Calibration of sensors

Giacomini Nicolò

December 27, 2023

Abstract

In this paper I describe the procedure to obtain the best results by cheap sensors in order that they are as close as possible to the results of the O_3 in the air measured by the reference stations. The sensors measure the value of the the O_3 in the air but also the temperature and the humidity in the air. All this values can be used in combination in order to get the most accurate value of O_3 in the air.

Studies show that both temperature and humidity affect on the amount of ozone (O_3) is in the air, but they do it differently. High temperatures are correlated to an high value of ozone in the air, and with an high effect. On the other hand, humidity also influences the amount of O_3 in the air, although the impact is not relevant as much as the temperature effect.

In the following experiment I will use the multi linear regression in order to get the best results. In the experiment I will build a model using the linear regression. I will take the data from the reference station and the sensor dataset in order to obtain the training set and the testing set. First I will train the model with the training set and then I will test the results. The main goal is to obtain approximately the same signal of the reference dataset starting from the sensor dataset.

1 Introduction

1.1 First evaluation of the datasets

In this experiment, initially, I have two dataset:

- Reference station dataset: the data of this dataset are the reference for the experiment. The unit of measure of the dataset is $\mu gr/m^3$.
- Sensor dataset: these data come from the sensors (whose cost is much lower than that of the reference stations) and they contain:
  - O_3 data: the dataset of the ozone in the air measured in $k\omega$.
  - Temperature: the dataset of the temperature measured in °C.
  - Humidity: the dataset of the humidity in the air measured by %.

The first part is to understand and plot the data. In the following plot we can see the reference dataset:

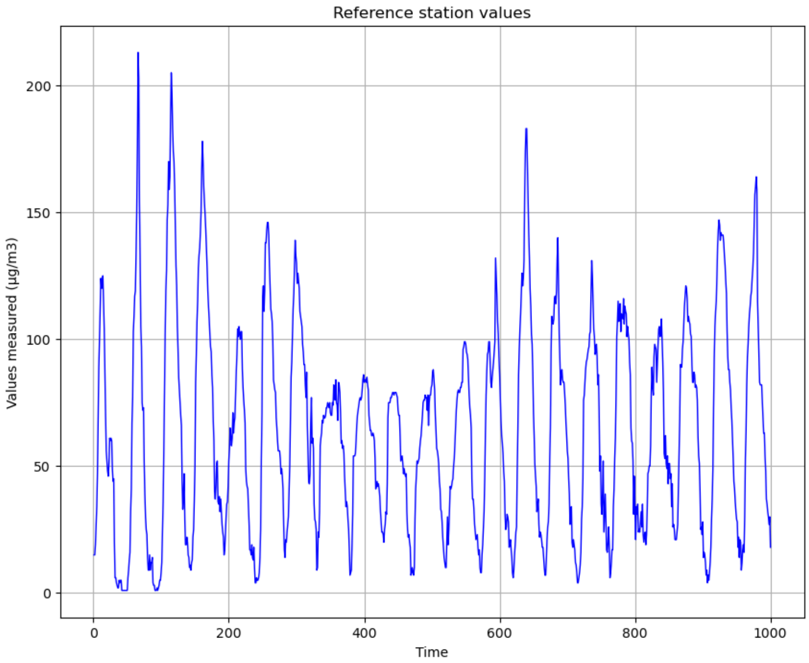

Figure 1: Reference dataset

It is possible to notice that the values of the O_3 show a periodic oscillation pattern over time. This could mean that the quality of air improves during the night time, while during the day, the quality decreases.

In the following plot we can see the O_3 dataset of the sensors:

