Forecasting Competition

Due: Saturday, August 10th

The aim of this competition is to motivate each of you to practice basic forecasting methods in time series analysis, and to perhaps research and implement your own, more novel, methods in order to claim the distinction of STAT 443 Forecasting Champion!! You will be asked to provide forecasts/predictions in five different scenarios. Individual performance in each scenario will be ranked across competitors, and your final competition score will be your average rank in the five scenarios. As such each scenario is equally important in determining the winner. By the due date **August 10th, 11:59 pm**, you must complete and turn in via Crowdmark and LEARN the following:

1. Five txt files (.txt), containing your forecasts for each scenario. Detailed instructions on how to construct each file are given below, and an `R` script is posted on LEARN with demonstrations of how to create each file. The file names for the five files should be of the form "Scenario1_lastnameIDnumber.txt" to "Scenario5_lastnameIDnumber.txt", where "lastname" is replaced by your last name, and ID number is replaced with your ID number. These are to be submitted on LEARN to the LEARN dropbox for the forecasting contest.

2. A report containing:

   (a) A description of your models/how you produced your forecasts.

   (b) Plots of the data along with your forecasts, as well as basic diagnostic of your model that support their effectivness/goodness of fit. In some cases I may ask specifically for certain plots to be displayed , e.g. plots of your forecasts with 95% confidence intervals. In the case of Scenario 2 you must only show plots of your forecasts for no more than 2 stocks. This is to be submitted on Crowdmark.

3. Supporting code that is well commented and may be easily run to reproduce your forecasts. The files will also be submitted on LEARN.

Your report should be short (no more than 2 pages per scenario, excluding plots), and should be written as if the audience is someone with basic time series knowledge, i.e. you do not have to introduce basic time series models. It should not contain many typos, and should be easily readable. The code can be in Python or R. If you wish to use another language please consult with me.

Note: The data used in each scenario has been obtained from publicly available sources, but has been privitized/transformed by me to increase the difficulty of finding the source. Nonetheless, I assume that it is possible to invert my transformation and find the source data. THIS IS NOT ALLOWED! In any case, you must also submit the supporting code to reproduce your forecasts, and if this is found to include knowledge of the source data, you will be disqualified, and receive a zero for this assessment. If your method involves any randomized procedure, you must use some version of a set.seed() function so that your results are reproducible.

1 Scenario 1: Hydrological Forecast

`hydrology_data.txt` contains 576 observations at a monthly resolution of the level of a body of water. Your task in this scenario is to produce 1-month to 24-month ahead forecasts of the series, as well as a 95% prediction intervals for the forecasts. If $\hat{x}_i$ $i = 1, \dots 24$ denotes your forecast of the true levels x_i $i = 1, \dots, 24$, your error in this scenario will be measured by

$$MSE = \sum_{i=1}^{24} (x_i - \hat{x}_i)^2.$$

Ranking among competitors will be determined according to lowest MSE. The .txt file in this case should contain one COLUMN of length 24 containing the forecasts. Prediction intervals will be scored using prediction interval scoring rules.

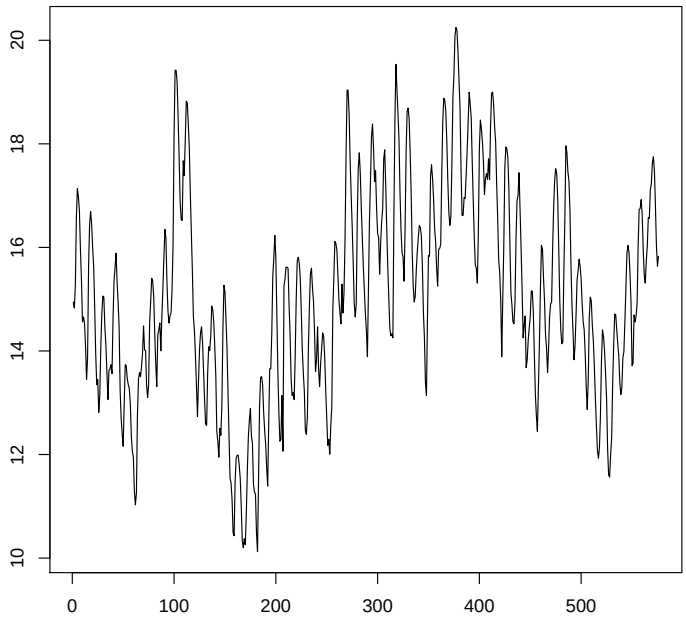

Figure 1: Hydrological time series.

