Water-level Predictions for Fraser River
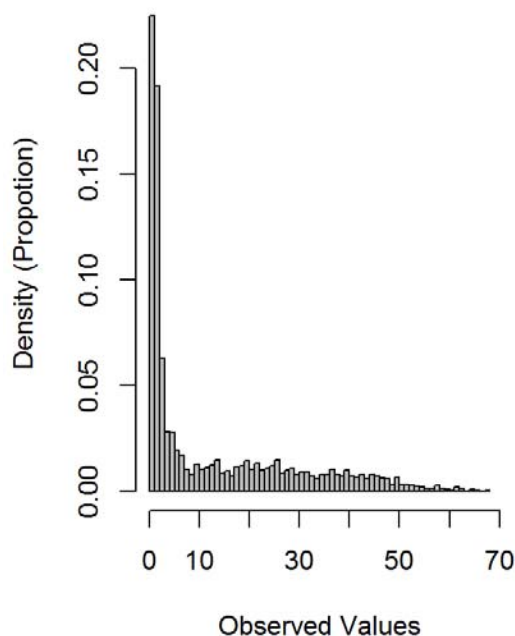
by Leshun Xu

18 Apr 2019

This is an example on how the GAM package works for the prediction of the water level of Fraser River in British Columbia, Canada. The data set is downloaded from the Canada government website (https://wateroffice.ec.gc.ca/download/report_e.html? dt=dd&df=ddf&md=1&ext=csv (https://wateroffice.ec.gc.ca/download/report_e.html?dt=dd&df=ddf&md=1&ext=csv) Please read the terms before you use the data).
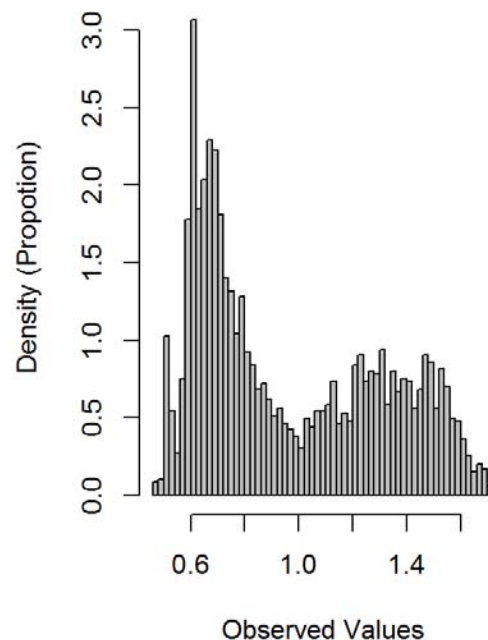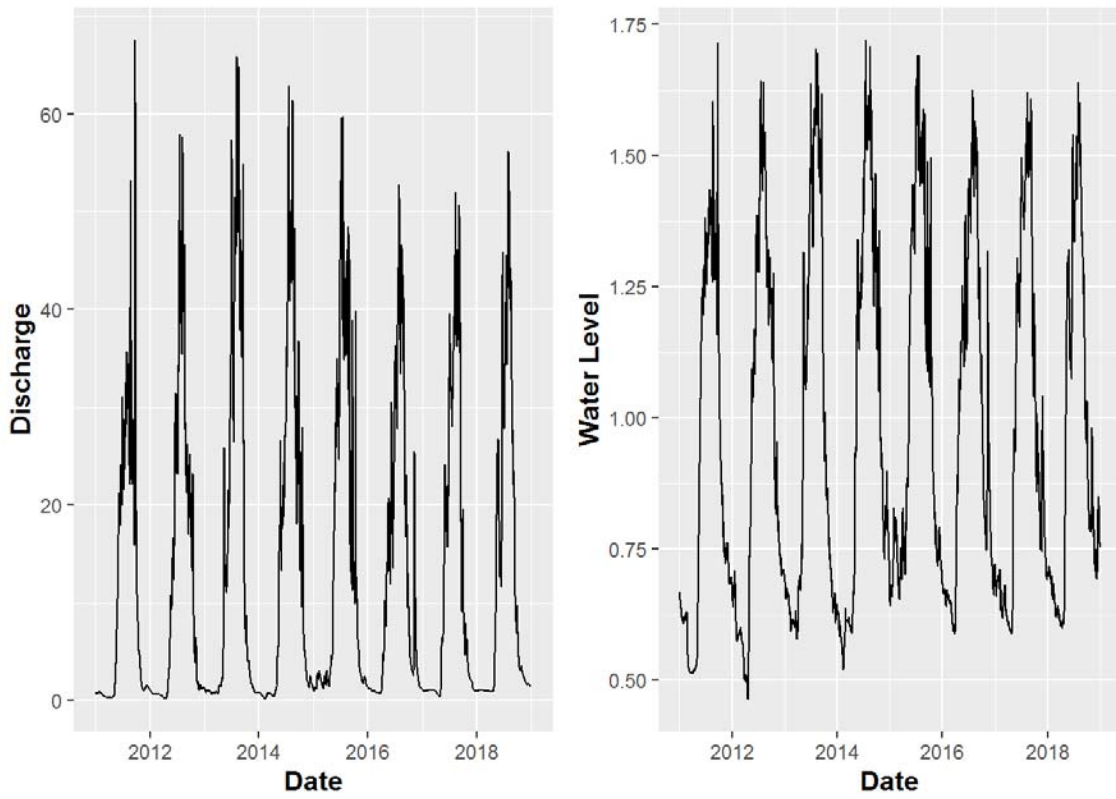
Overview of the data

