Projects Statistical Analysis of Networks and Systems (SANS-MIRI) Homework 2

By

Umberto Salviati

Universitat Politècnica de Catalunya

Barcelona, October 2023

Chapter 1

Visualize data

The first step I took was to analyze the data. To do this, I plotted four curves to see how the sensor data changed over time. I used PANDAS to read the csv data and handle it. In Figure 1.1, you can see the first four days plotted in four different graphs. The csv file contains the data organized in a table format, as illustrated in table 1.1

id	date	RefSt	Sensor_MLR_O3	Sensor_SVR_O3
0	10/05/2017 00:00	72.0	78.51	70.37
1	10/05/2017 01:00	60.0	66.49	59.77
2	10/05/2017 02:00	62.0	54.47	49.16
3	10/05/2017 03:00	87.0	71.58	65.94
4	10/05/2017 04:00	72.0	74.07	67.78

Table 1.1: CSV file head

The second step I take is to evaluate the quality of the in-situ calibration by computing the RMSE and R2 of the MLR and SVR estimates against the RefSt data. I perform these metrics for all the data points for each day of the in-situ calibrated sensor. The following graph shows this result. As we can observe form the graphs the best among the two sensors, before the denoising is the SVR sensor with and mean *RMSE* of 12.28 and a mean R^2 of 0.71(table 1.2).We can observe the performance and the trend of the *RMSE* and R^2 for each day in the graphs plotted in Figures 1.2 and 1.3 where all 127 days are visible.

ALL DATA POINT	RMSE	R^2
MLR	13.424	0.8764
SVR	**12.454**	**0.8936**

Table 1.2: Table with the value of the two metrics for all the set

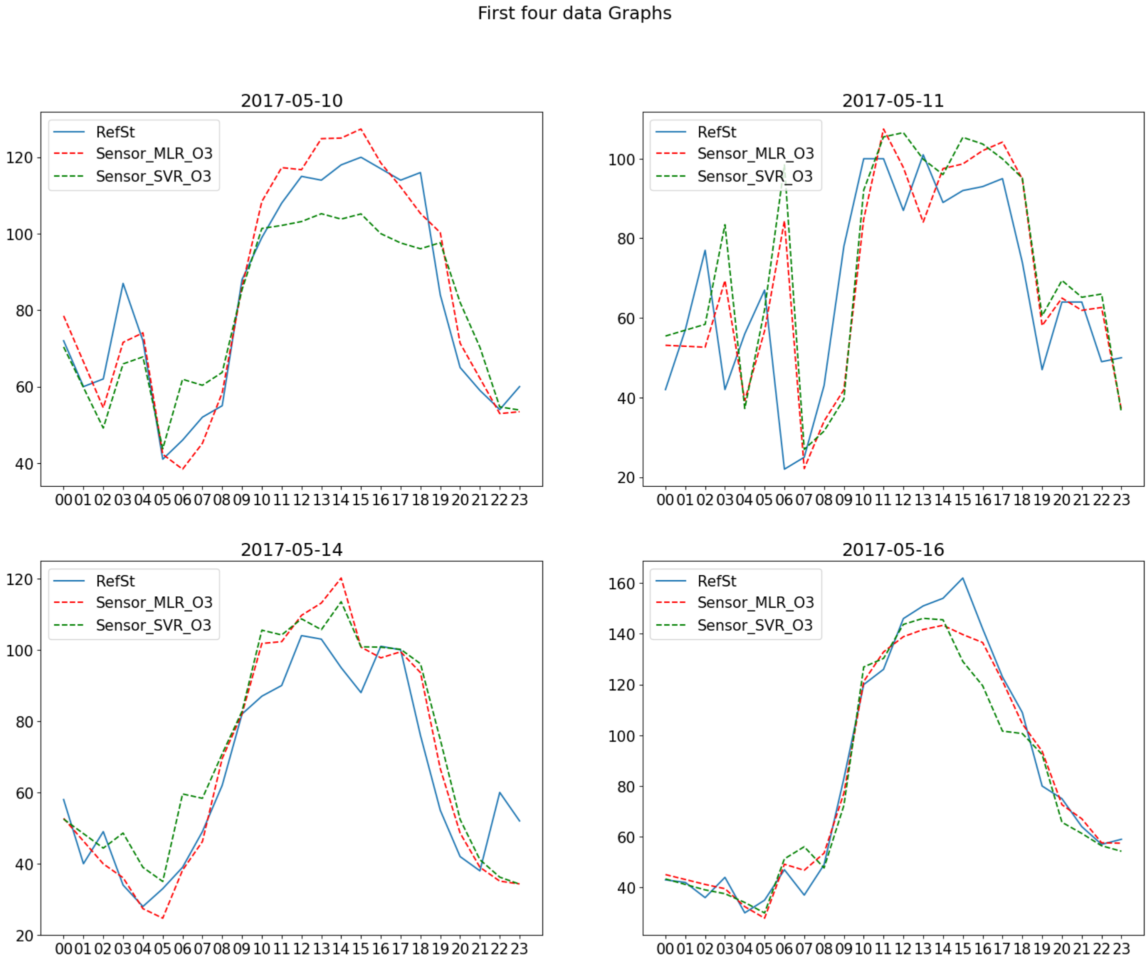

Figure 1.1: Plot of the first 4 days of the data

1.0.1 RMSE

The Root Mean Squared Error (RMSE) is calculated using the following formula(even though in my code I'm actually use the built in function of *sklearn*):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where:

- N is the number of observations or data points.
- y_i represents the observed values.
- $\hat{y}_i$ represents the predicted or estimated values.

