University of Liège

Project 1 - Exploratory data analysis

MATH2021-1 - High-dimensional data analysis

Yann Claes (s161317)
Gaspard Lambrechts (s161826)
François Rozet (s161024)

MSc in Data science and engineering

Academic year 2019-2020

1 Data

The chosen data set describes the evolution of major air pollutants and meteorological variables in 12 air-quality monitoring stations of Beijing. The hourly measured data spread from March 1st, 2013 to February 28th, 2017.

1.1 Variables

The pollutants measured are the concentrations ($\mu g\, m^{-3}$) of $PM_{2.5}$, PM_{10}[1], SO_2, NO_2, CO and O_3. The meteorological variables are the temperature (°C), the pressure (hPa), the dew point temperature (°C), the rain precipitation (mm) and the wind speed ($m\, s^{-1}$). The data set also features the wind direction as compass-like directions (N, S, W, NE, etc.).

With the assigned index, the time described by four variables (year, month, day and hour) and the station name, each line of the data set presents 18 columns/variables including 11 continuous.

1.2 Data handling

First of all, in order to greatly reduce the data quantity, it has been decided to limit the analysis to the measurements of the *Shunyi* station. Still, the number of rows/individuals in the data set (roughly 35 000) was much more important than asked. Therefore, it has been decided to sample randomly (and uniformly) 500 of them. However, for the sake of consistency, a *seed* has been set such that the sample is the same at each execution.

Secondly, text being hardly analyzable mathematically, the compass-like directions have been converted into angles (rad).[2]

Thirdly, the assigned index was replaced by a more genuine measure of time : the unix time stamp (i.e. the total number of seconds since January 1st, 1970).

Finally, as can be seen in the contingency table of `rain` (cf. Table 1), its continuous nature is quite disputable.

`rain` [mm]	0	0.1 to 0.8	1 to 1.6	15
Occurrence	483	12	4	1

Table 1 – Contingency table of the variable `rain`.

Indeed, it seems that the fact it has rained, whatever the quantity, is *abnormal*. Based on that observation, `rain` has been transformed into a binary indicator : `FALSE` if it hasn't rained and `TRUE` if it has.

1.3 Scientific question

Air pollution in cities becomes more and more concerning, especially for children. This data set might help to predict pollution peaks and warn the population about it.

[1] Particle matter of size up to 2.5, respectively, 10 µm.

[2] Unfortunately, this representation opposes 0 to 2π while they actually are equivalent. A possible way to correct that behaviour would have been to replace the angles by their sin and cos in the data set.

2 Missingness

Firstly, a plot displaying the global missingness of the selected data subset has been drawn using the `vis_miss` function, resulting in Figure 1. A total 1.8 percentage of missingness is observed.

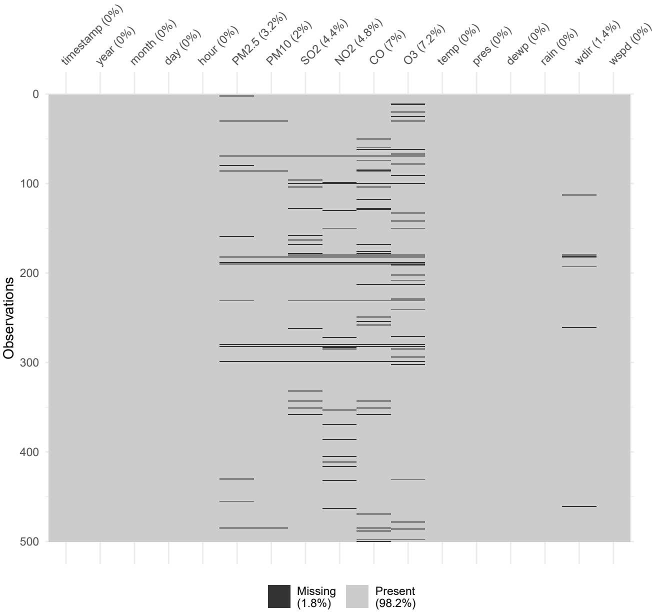


Figure 1 – Missingness in the data.

The missing values are mostly present in the pollutant concentrations, while only a few of them are located in the `wdir` variable.

To observe potential patterns between the missing variables, an upset plot has been derived using the `gg_miss_upset` function from the `naniar` library, leading to Figure 2.

At first glance, no real pattern can be distinguished as most (61) cases involve variables individually. Only 7 rows of the data set have missing values for all the pollutant concentrations. This could be the result of all sensors breaking down at the same time, due to unknown reasons, or of the central system failing to write the measurements to the database.

