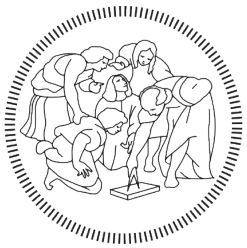


POLITECNICO
MILANO 1863

Pollution and weather report in Lombardy

Nonparametric Statistics course

Lorenzo Angiolini, Giulia Bergonzoli, India Ermacora, Lucia Gregorini

Academic year 2022-2023

Contents

1 Introduction

Atmospheric pollution has become a matter of concern as result of the dramatic consequences on population and human health of several dramatic pollution episodes that took place in the middle of the last century. As a consequence of the increasing deterioration of air quality especially in industrial and urban areas, the legislation developed standards for a number of gaseous pollutants and for total suspended particles.

In the field of air quality, the scientific progress of the last decades has detected the main properties of pollutant particles that make them so dangerous for human health. These features are directly related to dust capacity of penetration in the human respiratory system.
Airborne particulate matter represents a complex mixture of organic and inorganic substances that can be solid particles and liquid droplets, covering a wide range of diameters. They are classified based on them, into:
- TSP : Total Suspended Particles, with diameters that are generally $\leq 50 - 100\ \mu m$
- PM_{10} : inhalable particles, with diameters that are generally 10 μm and smaller
- $PM_{2.5}$: fine inhalable particles, with diameters that are generally 2.5 μm and smaller
Results from epidemiological studies contributed to demonstrate that PM10 was a parameter most appropriated for air quality purposes regulation than the existing one TSP, so we focused our attention on it.

Goals

- Compare the quantity of PM_{10} in different years to assess if there is an increase in the level of pollution. We would like to understand how this quantity changes exploiting the temporal dimension of the dataset. To achieve this we will perform different tests in a functional framework

- Exploit the possible relationship between weather covariates and PM_{10} level to gain more insight on what are the driving factors of pollution. The analysis will cover also the spatial dimension, in hope to provide suggestions and tools to limit the harm caused by these agents. The technique we will make use of is a GAM model. Moreover, it would be ideal to perform prediction based on weather forecast. To understand if this is reasonable we will use the prediction methods in the GAM framework together with a conformal prediction setting.

1.1 The Dataset

In 2018 the European Environment Agency established that in Europe 3.9 million people live in areas where the pollutants in the air exceed the daily fixed limit: among these, the 95% live in the North of Italy, reason why we chose to analyze this phenomenon in Lombardy.

The material we have based our analysis on comes form the official website of Arpa-Lombardia - the Regional Environmental Protection Agency- which has publish different datasets related to its meteorological monitoring activities and air and water pollution levels. This is a considerable amount of data, as it covers the history of the detections of several dozen years: we are speaking of millions of published records. In particular, we focused our attention on 2 datasets, one regarding air monitoring and the other regarding meteorological readings.

As for the former, as already specified in the introduction, we have analyzed - among the 7 possible species of pollutants monitored - only the PM10 particles, for which the daily average is provided. We decided to focus our initial analysis on the last available year, the 2021, in which 65 stations have measured the particle of interest. Latitude and longitude are provided for each station, they will be useful to reconstruct a spatial interpretation of pollution, and will therefore play a key role.

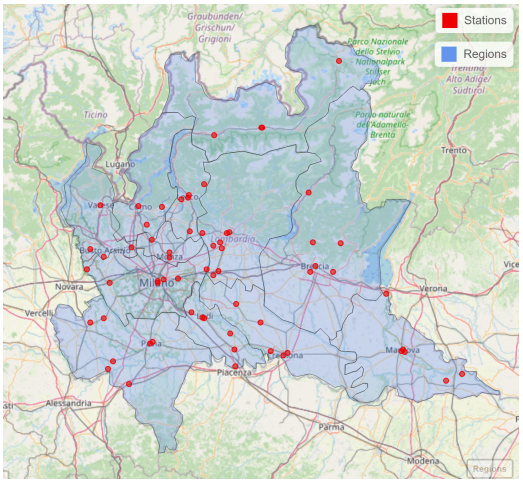


Figure 1: 65 stations measuring PM10 particles

In the second dataset 267 stations detect the following important variables: Temperature (°C), Rainfall (cm), Wind Speed (m/s) and Altitude (m). Is to be underlined that more variables were collected but those previously listed are the only ones we made use of.
At this point, since the locations of the meteorological stations differ from those of the air stations, we decided to associate to each air station the nearest meteorological station, thus assuming that weather information would remain almost unchanged at nearby locations, as is it reasonable to do in the real world.
Another issue was the presence of missing values which may have occurred due to malfunctions or sensor failures. Also some unrealistic observations, such as negative pollution levels or temperature above 50 degrees, were transformed into NAs in order not to impact the results by a transcription/detection error. These are common problem when working with real-world dataset, and there are various methods for resolving it. In some cases it's possible to ignore these values (such as in the computation of the mean of the plot of the functional data), but in other cases it became necessary have the complete time series, and so we decided to fill those gaps by taking the value of the previous day, since it's an estimate which makes sense for the data we are dealing with. Only stations were removed since the percentage of missing values was above 20% and therefore we think the estimated time series might not be reliable.

2 Exploratory data analysis

We begin our inspection with an exploratory data analysis through a frequentist approach, looking at both the pollution and meteorological dataset.

2.1 Air monitoring dataset

The time series of PM10 levels in the stations reveals a clear U-shaped pattern due to seasonal fluctuations, with values in the range from 0 $\mu g/m^3$ to 150 $\mu g/m^3$. It's clear that the pollution levels are higher during the colder months, primarily driven by increased heating and vehicular traffic in urban areas, while in the warmer ones measurements are distributed around low value. We then conclude that seasonality plays a key role in the definition of the pollution level and it will be an important aspect to be taken into account in defining our models. The data show a strong correlation among the different stations as they exhibit a very similar shape. The graph also displays a greater variability, as evidenced by the sharp peaks, which we will try to correlate with the presence of high values in the time series regarding weather information, since it is well known that meteorological factors heavily affect the PM10 concentration in the air. We will take into account this hypothesis and try to verify it more in details when we will build the model for our data.

