Faculty of Applied Science

MATH2021-1 High-dimensional statistics

Project 1 : Exploratory Data Analysis

Teacher :
Gentiane HAESBROECK

Group :
Romain LAMBERMONT
Arthur LOUIS

Master in Data Science
October 26, 2022

Contents

List of Figures

List of Tables

1 Presentation of the data

1.1 Discussion on the data

The data we used in this project is a subset of the data collected by the ENEA (National Agency for New Technologies, Energy and Sustainable Economic Development) alonside a road in a polluted area of Italy. This dataset is available on the UCI repository[1]. The data was harvested using a multicensor device and reference analyzers. The data was collected between March 2004 and February 2005.

There are 5 couples of variables in the dataset, each couple is composed of a variable measured by the reference analyzer and the corresponding variable measured by the multicensor device. The values represent the hourly average concentration of each variable. In addition to the variables couples, we have 3 other variables representing the temperature and the humidity (both relative and absolute). The different values are stored in the following columns of our dataset :

- `CO(GT)` : concentration of CO in the air (in mg/m^3)
- `PT08.S1(CO)` : average sensor response (nominally CO targeted)
- `NMHC(GT)` : concentration of non-methane hydrocarbons in the air (in μg/m^3)
- `C6H6(GT)` : concentration of benzene in the air (in μg/m^3)
- `PT08.S2(NMHC)` : average sensor response (nominally NMHC targeted)
- `NOx(GT)` : concentration of NOx in the air (in parts per billion)
- `PT08.S3(NOx)` : average sensor response (nominally NOx targeted)
- `NO2(GT)` : concentration of NO_2 in the air (in μg/m^3)
- `PT08.S4(NO2)` : average sensor response (nominally NO_2 targeted)
- `PT08.S5(O3)` : average sensor response (nominally O_3 targeted)
- `T` : temperature (in °C)
- `RH` : relative humidity (in %)
- `AH` : absolute humidity

On top of that, we created a binary indicator per variable measured by the reference analyzers which is equal to 1 when the measured value is above the median of the variable and 0 otherwise. The binary values are stored in the following columns :

- `HIGH_CO` : binary indicator for CO (above/under median)
- `HIGH_NMHC` : binary indicator for NMHC (above/under median)
- `HIGH_C6H6` : binary indicator for benzene (above/under median)
- `HIGH_NOx` : binary indicator for NOx (above/under median)
- `HIGH_NO2` : binary indicator for NO2 (above/under median)

1.2 Link between the variables

The values for each couple are obviously going to be quite correlated, as they are measuring the same thing. Furthermore, the values for the binary indicators are going to be highly correlated with the values of the corresponding measurements from the reference analyzers (as the binary indicator is equal to 1 when the value is above the median and 0 otherwise).

[1] https://archive.ics.uci.edu/ml/datasets/air+quality

2 Information about missing data

We have a total of 2.1% of missing values but this number is overestimated because the binary indicators are taken into account. The real ratio is 1.7% without this indicators. The missing values are due to hardware problems related to the measuring instruments and to the fact that the data was collected in a real environment.

