SUPPLEMENT B

AUTOMATED PCA ANALYSIS AND ERROR DISCRIMINATION

We are going to take a specific case, in this case we analyze 10 silicon-based sensors with a size of 5m x 5m as stated in Supplement A. In this case for the first database which is the silicon-based sensor we use 10 sensors and for each one a measurement of 243 spectra in total was made, our first database we have 2430 spectrums.

A through a script developed in Matlab we create a single database with all spectrums. Figure 1-B shows the 2430 spectra divided into a group of 243 by each sensor.

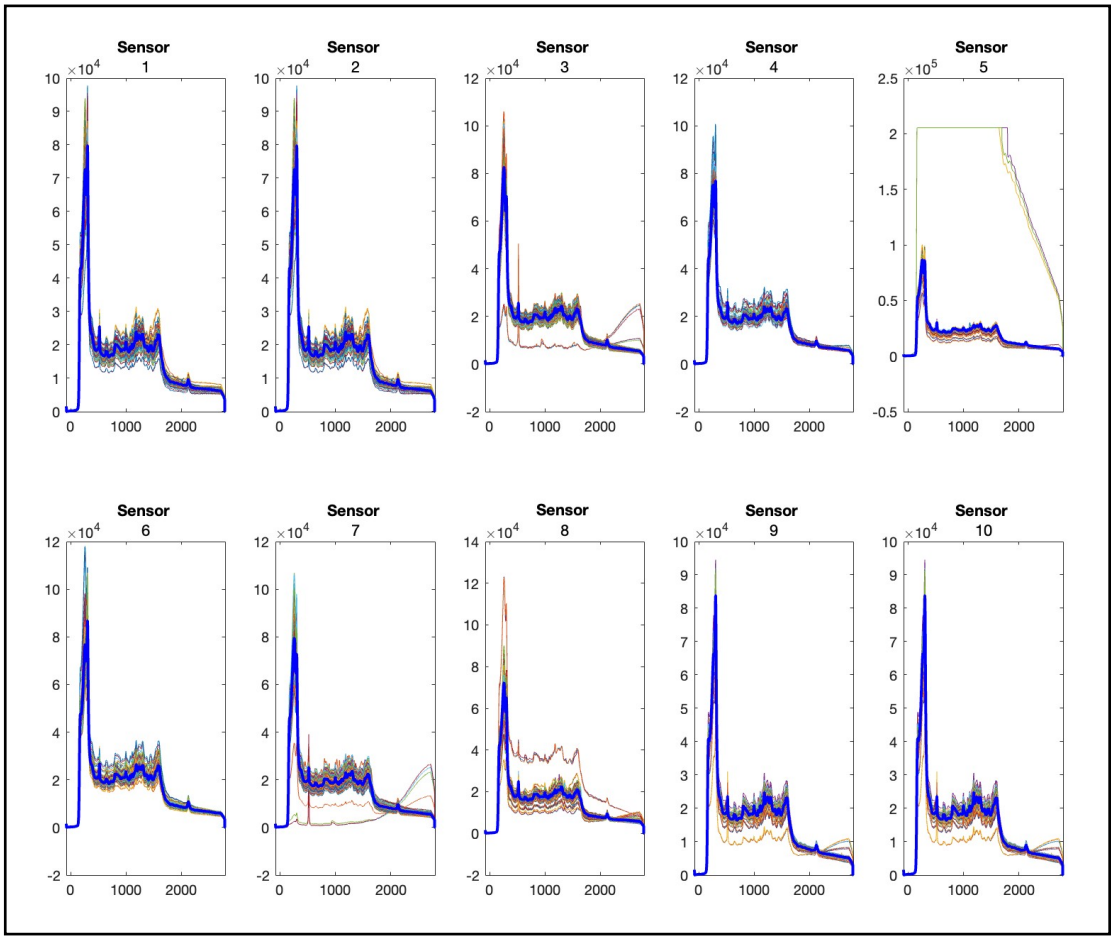

Fig 1-B. Spectra of the 10 sensors with their respective mean in blue line.

Figure 1-B shows that each sensor has certain errors or certain measurements that stand out from the original measurement or usual behaviour. This is known as instrumental error in our case error

at a specific point of the sensor that saturates our instrument; this is because the sensor at that measurement point has a scratch or an impurity, in the same way it happens in the opposite case when the signal is very low.

On the other hand, the averages of the sensors correspond to the behavior of the sensor. By normalizing the averages, it can be observed that they all correspond to the range of similar measures among all. As can be seen in figure 2-B.