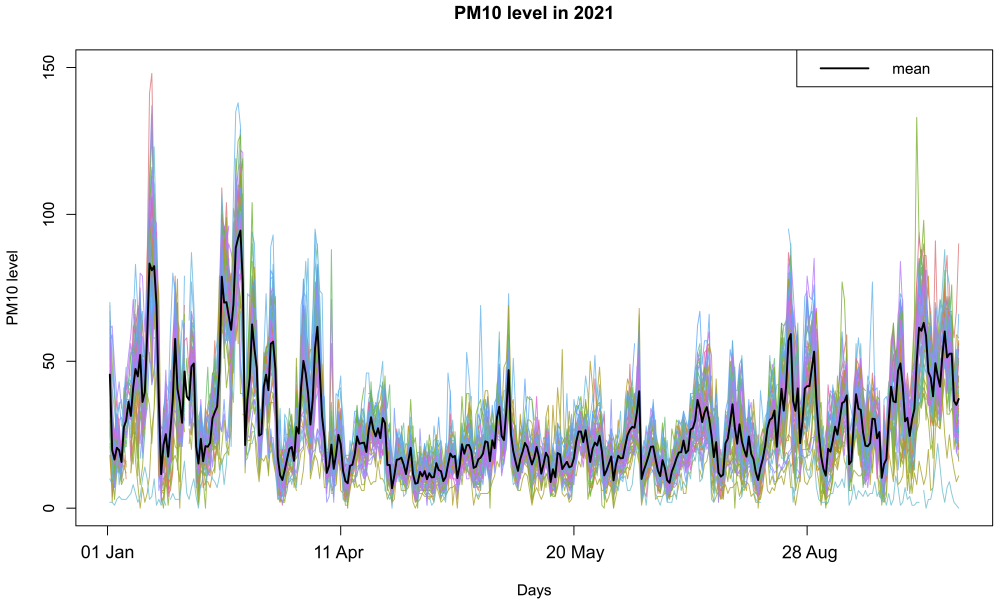


Figure 2: PM_{10} in 2021 ($\mu g/m^3$)

In order to have a continuous series during the months when pollution is greatest, i.e., the coldest months, we decided to consider years starting from September, recording in these way the highest values of pollution in the first part of the year (months from September to March) and the lowest values in the last part of the year (from May to August). This partition will be used for all remaining analysis.

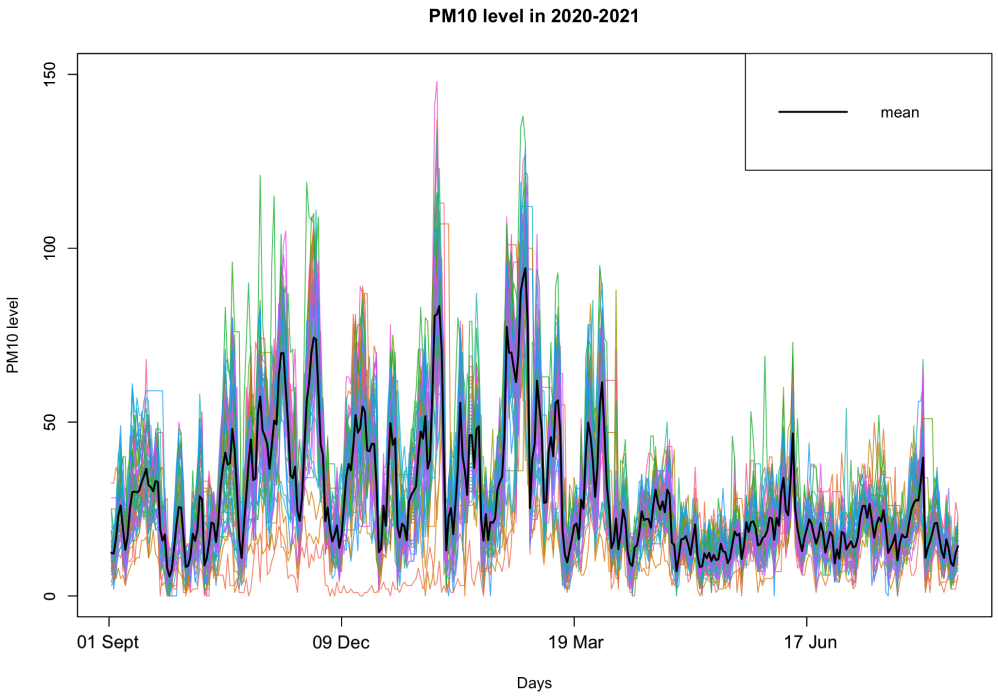


Figure 3: PM_{10} in 2020/2021 ($\mu g/m^3$)

2.2 Meteorological readings dataset

Later on, we considered the dataset regarding the measurements related to the weather parameters. We focused our attention on three of them: Temperature, Rainfall and Wind Speed. Before going into detail on each of them, we note that, once again, the trend of the 65 stations is strongly correlated in all three quantities.

From the plot of the time series regarding the Temperature we see that, as we expected, lower values are recorded in the colder months, and higher values in the warmer ones, thus going to build an opposite trend if compared to the shape characterizing the pollution. Although one would think that, given the exactly opposite trends in the time series regarding pollution and temperature, high temperatures would favor lower pollution, the issue is actually more complicated, since the warmer months also coincide with less domestic heating, which is one of the main causes of particulate production. However, we will investigate the relationship between these variables later in the research.

On the other hand, rainfall measurements show no periodic pattern: most days (70%) have zero rainfall and peaks are well distributed throughout the year. This variable will enter into our analysis since it has considerable effectiveness when purifying the atmosphere of particulate matter, especially with continuous rains. In some parts of our analysis it will be treated as a binary variable, given the huge number of days when it is completely absent.

Lastly, considering Wind Speed, we still see an absence of periodic pattern, with well-distributed ups and downs and almost no days windless. It is well knows that wind carries air contaminants

away from their source, causing them to disperse. In general, the higher the wind speed, the more contaminants are dispersed and the lower their concentration.

