High-dimensional statistics

Project 1: Exploratory data analysis

Corentin Merle s162662
Jad Akkawi s216641

Data Science and Engineering, bloc 2
Liège Université
School of Engineering
Academic year 2022-2023

1 Data

In order to have a data set responding to all imposed constraints we choose the `Air Quality Data Set` dataset From Saverio De Vito, ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development[1]. It contains sensor output (hourly averaged sensor response) of a multi-sensor device that measures air pollutants concentrations and their corresponding reference values from a measuring station (ground truth). The purpose of this Dataset was to create a neural network that calibrates sensor output to obtain accurate concentrations using fusion algorithms.

1.1 Description and Preprocessing

The dataset initially contains 9358 observations, thereby we selected the 500 first ones (the observations of March 2004) as new dataset. We choose to select 500 consecutive data-points in order to keep the time consistency as we have attributes **Date** and **Time**. Moreover, our sample has 15 quantitative variables ($\frac{500}{15} = 33.33 > 5$, no flat data) which are listed below,

- `Date` $(DD/MM/YYYY)$
- `Time` $(HH.MM.SS)$
- `CO.GT.` True hourly averaged concentration CO in mg/m^3 (reference analyzer)
- `PT08.S1.CO` (tin oxide) hourly averaged sensor response (nominally CO targeted)
- `NMHC.GT.` True hourly averaged overall Non Metanic HydroCarbons concentration in $\mu g/m^3$ (reference analyzer)
- `C6H6.GT.` True hourly averaged Benzene concentration in $\mu g/m^3$ (reference analyzer)
- `PT08.S2.NMHC.` (titania) hourly averaged sensor response (nominally $NMHC$ targeted)
- `NOx.GT.` True hourly averaged NO_x concentration in Parts per billion ppb (reference analyzer)
- `PT08.S3.NOx.` (tungsten oxide) hourly averaged sensor response (nominally NO_x targeted)
- `NO2.GT.` True hourly averaged NO_2 concentration in $\mu g/m^3$ (reference analyzer)
- `PT08.S4.NO2.` (tungsten oxide) hourly averaged sensor response (nominally NO_2 targeted)
- `PT08.S5.O3.` (indium oxide) hourly averaged sensor response (nominally O_3 targeted)
- `T` in $°C$
- `RH` Relative Humidity (%)
- `AH` Absolute Humidity `AH_bin` Absolute Humidity (Quantitative → binary)

However there is no binary variables. Therefore we decided to transform the **Absolute Humidity** variable into a binary variable using 0.75 as threshold to distinguish low and high humidity level. Finlay we replace missing values which were set to -200 by NA. The data preprocessing is performed with the R script named `data_pre_processing.R` and the new dataset file is saved as `preProcessAirQuality.csv`. Additional information about the data collection are provided in Appendix A.

[1] https://archive.ics.uci.edu/ml/datasets/Air+Quality

1.2 Missing data

In the R script `missing_data.R` we performed a missing data analysis in order to observe the distribution of missing values and also to ensure that the missingness rate is superior to 1%. As we can see on Figure 1 the missingness rate of our dataset is equal to 2.2% ($> 1\%$).

