REPORT ABOUT THE ASSIGNMENT

MATHIEU GUIGUE

1. Introduction

2. Part 1: Data averaging

2.1. **Time series.** 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.** Using an Allan representation of the data, I can estimate what the typical period where the measurements fluctuations are Gaussian is. It seems that the value at which non Gaussian fluctuations appear is defined for all three variables to be about 60 samples. For safety, I will use 50 samples for averaging.

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

The averaged data (represented on Figure **??**) are then saved in a CDF file.

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.

Here I have used the scipy clustering package[1]. Figure **??** shows the distance between the datapoints and each cluster center for various numbers of clusters.

3.2. **Determination of the optimal number of clusters.**

Date: March 21, 2018.

[1] Scipy clustering package

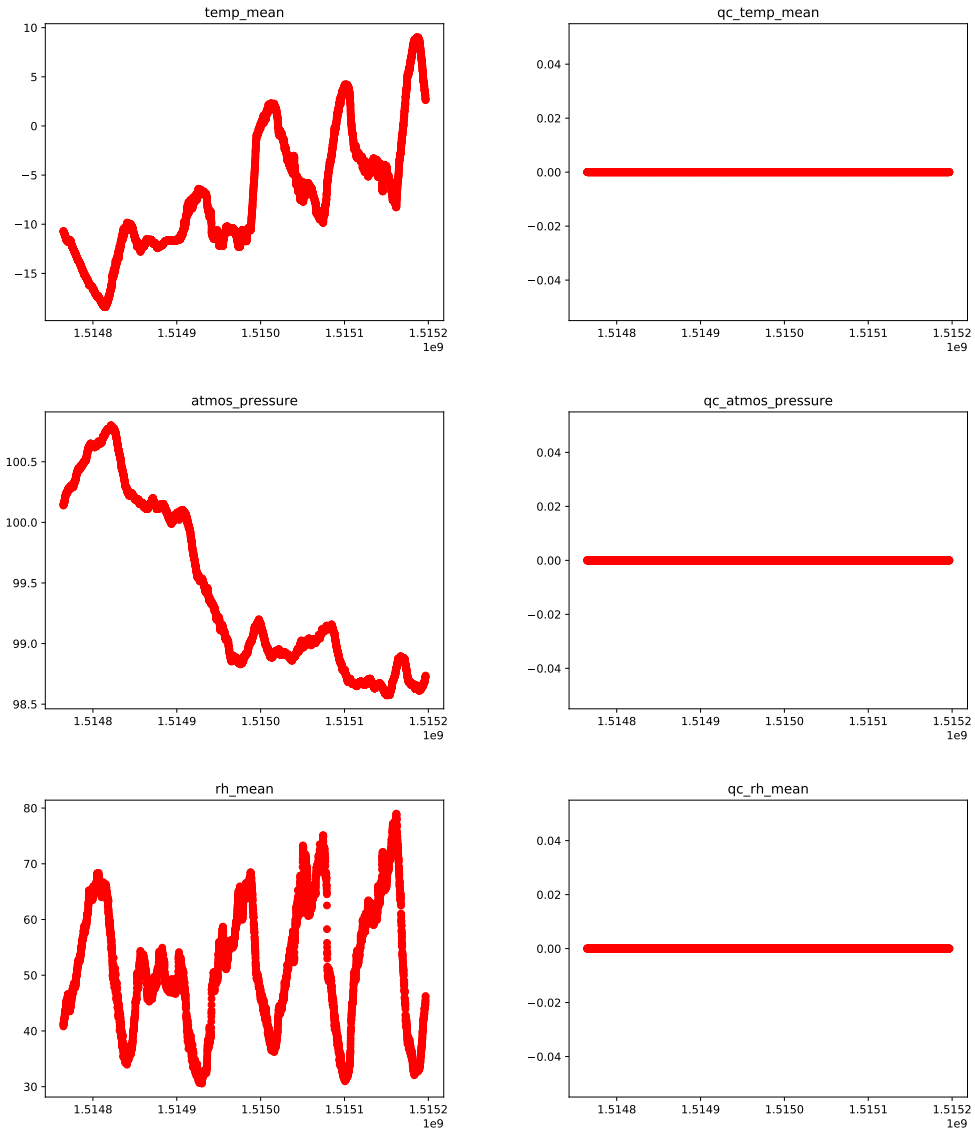

FIGURE 1. Time series of the mean temperature, mean relative humidity and atmospheric pressure (left) and the associated quality check value (right) as a function of time.