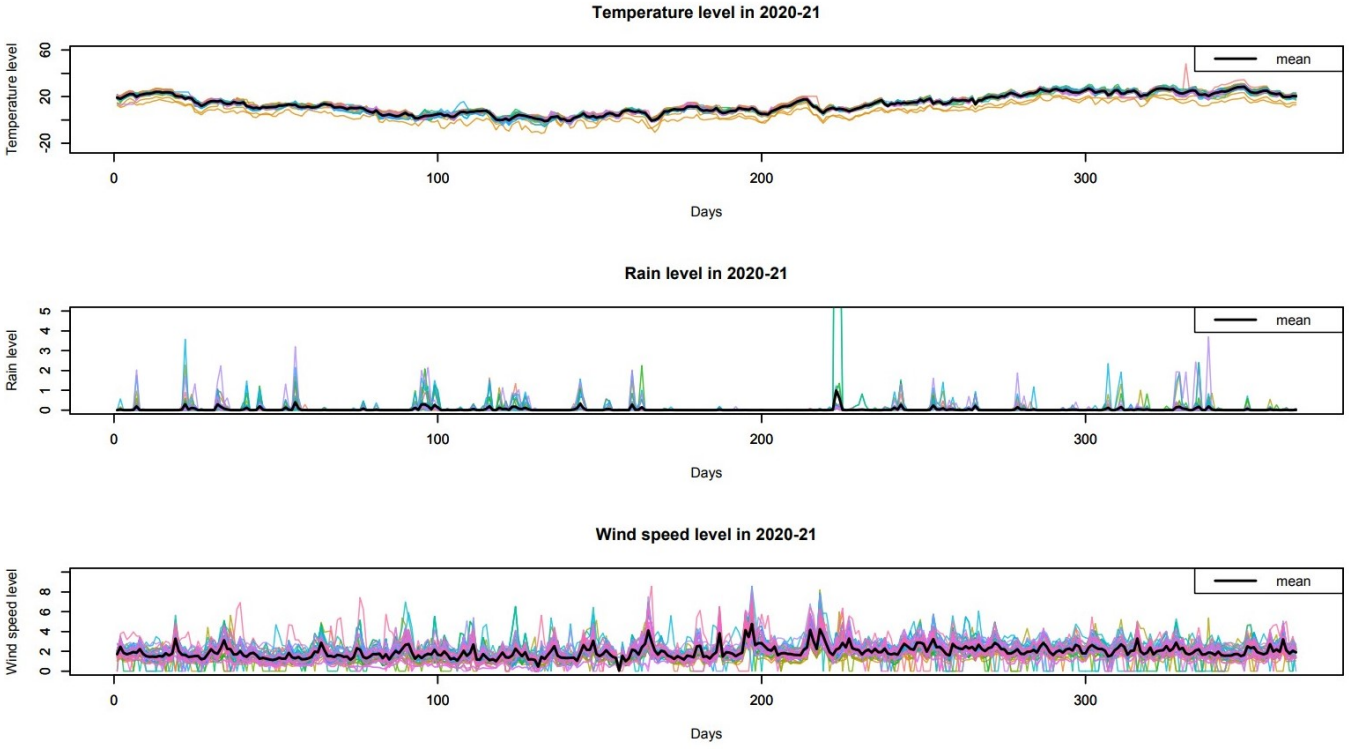


Figure 4: 2020/2021 levels for temperature(°C), rain(mL) and wind speed(m/s)

3 Depth measures

The first step to proceed after seeing the distributions is to look for possible outliers, or in other words stations that behave in a different way. We will use depth measures to create a nonparametric and more robust environment. Since we are dealing with functional data, the meaning of outliers is a bit different than the traditional univariate setting. We will look for magnitude outliers, which differ from the bulk of the distribution because of the amplitude of the curve. Moreover, we are especially interested in the presence of shape outliers, for which the difference is in the phase of the function.
For the amplitude outliers we utilize the functional boxplot which outputs the stations in figure 5. 10 stations are flagged as outliers: it is clear they correspond to either very high or very low pollution levels. The explanation is pretty straightforward, as low PM_{10} are associated to stations found on high altitudes, like Moggio and Bormio. On the other hand greater concentrations refer to zones with a high population density, such as Milano and Bergamo. Since we have information on the altitude of the stations and their geographical location we will not remove these curves from our analysis, but instead we will try to explain them using these variables.

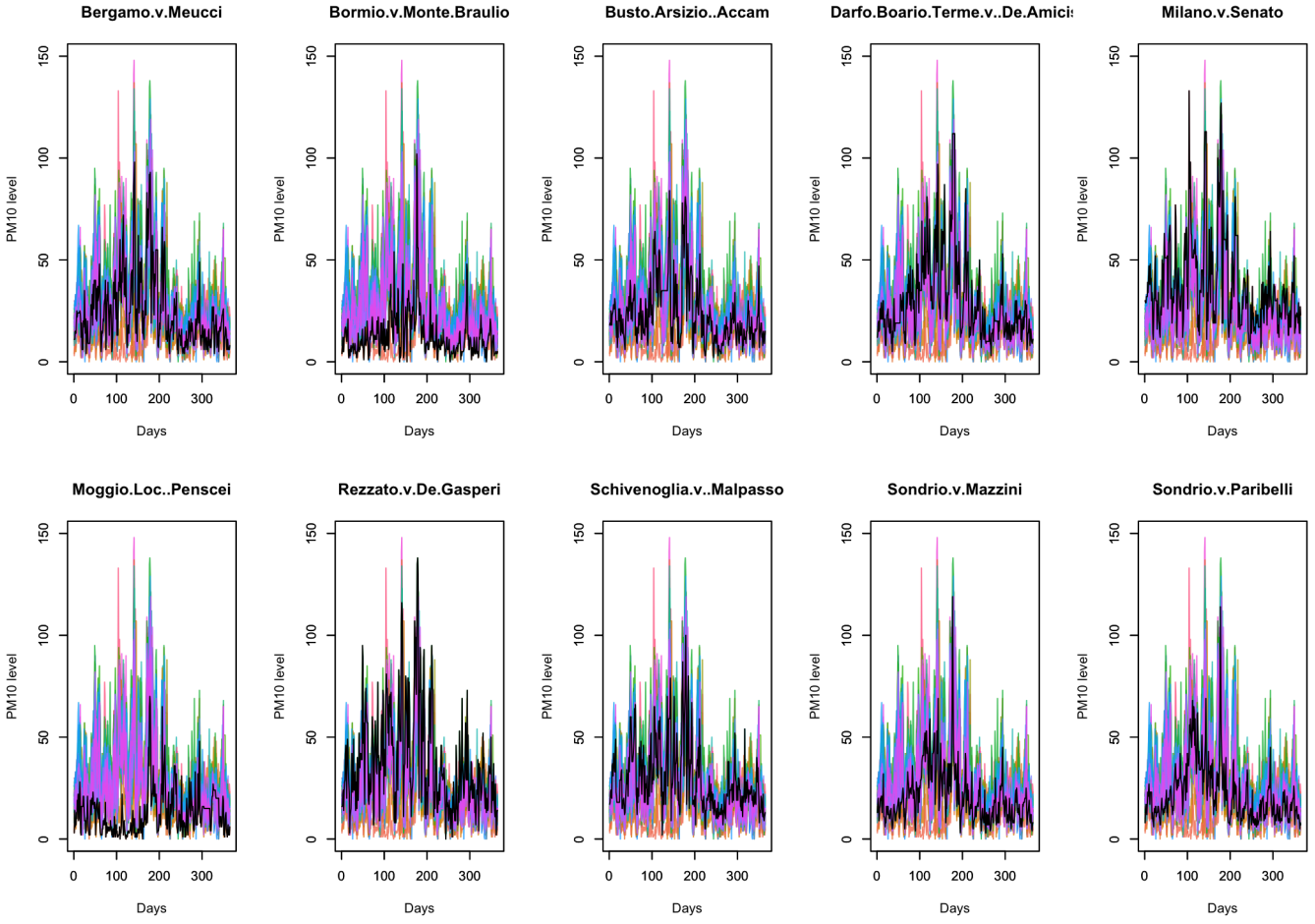


Figure 5: Magnitude outliers

Concerning the shape outliers, we use the outliergram to detect them. Almost all the stations belong to the safe area besides one, which is slightly off and does not represent a big concern. This makes some good sense as the trend of the time series is common as it refers to the seasonal periodicity.

