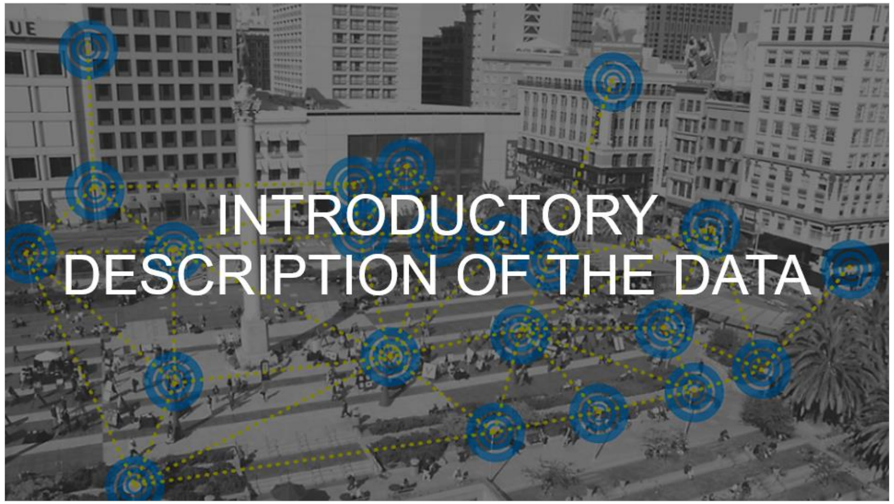
Predicting Air Pollution with Machine Learning

Gabriele Pinto
MADAS – Collegio Carlo Alberto

OUTLINE


INTRODUCTORY
DESCRIPTION OF THE DATA

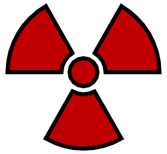

CONTEXT AND
OBJECTIVES


DATA PRE-PROCESSING

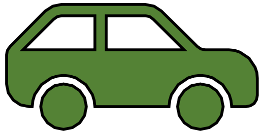

RESULTS

INTRODUCTORY DESCRIPTION OF THE DATA

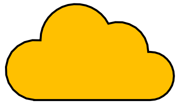
What data we have



- 41,288 (~40k) observations from pollution sensors (daily and hourly)



- 9,183,475 (~9 million) observations from traffic sensors (with timestamp)



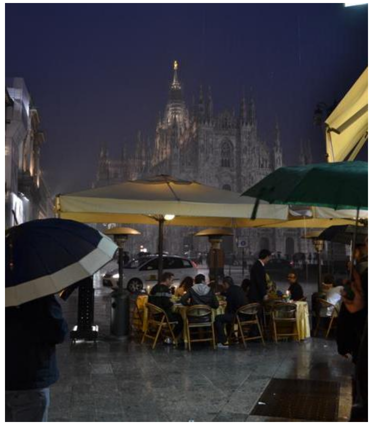
- 39, 167 (~40k) observations from weather sensors (hourly)

Where and When ?

- MILAN (Italy)
- 2 Months (between 01 Nov 2013 and 31 Dec 2013)

Weather Sensors

- **Features**
  - Unit measurement
  - Value of the measurements
  - Location



Sensor_type	Number of sensors
Atmospheric Pressure	1
Global Radiation	1
Net Radiation	1
Precipitation	4
Relative Humidity	5
Temperature	6
Wind Direction	6
Wind Speed	6

Traffic Sensors

- **Features:**
- Detailed information on type of vehicle coming from the matching of the palette with the “motorizzazione civile” dataset:
  - Size of the vechicle (width)
  - Engine (diesel, oil, electric…etc)
  - Environmental class (euro 0,1 ..5 etc)
  - Location

	Number of sensors
Sensor type	
Pedaggio	36
Tpl	6

Pollution sensors

- **Features**
  - Unit measurement
  - Value of the measurements
  - Location

	Number of sensors
Sensor_type	
Ammonia	1
Benzene	4
BlackCarbon	2
Carbon Monoxide	4
Nitrogene Dioxide	8
Ozone	1
Ozono	2
PM10 (SM2005)	3
PM2.5	2
Sulfur Dioxide	1
Total Nitrogen	8

CONTEXT AND OBJECTIVES

PROBLEM

Pollution sensors are costly!

And not real- time available.

While traffic and weather data are much less costly and are already adopted for other uses.

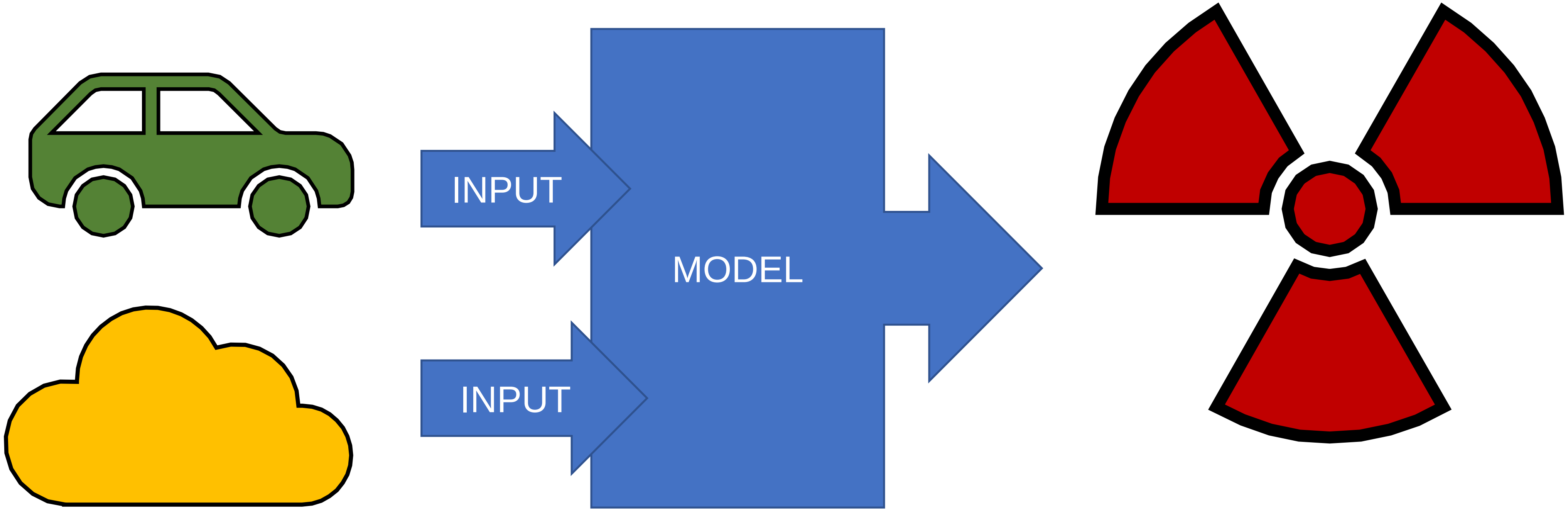
Thus, it might be a good idea to use traffic and weather sensors to predict pollution…..

Furthermore…..

Pollution predictions could be useful for..

- Inform policymakers to:
  - Adopt preventive actions
  - Policy planning

  and……
- Make **better informed decisions for** (hopefully) **Greener city!**



Objective of the modelling task:

Using Weather and Traffic Sensors to predict Pollution !

