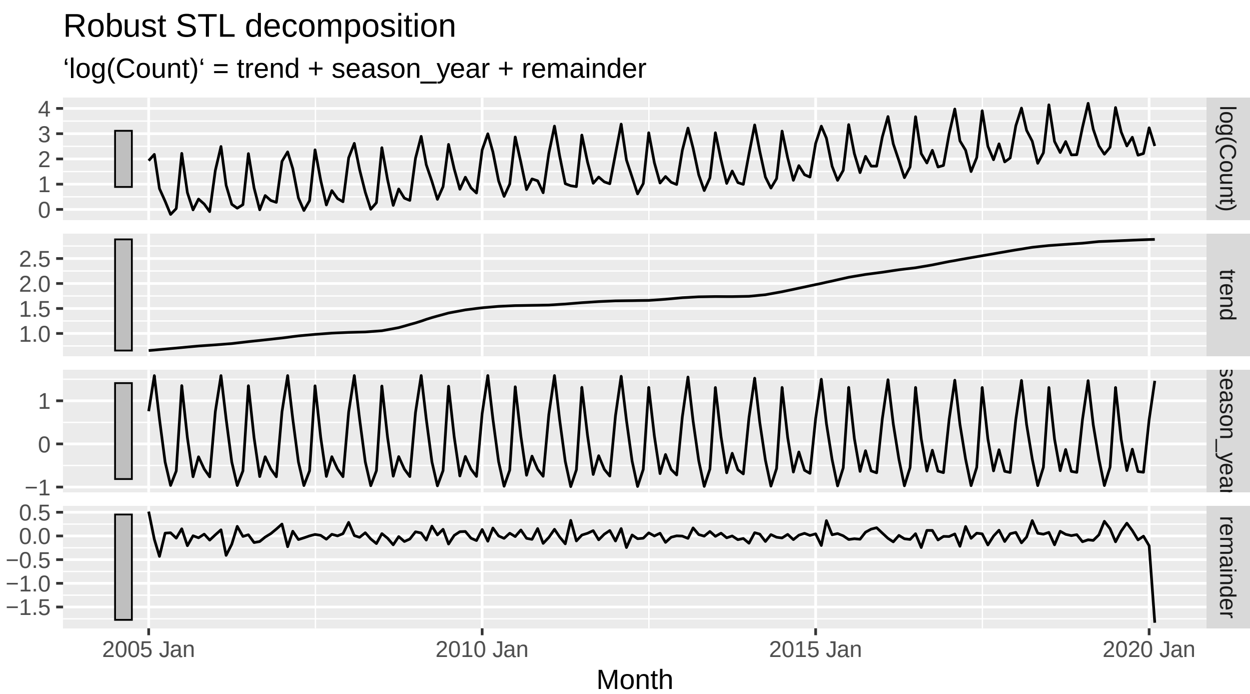

Figure 5:

3. You have been asked to provide forecasts for the next three years for the Chinese student arrivals series assuming that the travel bans will be lifted soon and travel will commence as normal in July 2020.

   Consider applying each of the methods and models below. Comment, in a few words each, on whether each one is appropriate for forecasting the data. No marks will be given for simply guessing whether a method or a model is appropriate without justifying your choice.

   Start your response by stating: **suitable** or **not suitable**.

   (a) Seasonal naïve method.

   (b) An STL decomposition combined with the drift method to forecast the seasonally adjusted component.

   (c) An STL decomposition on the log transformed data combined with an ETS to forecast the seasonally adjusted component.

   (d) Holt-Winters method with damped trend and multiplicative seasonality.

   (e) ETS(A,A,A).

   (f) ETS(M,A,M).

   (g) ARIMA(1,12,4).

   (h) ARIMA(3,2,1)$(1,1,0)_{12}$ on the log transformed data.

   (i) ARIMA(3,1,1)$(2,1,0)_{12}$ on the log transformed data.

   (j) Regression model with time and Fourier terms.