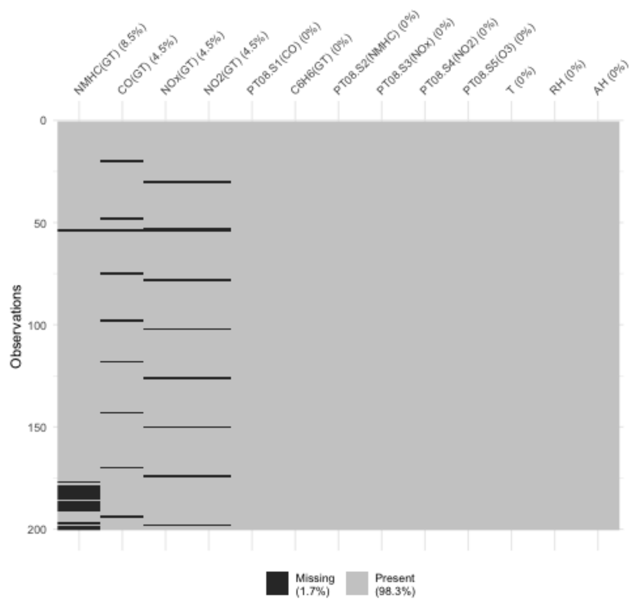


Figure 1: Missing data

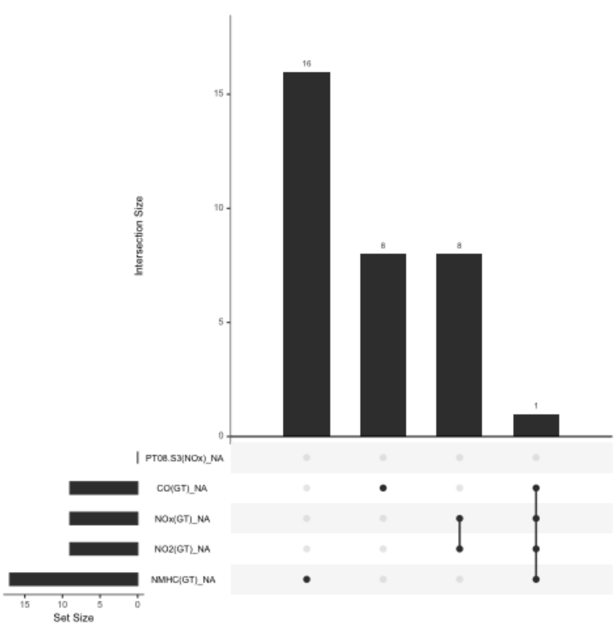


Figure 2: Missing data heatmap

To handle the missing values we will use for the rest of the project the complete case strategy. Indeed, replacing missing data by the mean would not be a good idea because it would change the correlation structure of the data which is important for the next steps. We only have 33 lines with missing values so we can afford to remove them keeping the ratio $\frac{n}{p} > 5$ as we have a total of 163 lines and 21 variables, we keep $7.8 > 5$.

Using the two figures here above (genereated using the functions `vis_miss` and `gg_miss_upset`) we can see that the missing values are not randomly distributed. Indeed, when there's a failure on the CO sensor, the NOx and NO_2 sensors are also failing a little bit after. We can also see that when the NOx sensor is failing, the NO_2 sensor is also failing, this is quite logical because NO_2 is a kind of NOx. Around the 175th observation, we can see the non methane hydrocarbons sensor failing and never quite coming to its original state. All these observations are confirmed by the second figure as the NMHC measure is the most missing well over the CO sensor or the couple NOx/NO_2 which is failling the same number of times and always together.

3 Exploratory data analysis

3.1 Statistical analysis

	n	mean	sd	median	trimmed	mad
CO(GT)	191	2.748691	1.596801	2.50	2.560784	1.33434
PT08.S1(CO)	200	1339.000000	255.446559	1332.50	1328.356250	233.50950
NMHC(GT)	183	160.158470	139.745774	122.00	138.448980	118.60800
C6H6(GT)	200	12.254000	8.274006	11.05	11.318750	7.33887
PT08.S2(NMHC)	200	1016.950000	281.940276	1017.50	1005.556250	278.72880
NOx(GT)	191	175.842932	94.999980	161.00	168.830065	85.99080
PT08.S3(NOx)	200	1003.195000	278.431170	945.00	976.950000	234.99210
NO2(GT)	191	115.612565	34.357971	119.00	116.549020	35.58240
PT08.S4(NO2)9	200	1671.040000	305.901187	1622.50	1641.750000	237.21600

	min	max	range	skew	kurtosis	se
CO(GT)	0.5	8.1	7.6	1.0755547	0.9704424	0.1155405
PT08.S1(CO)	831.0	2040.0	1209.0	0.3506958	-0.0955406	18.0627994
NMHC(GT)	7.0	685.0	678.0	1.3200143	1.4422105	10.3303049
C6H6(GT)	1.0	39.2	38.2	1.0086394	0.8567735	0.5850606
PT08.S2(NMHC)	501.0	1754.0	1253.0	0.3211184	-0.3339212	19.9361881
NOx(GT)	16.0	478.0	462.0	0.6699740	-0.0263597	6.8739573
PT08.S3(NOx)	537.0	1918.0	1381.0	0.9745243	0.9172654	19.6880569
NO2(GT)	28.0	194.0	166.0	-0.2179968	-0.4405303	2.4860555
PT08.S4(NO2)	1134.0	2679.0	1545.0	0.9116828	0.7869557	21.6304804