2.1 Missing data mechanism

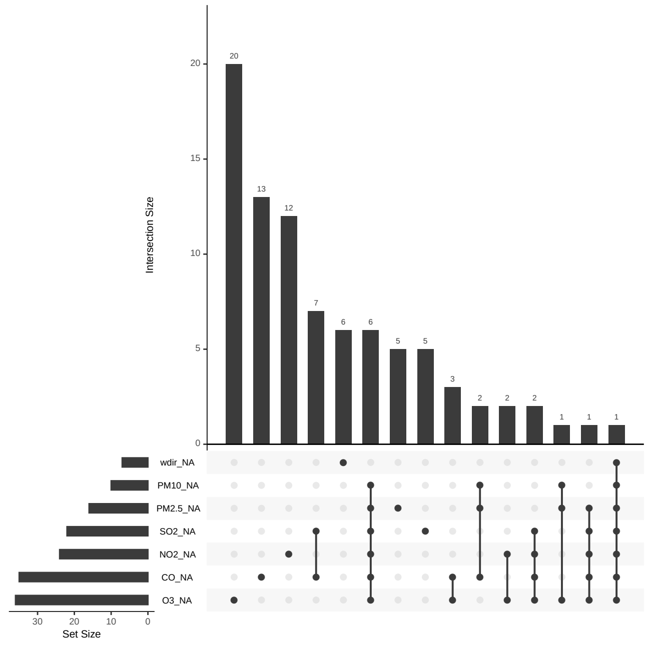


Figure 2 – Upset plot of the missing data.

In order to spot the potential reasons for which a variable could be missing, the following strategy has been used : for each pair of variables X and Y where X has missing values, the z-score of the mean of Y when X is missing has been computed. This z-score gives an insight on the potential influence of the variable Y on the missingness of X.

Indeed, if the missingness of X is independent of Y, m^*, the mean of the n values of Y for which X is missing, should be an unbiased estimator of the actual mean of Y. Therefore, assuming that Y is normal,

$$P\left(\frac{m^* - \mu}{\frac{\sigma}{\sqrt{n}}} \in \left[-u_{\frac{p}{2}}, u_{\frac{p}{2}}\right]\right) = 1 - p\,,$$

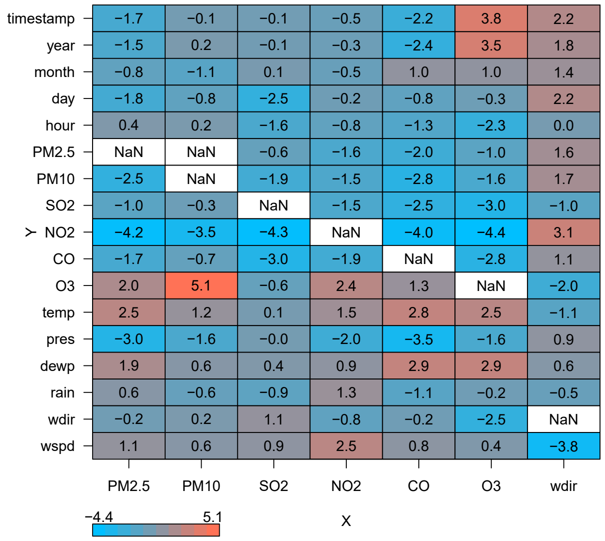
i.e. the z-score of m^* should belong to the confidence interval $\left[-u_{\frac{p}{2}}, u_{\frac{p}{2}}\right]$ with a probability $1-p$, where u_x is such that

$$P\left(\mathcal{Z} > u_x\right) = x, \quad \mathcal{Z} \sim \mathcal{N}(0,1)\,.$$

For example, taking $1-p$ as 95%, the confidence interval becomes $[-1.96, 1.96]$.

Therefore, high (or low) z-scores might be clues to relations between Y and the missingness of X, but only if Y follows a normal distribution. Unfortunately, it is not the case of any pollutants (cf. section 3.1), which might explain such extremes scores one can see in Figure 3.

Furthermore, if most variables present high value z-scores, only a few possible explanations could be drawn :



	PM2.5	PM10	SO2	NO2	CO	O3	wdir
timestamp	−1.7	−0.1	−0.1	−0.5	−2.2	3.8	2.2
year	−1.5	0.2	−0.1	−0.3	−2.4	3.5	1.8
month	−0.8	−1.1	0.1	−0.5	1.0	1.0	1.4
day	−1.8	−0.8	−2.5	−0.2	−0.8	−0.3	2.2
hour	0.4	0.2	−1.6	−0.8	−1.3	−2.3	0.0
PM2.5	NaN	NaN	−0.6	−1.6	−2.0	−1.0	1.6
PM10	−2.5	NaN	−1.9	−1.5	−2.8	−1.6	1.7
SO2	−1.0	−0.3	NaN	−1.5	−2.5	−3.0	−1.0
NO2	−4.2	−3.5	−4.3	NaN	−4.0	−4.4	3.1
CO	−1.7	−0.7	−3.0	−1.9	NaN	−2.8	1.1
O3	2.0	5.1	−0.6	2.4	1.3	NaN	−2.0
temp	2.5	1.2	0.1	1.5	2.8	2.5	−1.1
pres	−3.0	−1.6	−0.0	−2.0	−3.5	−1.6	0.9
dewp	1.9	0.6	0.4	0.9	2.9	2.9	0.6
rain	0.6	−0.6	−0.9	1.3	−1.1	−0.2	−0.5
wdir	−0.2	0.2	1.1	−0.8	−0.2	−2.5	NaN
wspd	1.1	0.6	0.9	2.5	0.8	0.4	−3.8

Figure 3 – Z-score of the mean of Y when X is missing.