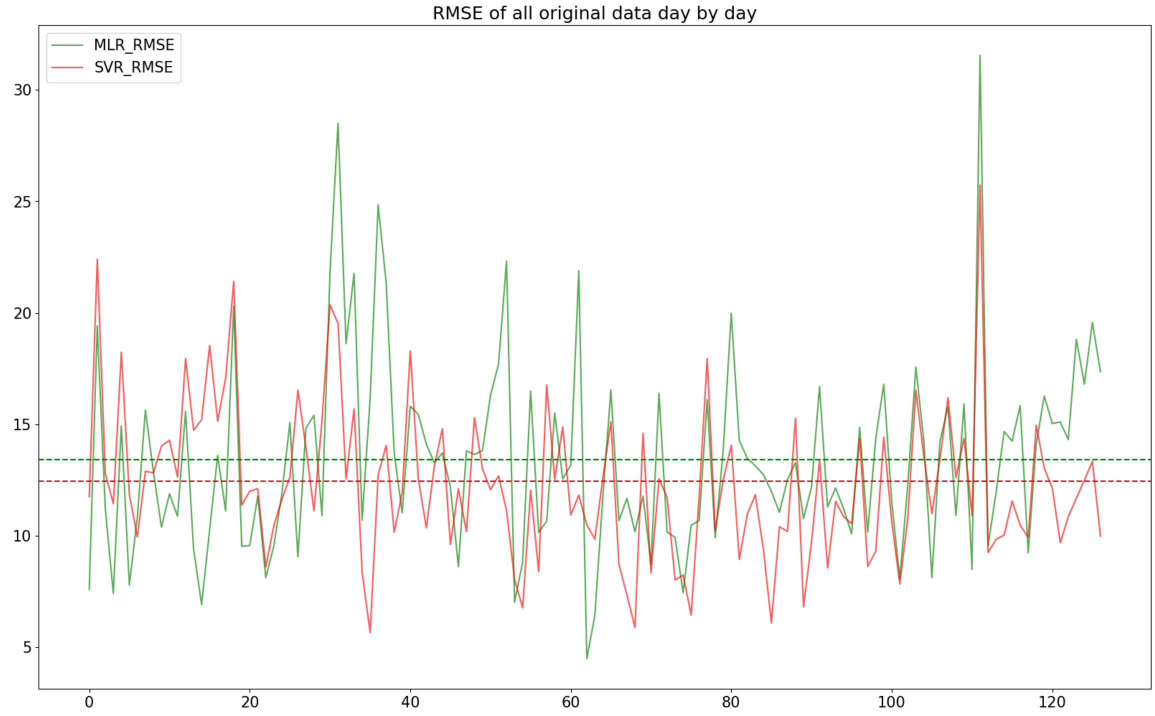

Figure 1.2: Plot of the day RMSE of all data, the orizionatal lines represents the RMSE for the enitre set

In particular the RMSE is a metric that quantifies how well a model can predict a variable. RMSE is measured in the same units as the predicted variable, which makes it easy to understand. By squaring the errors in RMSE, it gives more weight to larger errors, which can be very important in situations where large errors are more serious. RMSE is a real number that is always non-negative. It can be any real number that is zero or greater. A smaller RMSE means that the model's predictions are closer to the true values. A perfect model would have an RMSE of zero.

1.0.2 R^2

The coefficient of determination R^2 is calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where:

- n is the number of observations.
- y_i represents the observed values.

