REPORT ABOUT THE ASSIGNMENT

MATHIEU GUIGUE

1. Introduction

Meteorological data such as the atmospherical pressure, the relative humidity and the temperature have been provided over several days. The goal of this study is to extract relationships between these variables.

NetCDF4 files containing the data collected over a day have been given. The first part of the assignment consists in reducing the amount of data and especially removing the statistical fluctuations in the data that don't bing any useful piece of informations. A second part aims at finding clusters in the parameters space corresponding to locations where the parameters share similar behaviors.

2. Part 1: Data averaging

2.1. **Time series.** Figure 1 shows the data time series as extracted from the files. We have 5 days of data, with datapoints each 60 seconds. We verify that the data quality checks are good for all of these data.

Each file corresponds to a day of data, one data point every 60 seconds. Each day has its own time stamp (number of seconds since midnight) and a hardcoded-modification of the reader function is required to get things working.

2.2. **Estimating the optimal average period.** In the assignment a data averaging over 5 minutes (5 points) is requested. This aims at reducing the measurements fluctuations which are not bringing informations to the analysis (a.k.a. white noise fluctuations).

We can try to estimate the optimal value for such averaging. Using an Allan[1] representation of the data, I can estimate what the typical period where the measurements fluctuations are Gaussian is. Figure 2 shows the Allan Standard Deviation as a function of the averaging period. For short averaging period, the deviation decreases as the fluctuations are purely white. At some point, the deviation stops decreasing: this can be interpreted as this is the typical moment at which fluctuations are not Gaussian anymore. The corresponding averaging period would then be the optimal averaging period. It seems that the value at

Date: March 21, 2018.