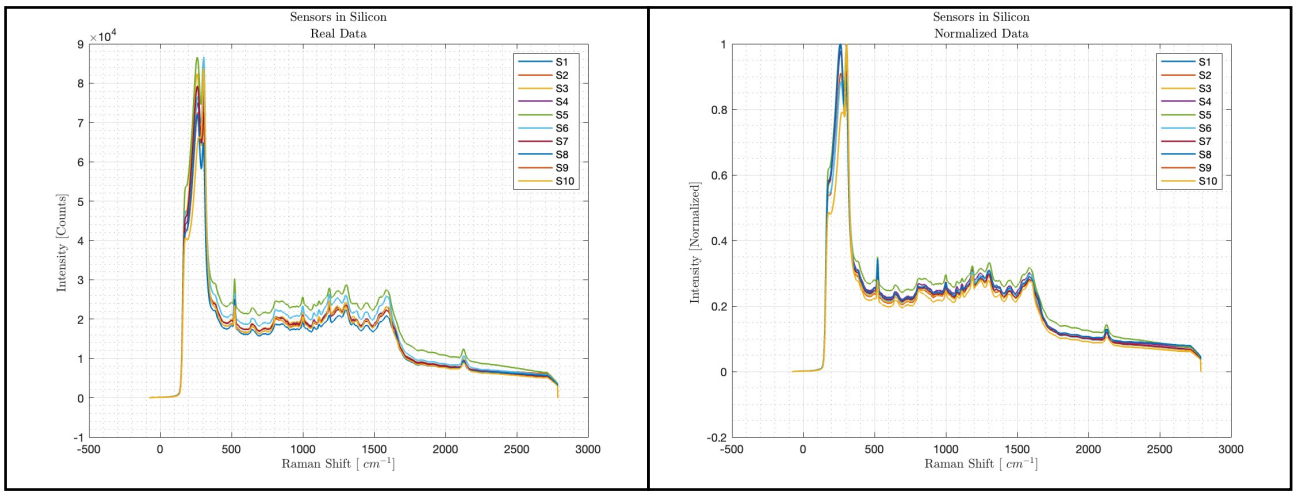

Fig 2-B. The average of the spectroscopic with the real data on the left, the normalized data on the right.

The previous figure 2-B, shows that once normalized the data have the same distribution and their intensity in the resonances are the same.

This is done in the same way with the gold sensor and a first PCA is made between the silicon and gold. As shown in Figure 3-B, we have components that leave the limits and are away from the group, so they represent the aforementioned experimental errors.

The variance explained shows that the large amount of data can be represented in the first main component with 70%, which means that the data are dispersed, so a cleaning of the 4860 spectra must be made; 2430 silicon and 2430 gold.

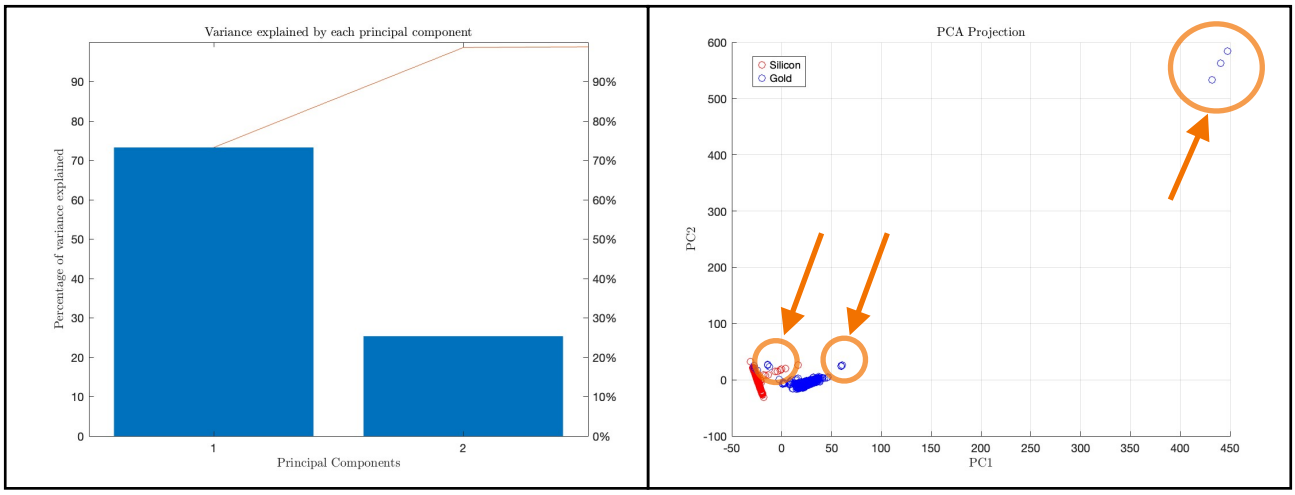

Fig 3-B. On the left side is shown the varied explained of the PCA and on the right side is shown the PCA with the dispersed data.