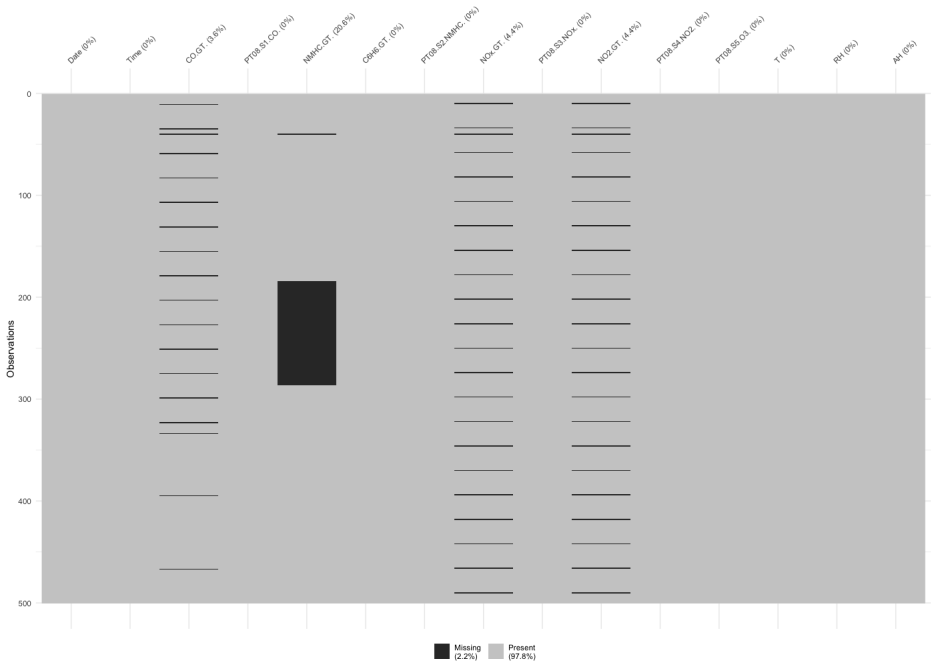

Figure 1: Missing matrix

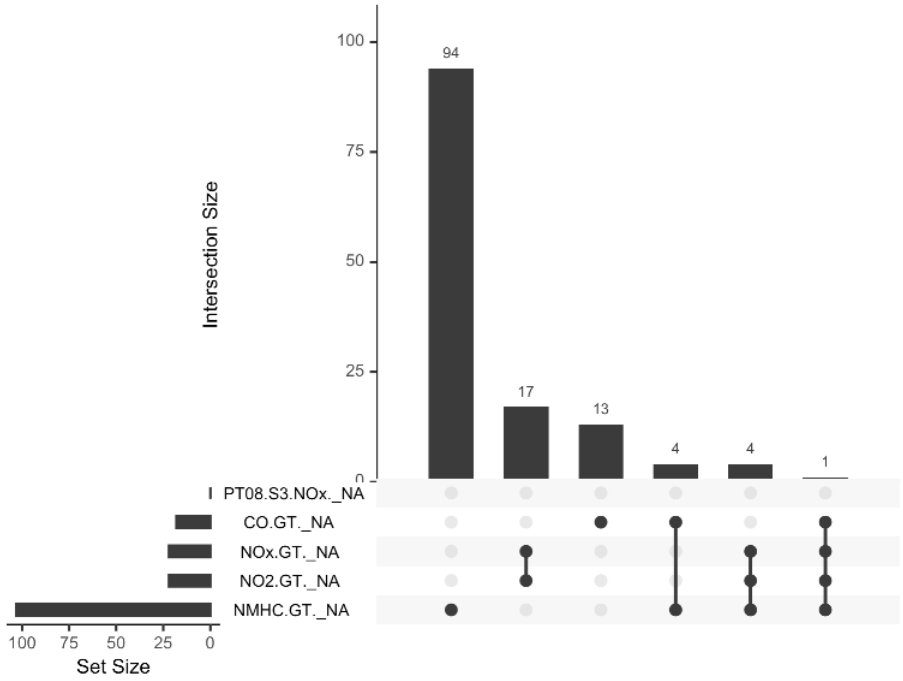

Figure 2: Intersection of missing values

We can see from Figure 2 that from the 133 missing cases, 103 of them are "non metanic hydrocarbure", and NO_x and NO_2 are missing each 22 times and finally CO is missing 18 values, the missingness of these values occasionally intersect in some observations.

In order to infer on the cause of the values being missing, we plot on Figure 3 to 6 those variables with the time variable to visualize location of missing points.

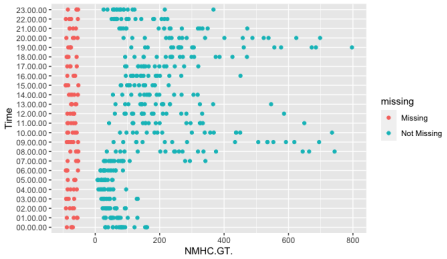

Figure 3: NMHC.GT vs Time

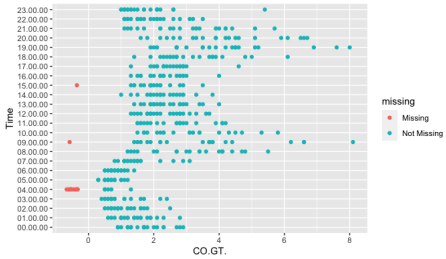

Figure 4: CO.GT vs Time

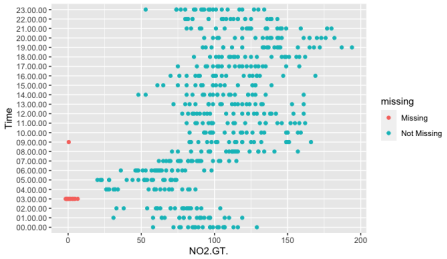

Figure 5: NO2.gt vs Time

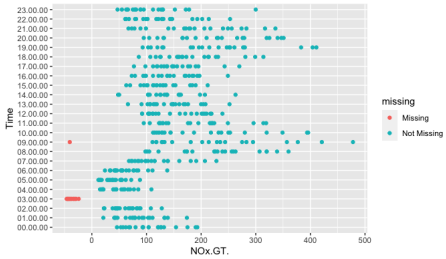

Figure 6: NOx.GT vs Time

We can clearly observe that during the month of march at 3:00 AM the reference center did not measure any data regarding NO_2 and NO_x maybe because the machines were resetting (22 missing value each). Same goes for CO values missing but at 4:00 AM (18 missing values). Thus it seems to be a MAR mechanism as the equality $P(M = 1|X_{obs}, X_{mis}) = P(M = 1|X_{obs})$ is verified for those variables.
While regarding the ground truth $NMHC$, we realize that from 18th till 20th of Mars these elements weren't measured (103 missing values). Thus it seems to be a MCAR mechanism as the it could be due to the measuring device that might have ran out of batteries.
Finally, these 4 variables were missing collectively on 12/03/2004 at 9:00 AM.

Fortunately, 11 attributes out of 15 do not contain missing values which can be beneficial in order to impute missing values after correlation analysis. On Figure 7 and 8 we can observe respectively the complete and pairwise correlation matrices. As we can see that there are strong correlation between most of the attributes. More specifically:

- CO.GT.$\sim$ C6H6.GT. with correlation equal to 0.9810
- NMHC.GT.$\sim$ C6H6.GT. with correlation equal to 0.8621
- NOx.GT.$\sim$ CO.GT. with correlation equal to 0.9575
- NO2.GT. $\sim$ PT08.S2.NMHC. with correlation equal to 0.9003

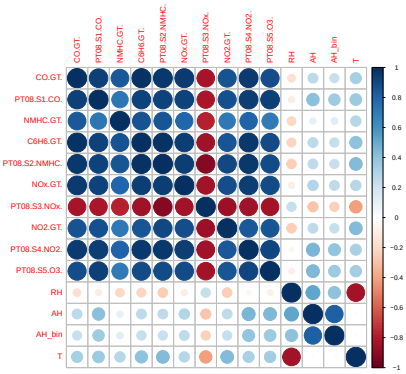

Figure 7: Correlation (complete)