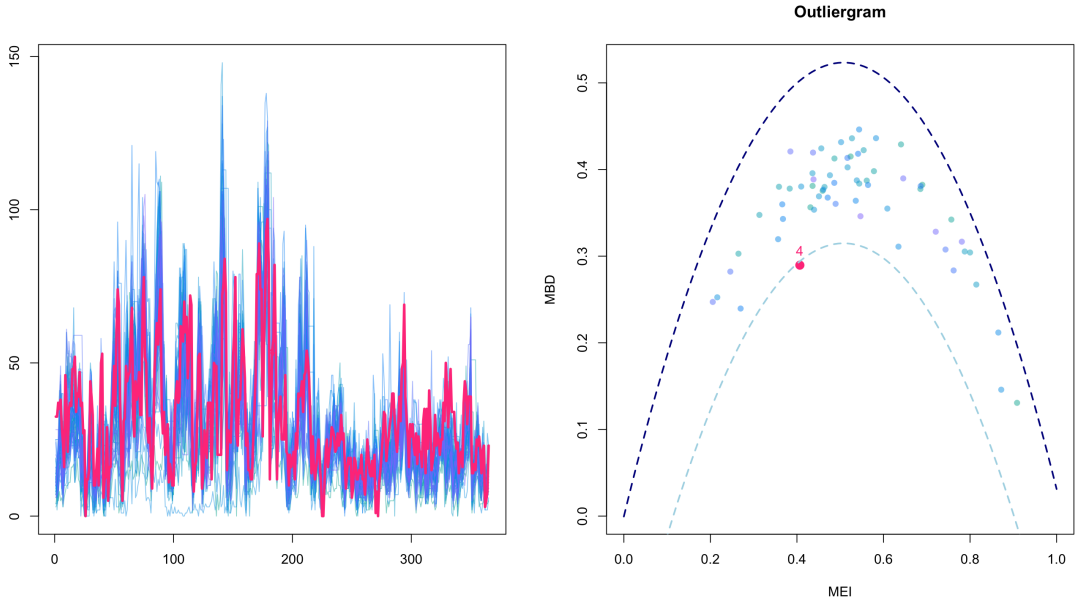
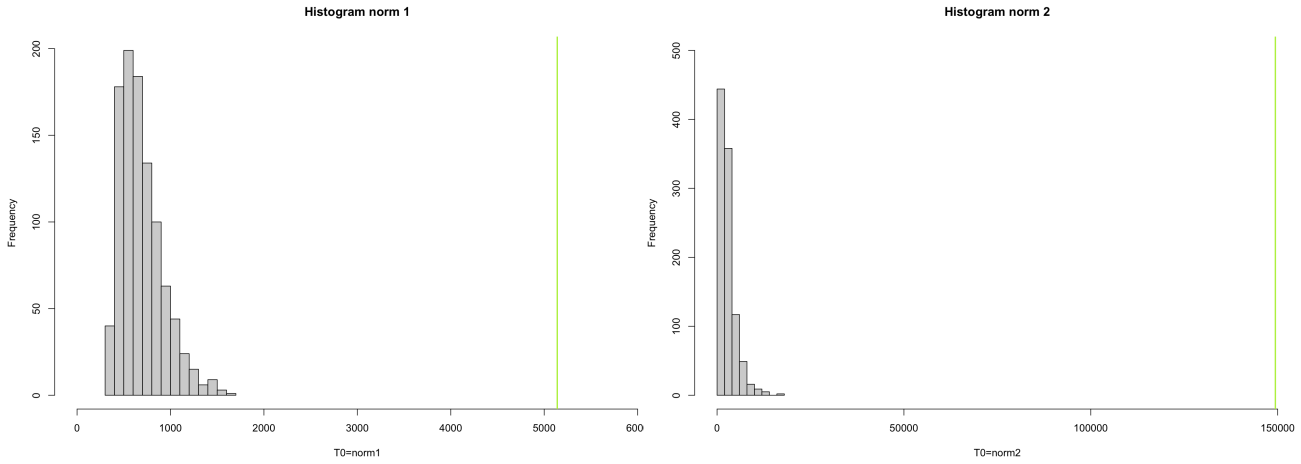


Figure 6: Shape outliers

4 Permutational tests

In this section we want to make use of permutational tests to compare distributions of PM_{10} over different years. In particular, we will report here the results for the temporal extrema of the dataset, namely 2016 and 2021, but similar results can be easily replicated for all the available years. Bear in mind that we will work in a functional setting, treating the yearly time series produced by each of the station as the i-th datum. This seemed like a proper approach to take care of the temporal dimension without modifying the information provided to us. A weak point resulting from this approach is of course the fact that we cannot assume the stations to be independent, as they for sure share spatial information, but the insight gained from this technique is still relevant. The test we set up is the following:



(a) Histogram using norm 1: p-value=0

(b) Histogram using norm 2: p-value=0

Figure 7: Results of the permutational test

In figure 7 we present results. It is clear that the p-value is 0 and the value of the test statistics with no permutation is far from the ones obtained exchanging stations. We cannot say that the two distributions are the same. Even though we get this result, when comparing the curves in figure 9a it appears little to no difference in the mean of the two distributions, but rather in the different peaks. We think therefore the real meaning of this p-value is that distributions are different, but not in what we are interested in. Our goal is indeed to understand if there is an increase in pollution over time, and consequently in the mean of our distributions. From a qualitative point of view, it does not seem there is one. To provide support to this hypothesis, we conudcted a piecewise permutational test to understand in which parts of the domain the curves are different.

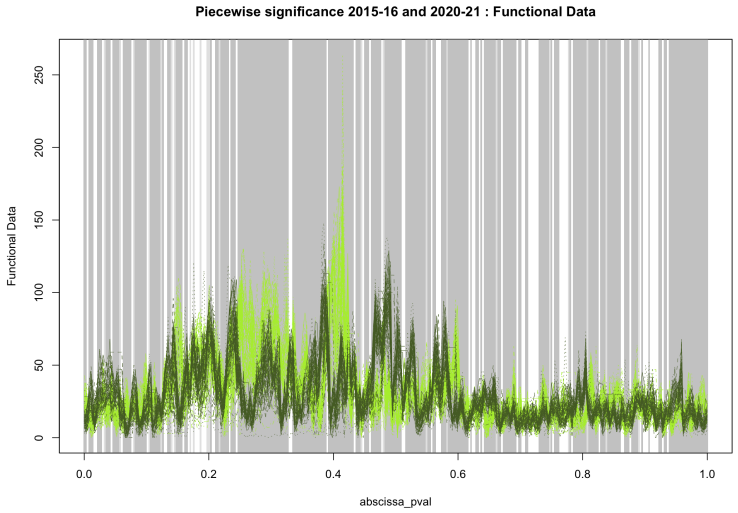
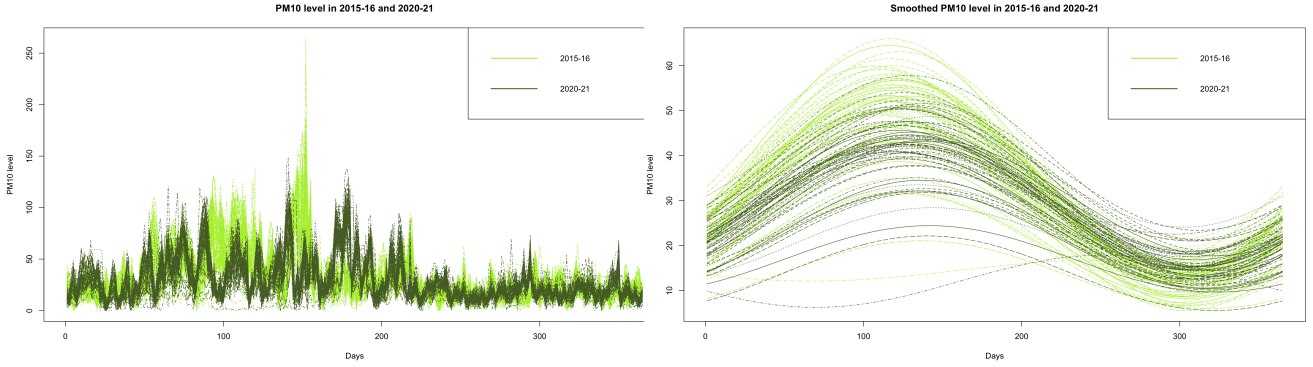