Wind speed and direction Although the wind speed does not follow a normal distribution, the observed z-score for Y as the wind speed and X as the wind direction is quite remarkable since a -3.8 z-score is reserved to abnormally low values. Indeed, by looking at the boxplots drawn at Figure 4, it could be deduced that missing values of wind direction are the logical results of a barely present wind, i.e. a wind whose speed is almost null.

Temperature, dew point and pressure As one can see in Figure 3, these three variables present multiple abnormal z-scores. If it is hardly possible to induce any direct link from that observation, it is however not risky to say there probably exists relations between the missingness of pollutants such as CO and O_3 and the temperature since temperature-caused troubleshoot are quite current.

Furthermore, because of the linear correlation between temperature, dew point and pressure (cf. section 3.2) it is perfectly logic to observe such visible correlation between their z-scores. Eventually, it is not possible to conclude anything with certitude about the missing data mechanism. However, it seems almost certain that it is not *MCAR* and *MAR* seems to hold quite well to observations.

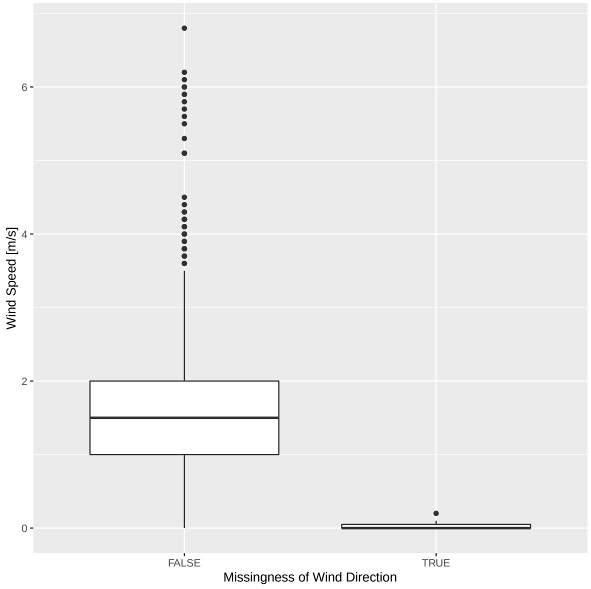


Figure 4 – Wind speed boxplots according to the missingness of the wind direction.

2.2 Dealing with missing data

For the exploratory part, it has been chosen to perform a *complete case analysis*, i.e. to delete all rows containing missing data that is to remove 86 rows. This strategy ensures to analyze a data set without any statistical artefact like those unconditional or conditional imputations would inevitably introduce.

3 Exploratory analysis

3.1 Univariate analysis

First of all, the univariate statistics, listed in Table 2, of relevant variables have been computed using the `summary` function of `R`.

	`PM2.5`	`PM10`	`NO2`	`SO2`	`CO`	`O3`	`temp`	`pres`	`dewp`	`wspd`
Units	$\mu g\,m^{-3}$						°C	hPa	°C	$m\,s^{-1}$
Min	3.0	5.0	1.0	4.0	100	1.0	−13.6	991.7	−25.9	0
Median	58.0	83.5	7.0	46.0	900	34.0	11.75	1014.1	0.35	1.5
Mean	86.5	108.3	15.4	49.6	1295	49.8	11.6	1014.6	1.1	1.7
Max	470.0	485.0	192.0	187.0	7000	293.0	37.9	1035.7	25.0	6.8
Std	82.8	90.0	20.9	31.4	1148	54.4	11.2	9.9	13.1	1.15

Table 2 – Univariate statistics summary.

At first glance, one can notice that the median and mean of the pollutants are quite dissimilar while those of the temperature, pressure and dew point are pretty close. That observation is strengthened by the histograms in Figure 13. Indeed, the pollutants do not seem to follow normal (nor symmetrical) distributions, while the temperature, pressure and dew point seem to. As far as the wind speed is concerned, it seems to follow a Poisson-like distribution.

3.2 Multivariate analysis

In order to get an overall visualization of the potential correlation between some of the variables, a correlation plot has been derived. By sticking to this correlation only, non-linear dependencies between variables could be missed. These other dependencies will be discussed later.

As one can see in Figure 5, there is a positive correlation between all the pollutant concentrations, except ozone (O_3). Indeed, the latter has little anti-correlation with the CO concentration and stronger anti-correlation with the NO_2 concentration.

