Projects Statistical Analysis of Networks and Systems (SANS-MIRI) Homework 3

By

Umberto Salviati

Universitat Politècnica de Catalunya

Barcelona, December 2023

Contents

List of Figures

Chapter 1

Exploring data

I started by exploring the data to understand its characteristics and relationships. I used PANDAS to manipulate and analyze the data. I also plotted some graphs to visualize the patterns and trends. The ozone sensor operates as a voltage divider, which means it can be modeled as a variable resistor. Therefore, the first plot presented in Fig.1.1 is the ozone sensor resistance (kΩ) as a function of time, and it is compared with the reference data.

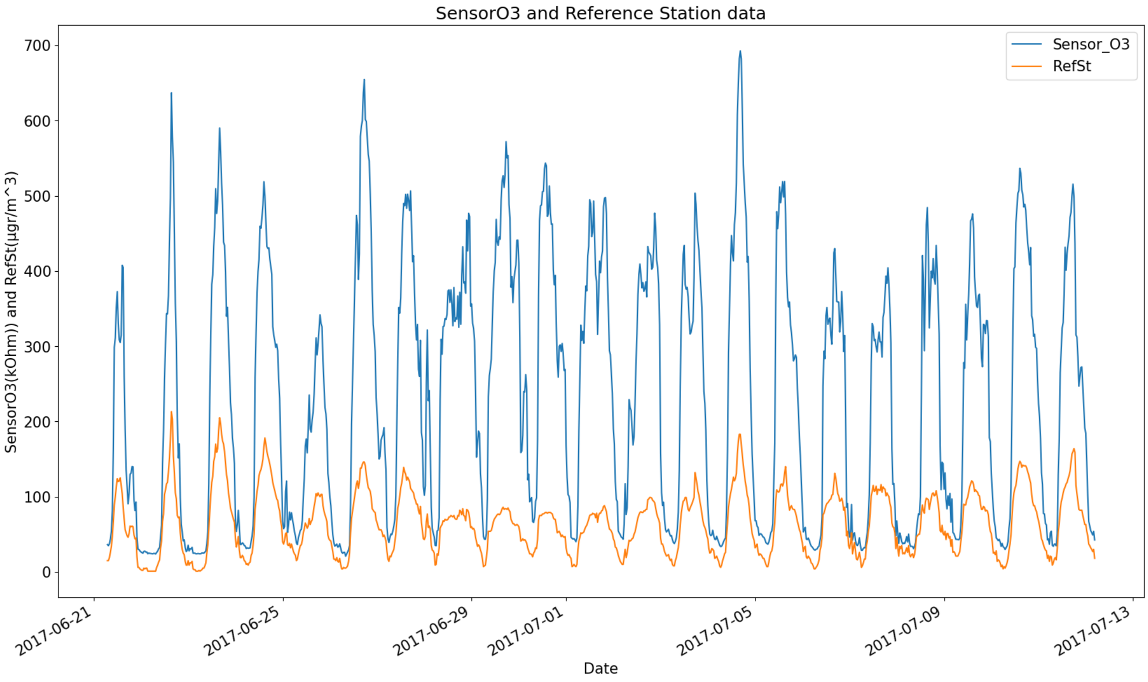


Figure 1.1: Reading of the ozone sensor and the reference station as function of time

To examine the linear relationship between the reference data and the sensor data, I create a scatter-plot with the x-axis representing the ozone sensor data in kΩ and the y-axis representing the reference data in μgr/m^3.

To compare the two signals, I normalize them by subtracting their mean and dividing by their standard deviation. I calculate the mean, μ_{sensor}, and the standard deviation, σ_{sensor}, of the training set and apply the following formula to each sample in the training set:

$$\bar{x}_{\text{sensor}_j} = \frac{x_{\text{sensor}_j} - \mu_{\text{sensor}}}{\sqrt{\sigma^2_{\text{sensor}}}}$$

where $\bar{x}_{\text{sensor}_j}$ are the normalized sensor data and $j = 1, \dots, K_1$. I verify that the normalization was done correctly by printing the mean and the standard deviation of the normalized data and confirming that they are 0 and 1, respectively.

The ideal scatter-plot would show a linear pattern with a 45-degree slope. However, due to measurement errors, the data points are scattered in a cloud around the line of 45 degrees as we can see in the scatter plot 1.2. As we can notice there is a strong positive linear correlation between the data.

