Nonparametric Statistics course - a.y. 2022/2023

Pollution and weather report in Lombardy

Lorenzo Angiolini, Giulia Bergonzoli,
India Ermacora, Lucia Gregorini

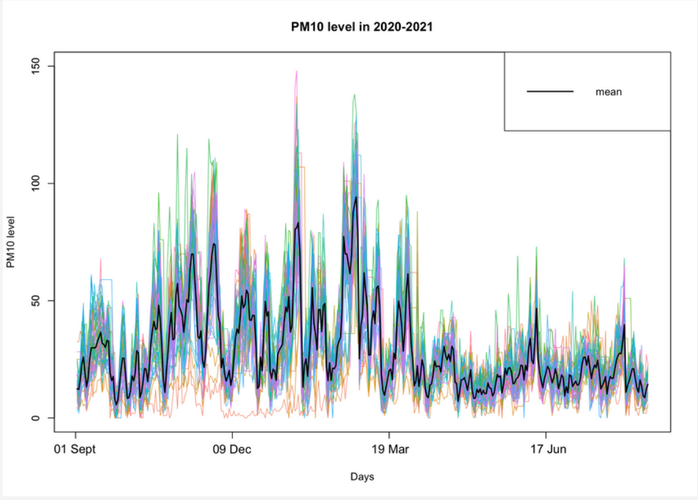
Final presentation

February 17 th 2023

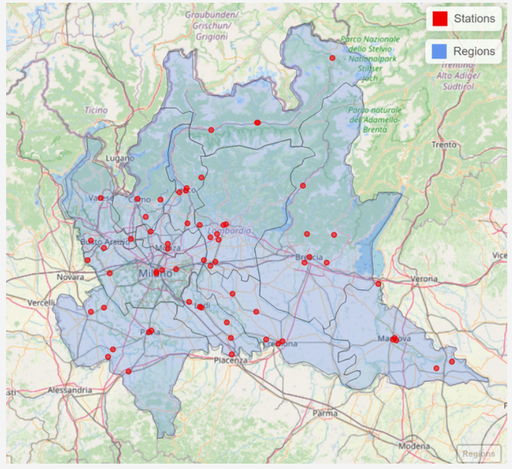
DATASET OVERVIEW

Response variable

PM10 level

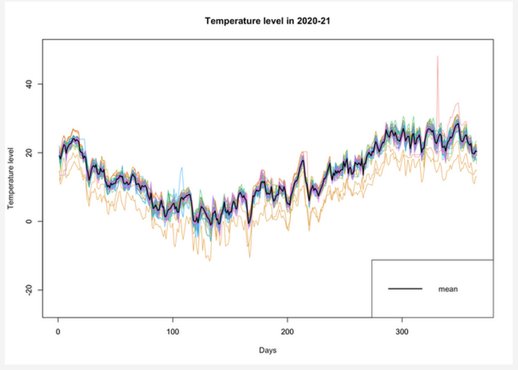


Spatial information

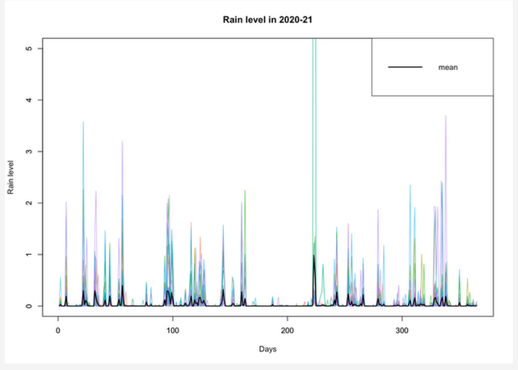


Altitude

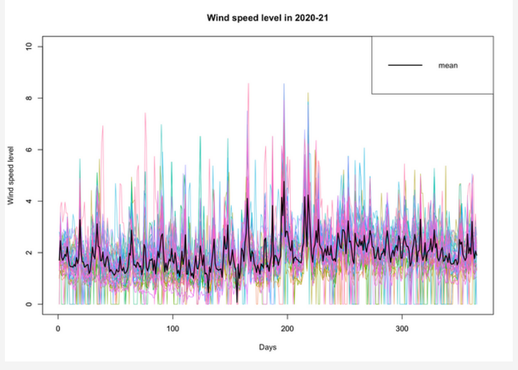