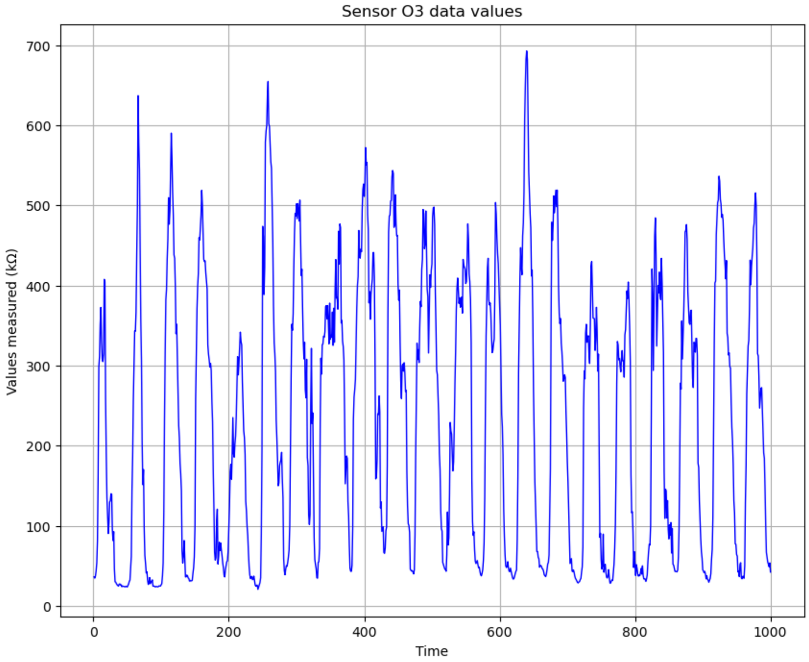

Figure 2: O_3 dataset of the sensors

Also in this case the values follow a periodical pattern, for the same reason explained before. It is important notice that the scale is completely different from the reference station, because the units of measure are different. In the following plot we can see clearly the difference between the two datasets:

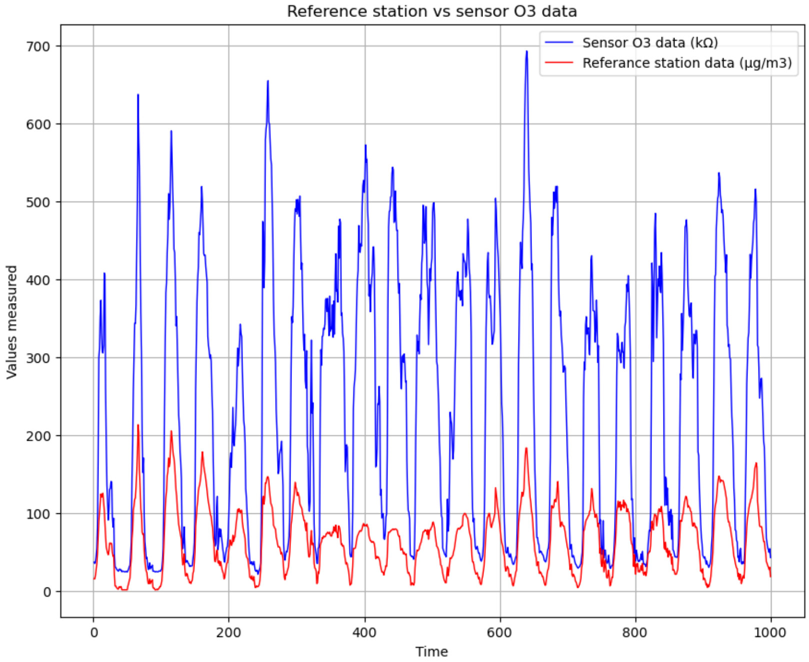

Figure 3: Comparison of the datasets

From this plot I get the following value for the RMSE and R^2 indexes:

- $RMSE = 217.8177$
- $R^2 = -25.9870$

This values are very bad, but they are useful to understand how the final model can improve the results.

In order to compare better the dataset, I normalize both dataset using this formula:

$$\bar{x}_{value_i} = \frac{x_{value_i} - \mu_{values}}{\sqrt{\sigma^2_{values}}} \tag{1}$$

This formula allows to get a dataset with mean equal to 0 and a variance equal to 1. Thanks to the normalization it is possible to compare dataset that using different scales.

