# Gas sensor array temperature modulation

*Mame Diarra Toure-Imane Alla: Binome 20*

*11/4/2019*

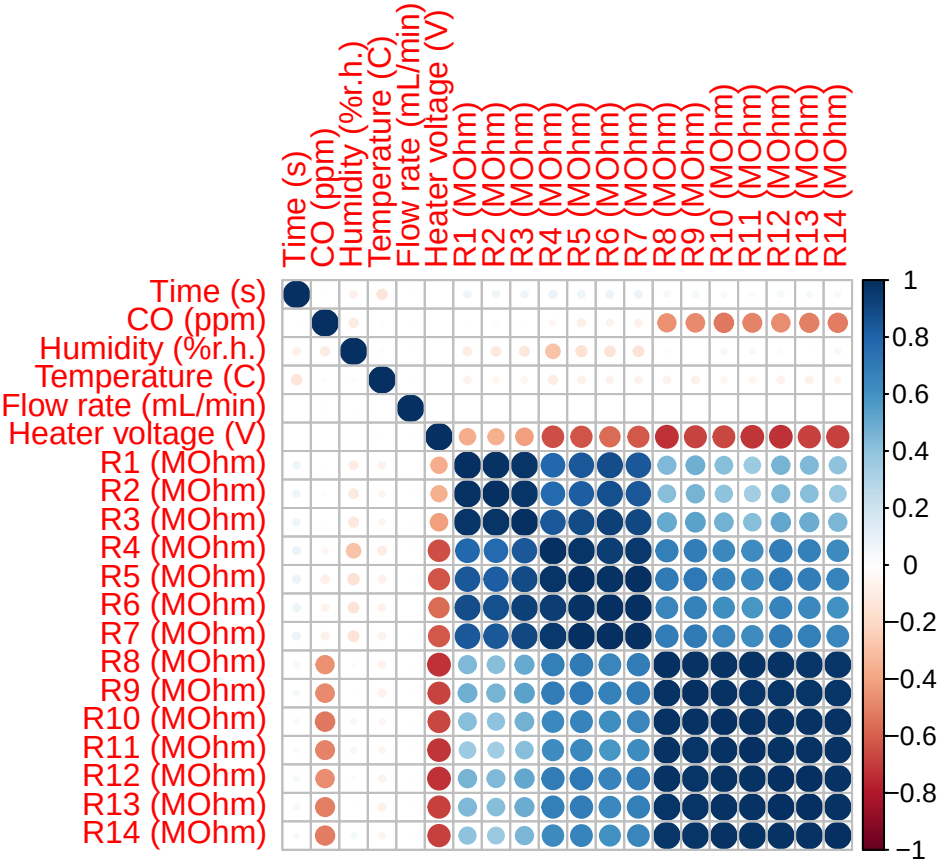
## Introduction

Poor air quality has become a global concern. No matter who we are, where we live or the state of our health, the quality of the air we breathe each day affects us. Even when we can't see it or smell it, air pollution can still be a threat. There are many different types of air pollutants – particulate (PM2.5, PM10), greenhouse gases (CO2, CH4) and toxic gases (Volatile Organic Compounds (VOCs), CO, NOx, SOx, H2S). Metal oxide semiconductor gas sensors are utilised in a variety of different roles and industries. They are relatively inexpensive compared to other sensing technologies, robust, lightweight, long lasting and benefit from high material sensitivity and quick response times. They have been used extensively to measure and monitor trace amounts of environmentally important gases such as carbon monoxide and nitrogen dioxide.

## Description of the dataset

Our data sets contains about 4 millions observations of the response of 14 Mox gaz sensor when exposed to differents concentrations of carbon monoxide(CO), humidity, temparatures, heat voltage and flow rate. The dataset is presented in 13 text files, where each file corresponds to a different measurement day. The experiment was to study the response of the sensors to different stimulis. So they wanted to determine which sensors should be trusted by comparing their response to different concentrations. So we are going to study the response of thoses sensors. We combined all measurements in one table in order to study our dataset. We then plot the correlation matrix to have a first glimpse of our dataset.