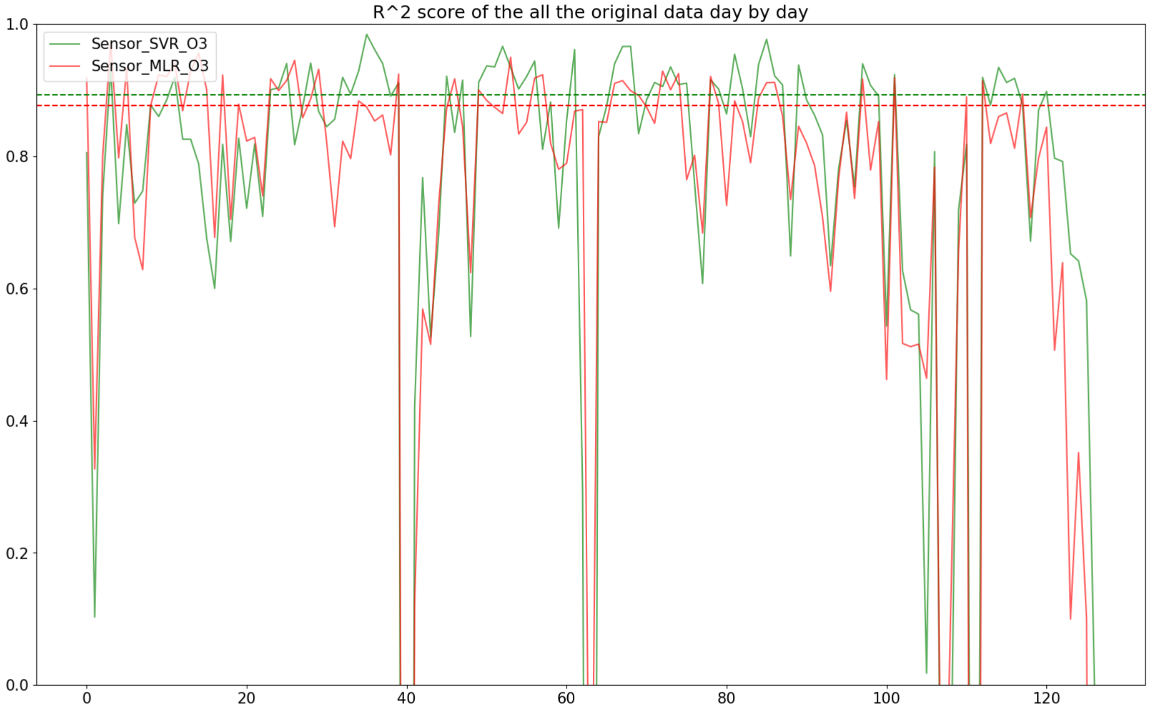

Figure 1.3: Plot of the day R^2 of all data, the orizionatal lines represents the R^2 for the enitre set

- $\hat{y}_i$ represents the predicted or estimated values from the model.
- $\bar{y}$ is the mean of the observed values.

R-squared, or R2, is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It is a scale from minus infinity to 1, where 1 indicates a perfect fit, meaning that all variability in the dependent variable is explained by the independent variables.

In simpler terms, R2 helps assess how well the independent variables in your model predict the variation in the dependent variable. A higher R2 suggests a better fit of the model to the data.

For your specific case, where R2 values are higher for data denoised by Support Vector Regression (SVR), it indicates that the SVR model is more effective at explaining the variance in your data compared to other methods or the original data. This suggests that SVR is providing a better fit to the denoised data.

1.0.3 Data Considering just the complete days

Now I am repeating the same plots for the data of complete days1.4. In this case, it is not possible to evaluate the difference in the first 4 graphs since there was no day in which the data were missing. The first day recorded with missing data is indeed the not complete `2017-05-17`. The excluded days in this new set of complete days are a total of 23, reducing the number of elements from 127 to 104.

