Forecasting Exam 2020: solutions

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.)*

Deduct marks for each major thing missed, and for each wrong statement. In general, be relatively generous if the answer makes sense and contains the main ideas.

1. The disadvantage of using a test set for choosing a forecasting model is that it uses only a small proportion of the data.

- This is true. $\boxed{1}$
- Alternative is to use information criteria - use all data $\boxed{1}$
- These are not useful/helful in all situations, e.g., for comparing across model classes or ARIMA with different orders, etc. $\boxed{1}$
- Another alternative is to use cross-validation. $\boxed{1}$
- If used across many series and across various forecast horizons it can be slow. In this case may be test sets across the many series will be useful. $\boxed{1}$

2. *The best forecasting models adapt rapidly to changes in the trend and seasonal patterns.*

- This is not necessarily true. $\boxed{1}$
- Good models need to be adaptive, but it is possible to overfit if the model adapts too quickly to changes in the data. $\boxed{2}$
- Models that don't adapt to changes will give biased forecasts, while models that adapt too quickly to changes will have inflated variances. $\boxed{2}$

3. *With STL decompositions and ETS models, we always need to transform our data before estimating the components.*

- This is not true. $\boxed{1}$
- STL is an additive decomposition and therefore we need to take a transformation before we apply the decomposition if the data has monotonically changing variance. $\boxed{2}$
- ETS models can have multiplicative components and therefore we do not need to take a transformation as these will take care of this. $\boxed{2}$

4. *The mean of a stationary AR(3) process*

- This is not true. $\boxed{1}$
- The mean in related to the constant but it is not equal to this $\boxed{2}$
- The actual mean can be computed like this. $\boxed{2}$

$$
\begin{aligned}
y_t =& c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t \\
E(y_t) =& c + \phi_1 E(y_{t-1}) + \phi_2 E(y_{t-2}) + \phi_3 E(y_{t-3}) + E(\varepsilon_t) \\
\mu =& c + \phi_1 \mu + \phi_2 \mu + \phi_3 \mu \\
\mu =& \frac{c}{1 - \phi_1 - \phi_2 - \phi_3}
\end{aligned}
$$

5. *ARIMA models are better than ETS models because there are more possible models available.*

- This is not true. $\boxed{1}$
- There are more ARIMA models available, but that does not make them a better class of models. $\boxed{1}$
- ETS models allow for multiplicative behaviour, while ARIMA models do not. $\boxed{2}$
- ARIMA models handle some types of data (e.g., cyclic, stationary) which ETS models cannot handle. $\boxed{1}$

6. *Regression models are not very useful for forecasting because you have to forecast all the predictors as well.*

- This is not true. $\boxed{1}$
- Sometimes the predictors are known such as trend, seasonality, public holidays, other calendar effects, lagged predictors, etc. $\boxed{2}$
- When the predictors are unknown in the future, you do have to forecast them, or create scenario forecasts. $\boxed{2}$

[Total: 20 marks]

— END OF SECTION A —

SECTION B

1. Marks are allocated for the four main features, trend, seasonality, two-peaks, COVID-19.

   Time plot:

   - the series is trending. $\boxed{1}$
   - strong seasonality with seasonal variation increasing with the level of the series, i.e., multiplicative seasonality. $\boxed{1}$
   - last two peaks seem lower than previous years.

   Seasonal plot:

   - There are two peaks, one in February, due to the beginning of year, and one in July, mid-year break. $\boxed{1}$
   - The drop in February 2020 shows the effect of COVID-19 and the travel bans. $\boxed{1}$

   Subseries plot:

   - Effect of COVID-19 and travel bans clearly shown in the February 2020.
   - The average and growth of the peak season arrivals of February and July are almost identical. Flatter growth shown across May-Jun and Nov-Dec.