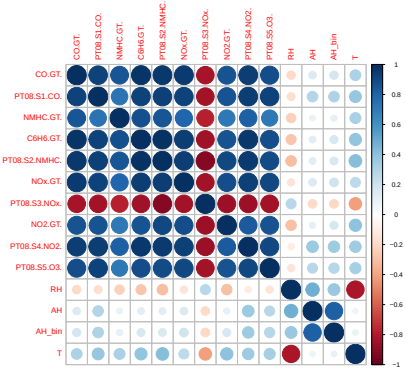

Figure 8: Correlation (pairwise)

1.3 Missing values imputation

In order to impute missing values a linear regression model was done for each case using the most correlated variable. We will now explore graphically the changes in the quantiles, mean and variance of the imputed variables.

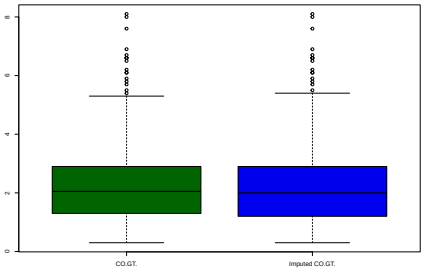

Figure 9: Box plot CO

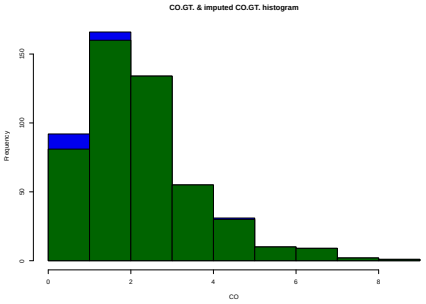

Figure 10: Histogram CO

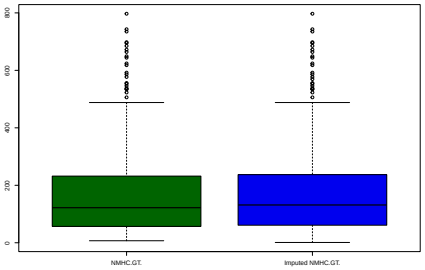

Figure 11: Box plot NMHC

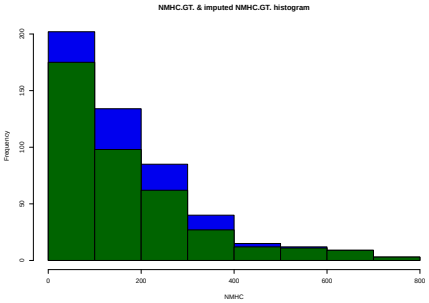

Figure 12: Histogram NMHC

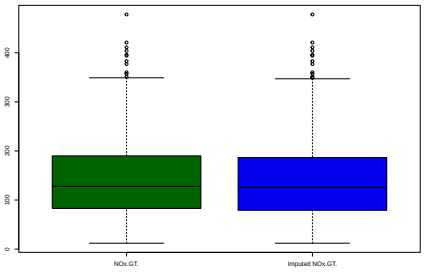

Figure 13: Box plot NOx

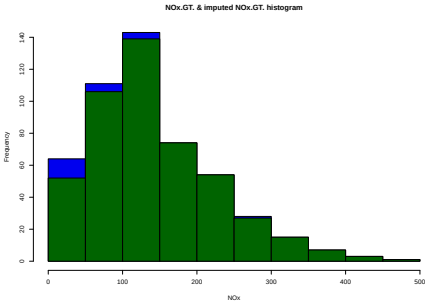

Figure 14: Histogram NOx

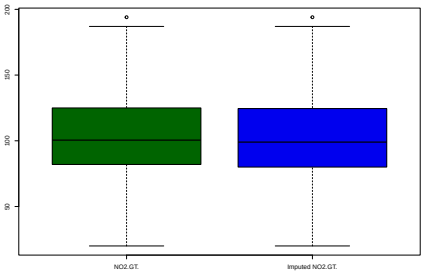

Figure 15: Box plot NO2

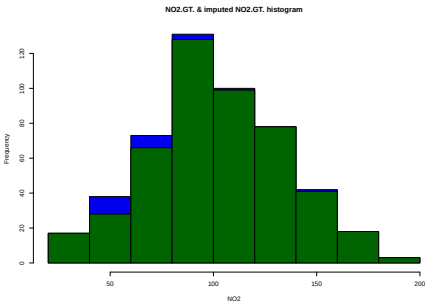

Figure 16: Histogram NO2

We can see that changes aren't major in the distribution of most variables. While we can see some changes in the histogram of NMHC.GT variable and most likely this is due to the fact that it has the most missing values and the lowest max correlation found (0.8621). In all cases the correlation was high enough to not cause major changes in the distribution. The new dataset is stored in the file `cleanAirQuality.csv`.

2 Exploratory data analysis

Now we have clean and complete data to analyse. In file `EDA.R` we can start by having a look at a summary of the data on Table 1. Everything seems right to start the exploratory data analysis.

