**Watershed challenge**

**Important points to take into account:**

This challenge shouldn't take more than 5 hours, so we don't expect perfect answers. The difficulty will grow, so try to answer as much as you can (it is not mandatory to complete the whole challenge).

We will only accept Jupyter notebook/lab files and they need to run!

**Motivation**

**Note**: the following paragraph is only for understanding the problem and why it is relevant. It is not necessary to fully understand the background, the challenge itself will guide you. So read fast this information and follow with the challenge.

The heat waves are extreme meteorological events (extreme temperatures) which can impact negatively in our ecosystem. This situation can have multiple impacts; from very hot summer days and wildfires, to floods and alluviums. As the heat waves depend on atmospheric circulation phenomenas, which can be detected few days in advance, there is a big opportunity for trying to predict the occurrence of the negative impacts associated with it. This is key in the context of global warming, where the frequency of this events have grown in the last century (<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2012GL053361>). We can explore the opportunity of predicting heat waves, in particular the peak-flows events, using a Chilean monitoring network of hydro meteorological variables.Thanks to public institutions such as DGA and CR2, we have data from different stations, some of which measure temperature and precipitations, and others measuring the total volume of water that a hydrographic watershed contribute to a section of river.

Some of the key questions we could try to answer are:

* Has the frequency of heat waves events increased over the last years?
* Is there any relationship between heat waves and peak flow events?
* If so, can we correlate those events with the watershed's features?

**Instructions**

As we discussed above, we will try to predict extreme watershed events in Chile. For this we count with a public and real dataset from meteorological stations. Each row represents the measure of daily flux in a specific station. This measure will be associated with watershed's features (fixed) and daily temperature and precipitation measures of nearby stations.

The file flux.csv contains all the data used for this challenge. This database was produced by us and synthesizes flux, temperature and precipitation data.

**Note about the database**: the stations which measure flux and the stations that measure temperature and precipitation are not located at the same place. For building this database, we took the watershed's upstream polygon and find the temperature and precipitation stations inside this polygon, and we calculated the average over that variables. In this way, every flux measure will be accompanied of a unique temperature and precipitation measure.

The database contains the following variables between others:

* station\_code: station code
* station\_name: name of the watershed
* date: date of measurement
* flux: water flux for that day
* avg\_precip: average precipitation for that day in that watershed
* avg\_max\_temp: maximum average temperature for that day in that watershed

**Challenge**

1. Download the file flux.csv from github (compressed as flux.csv.zip).
2. Perform an EDA over flux.csv file.
3. Plot flux, temperature and precipitations:
   * a) Write a function that plot a time series of a specific variable (flux, temp, precip) from a station. Should look like this:
   * def plot\_one\_timeserie(cod\_station, variable, min\_date, max\_date):
   * b) Now write a function that plots the 3 variables at the same time. As the variables are in different scales, you can normalize before plotting them. Should look like this:
   * def plot\_three\_timeseries(cod\_station, min\_date, max\_date):
4. Create three variables called:
   * flux\_extreme
   * temp\_extreme
   * precip\_extreme

This variables should take the value of 1 when that variable in a specific day was extreme. Being extreme could be considered as being greater than expected. For example, a flux can be considered as extreme (value 1) when is over the 95 percentile of the flux distribution for that specific season, and takes the value 0 otherwise. Taking into account the seasonality of that variables is very important, because could be considered as extreme in wintertime, but it’d be a normal temperature for summertime.

Do you consider this a good way of capturing extreme events? Or you would have used a different method? Which one?

1. Plot the variable flux\_extreme. Are there any different behaviours among different watersheds?
2. Plot the percentage of extreme events during time. Have they become more frequent?
3. Extreme flux prediction. Train one or many models (using your preferred algorithms) for estimating the probability of having an extreme flux. Feel free to create new features or use external variables. Some of the discussion we would like to see: Which data can be used and which cannot? Of course, we cannot use future data, but what about data from the same day? Or from the previous day?

Everything depends on how you propose the model use. Make a proposal on how you would use the model in practice (for example, once trained, the model will predict next day probability). Depending on your proposal, set constraints about which variables you can or cannot use.

1. Analyze the model results.
   * a) What is the performance of the model? Which metrics you consider are the best suited for this problem? What are the most important variables? What do you think about the results?
   * b) If we wanted to identify at least 70% of the extreme flux events, which are the metrics of your model for that threshold? It is a useful model?
2. Upload your work to a public repository (MIT licence desired) and send the link to us:
   * **To**: [mariapaz.salvatierra@bain.com](mailto:mariapaz.salvatierra@bain.com)
   * **Cc**: [aline.andrade@bain.com](mailto:aline.andrade@bain.com)
   * **Subject**: “Watershed Challenge”