Temperature



Rainfall



Wind speed



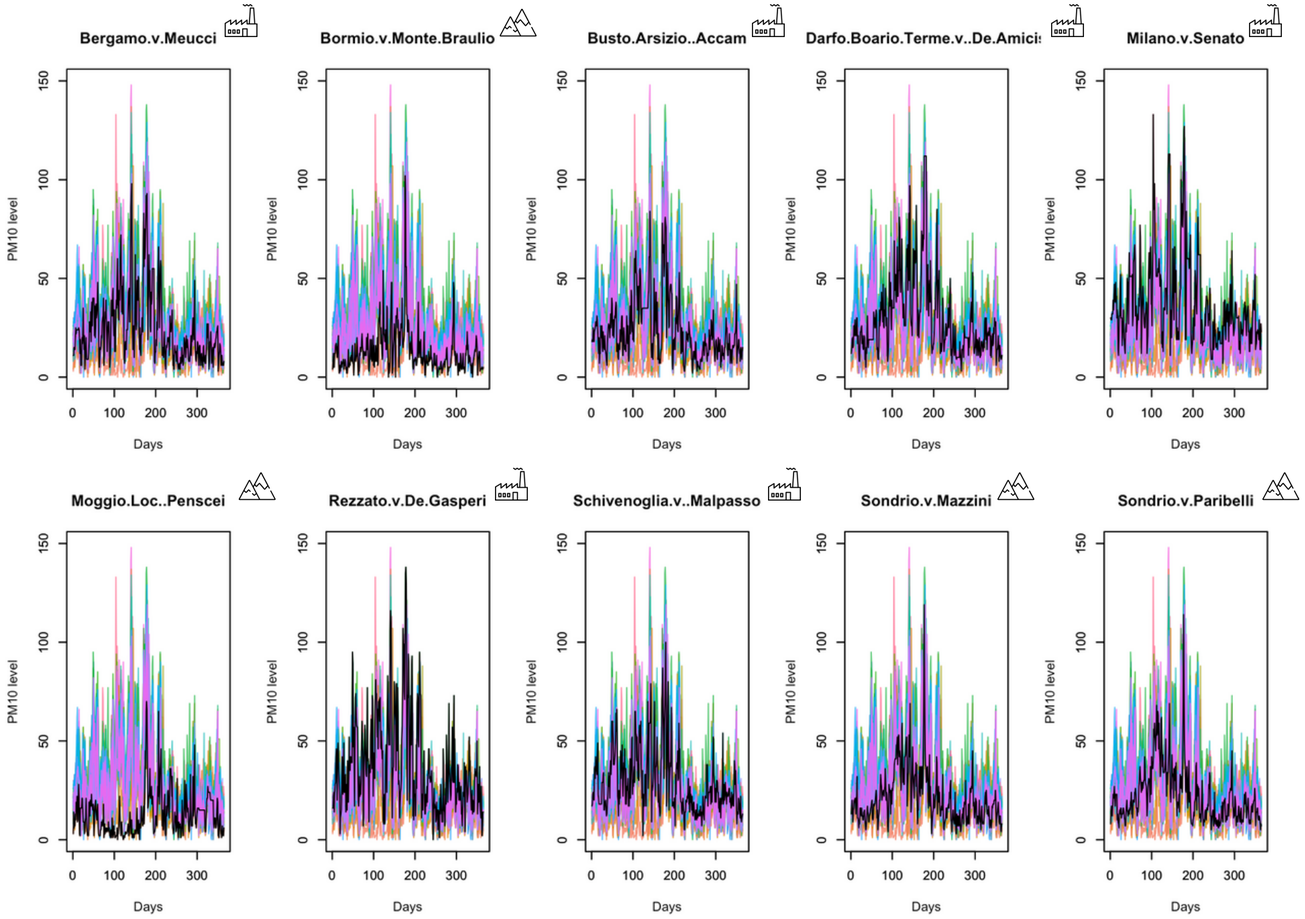
DESIRED GOALS

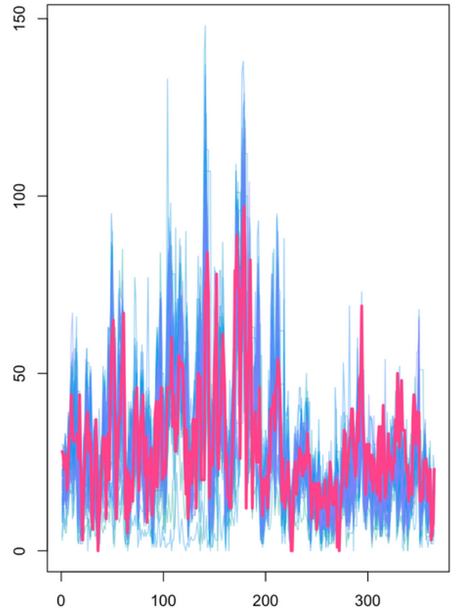
We will investigate on the **increase** of pollution level over the years, exploiting the temporal dimension of the dataset.

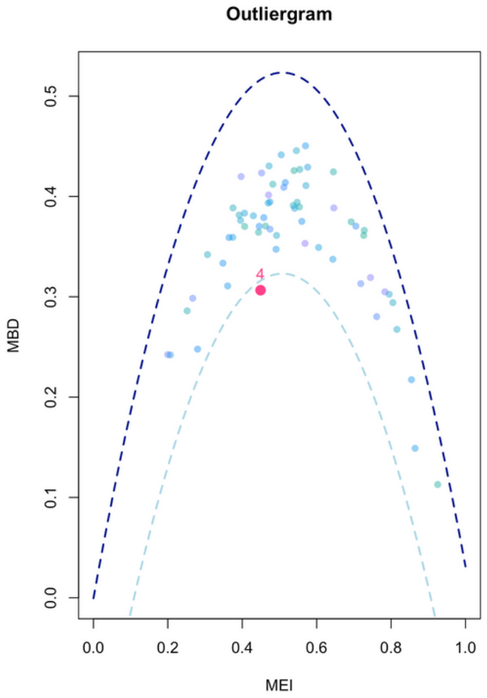
We want to exploit **weather** and **spatial** information to describe and possibily predict the concentration of PM10 over time, in order to **monitor** the pollution level.