Table 1: Data Summary

Statistic	N	Mean	St. Dev.	Min	Max	Unit
Date	500	11.668	6.031	1	22	Day
Time	500	12.468	6.947	1	24	hour
CO.GT.	500	2.252	1.365	0.300	8.100	mg/m^3
PT08.S1.CO.	500	1,223.038	242.831	818	2,040	Sensor Response
NMHC.GT.	500	172.733	148.746	1.022	797.000	$\mu g/m^3$
C6H6.GT.	500	9.943	7.118	0.600	39.200	$\mu g/m^3$
PT08.S2.NMHC.	500	935.424	261.010	457	1,754	Sensor Response
NOx.GT.	500	140.889	83.224	12.000	478.000	ppb
PT08.S3.NOx.	500	1,031.834	269.595	537	1,935	Sensor Response
NO2.GT.	500	101.235	32.749	20.000	194.000	$\mu g/m^3$
PT08.S4.NO2.	500	1,571.402	282.580	1,050	2,679	Sensor Response
PT08.S5.O3.	500	1,028.344	384.831	341	2,359	Sensor Response
T	500	14.400	4.488	6.100	29.300	C°
RH	500	50.044	14.360	14.900	83.200	%
AH_bin	500	0.604	0.490	0	1	Low/High Humidity (binary)

2.1 Variables distributions, histograms and curve fittings

We can skip **DATE** and **TIME** variables because we know well that they originate from a uniform distribution since observations are recorded every day and every hour.

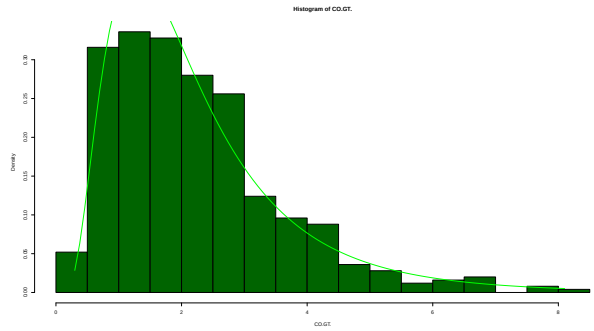

Figure 17: CO.GT.

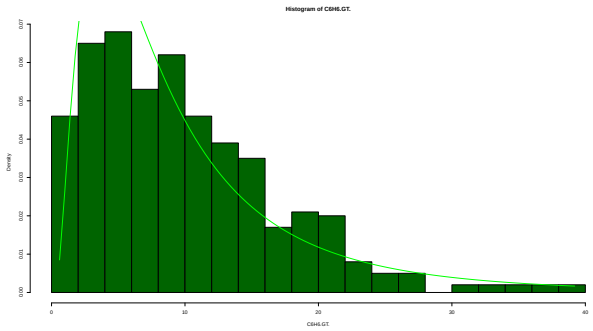

Figure 18: C6H6.GT.

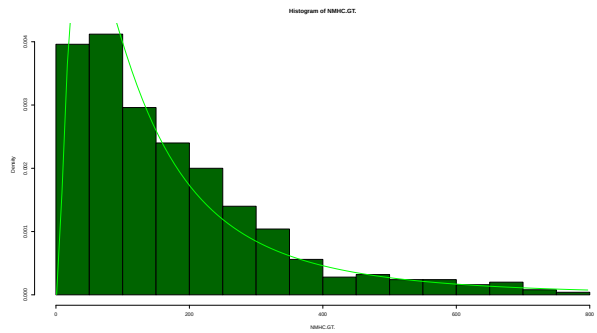

Figure 19: NMHC.GT.

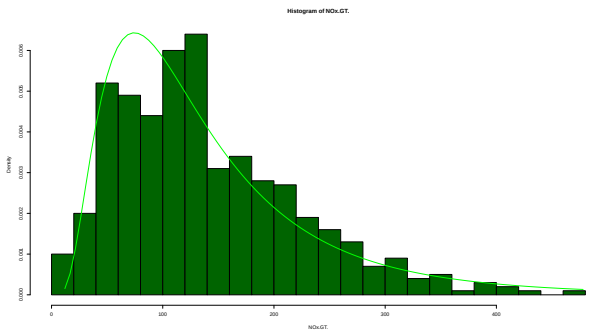

Figure 20: NOx.GT.

CO.GT. seem to fit really well a log Normal distribution, with a sharp peak at the left and a tail spreading far to the right of the graph. The same could be said to the NMHC.GT., C6H6.GT., NOx.GT.

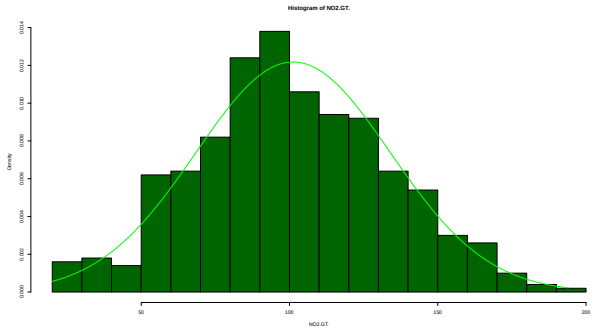

Figure 21: NO2.GT.

For the last ground truth concentration which is NO2.GT., the peak seems to be at more centered and its obvious we have 2 tails, therefor it fits more a Normal distribution.(apparent symmetry)

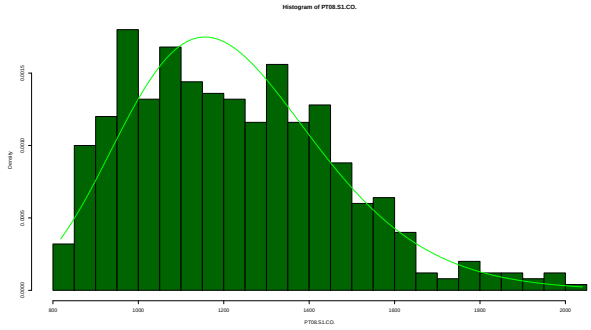

Figure 22: PT08.S1.CO.