Figure 8: Piecewise p-value function

Results are shown in figure 8: the p-value function is very fragmented, alternating between sections where the p-value is high and low, providing evidence to our initial guess. Indeed if the curves were different in mean the p-value function would be uniform over low values. To further provide a more rigorous approach, we decided to smooth the curves using a truncated expansion in Fourier basis to extract the fundamental shape of the curve and leave out the peaks that seem to be causing this big difference. This method allows to construct a Fourier expansion whose coefficients are determined minimizing the approximation error on the curves. We will keep just a few basis in order to recreate the annual periodicity. The smoothed curves are shown in figure 9b: it is clear we are capturing only the backbone of the data removing the peaks we are not interested into for the moment. After smoothing the curves we conducted the same permutational tests as before, including also the piecewise one



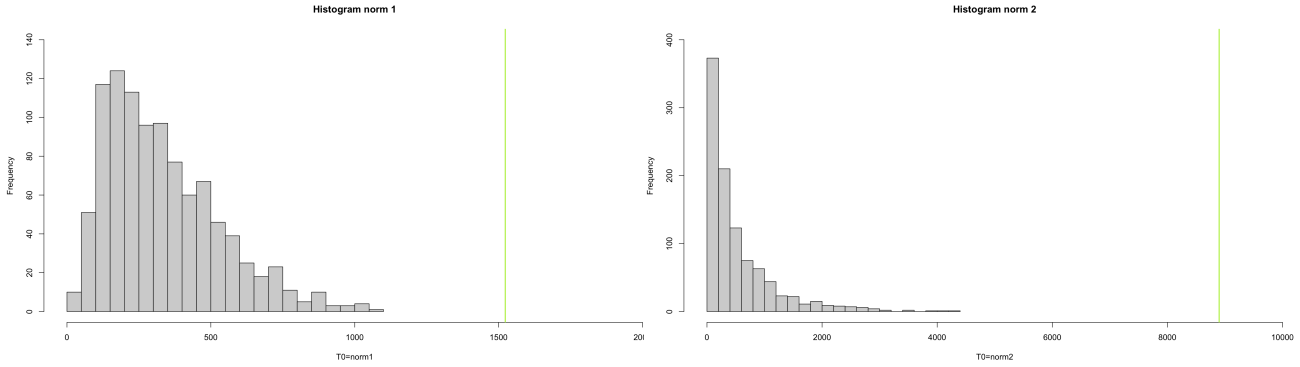
(a) PM_{10} time series in 2016 and 2021

(b) Smoothed PM_{10} time series in 2016 and 2021

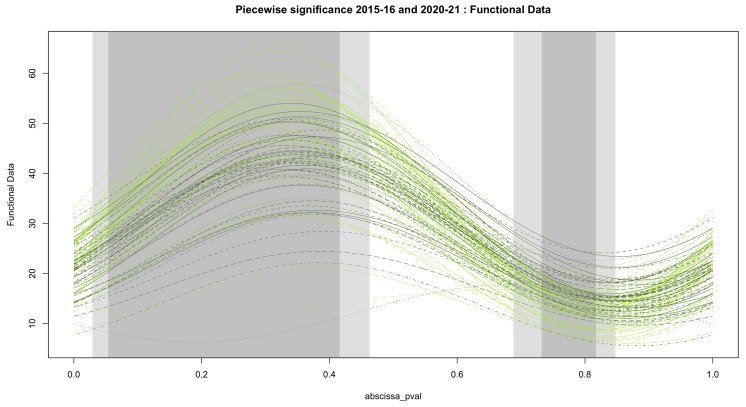
Figure 9

Results are shown in figure 10. Again the p-value for the global test is 0, even though the original test statistics is closer to the permuted ones. We gain further insight when looking

at the piecewise p-value function: it is only in the last part of the year that we can assume the distributions are different. Indeed in the first months of the year we cannot refuse H_0, concluding the smoothed distributions are the same. The p-value equal to 0 is driven by the strong difference in the last 3 months of the year, rather than a global one. Also by visual inspection it seems the curve of 2016 have a mean which is greater than 2021. The conclusion is that we think the difference of the two years is not in the mean and in the general trend, but mainly in the peaks that could be driven by external factors.



(a) Histogram using norm 1 on smoothed data: p-value=0 (b) Histogram using norm 2 on smoothed data: p-value=0



(c) Piecewise p-value function on smoothed data

Figure 10: Results of the permutational tests on smoothed data

A possible way to better understand how PM_{10} behaves is to include it in a single model with the weather covariates we are given. In order to understand if there is a correlation between these variables we resort to a nonparametric measure, which is the Spearman index. Given the model we will try later, this seems like a proper procedure: this technique gives information on monotonic correlation, rather than on linear relationship. The Spearman correlation matrix is shown in figure 11. The row we are interested in is the last one: it is clear there is a strong correlation especially with temperature and wind. Again we remind this index has been computed on functional data, therefore we are comparing the yearly distribution of these functions. Moreover, all the indexes are negative, indicating an opposite trend in the distributions (see also figure 4), which is expected. Also rain is negatively correlated, even though the value is smaller in comparison. We conclude it is worthy to construct a model to exploit these correlations: in

particular we will use a Generative Additive Model.

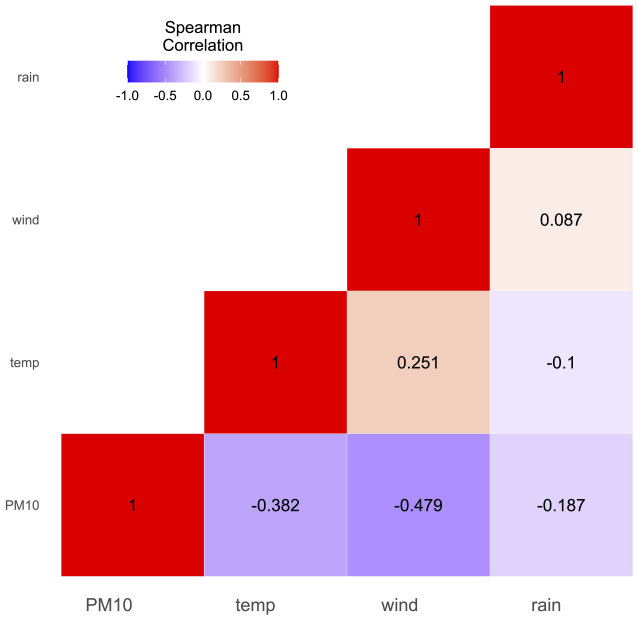


Figure 11: Spearman correlation matrix

5 GAM

In this section we want to exploit the relationship between the covariates and the concentration of PM_{10} to gain some insight and possibily predict the pollution level. The best nonparametric method we came up with is a Generalized Additive Model taking care of the covariates in a smart and straightforward way. In this sense data will be treated in a different manner with respect to the previous section, as we will partially lose the functional approach. Indeed we will regress the log-PM_{10} at time t and station i over the corresponding covariates in the same conditions. This method allows to consider independent basis over each direction we regress on: this characteristic, that simplifies the interactions between covariates, is actually very advantageous in our case. Indeed we will be able to explore the different dimensions independently, useful for the interpretation we will give to the coefficients of the model. To be more specific, we will make use of a semiparametric model including both a nonparametric smoothing section and a linear parametric one. Let us write the model and then explain for each term the choices we made.