The attached .csv file contains the daily discharge and the daily water level of a part of Fraser River, from 04-Jul-1978 to 29-Jan-2019. The data from 01-Jan-2011 to 31-Dec-2018 is extracted out as the data set used in this example. The following graphs show the distributions and the pattern of the values within the eight-year period.
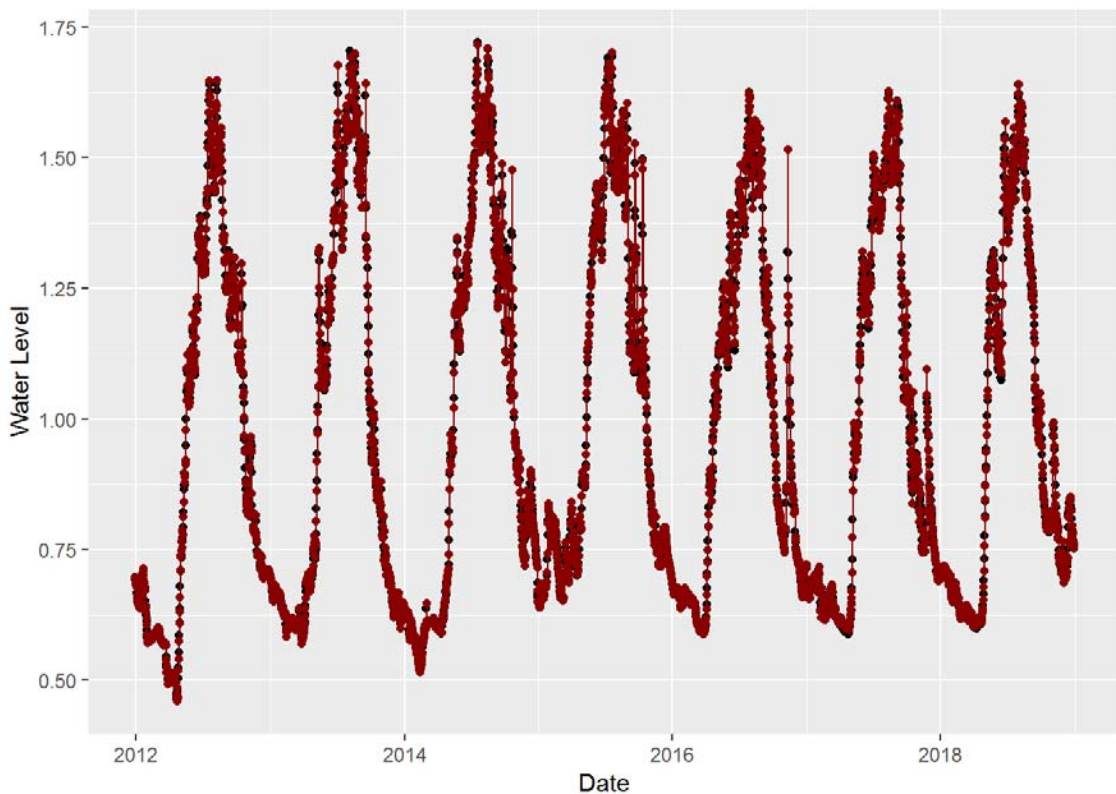
GAM model for the water level