10 marks

Total: 20 marks

— END OF SECTION B —

SECTION C

The following R code and output concern models for the monthly student arrivals from China plotted in Figure 1.

```
fit <- ch_edu_arrivals %>%
  model(
    ets = ETS(Count),
    dcmp = decomposition_model(STL(log(Count)), ETS(season_adjust))
  )
```

```
fit %>% select(ets) %>% report()
```

```
## Series: Count
## Model: ETS(M,A,M)
##   Smoothing parameters:
##     alpha = 0.19288
##     beta  = 0.0066471
##     gamma = 0.292
##
##   Initial states:
##    l[0]     b[0]     s[0]    s[-1] s[-2]    s[-3]  s[-4]  s[-5]    s[-6]    s[-7]
##  3.4222 0.046599 0.31423 0.35995 0.467 0.27169 0.7335 2.5698 0.34985 0.28221
##    s[-8]  s[-9] s[-10] s[-11]
##  0.43686 1.2184 3.4074 1.5891
##
##   sigma^2:  0.0329
##
##     AIC     AICc      BIC
##  981.74  985.47 1036.21
```

```
fit %>% select(ets) %>% components() %>%
  autoplot() + labs(subtitle = "Components")
```

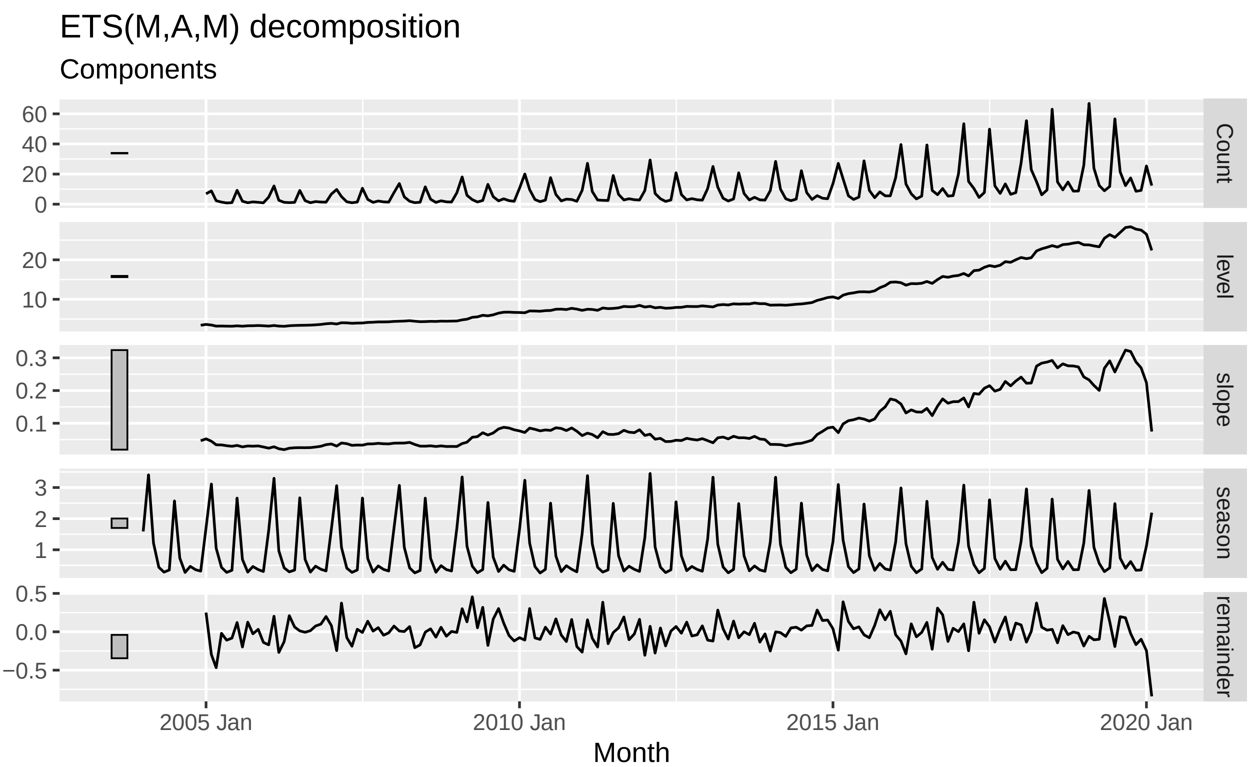

Figure 6:

```
fit %>% select(ets) %>% gg_tsresiduals()
```

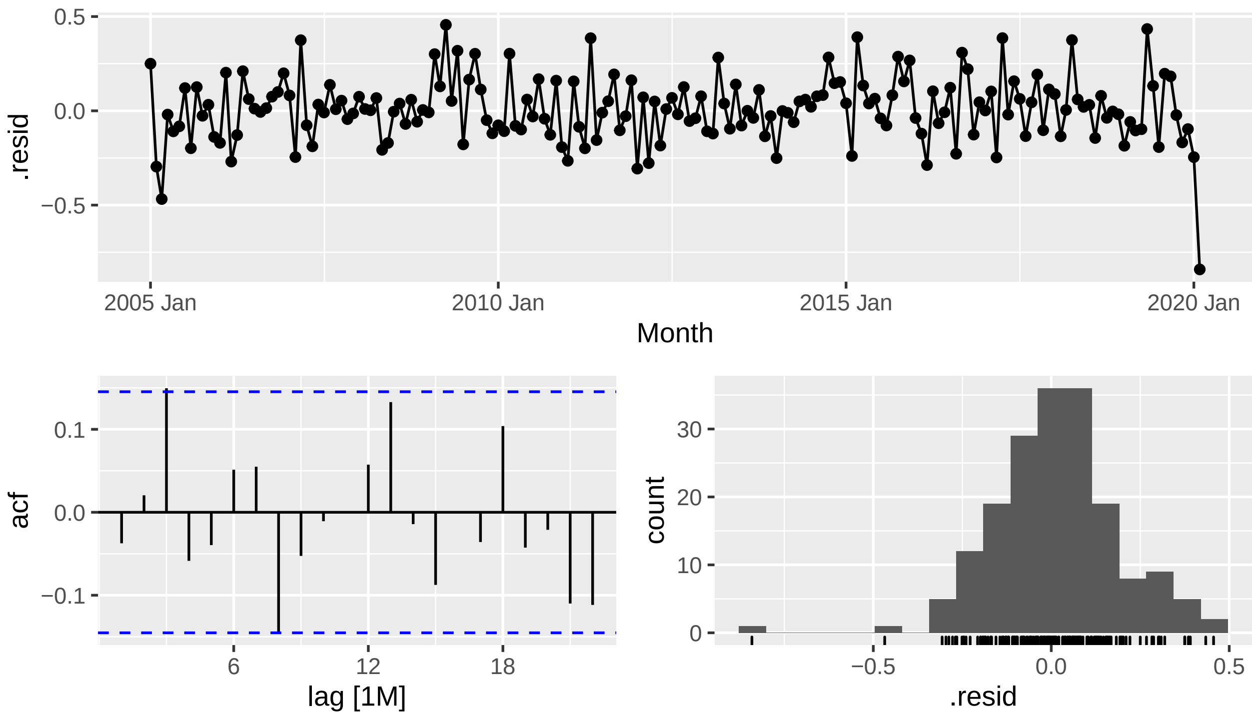

Figure 7:

1. Briefly describe the `fit` object.

   2 marks

2. Write down the estimated ETS model in full.
   (You can write out your answer elsewhere and upload it as an image if you prefer).

   4 marks

3. Comment on Figure 6 and how this relates to the estimated ETS model.

   5 marks

4. Comment on Figure 7.

   2 marks

5. Use the following output to produce forecasts and 95% prediction intervals for 1-step and 2-steps ahead. Give details for how the point forecasts can be obtained from the components. You must show your full workings. (You can write out your answer elsewhere and upload it as an image if you prefer).