DATA EXPLORATION

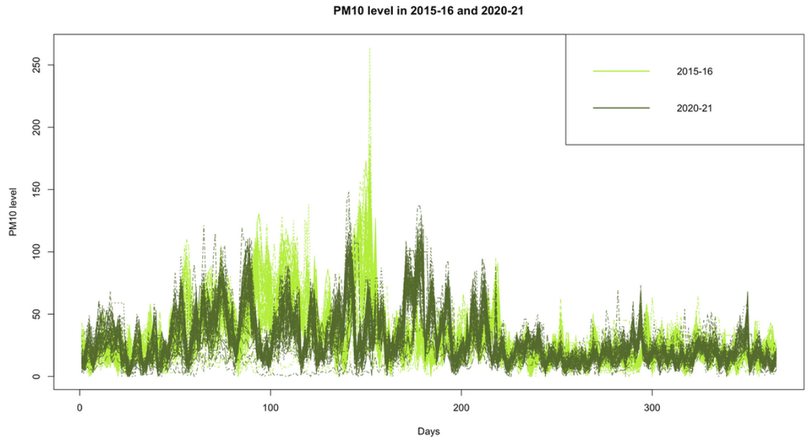
Shape and magnitude outliers

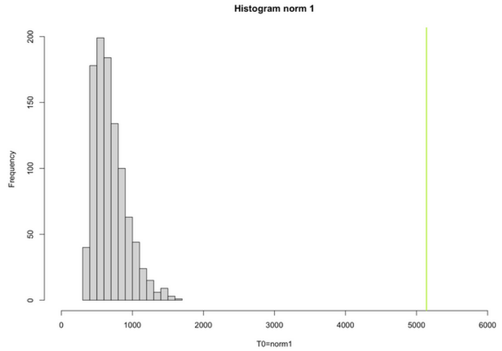


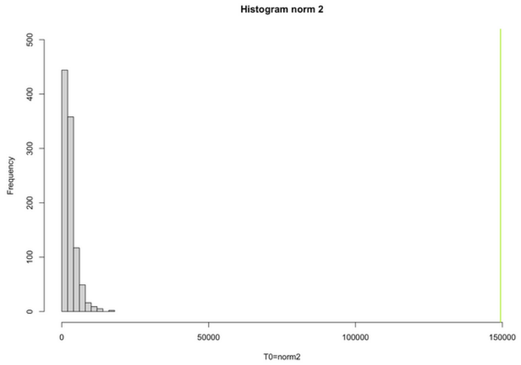




TESTING FOR POLLUTION INCREASE

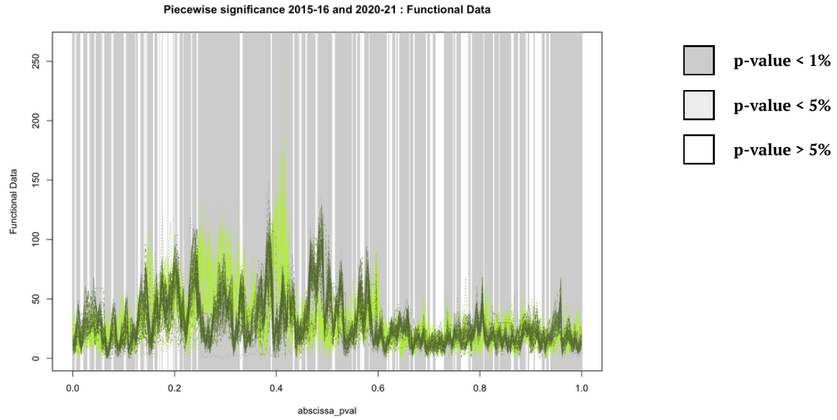


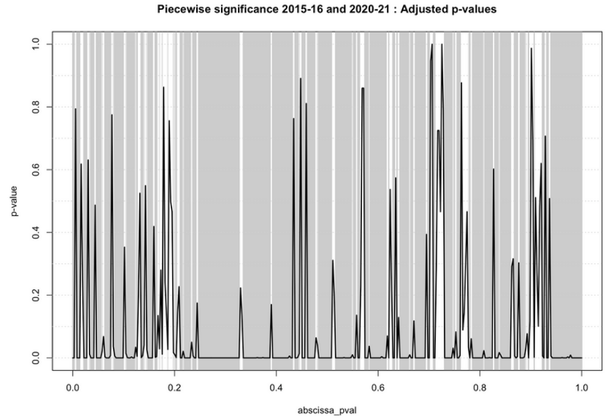




A p-value of 0 indicates the distributions are **not the same**. This might be due to the different **peaks** in the time series

Interval Wise Testing

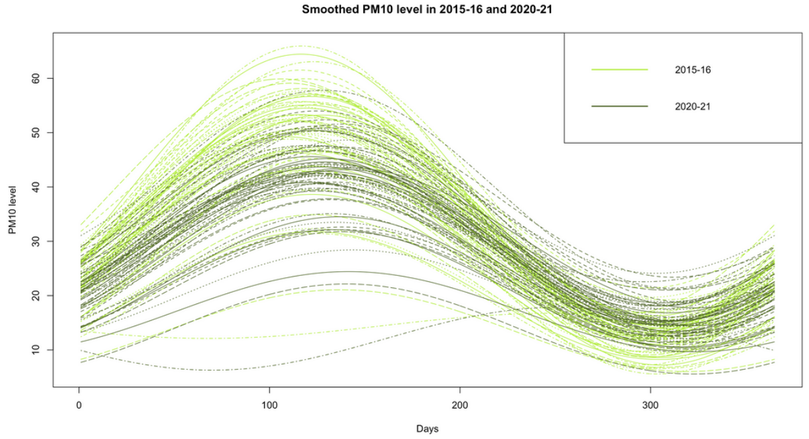


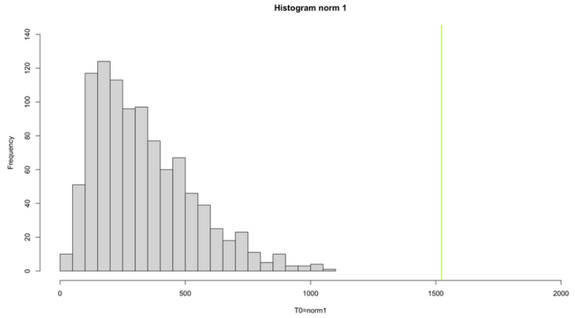


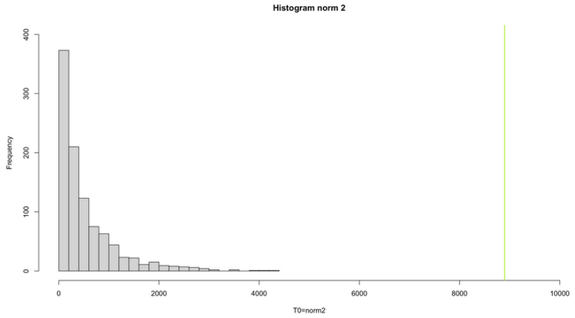
The **fragmented** behaviour of the adjusted p-value function on the domain confirms the hypothesis

