Exploratory Data Analysis

Prediction of CO (ppm) particles in a gas chamber.

Dataset Description

Time Series, Multivariate

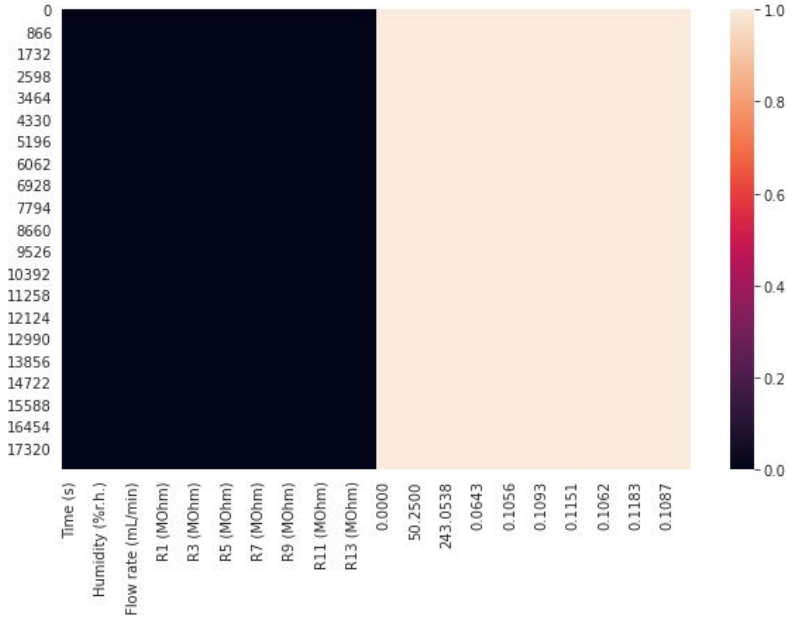
The dataset consist on the reading of 14 temperature modulated mox sensors. Each experiment consists of 100 measurements: 10 experimental mixtures uniformly distributed in the range of 0 - 20 ppm and 10 replicates per concentration.

At the beginning of each experiment the gas chamber is cleaned by 15 mins using a stream of 240 mln/min. And after that the gas mixtures are released at 240 mln/min, thus assuming the flow rate is constant.

A single experiment lasted 25 hours (100 samples x 15 minutes / sample) and was replicated on 13 working days.

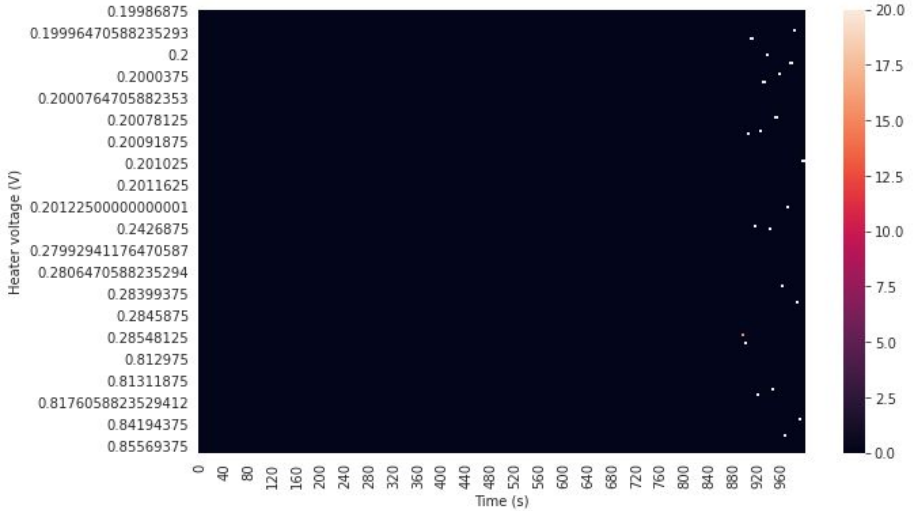
Data Structure and missing values

In the following chart the white-like color represent the missing values, we can appreciate that after the R14 there is a series of empty columns. These columns are irrelevant for the analysis so they get removed from the data.



