Project
APDSE 2022/23

1. Introduction

Consider the weather station data provided from the region of Bombarral/Cadaval in the center west of Portugal. The information covers part of 2020 and 2021, and is obtained from several weather stations from different providers.

The data is distributed in several CSV files:

- data/datasources_coordinates.txt (the datasources with coordinates)
- data/datasources.txt (the datasources, with extra information about provider and type)
- data/series.txt (the metadata about series)
- data/series_XYZ.txt (the raw data for the series for the datasource XYZ)

A DataSource has an identifier, name, an external identifier, and coordinates. Additionally, it is associated with each DataSource its type and the corresponding provider. Each DataSource can provide information about several variables, i.e. time series.

A Series has an identifier, the corresponding DataSource identifier, its type (1-Raw Series, 2 – Derived Series), some properties namely if it is active, and the corresponding variable being measured. Every variable has an identifier, a datatype identifier (identifying the underlying physical variable being measured), a series period identifier characterizing the period of observation for each value, and the form of aggregation of periods in that period (the calculation method); all these are combined in a standard name of the form CALCULATION_METHOD(DATATYPE PERIOD). The symbol of the unit of the datatype as well as several text columns are provided in file series.txt

The raw data for each datasource contains in a row the series identifier, a uniquely generated sequential identifier, the data timestamp, the data real value, and the creation time.

All the files may have additional columns that either are easily interpretable or that can be discarded.

You should deliver:

- A commented JupyterLab notebook with clear identification of the following steps

2. Data Wrangling

The following data wrangling tasks are fundamental to be able to prepare the data for analysis. They might not be complete, and if needed, you should complete them in the way you deem necessary. If for some reason you are not able to process the complete data in your computer, divided it in batches so that it becomes possible to execute the code, and store the information in extra files.

2.1 Decide the format for your data (long, wide, etc.);

2.2 Add extra columns for simplifying understanding of the series variable, namely by splitting the series variable standard name of the form CALCULATION_METHOD(DATATYPE PERIOD) into its parts, namely by creating text columns with: the calculation method, the datatype, and the period;

2.3 Additionally, check If there are any duplicates, and try to understand why, suggesting an appropriate duplicate removal strategy;

2.4 Perform an analysis of missing values in the way you find more convenient (see also below);

2.5 Define a function that allows you to plot easily the minimum, mean, maximum values in period in a given time span (e.g. hourly, daily, weekly), for a particular series identified by its series_id or any other key.

3. Data Interpolation

You will realize that there is some missing data, and therefore you will explore several alternative forms of filling in the information. We will proceed in two stages

3.1 Filling missing information using interpolation methods (see for instance https://docs.scipy.org/doc/scipy/reference/interpolate.html for several methods besides linear interpolation), at the finest weather time resolution (typically, 15 minutes). This should be implemented in a function that besides the data it receives a parameter defining the maximum consecutive missing values that you are willing to allow in order to apply the desired interpolation method;

3.2 Aggregate the previous data hourly;

3.3 If there are missing values at the hourly value, you can employ the same or other interpolation methods to construct a new derived series, also with the control parameter of maximum hours that you are willing to consider;

3.4 Characterize daily each of your series with the number of missing values, or by another form that permits the filtering or visualization of series depending on its "quality".

4. Data Understanding

Perform an exploratory analysis of the data, regarding value distribution of the variables, and the variability due to seasons.

4.1 Perform the analysis globally and for each individual station, for the major weather variables: **temperature**, **relative humidity**, **wind speed**, **precipitation**, and **leaf wetness**; Check the correlation of these weather variables for each individual station.

4.2 Perform correlation analysis among the several weather stations and major weather variables (temperature and humidity)

5. Calculating Leaf Wetness

Major variables for plant disease modeling and fungi development are leaf wetness duration, air temperature, relative humidity, and wind speed. Most weather data providers do not make available leaf wetness duration, therefore it is necessary to estimate this variable. Some methods are very simple, and you find several of them in the folder “papers”.

5.1 Perform UMAP (or t-SNE) dimension reduction for the data, trying to determine if data is separable regarding the cases where leaf wetness = 0 versus leaf wetness > 0.

5.2 Use RH, DPD, and CART/SLD methods to determine if in a period of 15 minutes leaf_wetness > 0 (see the papers);

5.3 Construct **one of the two types of models** using random forests (you can try other methods, if you have time), to determine if the leaf is wet or not in a 15 minute period (i.e. leaf_wetness > 0), and the other to estimate the leaf wetness duration in minutes;

5.4 Evaluate the previous methods using the standard methods and metrics against the ground-truth, for the classification case (leafwetness > 0) and regression case (estimate leaf wetness duration).