   3 marks

```
fit %>% select(ets) %>% components() %>% tail(14)
```

```
## # A dable: 14 x 7 [1M]
## # Key:      .model [1]
## # :         Count = (lag(level, 1) + lag(slope, 1)) * lag(season, 12) * (1 +
## #   remainder)
##    .model     Month Count level  slope season remainder
##    <chr>      <mth> <dbl> <dbl>  <dbl>  <dbl>     <dbl>
##  1 ets     2019 Jan  25.7  23.8 0.242   1.21    -0.185
##  2 ets     2019 Feb  66.9  23.8 0.232   2.91    -0.0587
##  3 ets     2019 Mar  24.0  23.5 0.216   1.08    -0.104
##  4 ets     2019 Apr  12.4  23.3 0.200   0.563   -0.0975
```

```
##  5 ets    2019 May  8.91  25.5 0.268   0.298    0.434
##  6 ets    2019 Jun 11.7   26.4 0.291   0.418    0.132
##  7 ets    2019 Jul 56.7   25.7 0.257   2.48    -0.192
##  8 ets    2019 Aug 21.5   26.9 0.291   0.732    0.197
##  9 ets    2019 Sep 12.4   28.2 0.324   0.404    0.183
## 10 ets    2019 Oct 17.4   28.4 0.319   0.622   -0.0225
## 11 ets    2019 Nov  8.57  27.8 0.287   0.341   -0.168
## 12 ets    2019 Dec  9.16  27.5 0.269   0.351   -0.0967
## 13 ets    2020 Jan 25.4   26.5 0.224   1.12    -0.245
## 14 ets    2020 Feb 12.3   22.4 0.0747  2.19    -0.841
```

```
fit %>% select(ets) %>% forecast(h = 2)
```

```
## # A fable: 2 x 4 [1M]
## # Key:     .model [1]
##   .model    Month      Count .mean
##   <chr>     <mth>     <dist> <dbl>
## 1 ets     2020 Mar  N(24, 19)  24.3
## 2 ets     2020 Apr N(13, 5.5)  12.7
```

6. Describe the alternative model estimated via the `decomposition_model()` function and comment on the forecasts plotted in Figure 8.

4 marks

```
fit %>%
  forecast(h = "3 years") %>%
  autoplot(ch_edu_arrivals, level = NULL) + ylab("Thousands of students")
```

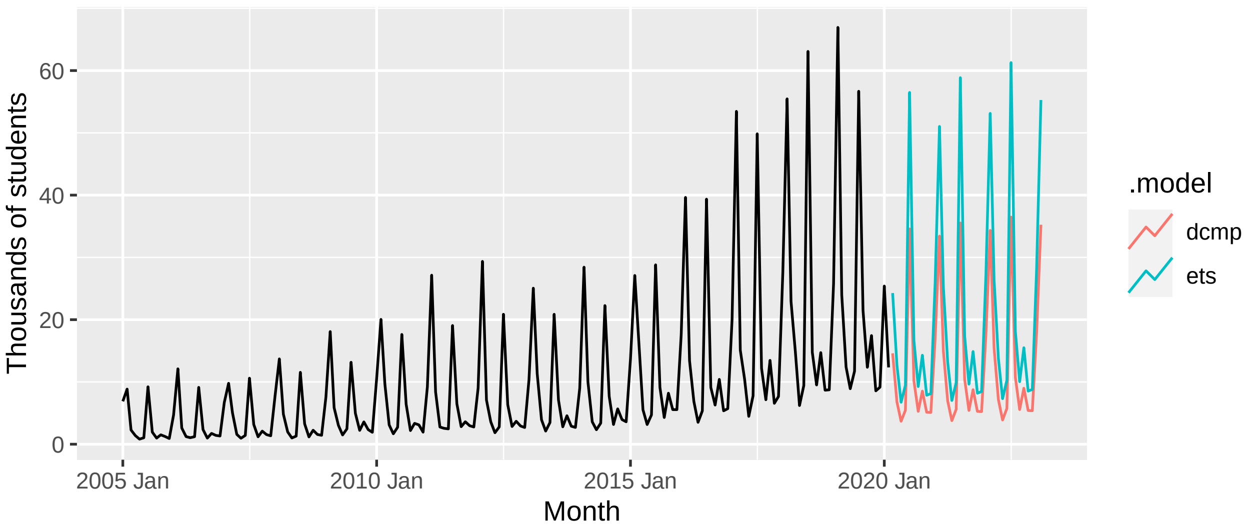

Figure 8:

Total: 20 marks

— END OF SECTION C —

SECTION D

Melbourne has a large number of sensors around the city which capture the number of pedestrians passing by. If we average the daily total pedestrians captured by each sensor, we get a measure of the pedestrian activity in the city each day.

Figures 9 and 10 relate to the daily average number of pedestrians across the City of Melbourne from 2019-01-01 until 2020-05-24.

