POLITECNICO DI MILANO
Master Degree in Mathematical Engineering
Industrial and Information Engineering
Department of Mathematics

Air Quality forecasting

Bayesian Statistics Project Report

Group members:
Datena Amedeo
Dell'Atti Martina
Palummo Alessandro

Academic Year 2020-2021

1 Wiseair and data measurement

1.1 Wiseair

The measure of air quality in Lombardy is carried out by ARPA using accurate sensing stations. These stations, however, are expensive and bulky, so only a few of them are deployed on the territory. Only two of these stations are present in Milan.

However, two stations are not enough to capture the phenomenon in way that is useful for a single citizen living in a specific area of the city, Since the concentration of particulate matter (PM) varies at an hyperlocal scale.

Wiseair is a startup company that has designed a low-cost sensor, meant to be user-friendly and easy to install: Arianna. More than 100 citizens preordered one, and 50 of these sensors have been distributed as of July 2020, thanks to the involvement of citizens.

Below is a map of Milan with the locations of installed and functioning Ariannas.

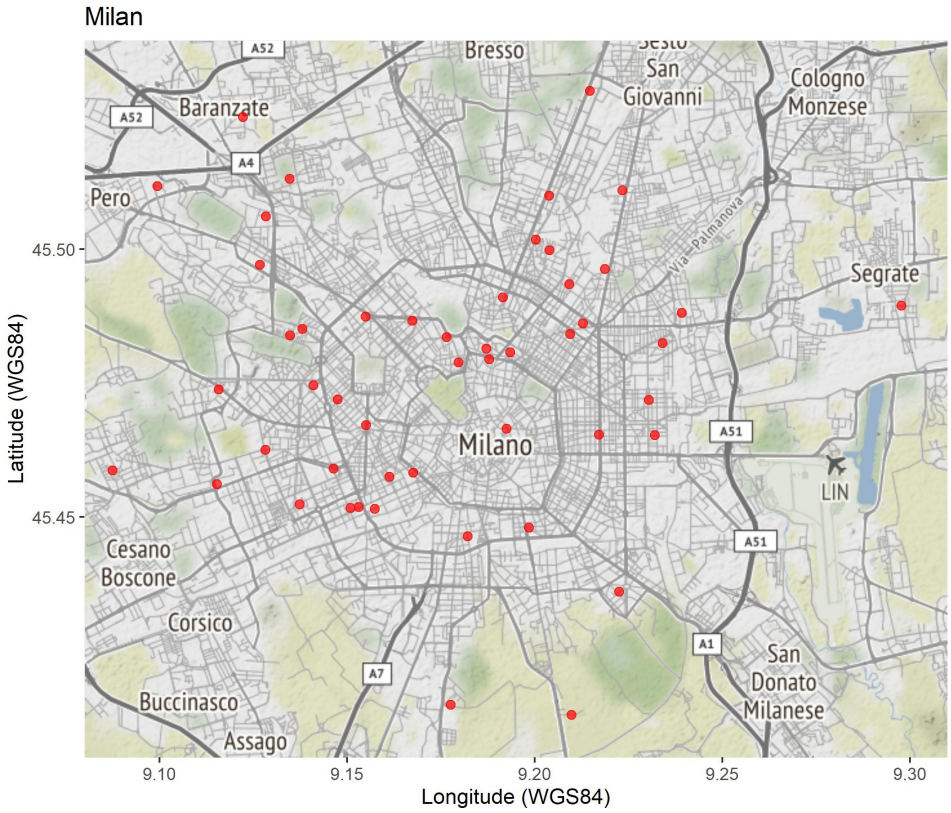


Figure 1: locations of installed Ariannas in the city of Milan

1.2 Objective

Wiseair's goal is to provide citizens with useful information about air quality. To achieve this, the information coming from the Ariannas needs to be interpolated on areas where sensors are not present. Also, an air quality forecast would be useful, for example for planning outdoor activities.

1.3 Data measurement

An Arianna sensor measures the concentration of PM once per hour or more frequently. The sensor also measures humidity and temperature with each PM measurement. It has been observed that the measurement accuracy is negatively affected by adverse weather conditions, so raw data can be filtered and processed so that outliers can be detected and removed, also using humidity and temperature information.

2 Database structure

Each observation in the dataset is a measurement from and Arianna station.