In the next plot is possible to see how the data are following the same function:

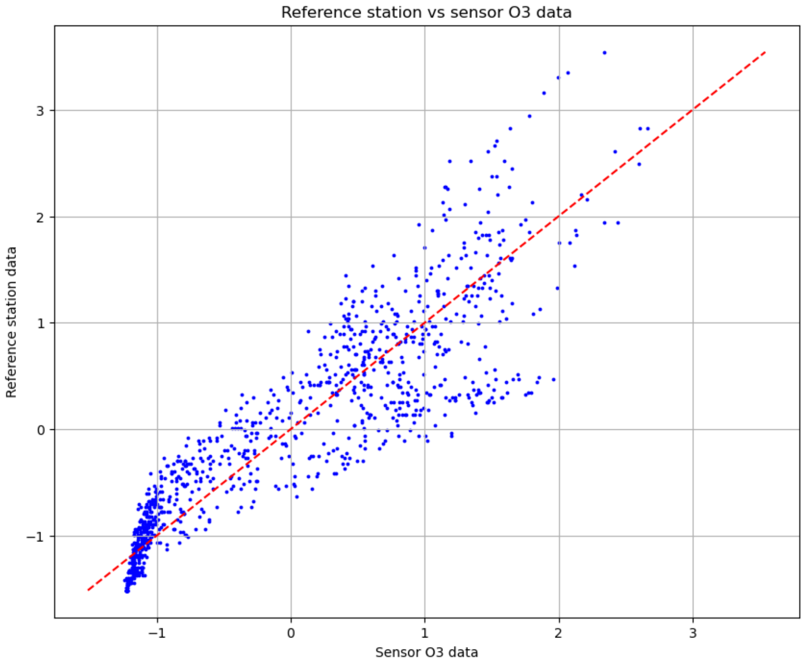

Figure 4: Scattering plot. The x-axis represents the sensor values, while the y-axis represents the reference station values. The red line represents the function y=x, i.e. where the values of the sensors are equal to the value of the reference stations

By the scatter plot it is possible to notice that the values are following more or less the same pattern, but more the value are high more the difference increases. The main goal of the linear regression will be to make the difference between the values obtained by the sensors and the values of the reference stations, smaller as possible.

1.2 Comparing with the temperature and humidity

In order to check how the temperature and humidity impact on the measure of the O_3, I normalize the dataset of the temperature and humidity and I compare the dataset with the reference stations dataset and the O_3 values of the sensors, both normalized. For the normalization of the temperature and humidity I use the formula 1.

Temperature comparison In the following plot, it is possible to notice the comparison between the values of the O_3 of the sensors and the values of the temperature measured by the sensors.

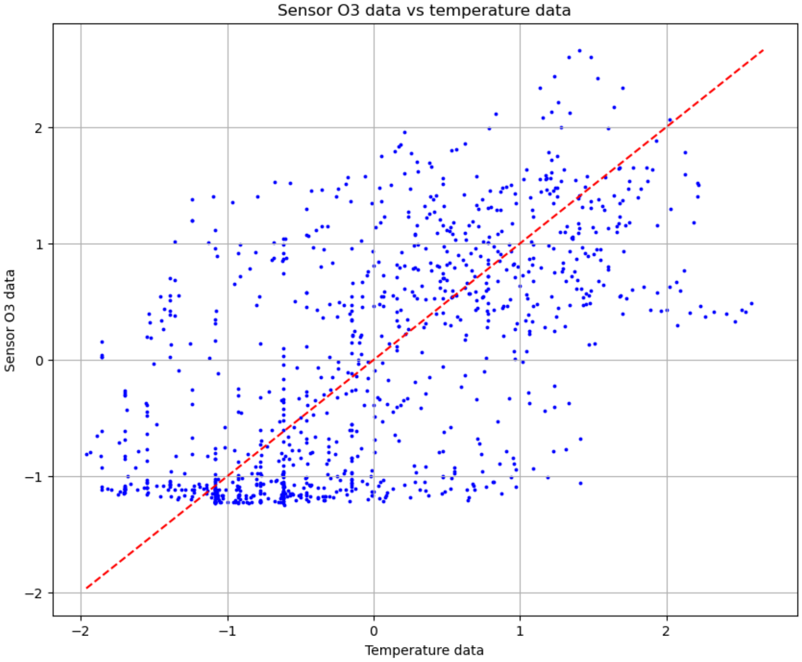

Figure 5: Comparison of the datasets

By this plot is possible to notice that there is not so much correlation, but the cloud of dots is in the center of the graph. We cannot draw any conclusions by the plot.

