POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

# Spatio-Temporal Models for Particulate Matter in the Po Valley

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

**Author:** Michela Frigeri

**Advisor:** Prof. Alessandra Guglielmi

**Co-advisor:** Prof. Giovanni Lonati

**Academic year:** 2021-2022

## 1. Introduction

In the Po valley the air quality problem became of great significance, especially in the last years (see [2]). Moreover, as can be seen in [4], this area is classified as one of the most polluted in Europe, due to both geographical and anthropic factors; see also [8]. The massive presence of air pollutant has a negative effect over the population life expectancy, increasing the natural mortality rate, as reported for instance in [9]. Hence, there is a need for deeper analysis of the phenomenon. In this paper the attention will be focused over a single atmospheric pollutant, namely the particulate matter (PM) having diameter less or equal to $10\mu m$, which is referred as $PM_{10}$. Indeed particulate matter is one of the most diffused air pollutants in the Po valley, because of many sources producing these substances such as vehicle exhaust, industries, residential heating etc. (see [5]). This project aims at defining appropriate models to spatially describe the $PM_{10}$ concentrations trend, together with more complex models which will provide a cluster estimate of the recording stations. All the models will be defined and implemented using the Bayesian approach. The $PM_{10}$ pollution problem is then approached as summarized in the following sections.

## 2. Air Pollution in the Po Valley

The problem of air pollution is of paramount importance to understand and analyze the quality of living in a certain area. By air pollution, we mean the release by human activities of gases and particulates into the atmosphere. The most common air pollutants are: carbon monoxide, sulfur dioxide, chlorofluorocarbons (CFCs), nitrogen oxides, photo–chemical ozone and smog, particulate matter, or fine dust, characterized by their diameter size ($PM_{10}$ and $PM_{2.5}$). In the following work, the statistical analysis will focus only on $PM_{10}$ air-pollution, since this pollutant is massively present in the Po valley, and it has not decreased even in the COVID19 lockdown period (see [2, 6, 7]). Population density and preponderance of industries are the main responsible for PM emissions in this area (see [5]), while its shape and climate hampers the exchange of air masses, favouring pollutants built-up. Air pollution has irreparably serious effects over the quality of human life. In fact, all air pollutants have a significant association with respiratory mortality (see [9]).

## 2.1. The $PM_{10}$ Dataset

The data considered for this work come from the air quality monitoring network of the Po Valley. This network is managed by Regional Agencies of Environmental Protection (ARPA). The concentrations of pollutants are collected by fixed monitoring stations, distributed across the whole Po valley area. Each station collects data about many pollutants, in particular: benzene (C6H6), nitrogen dioxide (NO2), particulate matter (PM10, PM2.5) and ammonia (NH3). The data we will use, i.e. particulate matter (PM) concentrations, are collected as daily averages in all regions. Data concerning $PM_{10}$ pollution over time are always intended with the proper measurement unit $\mu g/m^3$. The unit of measurement will be generally omitted in this summary, unless it is necessary. In our dataset, daily values of $PM_{10}$ concentration have been collected from 2014 to 2020 by a total of 201 stations located in four regions: Emilia Romagna, Lombardia, Piemonte and Veneto. Each recording station is uniquely identified through an ID and a specific station-name. Moreover the stations are characterized by additional information (area, type, zoning, altitude and geographical coordinates); see Section 2.2 for more details. Figure 1 shows the $PM_{10}$ time-series recorded in Veneto, as an example of the general trend of data. Looking at the time-series, we can see a peculiar U-shape repeating with annual frequency over the whole time horizon. This pattern might be explained by the specific nature of data, which are strongly influenced by seasonal phenomena such as domestic heating or massive vehicle use.

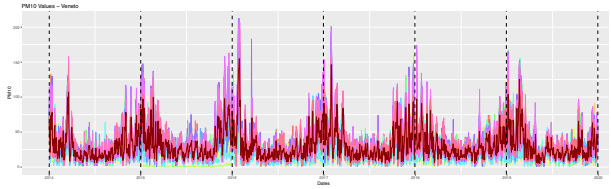


Figure 1: Time series of $PM_{10}$ values registered in Veneto from 2014 to 2020. The dashed vertical lines denote the 31st December of each year.