$$\begin{aligned} log\, PM10_{t,i} \sim \beta_0 + \beta_1 \cdot wind_{t,i} + \beta_2 \cdot altitude_i + \beta_3 \cdot rainfall_{t,i} + \\ f_{\text{cyclic}}(t) + f_{\text{smoothing}}(temperature_{t,i}) + f_{\text{smoothing}}(latitude_{t,i}, longitude_{t,i}) \end{aligned} \quad (1)$$

$$t = 1, \ldots, 365 \quad i = 1, \ldots, 65$$

- $log\, PM10_{t,i}$: logarithmic transformation applied to the measurement of the i-th station at time t. Its unite measure is $\mu g/m^3$.
- Linear regression part
  - β_0: constant term
  - $\beta_1 \cdot wind_{t,i}$: wind speed has been treated linearly since also when using more complicated nonparametric smoothing function the behaviour resembled a linear one. Its unite measure is m/s.
  - $\beta_2 \cdot altitude_i$: constant value for each station. Its unite measure is m.
  - $\beta_3 \cdot rainfall_{t,i}$: the rain information has been converted to a binary variable (rain/no rain). We believe the key information is not in the quantity of rain that falls but rather in the occurrence of the event.
- $f_{\text{cyclic}}(t)$: The time information will be dealt with a cubic cyclic spline. The benefit of this basis is in the extrema of the domain, that normally do not have any constraint. In this case instead they are bound to coincide, to replicate the annual periodic behaviour that the time series presents. This allows for a more correct modeling of time, as this process possesses an intrinsic seasonality we need to take care of.
- $f_{\text{smoothing}}(temperature_{t,i})$: when missing further information, as a first approach we used a cubic natural spline which is the standard. Its unite measure is °C.
- $f_{\text{smoothing}}(latitude_{t,i}, longitude_{t,i})$: the spatial information had to be dealt with in a different way. Indeed treating latitude and longitude independently would be too naive, in the sense

that it is the interaction of these two terms that actually determines the location of the station, and consequently the relevant information. For this reason we are making use of a tensor product cubic spline, which creates a 2-D grid of smoothing coefficients. Since latitude and longitude have the same unit measure, we do not require a te spline, as the smoothing one will take care of producing the grid we need.

5.1 Results

This model reaches a R^2 of 0.431 meaning the model explains almost half of the variability, which is a reasonable and satisfying result on real data. Moreover, all the included covariates are significant with a very low p-value. Let us now look at the estimated coefficients numerically and graphically in order to understand how the covariates affect the output.
The time information is dealt with successfully, as the non-linear shape is coherent with the seasonal trend. Indeed we have an increase in autumn and winter, while with spring the trend starts decreasing reaching the global minimum during summer. The smoothing is located in a range between -1 and 1, therefore the impact on the corresponding estimated $logPM_{10}$ is quite important. This estimate will basically catch the trend of the time series, onto wich the other covariates will build further.

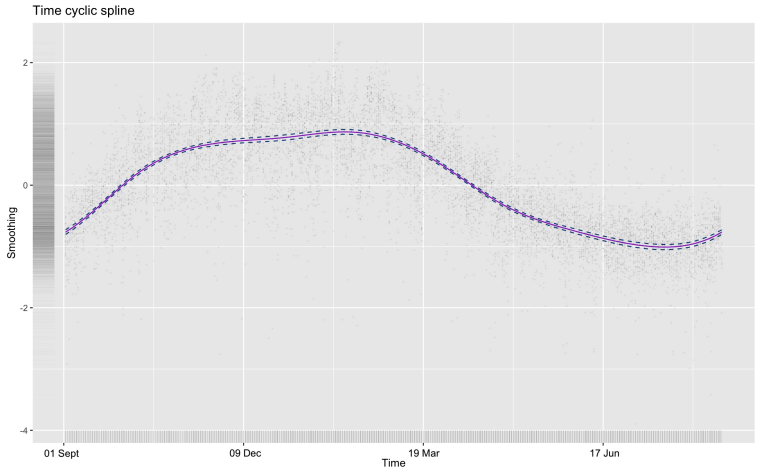
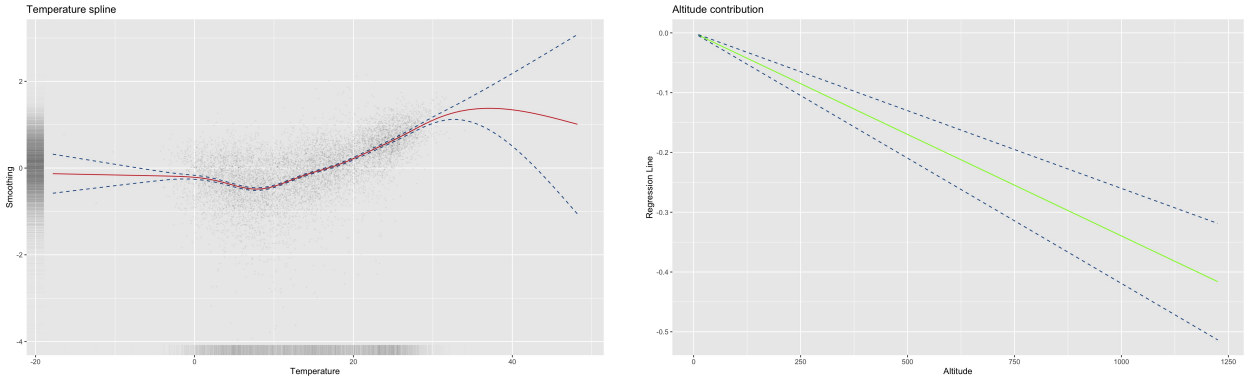


Figure 12: Time cyclic spline smoothing

Results for temperature smoothing are shown in figure 13a. The non-linear behaviour could be approximated with a constant line before a threshold of 7-8 °C, and a linear growth after this point. Remember here we are keeping all the other covariate fixed, so this means that in the same conditions a higher temperature leads in general to higher level of pollution, as the function is increasing. Again the smoothing is significant as it covers the range 0-1.
The β_2 associated with altitude is negative: $\beta_2 = -3.397 \cdot 10^{-4} \mu g/m^3 \cdot m^{-1}$ (figure 13b). As expected stations with higher altitude are associated with lower values of PM_{10}: in general pollution is greater near big cities which are in plain.

