Forecasting: Exam assignment 2022

IÉSEG School of Management

Practical issues

1. The deadline for handing in the paper is 2022-04-30 (midnight). Papers that are submitted after the deadline will not be accepted. In case of discussion, the time stamp in the email will be used as a criterion (unless grave circumstances can be argued).
2. Send the paper, the R scripts, and the data for the second exercise, by email.
3. Please put your name on the report, and number the pages.

Data sets

The data for the first exercise are in the Excel spreadsheet DataSets2022.xlsx, to be found on IÉSEG Online. The "content" tab in this file contains a description of the characteristics of the data sets, and a reference to the source.

Once you've assigned a working directory, you can load the data sets. These are stored in the other tabs of the Excel spreadsheet. For example, to load the Fatalities data set, you can use the following code:

```
setwd("C:/Temp") # Specify you own working directory here.
data <- read_excel("DataSets2021.xlsx", sheet="Fatalities")
Fat <- ts(data[,2], frequency = 1, start = 1965)
```

Other data sets in the Excel spreadsheet can be loaded in a similar way, by referring to the correct tab in the read_excel command (you need the readxl library for this).

Exercise 1

The data set Airpass_BE contains international intra-EU air passenger transport by Belgium and EU partner countries, from January 2003 to October 2021.

Split the time series in a **training set from January 2003 up to December 2017** and a **test set from January 2018 up to February 2020**. Use the training set for estimation of the methods/models, and use the test set for assessing their forecast accuracy.

The remaining observations (March 2020 - October 2021) do not belong to the training/test set, but we keep them for later reference.

1. Explore the data using relevant graphs, and discuss the properties of the data. Include and discuss a time series plot, a seasonal plot, a seasonal subseries plot and a (P)ACF plot.

2. Discuss whether a transformation and/or any other adjustment of the time series would be useful. If so, apply the most appropriate transformation and/or adjustments. Also, report the optimal Box-Cox lambda value that could be used to transform the time series. Clarify how you will proceed with the transformation in the remainder of the exercise.

3. Create forecasts using the seasonal naive method. Check the residual diagnostics (including the Ljung-Box test) and the forecast accuracy (on the test set).

4. Use an STL decomposition to forecast the time series. Use the various underlying forecasting methods for the seasonally adjusted data (naive, rwdrift, ets, arima). Check the residual diagnostics and the forecast accuracy and select the best performing STL decomposition.

5. Generate forecasts using ETS. First select the appropriate models yourself and discuss their performance. Compare these models with the results of the automated ETS procedure. Check the residual diagnostics and the forecast accuracy for the various ETS models you've considered. Present the parameters of the final ETS model and show the forecasts in a graph.

6. Generate forecasts using the `auto.arima` procedure. Present the estimated model using the backward shift operator. Include the parameter estimates. Check the residual diagnostics and the forecast accuracy. Discuss your results, and if necessary compare these with other possible ARIMA models (e.g. if small changes in the model specification improve the properties of the residuals and/or the forecast accuracy).

7. Compare the different models (naive, STL, ETS, ARIMA) in terms of residual diagnostics and forecast accuracy. Present the results in a summary table. Analyse your results and select your final model.

8. Generate out of sample forecasts up to December 2022, based on the complete time series **(January 2003 - February 2020)**. Present your results.

9. Now consider the last observations in the time series (March 2020 - October 2021). They correspond to the COVID pandemic times. What do you learn about the impact of the pandemic on air passenger transport between Belgium and other EU countries, based on the data and your final forecasts?

Exercise 2

For this exercise, find a recent and relevant time series to forecast. The time series should be original (not from `R` packages!), recent, sufficiently long and must include a seasonal component.

The time series is analysed according to a carefully selected forecasting process, using at least two techniques (other than the naive methods) that have been discussed during the lectures, and results should be compared. **Just providing `ets` and `auto.arima` results is not sufficient.** Describe your approach, and motivate your choices. Present a full analysis and description of the time series forecasting process.

General guidelines

- **Note that a fully elaborated paper is expected. The provided `R`-code will be checked only in case of doubt. What counts is your description, analysis and interpretation in the paper. A collection of R code and output is not considered a fully elaborated paper.**
- In each step of the exercises, discuss your results and explain your choices. Use additional tables and graphs to clarify your answer. Just providing `R` code and corresponding output is not sufficient to pass. You'll need to show some understanding of the course content.
- Please make sure that you analyse the correct data. This is very easy to verify in the first step of your analysis. I do not grade exercises on wrong data.
- Keep in mind that impudently copying text from the course material (including slides, R forecasting labs, solutions...) or from external sources is considered a very bad practice (it is plagiarism). Use references to sources whenever necessary. Cases of plagiarism will be reported to the program director.
- Please, read and re-read your own work before handing it in! Errors en inconsistencies make a paper very difficult to read.
- It is very useful to see the estimated parameters of your models, and your corresponding interpretation in the output. Therefore, include these in your answers.
- Do not include useless models. If the time series is trending, then a naive (non-seasonal) forecasts or a SES are not useful, so do not include them.
- If you split the data set in a training and test set, make sure that you assess forecast accuracy on the test set, not (only) on the training set or on the complete data set.

Good Luck, and Happy Forecasting!