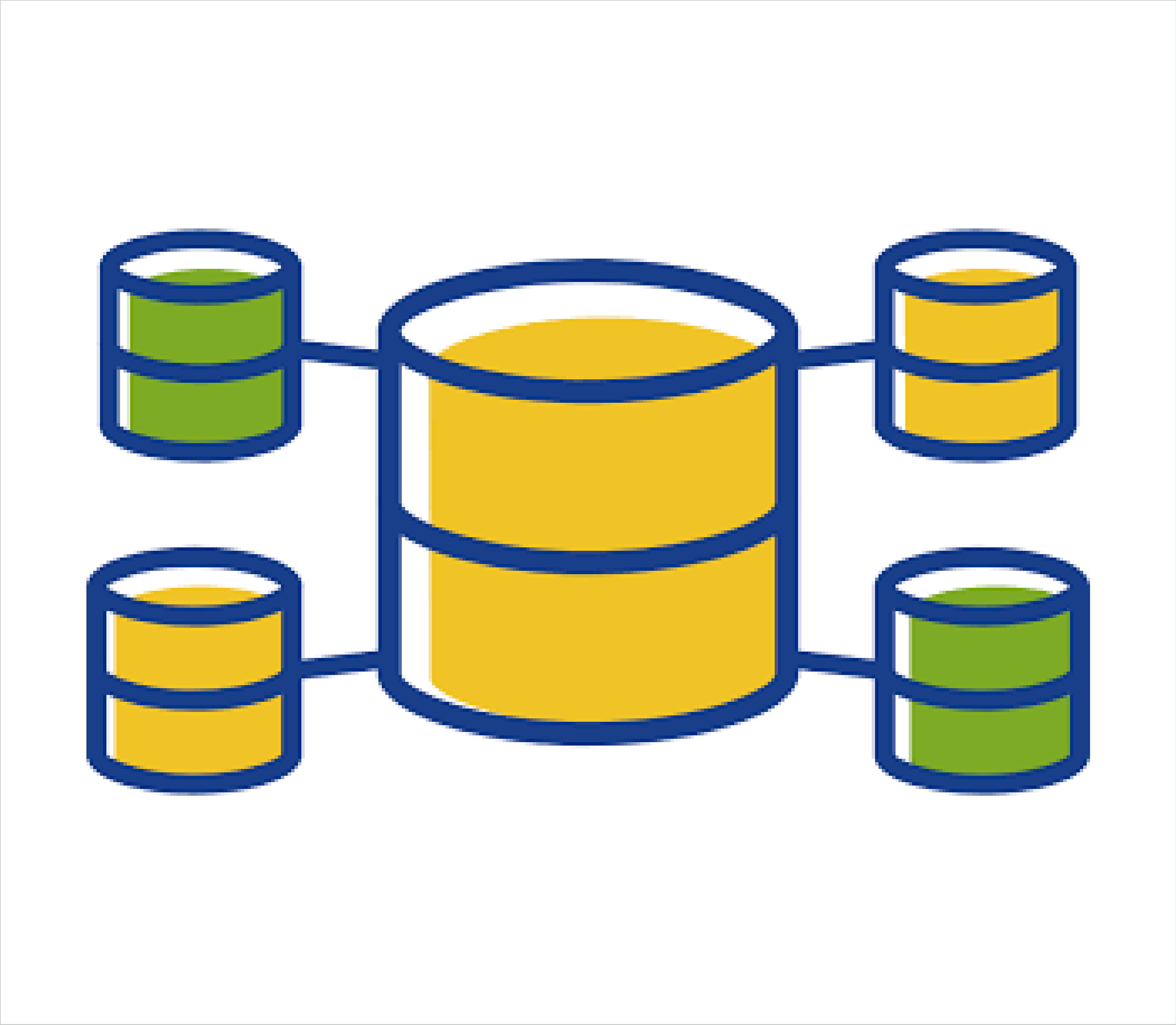
DATA PRE-PROCESSING

Data aggregation

Problem is that our data are heterogeneous in terms of **time** and **space.**

As a solution we aggregate all the data on <u>hourly observations</u>

While we average sensors based on their type (for pollution and weather only). And we add up information from traffic sensors.