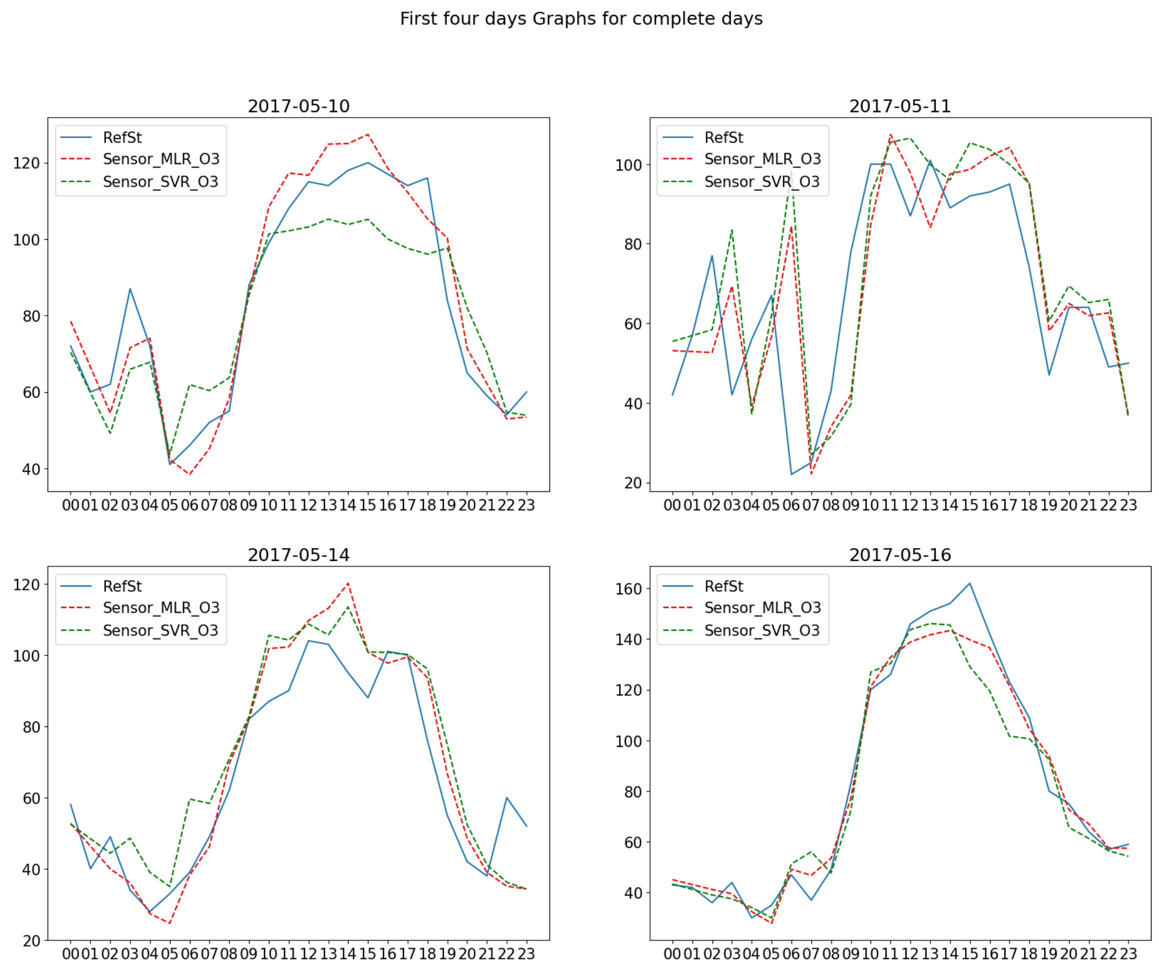

Figure 1.4: Plot of the first 4 days of the data of complete days

Now, as before, I include the two plots for RMSE1.5 and R21.6. As evident, RMSE and R2 do not change significantly. However, in the R2 graph, it is particularly noticeable the removing missing data leads to the exclusion of points that give to some day a negative value. Subsequently, total R2 and RMSE values are depicted in tables, serving as crucial indicators to verify the correctness of our denoising process.

COMPLETE DAYS	RMSE	R^2
MLR	13.424	0.7671
SVR	**12.454**	**0.7753**

Table 1.3: Table with the value of the two metrics for the complete days

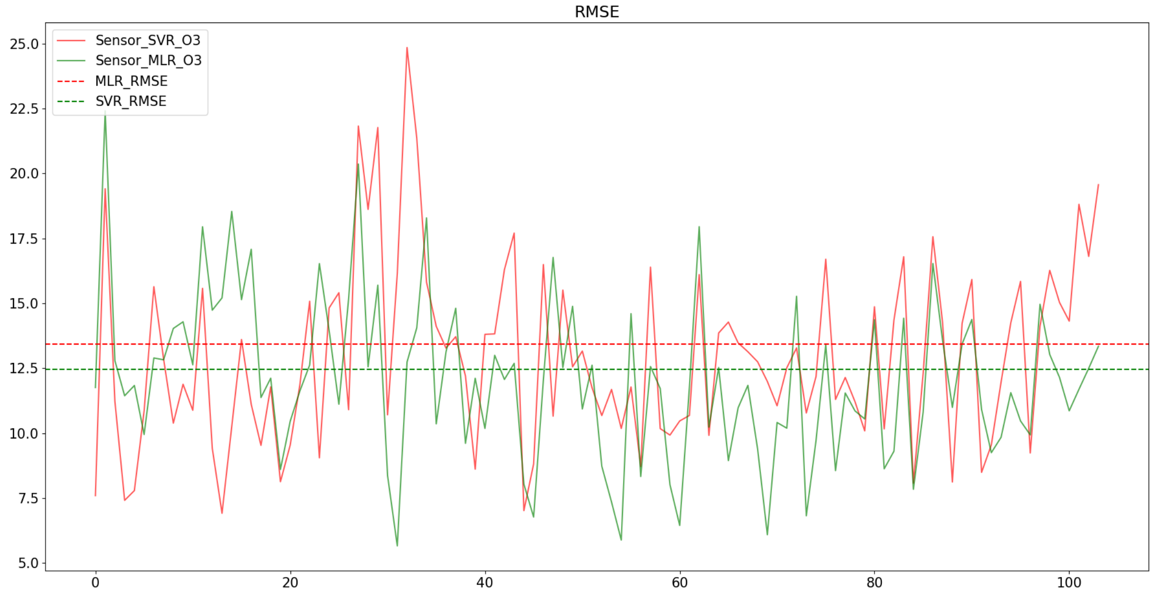

Figure 1.5: Plot of the day RMSE of all data, the orizionatal lines represents the RMSE for the complete days

1.0.4 Comparision

Comparing these two LCS in both the complete days and the entire dataset, it is evident that SVR has more effectively represented the reference data. Particularly, observing the RMSE, the value of 12.454 for SVR is lower than that of MRV (13.424).