- The variable "PotID" identifies which sensor has taken the measurement. The ID is useful to geolocalize the measurement.
- The variable "created_at" gives the time coordinate for the measurement, up to the second. The format is "yyyy-mm-dd hh:MM:ss".
- There are 4 variables related to air quality measurement, one for each of the following particulate matter types: PM_1,$PM_{2.5}$,PM_{10}. These types refer to the thickness of the measured particulate (for example, the variable $PM_{2.5}$ refers to particles of less than 2.5 µm. The PM concentrations are measured in $\mu g\, m^{-3}$.
- The variables "temperature" and "humidity" record the temperature and humidity measurements made by the Arianna station.

2.1 Time irregularity

The measurement frequency of the Arianna station can be set by the user and the stations do not transmit their measurement if an internet connection is not available. This means (1) that the measurement frequency varies among stations and (2) most stations do not perform measurement at a regular time frequency: for example, many stations stop transmitting data at night. This introduces time irregularity in the time series. Another issue is that different stations have been installed at different times: the time series for each Arianna starts at a different time.

Time irregularity is an issue to be solved before performing univariate or multivariate analyses, and each type of analysis might require a different approach to the issue.

Time irregularity can be solved in multiple methods. The data could be smoothed using spline curves, producing continuous functions that can be sampled at regular time intervals.

—— Originally we had data with a non regular frequency, i.e. the frequency of measurements of a sensor is different from others. For example, some pot-id make two measurements for hour, other pot-id make five measurements for hour. So we have performed an hourly average of the measured quantities in order to have equally spaced points. We choose to aggregate hourly because of two reasons:

- This choice does not remove eventual patterns during the day, differently from the case in which we tried to aggregate taking daily averages.
- Weather data (rain level, wind speed) from ARPA are available with hourly frequency so we needed to average available data hourly before integrating further information from ARPA sensors.

created_at	pot_id	pm1SPS	pm2p5SPS	pm4SPS	pm10SPS	temperature_sht	humidity_sht	latitude	longitude	weekend	wind	rain
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

