# BINOMIAL_30_MAKAROFF_PAVARD

*Nicolas Makaroff & Mathieu Pavard*

*22/11/2019*

With this study, we aim to provide an insight on the atmosphere quality in Beijing, capital of China. The air-quality is said the improve by officials. The air-quality is monitored with $PM_{2.5}$ which is the level of micro-particule under $2.5\mu$ meter. We will look into the data and use our solution to argue weither or not inhabitants of the megapole can be reasured that there is a real improvement. There exists a solution to this problem given by chinese official where the air-quality is said to have improve in year 2016. We are looking at the concentration of $PM_{2.5}$ in the air. Mesuring the performance will be based on an R-squared error named accuracy. We are facing a supervised problem. As help to conduct our study, we will refer to the paper "Cautionary tales on air-quality improvement in Beijing" [1] where information on the nature of the data set is given with intel about meteorological insight. The data is provided by the UCI machine learning repository. The data is of type time-serie and multivariate. We have 12 data set from different sites in Beijing where the data provided runs from March the $1^{st}$ 2013 to February the $28^{th}$ 2017. 17 covariates/attributes are available. Let's start diving into the data. The data follows the same structure for each file available. Therefore, in order to gain insights on the data, we will consider as exemple only the Aotizhonxin site.

We transformed the data to ease the statistic description. We dropped the *No* and *station* columns. Following [**1**], we changed the year attributes with seasonal categories.

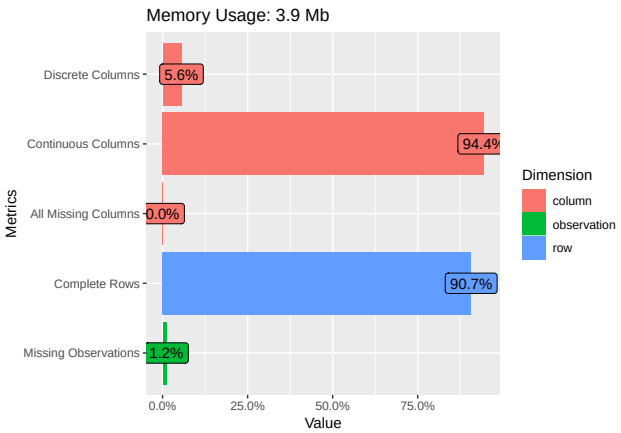


Figure 1: Description of the data of the Aotizhongxin data set

The data we were provided contain many types. Most of them are floats or integer but we also have categorical value type which we will have to convert as convey the figure **1**. For every covariates that are monitored, we are facing missing values $NA$. For our variable of interest, it is 2.6 % of all values available for the station Aotizhongxin.

Most of our concern on this first dive is to deal with NA values and how to handle them. The figure **2** shows the percentage of missing values for each attribute. Logically, none of the time and date related variables are missing but some of the measured ones are. Our value of interest PM2.5 misses 2.64 % of its values. As it will be necessary to use wind direction attribute, temperature, atmospheric pressure, rain, wind speed, relative humidity, we also have to be concerned about their missing values.

CO missing values rate is very high and goes beyond the common 5% threshold. We will probably have to leave that attribute.

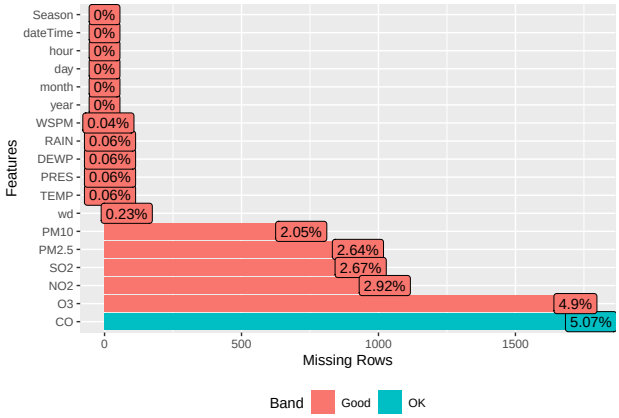


Figure 2: Missing values distribution for Aotizhongxin data

To remove categorical data, we encoded non-numeric values into integers.

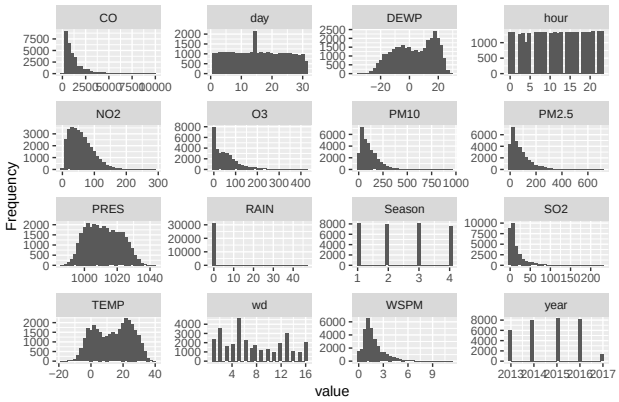


Figure 3: Distribution of all variables

The figure **3** gives us intell about the interpretability of the variables. For each variables we plotted the frequency of appearence of the possible values related. Continuous attributes are more or less interpretable. The more they look like a gaussian distribution the better it is.