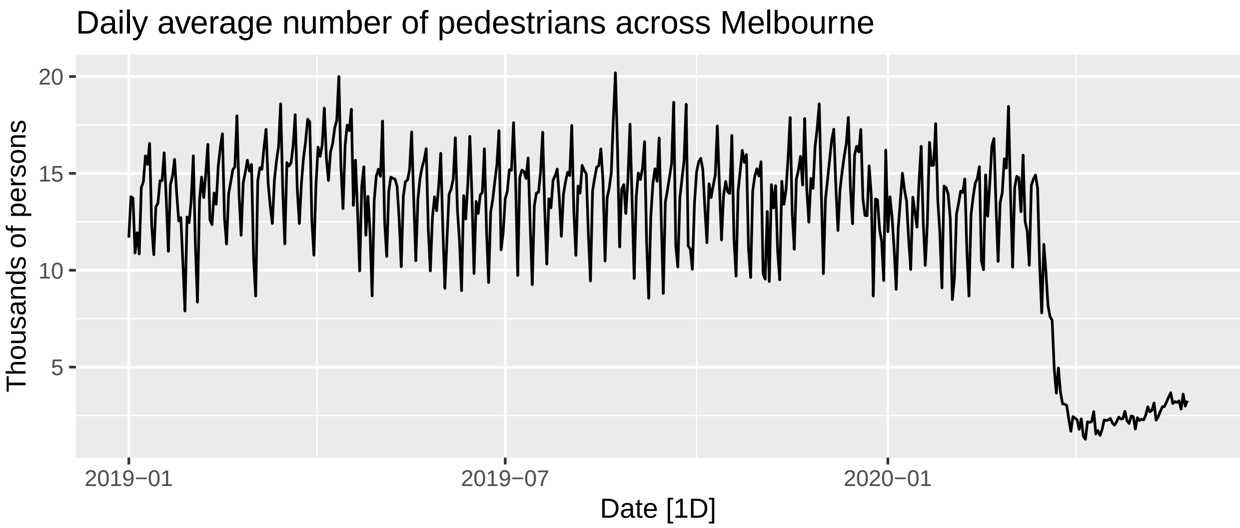

Figure 9:

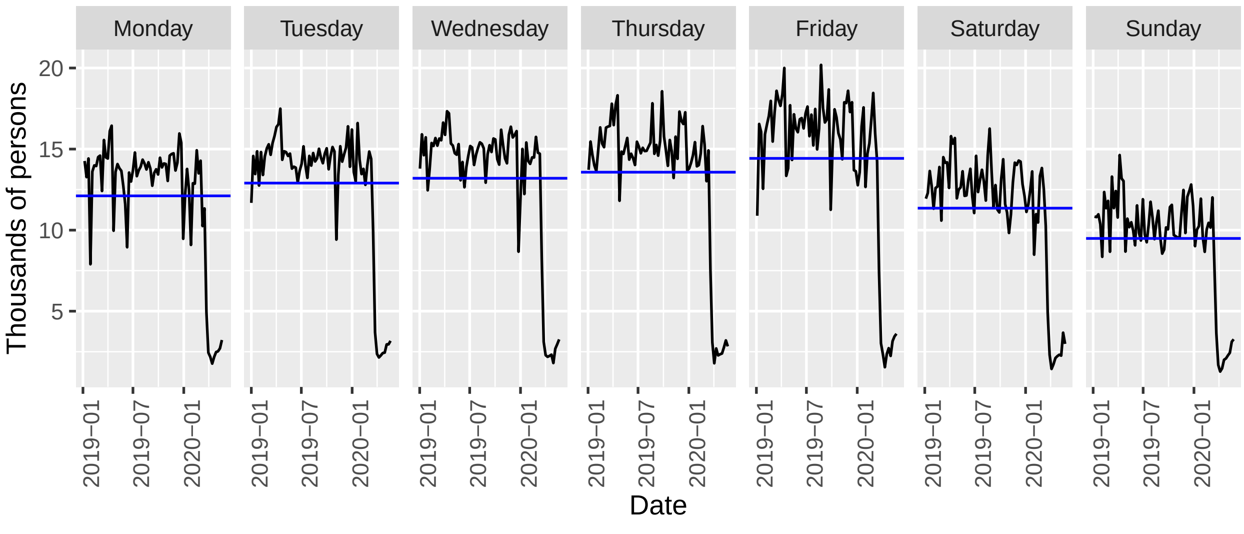

Figure 10:

1. Describe the important features of the time series.

   4 marks

2. Study Figure 11 and specify a plausible ARIMA model. Justify your choices.

   4 marks

```
average_daily_walkers %>% gg_tsdisplay(difference(Count, 7), plot_type = "partial")
```