TESTING FOR POLLUTION INCREASE

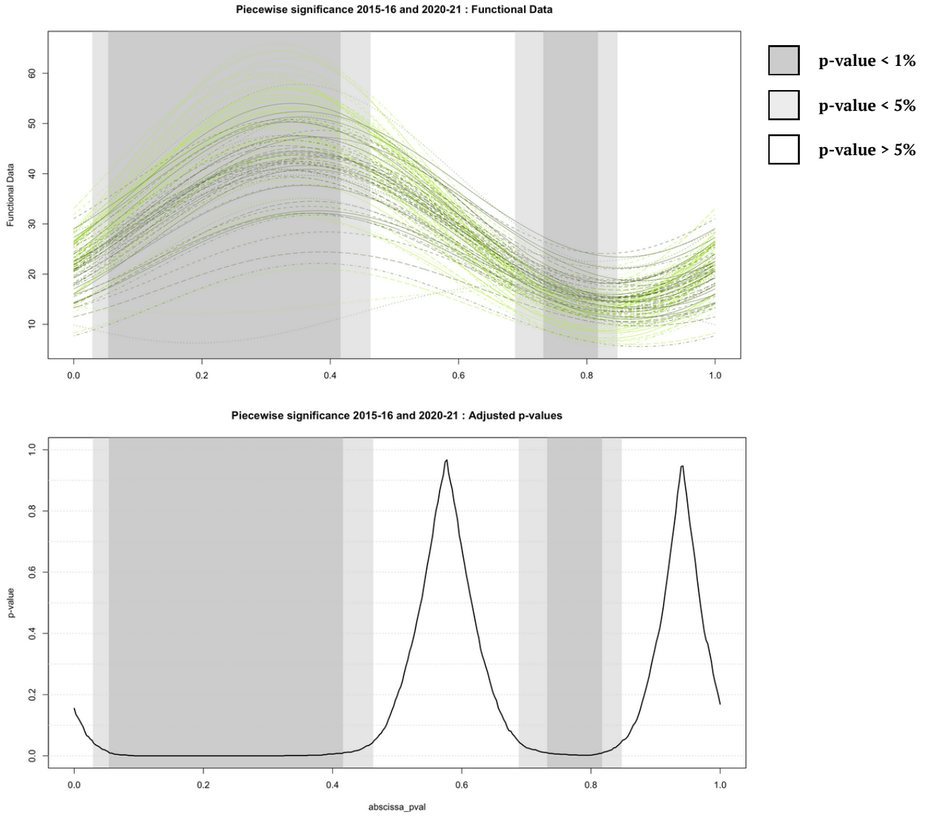
Smoothing







Interval Wise Testing



Smooth of the curves using annual **Fourier basis**

→

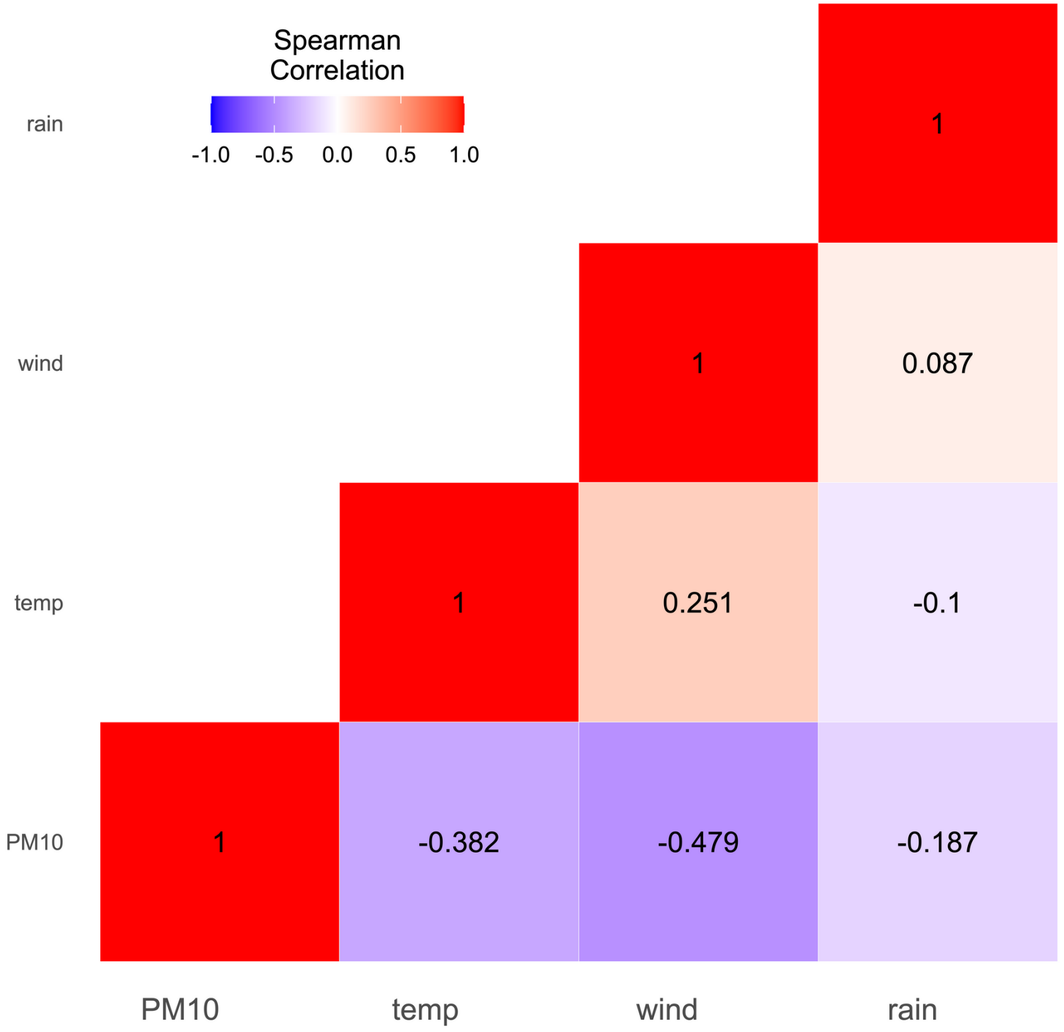
The p-value does not change: the temporal information is not sufficient on its own to explain the **entire variability**

→

To model peaks **weather** information might be relevant

WEATHER INFLUENCE

Spearman matrix



MODELING PM10 LEVEL

GAM model

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

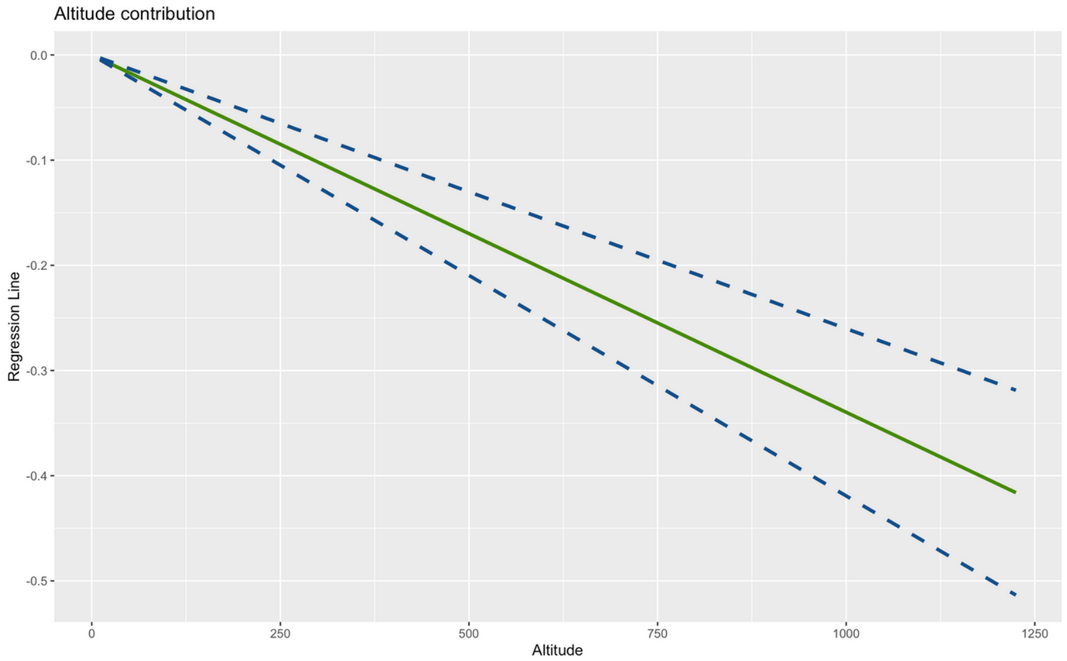
$$t = 1, \ldots, 365\ days \quad i = 1, \ldots, 65\ stations$$