Heatmap of CO readings.

At the beginning of each experiment, the gas chamber is cleaned with a stream of synthetic air, in the next chart we can appreciate that there are no readings of CO the first 900 seconds, corresponding to the 15 minutes cleaning.



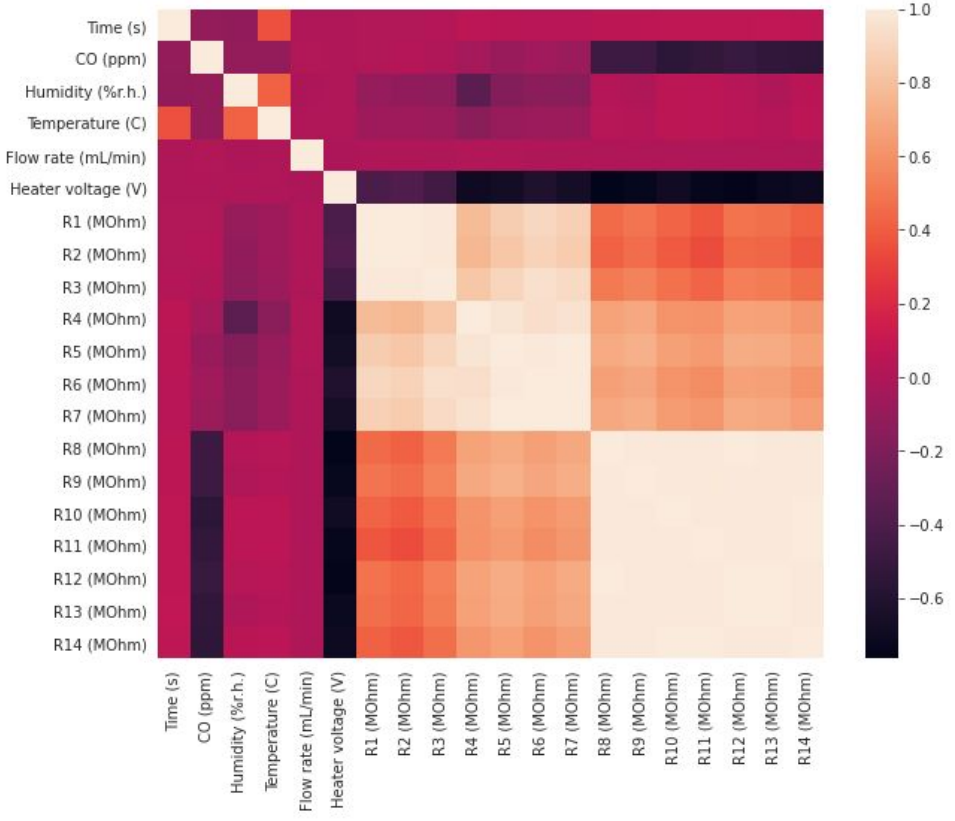
Distribution of numeric values.

By looking at the distribution of the numerical values we can make the following assumptions:

- The CO (ppm) is distributed evenly.
- The Flow rate is constant-like.
- And the humidity is gaussian-like distributed,
- The sensors are left-skewed indicating noise and outliers.