[1] Allan Variance wiki page

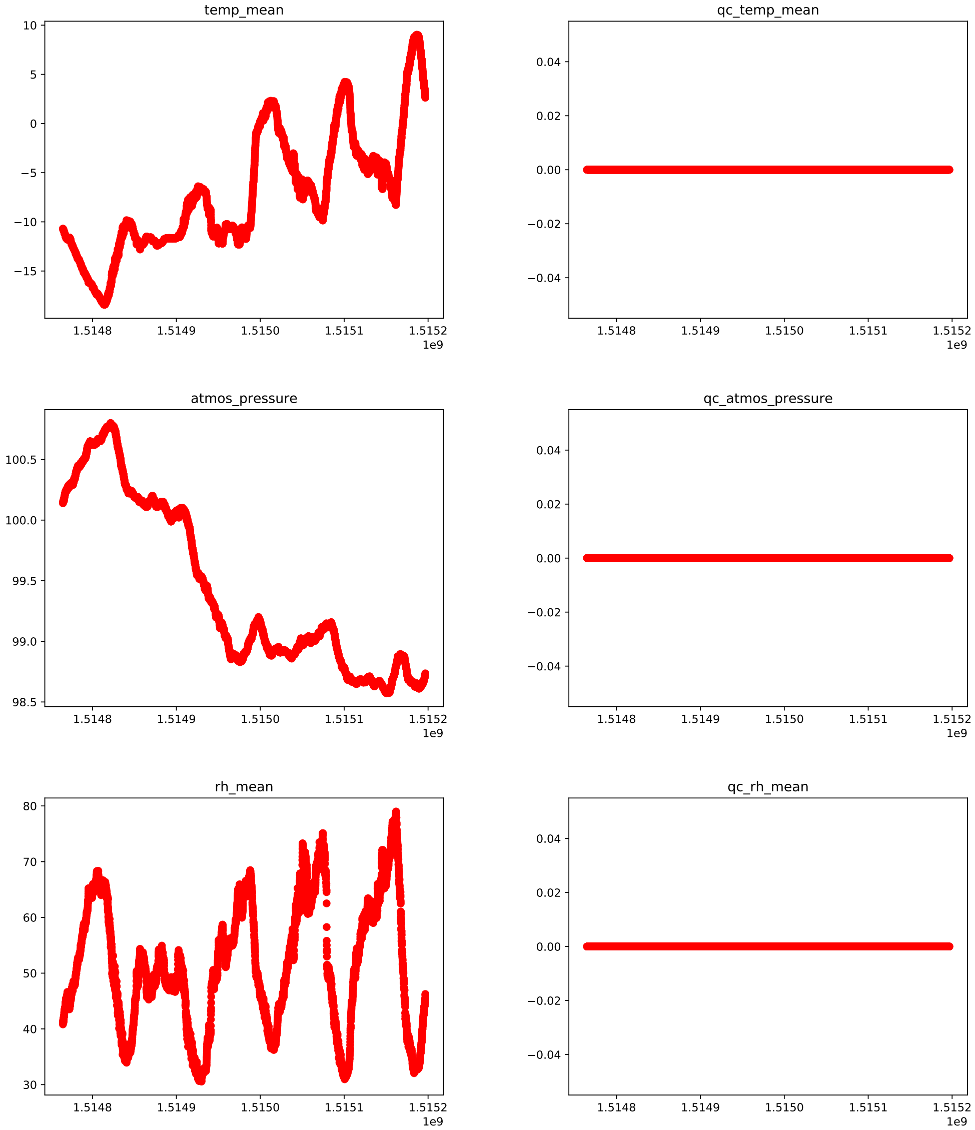

FIGURE 1. Time series of the mean temperature, mean relative humidity and atmospheric pressure (left) and the associated quality check value (right) as a function of time.

which non Gaussian fluctuations appear is defined for all three variables to be about 50 samples.

To-Go-Further 1. *Implement an algorithmic way to extract from the Allan Standard Deviation the number of measurements to average (something like a minimum finder).*

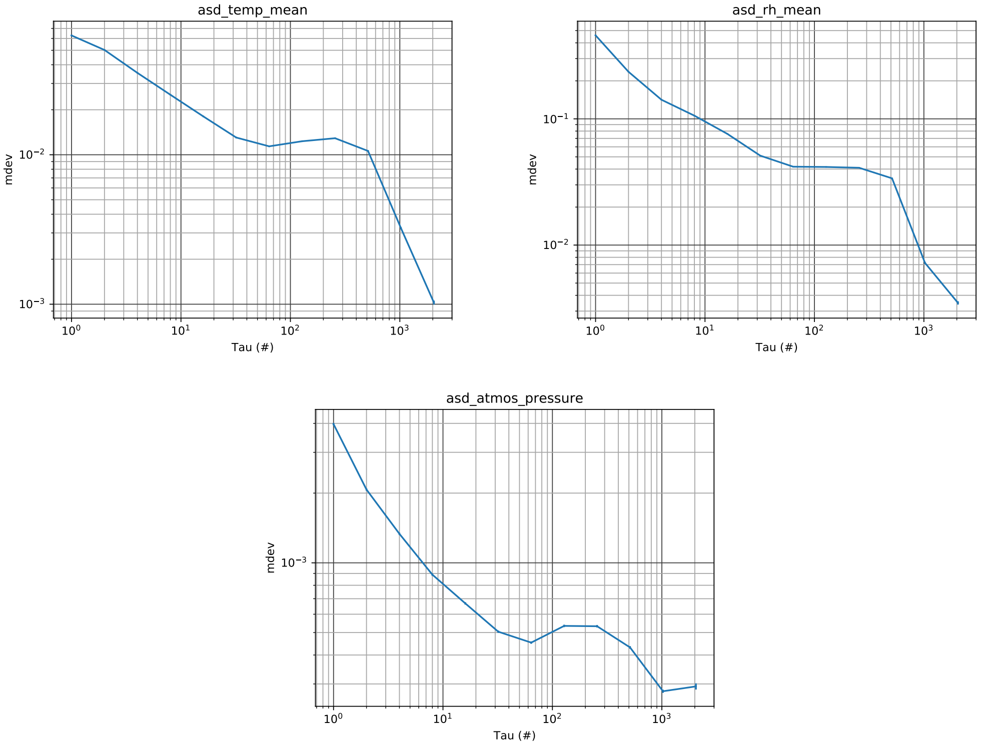

Figure 2. Allan Standard Deviation of the mean temperature, mean relative humidity and atmospheric pressure as a function of the period (in number of samples).

