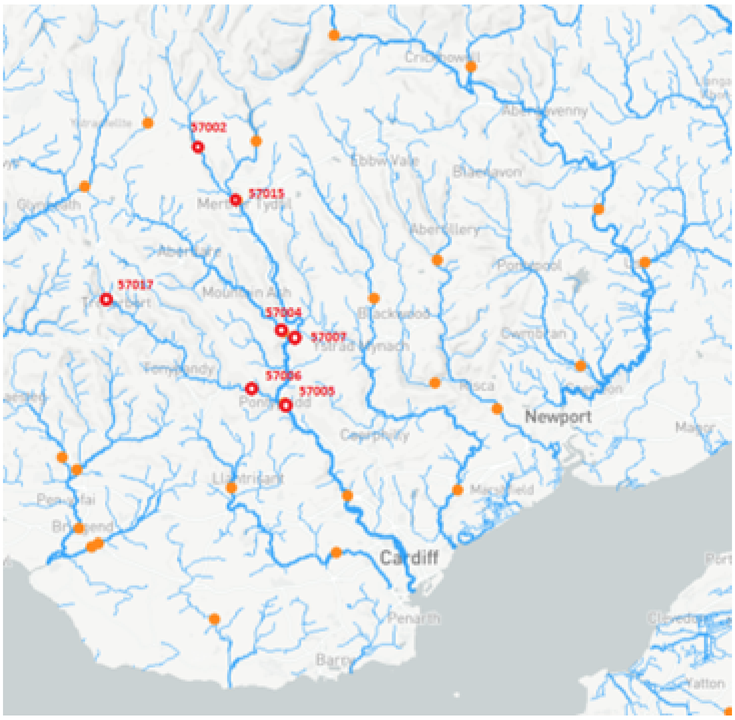
Forecasting next day river levels

MSc Dissertation Project

Consider a river catchment with multiple flow and rain gauges, such as the Taff catchment below (gauges in red). How well can we use all the data available to forecast the river level a day ahead?

Using daily flow and rainfall data from the National River Flow Archive (maintained by the UK Centre for Ecology and Hydrology), the goal is to build and test a model to forecast next day river levels. You will need to model the dependencies of flow measurements from the same catchment, the dependency of flow on rainfall, and the lag between rainfall and river flow.



1 Data acquisition and visualisation

Your first task is to acquire some data, clean it, and then use graphs to describe it.

For the data go to The National River Flow Archive at `https://nrfa.ceh.ac.uk/`. Choose a river catchment with at least half-a-dozen *active* gauges, and download:

- Daily flow data;
- Daily rainfall data;

Let me know which catchment you choose, because there are multiple people working on this project and you should have different data.

Be aware that there are gaps in the data and the recording periods for different gauges may be different, and the recording periods for river flow and rainfall at the same gauge may also be different. You should arrange the daily flow and rainfall data into a single data-frame with variables `date`, `gauge`, `gdf` and `cdr`. This step needs to be *reproducible*, that is, another researcher reading your report should be able to start with the same raw data and produce the same cleaned data. The best way to do this is to provide a script that takes the original data files and produces the desired data frame. If you find it necessary to use a text editor or spreadsheet to modify the original data files, then the changes made need to be catalogued precisely.

For cleaning and plotting the daily flow and rainfall data you can use either R or Python. In R the tidyverse packages `readr`, `dplyr` and `lubridate` have a lot of useful functions. When visualising the data, you should be looking for the following (not an exhaustive list):

- The distribution of flow and rainfall at each gauge;
- Relations between the flows at each gauge;
- Temporal dependence in the flows;
- Relations between the flows at each gauge and rainfall;
- Seasonal patterns and/or trends;
- Any anomalies.

Because of the natural variation in the data and the time it takes rainfall to reach the river, for many of these you will be better off using *monthly averages* rather than daily data. As for the data cleaning, the data visualisation should be reproducible, with the code used to produce the figures in your report given in an appendix.

2 Linear model

Suppose we are using data from the river Taff and its tributaries, as per the figure above. Let the response be $y(t)$ be the flow at Gauge 57005 on day t. Predictors include flow at Gauges 57002, 57015, 57004, 57007 and 57006 on previous days: call these $f_2(s), f_{15}(s), \ldots, f_6(s)$ for $s = t-1, t-2, \ldots$. Other predictors are rainfall on the catchments for 57002, 57015, 57004, 57007 and 57006 on previous days: call these $r_2(s), r_{15}(s), \ldots, r_6(s)$ for $s = t-1, t-2, \ldots$. Then to start consider linear models of the form

$$\begin{aligned}
y(t) &= \alpha_0 \\
& \quad +\beta_{2.1} f_2(t-1) + \beta_{2.2} f_2(t-2) + \cdots \\
& \quad +\beta_{15.1} f_{15}(t-1) + \beta_{15.2} f_{15}(t-2) + \cdots \\
& \quad \vdots \\
& \quad +\gamma_{2.1} r_2(t-1) + \gamma_{2.2} r_2(t-2) + \cdots \\
& \quad \vdots \\
& \quad +\epsilon(t)
\end{aligned}$$

where the $\epsilon(t)$ are i.i.d. normal errors.

Depending on how well that works you should also consider replacing α_0 with seasonal intercepts (or possibly deseasonalise the data first), interactions between the variables, non-linear transforms such as $\log(x)$ or $\log(1+x)$, etc.

3 Machine learning model

The linear model should give you a feel for what variables are useful for forecasting flow. Your final task is to choose a machine learning approach, use it to forecast day ahead flow, and compare the results to those from the linear model.