Air Quality Forecasting

Datena Amedeo, Dell'Atti Martina, Palummo Alessandro

November 20, 2020

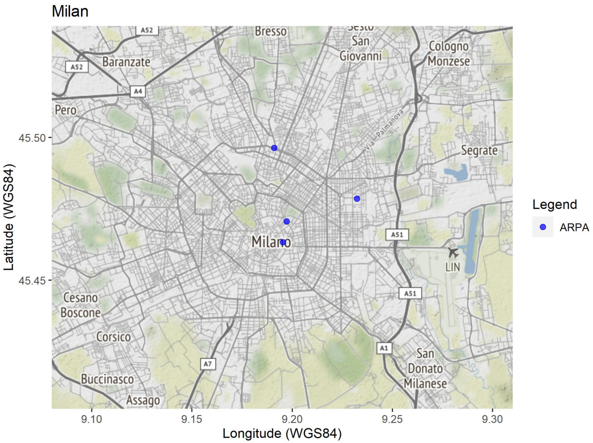
POLITECNICO
MILANO 1863

Wiseair

Wiseair is a startup born in 2018 from some students of Politecnico di Milano, who focused on the problem of **air quality**.
They designed low-cost sensor called **Arianna**

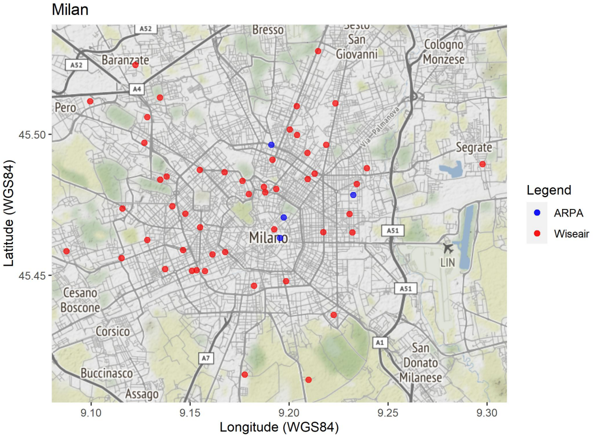
Wiseair

ARPA has installed four air quality measurement stations in Milan.
There are > 50 Arianna stations.



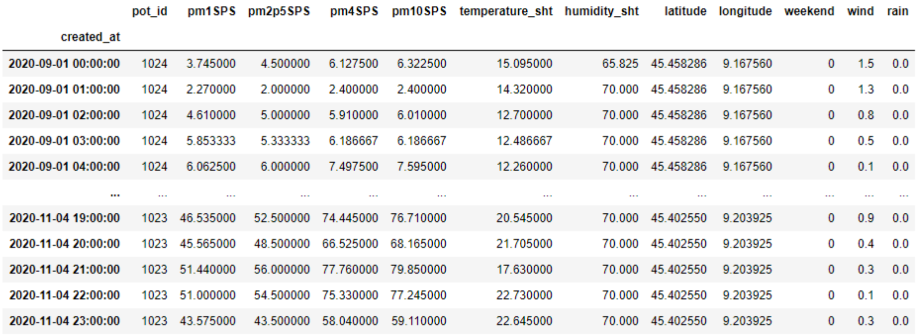
Wiseair

ARPA has installed four air quality measurement stations in Milan. There are $>$ 50 Arianna stations.



Dataset

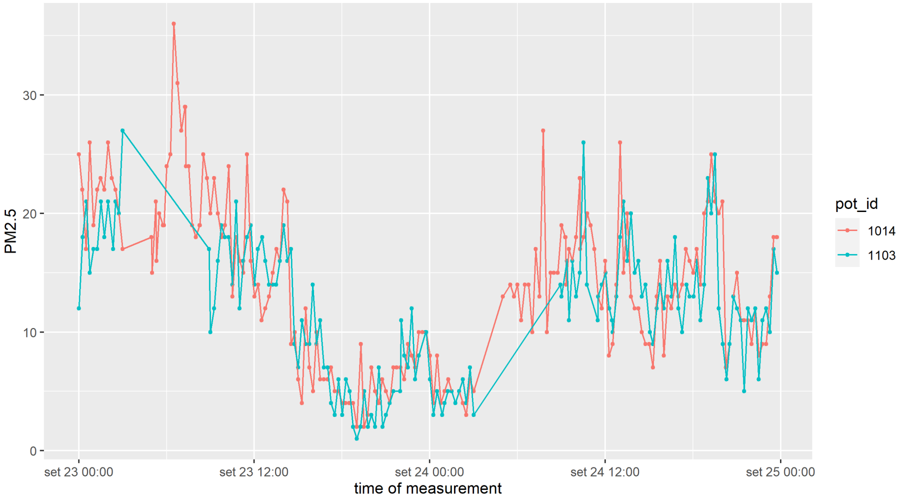
Wiseair own 68 air quality sensor in Milan at the moment, active from July 2020, and they are increasing day by day thanks to the involvement of citizens who installed them. Our data go from September 2020 to November 2020