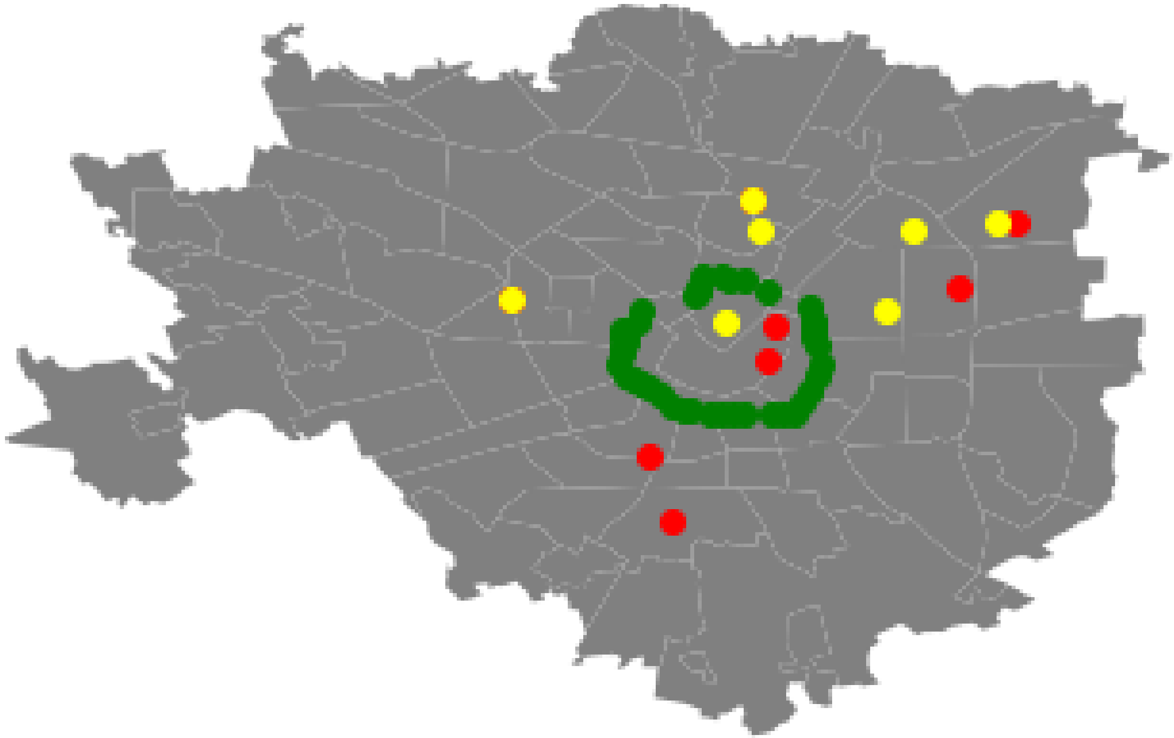


Sensors Location

Pollution Sensors

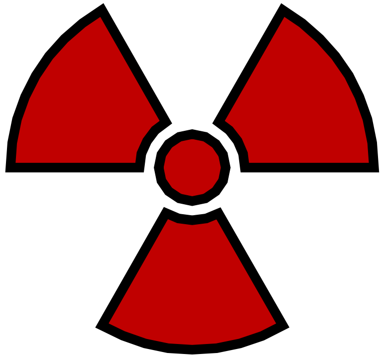
Traffic Sensors

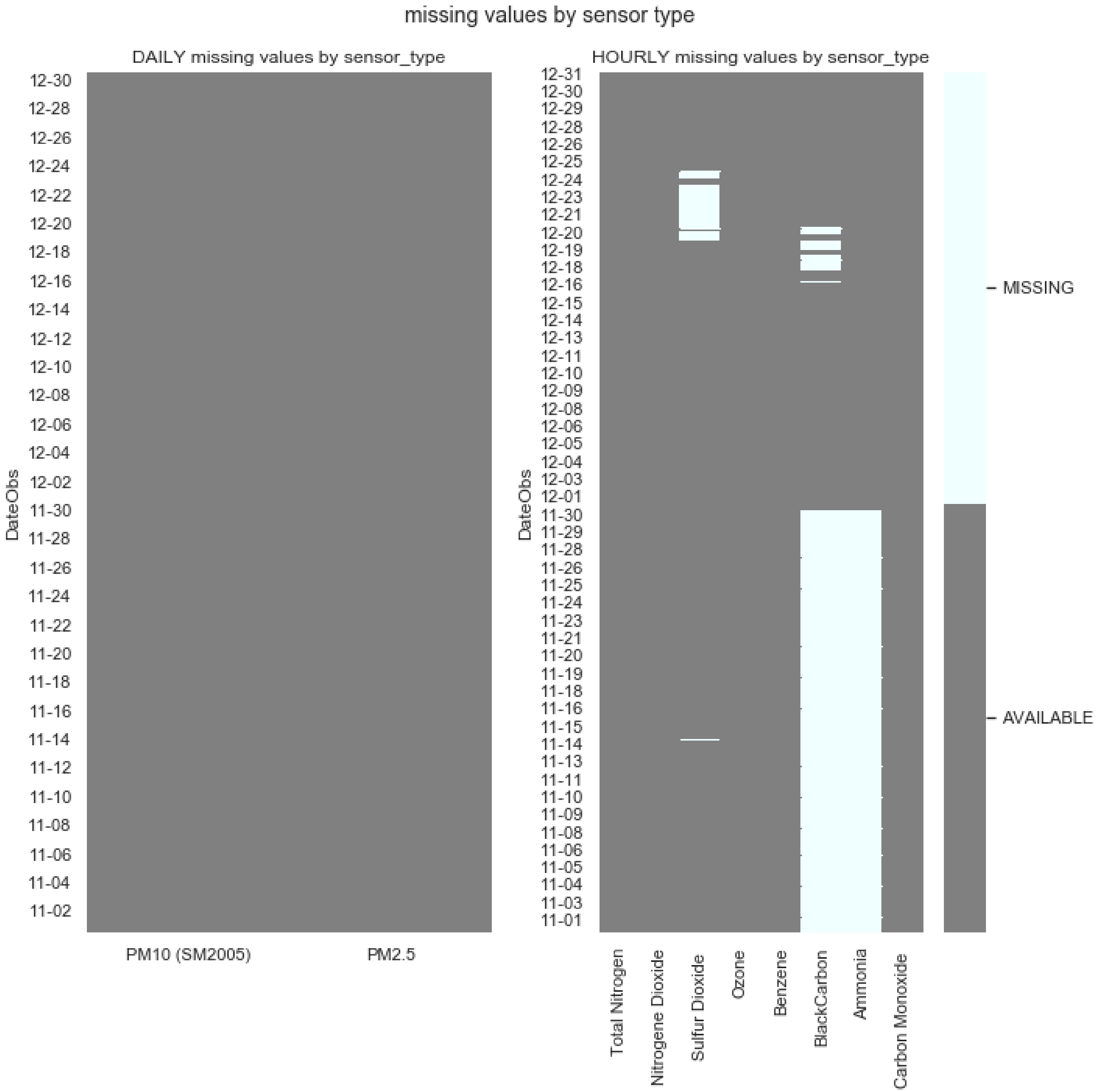
Weather Sensors



Data imputation

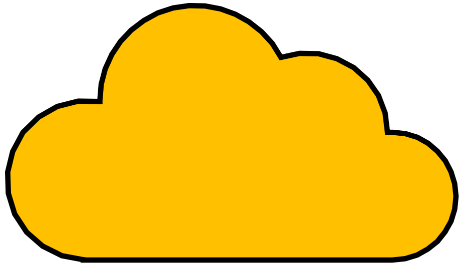
We have several missing values in the pollution dataset we decide to impute using random forest.

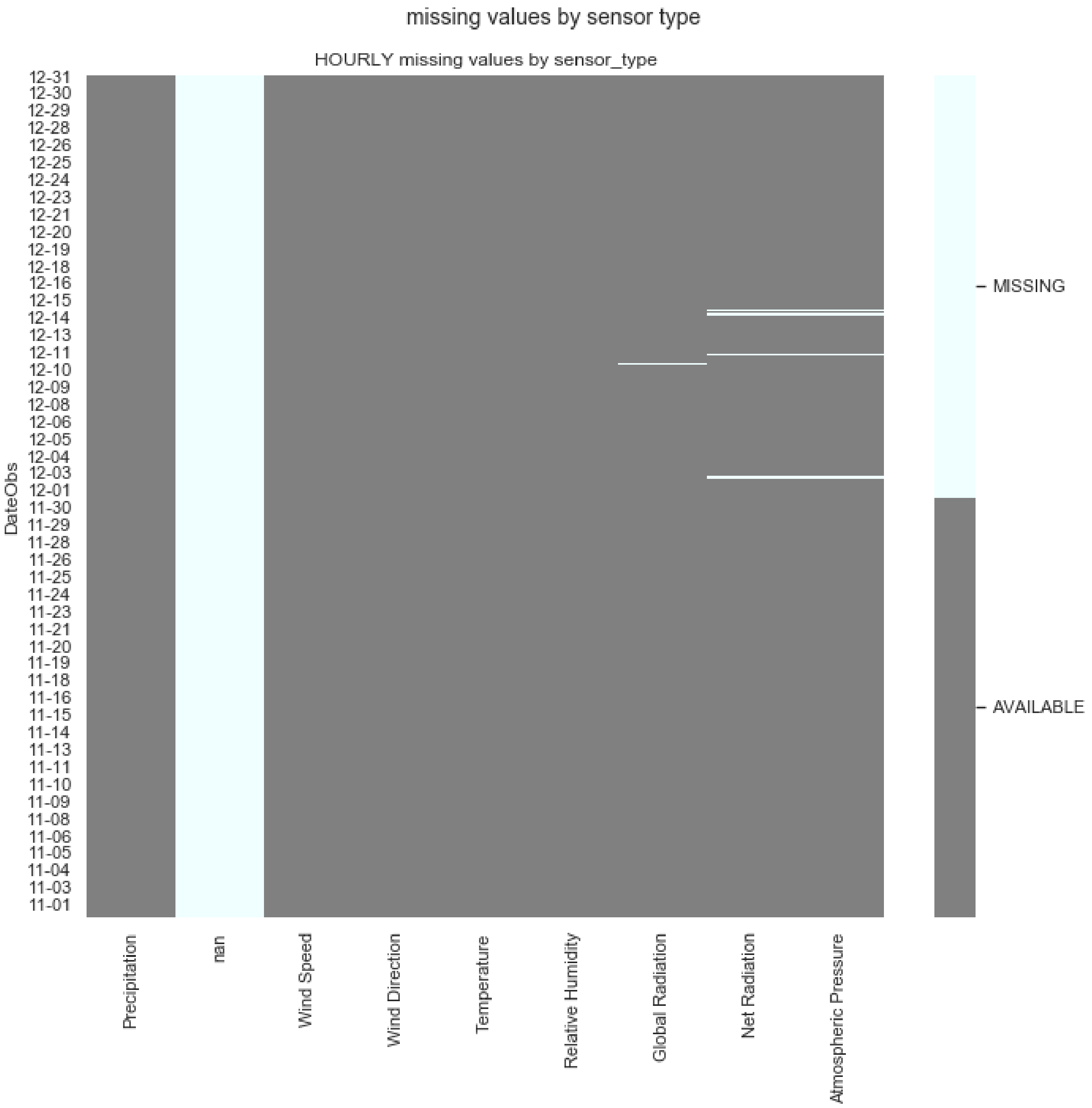




Data imputation

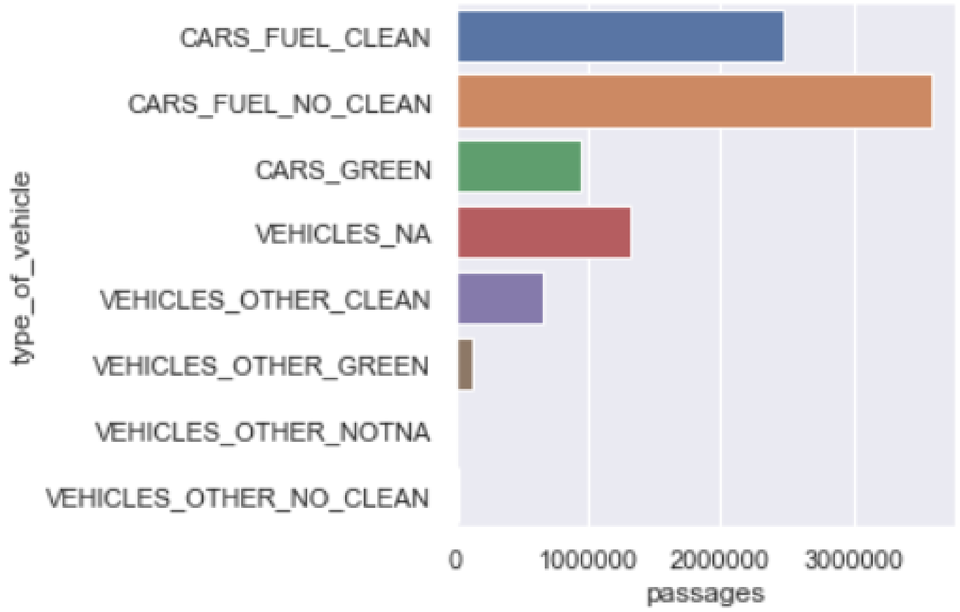
The weather dataset has only few missing values we impute as well using random forest. We also have a sensor for which no obs neither details are available (we decide to drop it from the dataset).