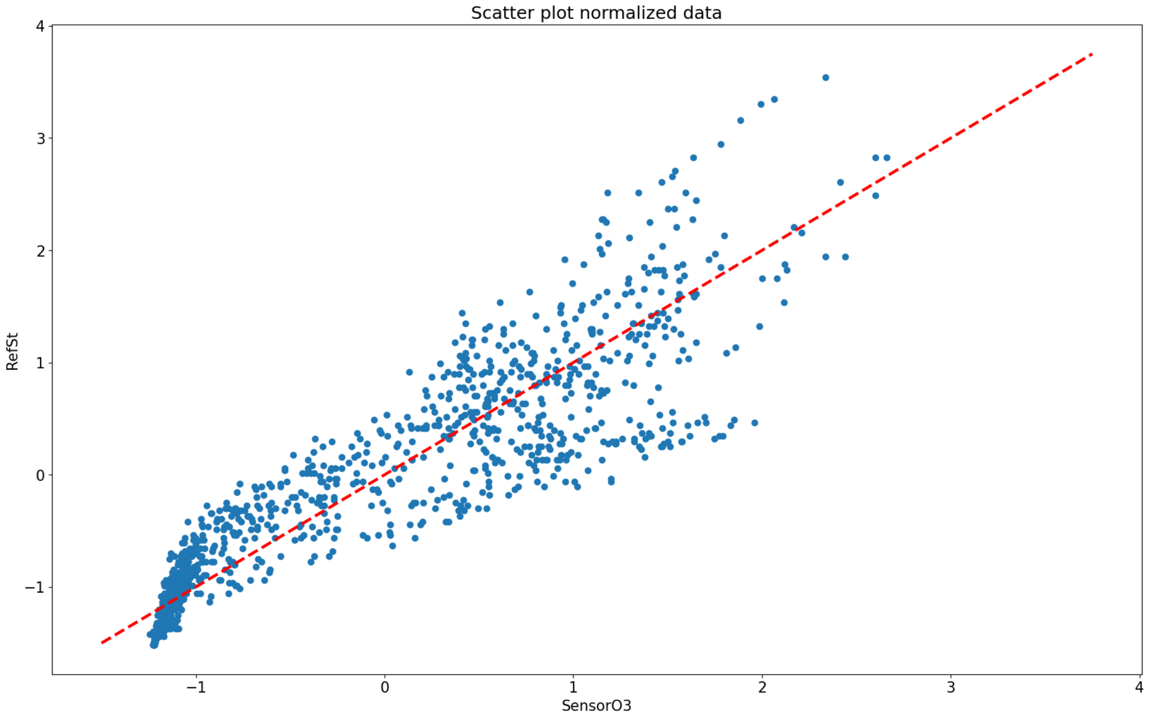
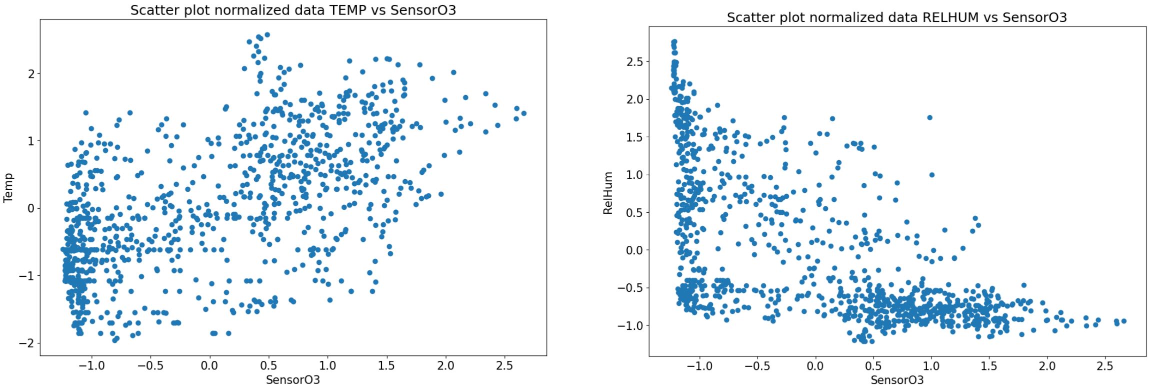


Figure 1.2: Scatter plot to see the linear correlation between sensor O3 and Reference station