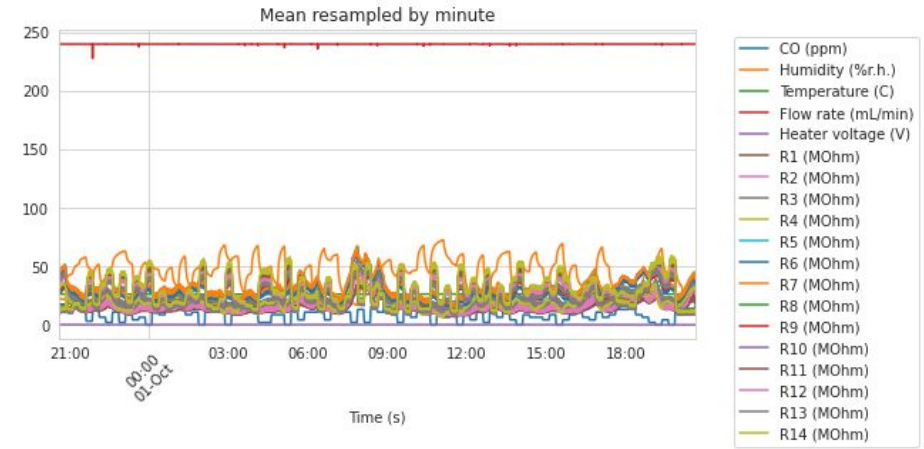
Linear Correlation Heatmap

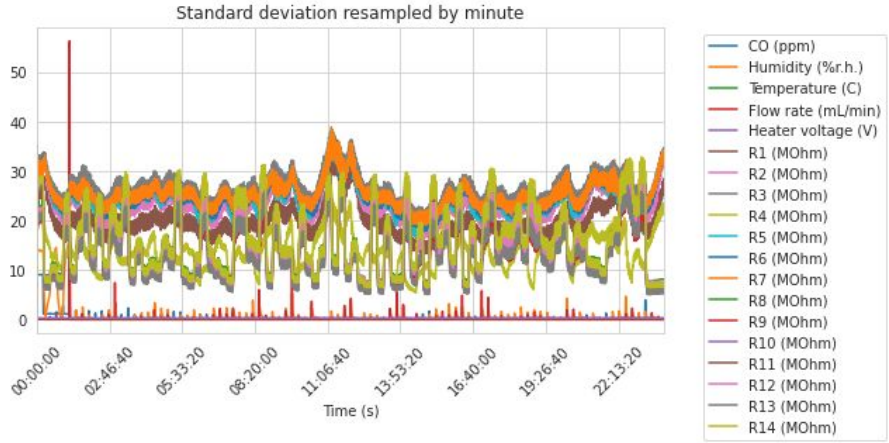
In the next heatmap, all the variables are tested against each other for correlation. It's visible a strong negative correlation between the target (CO ppm) and the readings from the R8 to R14 sensors.



Central tendencies over time

The relevant information about these graphs is the variance of the mean and the standard deviation over time, the mean of the CO (ppm) is not totally constant but can be time-modelled.



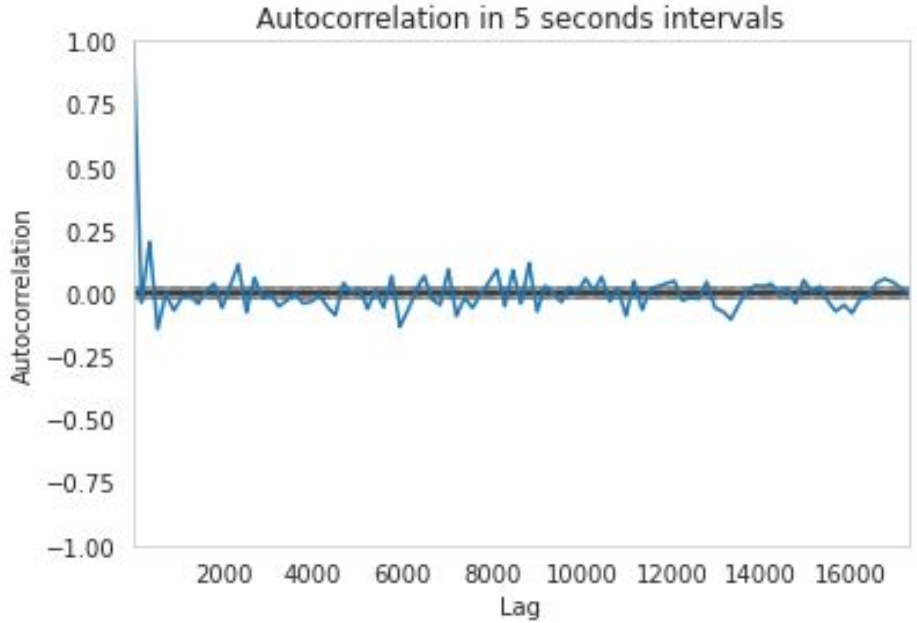


Autocorrelation of target.