In this example, we only focus on the prediction of water levels. Since we have the history data, in the following model, we use lag_i (i = 1, 2, 3, 4, 60, 362) to stand for the water level i day(s) before the predicted date. The $month$s are translated into numbers between 1 and 12. We introduce the following GAM model:
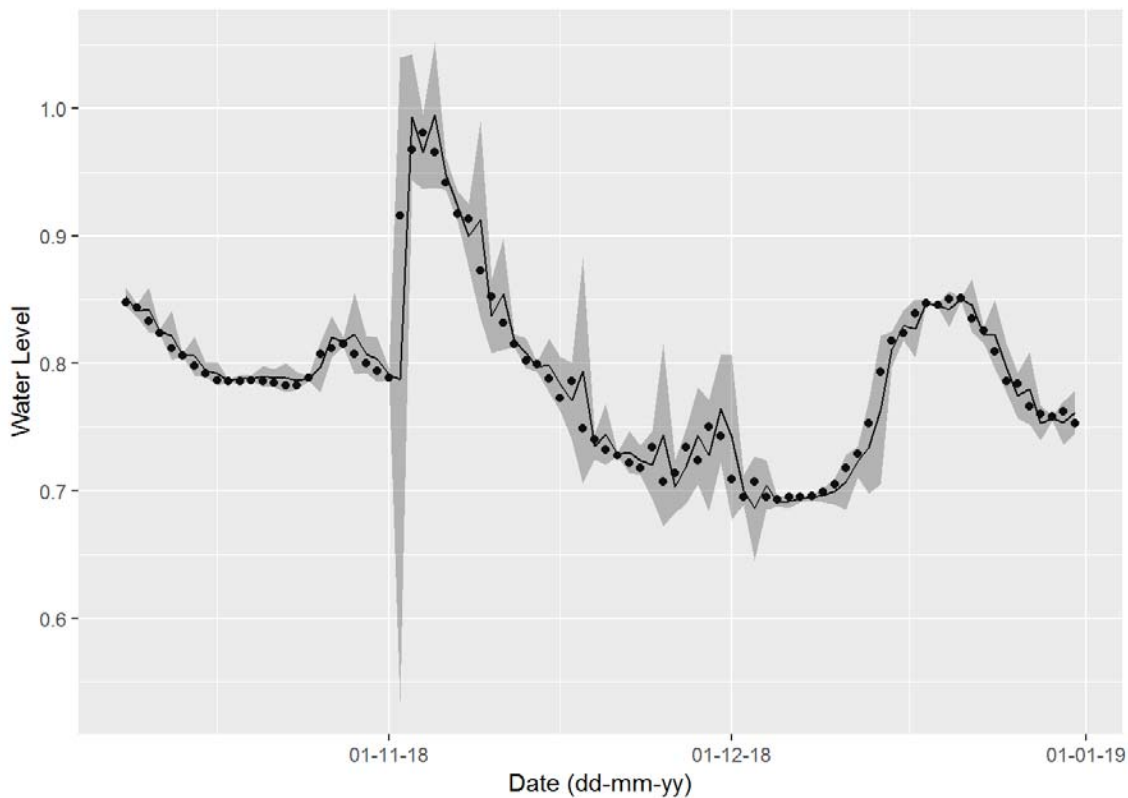
$$Value = \beta_0 + f_0(month, 6) + f_1(lag_1, 8) + f_2(lag_2, 8) + f_3(lag_3, 8) + f_4(lag_4, 8) + f_{60}(lag_{60}, 8) + f_{362}(lag_{362}, 8) + \epsilon,$$