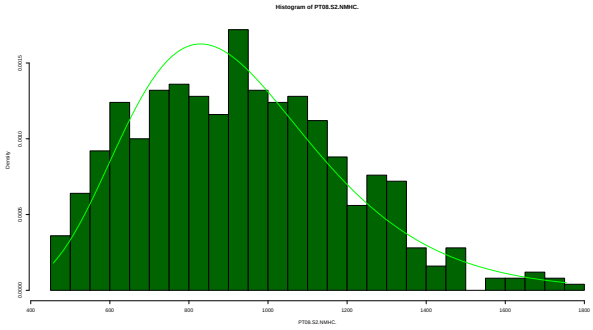

Figure 23: PT08.S2.NMHC.

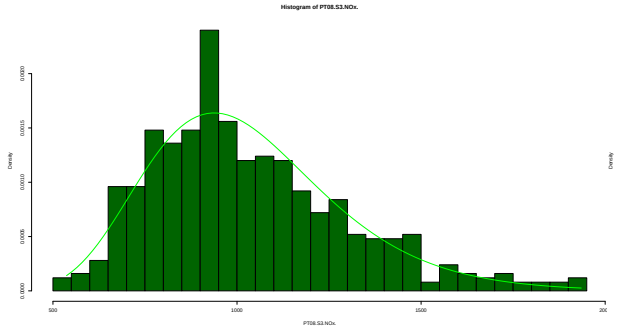

Figure 24: PT08.S3.NOx.

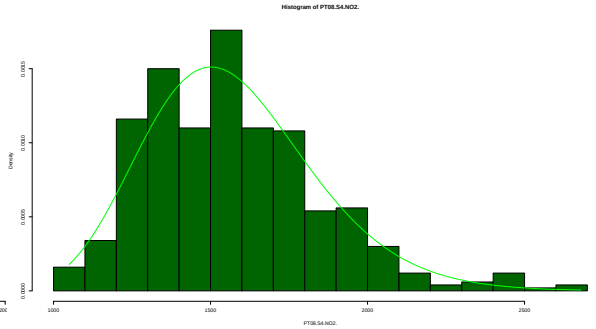

Figure 25: PT08.S4.NO2.

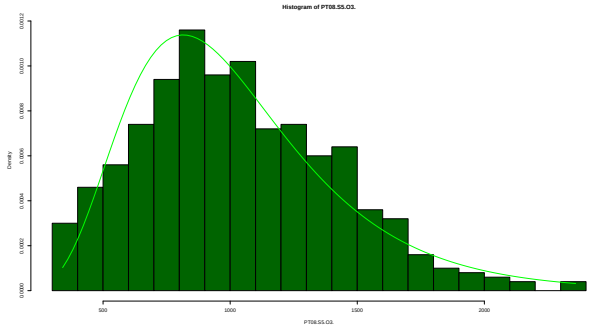

Figure 26: PT08.S5.O3.

Now regarding hourly response of sensors: PT08.S1.CO, PT08.S2.NMHC, PT08.S3.NOx, PT08.S4.NO2 and PT08.S5.O3. Its clear that the distribution is much wider than that of GT concentrations due to differences in units and measuring methods. Still, log normal seems to be a good fit for the sensor response due to peaks to

the left and tails at the right. Given that distribution is more spread out, the peaks may not be extremely to the left but at least there is a clear undoubted skewness to the right (positive skew)

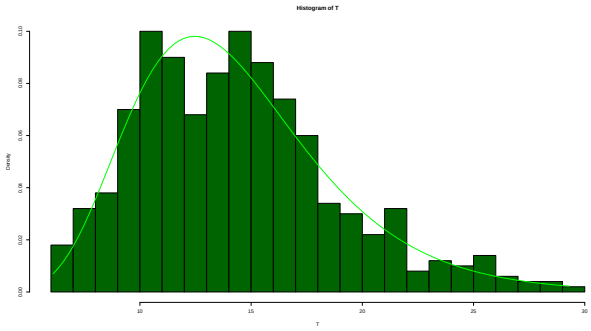

Figure 27: T

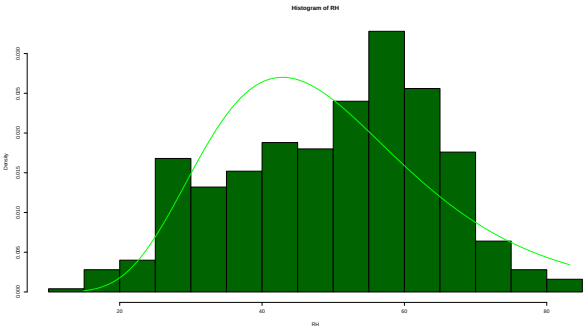

Figure 28: RH

Moreover we see the positive skew in the Temperature variable and the negative skew in the relative humidity Variable.

Finally we can deduce from the summary Table 1 that the binary variable **AH_BIN** have 60,4 % of high humidity (1) instances and thus 39,6 % of low humidity instances.