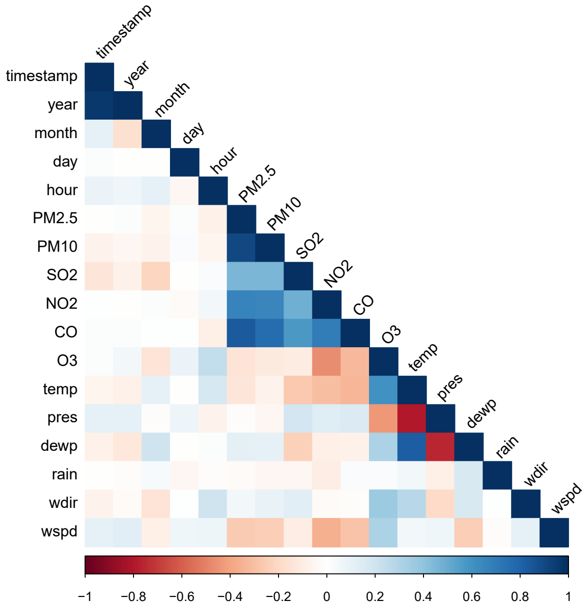


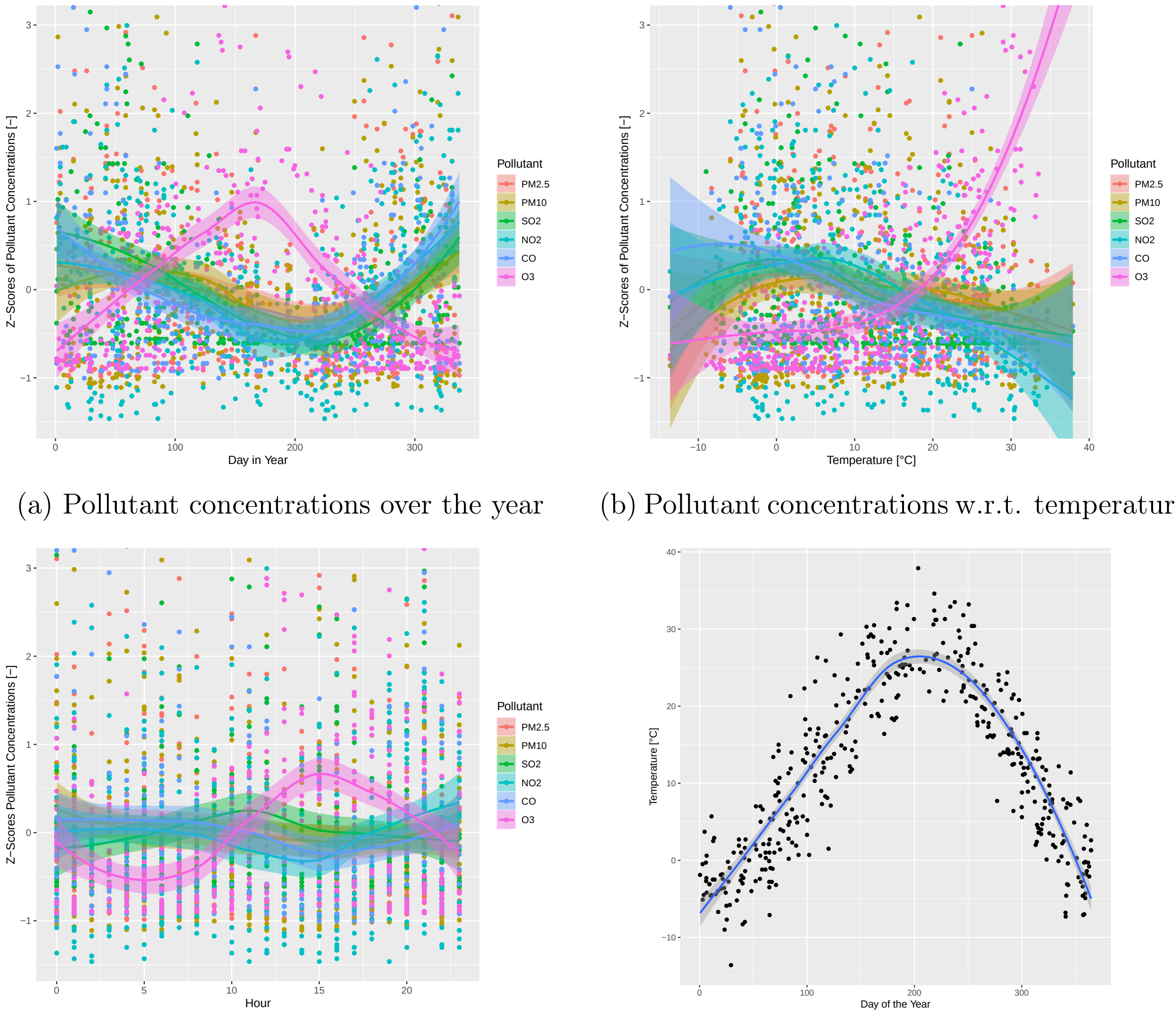
Figure 5 – Correlation plot of the variables.

This *anti*-correlation can be explained by the following chemical equilibrium :

$$NO_2 + O_2 + UV + Heat \rightleftharpoons NO + O_3 \tag{1}$$

This effect is even more visible on the scatter matrix of Figure 14 (where only 100 rows have been sampled in order to make the plot readable).

Also, the pollutant concentrations are slightly dependent on the temperature, especially for NO_2 and O_3. As far as these two pollutants are concerned, it may be explained by the chemical equilibrium given above. This behaviour can be observed in Figure 6b[3].