	pot_id	pm1SPS	pm2p5SPS	pm4SPS	pm10SPS	temperature_sht	humidity_sht	latitude	longitude	weekend	wind	rain
created_at												
2020-09-01 00:00:00	1024	3.745000	4.500000	6.127500	6.322500	15.095000	65.825	45.458286	9.167560	0	1.5	0.0
2020-09-01 01:00:00	1024	2.270000	2.000000	2.400000	2.400000	14.320000	70.000	45.458286	9.167560	0	1.3	0.0
2020-09-01 02:00:00	1024	4.610000	5.000000	5.910000	6.010000	12.700000	70.000	45.458286	9.167560	0	0.8	0.0
2020-09-01 03:00:00	1024	5.853333	5.333333	6.186667	6.186667	12.486667	70.000	45.458286	9.167560	0	0.5	0.0
2020-09-01 04:00:00	1024	6.062500	6.000000	7.497500	7.595000	12.260000	70.000	45.458286	9.167560	0	0.1	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
2020-11-04 19:00:00	1023	46.535000	52.500000	74.445000	76.710000	20.545000	70.000	45.402550	9.203925	0	0.9	0.0
2020-11-04 20:00:00	1023	45.565000	48.500000	66.525000	68.165000	21.705000	70.000	45.402550	9.203925	0	0.4	0.0
2020-11-04 21:00:00	1023	51.440000	56.000000	77.760000	79.850000	17.630000	70.000	45.402550	9.203925	0	0.3	0.0
2020-11-04 22:00:00	1023	51.000000	54.500000	75.330000	77.245000	22.730000	70.000	45.402550	9.203925	0	0.1	0.0
2020-11-04 23:00:00	1023	43.575000	43.500000	58.040000	59.110000	22.645000	70.000	45.402550	9.203925	0	0.3	0.0

Time irregularity

Each Arianna station has its own measure frequency. Some stations stop transmitting at night.



Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement

Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement
- **created-at** gives the time coordinate for the measurement, up to the second. The format is ”yyyy-mm-dd hh:MM:ss”

Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement
- **created-at** gives the time coordinate for the measurement, up to the second. The format is "yyyy-mm-dd hh:MM:ss"
- **pm1SPS**, **pm2p5SPS**, **pm10SPS** are the Particular Matter, pollutant of the air, which have dimensions, respectively, less or equal than $1\mu m$, $2.5\mu m$, $10\mu m$. They are measured in $\mu g/m^3$

Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement
- **created-at** gives the time coordinate for the measurement, up to the second. The format is "yyyy-mm-dd hh:MM:ss"
- **pm1SPS**, **pm2p5SPS**, **pm10SPS** are the Particular Matter, pollutant of the air, which have dimensions, respectively, less or equal than $1\mu m$, $2.5\mu m$, $10\mu m$. They are measured in $\mu g/m^3$
- **temperature-sht, humidity-sht** are the temperature and the humidity for each sensor's measurement of Arianna

Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement
- **created-at** gives the time coordinate for the measurement, up to the second. The format is ”yyyy-mm-dd hh:MM:ss”
- **pm1SPS**, **pm2p5SPS**, **pm10SPS** are the Particular Matter, pollutant of the air, which have dimensions, respectively, less or equal than $1\mu m$, $2.5\mu m$, $10\mu m$. They are measured in $\mu g/m^3$
- **temperature-sht, humidity-sht** are the temperature and the humidity for each sensor's measurement of Arianna
- **latitude, longitude**

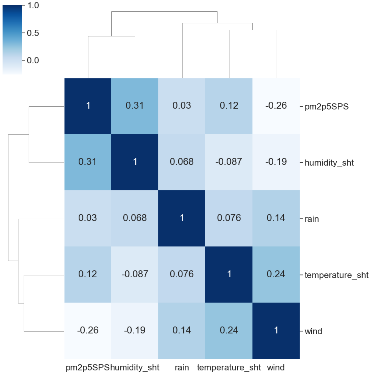
Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement
- **created-at** gives the time coordinate for the measurement, up to the second. The format is "yyyy-mm-dd hh:MM:ss"
- **pm1SPS**, **pm2p5SPS**, **pm10SPS** are the Particular Matter, pollutant of the air, which have dimensions, respectively, less or equal than $1\mu m$, $2.5\mu m$, $10\mu m$. They are measured in $\mu g/m^3$
- **temperature-sht, humidity-sht** are the temperature and the humidity for each sensor's measurement of Arianna
- **latitude, longitude**
- **weekend** is a dummy variable which indicate if the measurement is in a day of the week or not