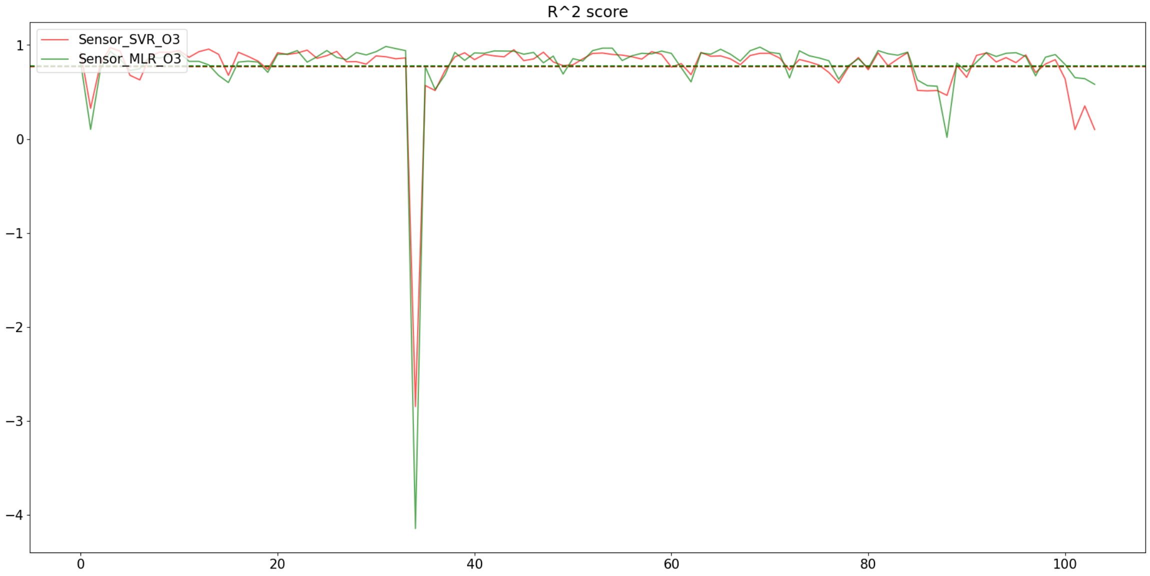

Figure 1.6: Plot of the day R^2 of all data, the orizionatal lines represents the R^2 for the the complete days

Chapter 2

Denoise

I am now applying Singular Value Decomposition (SVD) in a denoising application using our set of complete days. The reference station daily data serve as a database of reference(as in the eigenfaces proglem), and the sensor data represent the faces to be denoised. I will compute the RMSE for each possible k and select the one with the lowest RMSE.

To calculate the denoised data, I will use the formula:

$$\tilde{x}_i = \Psi + U_k \Sigma_k U_k^T \hat{x}_i$$

where $\hat{x}_i = x_i - \Psi$, Ψ is the average of reference station daily data, and U_k comes from the SVD of the reference data after subtracting the average.

This process involves obtaining the denoised representation $\tilde{x}_i$ for each data, comparing it with the original data, and selecting the k that minimizes the RMSE. This method allows for effective denoising of sensor data using SVD. Given that the Singular Value Decomposition (SVD) yields 24 singular values, the parameter k ranges from 1 to 24. Choosing $k = 1$ results in the most pronounced denoising effect, as it corresponds to the application of the Eckart–Young theorem. This theorem suggests that the best k-rank approximation of matrix X is achieved by considering the singular vectors related to the k largest singular values. In the case of $k = 24$, the opposite effect occurs compared to the previous scenario — instead of denoising, the signal would essentially remain unchanged from the original data. The concept is vividly illustrated in Figures 2.1 and 2.1 , where $k = 1$ results in a dashed line resembling the mean, indicating a significant denoising effect. Conversely, for $k = 24$, the reconstructed signal perfectly matches the original data, demonstrating no denoising and preserving the full signal integrity.

Observing the obtained plots by graphing the RMSE data on the plane with k on the x-axis and RMSE on the y-axis, one can observe a U-shaped trend where

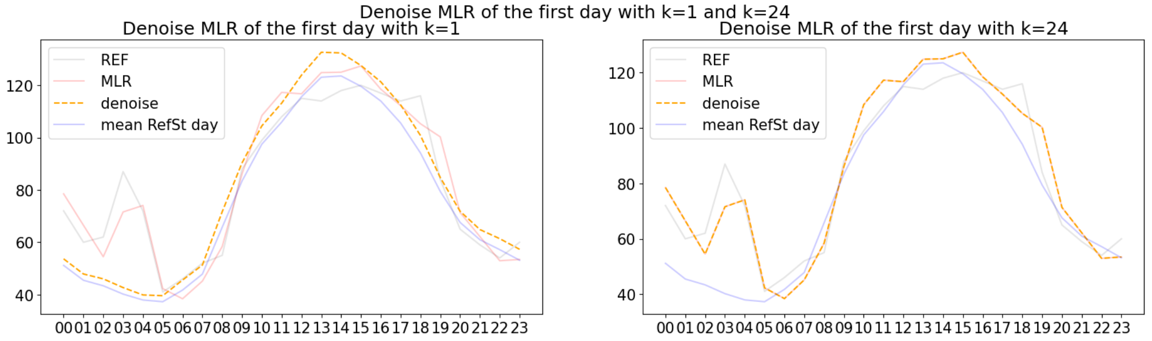

Figure 2.1: Plot of a random day of MLR data where we can see the effecto of k=1 an k=24