(a) Pollutant concentrations over the year

(b) Pollutant concentrations w.r.t. temperature

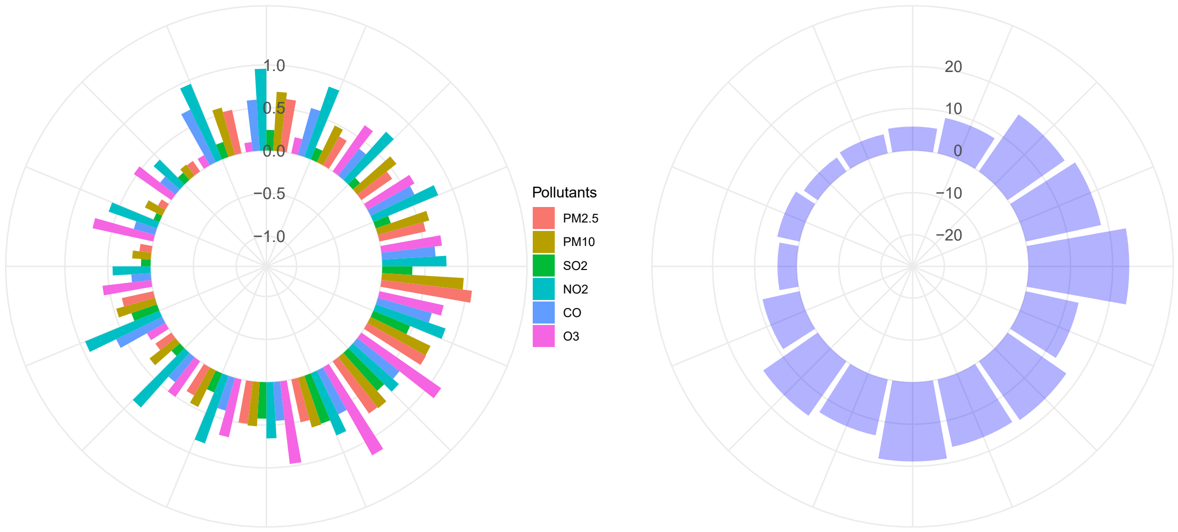
(c) Pollutant concentrations over the day

(d) Temperature evolution over the year

Figure 6 – Relation between temperature, pollutant concentrations and time.

It could suggest that the temperature causes or avoids the apparition of some pollutants, but that could also be due to a confounding factor, such as the period of the year or the wind direction which are, indeed, correlated with the temperature and pollutants as one can observe in Figures 6a, 6d and 12.

[3] A locally estimated scatter plot smoothing estimator with confidence intervals for the conditional mean has been plotted in order to help the eye in seeing patterns in these scatter plots. A drawback about these smoothing curves is that they do not take continuity into account. Indeed, smoothing over a whole year should give a periodic curve.



(a) Median pollutant concentrations w.r.t. wind direction

(b) Median temperatures w.r.t. the wind direction

Figure 7 – Circular plot of the temperature and pollutants w.r.t. the wind direction.

Another remarkable correlation is the one between the temperature and the dew point[4]. The dew point is strongly linearly correlated to the temperature. Actually, that relation is verified by the thermodynamic theory, since

$$T_{dp} = T - \frac{100 - RH}{5} \qquad (2)$$

where T_{dp} is the dew point, T the temperature and RH the relative humidity.

Finally, the pressure is highly anti-correlated to the temperature, yet no convincing explanation could be found.

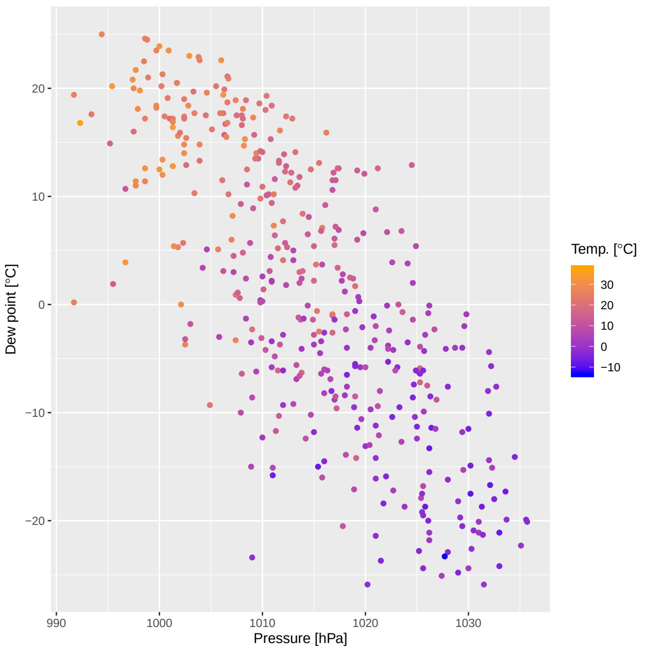


Figure 8 – Evolution of the dew point with respect to temperature and pressure.

3.3 Qualitative analysis

As no qualitative variable could be found in the data set to conduct this analysis, it has been decided to analyze the impact of the binary indicator `rain` on the pollutant concentrations. This yielded the boxplots of Figure 9.