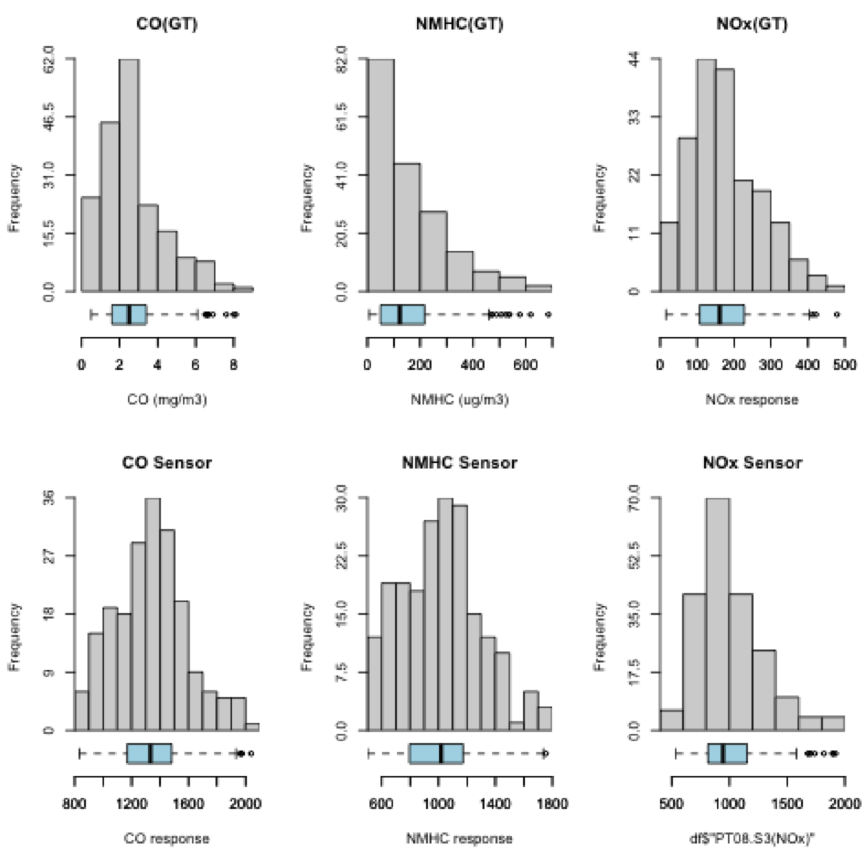
	n	mean	sd	median	trimmed	mad
PT08.S5(O3)	200	1233.2450000	389.2906253	1204.5000	1222.0375000	384.734700
T	200	15.1965000	5.5702402	14.3000	14.8068750	5.189100
RH	200	49.8030000	15.1352426	53.9000	50.6387500	15.270780
AH	200	0.8085450	0.1059962	0.8125	0.8092338	0.104375
HIGH_CO	191	0.7225131	0.4489355	1.0000	0.7777778	0.000000
HIGH_NMHC	183	0.9945355	0.0739221	1.0000	1.0000000	0.000000
HIGH_C6H6	200	0.6500000	0.4781665	1.0000	0.6875000	0.000000
HIGH_NOx	191	0.9947644	0.0723575	1.0000	1.0000000	0.000000
HIGH_NO2	191	0.5968586	0.4918179	1.0000	0.6209150	0.000000

	min	max	range	skew	kurtosis	se
PT08.S5(O3)	384.0000	2359.0000	1975.0000	0.2942286	-0.1157893	27.5270041
T	6.1000	29.3000	23.2000	0.6005822	-0.4670720	0.3938755
RH	14.9000	81.1000	66.2000	-0.4375140	-0.8600569	1.0702233
AH	0.5237	1.0945	0.5708	0.0279583	-0.2727186	0.0074951
HIGH_CO	0.0000	1.0000	1.0000	-0.9861019	-1.0329289	0.0324838
HIGH_NMHC	0.0000	1.0000	1.0000	-13.3067908	176.0326973	0.0054645
HIGH_C6H6	0.0000	1.0000	1.0000	-0.6242595	-1.6183168	0.0338115
HIGH_NOx	0.0000	1.0000	1.0000	-13.6039602	184.0313314	0.0052356
HIGH_NO2	0.0000	1.0000	1.0000	-0.3918179	-1.8561145	0.0355867

	n	mean	sd	median	trimmed	mad
HIGH_T	200	0.275	0.4476348	0.0	0.21875	0.0000
HIGH_RH	200	0.495	0.5012296	0.0	0.49375	0.0000
HIGH_AH	200	0.500	0.5012547	0.5	0.50000	0.7413

	min	max	range	skew	kurtosis	se
HIGH_T	0	1	1	1.0002574	-1.004432	0.0316526
HIGH_RH	0	1	1	0.0198512	-2.009579	0.0354423
HIGH_AH	0	1	1	0.0000000	-2.009975	0.0354441