The averaged data (represented on Figure 3) are then saved in a CDF file named sgpmetavgE13.b1.20180101.000000.cdf.

3. Part 2: Clustering of data

3.1. ***k*-means algorithm.** The goal here is to determine how the data are clustered. The features we will use to find these clusters are the average temperature, atmospheric pressure and relative humidity and the derivatives of these quantities. The clusters locations are determined using a k-means clustering algorithm that works in 2 steps that are repeated until convergence. Given a number of cluster to find and a initial guess of the clusters centers, it first assigns all the data points to a cluster by determining the closest cluster center to each datapoint. Then it calculates the average position of the obtained cluster.

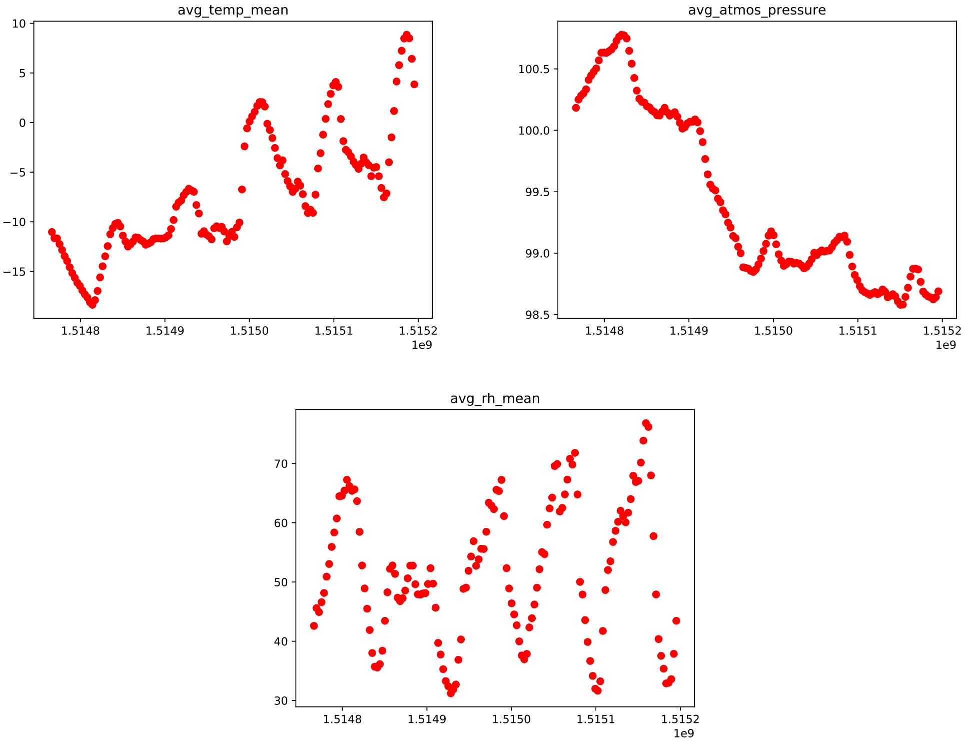

FIGURE 3. Time series of the average of mean temperature, mean relative humidity and atmospheric pressure as a function of time.

Here I have used the Scipy clustering package[2]. Figure 4 shows the distance between the datapoints and each cluster center for various numbers of clusters.

3.2. **Determination of the optimal number of clusters.** If the number of clusters we assume exists is too large, we encounter a over-fitting (or over-interpretation) issue, where statistical fluctuations are interpreted as characteristics that actually don't exist. The optimization of the number of clusters consists in determining the transition location when overfitting happens.

The derivative of the average temperature is calculated and used in this study. We divide the dataset in two subsets: 90 % goes as training dataset and the rest as testing set. A shuffling of the data is done, so the time correlations can be ignored. We then determine the clusters locations using the training set and we estimate the distance between the testing set data and the clusters centers. For several numbers of clusters, we repeat this procedure and extract the average distance. Figure 5 shows the results of this approach.