2.2 Correlation structure

For this section we start off by viewing some variables as time series:

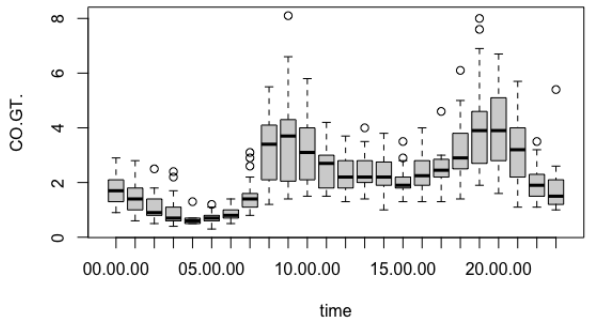

Figure 29: Time vs CO.GT

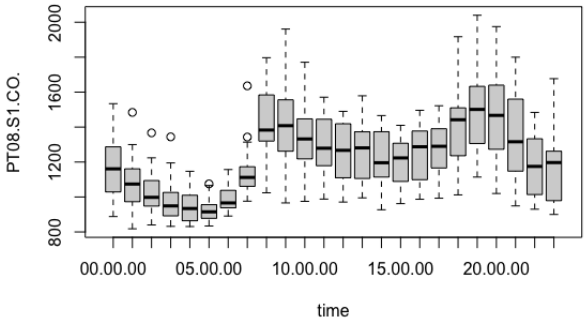

Figure 30: Time vs S1.CO

The reference values from the station and the sensor output seem to have the same daily pattern with peaks during rush hours of mornings and evenings. Similar behaviours is observed in other pollutants. This means that CO.GT, PT08.S1.CO, NMHC.GT., C6H6.GT., PT08.S2.NMHC., NOx.GT., NO2.GT., PT08.S4.NO2. and PT08.S5.O3. will have high correlation between each other. This can be seen in the pairs plot on Figure 35.

The only targeted molecule to have an opposite behaviour is NO_x:

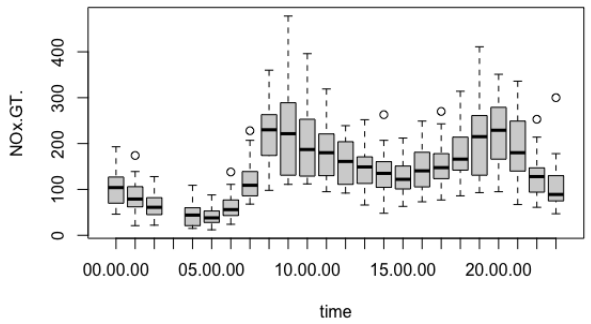

Figure 31: Time vs NOx.GT

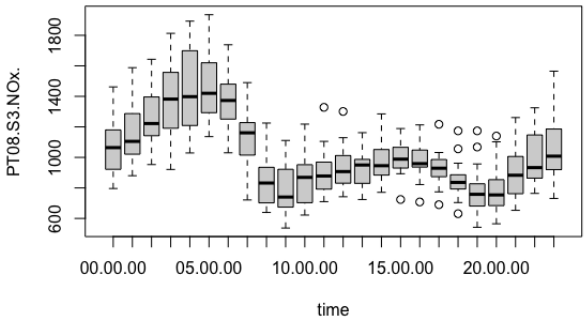

Figure 32: Time vs S3.NOx

However, the sensor head targeting NOx seems to work in an opposite manner by having a weak or dampened response when exposed to NO_x. This explains the negative correlation between this response and the rest of the concentrations/responses.

A noticeable correlation is between C6H6.GT. and PT08.S2.NMHC with a value of 0.9825:

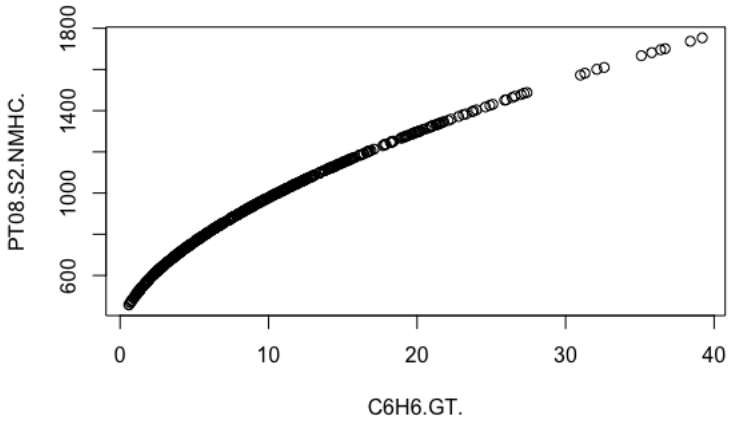

Figure 33: C6H6.GT. vs PT08.S2.NMHC

This supposedly measured variable of C6H6.GT is almost perfectly correlated with the S2.NMHC response if it weren't for the slight non-linearity. Our doubts are that one of these variables has been synthesized by the other.

We can also see that temperature and relative humidity variables are well correlated with each others while their correlations with the rest of the variables are week.