Features Selection in the Traffic Dataset

Grouping type of cars



- "VEHICLES_NA" the not available
- "CARS_FUEL_NO_CLEAN" Vtype==4 and EURO >=1 <=4" the fuel car EURO 0-EURO 5
- "CARS_FUEL_CLEAN" Vtype==4 and EURO >=5 " the fuel car over EURO 4
- "CARS_GREEN" Vtype==4 and EURO >4 " non-fuel car (electric, hybrid...etc...)
- "VEHICLES_OTHER_NO_CLEAN" "Vtype!= 0 and 4 and less than EURO 3"
- "VEHICLES_OTHER_CLEAN" "Vtype!= 0 and 4 and more than EURO 3"
- "VEHICLES_OTHER_GREEN" "Vtype!= 0 and non-diesel/fuel" (electric.hybrid etc...)
- "VEHICLES_OTHER_NOTNA" vehicles for which we have some information but do not enter in any of the above groups

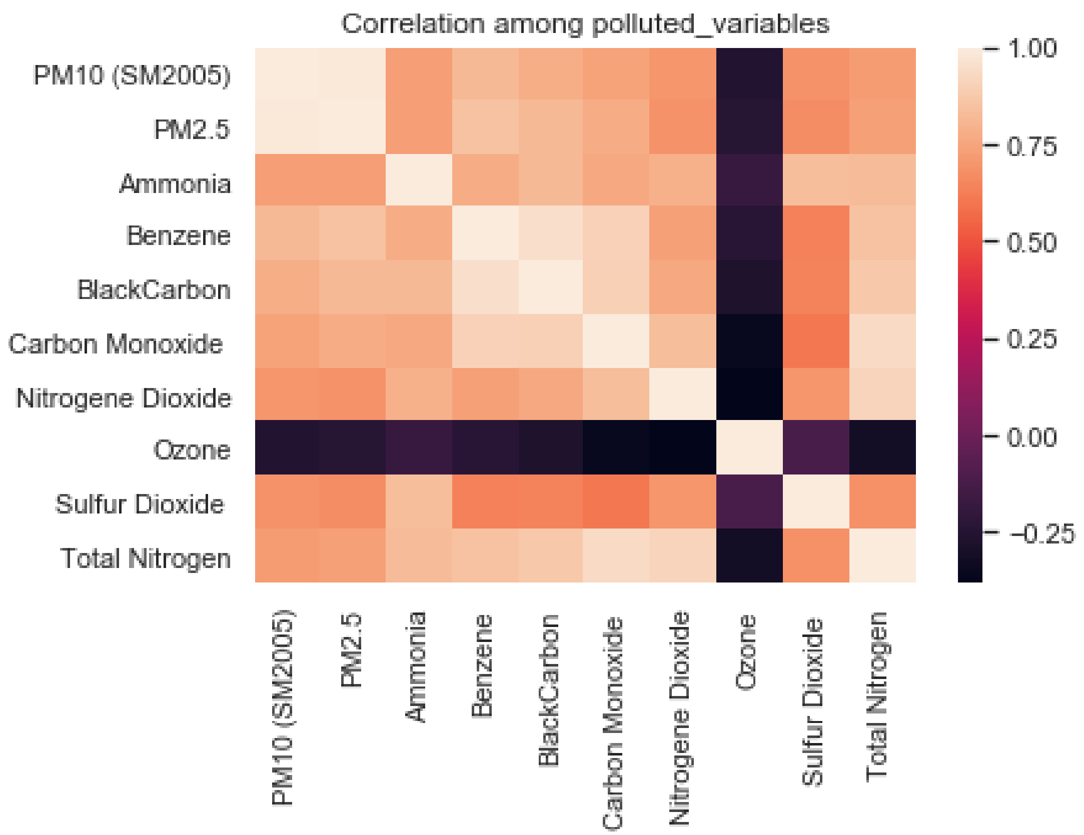
Other Features enginnering

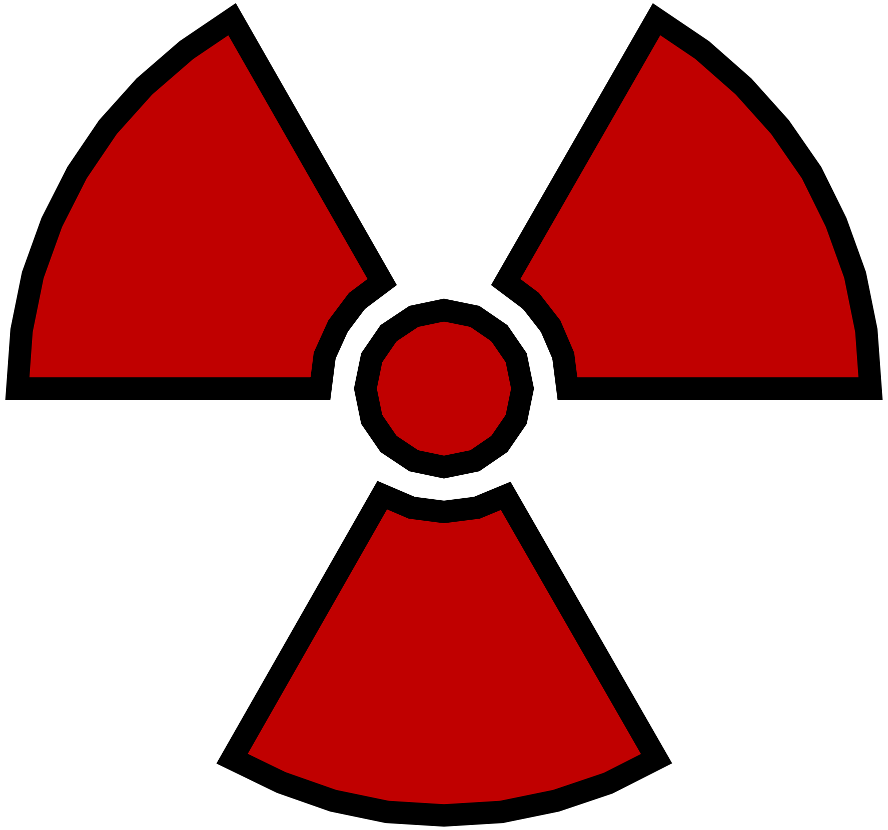
We end up with 18 Features

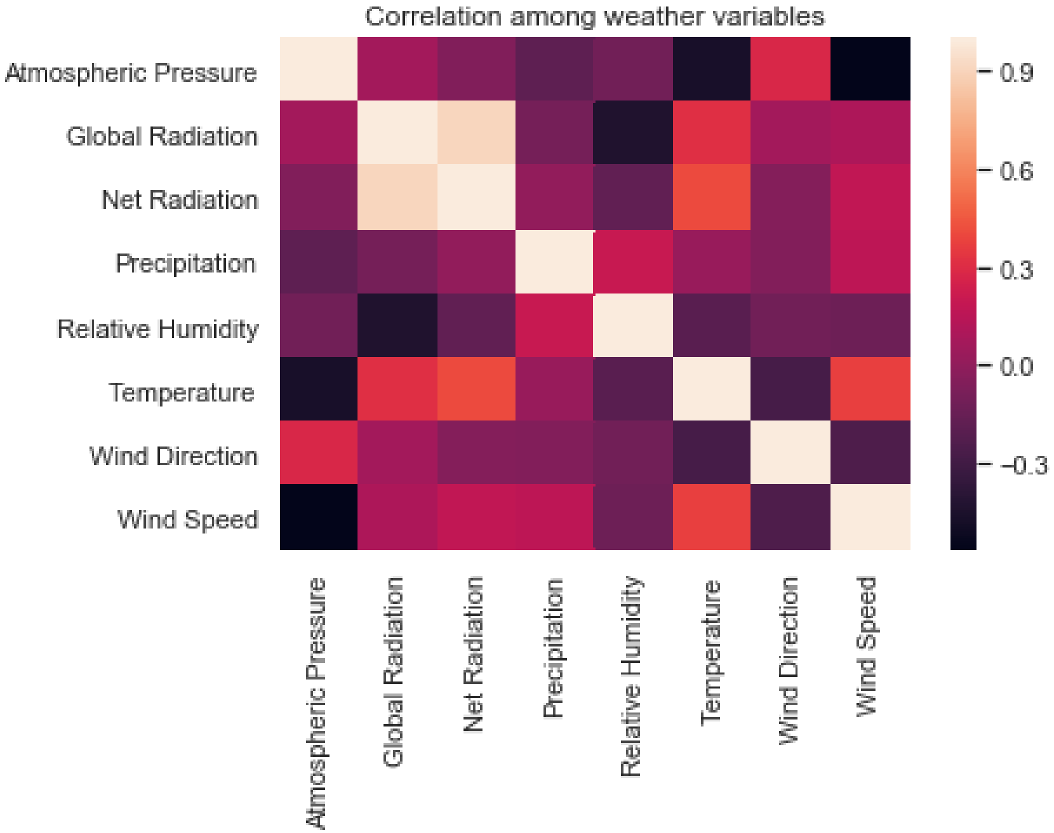
To which we add:

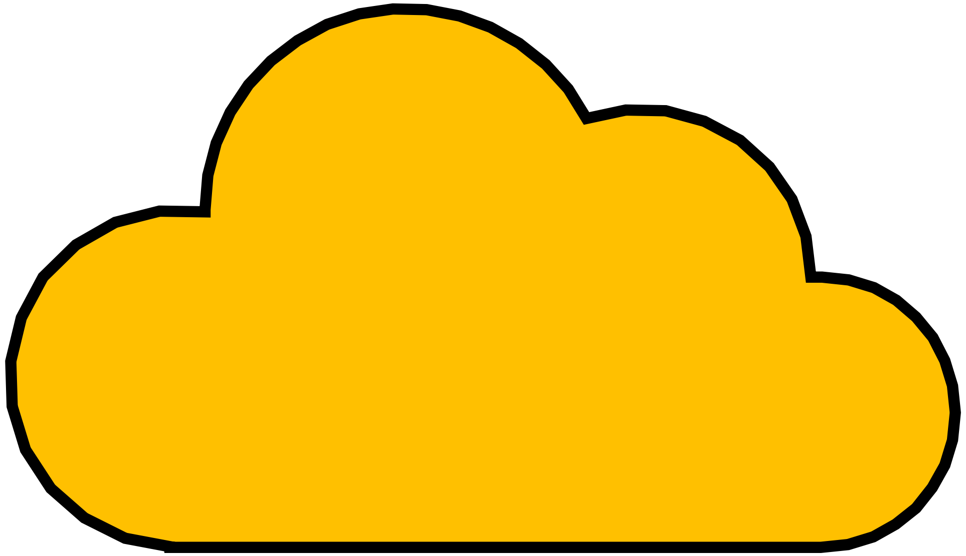
- Exponentially Smoothed variables (+ 18 features)
- 24 Hours var (ratio var between 0 and 24)
- Number of the day in the week (ratio var from 0 to 7)
- Weekend dummies (0-1)

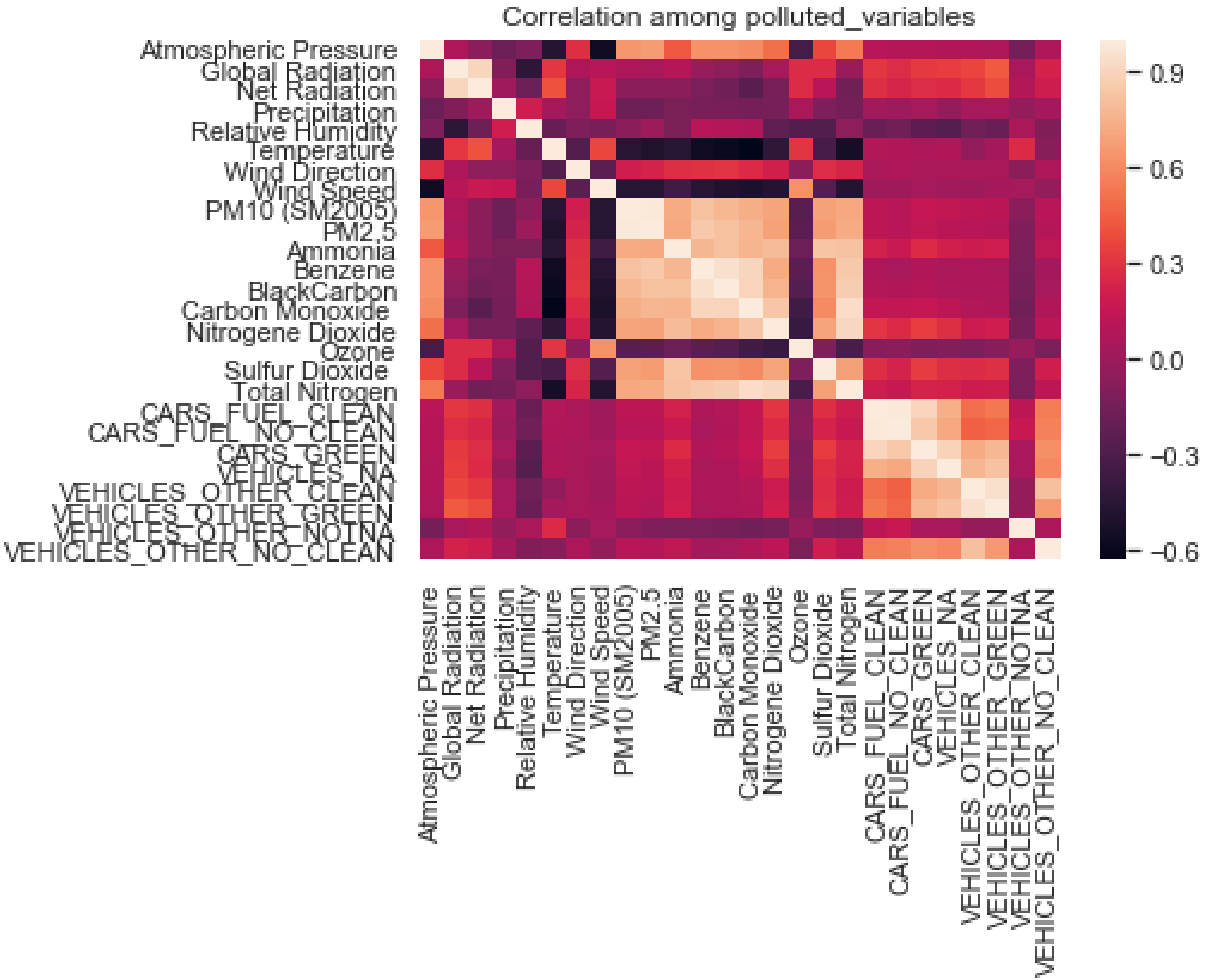
For a total of **36** features and **1464** observations


Correlation among polluted_variables
PM10 (SM2005)
PM2.5
Ammonia
Benzene
BlackCarbon
Carbon Monoxide
Nitrogene Dioxide
Ozone
Sulfur Dioxide
Total Nitrogen
1.00
0.75
0.50
0.25
0.00
−0.25

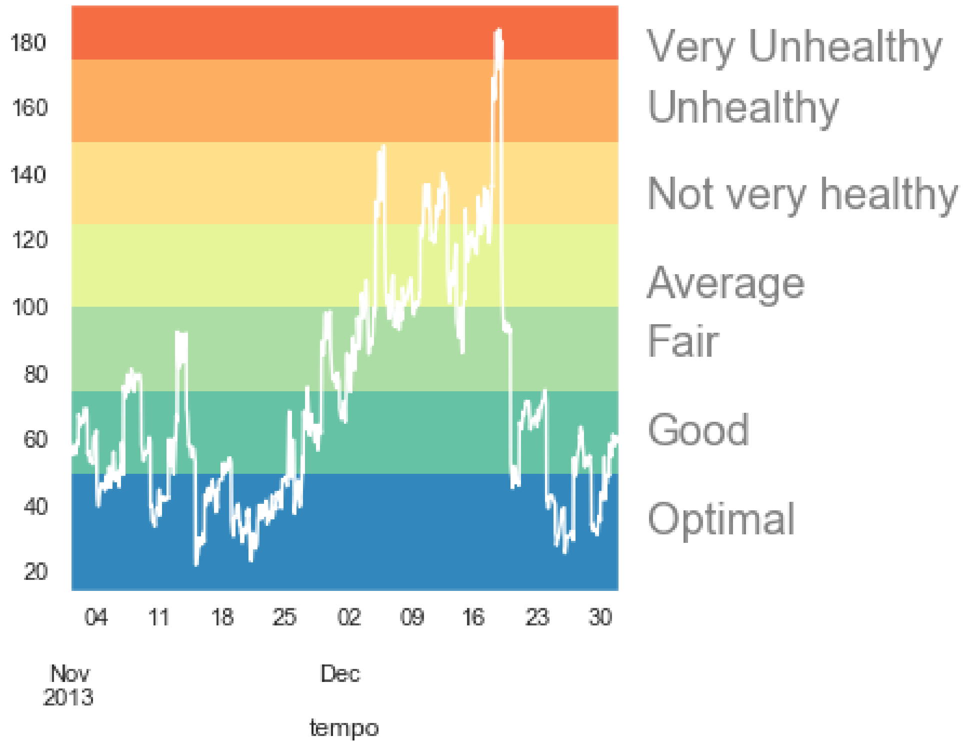



Correlation among weather variables
Atmospheric Pressure
Global Radiation
Net Radiation
Precipitation
Relative Humidity
Temperature
Wind Direction
Wind Speed
0.9
0.6
0.3
0.0
−0.3