In the following plot I show the graph obtained by the comparison between the reference dataset and the temperature measured by the sensors:

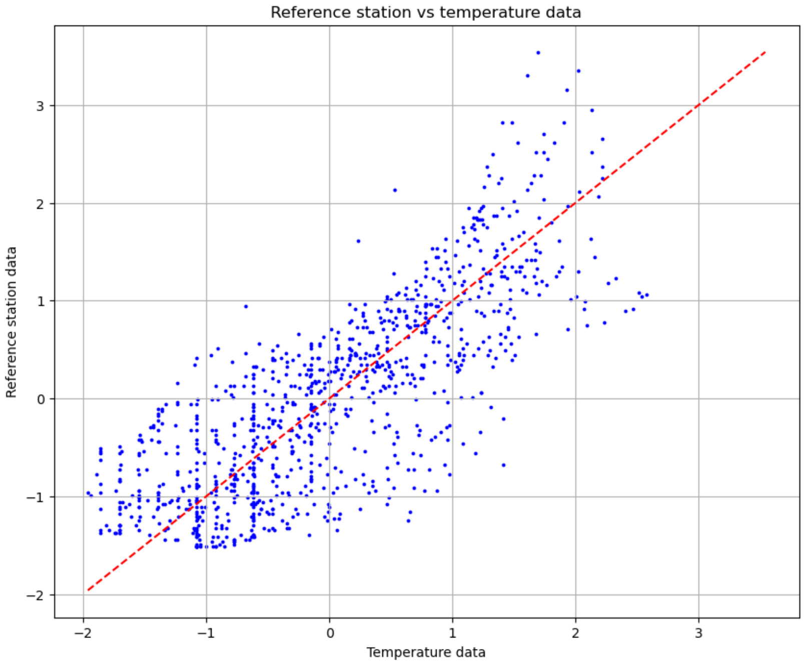

Figure 6: Comparison of the datasets

By the plot, we can see that the temperature has more or less the same trend. The cloud of dots surround the y=x function, however the variance is very high.

Humidity comparison In the following plot I show the relation between the sensor dataset (O_3) and the humidity:

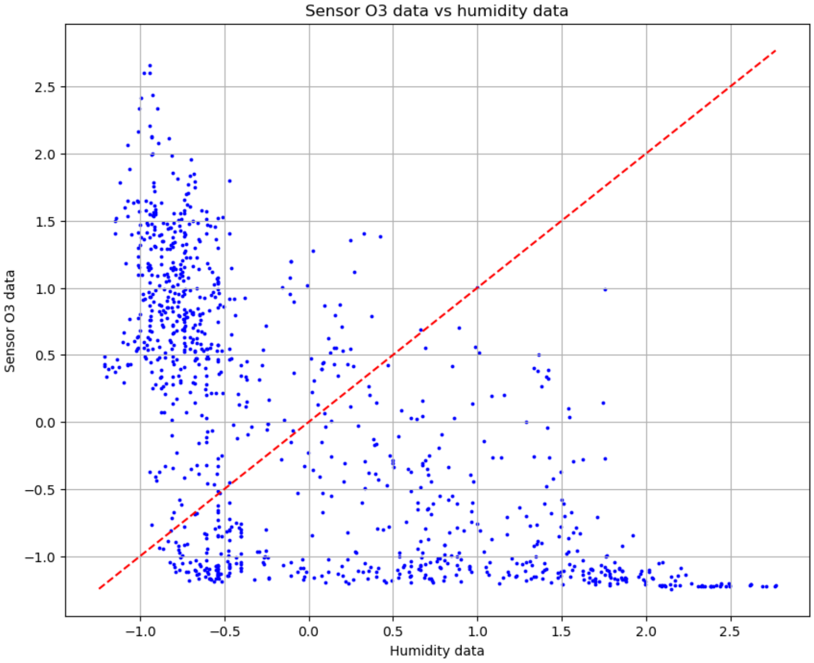

Figure 7: Comparison of the datasets

By the scatter plot is possible to claim that the values are very uncorrelated, in fact most of the dots are on the edge of the graph.