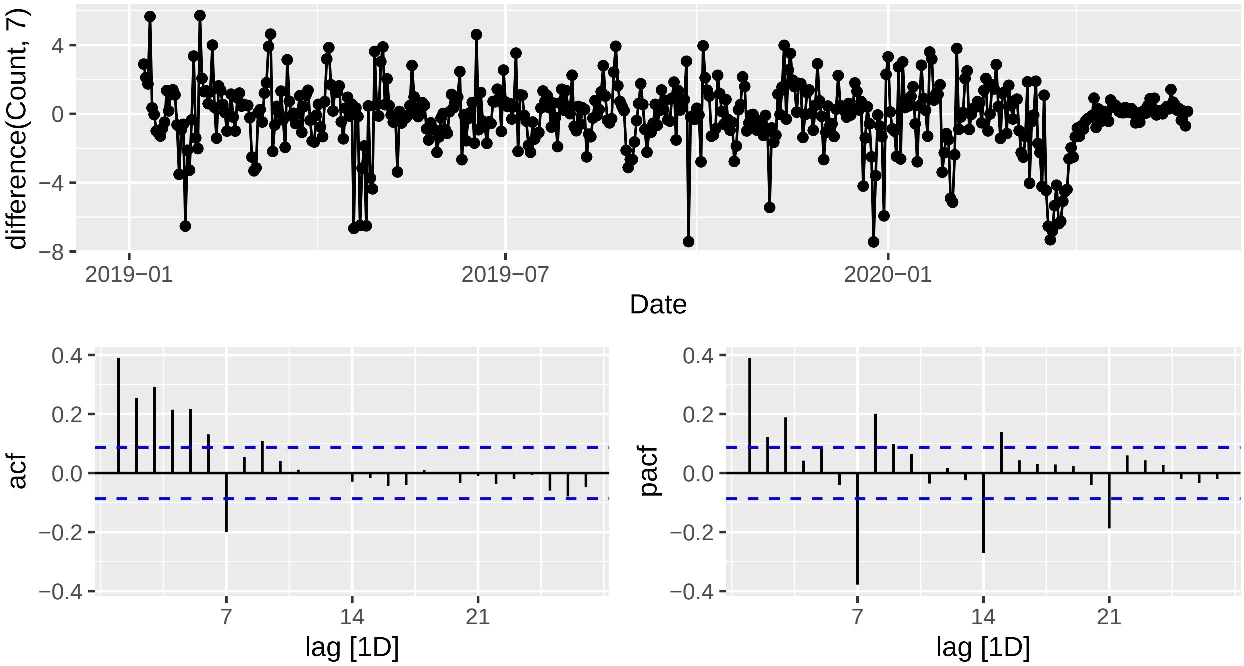

Figure 11:

3. Open the R file `Exam2020_for_students.R` provided to you in Moodle and run the first few lines to read in the Melbourne pedestrian data and create the `average_daily_walkers` tsibble object. Estimate the ARIMA model you have specified above. Check whether you are satisfied with the fitted model by performing some diagnostic checks of the residuals (clearly state any relevant parameters of any tests you may choose to conduct). Paste any relevant R output in the Moodle exam. (Further hints are included in the R file).

4 marks

4. A second ARIMA model is estimated using the following code. Briefly comment on the residuals and the forecasts generated from this model. Which of the two models do you prefer? Explain.

3 marks

```
fit_ARIMA_auto <- average_daily_walkers %>%
  model(ARIMA(Count, order_constraint = p + q + P + Q <= 3, approximation = FALSE))
fit_ARIMA_auto %>% report()
```

```
## Series: Count
## Model: ARIMA(1,0,1)(0,1,1)[7]
##
## Coefficients:
##          ar1      ma1     sma1
##       0.9769  -0.6132  -0.8459
## s.e.  0.0113   0.0385   0.0305
##
## sigma^2 estimated as 1.974:  log likelihood=-887.33
## AIC=1782.7   AICc=1782.7   BIC=1799.5
```

```r
fit_ARIMA_auto %>% gg_tsresiduals()
```