Correlation among polluted_variables
Atmospheric Pressure
Global Radiation
Net Radiation
Precipitation
Relative Humidity
Temperature
Wind Direction
Wind Speed
PM10 (SM2005)
PM2,5
Ammonia
Benzene
BlackCarbon
Carbon Monoxide
Nitrogene Dioxide
Ozone
Sulfur Dioxide
Total Nitrogen
CARS_FUEL_CLEAN
CARS_FUEL_NO_CLEAN
CARS_GREEN
VEHICLES_NA
VEHICLES_OTHER_CLEAN
VEHICLES_OTHER_GREEN
VEHICLES_OTHER_NOTNA
VEHICLES_OTHER_NO_CLEAN
0.9
0.6
0.3
0.0
−0.3
−0.6

Air Quality index Classes



$$I_{PM10} = \frac{\overline{V_{med24h_{PM10}}}}{V_{rif_{PM10}}} \times 100 \longrightarrow$$

Valore di riferimento
(50 µg/m³) valore limite giornaliero per la protezione della salute umana (D.M. 02/04/2002 n. 60)

$$I_{NO_2} = \frac{\overline{V_{\max h_{NO_2}}}}{V_{rif_{NO_2}}} \times 100 \longrightarrow$$

Valore di riferimento
(200 µg/m³) valore limite orario per la protezione della salute umana (D.M. 02/04/2002 n. 60)

$$I_{8hO_3} = \frac{\overline{V_{\max 8h_{O3}}}}{V_{rif_{8hO3}}} \times 100 \longrightarrow$$

Valore di riferimento
(120 µg/m³) valore bersaglio per la protezione della salute umana (D. Lgs 21/05/2004 n. 183)

IQA complessivo

$$I_{IQA} = \frac{I_{PM10} + \max(I_{NO2}, I_{O3})}{2}$$

- Adapted to hourly!

RESULTS

But which model we should use ?

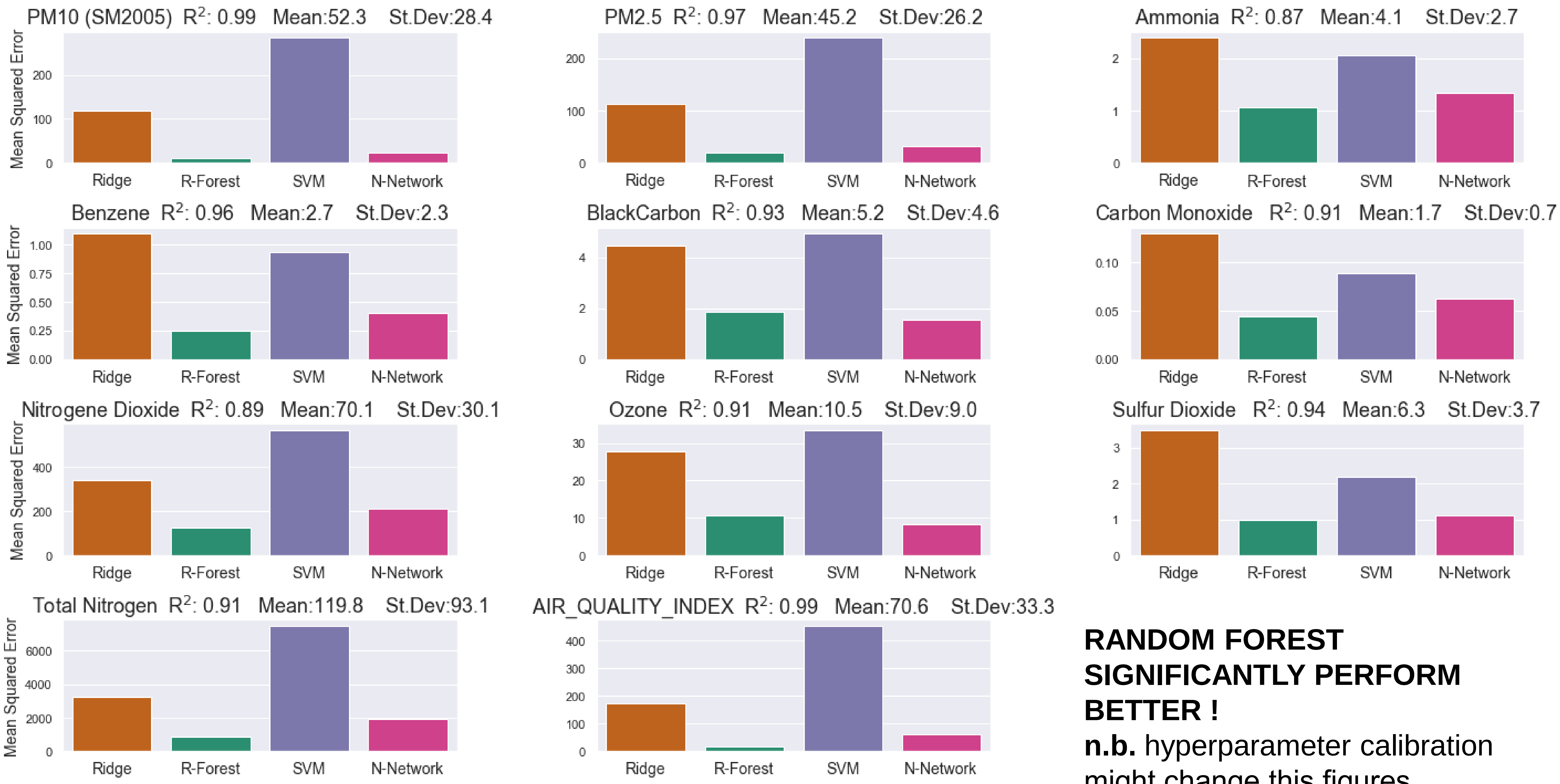
We opt for different models that include “auto” feature selections (or shrinking parameter that take into account var importances…)

1. Ridge
2. Random Forest
3. Support Vector Machine
4. Neural Networks

n.b.:
-we use random sampling with test size 20% (limitations discussed in last slide)

-no big gain from stratifies sampling (not reported here)

Performance of different models among pollutants



RANDOM FOREST SIGNIFICANTLY PERFORM BETTER !

n.b. hyperparameter calibration might change this figures…

A RANDOM FOREST

Approaches to predict Air Quality Index

1

“LA COMPLICATA”

Predict pollutants and then compute the corresponding AQI (using also “training” obs , since sampling is randomized).

2

“LA DIRETTISSIMA”

Predict directly AQI

3

“LA SEMPLICE”