2 Scenario 2: Financial Risk Forecast

The files `stock1.txt` to `stock40.txt` each contain 150 days of daily resolution log-differenced price data from several different stocks listed on the New York stock exchange. Your task in this scenario is to forecast (lower) 15% quantiles 10 steps ahead for each series (in other words, forcasts for Value-at-Risk). Your error will be measured as follows. Let $X_{i,j}$ denote the j'th observation from the ith log-differenced series. Letting $\hat{q}_{i,1}, \ldots, \hat{q}_{i,10}$ denote your quantile forecasts for the ith series, $i = 1, \ldots, 20$, we define

$$ERR = \frac{1}{400} \sum_{i=1}^{40} \sum_{j=1}^{10} (X_{i,150+j} - \hat{q}_{i,j})(0.15 - \mathbb{1}_{\{X_{i,150+j} \leq \hat{q}_{i,j}\}}).$$

For more information on this error measure/scoring rule for quantile prediction, see Gneiting and Raftery (2007), page 370. The .txt file in this case should contain 40 COLUMNS of length 10, the ith column containing the 10 quantile forecasts for stock i.

3 Scenarios 3 and 4: Imputation and Multivariate Time Series Forecasting

The file `beer.txt` contains a monthly resolution time series of the amount of beer produced in Australia. 30 observations in the middle of the times series have been removed and replaced with "NA" values. Your task for these scenarios is to:

1. Impute (predict) the missing values, and
2. Forecast the series `beer.txt` 24 steps ahead.

Your report should contain graphs with 95% prediction intervals for your imputations/forecasts. You may use the series in `car.txt` on car production, `steel.txt` on steel production, `gas.txt` on gas consumption, `electricity.txt`

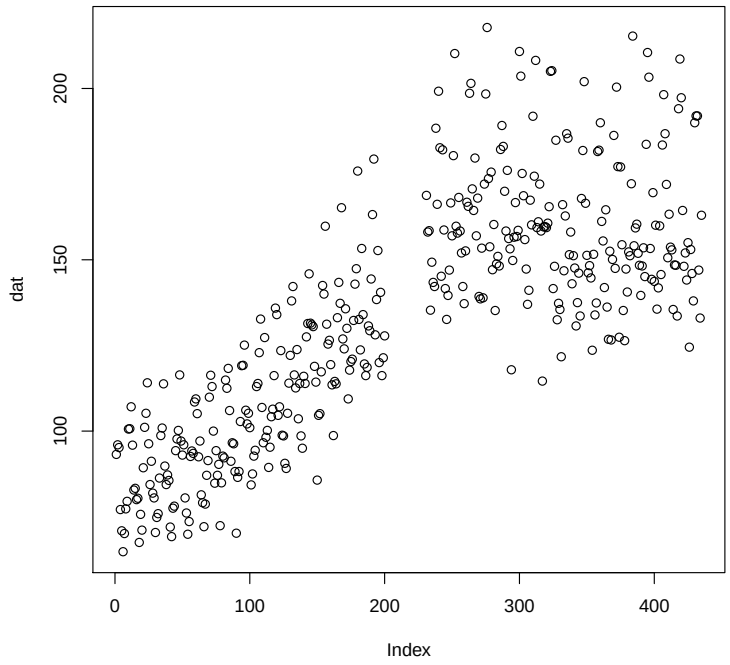

Figure 2: Plot of `beer` time series with missing segment displayed.