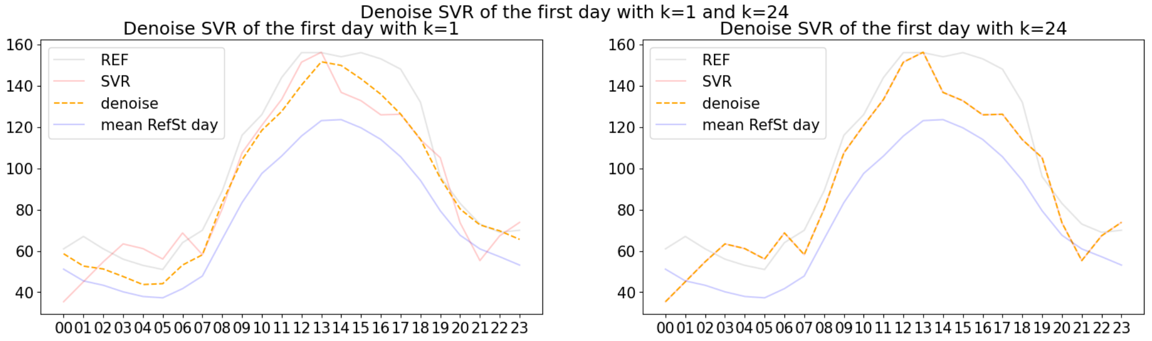

Figure 2.2: Plot of a random day of SVR data where we can see the effecto of k=1 an k=24

initially the values exceed 15. However, the tail of the graph reaches the RMSE obtained at the beginning, as described in Chapter One. This last observation further confirms that the signal reconstruction occurred correctly. In fact, with $k = 24$, as specified earlier, we have recreated the initial data, and consequently, the RMSE is exactly identical to the initial one.

As seen from the graphs, they exhibit a global minimum, particularly evident in Table 2.1, where both RMSE values are reduced compared to the initial one. Using the percentage difference formula expressed as

$$\text{Percentage Difference} = \left| \frac{A - B}{\frac{A+B}{2}} \right| \times 100$$

, we observe an improvement of approximately 3 to 9 percent. Upon closer inspection, we note that the trend of the two graphs is similar, but the SVR graph reaches the minimum earlier. This suggests that SVR requires a more substantial denoising compared to MLR. It's important to note that this difference is not pronounced, given that the MLR's k is only 2 units larger than SVR's.Once again, the data with