These attributes have very different scales. We will probably have to feature scale the covariates. Finally, many histograms are tail heavy. This means that they extend farther to the right of the median than to the left. This might make the learning process harder. It is the case for the $PM_{2.5}$ attribute.

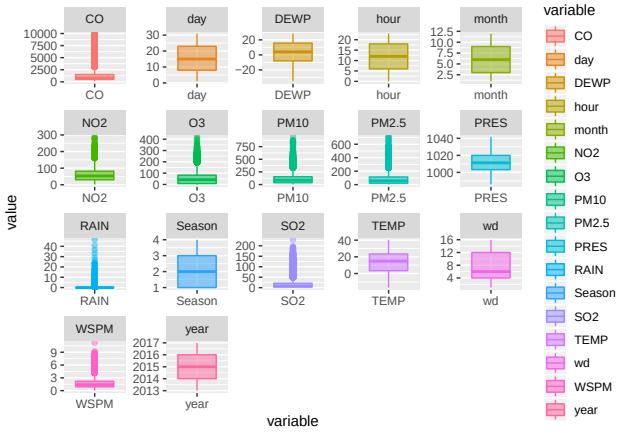


Figure 4: Boxplot of each variables

We should also take a look at the outliers. We use a boxplot (figure **4**) were outliers are represented by dash. They correspond to values that are over the third quartile added to 1.5 times the inter-quartile interval or inferior than the first quartile plus 1.5 times the inter-quartile interval.

Finally, we compute the standard correlation coefficient between every pair of attributes (figure **5**). Following [1], we know that the concentration of $PM_{2.5}$ is related to climate condition. We must conclude that the relation between the two isn't linear correlated as figure **5** shows none. It is naturally missing nonlinear relationship.

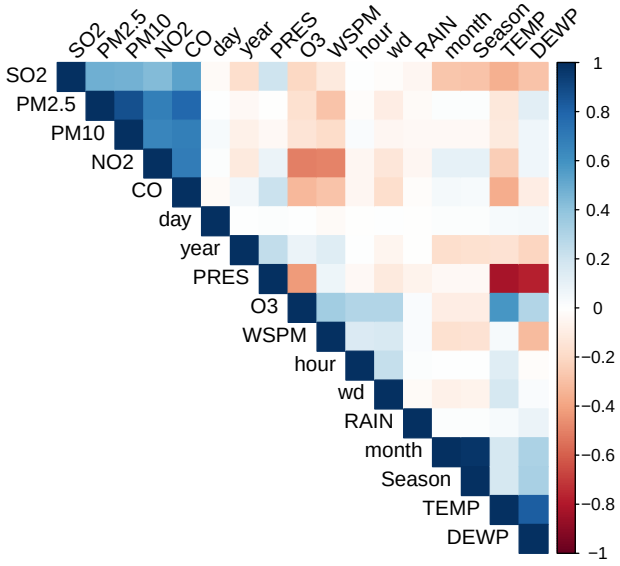


Figure 5: Standard correlation coefficient for every pair of attributes

Figure **6** shows how the season influence the level of $PM_{2.5}$ in all Beijing and to which extent the climate condition that are totally different between two season. For this figure we used a simple average.

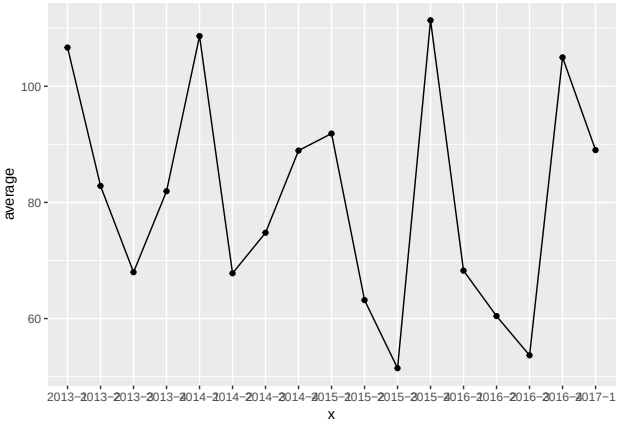


Figure 6: PM2.5 AVG per season in Beijing

[**1**] Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. and Chen, S.X. (2017) Cautionary Tales on Air-Quality Improvement in Beijing. Proceedings of the Royal Society A, Volume 473, No. 2205, Pages 20170457.