on electricity consumption and `temperature.txt` describing the monthly mean high temperatures, all from Australia, to improve your predictions. You may not use any additional information other than these given series. Letting $\hat{x}_1, \ldots, \hat{x}_{30}$ denote the predictions of the missing values $x_1, \ldots, x_{30}$, and $\hat{y}_1, \ldots, \hat{y}_{24}$ the forecasts of the future 24 values of the series $y_1, \ldots, y_{24}$, error will be measured in this case by

$$MSE(Scenario3) = \sum_{i=1}^{30} (x_i - \hat{x}_i)^2$$

and

$$MSE(Scenario4) = \sum_{i=1}^{24} (y_i - \hat{y}_i)^2.$$

You should provide 2 .txt files for these scenarios, which should each contain respectively 1 COLUMN of length 30 and 24, in which the 30 imputed values $\hat{x}_1, \ldots, \hat{x}_{30}$ are given in the file for Scenario 3, and the 24 forecasts $\hat{y}_1, \ldots, \hat{y}_{24}$ are given in the file for Scenario 4. In your report also produce and plot 95% prediction intervals for each of these imputations/forecasts. Prediction intervals will be scored using prediction interval scoring rules.

4 Scenarios 5: Long Horizon Pollution Forecasting

The files `pollutionCity1.txt`, `pollutionCity2.txt`, and `pollutionCity3.txt` contain standardized half hourly resolution measurements of the concentration of an air pollutant in three different cities over 53 days (time series of length 2544 for each city). Your task in this scenario is to forecast each series forward 1 to 336 half-hourly steps ahead, which corresponds to forecasting each series one week ahead, and produce 95% prediction intervals for these forecasts. If $\hat{x}_{i,1}, \ldots, \hat{x}_{i,366}$ denote your forecasts of pollution levels $x_{i,1}, \ldots, x_{i,366}$ in City i, $i = 1, 2, 3$, then your error in this scenario will be measured by

$$MSE(Scenario5) = \sum_{i=1}^{3} \sum_{j=1}^{366} (\hat{x}_{i,j} - x_{i,j})^2.$$

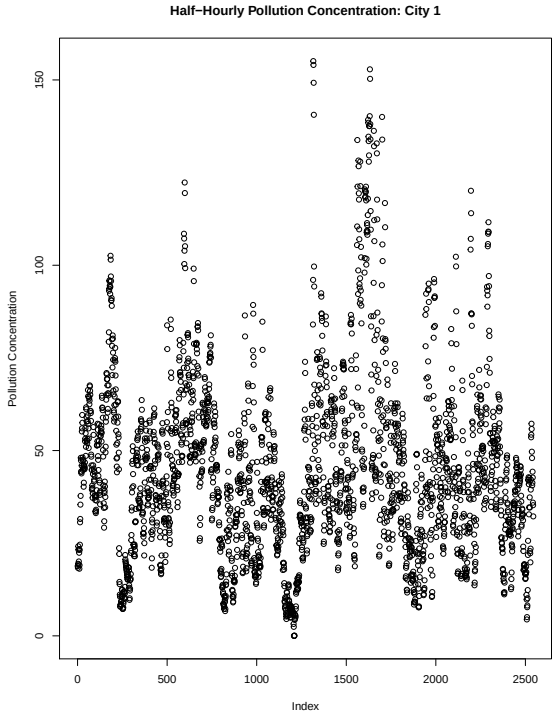

Figure 3: Plot of `pollutionCity1` time series.

The .txt file in this case should contain 3 COLUMNS, each with 336 rows, containing your forecasts for City i in column i. In your report also produce and plot 95% prediction intervals for each of these forecasts. Prediction intervals will be scored using prediction interval scoring rules.

5 Grading

Your grade on this assessment will largely be determined by your report, which will be graded as follows:

1. 40% Methods implemented are clearly explained and reasonable to use in each case
2. 40% Figures and model diagnostics are clear and reasonable.
3. 20% The report is readable and contains few typos

References

[1] T. Gneiting, and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* , *102*, 359–377.