Table 1: Summary statistics



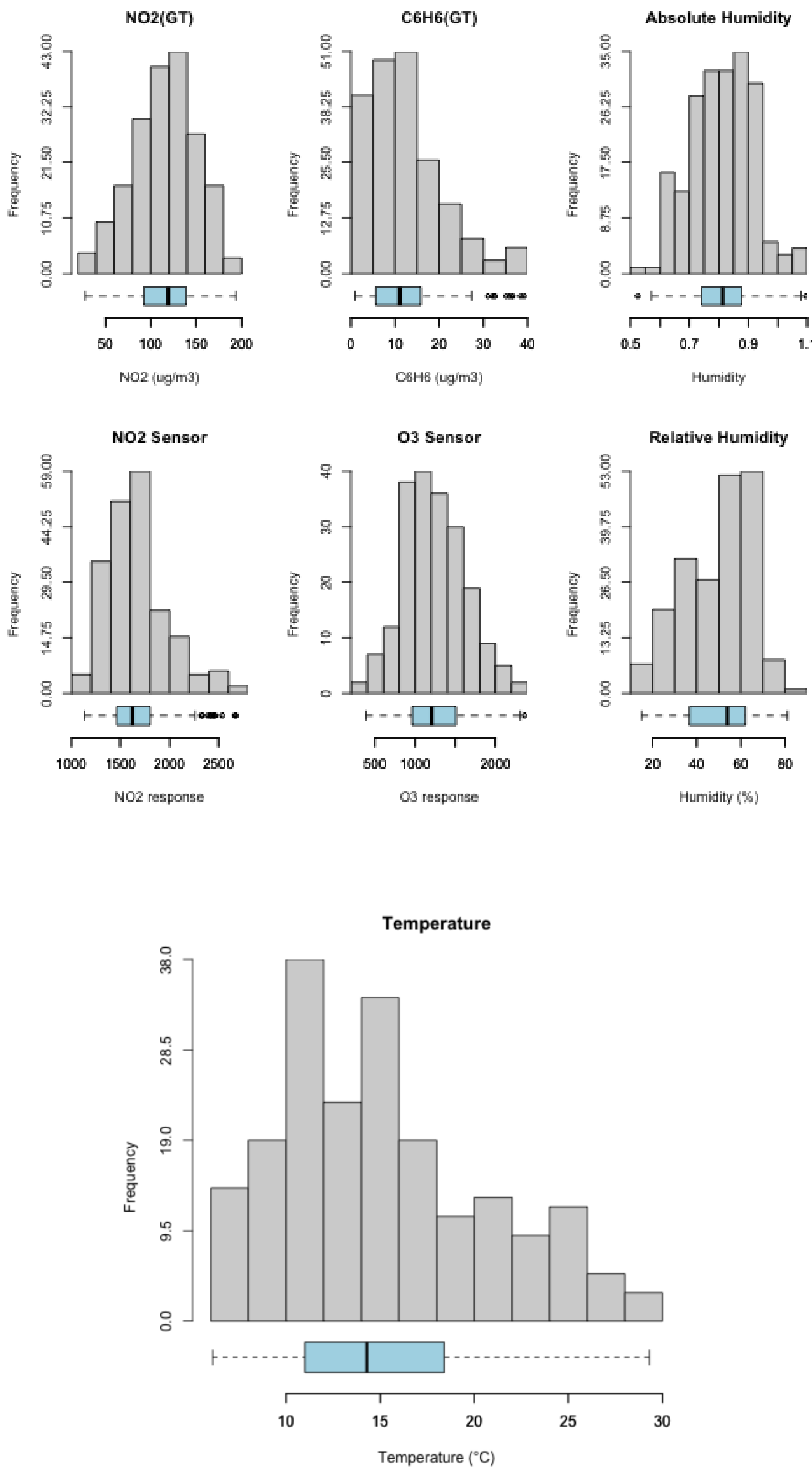


Figure 3: Summary of the data

3.2 Correlation structure of the data

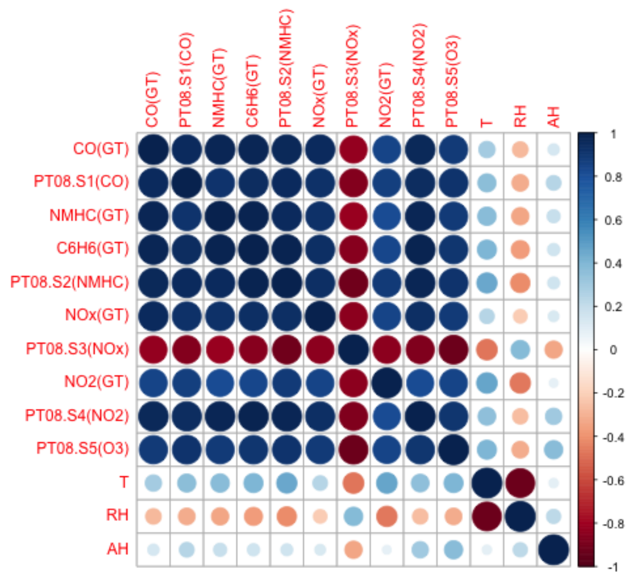


Figure 4: Correlation matrix

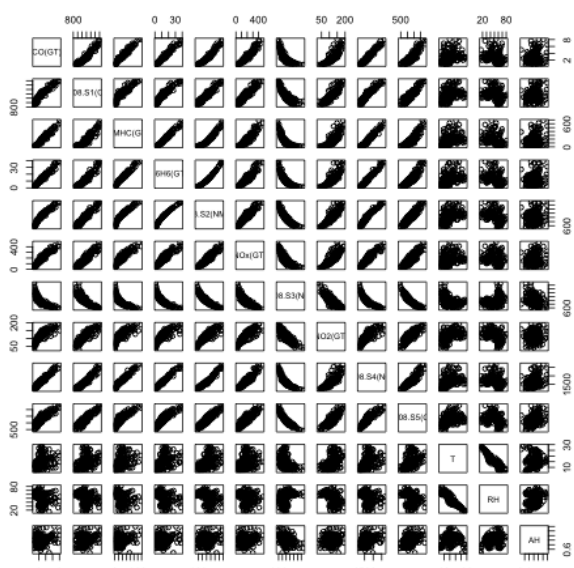


Figure 5: Scatter matrix

As can be seen in the two previous figures, the 10 first variables are correlated all together and the next 2 variables are also correlated together. The absloute humidity is isolated from the rest. This is a first clue that we can use to reduce the dimension of the data. We will use the PCA to reduce the dimension of the data. The only variable collected by the sensors that is not correlated in the