While some distributions might look different, for instance O_3 and NO_2, hasty conclusions should not be drawn, as only 17 rows with positive precipitation are present in the data

[4] The dew point is the minimum temperature to which air must be cooled in order to start condensing water vapor into liquid vapor.

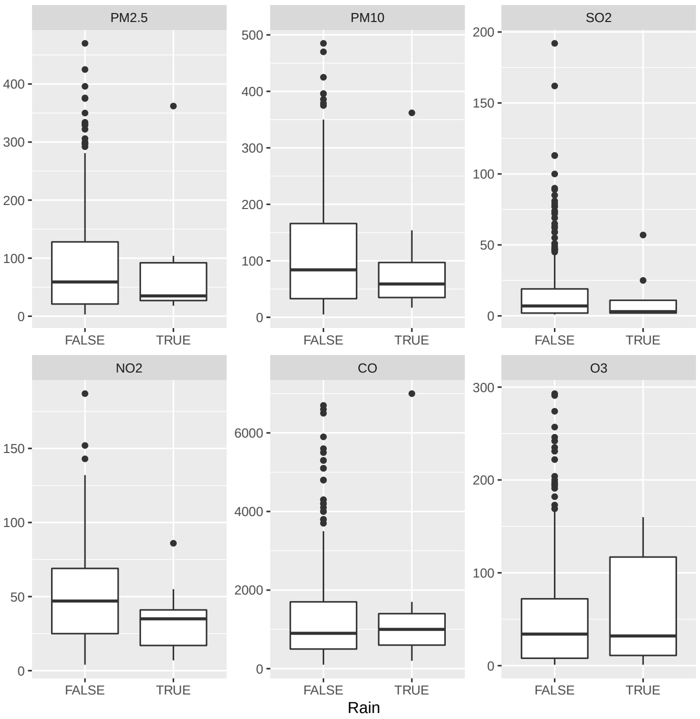


Figure 9 – Effect of the rain on the concentrations of the different pollutants

set. Indeed, there is no way of telling if both distributions would join or diverge if there were more rainy cases.

For most atmospheric variables, the exact same conclusion should be drawn. However, the distribution of temperature and dew point are quite different in the dry and rainy cases. Indeed, Beijing's climate is a *monsoon* climate which confirms the link between higher temperatures and rainy weather. Moreover, knowing equation (2), a rise of temperature and humidity should as well increase the dew point, as boxplots of Figure 10 show.

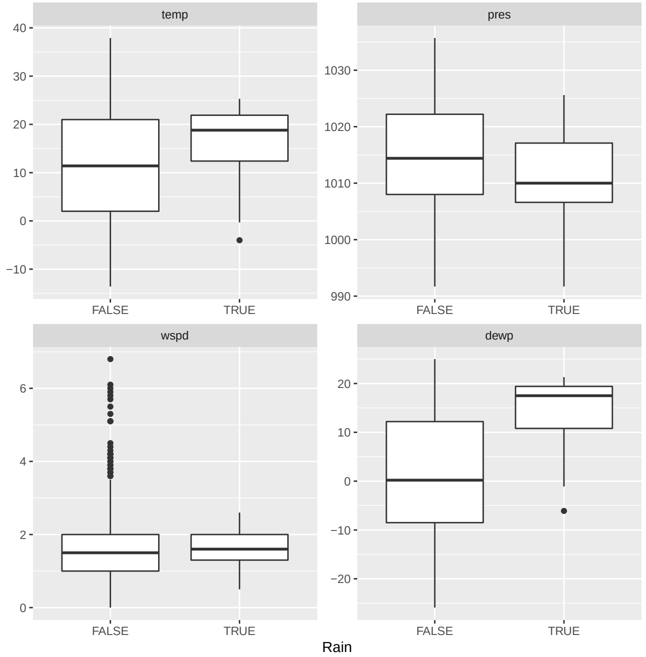


Figure 10 – Effect of the rain on the atmospheric variables

3.4 Outlying observations

A plot gathering the z-score distributions for all the quantitative variables (except for the temporal ones) has been drawn, yielding Figure 11. From this plot, one can see that the assumed normal variables (i.e. `pres`, `temp`, `dewp`) do not seem to have outliers. This is not the case for the remaining variables, which do not follow normal distributions.

