Forecasting: principles and practice

Lab Session 5 24 September 2014

Before doing any exercises in R, load the **fpp** package using `library(fpp)`.

1. Consider the monthly sales of product A for a plastics manufacturer for years 1 through 5 (data set `plastics`).
   (a) Plot the time series of sales of product A. Can you identify seasonal fluctuations and/or a trend?
   (b) Use an STL decomposition to calculate the trend-cycle and seasonal indices. (Experiment with having fixed or changing seasonality.)
   (c) Do the results support the graphical interpretation from part (a)?
   (d) Compute and plot the seasonally adjusted data.
   (e) Use a random walk to produce forecasts of the seasonally adjusted data.
   (f) Reseasonalize the results to give forecasts on the original scale.
   [Hint: you can use the `stlf` function with `method="naive"`.]
   (g) Why do the forecasts look too low?

2. For this exercise, use the monthly Australian short-term overseas visitors data, May 1985–April 2005. (Data set: `visitors` in `expsmooth` package.)
   (a) Use `ets` to find the best model for these data and record the training set RMSE. You should find that the best model is ETS(M,A,M).
   (b) We will now check how much larger the one-step RMSE is on out-of-sample data using time series cross-validation. The following code will compute the result, beginning with four years of data in the training set.

```
k <- 48 # minimum size for training set
n <- length(visitors) # Total number of observations
e <- visitors*NA # Vector to record one-step forecast errors
for(i in 48:(n-1))
{
  train <- ts(visitors[1:i],freq=12)
  fit <- ets(train, "MAM", damped=FALSE)
  fc <- forecast(fit,h=1)$mean
  e[i] <- visitors[i+1]-fc
}
sqrt(mean(e^2,na.rm=TRUE))
```

   Check that you understand what the code is doing. Ask if you don't.
   (c) What would happen in the above loop if I had set `train <- visitors[1:i]`?
   (d) Plot `e`. What do you notice about the error variances? Why does this occur?

(e) How does this problem bias the comparison of the RMSE values from (2a) and (2b)? (Hint: think about the effect of the missing values in `e`.)

(f) In practice, we will not know that the best model on the whole data set is ETS(M,A,M) until we observe all the data. So a more realistic analysis would be to allow `ets` to select a different model each time through the loop. Calculate the RMSE using this approach. (Warning: it will take a while as there are a lot of models to fit.)

(g) How does the RMSE computed in (2f) compare to that computed in (2b)? Does the re-selection of a model at each step make much difference?

3. Try a similar cross-validation approach on one of the other time series considered yesterday.

   (a) Does the `ets()` model selection via AICc give the same model as obtained using cross-validation?

   (b) Which model would you use in practice?