Forecasting Exam 2024: solutions

The exam contains FIVE questions. ALL questions 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.*

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 AIC is better than the MSE for selecting a forecasting model.*
    - The MSE on the training set only measures goodness of fit, not forecast accuracy. 1
    - The AIC is also a measure of goodness of fit on the training set, but includes a penalty for the number of parameters in the model. 1
    - The MSE on a test set is a measure of forecast accuracy, but the test set may be too small to be reliable. The MSE on a cross-validated set is better but is computationally expensive. 1
    - The AIC is asymptotically equivalent to one-step time series cross-validation. 1
    - The AIC can only be compared within the same model class, whereas MSE on a test set, or on a cross-validated set, can be used to compare different model classes. 1
2. *The Ljung-Box test is useful for selecting a good forecasting model.*
    - The Ljung-Box test is used to test whether the residuals from a model are white noise. 1
    - It should not be used to select a forecasting model. 1
    - If the residuals are not white noise, then the model is not capturing all the information in the data, and can potentially be improved. 1
    - A model that has been over-fitted may pass a Ljung-Box test, but be a poor forecasting model. 1
    - It may not be possible to find a model that passes the Ljung-Box test, especially with a long time series. But the forecasts from a model that fails the Ljung-Box test may still be good. 1
3. *The MAPE is better than the RMSE for measuring forecast accuracy because it is easier to explain.*
    - The MAPE is a percentage, so it is easier to interpret than the RMSE. 1
    - However, the MAPE may not make sense for the data, especially if the data contains zeros, small values, or negative values, or if the zero point is arbitrary. 1
    - The RMSE is scale-dependent so can't be used to compare forecasts across different data sets, whereas the MAPE is scale-independent. 1
    - The RMSE is optimized by the mean, so finding a model that minimizes the RMSE will give the best point forecasts. 1
    - The MAPE is not optimal for any meaningful characteristic of the forecasts. 1

4. *Dynamic regression models are only useful for forecasting if the predictors are deterministic or easy to forecast, and there is no serial correlation in the residuals.*

    - If the predictors in a regression are deterministic or easy to forecast, then the dynamic regression model can be used to forecast the response variable. 1
    - The dynamic regression model is designed to handle serial correlation in the residuals. 1
    - However, it may not be possible to find an ARIMA error model that captures all the serial correlation in the residuals. 1
    - If the predictors are not easy to forecast, the dynamic regression model can still be used for scenario forecasting, where the predictors are set to some plausible values. 1
    - Ex post forecasting can also be used for testing purposes, when known future values of all predictors are used. 1

5. *ETS and ARIMA models are largely interchangeable, so it does not matter which one we use.*

    - These model classes are not interchangeable. 1
    - Some of the simpler additive models are equivalent including SES and Holt's method with additive errors. 1
    - ETS models with multiplicative errors handle heteroskedasticity, whereas ARIMA models do not. 1
    - ARIMA models handle a much wider range of dynamics than ETS models. 1
    - ETS models cannot be used for stationary series. 1

6. *STL is not useful for forecasting because it only provides a decomposition of the data.*

    - STL is a decomposition method that separates a time series into trend, seasonal, and remainder components. 1
    - STL, on its own, is not a forecasting method. 1
    - However, STL is useful for forecasting because the decomposition can help us understand the data better. 1
    - STL can be coupled with a forecasting method applied to the seasonally adjusted data. 1
    - In this case, the seasonal component is usually forecast using the seasonal naive method, and then the seasonal component forecasts and seasonally adjusted forecasts can be added to obtain forecasts of the original data. 1

Total: 20 marks

SECTION B

Figures 1–3 relate to the daily use of public transport in Canberra, from July 2019 – March 2024. The variable plotted is the total number of passenger boardings each day on all forms of public transport except for school buses.

1. *Using Figures 1–2, describe the daily passenger boardings for public transport in Canberra. Canberra had two major periods of "COVID-19 lockdowns" where there was substantially reduced travel, and has four school terms per year. Carefully comment on the interesting features of both plots, and how the lockdowns, school terms and other holidays are evident.*

   - There is strong weekly seasonality with lower boarding numbers on weekends, seen in Figure 1 but more clearly in Figure 2. 1
   - Some Mondays (Fig 2) have lower boarding numbers, probably associated with long weekends. 1
   - Fig 1 shows there is also some annual seasonality, with low boarding numbers at the end of the year and the beginning of the next year (summer holidays). 1
   - The school holidays (incl Easter) have a noticeable effect with lower boarding numbers during the holidays/terms breaks. 1
   - The two COVID-19 lockdowns are visible in the first half of 2020, and the second half of 2021. 1

2. *For the STL decomposition shown in Figure 3, discuss what is shown in each panel. Why has a log transformation been used? Describe how COVID-19 lockdowns and holidays have affected the trend, seasonal and remainder components.*