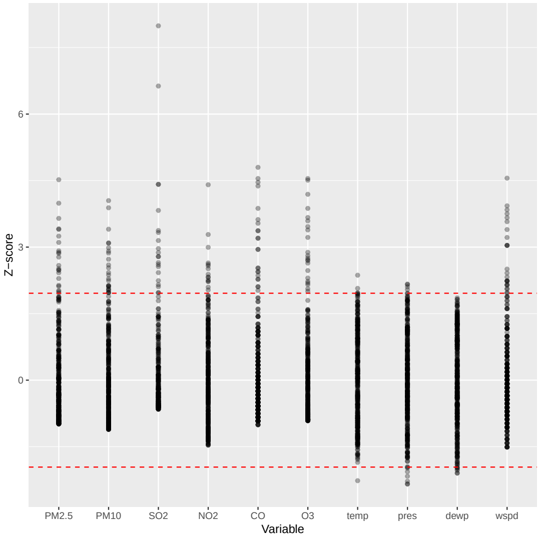
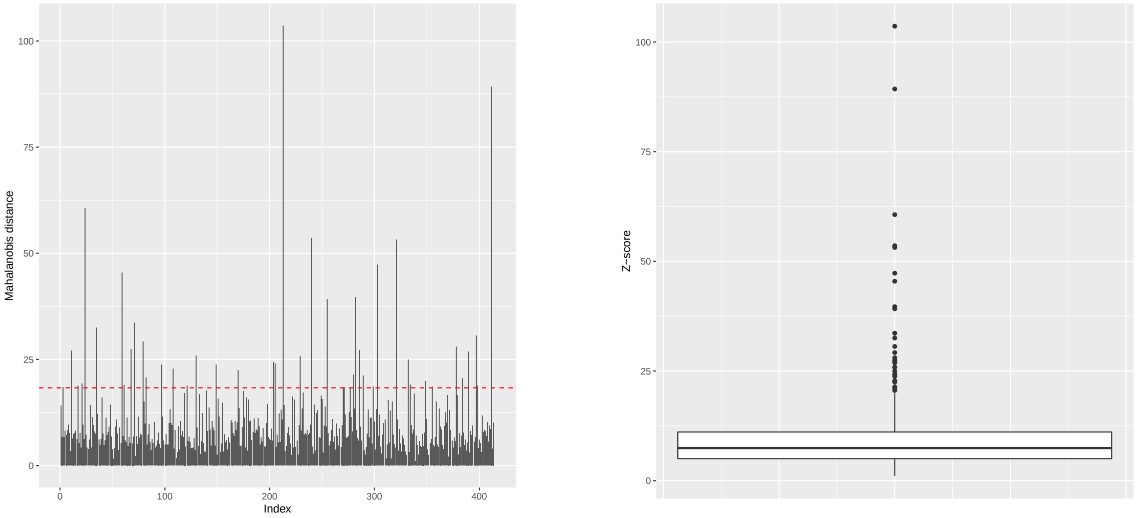


Figure 11 – Z-score distributions

The Mahalanobis distances have been computed through the `mahalanobis` function. Note that omitting the temporal variables may remove some dependencies between the variables, but this has been done to remain coherent with the z-score analysis done above. Results have been plotted in a bar plot, adding a line corresponding to the 95% quantile of the chi-squared distribution, resulting in Figure 12a. As most quantitative variables do not seem to follow normal distributions, interpretations should be relaxed with respect to the observed distances. Indeed, the number of outliers (42) is largely above 5% which confirms that pollutants do not follow normal distributions.



(a) Barplot

(b) Boxplot

Figure 12 – Mahalanobis distance plots.