where $f.(\cdot, k)$'s are functions expanded by cubic splines with k knots. The following graph implies that the model provides a good prediction for a training dataset (i.e. the eight-year period). The black dots represent true values, and the red points are predicted values.
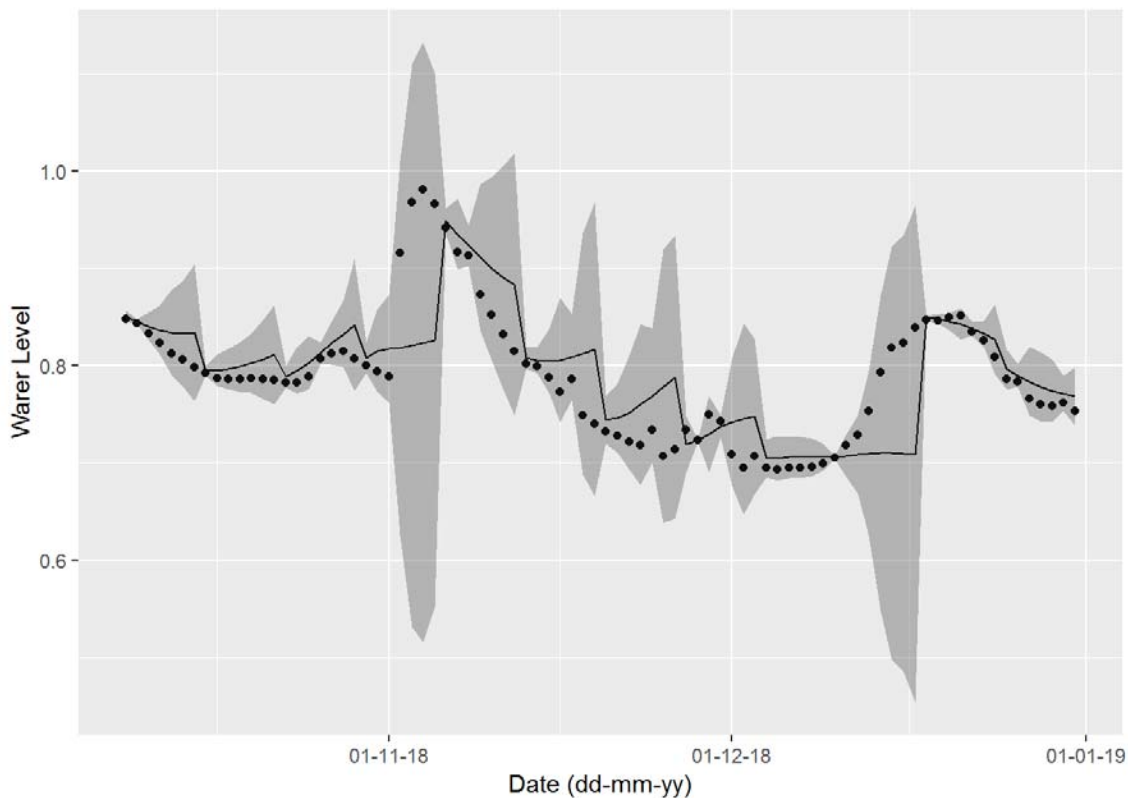
Prediction for tomorrow

Supposedly, we already know the water level of the previous days, including today, and we are going to predict the water level for tomorrow. It means the model has to be trained each day, and then it makes a prediction for the next day. In the following graph, 84 'tomorrows' are predicted. The solid line represents the predicted value; the black dots, true values; and the grey belt, the prediction intervals.

Prediction for the next 7 days

We use the water levels of today and the previous days to predict the following seven days without using any updated information. The prediction intervals cover all true values. The following graphs contains 12 seven-day predictions.

Remark to the model:

If we set "month" as a factor with 12 levels, we will get similar results, but with a little bit more freedom for the residuals. In this example, we will not go further into the discussion on overfitting, how to choose knots and lags, and whether treating "month" as a factor is good, etc.