same with the others comes from the NOx sensor. Indeed, we can clearly dark red spot in the first figure and a linear relation in the other way on the second.

The variables relative to the humidity and temperature are much less correlated with the sensors varibles but much more within themselves. We can thus clearly see two groups of variables in the correlation matrix.

3.3 Outlying observations using Mahalanobis distance

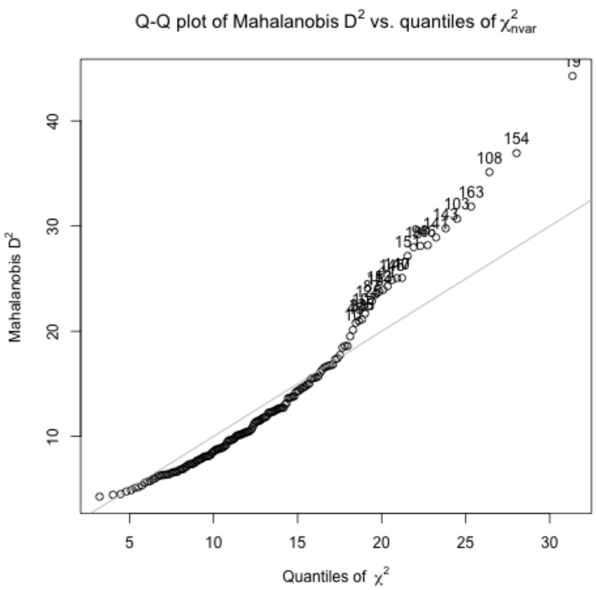


Figure 6: Mahalanobis distance outlier detection

Using the `outlier` function from `pscyh` package, we can clearly distinguish the outliers from the dataset, which are the points that are not close to eachother.