Figure 2: Portion of the processed dataset

3 Univariate Analysis

3.1 Exploration analysis

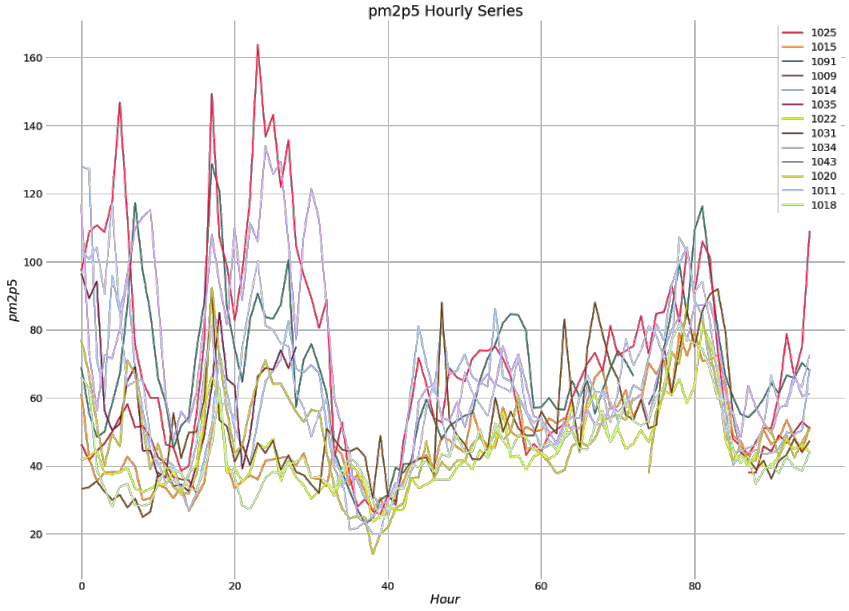


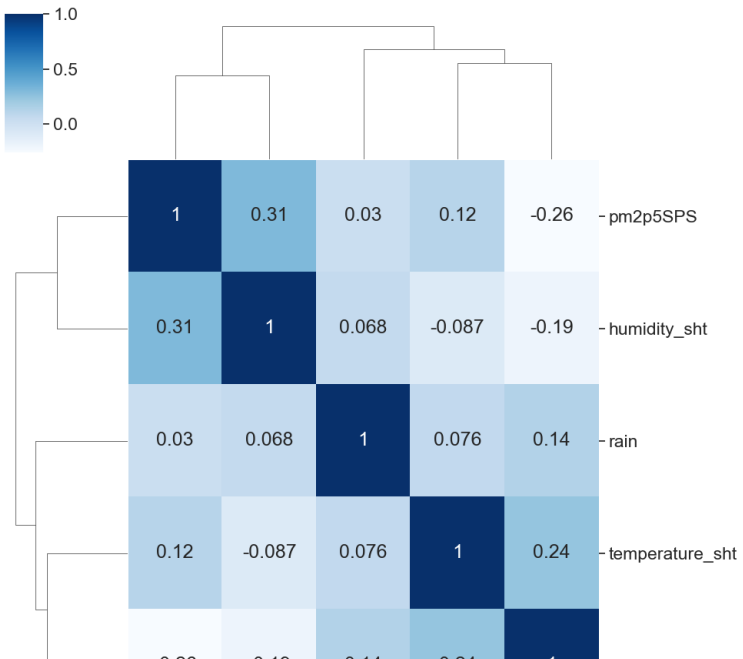
Figure 3: time series of pm2p5 in 48 hours for some pot-id

Wiseair own more than 60 air quality sensor in Milan at the moment and they are increasing day by day thanks to the involvement of citizens who install them.

In our dataset they are identified by the variable potid. We consider only one potid to make some initial exploration, in particular we consider the **potid 1091** because it is the pot with less missing values.

The precision of sensors is negatively affected by special weather conditions (e.g., heavy rain, humidity, wind speed, temperature). Therefore, we examine atmospherical condition in order to find some interplay with the variation of pm2pm5. In the following graph, we show a qualitative comparison in the effect of rain, wind,humidity and temperature with the variation of pm2p5 in 48h for the potid 1091.
As for the moment just observe that there does not appear to be a strong relationship between pm2p5 and rainy level or wind speed or temperature. A bit different is the nature of humidity. Even though the time series seems to have a natural periodicity, the action of humidity sometimes disrupts this behaviour, since there is a big peak of pm2p5 in correspondence of a continuous high level of humidity(in the end of October).

This can be seen also through the clustermap below, which plot a matrix dataset as a hierarchically-clustered heatmap . Motivated by these results, we will try to take into account this fact in our future analysis.



3.2 Stationarity

In order to check the possibility to consider our model as a structural time series, we check the stationary of our time series such that its statistical properties (such as mean, variance, autocorrelation, etc.) are all constant over time. To check whether a series is stationary, we can apply the Augmented Dickey fuller test. Since the p-value of the test is 0.020 we can reject the null hypothesis that the series is non-stationary.