In the next plot, we can see the relation between the reference dataset and the humidity dataset measured by the sensors:

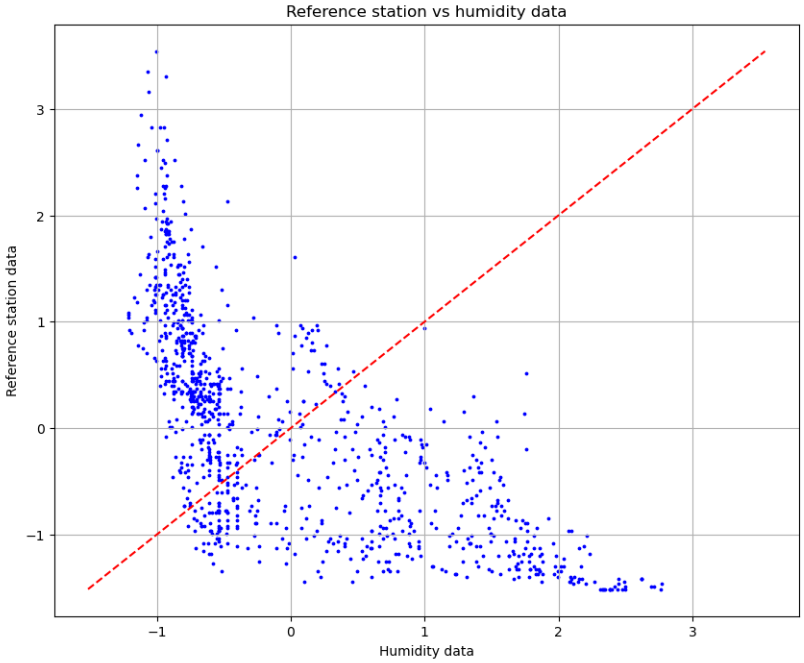

Figure 8: Comparison of the datasets

In this plot the result is similar to the previous graph. Most of the dots are very far from the line of the y=x function.

2 Multi linear regression

After to analyse the different dataset that we have available, we can use the dataset of the sensors in order to get the best prediction of the reference station dataset. In order to do this I will use a multi linear regression model, as follows:

$$\bar{y}_{RefSt_i} = \theta_0 + \theta_1 x_{O3_i} + \theta_2 x_{SensTemp_i} + \theta_3 x_{SensHum_i} + \epsilon_0 \tag{2}$$

where the $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$.

Multi linear regression is a statistical method used to analyze the relationship between multiple independent variables and a single dependent variable. In multi linear regression many predictors are considered from the model. The main goal is to find the best linear equation that represents the reference dataset from the sensors dataset. The model in multi linear regression is composed by different parameters θ that change the value based on the relevance that every component (O_3, temperature and humidity) has.

For build the model, first of all, I need to define four dataset:

- X training set: it is the dataset that allows to train the model. The data are taken randomly from the sensors data. For the modelling I use the 70% of the total dataset.
- X test set: it is the dataset that allows to test the model after the training. The data are taken randomly from the sensors data. I use the remaining 30% of the whole dataset.

- Y training set: it is the dataset used to the training. Every time the model test a new value using the X training set, then it is compared to the Y training set in order to check if the result is close or not. If not the model is correct and proceed with the next value to test. Also in this case, it is used the 70%. The data are taken randomly from the reference stations data.
- Y test set: it is used to validate the model in the testing phase. The data are taken randomly from the reference station data. These data represents the 30% of the whole dataset.

Notice that every value taken from the dataset is picked randomly in order to train and test better the model, and the datasets that I used, contain the normalized data.

2.1 Model and results

After the fitting of the model I get the following results for the coefficients θ:

- $\theta_0 = 64.3491$
- $\theta_1 = 26.7850$
- $\theta_2 = 16.0864$
- $\theta_3 = -0.7932$

Notice that θ_0 represents the intersection with the y-axis, when all the values are equal to zero, θ_1 (i.e. the coefficient of the O_3 sensor) is the component with the greatest relevance in the model, θ_2 value (i.e. the coefficient of the temperature) indicates that also the temperature has a main role in the model, however the θ_3 (i.e. the coefficient of the humidity) is negative and very small. The latter value indicates that the humidity has not a very high significance in the model, but, on the contrary, it goes against the trend of the model. We could predict this result by the plots in figure 7 and 8, where the plots are very uncorrelated with the O_3 values.