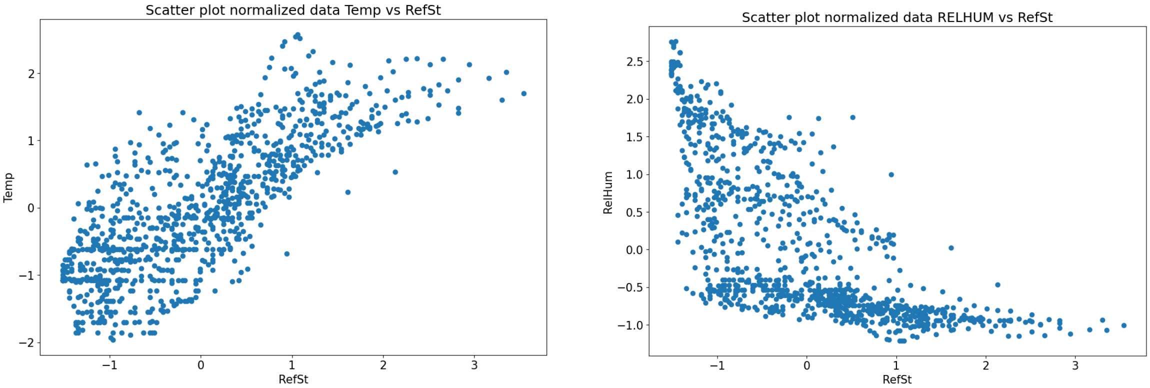
To check the correlation among the data, I plotted the scatter plot of the sensor O3 and the other features: the temperature and the relative humidity.



((a)) Scatter plot of the temperature ((b)) Scatter plot of the relative humidity

Figure 1.3: Side-by-Side scatter plots

From the two scatter plots 1.3 of the temperature and relative humidity data, there appears to be a correlation, although mild with the sensor. Upon a first look at the graphs, it seems that temperature has a linearly positive correlation. In contrast, for relative humidity, there seems to be a linear but weaker negative correlation.



((a)) Scatter plot of the temperature ((b)) Scatter plot of the relative humidity

Figure 1.4: Side-by-Side scatter plots respect to the reference values

Examining the scatter plots now, in comparison with the reference data, the same correlations are observed again, though more pronounced: thus, a positive correlation with temperature and a negative correlation with relative humidity.

Chapter 2

Sensor calibration using multiple linear regression

Now, I will calibrate the sensor using a multiple linear regression model with three features. This implies that:

$$\bar{y}_{\text{RefSt}_j} = \theta_0 + \theta_1 x_{\text{sO3}_j} + \theta_2 x_{\text{sTemp}_j} + \theta_3 x_{\text{sRH}_j} + \sigma_j^2$$

Where:

- $\bar{y}_{\text{RefSt}_j}$: Reference station reading of the obeservation at the jth time,
- x_{sO3_j}: The measurement of the O3 sensor at the jth time,
- x_{sTemp_j}: The measurement of the tempertature sensor at the jth time,
- x_{sRH_j}: The measurement of the relative humidity sensor at the jth time,
- σ_j^2: The error at the jth time,
- $\theta_0, \theta_1, \theta_2, \theta_3$: Regression coefficients.

I implemented this using scikit-learn developed in Python. I utilized the built-in function to divide and shuffle the dataset. The training set was designated as 70% of the dataset, while the test set comprised 30% of the data.

$$y^*_{\text{predicted}_j} = \hat{\theta}_0 + \hat{\theta}_1 x_{\text{sO3,predicted}_j} + \hat{\theta}_2 x_{\text{sTemp,predicted}_j} + \hat{\theta}_3 x_{\text{sRH,predicted}_j}$$

To perform the multiple linear regression, our objective is to minimize the following function:

$$\min_{\theta} \|y - \Phi(x)\theta\|^2 = \sum_{i=1}^{m} (y_i - \theta^T \phi(x_i))^2$$

The output of our implementation will consist of the comprehensive set of θ values that define the parameters for the approximated function.Where the function of the our linear regression predictor can be represented as:

$$y^*_{\text{predicted}_j} = \hat{\theta}_0 + \hat{\theta}_1 x_{\text{sO3}_j} + \hat{\theta}_2 x_{\text{sTemp}_j} + \hat{\theta}_3 x_{\text{sRH}_j}$$

Table 2.1: Evaluation Metrics for the Predictor

Metric	Training Set	Test Set
RMSE	0.8897	0.9091
R2	13.8594	12.7577

In particular, this function approximates the real data. To evaluate our predictor, we can use the RMSE and R2 metrics, as shown in Table 1. In this table, we can observe different data for the training and test sets. The metrics that are more important to look at are associated with the test data, representing the real error of our predictor, since the training set is not used to create the prediction function.