2. - The panels show STL decompositions, where the lower three panels show trend-cycle, seasonal and remainder components. $\boxed{1}$
   - As STL is an additive decomposition a log transformation is first applied to account for the multiplicative seasonality. $\boxed{1}$
   - The default setting for `trend(window)` and `season(window)` are 21 and 13 respectively. These seem to be appropriate here with both the trend and the seasonal components smoothly changing over time. $\boxed{1}$
   - The reminder in both decompositions shows a large value for Feb 2020 due to the travel ban restrictions related to COVID-19. $\boxed{1}$
   - Setting `robust=TRUE` removes its effect from the trend and makes the remainder value even larger. $\boxed{1}$
   - For forecasting, the `robust` setting doesn't make much difference as it hardly affects seasonality, and the seasonally adjusted values will therefore be almost identical either way. For analysing the trend, we know that the downturn is likely to continue, so it should show up in the trend component. In that case, it is probably better not to use `robust=TRUE`, and instead to choose a smaller trend window. *(Any sensible remarks along these lines are ok.)* $\boxed{1}$

3. (a) *Seasonal naive.*

   Not suitable. Data needs to be transformed and is also trending. $\boxed{1}$

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

   Not suitable. Data needs transformation before STL is implemented. Also drift will not capture the change in the trend due to COVID-19. $\boxed{1}$

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

Suitable. Log transformation will take care of the multiplicative seasonality and ETS trend should update enough to deal with change in trend. $\boxed{1}$

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

Suitable, both multiplicative seasonality and damped trend would be appropriate. $\boxed{1}$

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

Not Suitable. Some multiplicative component would be required. $\boxed{1}$

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

Suitable. Accounting for multiplicative components. $\boxed{1}$

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

Not suitable. Not a feasible model. $\boxed{1}$

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

Not suitable. Too many differences. $\boxed{1}$

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

Suitable. Order of differencing seems fine. Of course you will need to check the residuals. $\boxed{1}$

(j) *Regression with time and Fourier terms.*

Not suitable. Seasonality is changing so need transformation. Also need something like a step variable, but probably hard to estimate its coefficient successfully with only limited data. $\boxed{1}$

[Total: 20 marks]

— END OF SECTION B —

SECTION C

1. fit is a mable (model table) containing two models for the Chinese education series: an ETS model and a model combining an STL decomposition and ETS. $\boxed{2}$

2. The full estimated model is:

$$\begin{aligned} y_t &= (\ell_{t-1} + b_{t-1})s_{t-m}(1+\varepsilon_t) \\ \ell_t &= (\ell_{t-1} + b_{t-1})(1 + 0.1929\varepsilon_t) \\ b_t &= b_{t-1} + 0.0066(\ell_{t-1} + b_{t-1})\varepsilon_t \\ s_t &= s_{t-m}(1 + 0.292\varepsilon_t) \end{aligned}$$

$\varepsilon_t \sim N(0, 0.0329)$. $\boxed{4}$

3. Figure 6 shows the data in the first panel and below that the estimated components of the ETS model.

   - The level is adjusting over time, with a smoothing parameter of $\alpha = 0.2$, especially showing an increase after 2015. $\boxed{1}$
   - Although the slope coefficient is relatively low $\beta = 0.0066$ the slope is adjusting, especially showing an increase after 2015 (not much change before that). $\boxed{1}$
   - It is clearly visible that the seasonal component is rapidly changing/adjusting with a coefficient of $\gamma = 0.292$. $\boxed{1}$
   - The level and trend show a dip at the end showing the model adjusting and trying to accounting for the travel bans and COVID-19. The last peak of the seasonal component is also lower than previous peaks. The remainder shows a large residual due to the travel bans and COVID-19. $\boxed{2}$

4. Residuals look like WN, with no significant autocorrelation left over. They are close to normally distributed. The effect of the travel bans and COVID-19 is visible by the large residual for Feb 2020. $\boxed{2}$

5. Point forecasts:

$$\begin{aligned} \hat{y}_{Mar,2020} &= (22.4 + 0.0747) * 1.08 = 24.273 \\ \hat{y}_{Apr,2020} &= (22.4 + 2 * 0.0747) * 0.563 = 12.695 \end{aligned}$$ $\boxed{2}$

   Forecast intervals:

$$\begin{aligned} \text{Mar 2020} &: 24.273 \pm 1.96 * \sqrt{19.4} = (15.64, 32.906) \\ \text{Apr 2020} &: 12.695 \pm 1.96 * \sqrt{5.5} = (8.098, 17.292) \end{aligned}$$ $\boxed{1}$

6. The alternative model dcmp fits an STL decomposition on the log() transformed data, then the seasonal component is projected using a seasonal naïve and the seasonally adjusted series using an ETS model. It seems that the dcmp adjusts more severely to the break in level and seasonal component compared to the ETS model. $\boxed{4}$

[Total: 20 marks]