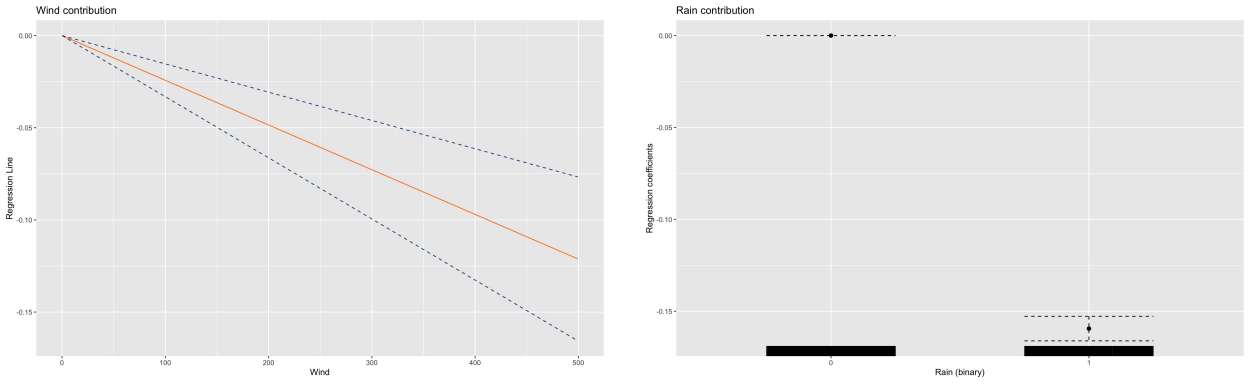


(a) Temperature cubic spline smoothing

(b) Altitude linear contribution

Figure 13

Wind speed has a positive effect on reducing PM_{10} level: the associated $\beta_1 = -2.425 \cdot 10^{-4} \mu g/m^3 \cdot (m/s)^{-1}$ (see figure 14a). It makes sense that with stronger winds the particles get scattered more easily, resulting in a better air quality. The contribution is smaller comparing it to the other covariates.
The presence of rain also decreases the PM_{10} value, as $\beta_3 \approx -0.159 \mu g/m^3 \cdot (ml)^{-1}$. This is also a well known effect of the rain which contributes in cleaning the air and dissolving pollutant particles.



(a) Wind speed linear contribution

(b) Rain (right)/no rain(left) contribution

Figure 14

In figure 15 we can observe the spatial estimate we get from our GAM model: the most important observation is that in the north of Lombardia the associated pollution level is lower, which makes sense as it is the area with mountains. Instead near the big cities we have higher estimated values for PM_{10}. The resulting map is quite satisfying and provides more insight on where pollution is greater and should be kept more under control.

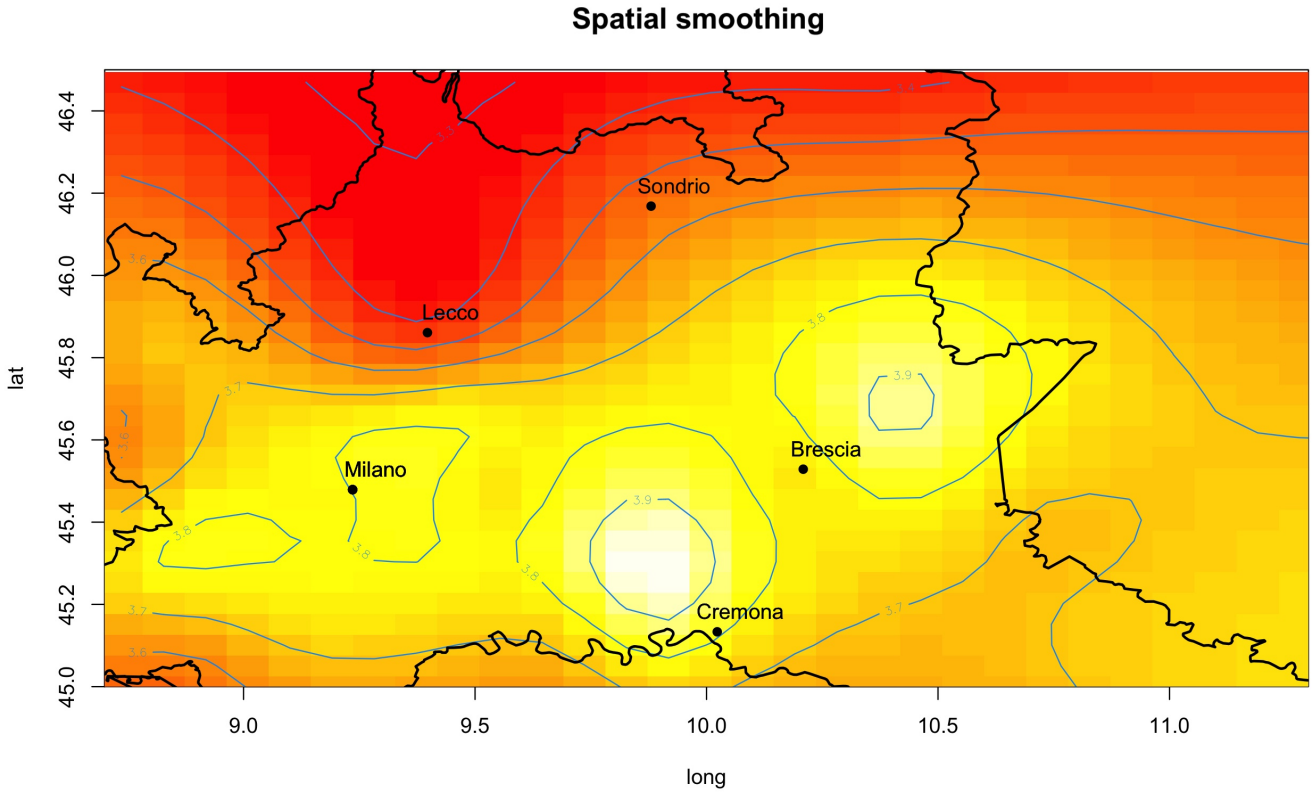


Figure 15: Contour plot

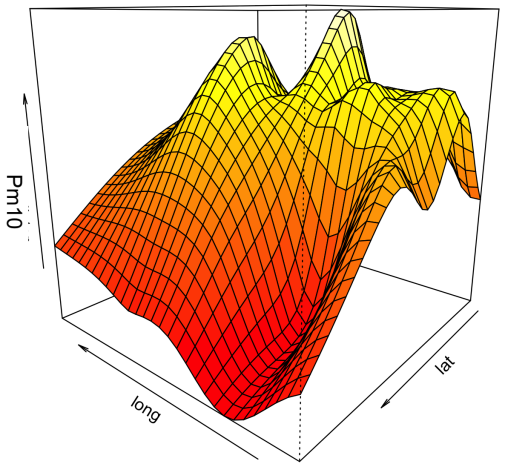


Figure 16: 3D plot

5.2 Dummy for the year

The second GAM model we propose also includes a dummy on the year, which can provide some answers to our first research question. The resulting model is

$$log\ PM10_{t,i} \sim \beta_0 + \beta_1 \cdot wind_{t,i} + \beta_2 \cdot altitude_i + \beta_3 \cdot rainfall_{t,i} + \boldsymbol{\beta_4} \cdot Year \\ f_{\text{cyclic}}(t) + f_{\text{smoothing}}(temperature_{t,i}) + f_{\text{smoothing}}(latitude_{t,i}, longitude_{t,i}) \quad (2)$$

$$t = 1, \ldots, 365 \quad i = 1, \ldots, 65$$

where $\boldsymbol{\beta_4}$ is a vector of dimension $\#years - 1$ and year is a categorical variable. The results for the variables also in model 1 are basically the same. In figure 17 we can observe the estimated coefficients associated to the year, which are all significant. It seems our initial guess is confirmed in the estimates: after 2015/2016 there is an increase in pollution, while PM_{10} level decreases in the COVID years.