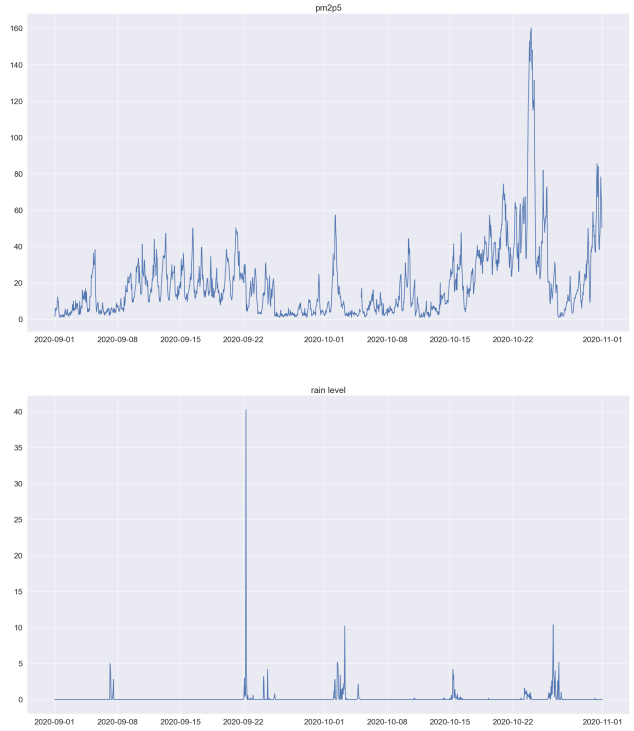


Figure 4: (a) Plot of rain vs pm2p5

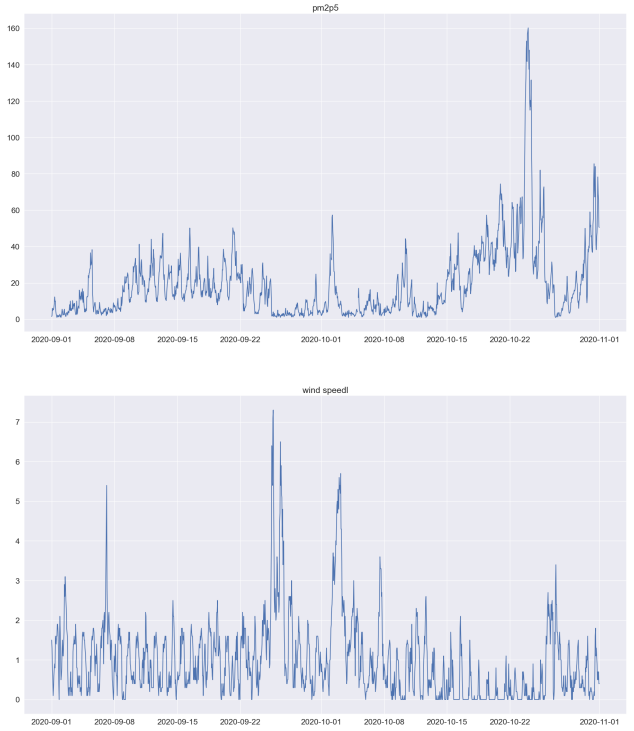


Figure 5: (b) Plot of wind speed vs pm2p5

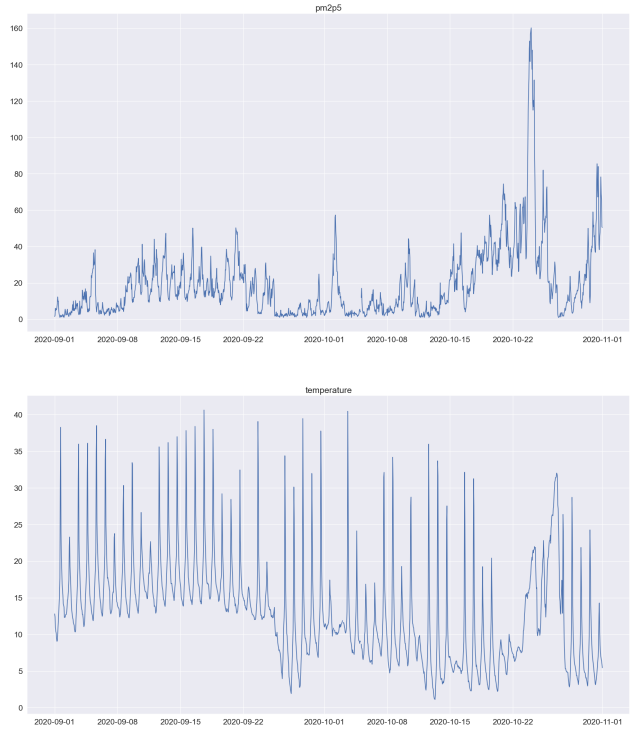


Figure 6: (c) Plot of temperature vs pm2p5

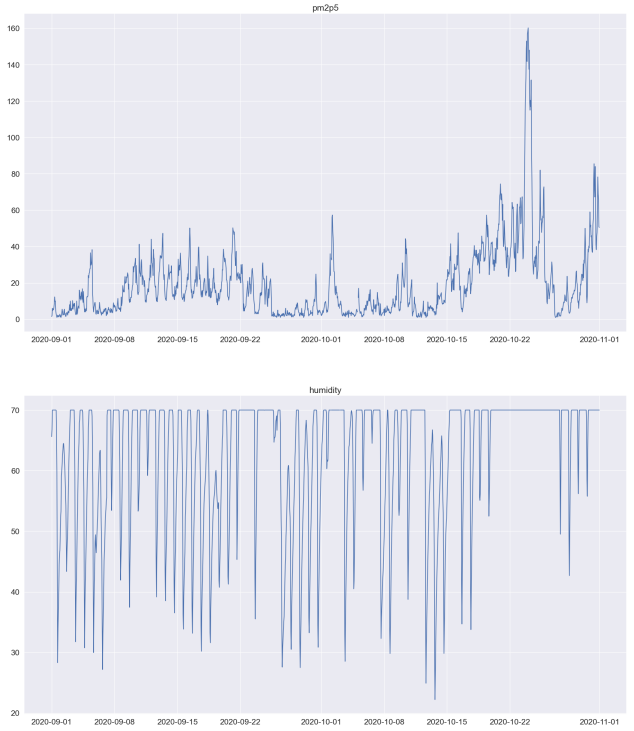


Figure 7: (d) Plot of humidity vs pm2p5

4 Model

Autoregressive time series models are central to stationary time series data analysis. Therefore, as a first