— END OF SECTION C —

SECTION D

1. **Time plot:**

    - Strong seasonal component $\boxed{1}$
    - significant break in the level towards the end due to COVID-19 and the social distancing measures $\boxed{1}$
    - significant break in the variance together with the level towards the end due to COVID-19 and the social distancing measures $\boxed{1}$

    Subseries plot

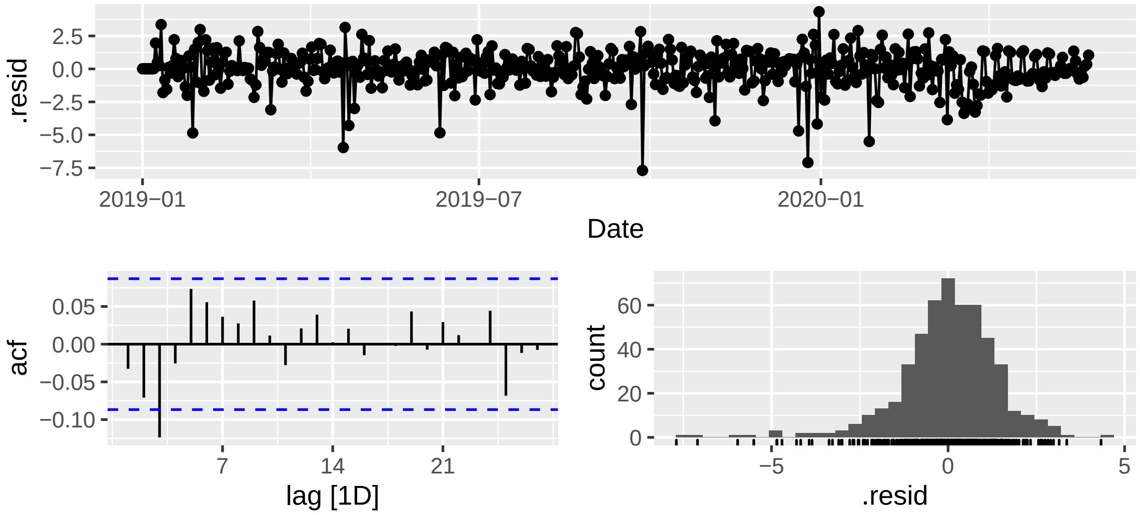
    - Maximum average pedestrian traffic happens on a Friday. Weekends have lower pedestrian traffic than weekdays with Sunday being the lowest. $\boxed{1}$

2. - Seasonal component. ACF shows 1 spike, PACF exponential decay. Hence, I choose PDQ(0,1,1) $\boxed{2}$
    - Non-seasonal. PACF shows 3 spikes (maybe 5), ACF decaying. Hence I choose pdq(3,0,0) $\boxed{2}$

3. We expect something like the following output.

```
## Series: Count
## Model: ARIMA(3,0,0)(0,1,1)[7] w/ drift
##
## Coefficients:
##          ar1     ar2     ar3     sma1  constant
##       0.4543  0.1766  0.2633  -0.8241   -0.0140
## s.e.  0.0439  0.0470  0.0431   0.0344    0.0115
##
## sigma^2 estimated as 2.028:  log likelihood=-893.01
## AIC=1798   AICc=1798.2   BIC=1823.3
```

Correct model and pasted estimation output. $\boxed{1}$



- There are a bunch of residuals after the COVID-19 date that show some significant pattern. These may indicate that the model has not best dealt with this.
- ACF has one significant spike at lag 4 (its value fairly low). However I would change my model to capture this. Possibly introduce an MA component to the model. $\boxed{1}$
- Histogram shows a longer left tail due to COVID-19 but otherwise fairly satisfied with assuming normality $\boxed{1}$

```
## # A tibble: 1 x 3
##   .model                                     lb_stat lb_pvalue
##   <chr>                                        <dbl>     <dbl>
## 1 ARIMA(Count ~ pdq(3, 0, 0) + PDQ(0, 1, 1))    20.1    0.0176
```

Ljung-Box for lag 14 with dof=5 rejects the null of WN at 5% level of significance. $\boxed{1}$

4.
- The residuals look more like white noise, although there is still pattern after the COVID-19 break.
- The point forecasts are flat and too seasonal, having not properly accounted for the COVID-19 break.
- The width of the prediction intervals are too large and have negative lower bounds, which cannot be the case for the number of pedestrians on Melbourne streets. Taking logs would avoid this. $\boxed{3}$