Predict AQI classes using AQI scores

4

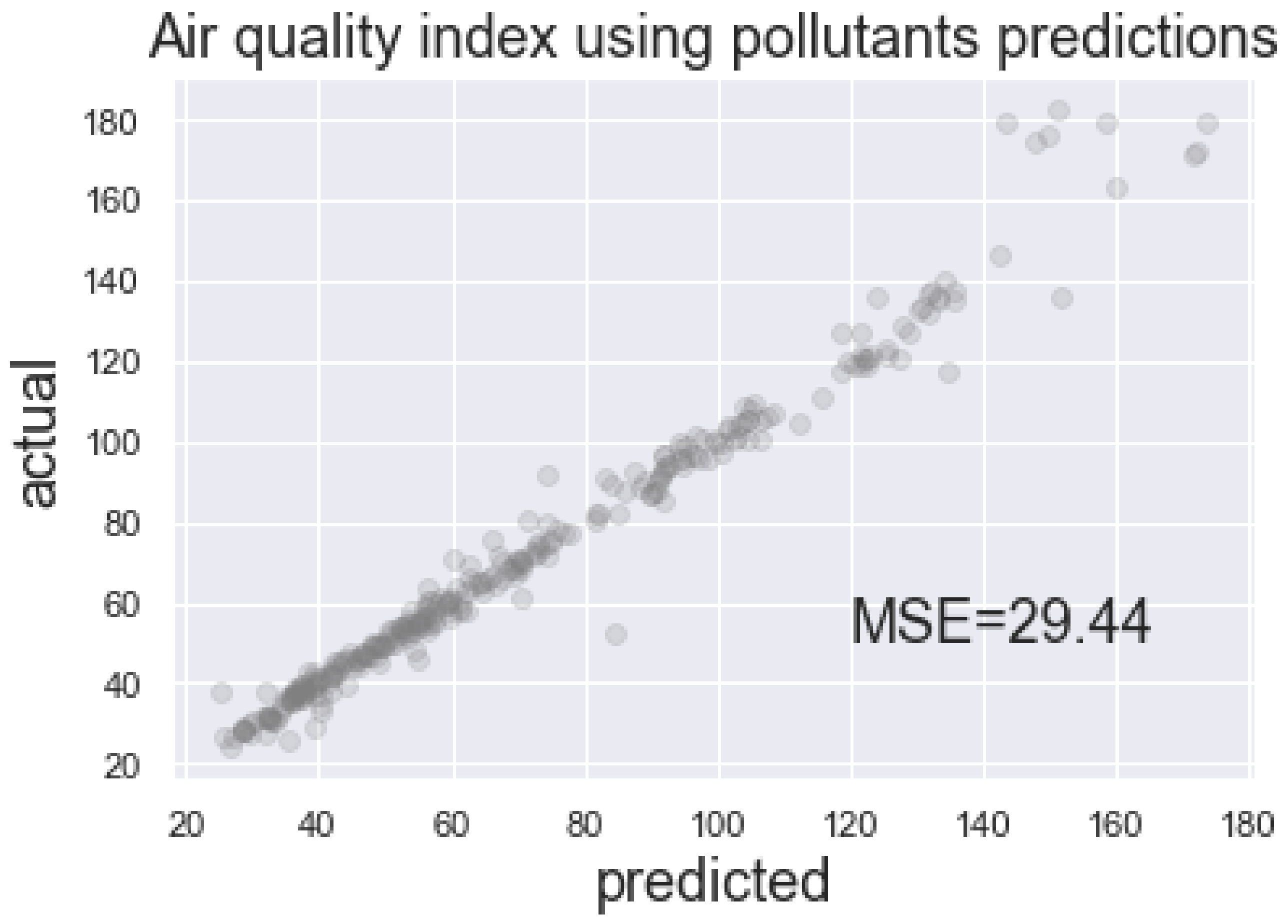
“LA CLASSICA”

Classify AQI classes (**Classification Task)**

“LA COMPLICATA”

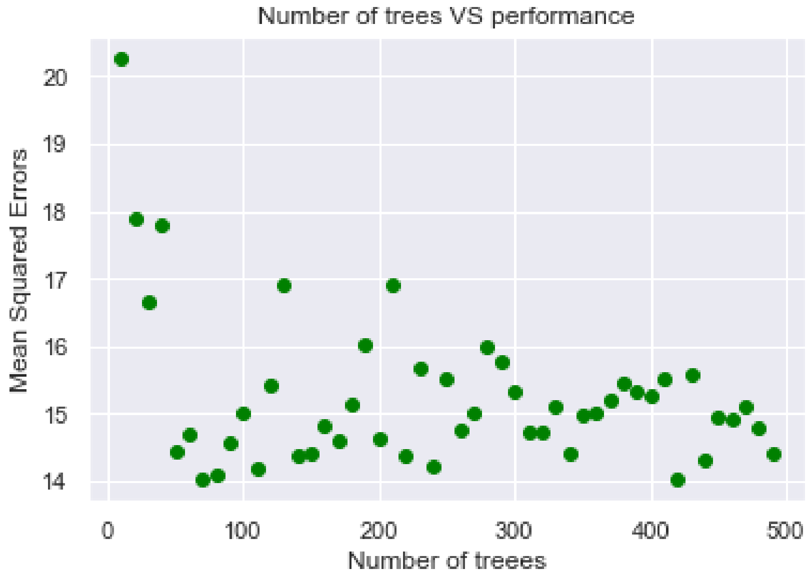
We take results from the pollutants predicted by random forest… and then compute the implied AQI

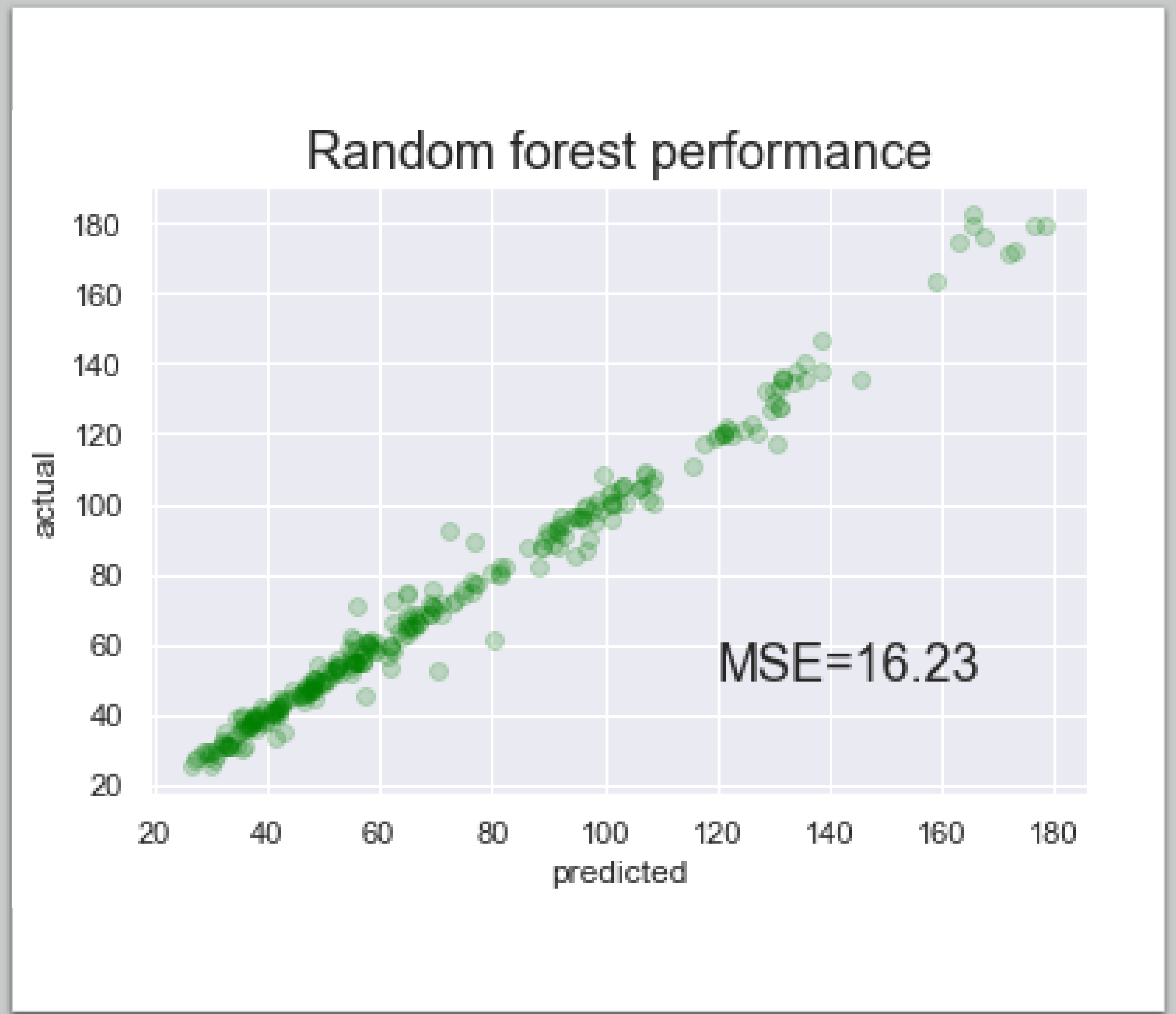
n.b. please note that this way we also use obs from the test set to compute AQI



“LA DIRETTISSIMA”

We predict directly the AQI using Random Forest. (with 100 trees).