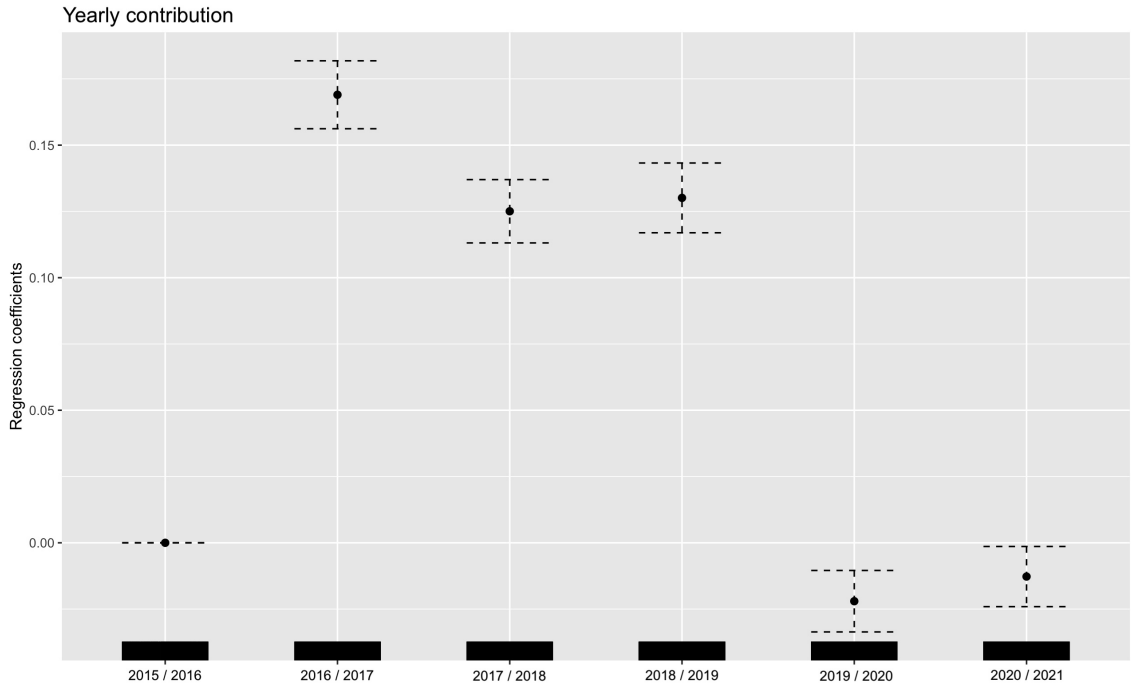


Figure 17: $\boldsymbol{\beta_4}$ associated to Yearly categorical

6 Conformal prediction

To validate our model we resort to Conformal prediction techniques applied in the GAM framework. In particular, we will use the last 4 months of 2021 to predict PM_{10} concentration, to understand how well the model behaves. This technique allows to produce conformal bands around the pointwise prediction produced by the GAM model, with a given confidence we chose as 95%.

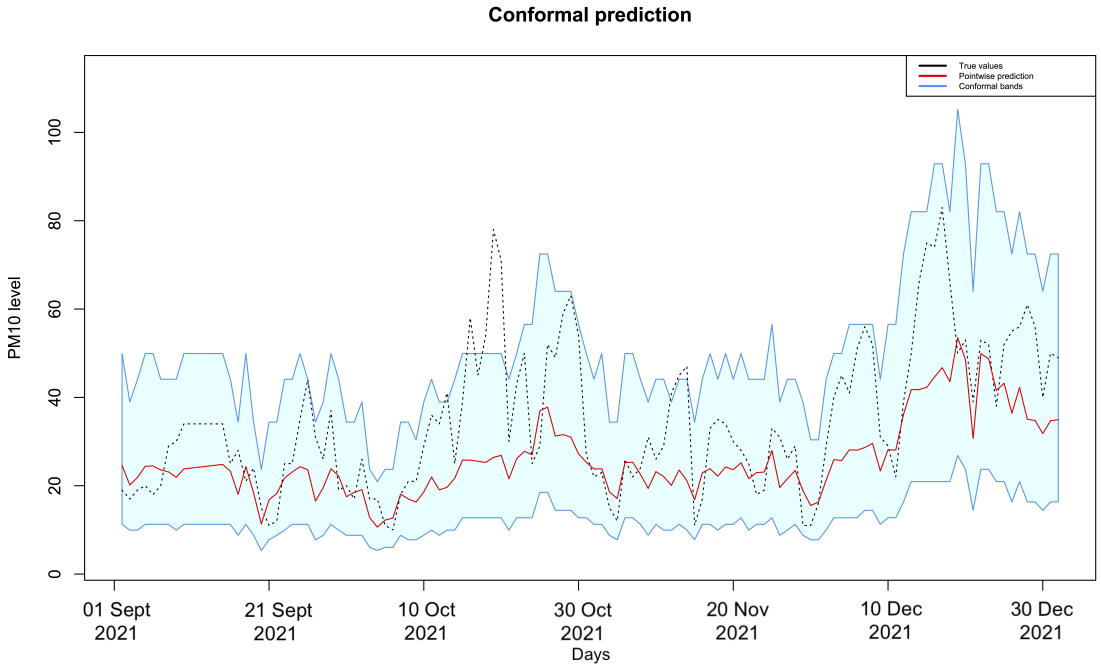


Figure 18: 95% conformal prediction bands for September-December 2021

Results are shown in figure 18. Let us first comment on the pointwise prediction to evaluate performances of the GAM model: we are able to capture the seasonal trend, which seems to be the main component driving pollution level. Not only that, but also some of the peaks are well modeled, even if most of the up and downs cannot be so easily described. It becomes clear that 95% prediction bands instead provide very satisfying regions that include almost all the observations. This framework could be utilized having weather forecast under hand to provide an effective range for the PM_{10} level.

7 Conclusions

Our analyisis was able to provide some interesting insights. Regarding the possible increase of pollution level over the years, we can create an outline of the years with two different tendencies. After 2016 there is an increase in PM_{10} concentration, which remains throughout the years until the COVID pandemic, which caused a decrease, surely because of the fact that most of the pollutant activities stopped. This evidence was provided by the permutational tests and the GAM model showing these trends. Morevoer, this approach allowed to detect the weather factors that contribute in determining pollution. The general trend is of course dictated by the period of the year, as the coldest months are in general associated to higher level of PM_{10} because of domestic heating. On top of this, temperature is a negative factor, as higher temperature on the same day often prevent pollution to dissolve. Strong wind speed and rain instead positively contribute to reducing $PM10$ concentration. In the end, with conformal prediction we successfully provide a setting in which weather forecast can be used to estimate with some precision the future pollution levels.