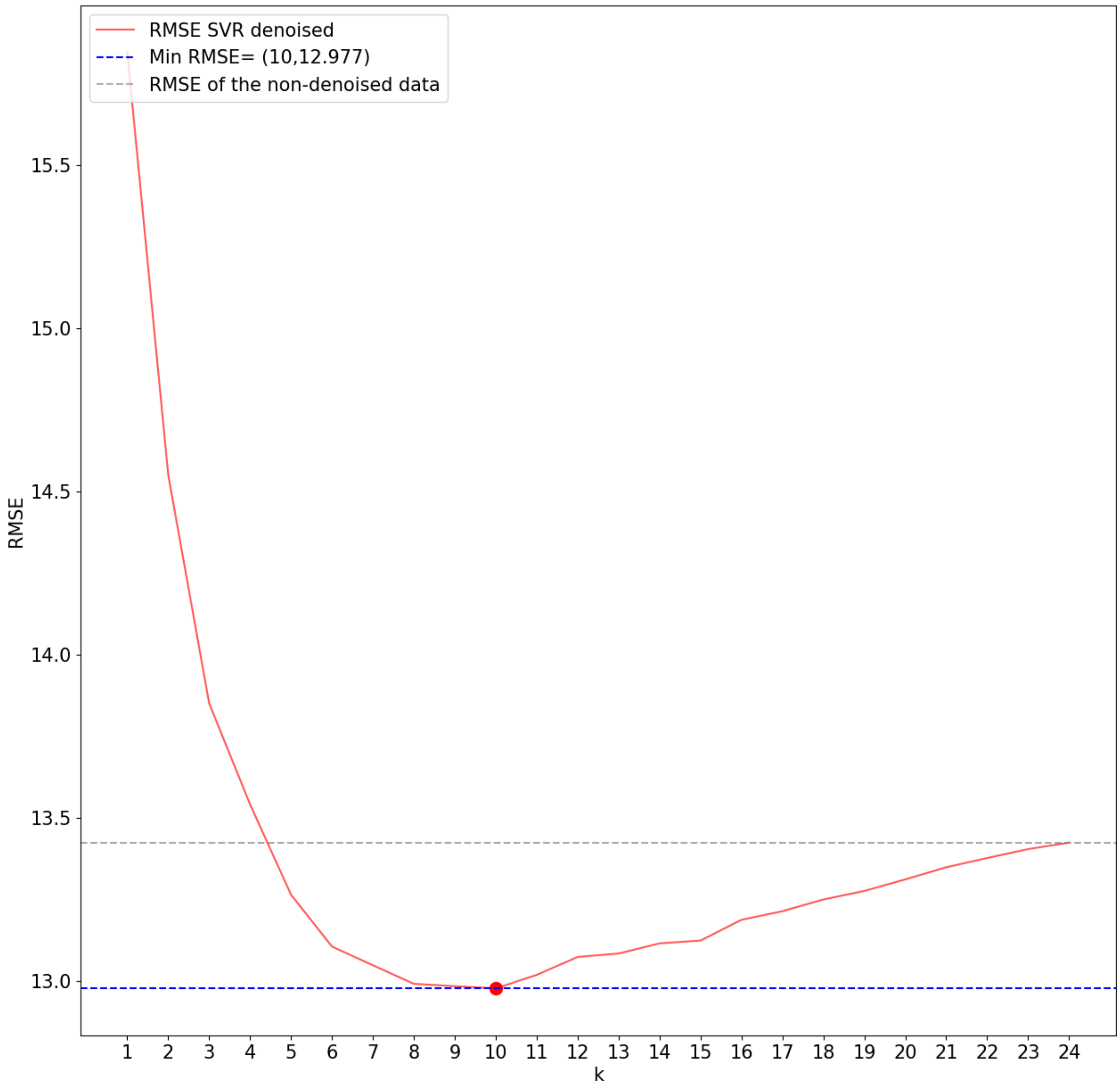

Figure 2.3: Plot of the RMSE for each k, orizontal line is the initial RMSE

SVR, featuring an RMSE of 11.39, has proven to represent the reference data more effectively compared to the other sensor.

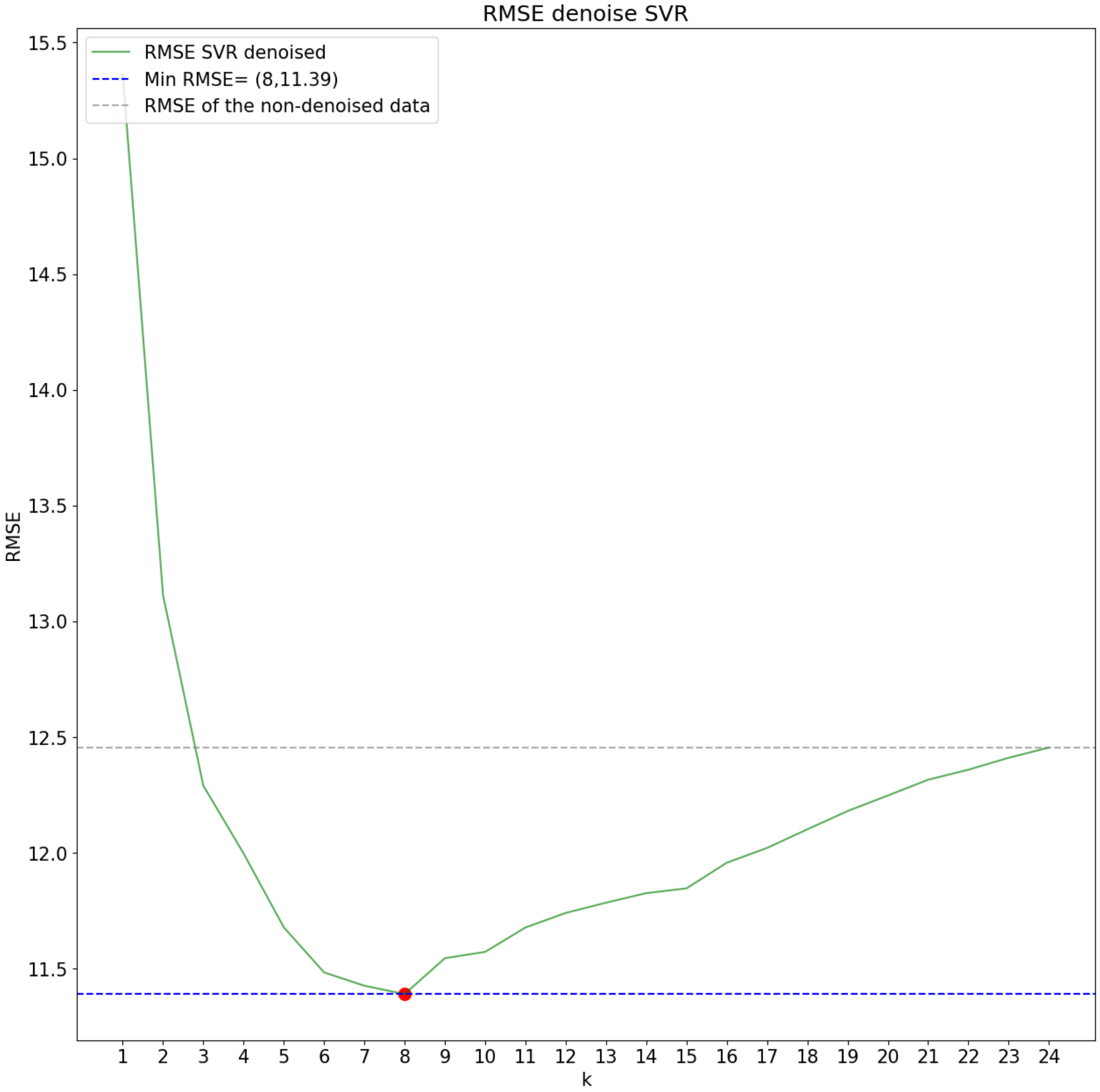

Figure 2.4: Plot of the RMSE for each k, orizontal line is the initial RMSE

	k min	RMSE min	initial RMSE	Percentage Difference %
MLR	10	12.977	13,424	3,386
SVR	8	**11.39**	12,454	8,925