Dataset

- **pot-id** identifies which sensor has taken the measurement. The ID is useful to geolocate the measurement
- **created-at** gives the time coordinate for the measurement, up to the second. The format is "yyyy-mm-dd hh:MM:ss"
- **pm1SPS**, **pm2p5SPS**, **pm10SPS** are the Particular Matter, pollutant of the air, which have dimensions, respectively, less or equal than $1\mu m$, $2.5\mu m$, $10\mu m$. They are measured in $\mu g/m^3$
- **temperature-sht, humidity-sht** are the temperature and the humidity for each sensor's measurement of Arianna
- **latitude, longitude**
- **weekend** is a dummy variable which indicate if the measurement is in a day of the week or not
- **wind, rain** are the wind speed and the rain level measured by ARPA sensors

Data Exploration

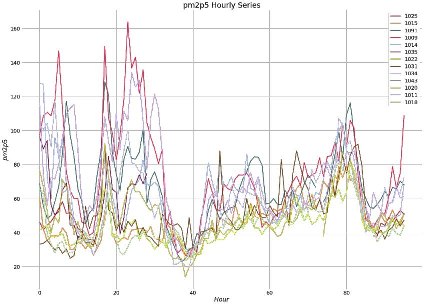


From domain knowledge we know that sensors tends to measure higher PM values if humidity increases a lot. The clustermap, which plot a matrix dataset as a hierarchically-clustered heatmap, put in evidence the correlation between humidity and pm2p5 values, which is higher w.r.t. the correlation between the other features.

Goal

Our main **goal** is to predict future PM values.
We will tackle the problem in two steps:

- **Univariate analysis**: focus on a single time series without taking into account any spatial correlation between different sensors.
- **Multivariate analysis**: derive a vectorial model which uses information from all the sensors, this time taking into account the spatial correlation between them.



The model

The first proposed model for Univariate Analysis is an AR(2) model:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

where ϵ_t is a sequence of uncorrelated error terms and the ϕ_i are constant parameters.

LIKELIHOOD

$$y|\underline{\phi}, \beta \sim \mathcal{N}(\underline{\phi}^T \beta, \sigma^2)$$

σ^2 not known

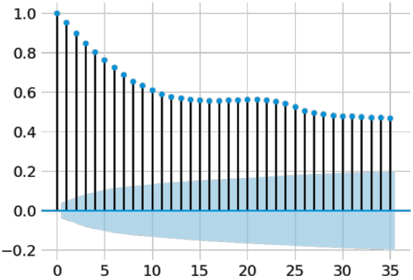
PRIORS

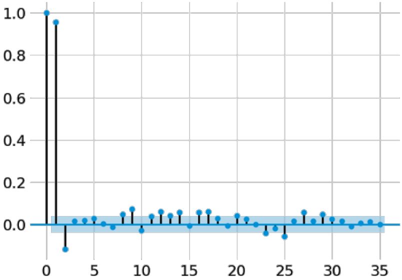
$$\underline{\phi}|\sigma^2 \sim \mathcal{N}_2(\underline{\mu}_0, \sigma^2 B_0)$$

$$\sigma^2 \sim inv\Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$\mu_0, B_0, \nu_0, \sigma_0^2$ fixed

Why AR(2)?





ACF tends to zero only asymptotically, while PACF drops to zero at lag 2. An autoregressive model of order 2 seems the most appropriate choice.

The model

Since the model is a standard conjugate linear model,
the POSTERIORS are:

- $\underline{\phi}|Y \sim \mathcal{N}_2(\mu_n, \sigma^2 B_n)$

$$\mu_n = \mu_0 + B_0\Phi[\Phi^T B_0\Phi + I_n]^{-1}(Y - \Phi^T\mu_0)$$
$$B_n = B_0 - B_0\Phi[\Phi^T B_0\Phi + I_n]^{-1}\Phi^T B_0$$

- $\sigma^2|Y \sim inv\Gamma\left(\frac{v_n}{2}, \frac{v_n\sigma_n^2}{2}\right)$

$$v_n = v_0 + n$$
$$\sigma_n^2 = \frac{1}{v_n}\left[v_0\sigma_0^2 + (Y - \Phi^T\mu_0)^T[\Phi^T B_0\Phi + I_n]^{-1}(Y - \Phi^T\mu_0)\right]$$

Future model development

We do not expect that a simple AR(p) model will work. This is just a first attempt to understand the structure of the time series. Next steps:

- try to fit more "complex" classical TS models as ARMA/ARIMA
- include seasonality in the model by fitting a SARIMA model (we have spotted a daily seasonal component)
- pass to a Dynamic Linear Model formulation (poi spiego meglio...)
- extend the scalar case to the vectorial case, introducing spatial correlation between different sensors

Bibliography

- Time Series Modeling, Computation,and Inference. West Mike, Prado Raquel
- Quadro di riferimento ambientale, Componente Atmosfera - Istituto superiore per la protezione e la ricerca ambientale(ISPRA)