4 Correlation analysis with data reduction

4.1 Choice between PCA and t-SNE

We ave a lot of Correlation in our dataset so the usage of PCA is recommanded. We can see on the scree plot that we have to keep 2 components for our analysis. We can make 2 subsets of our data frame: 1 with the 2 components Temperature and Relative humidity and an other subset with the other components. The Absolute humidity is not used because there isn't any correlation with the other components. If we keep 2 components (1 of each group) we have more than 95% of the variance explained. We can see that the 2 components are not correlated with each other. We can use the 2 components for our analysis.

4.2 2D plot of the data

On the graph below we can see the 2 composents of the PCA. We can say that they are not correlated with each other because the arrows are not aligned.

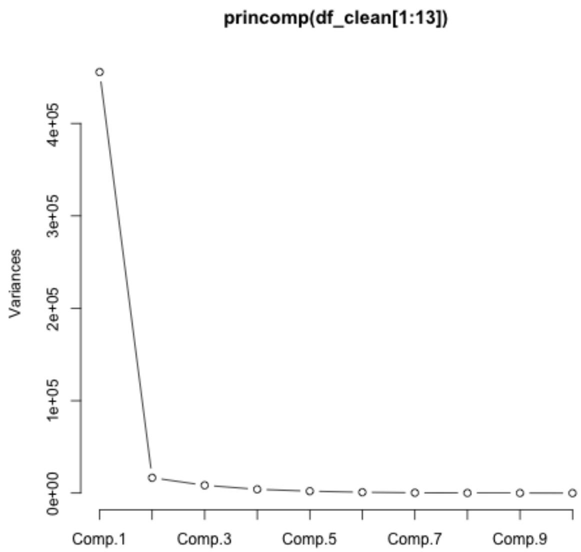


Figure 7: Scree plot of the PCA

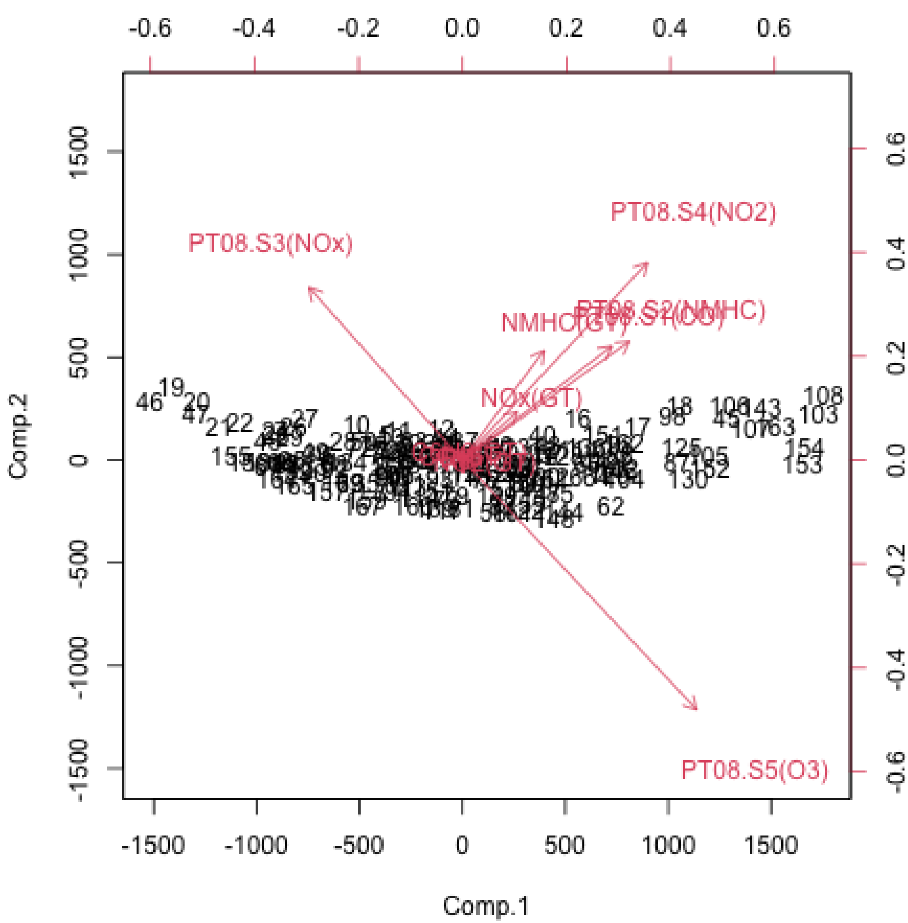


Figure 8: 2D plot of the data