Table 2.1: Table with the value of RMSE min and K min

Chapter 3

Singular Value threshold

Now, the objective is to estimate the value of κ without having access to the reference data, making it impossible to calculate the RMSE for each k. Instead, I will employ the method proposed in the Gavish et al. paper. In this method, assuming the recovery of data using the centered matrix $\tilde{Y}_R = [\tilde{y}_{R1}, \ldots, \tilde{y}_{RM}] \in \mathbb{R}^{D \times M}$, where each $\tilde{y}_{Ri}$ represents a day of reference station data, the threshold singular value is given by:

$$\tilde{\sigma} = w(\beta) \cdot \sigma_{\text{med}}$$

Here, σ_{med} is the median of the singular values, and $w(\beta)$ is determined as:

$$w(\beta) \approx 0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43$$

where $\beta = \frac{D}{M}$, with D being the dimension of daily data and M the number of days used by nearby reference stations. Finally, κ corresponds to the number of singular values greater than the threshold $\tilde{\sigma}$. The curve of singular values obtained is depicted in Figure 3.1 and Figure 3.2. Notably, following this method, I find that both κ values obtained are equal to 7, applicable to both the data initially denoised with MLR and SVR.

Given that this method involves approximating the low-rank value κ in the absence of a reference instrument, I repeat the denoising step without incorporating reference data. Subsequently, I calculate the RMSE using the κ value from Gavish et al. Now, in the table 3.1, we can observe a comparison between the initial RMSE and the RMSE obtained with the experimental κ value.

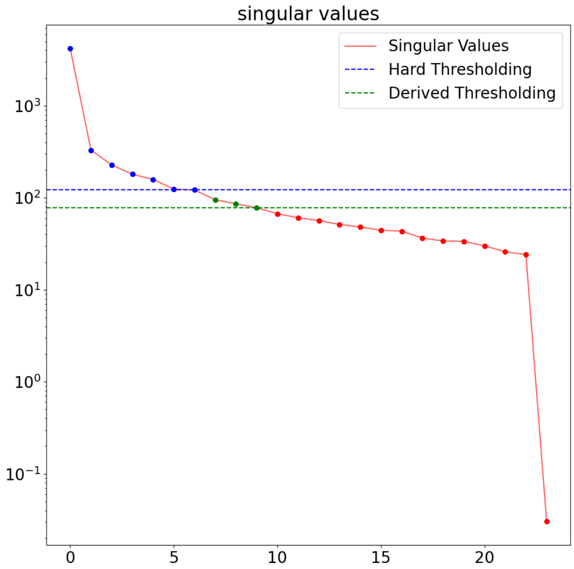

Figure 3.1: Singualar Value of MLR plot in a log scare. Hard threshold is the calculated one while the derived one is from the chapter 2

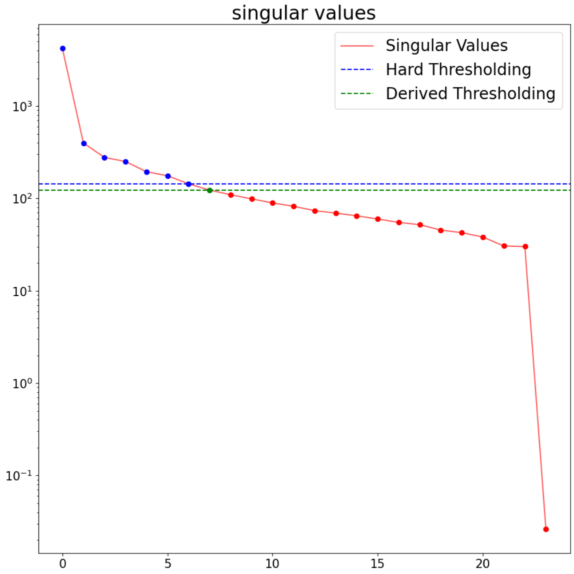

Figure 3.2: Singualar Value of SVR plot in a log scare. Hard threshold is the calculated one while the derived one is from the chapter 2

	RMSE Gavish Method	RMSE min	initial RMSE
MLR	13,464	12,977	13,424
SVR	11.956	**11,38**	12,454

Table 3.1: Result of the experiment

Chapter 4

Consluion

Finally, we can say that we have achieved the goal of this homework, as thanks to the implementation in Python, we managed to effectively denoise our data both with and without reference data. Applying SVR resulted in a final RMSE of approximately 11.956, which improved compared to the initial 12.454. Even though there was a minimal and negligible increase in MLR, we can state that Gavish's method, without the use of reference data, was still effective. The reason why MLR was not denoised effectively with Gavish's method could be attributed to the suboptimal threshold we used. Without reference data, the Gavish et al. method provides only a possible threshold, which may not be optimal.

Focusing on the comparison of the two sensors, we can confidently say that between the two, the LCS sensor is preferable when denoised with a non-linear model, namely SVR. Throughout our processes of denoising and reconstructions, SVR consistently proved to be the best at representing reference data.