Intervals of 5 seconds

In the following plots we are going to visualize the autocorrelation of a variable against a previous observation of the same. The black horizontal lines represent the significance level.

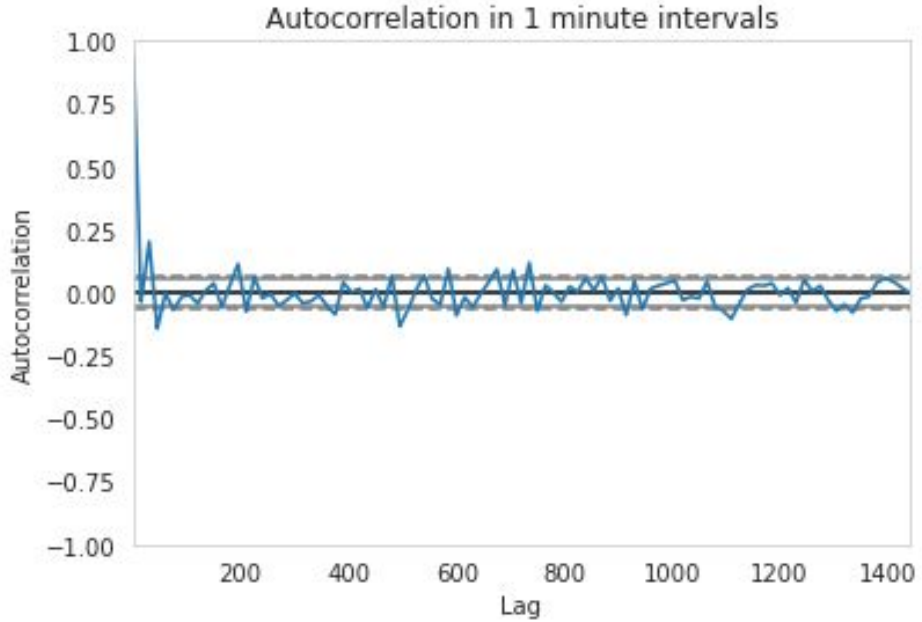
The current plot is sampled at intervals of 5 seconds per lag, the oscillation of the series indicate a correlation in time.



Autocorrelation of target.

Intervals of 1 minute.

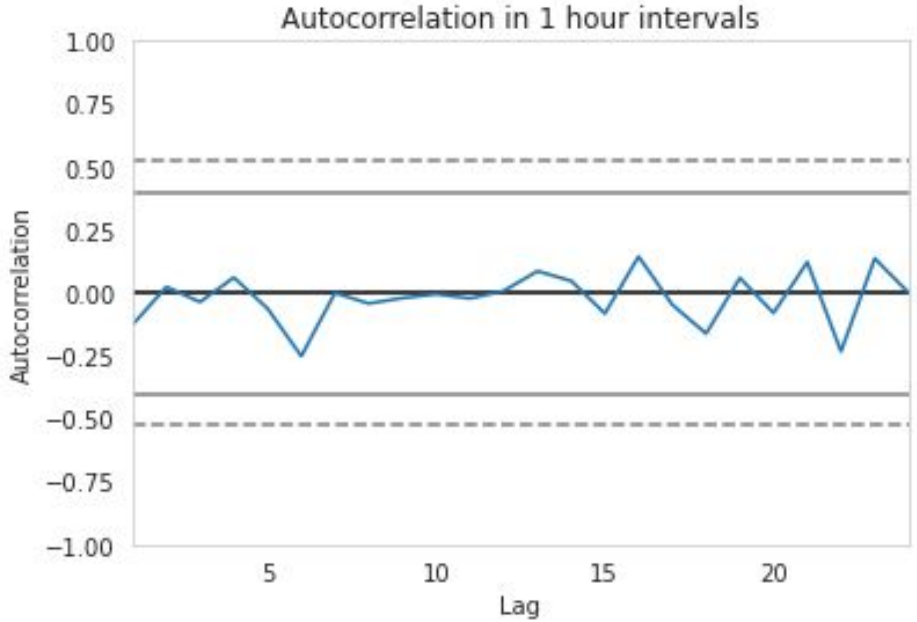
The more large interval of sampling decreased the cyclic component, now the autocorrelation between observations is decreased under the significance level.