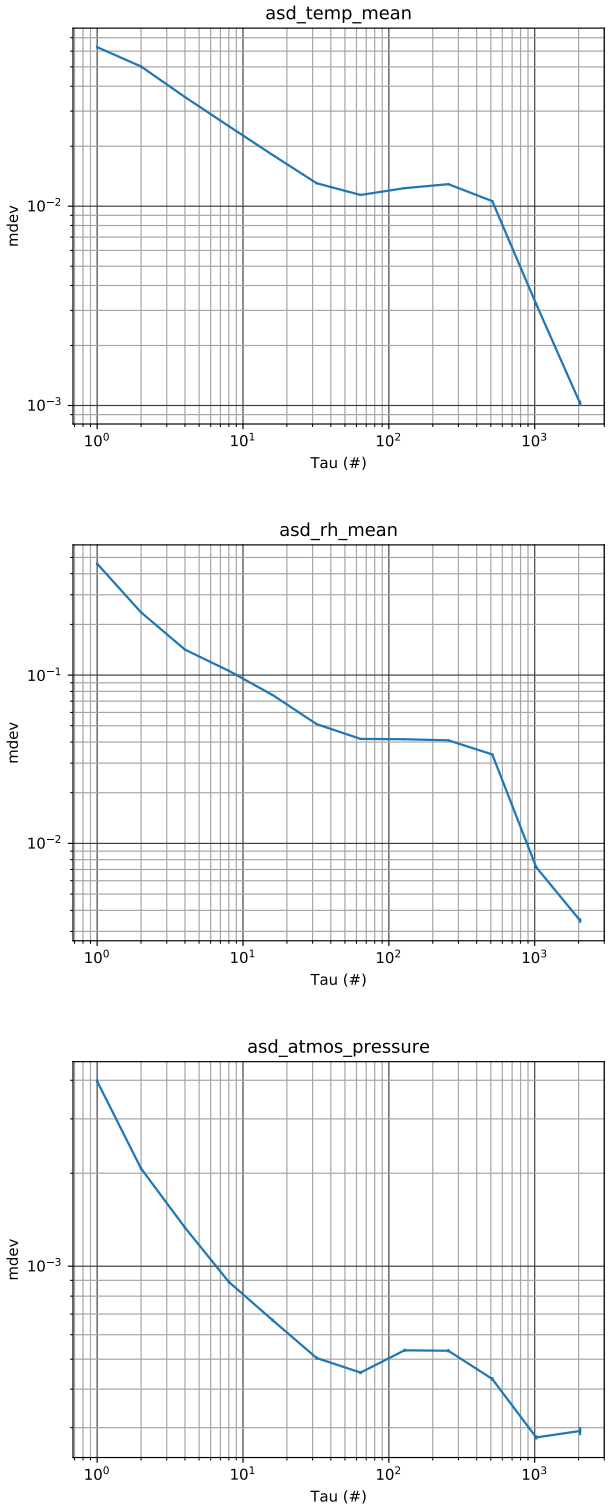

FIGURE 2. Allan Standard Deviation of the mean temperature, mean relative humidity and atmospheric pressure as a function of the period (in number of samples).