2.3 Outlying observations

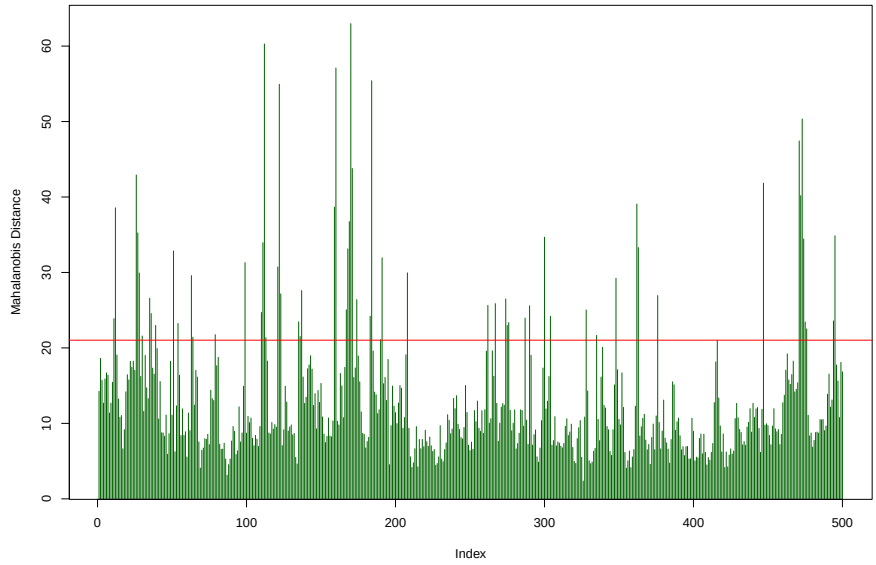

Figure 34: Mahalanobis

The Chi-Square cut-off value used in the plot 34 does not really reflect the outlying values because chi-square threshold is based on the normality assumption of the variables and, as explained earlier, only one variable out of the 12 has a distribution similar to that of a Normal, the rest resembles more a Log-Normal distribution. This is why we have around 60 out of 500 observations considered as outlying when the threshold of Mahalanobis-distance is at 21.5. While we should expect around 25 observation to be outlying when using the 95% threshold.

3 Dimension reduction

As correlation is high between variables (Figure 35) we use PCA in order to reduce the data dimension. This section has been achieved thanks to the file `dim_reduc.R`.

Since we have different units and different scales of values we opted a PCA on the covariance matrix over the **standardized** data which is equivalent to the correlation matrix.

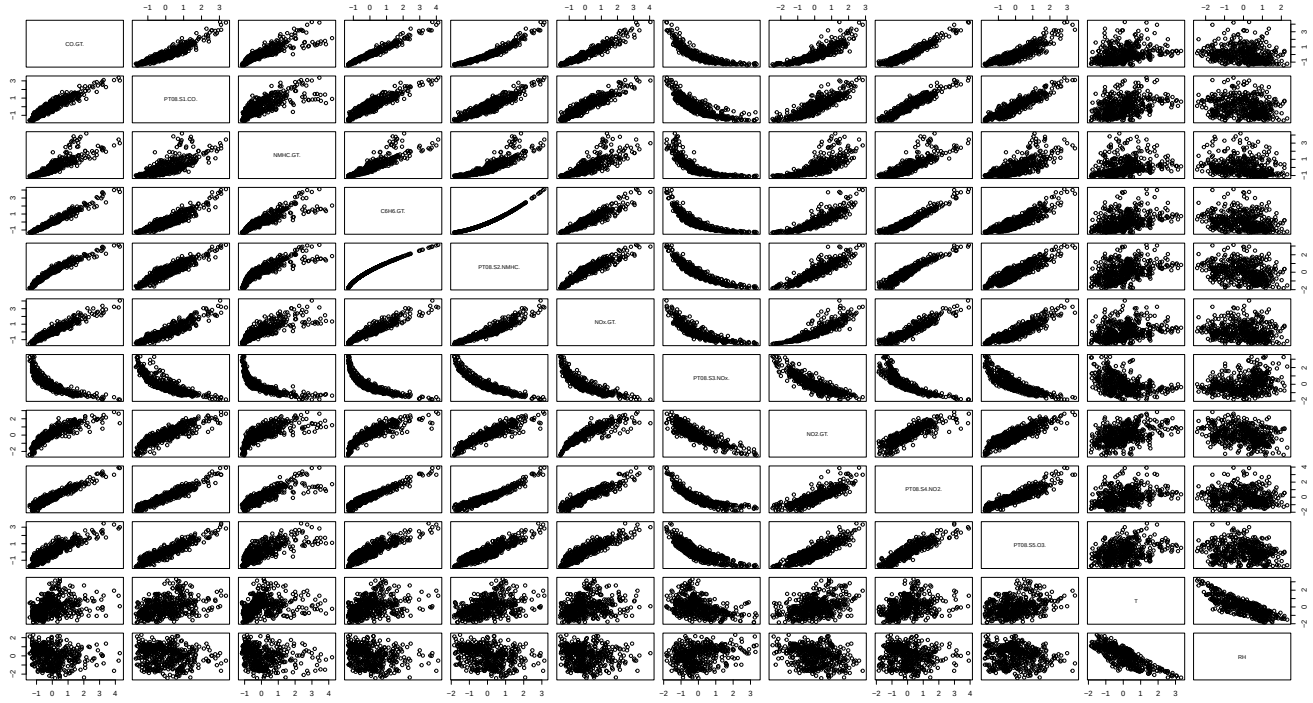

Figure 35: Pairs plot

We use the rule of thumb to decide on the number of principal components. By observing the scree plot on Figure 36 we see that we can keep only two principal components, the first represents 9.23 of the variability and the second represents 1.62 of the variability of the data. Furthermore on Figure 37 we can observe the correlation circle and how variables are described by the two principal components. It confirms the discussion at Section 2.2 and observations of Figure 35. In fact we can see that T and RH are mainly described by the same component but they are inversely proportional, and CO.GT., PT08.S1.CO., NMHC.GT., C6H6.GT., PT08.S2.NMHC., NOx.GT., PT08.S3.NOx., NO2.GT., PT08.S4.NO2. and PT08.S5.O3. are mainly described by the same component and proportional except PT08.S3.NOx. which is inversely proportional.