Autocorrelation of target.

Intervals of 1 hour

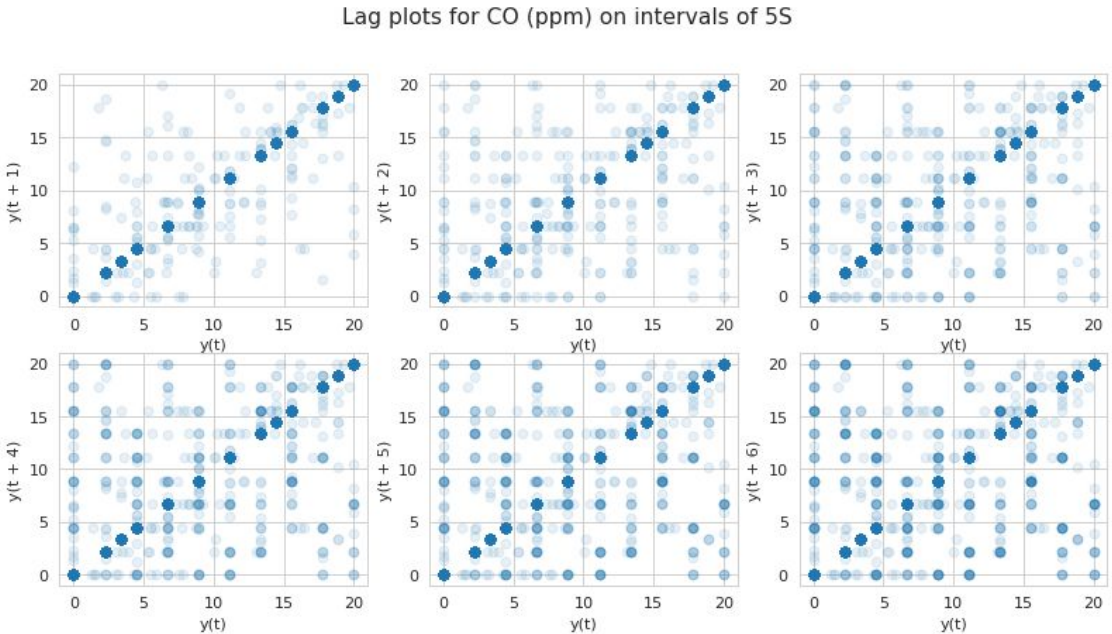
Here the data is sampled on intervals of an hour, corroborating the previous statement. Now the autocorrelation between observations of the same variable is irrelevant.



Lag Plot - 6 intervals (5 S)