Here I show the final model compared with the reference dataset:

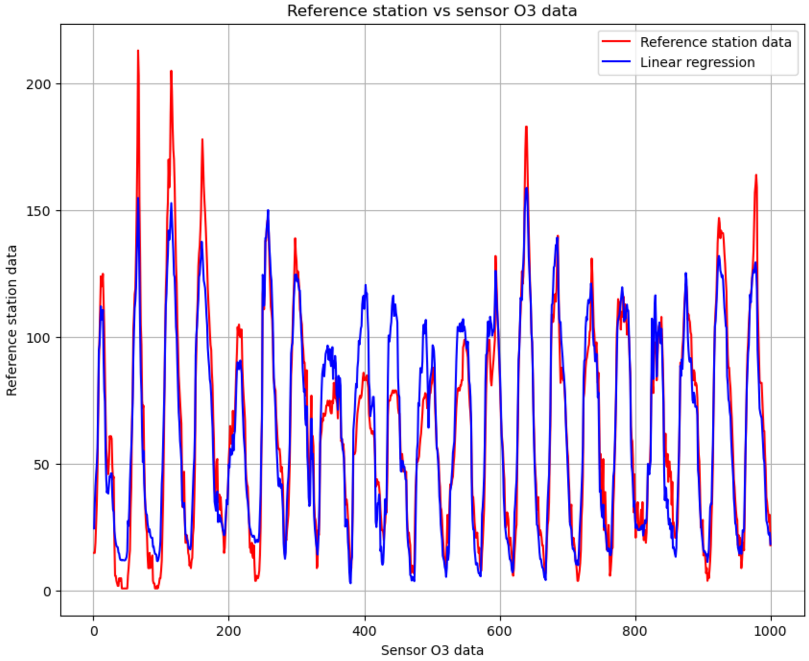

Figure 9: Comparison of the datasets

From the plot we can claim the model is quite good. In addition we can check also how the RMSE and the R^2 indexes are impreved:

- $RMSE = 13.0098$
- $R^2 = 0.8908$

The RMSE is definitely decreased. Moreover, this model has a confidence of 89%, a very good result.

Here I plot also the scatter plot of the model compared to the reference dataset:

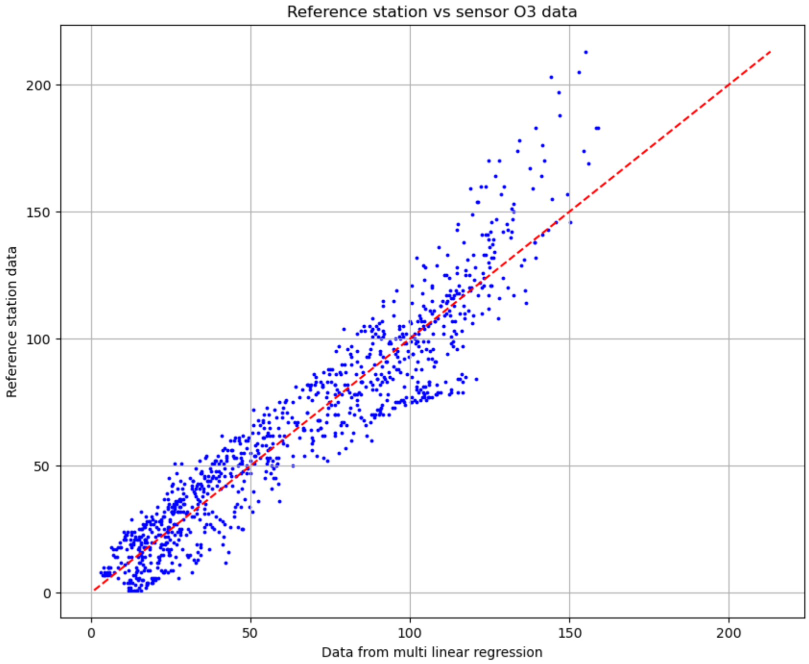

Figure 10: Comparison of the datasets

The scatter plot shows that the dots are closer to the y=x function than it was initially.

3 Conclusion

The results highlight how well the model matches the values from the reference station, showing a strong accuracy in its predictions. The utilization of multi linear regression proved to be an effective technique, that provides a robust approximation of the dataset. From this outcome I verified how practical and useful this method can be. This experience taught me how to apply multi linear regression, and it improved my confidence in using it for future analyses.