Statistical Analysis Plan

Spatiotemporal modeling and predicting of average flows in Scotland based on generalized additive models

Mengran Li

Population

Daily river flows in Scotland.

Objectives

- To detect the data pattern of average daily flows over space.
- To detect the data pattern of average daily flows over time.
- To predict the average flows and evaluate the efficiency of the covariates in forecasting.

Data Collection

The daily flow data and gauge station information are provided by the Scottish Environment Protection Agency (SEPA) and the National River Flow Archive(NRFA). The data contains observations of 64 gauge stations during 01/01/1989 - 31/12/2015.

Variables Under Consideration

- **Flow**: average daily flow (m^3/s); **main response variables**
- **Time**: temporal effect, 1989 - 2015; **covariate**
- **Season**: seasonal effect, days within a year, 1 - 365; **covariate**
- **Space**: spatial effect, longitude and latitude, (easting, northing); **covariates**
- **Catchment.Area**: catchment area of the measured river (km^2); **covariate**
- **Max.Altitude**: max altitude inside a catchment; **covariate**

Missing Data Procedures

The proportion of missing values will be checked and the missing values will be imputed using linear interpolation separately for months.

Summaries to be presented

- For each gauge station, the river main characteristics will be displayed: catchment area, max elevation, mean flow (se), min flow, max flow, mean flow/catchment area etc.
- A kernel smoothed method will be employed to detect the trend of mean flows and smoothed lines will present the data patterns.
- A map will show the positions of gauge stations and the colors represent the mean flows which indicate the spatio patterns.

Models to be fitted

Candidate models

- Linear models
- GAMs

The models will be applied in one or several rivers for the initial exploration respectively and then in the whole dataset.

Variable sets

- **Base model**

Catchment.Area + Max.Altitude

- To consider **temporal** and **spatial** effects

Time + Season + Space + Catchment.Area + Max.Altitude

- To add **interaction** between **trend** and **season**.

Time + Season + Space + Time×Season + Catchment.Area + Max.Altitude

- To consider **interaction** between **temporal** and **spatial** effects.

Time + Season + Space + Time×Season + Time×Space + Season×Space + Catchment.Area + Max.Altitude

Model Selection

- GAMs
  - Choose the numbers of basis functions
  - A smoothness penalty is imposed on the spline coefficients.
  - Smoothing parameters are chosen based on the AIC/BIC/SIC/KL loss etc.
  - For each of four type models mentioned in *Variable sets*, one best model is selected.

Model diagnosis, prediction and comparasion

For the four linear models and the selected four GAMs,

- One of AIC/BIC/SIC/KL loss etc. is used to check the nested models.
- Likelihood ratio is employed to test the significance of covariates.
- The estimated parameters and fitted values with 95% confidence intervals will be plotted and compared.
- Residual correlations are investigated by means of empirical variograms and autocorrelation plots.
- The predictions of whole scottish mean flows via Kriging will be drawn in a map.
- Root mean square prediction error: leave-one-out cross validation to evaluate the performance of the selected models.

$$y_i = x_{1i} + x_{2i} + t_i + d_i + z_i,\ i = 1, \ldots, n.$$