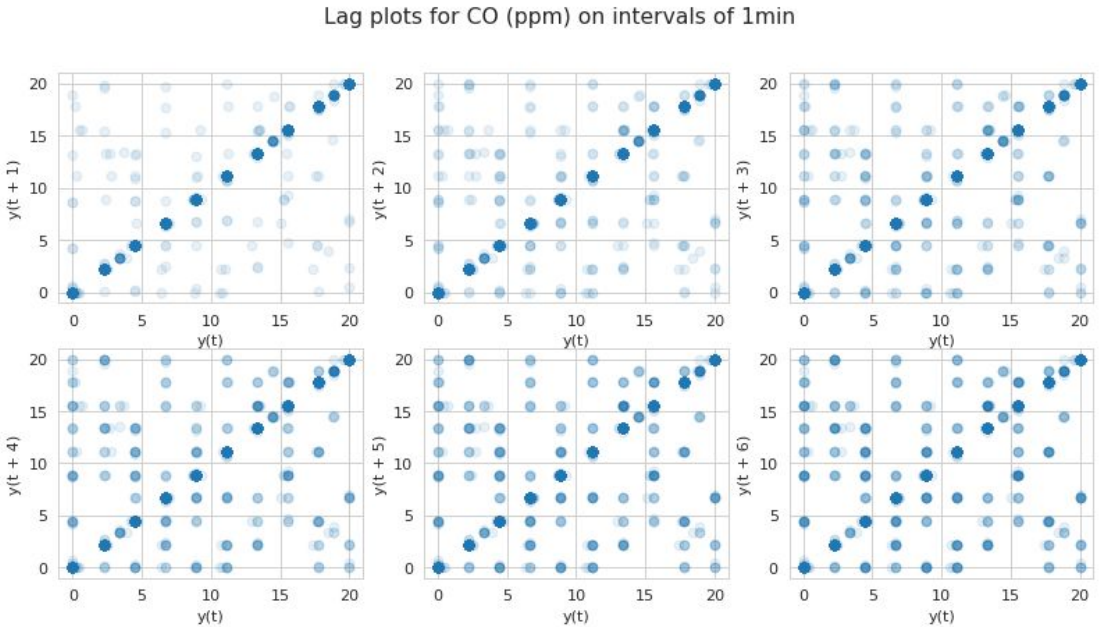
Another way to visualize the correlation between previous observations in the time series is through fixed lag scatter plots.

In the present chart we can see a correlation between the target and previous observations of itself.



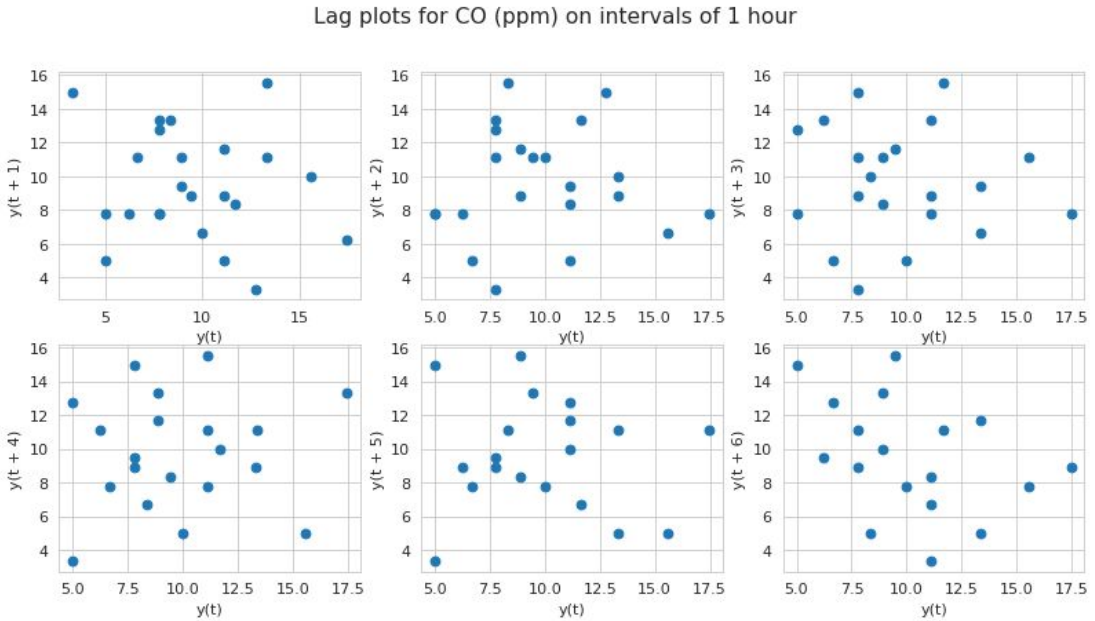
Lag Plot - 6 intervals (1 Min)

As the sampling intervals get increased the correlation tend to decrease, in the one minute intervals this degradation is not that much.