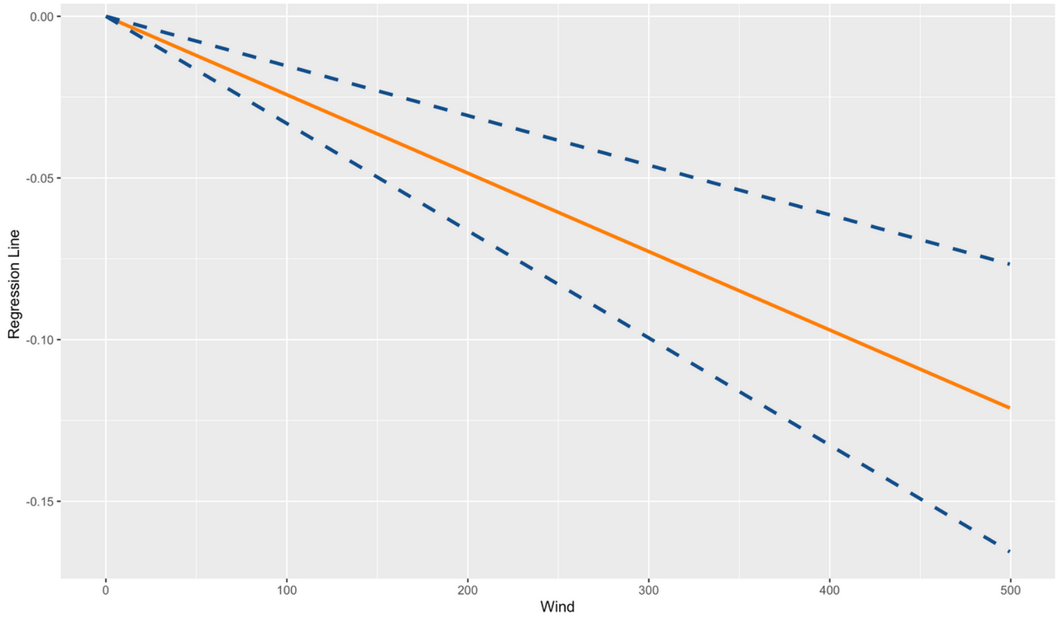
- β = coefficients of parametric part
- f_{cyclic} = cubic cyclic spline
- $f_{\text{smoothing}}$ = cubic natural spline

Reaching an adjusted R^2 of **0.441**

MODELING PM10 LEVEL

GAM model

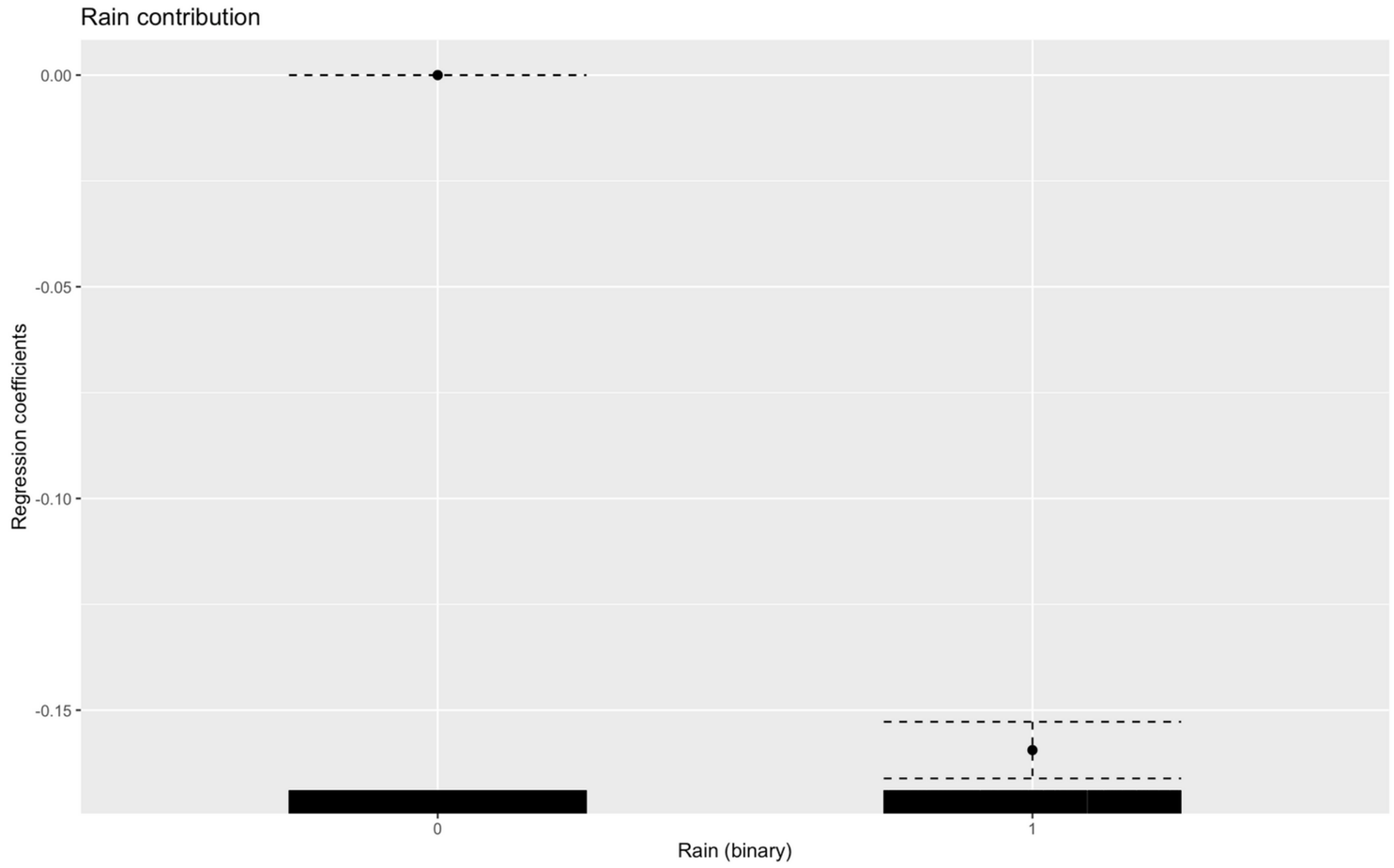




Both **wind** and **altitude** have a positive effect on **reducing** the PM10 level