A Figures

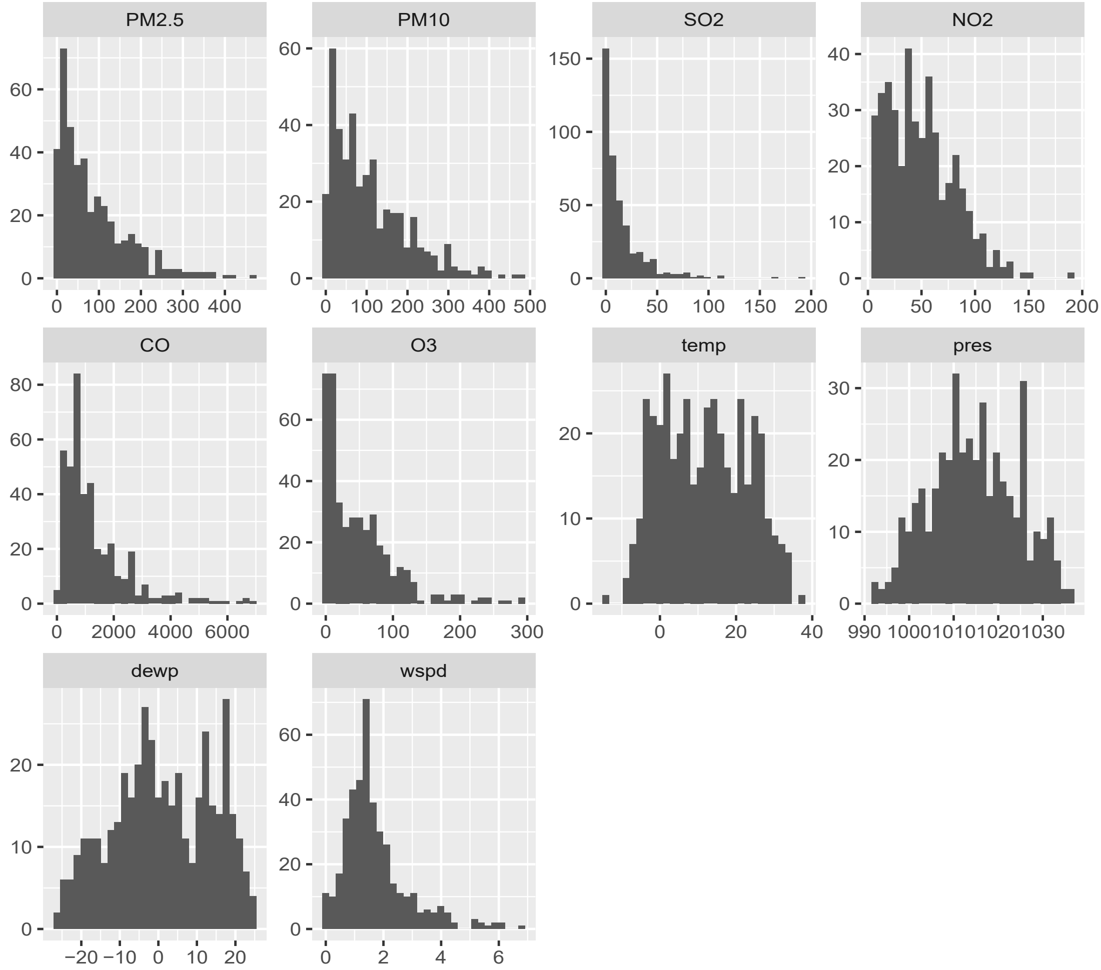


Figure 13 – Histograms of the different variables.

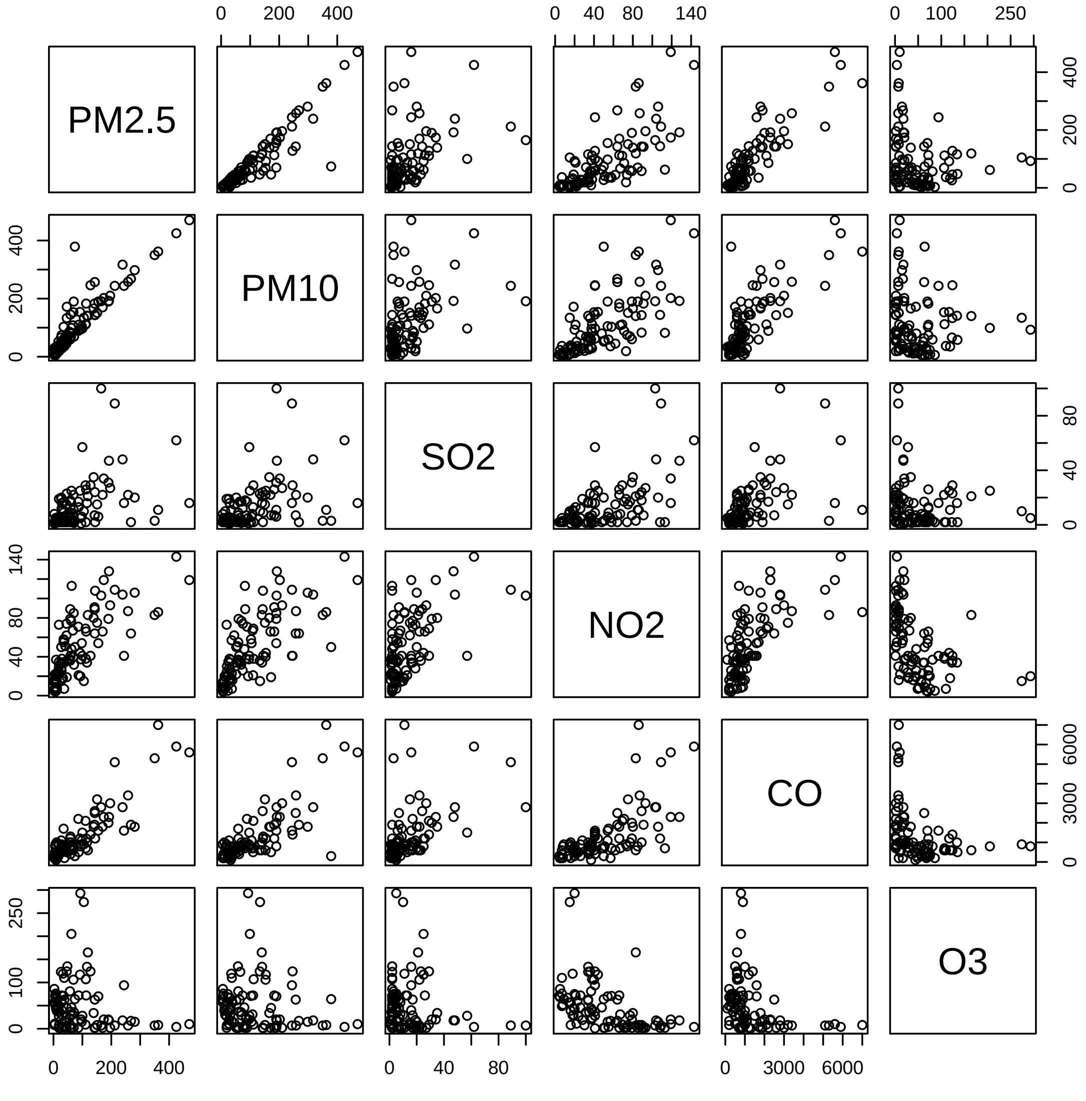


Figure 14 – Scatter plots matrix of the different pollutant concentrations.