Lag Plot - 6 intervals (Hour)

At intervals of one hour the autocorrelation of the variable is non-existent, this information determine the sampling intervals that can be used for forecasting.



Augmented Dickey-Fuller Test

The ADF test is a common statistical technique to determine whether a variable is “stationary” or not. It is based in the unit root test of current and lagged observations.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e_t$$

Null Hypothesis (p-value > 0.05): The variable is not stationary.

Alternative Hypothesis (p-value < 0.05): The variable is stationary.

Augmented Dickey-Fuller Test

The following table summarizes the p-values obtained after performing the ADF test on the variables from our dataset.

The only non-stationary variable is the temperature.

Summary of ADF		
Feature	**p-value**	**result**
CO (ppm)	0.00000	Stationary
Humidity (%r.h.)	0.00000	Stationary
Temperature (C)	0.22748	No Stationary
Flow rate (mL/min)	0.00000	Stationary
Heater voltage (V)	0.00000	Stationary
R1 (MOhm)	0.00663	Stationary
R2 (MOhm)	0.00016	Stationary
R3 (MOhm)	0.00061	Stationary
R4 (MOhm)	0.00002	Stationary
R5 (MOhm)	0.00017	Stationary
R6 (MOhm)	0.00024	Stationary
R7 (MOhm)	0.00020	Stationary
R8 (MOhm)	0.00000	Stationary
R9 (MOhm)	0.00000	Stationary
R10 (MOhm)	0.00000	Stationary
R11 (MOhm)	0.00000	Stationary
R12 (MOhm)	0.00000	Stationary
R13 (MOhm)	0.00000	Stationary
R14 (MOhm)	0.00000	Stationary

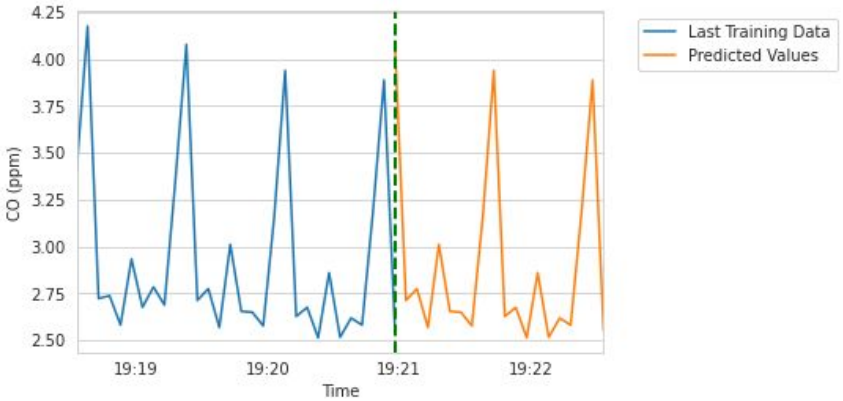
Model Selected: LSTM NN

Because time is relevant in the prediction of the CO ppm particles between the cycles of the sensor's heaters, the Long Short Term Memory Neural Network (LSTM NN) suits for the case.

Features Selected:

- R8 (MOhm)
- R9 (MOhm)
- R10 (MOhm)
- R11 (MOhm)
- R12 (MOhm)
- R13 (MOhm)
- R14 (MOhm)

Model Results



In the previous chart we can visualize the predictions made by the LSTM model, the green-dotted line represents the last observed value in the training data, the yellow line represents the forecasted values.

Root Mean Square Error (RMSE):

2.53