MODELING PM10 LEVEL

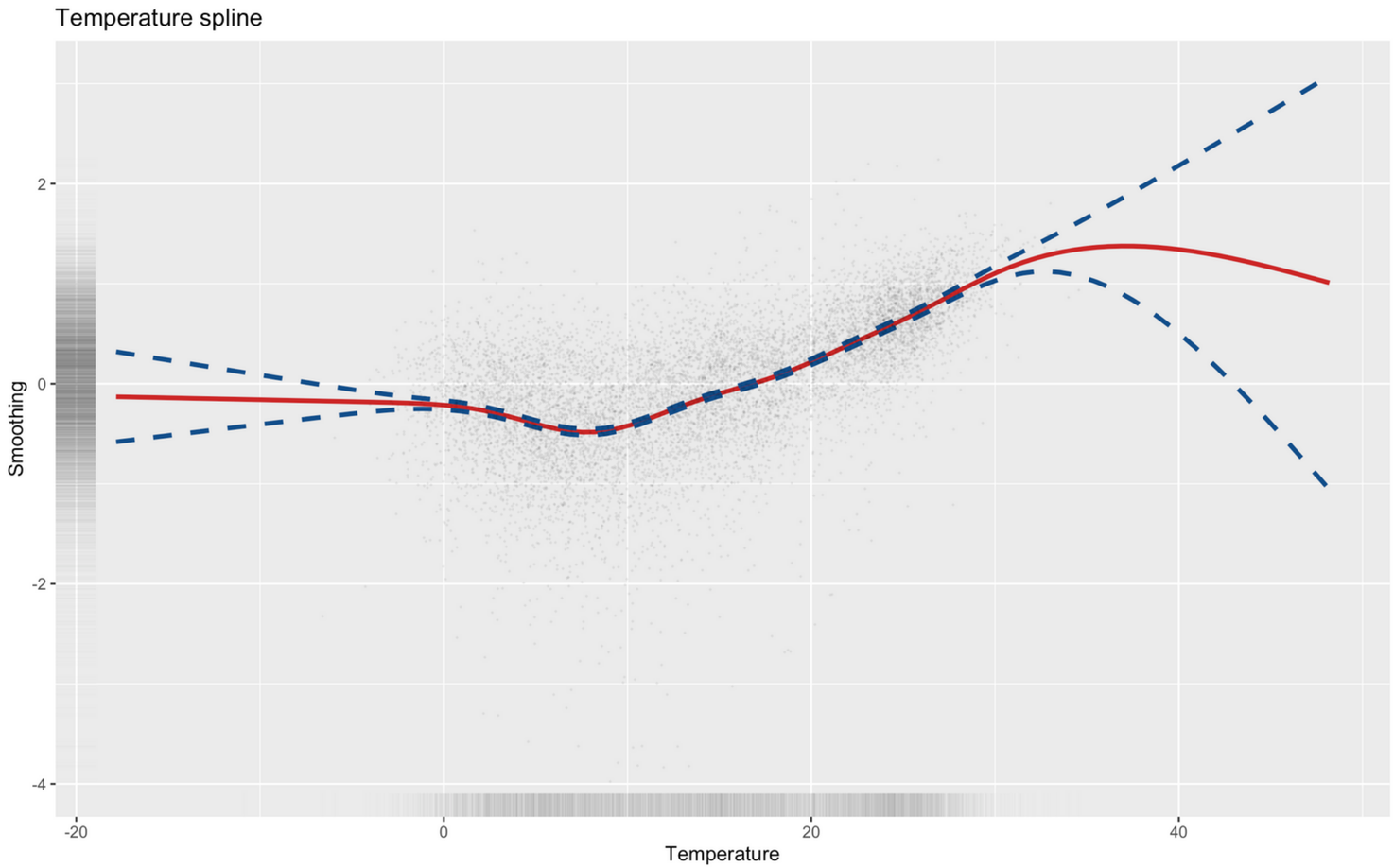
GAM model

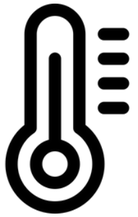


Rain contributes in modeling the **downward** spikes

MODELING PM10 LEVEL

GAM model

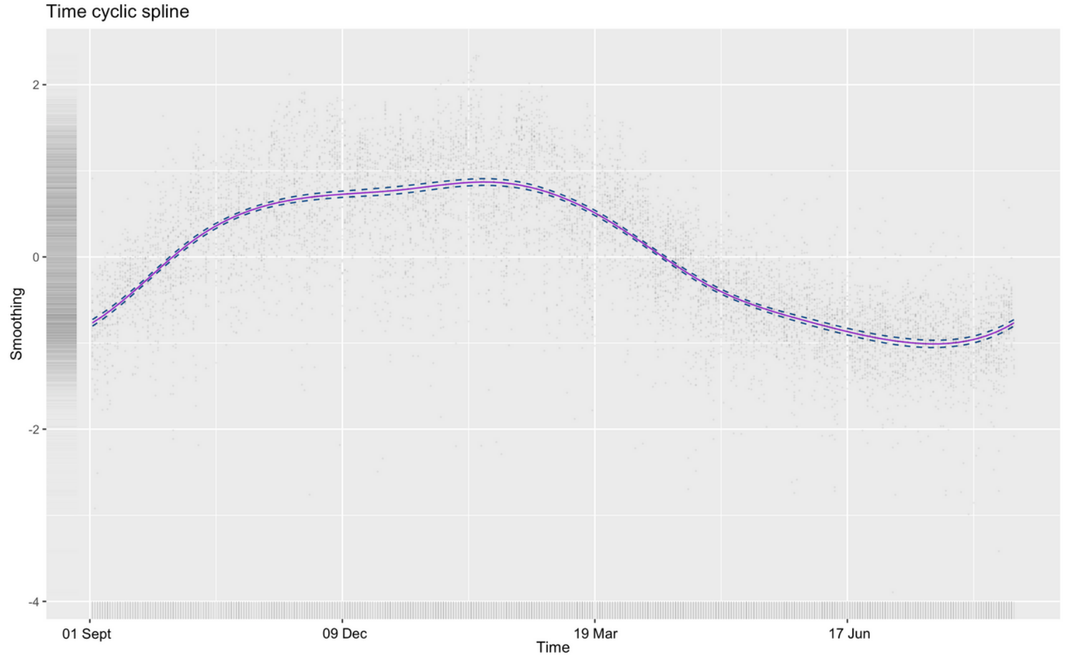


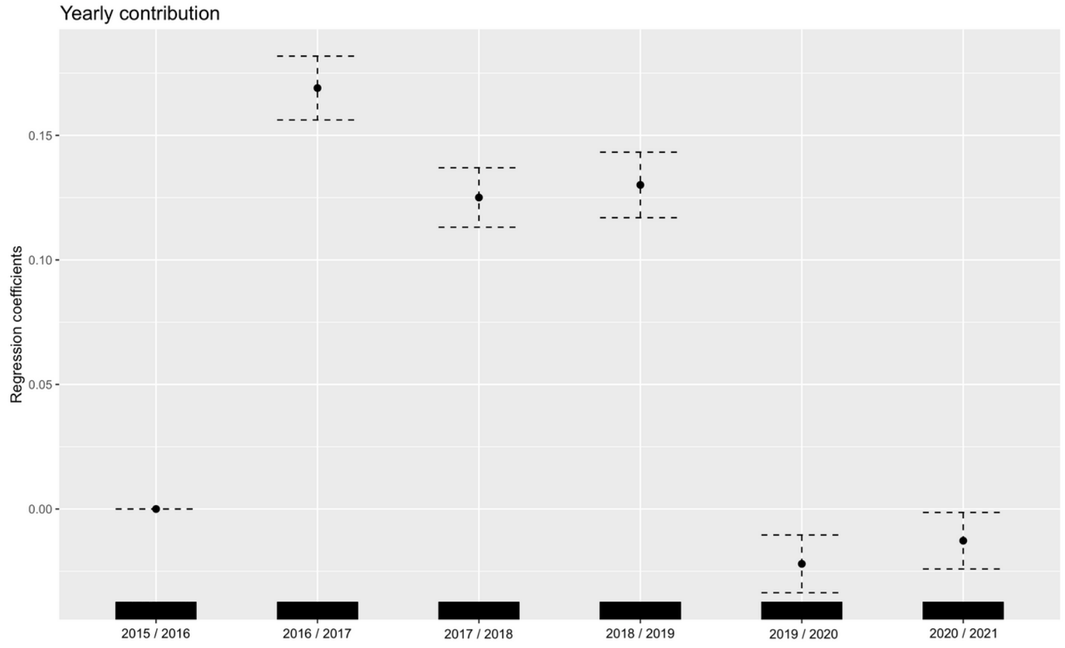


High temperatures lead to a **nonlinear increase** in pollution

MODELING PM10 LEVEL

GAM model

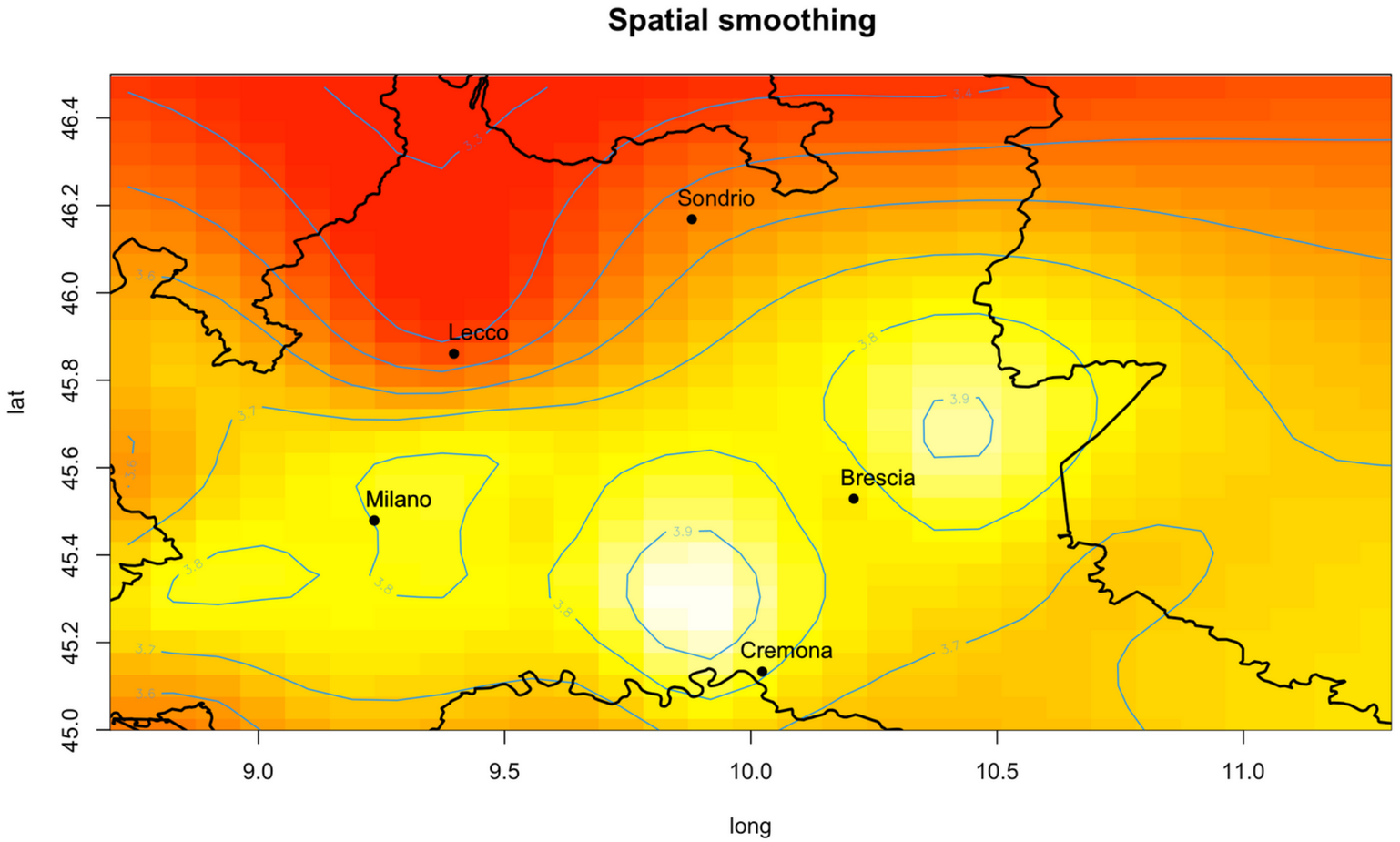


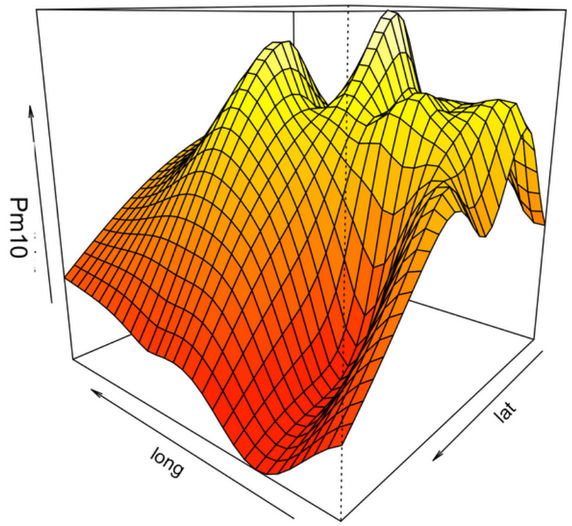


- The non-linear shape of the time series, captured by the cyclic spline, is coherent with the **seasonal trend**