We use the Mahalanobis distribution which is a metric that measures the distance between a point and a multivariate distribution. Unlike the Euclidean distance, which simply measures the distance in a straight line between two points in space, the Mahalanobis distance takes into account the distribution of the data (such as variance and correlations between variables)

The Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}$$

Where:
x Is the data vector (the point you want to measure).
μ Is the average vector (the average of the multivariate distribution).
S Is the covariance matrix of the data.
S^{-1} Is the inverse of the covariance matrix.
$(x-\mu)^T$ Is the transposed vector of $(x-\mu)$

This distribution will allow to calculate the pointing of a data to a reference distribution. In our case we will occupy a maximum threshold and a minimum threshold to select our points of interest.

In figure 4-B, the distribution of the data is observed in the upper left part and the maximum and minimum limits are placed to

focus on the data that we are going to use and we are going to clean the silicon data. Which will be used as a database.

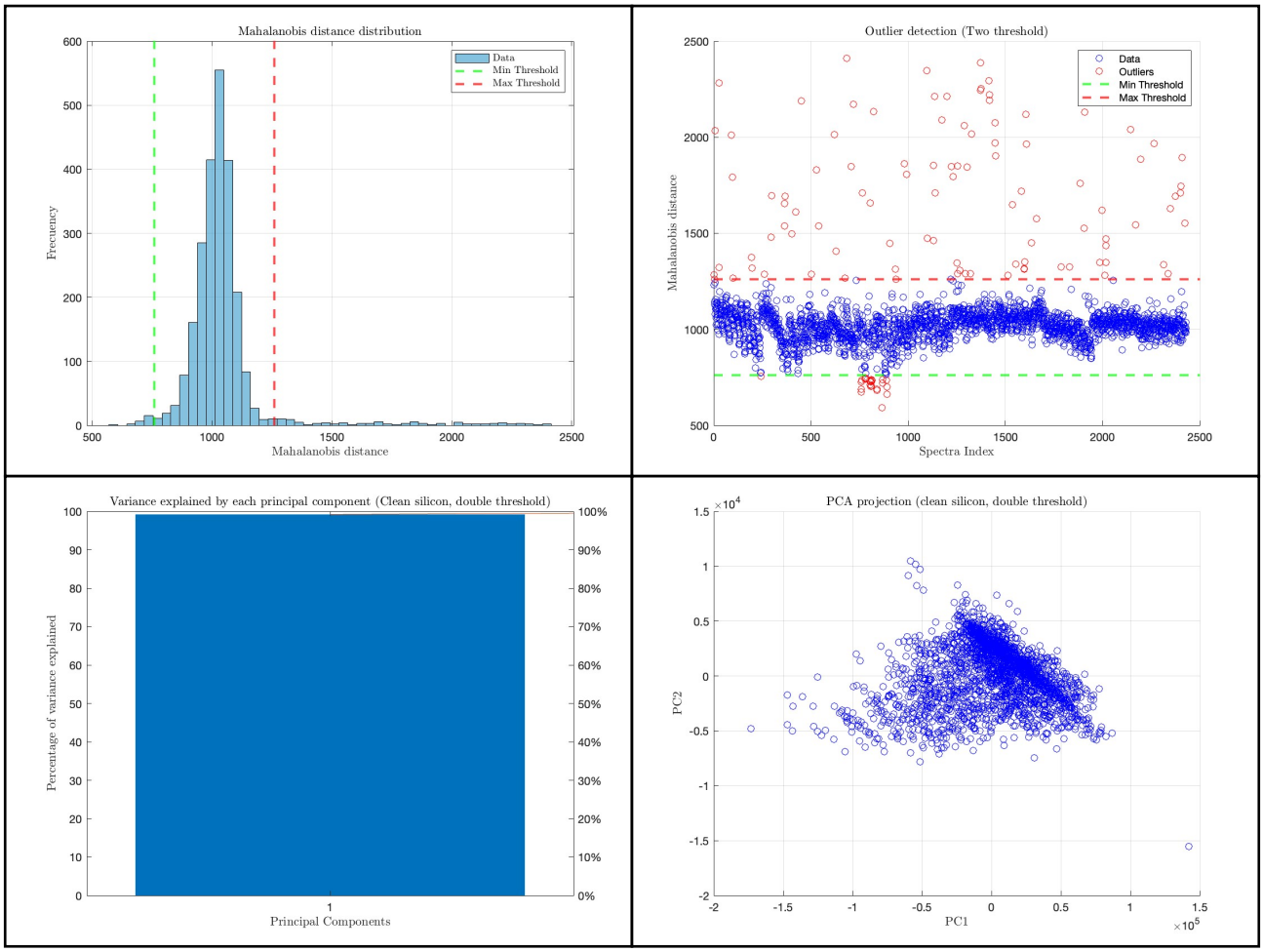

Fig 4-B. Distribution data, outliners, variance and PCA.