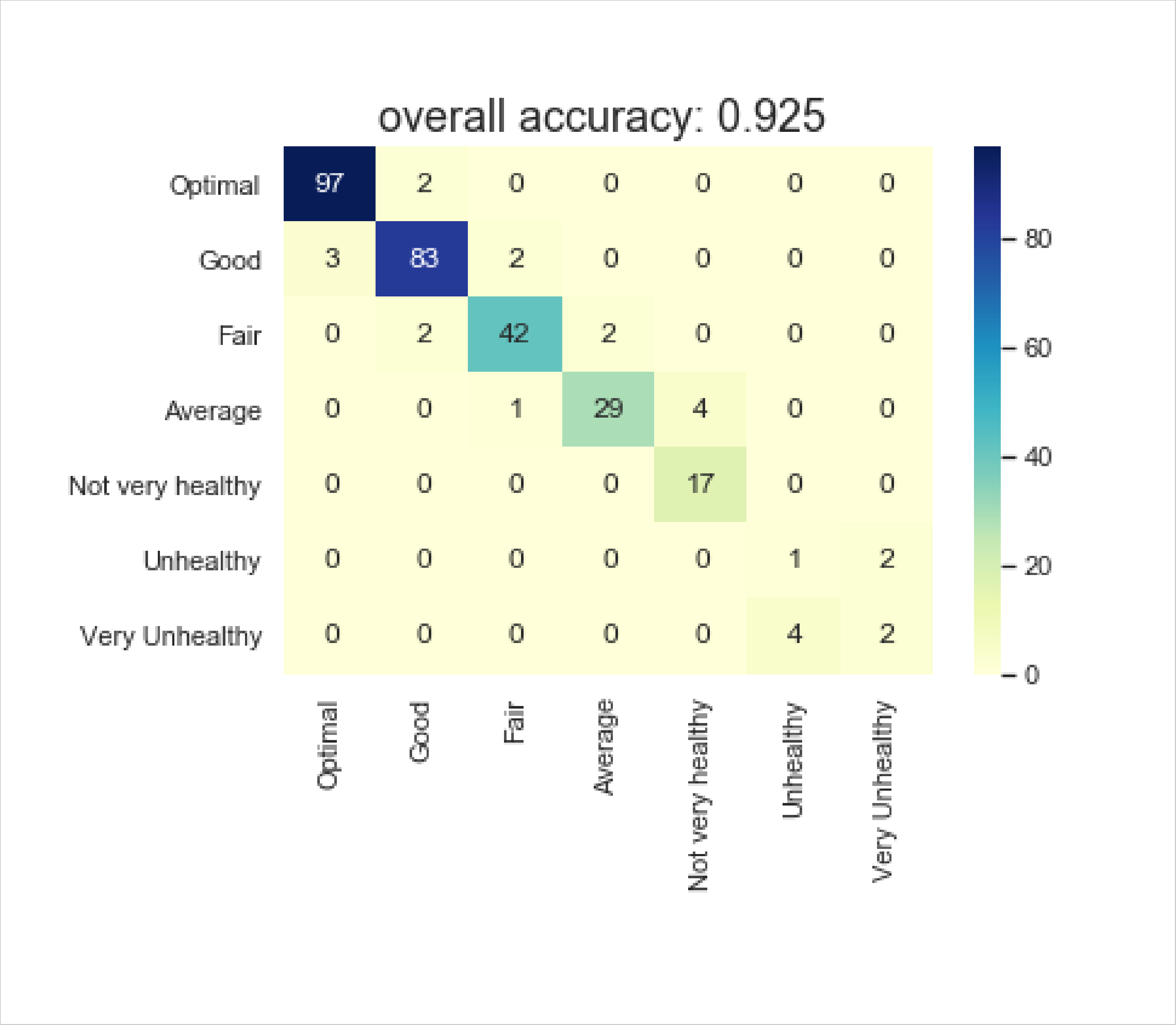


"LA SEMPLICE"

We use the results from a Random Forest Regressor, to compute the corresponding labels.

(i.e. rows are actuals, columns are predicted)

Benchmark is ~0,35 (frequency of "optimal)

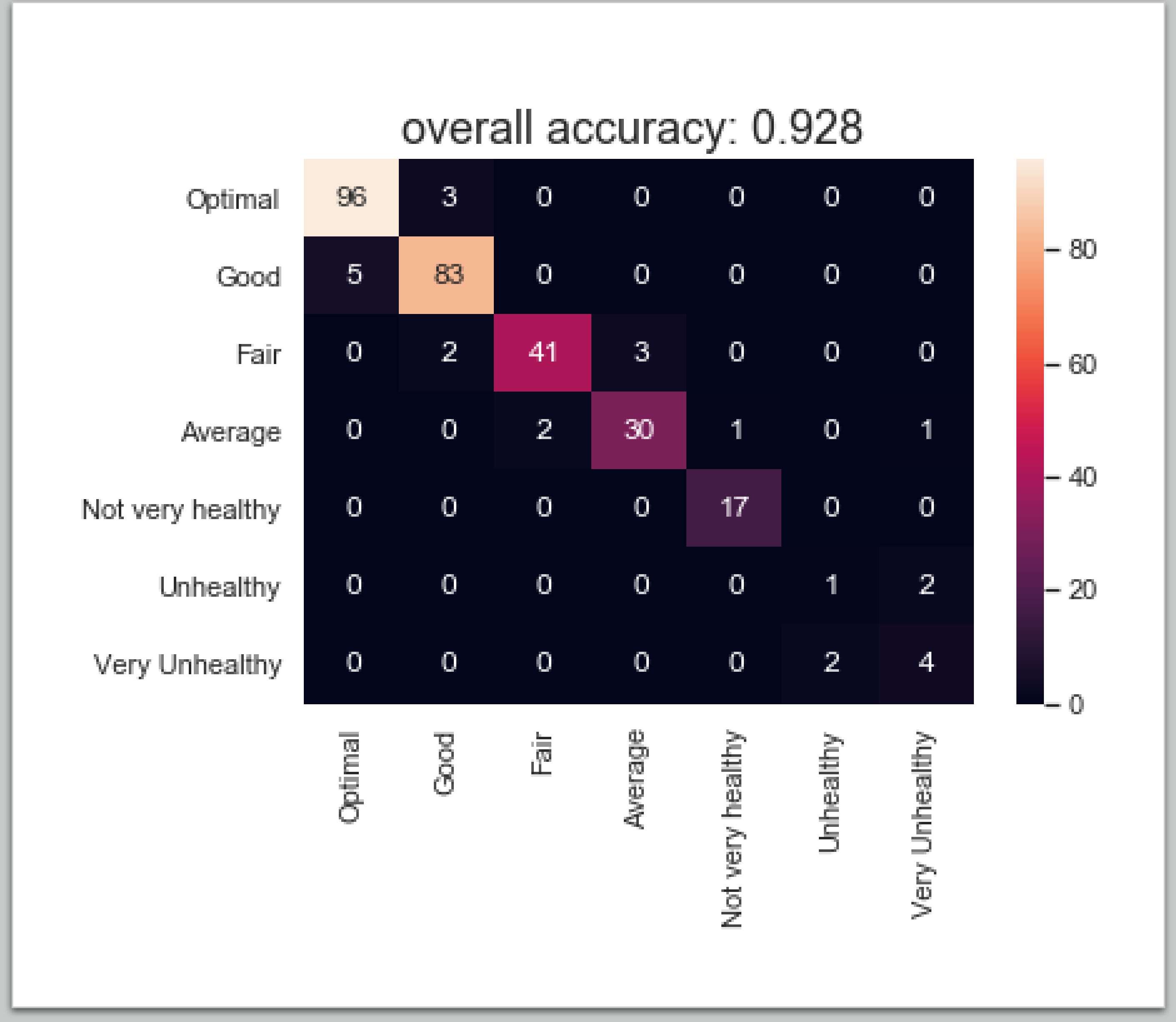


“LA CLASSICA”

We predict the class using a Random Forest Classifier.

Results are quite similar as before.

(i.e. rows are actual, columns are predictions)



Summing up…..

We manage to get:

MSE: ≈15
(benchmark 1100)

ACCURACY: ≈ 92%
(benchmark 35%)

Everything nice BUT…

Some important limitations

The results we show cannot be properly interpreted as "out of sample" test errors for several reasons:

1. In the imputation phase we used all the variables (both from train and test set).
2. Smoothed variables incorporate info from both test and train set
3. To get appropriate estimates for the performance of the model, the test should be sampled considering time constraint of the variables (i.e. should be taken from a period of observations **after** the train set).

*Test error is the error you get when you run the trained model on a set of data that it **has <u>previously never been exposed</u> to**. This test error can be used to estimate the accuracy of the model before it is shipped to production.*


Thank you. Thank you so much.