   - The top panel shows the transformed data, which is the log of the original data. 1
   - The next panels show the trend, weekly seasonality, annual seasonality and the remainder components. 1
   - The log transformation is used to stabilize the variance of the data. 1
   - The lockdowns are seen in the two large dips in the trend. 1
   - The holidays have mostly been put into the annual seasonal component and the remainder component. 1

3. *You have been asked to provide forecasts for the next four weeks for daily passenger boardings. Consider applying each of the methods and models below to the data from March 2022 onwards. Comment, in a few words each, on whether each one is appropriate for forecasting the next two weeks of data, and what features of the data each would miss. No marks will be given for simply guessing whether a method or a model is appropriate without justifying your choice.*

Be flexible in marking, especially around the challenge with capturing term breaks and holidays.

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

(a) *Seasonal naïve method using weekly seasonality.*

   - Maybe suitable, but will miss the annual seasonality and holidays. 1

(b) *Naïve method*

   - Unsuitable. Won't even capture weekly seasonality. 1

(c) *An STL decomposition on the log transformed data combined with an ARIMA to forecast the seasonally adjusted component, and seasonal naïve methods for both seasonal components.*

   - Might be suitable, but will miss holidays that do not have a fixed date. 1

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

   - Might be suitable, but will miss annual seasonality and holidays. 1

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

- Unsuitable. Seasonality is not additive. Also misses annual seasonality and holidays. 1

(f) *ETS(M,A,M) with annual seasonality.*

- Unsuitable. ETS won't handle annual seasonality, and model doesn't capture weekly seasonality. 1

(g) *ARIMA(2,4,2)(1,1,0)$_7$ applied to the log transformed data.*

- Unsuitable. Far too many differences. 1

(h) *ARIMA(1,0,1)(1,1,0)$_7$ applied to the log transformed data.*

- Might be suitable, but will miss annual seasonality and holidays. 1

(i) *Regression with time and Fourier terms for both weekly and annual seasonality.*

- Unsuitable. Misses serial correlation, heteroskedacity, and the holidays that do not have a fixed date. 1

(j) *Dynamic regression on the log transformed data with Fourier terms for the annual seasonality and a seasonal ARIMA model to handle the weekly seasonality and other dynamics.*

- Might be suitable, but will miss holidays that do not have a fixed date. 1

Total: 20 marks

SECTION C

1. *An ETS model is fitted to the time series, using only data from March 2022. Write down the equations for the model, including specifying the values of all model parameters.* 3

$$\begin{aligned} y_t &= \ell_{t-1}s_{t-m}(1+\varepsilon_t) \\ \ell_t &= \ell_{t-1}(1+\alpha\varepsilon_t) \\ s_t &= s_{t-m}(1+\gamma\varepsilon_t) \qquad \varepsilon_t \sim N(0,\sigma^2) \end{aligned}$$

   where $\alpha = 0.208$, $\gamma = 0.0001$, $\sigma^2 = 0.0338$. 2

2. *What features of the data has the model ignored?*
   - annual seasonality 1
   - holidays 1

3. *What does the value of γ tell you?*
   - seasonal pattern hardly changing over time. 1

4. *The data, remainder (i.e., residuals), and estimated states, are shown below for the last week of observations, along with the forecasts for the next day. Show how the forecast mean and variance have been obtained, and give a 95% prediction interval for this day.*
   - Forecast mean: $\hat{y}_{T+1|T} = \ell_T s_{T+1-m} = 48.1 \times 1.17 = 56$ 3
   - Forecast variance is 106. So a 95% prediction interval is 2

$$56 \pm 1.96 \times \sqrt{106} = (36, 76)$$

5. *Some plots of the residuals are shown in Figure 4. Discuss what these tell you about the model?*
   - Significant serial correlation remaining in residuals. 1
   - Lots of outliers, probably associated with holidays. 1
   - Residuals are not normally distributed. 1
   - The model has not captured the holiday effects, or the time series dynamics in the data. 1
   - The model will give prediction intervals with incorrect coverage. 1

6. *If you conducted a Ljung-Box test of the residuals using 14 lags, what do you think the p-value would be? Why?*
   - Close to zero 1
   - Due to significant serial correlation. 1

Total: 20 marks

SECTION D

1. *An ARIMA model is fitted to the time series, using only data from March 2022. Write down the equation for the model using backshift notation, including specifying the values for all model parameters.* 3

$$(1-\phi_1 B)(1-\Phi_1 B^7)(1-B^7)y_t = (1+\theta_1 B)\varepsilon_t$$

where $y_t = \log(\text{Boardings})$, $\phi_1 = 0.8165$, $\theta_1 = -0.5920$, $\Phi_1 = -0.4574$, $\varepsilon_t \sim N(0, 0.1045)$. 2