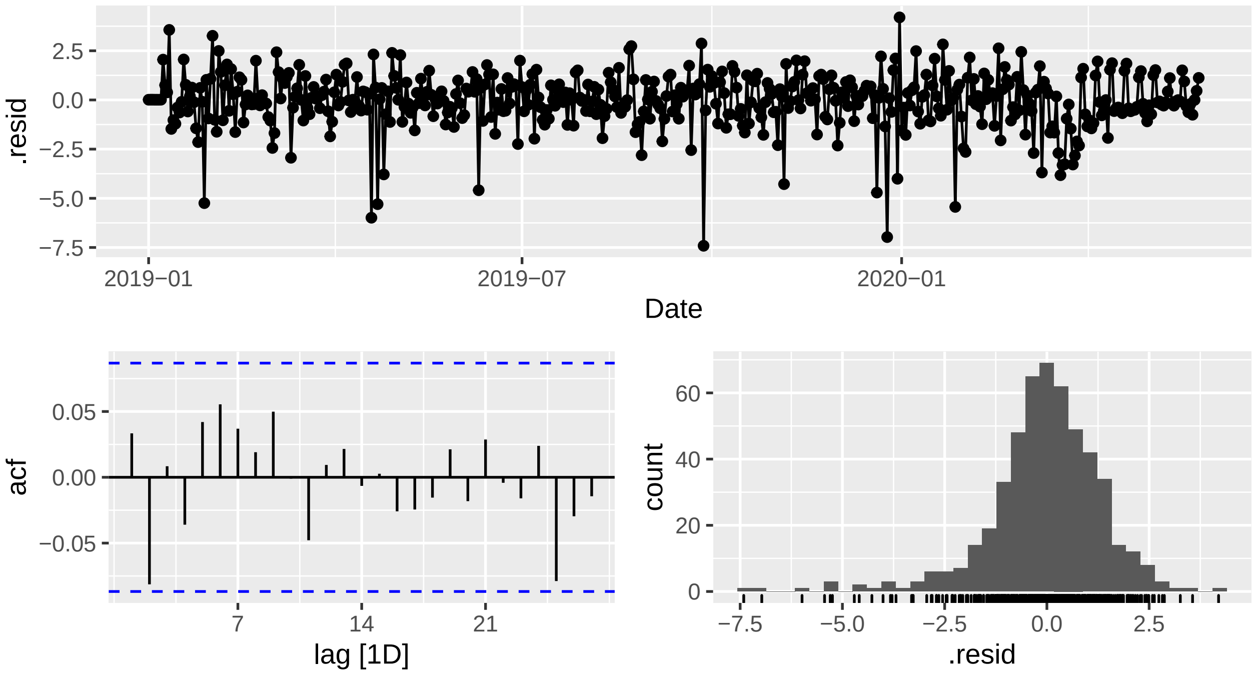

Figure 12:

```r
fit_ARIMA_auto %>%
  forecast(h = "28 days") %>%
  autoplot(average_daily_walkers)
```