As seen in the figure 4-B, the PCA of silicon is concentrated and we no longer have data very separate from the concentration of population data. 128 points were eliminated outside the range [760.00, 1260.00] out of a total of 2430.

In the same way we use the same Mahalanobis technique to clean the sensor data that has the gold layer, as shown in figure 5-B, 1041 points were eliminated outside the range [1000.00, 1500.00] out of a total of 2430.

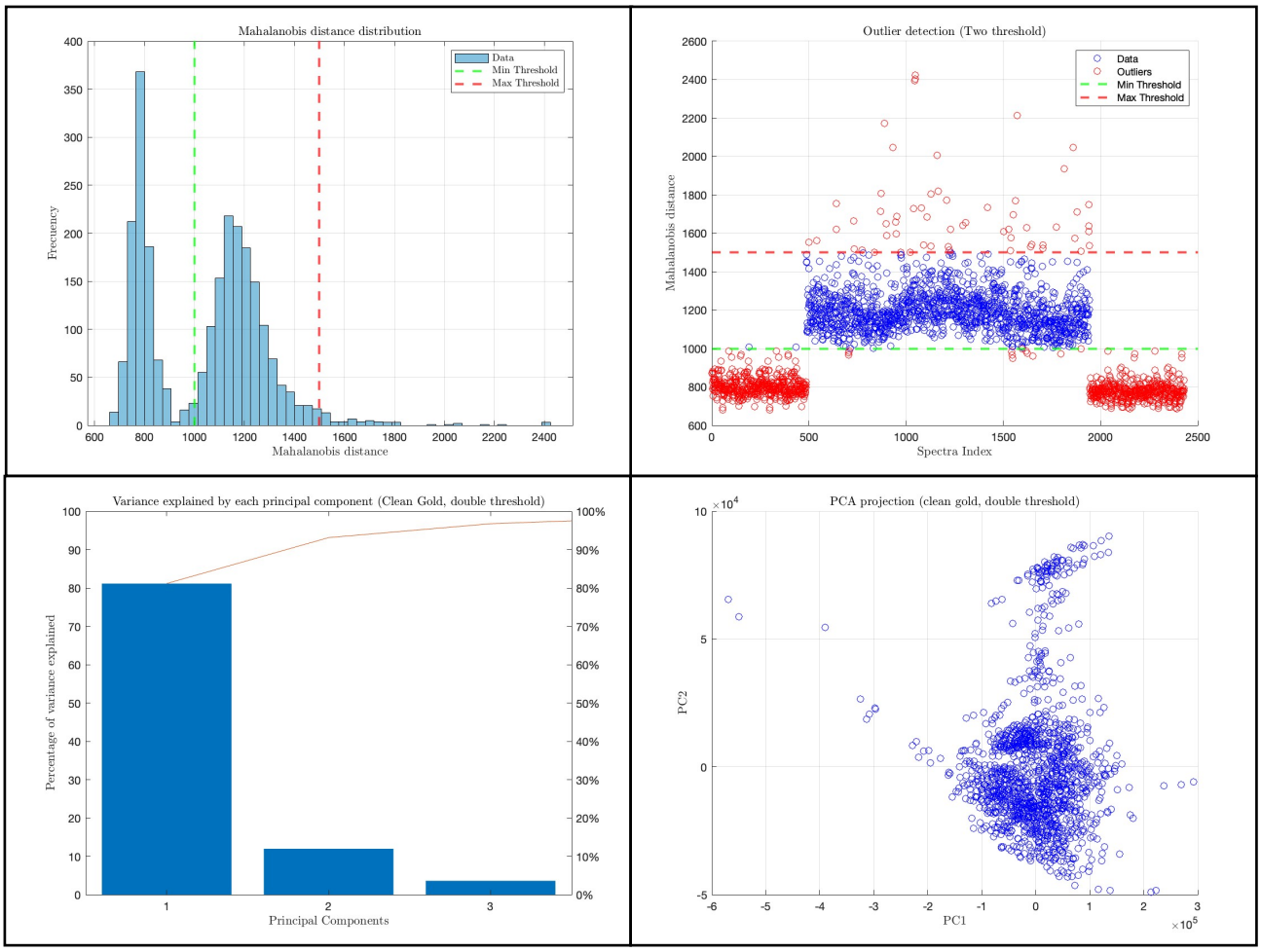

Fig 5-B. Distribution data, outliners, variance and PCA.