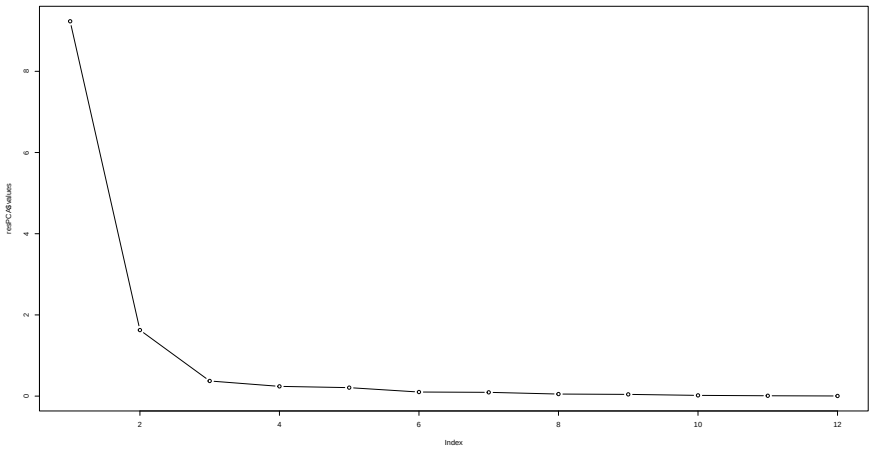

Figure 36: Scree plot

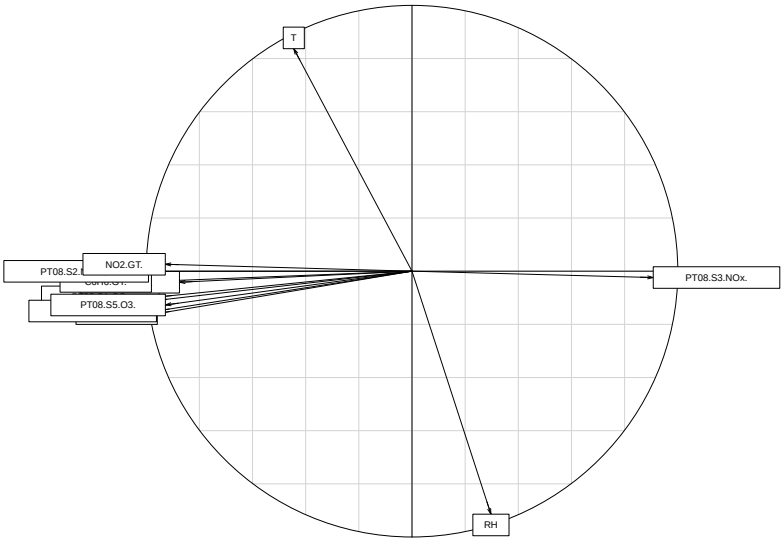

Figure 37: Correlation Circle

Finally by representing quantitative variables with respect to the first two principal components and the binary variable with two different colors we get the 2D plot at Figure 38. We observe 2 things: First, the two subsets of observations created from the binary variable do not form separate clusters, maybe we have not chosen a convenient threshold when creating this variable. Second, variability of the observations is clear in both orthogonal directions meaning our plane has truly captured the sample's dispersion.

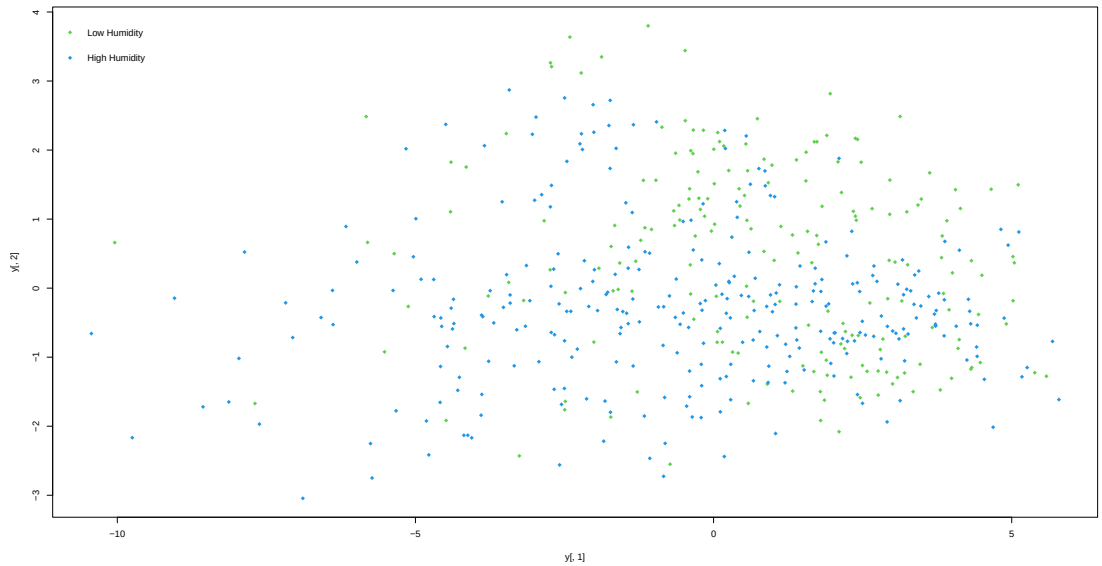

Figure 38: First Principal Plan

Appendix

A Dataset Information

- The hourly averaged responses are measured with an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device.
- The device was located on the field in a significantly polluted area, at road level,within an Italian city.
- Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on-field deployed air quality chemical sensor devices responses.
- Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.
- Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities.
- Missing values are tagged with -200 value.
- This dataset can be used exclusively for research purposes. Commercial purposes are fully excluded.