It's crucial to notice that, in our case, the test set has better values than the training set. This could be a result of the randomness in the test set, but it would be overfitting to the test set if we modify our data and procedure to make the training set lower than the test set. So, it's more likely just a random occurrence.

Now, to visualize the data more clearly, I present in Figure 2.1 the supports of the graphs of true values (reference station) and predictions. As you can see, unlike Figure 1.1, the data now represents the real data more accurately.

Finally, to check the correlation between the predicted data and the reference station data, I will plot the scatter plot of it in Figure 2.2. This plot clearly shows that the predictions are strongly positively correlated. We can observe that, among all the comparisons we have conducted to check the correspondence, this is the strongest correlation.

2.0.1 Observations

Since we trained our machine learning model using normalized data, this procedure allows us to visualize the coefficients and compare them.

From the bar graph 2.3, we can observe that, as previously noted, the data do not influence the model in the same way. It's interesting to note how obviously the sensor O3 is the most influential, but also the temperature still contributes to the

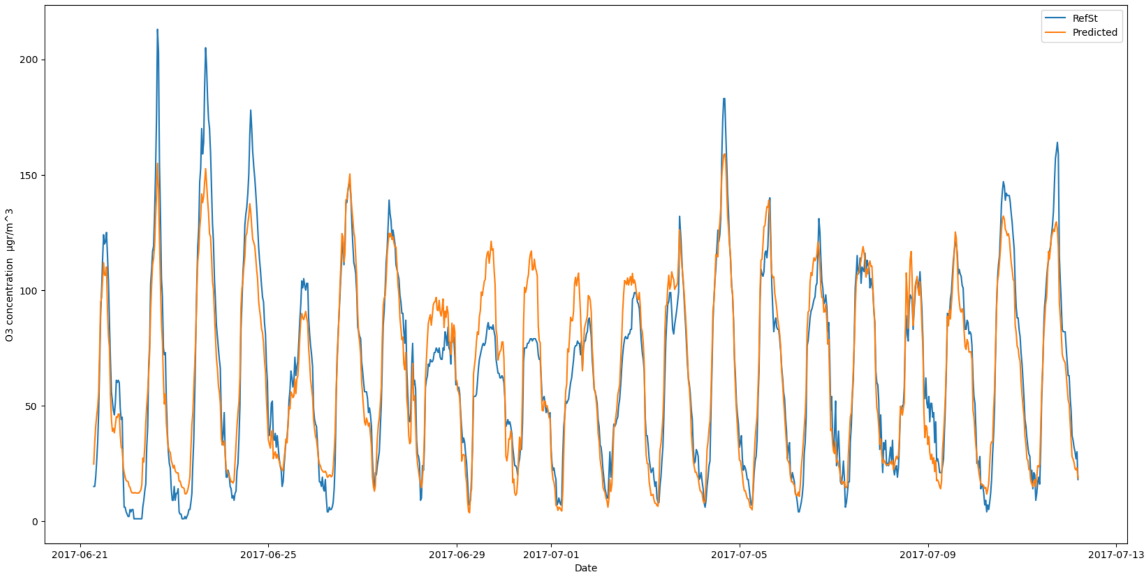


Figure 2.1: Comparison of True Values and Predictions

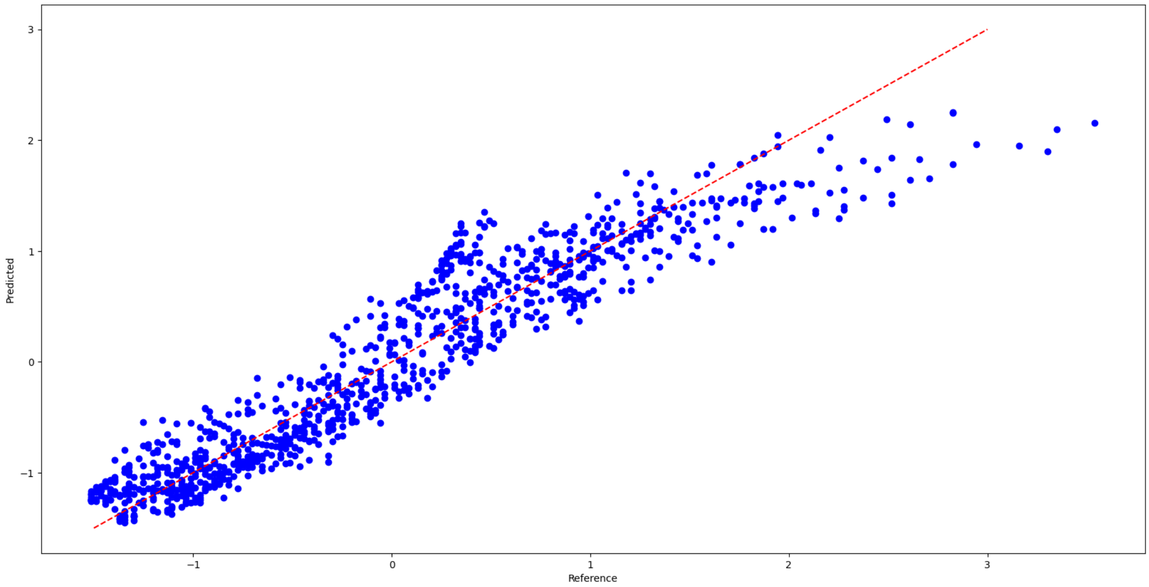


Figure 2.2: Scatter plot of Predicted Data vs. Reference Station Data

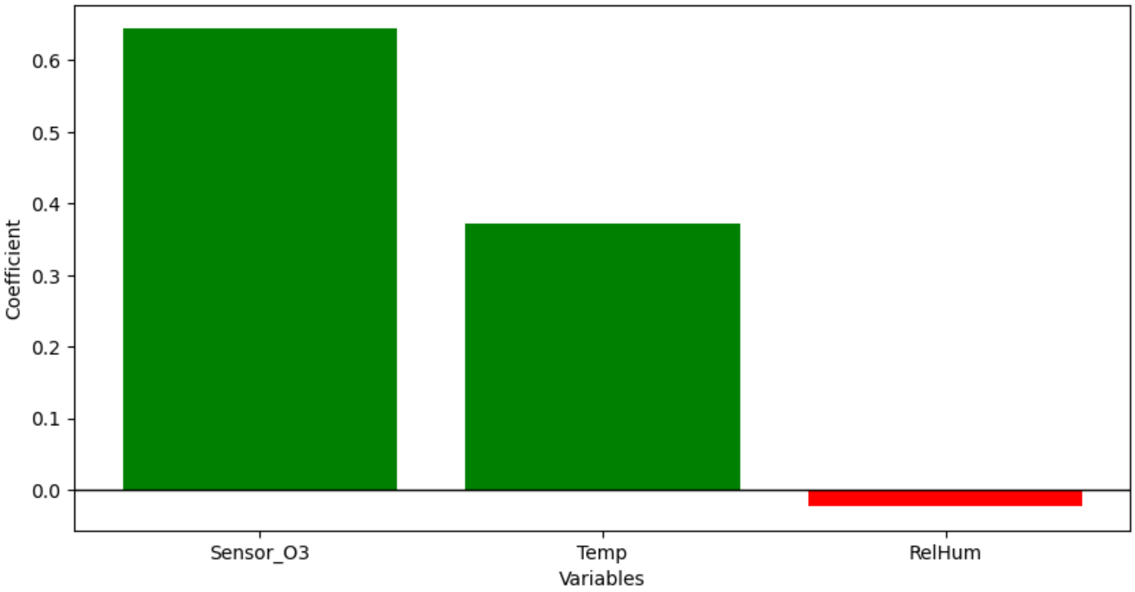


Figure 2.3: Bar Plot of Coefficients

prediction. On the other hand, relative humidity, as shown in the scatter plot, has little to no influence on the prediction. We notice that even if the influence is very little, it is negative, as we stated in the previous observation of the scatter plot.

Chapter 3

Conclusion

In conclusion, we can say that the objective of the homework was archived since we applied the theory studied in class. Utilizing Python for implementation, we thoroughly investigated the correlation among the dataset, enhancing our comprehension of which data exhibits a stronger correlation with the reference station data. The development of a predictor through linear regression proved to be effective, providing an accurate representation of the dataset. Notably, the resulting function achieved an impressive R^2 value surpassing 0.9 and an RMSE of 12.57 $\mu g/m^3$.