2. *This model suggests that the weekly differences of the logged data are stationary. What aspects of the model lead to that conclusion?*
   - The fitted model has $D = 1$ for weekly differences, but $d = 0$ so no ordinary differencing, and it is applied to the logged data. So if we apply these differences to the logged data, we should have stationary data. 3

3. *Let the observed data be given by y_t, and define $x_t = \log(y_t) - \log(y_{t-7})$. The x_t series is shown in Figure 5. What features of these plots suggest that x_t is stationary?*
   - The ACF and PACF both decay to zero quickly. 1
   - The time plot is consistent with a zero mean and constant variance. 1

4. *The residuals shown in Figure 6 show a lot of large outliers. What might be causing these outliers?*
   - These are related to holidays, or the start and end of holiday periods, as these have not been captured by the model. 2

5. *The residuals shown in Figure 6 are clearly not white noise, and are not normally distributed. What features of the plots suggest this?*
   - The ACF shows significant serial correlation, so not white noise. 1
   - The time plot shows large outliers, so not normally distributed. 1

6. *If you produced forecasts from this model, how reliable do you think the point forecasts and prediction intervals would be? Why?*
   - The point forecasts will be biased around holiday times, as the model has not captured the holiday effects. 1
   - The prediction intervals may have inaccurate coverage due to the non-normality of residuals, and the serial correlation present in the residuals. 2

7. *Another model is found with 10 parameters, a slightly lower AICc value, but with no improvement in the residuals. Would you prefer this model to the one you have already fitted? Why or why not?*
   - No. Small differences in AICc don't matter much, and it doesn't solve the problems of the existing model. 2
   - Also a model with a large number of parameters is more difficult to estimate. 1

Total: 20 marks

SECTION E

1. *A dynamic regression model is fitted to the time series, using only data from March 2022. Write down the equations for the regression part of the model. There is no need to give numerical values for the coefficients.*

$$y_t = \sum_{i=1}^{10} \left[\beta_i \sin\left(\frac{2\pi i t}{365}\right) + \gamma_i \cos\left(\frac{2\pi i t}{365}\right) \right] + \eta_t$$

where $y_t = \log(\text{Boardings})$ and η_t is an ARIMA(0,0,1)(1,1,2)$_7$ process. 5

2. *Which coefficients in the model relate to annual seasonality, and which coefficients relate to weekly seasonality? What does the remaining coefficient handle?*

   - All coefficients associated with fourier terms relate to annual seasonality. 1
   - The `sar1`, `sma1` and `sma2` coefficients relate to weekly seasonality. 1
   - The remaining `ma1` coefficient handles any remaining serial correlation. 1

3. *The model could be improved by changing the number of Fourier terms used. Explain how you could determine the optimal number of Fourier terms to use.*

   - Start with no Fourier terms ($K = 0$), fit the model, and compute the AICc value. 1
   - Increase the K parameter by 1 and recompute the AICc value. 1
   - Repeat until the AICc value starts to increase. Pick the model with smallest AICc value. 1

4. *It is thought that days on which rain is forecast may have fewer passengers using public transport. How could you incorporate this information into your forecasts?*

   - Obtain rainfall forecasts from a weather service, and use these to determine when rain is forecast. 1
   - Add a binary variable to the regression model, with 1 on days when rain is forecast, and 0 otherwise. 2

5. **ETC3550 only** *To compare all the models used so far, you decide to use a test set comprising the observations from January 2024 to March 2024. Explain how you would use this test set to compare the models, and what you would look for in the comparison. Discuss whether this is the most reliable way to compare the models.*

   - Compute a training set comprising the data up the end of 2023. 1
   - Fit all the models to this training set, and compute forecasts for the period of the test set. 1
   - Compute the forecast accuracy on the test set using a suitable measure, such as the RMSE. 1
   - Looking for the smallest value of the forecast accuracy measure. 1
   - This is not the most reliable way to compare the models, as the test set is small, and the models may not be stable over time. 1
   - Better to use time series cross-validation, so you measure accuracy across multiple test sets. 1

6. **ETC5550 only** *You decide to compare all the models used so far, in a time series cross-validation comparison. Explain what this means, and why it is a useful way to compare the models in this way.*

   - Time series cross-validation involves using many training/test set splits, with each training set a little larger than the previous one. 1
   - Fit all the models to each training set, and compute forecasts for each test set. 1
   - Compute average forecast accuracy on test sets using a suitable measure, such as the RMSE. 1
   - Looking for the smallest value of the forecast accuracy measure. 1
   - A reliable way to compare models, as they are tested on a variety of different test sets. 1
   - It also allows for us to compute accuracy for a specific forecast horizon. 1

Total: 20 marks