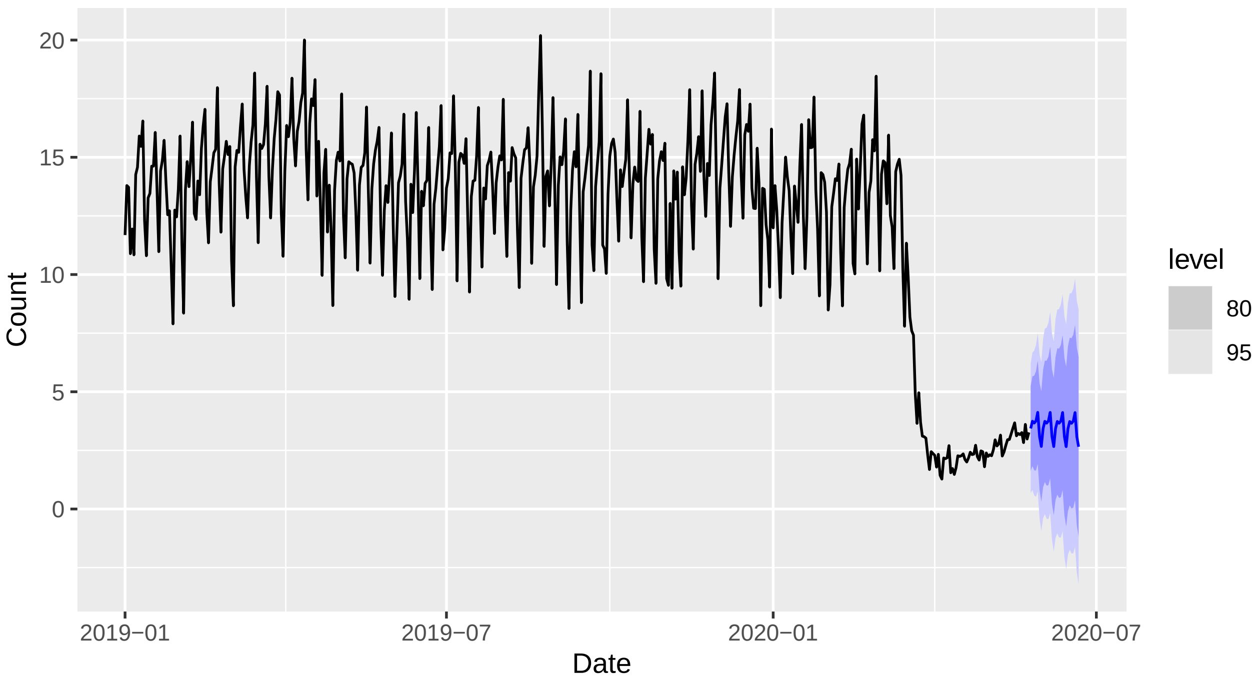

Figure 13:

5. Write out the estimated model from Q4 in full, first using backshift notation, then using subscript notation. Use the information below to generate forecasts for 1 and 2-steps ahead. Also generate a prediction interval for 1-step ahead. You must show all workings. (You can write out your answer elsewhere and upload it as an image if you prefer).

5 marks

```
## # A tsibble: 14 x 3 [1D]
##    Date       Count    .resid
##    <date>     <dbl>     <dbl>
##  1 2020-05-11  2.71 -0.190
##  2 2020-05-12  2.96 -0.289
##  3 2020-05-13  2.96 -0.0658
##  4 2020-05-14  3.20 -0.00536
##  5 2020-05-15  3.45 -0.0721
##  6 2020-05-16  3.68  1.51
##  7 2020-05-17  3.12  0.943
##  8 2020-05-18  3.22 -0.317
##  9 2020-05-19  3.17 -0.625
## 10 2020-05-20  3.26 -0.201
## 11 2020-05-21  2.83 -0.752
## 12 2020-05-22  3.61  0.000744
## 13 2020-05-23  2.98  0.464
## 14 2020-05-24  3.27  1.12
```

Total: 20 marks

— END OF SECTION D —

SECTION E

You will now analyse the data from one of the sensors in Melbourne. Use the code provided in the file `Exam2020_for_students.R` (on Moodle) to create your series. (Each student will have a different series.)

1. Fit a regression model to the log data, with Fourier terms for the weekly seasonality and a piecewise trend having breaks at 11 March 2020 and at 1 April 2020. The following code can be used to fit this model.

   ```
   fit <- myseries %>%
     model(
       TSLM(log(Count) ~ fourier(K = 3) +
              trend(knots = as.Date(c("2020-03-11", "2020-04-01")))
       )
     )
   ```

   Write out the fitted model for the daily counts for your sensor. Interpret the trend coefficients. (You can write out your answer elsewhere and upload it as an image if you prefer)

   5 marks

2. Suppose you used a training set to the end of February 2020, and a test set consisting of the remaining data, and compared the accuracy of several possible models on the test set. Would the results help you to choose an appropriate forecasting model? Explain.

   2 marks

3. Why is it impossible to use time series cross-validation with this model?

   2 marks

4. How could you decide where to place the knots with this model?

   2 marks

5. Plot the residuals from the model using the following code

   ```
   fit %>% gg_tsresiduals()
   ```

   Upload the plot in moodle. Explain how you might improve the model based on this plot.

   3 marks

6. What makes the model unrealistic for forecasting more than a few weeks ahead?

   2 marks

7. Suppose the government decided to remove all COVID19 restrictions from 5 July, how would you allow for this in your forecasts?

   4 marks

ETC5550 students only

- Please submit an Rmarkdown file containing the analysis described in Q1 and Q5 of Section E, including at least one model that is better than the one described here. [5 bonus marks]

Total: 20 marks

— END OF SECTION E —