[2] Scipy clustering package

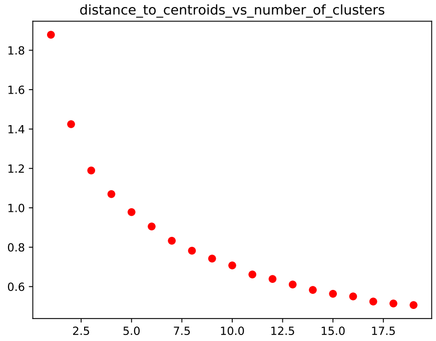

FIGURE 4. Distance to the cluster centers as a function of the number of clusters for the entire dataset.

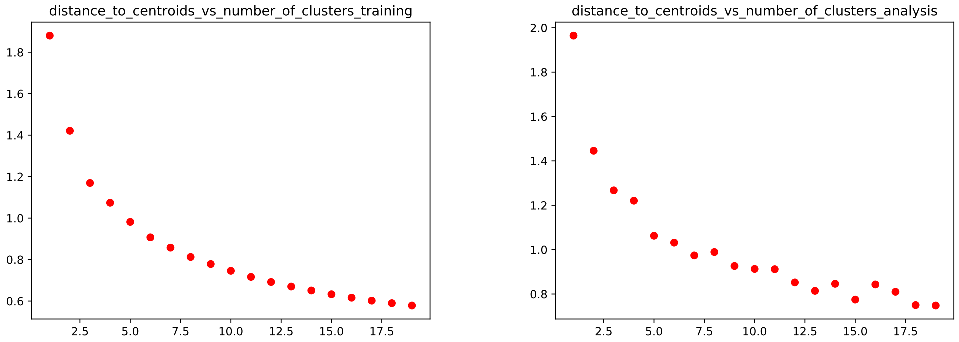

FIGURE 5. Distance to the cluster centers as a function of the number of clusters for a 5 minutes averaging: the plot on the right corresponds to the training dataset and the right to the testing dataset.

We can see that, as we can expect, the more cluster the better the algorithm is performing on the training set. A similar behavior is observed for the testing data, but the distances are generally larger for this set.

The best value for the number of clusters could be when the improvement rate of the algorithm in classifying the data decreases because it starts to explain statistical fluctuations by using clusters. On the figure 5, this corresponds to a change in the slope of the distance to centroids. On the training set, this phenomenon appears around $k = 6$. A similar inflection point appears for the testing set around $k = 6$. The optimal number of clusters seems to be about 6.

As a comparison, the figure 6 shows the distance to the cluster centers as a function of the number of cluster, but for an averaging of the data of 50 minutes (instead of 5, as

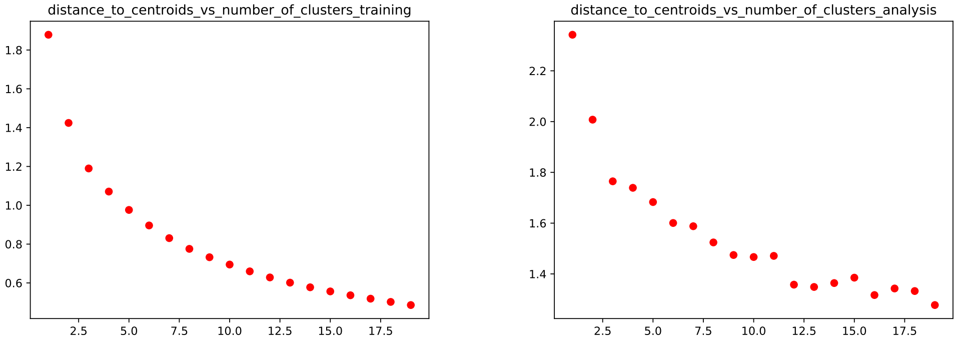

Figure 6. Distance to the cluster centers as a function of the number of clusters for a 50 minutes averaging: the plot on the right corresponds to the training dataset and the right to the testing dataset.

requested by the assignment). The same conclusions can be drawn. However a difference in the absolute values of these distances is observed, this would required some investigations to understand what is causing this.