proposal, we choose an AR(2) model for the univariate case and we are going to extend this model with more complicated models for the multivariate case. We make this choice focusing on the analysis of the autocorrelation (ACF) and partial autocorrelation plot (PACF) of the pot 1091: since ACF tails off and PACF cuts off after one lag, then we are dealing with an AR(2) model (see figure 9 and 10).

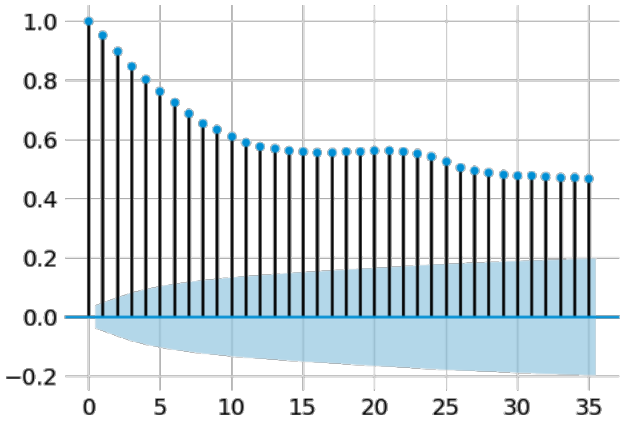


Figure 9: Autocorrelation plot for the 1091 pot series

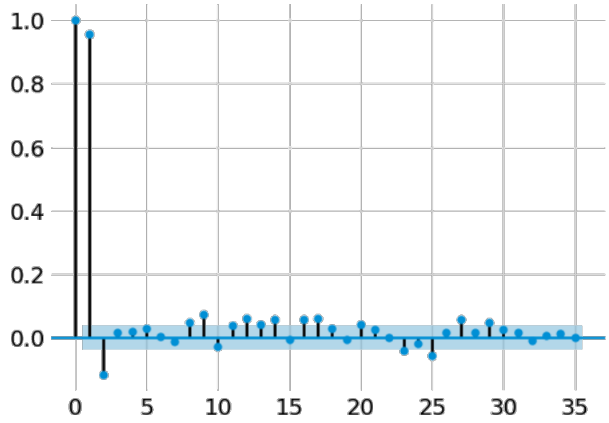


Figure 10: Partial autocorrelation plot for the 1091 pot series

As a first proposal, we assume a conjugate model to explain the AR(2) structure of the time series:

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

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

LIKELIHOOD

$$y|\underline{\phi}, \beta \sim N(\underline{\phi}^T \beta, \sigma^2)$$
$$\sigma^2 \text{ not known}$$

PRIORS

$$\underline{\phi}|\sigma^2 \sim \mathcal{N}_2(\underline{\mu}_0, \sigma^2 B_0)$$
$$\sigma^2 \sim inv\Gamma\Big(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\Big) \qquad \text{with } \mu_0, B_0, \nu_0, \sigma_0^2 \text{ fixed}$$

B_0 is a 2×2 matrix, μ_0 is any vector in $\mathcal{R}^2$. Because the model is a standard conjugate linear model, posteriors are:

POSTERIORS

$$\underline{\phi}|Y \sim \mathcal{N}_2(\mu_n, \sigma^2 B_n)$$
$$\sigma^2|Y \sim inv\Gamma\left(\frac{v_n}{2}, \frac{v_n \sigma_n^2}{2}\right)$$

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

Unfortunately can happen that sensor does not send any data. From our perspective this translates in the presence of missing values in the time series. For this univariate analysis pot 1091 has only the 0.012% of value

missing, hence in this case is not a particular problem. Morover missing data are distributed over the all considered period (we have missing values in correspondence of just few hours) and do not span a large interval of time.