5. ARIMA(1,0,1)(0,1,1)[7]:

$$\begin{aligned}
&(1-\phi_1 B)(1-B^7)y_t = (1+\theta_1 B)(1+\Theta_1 B^7)\varepsilon_t \\
&(1-\phi_1 B - B^7 + \phi_1 B^8)y_t = (1+\theta_1 B + \Theta_1 B^7 + \theta_1\Theta_1 B^8)\varepsilon_t \\
y_t &= \phi_1 y_{t-1} + y_{t-7} - \phi_1 y_{t-8} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \Theta_1\varepsilon_{t-7} + \theta_1\Theta_1\varepsilon_{t-8} \\
&= 0.9769 y_{t-1} + y_{t-7} - 0.9769 y_{t-8} + \varepsilon_t - 0.6132\varepsilon_{t-1} - 0.8459\varepsilon_{t-7} + 0.5187\varepsilon_{t-8}
\end{aligned}$$

$\boxed{2}$

$$\begin{aligned}
\hat{y}_{T+1|T} &= 0.9769 y_T + y_{T-6} - 0.9769 y_{T-7} - 0.6132 e_T - 0.8459 e_{T-6} + 0.5187 e_{T-7} \\
&= 0.9769 * 3.27 + 3.22 - 0.9769 * 3.12 - 0.6132 * 1.12 - 0.8459 * (-0.32) + 0.5187 * 0.94 \\
&= 3.43 \\
\hat{y}_{T+2|T} &= 0.9769 \hat{y}_{T+1|T} + y_{T-5} - 0.9769 y_{T-6} - 0.8459 e_{T-5} + 0.5187 e_{T-6} \\
&= 0.9769 * 3.43 + 3.17 - 0.9769 * 3.22 - 0.8459 * (-0.62) + 0.5187 * (-0.32) \\
&= 3.74
\end{aligned}$$

$\boxed{2}$

Prediction interval for 1-step ahead $\boxed{1}$

$$3.43 \pm 1.96\sqrt{1.97} = [0.68, 6.19]$$

[Total: 20 marks]

— END OF SECTION D —

SECTION E

1.
$$y_t = \beta_0 + \beta_1 t + \beta_2 (t - \tau_1)_+ + \beta_3 (t - \tau_2)_+ + \sum_{k=1}^{3} \left[\alpha_k \sin\left(\frac{2\pi k t}{52.18}\right) \gamma_k \cos\left(\frac{2\pi k t}{52.18}\right) \right] + \varepsilon_t,$$

where τ_1 corresponds to 11 March 2020 and τ_2 corresponds to 1 April 2020. $\boxed{2}$

- β_1 is the trend prior to 11 March. $\boxed{1}$
- $\beta_1 + \beta_2$ is the trend between 11 March and 1 April 2020. $\boxed{1}$
- $\beta_1 + \beta_2 + \beta_3$ is the trend after 1 April 2020. $\boxed{1}$

2. No $\boxed{1}$

The structural breaks in March and April means that any model which fits well to the end of February may perform badly in March and April. $\boxed{1}$

3.
- The performance of models before the structural breaks is not really important. $\boxed{1}$
- It is impossible to fit the model before the structural break as the β_2 and β_3 coefficients are not estimable. $\boxed{1}$

4.
- We could try moving the knots or adding new knots and minimizing the AICc statistic. $\boxed{2}$

5.
- The plot will show some autocorrelations in the ACF, and possibly an outlier in the histogram. $\boxed{1}$
- We could add an ARIMA error term to the model to handle the autocorrelations $\boxed{1}$
- We could add additional knots where the residuals appear to show nonlinear trend. $\boxed{1}$

6.
- The structural breaks are due to policy changes associated with COVID19. Additional policy changes will probably change the trend again. $\boxed{2}$

7.
- The simplistic approach would be to add another knot at 5 July, and assume all the trend changes would cease to have any effect after that date. This could be achieved by setting the piecewise linear predictors associated with β_2 and β_3 to be zero after 5 July. So then the long-term trend from before 11 March would be the only trend term continuing to have an effect after 5 July. $\boxed{2}$
- However, it is impossible to know how people would respond, but it is unlikely things would return to normal. Therefore we would probably need some kind of judgmental adjustments to the forecasts to allow for scenarios around human responses to the policy changes. $\boxed{2}$