Figure 7 shows the 2D representations of the data with the centroids location using $k = 6$. We can see that the locations of the centroids make sense, especially on some plots.

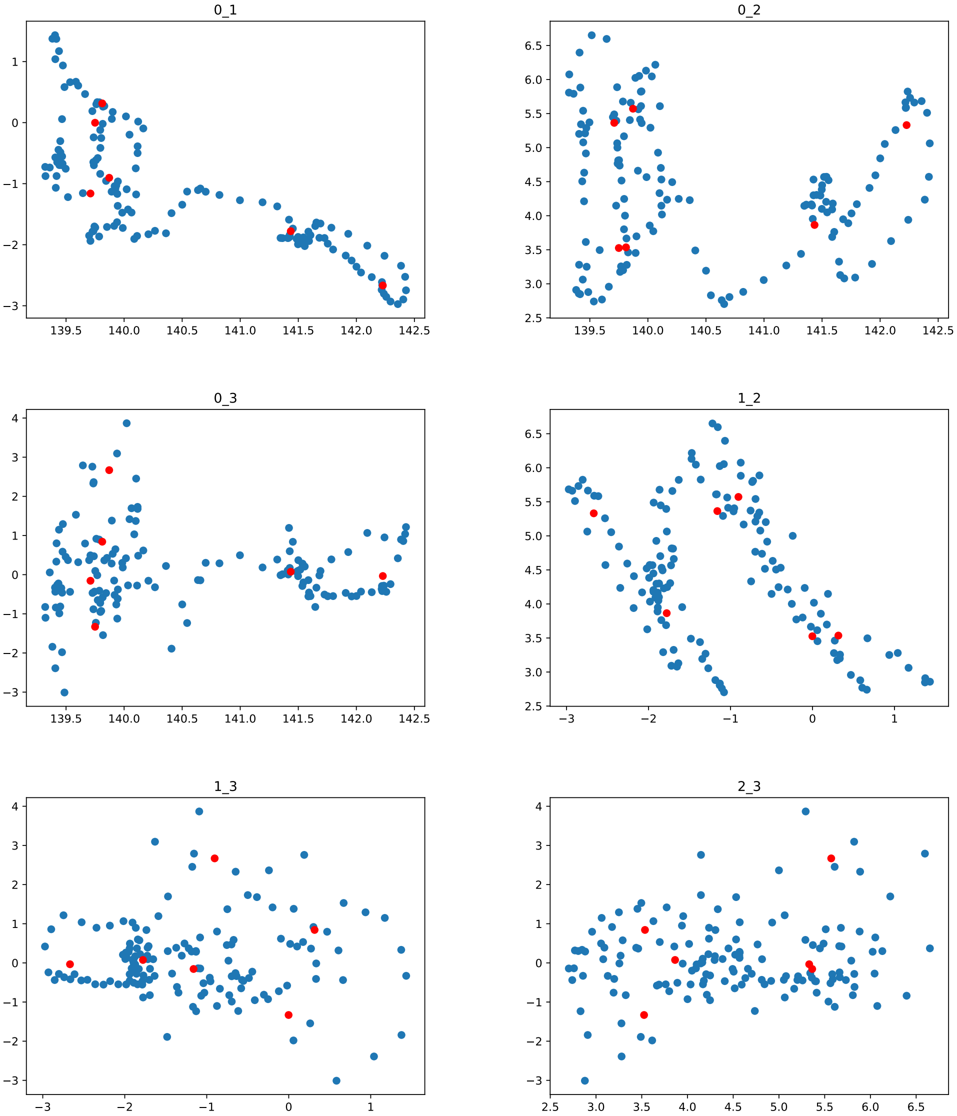

Figure 7. Data (blue) and 6 centroids (red) obtained by the k-means algorithm.