## Correlation matrix



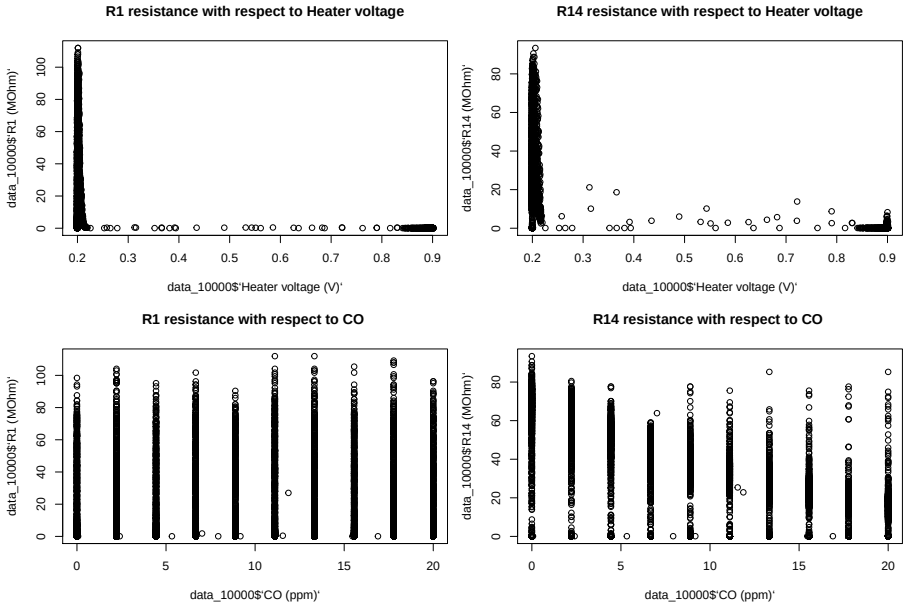

We see in this corrplot that the variables R1,. . . .,R14 are highly correlated. To be more specific the variables R1,. . . .,R7 are strongly correlated to each other but less correlated with the variables R8,. . . .,R14. And in opposition the variables R8,. . . .,R14 are strongly correlated to each other but less correlated with the variables R1,. . . .,R7. We also see that the heater voltage is negatively correlated with R1,. . . ,R14

(in an increasing way). Adding to that the CO concentration is negatively correlated with the variables R8,...R14.Hence, we except that we will need to proceed to variable selection in order to have a good model

## Some plots

To visualize more the distinction between the two groups , let's take one from each and compute the scatter plot with heater voltage and CO concentration.

```
data_10000 <-dat_csv[sample(nrow(dat_csv),size=10000),]
```



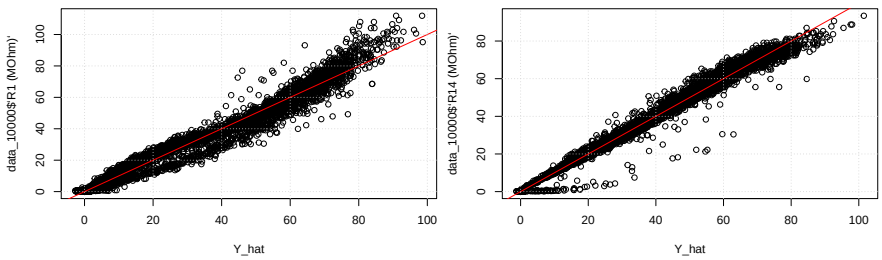

As we see there are more variation in the plot of R14 according to heater voltage than there is in the plot of R1 according to heater volatge. Thats understandable since heater is more correlated to R14 than it is to R1 The same conclusion can be drawn with the plot of R14 and R1 according to CO concentration.

## Regression problem

As we've see it previously with the corrplot the variables R1,....,R14 are highly correlated and there are also strongly correlated with the heater voltage. However only R8,...,R14 are correlated (negatively ) with the CO concentration. So the problem here is to decide whch variable From R1,..,R14 should be our target variable. Indeed we see that the sensors can be divided into 2 groups: Group 1 R1,...,R7 and group 2 R8,...,R14 The group 2 is the one correlated with the CO concentration and the heater voltage So should we just pick on in the 14 variables to be our target variables or do a multiple target regression?

## target variable

For starters we are going to pick randomly a target variable from the group 1. Using that target variable we compute a simple linear regression which consider to be our basline model Then we do the same with a target variable of the second group. We are going to use a random sample of 10000 observations to compute the linear regressions




So we see that the regression with R14 as a target variable has a a better adjusted rsquare and a smaller error than the model with R1 as a target variable We then decided to use R14 as our target variable. Our baseline model will be used to compare the goodness of our future more complex models. Indeed more complex models should give us a better results than our baseline model.