- Our initial guess is confirmed by the **dummy** variable on the year

MODELING PM10 LEVEL

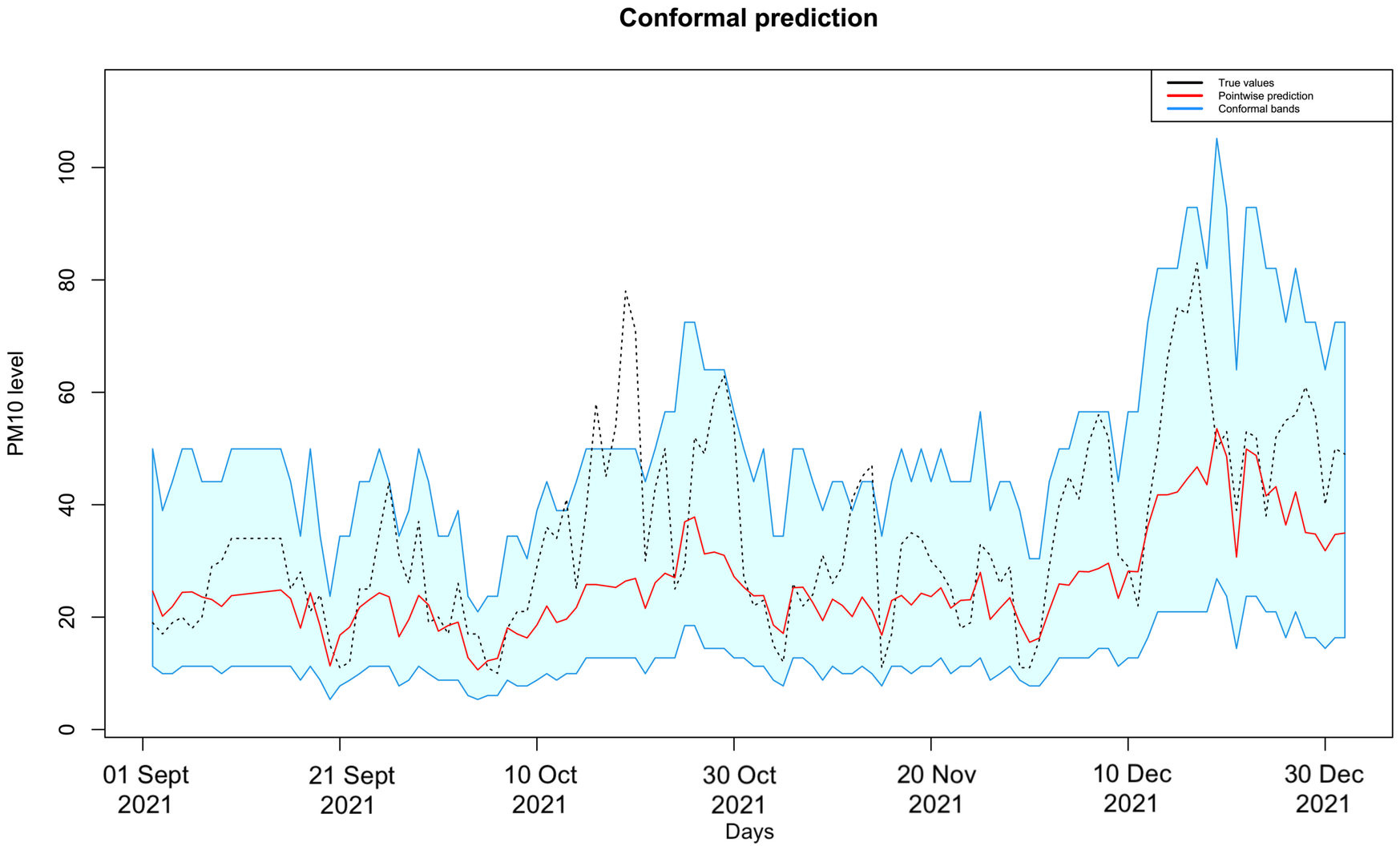
GAM model





CONFORMAL PREDICTION

Milano città studi station



Conclusions

Pollution **increases** after 2016 while the COVID years show an **opposite trend**; temporal information is not enough on its own to understand PM10 concentration

Weather forecast is crucial to estimate PM10 trends. Together with the spatial location and time coordinates we can produce a **reliable estimate** on the pollution level

Thanks for your attention!

- ARPA Lombardy
  https://www.dati.lombardia.it/stories/s/auv9-c2sj
- European Environment Agency
  https://www.eea.europa.eu/publications/2-9167-057-X/page021.html
  https://www.eea.europa.eu//publications/air-quality-in-europe-2022
- Will Media
  https://willmedia.it
- Podvin, Alexandre. "R-package for Interval-Wise Testing Procedure"
- Feng, Cindy. "Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada." Spatial statistics 49 (2022): 100526.
- GAM model
  https://noamross.github.io/gams-in-r-course/chapter1