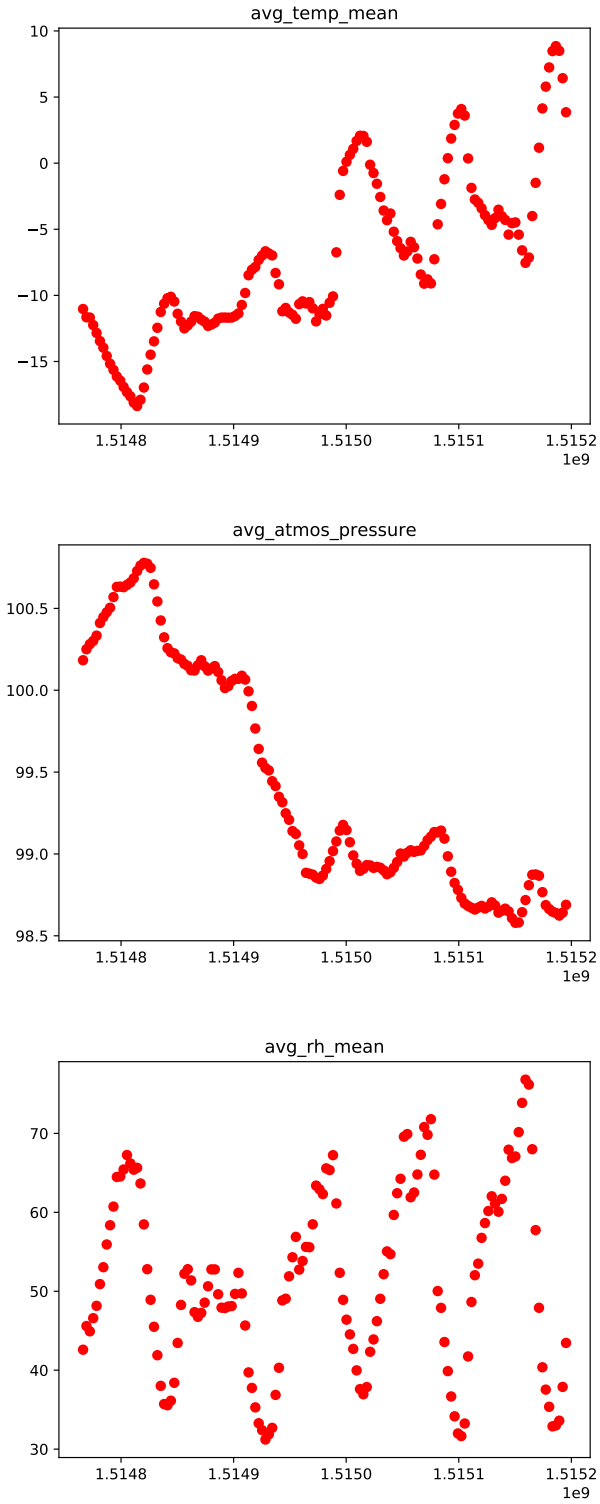

FIGURE 3. Time series of the average of mean temperature, mean relative humidity and atmospheric pressure as a function of time.

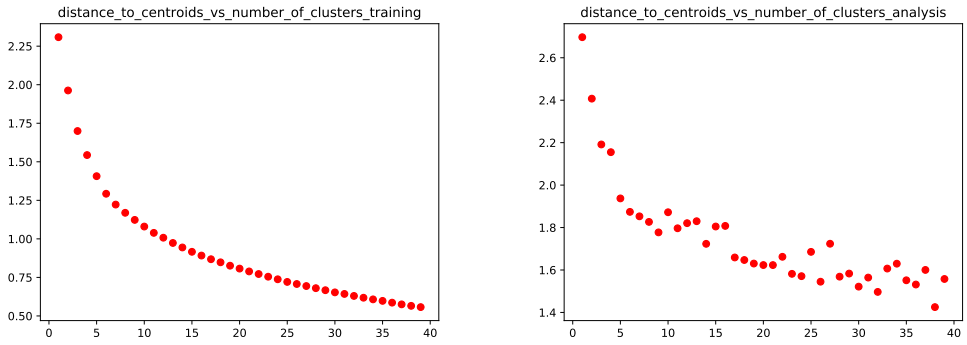

FIGURE 4. Distance to the cluster centers as a function of the number of clusters: the plot on the right corresponds to the training dataset and the right to the testing dataset.