Data provided by the four ARPA organizations presented many issues that we have addressed before proceeding with the analysis. The more relevant problems concern multiple observations taken in the same day by the same station (probably due to wrong data management) and the presence of zero-valued registrations, which are unrealistic in the air pollution field. The presence of missing values was dealt with the Bayesian approach, in which they will be treated as sort of 'holes' in the $PM_{10}$ time series. Then the missing concentrations will be simulated according to the model we used. Other minor issues, such as typos or missing specifications, were easily corrected. Note that, in the rest of the work, only data recorded in 2018 have been considered, being the year that best guaranteed stationarity of the historical series. After this data "polishing", the logarithmic transformation was applied to all $PM_{10}$ concentration values. Indeed, it is common use to log-transform the PM concentrations in the atmospherical research field, since the data are strictly positive and the PM concentrations are often assumed to be log-Gaussian distributed (assumption that we make also here).

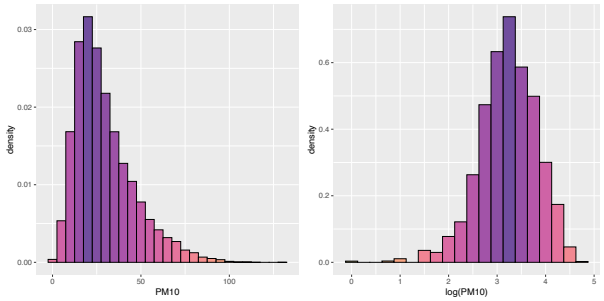



Figure 2: Histograms of PM10 concentrations in Lombardy before (left) and after (right) the logarithmic transformation.

Figure 2 reports the histograms of PM10 concentrations before and after the log-transformation, showing the data provided by ARPA Lombardia as an illustrative example.

## 2.2. Explorative data analysis

We report here explorative data analysis (EDA) for the $PM_{10}$ daily average log-concentrations recorded in Emilia-Romagna during 2018. Analogous results have been found for the other regions, leading to similar comments. A graphical analysis of the more relevant covariates is presented here, in order to enlighten peculiar patterns that may be useful for the definition of the model. The data will be treated as time-series, assuming the log-concentrations of $PM_{10}$

as a function of time. Each station will be represented by a single time series reporting its daily registrations over the 365 days under study. As displayed in Figure 3, the specific area in which each station is located is classified as *urban, suburban* or *rural*. Urban stations are located in a completely built area, while suburban stations are located in a zone having both built-up areas and not urbanized areas. Rural stations, instead, are located in a zone that can not be labeled neither as urban or suburban, presenting very little urbanization.

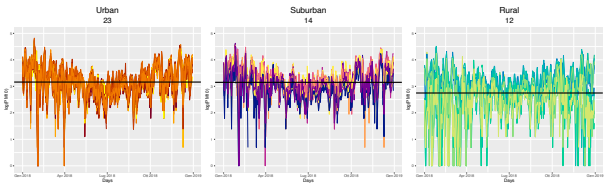



Figure 3: Time series of log-$PM_{10}$ in Emilia-Romagna recorded by each station, grouped according to area characterization.

Note that the U-shaped behaviour characterizing the PM pollution level is emphasized in the urban and suburban areas, while the rural stations seem to follow a different trend. Figure 4 shows the log-$PM_{10}$ time series in the three different types of stations: *traffic, industrial* or *background*.


Figure 4: Time series of log-$PM_{10}$ in Emilia-Romagna recorded by each station, grouped according to type characterization.

Traffic stations are located in specific sites, where the air pollution is mainly due to vehicle traffic emissions. Industrial stations are located sufficiently close to industrial areas, such that it can be assumed that recorded pollutants mainly derive from industrial sources. Background stations are positioned such that we can assume there is no specific additional source of pollution, but they should record the "background" level of $PM_{10}$. Finally, the altitude at which the the stations are located plays a fundamental role in the phenomenon explanation. Indeed, the concentrations of PM seem to decrease when the altitude is increasing, delineating an inverse relation between the recorded level of air pollution and the height at which the station is located. Figure 5 reports the data outline in four qualitative altitude-driven groups. This discretization of the altitude factor (which is a continuous variable) groups the stations in 4 classes defined by first, second and third empirical quartile, computed with respect to the empirical distribution of altitude values.

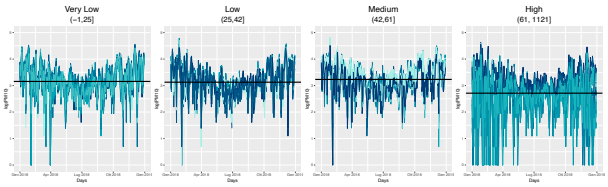



Figure 5: Time series of log-$PM_{10}$ in Emilia-Romagna recorded by each station, grouped into 4 classes of station altitude.

Dealing with time series entails very often the presence of some periodical pattern, and the presence of such phenomena must be investigated. At first, we compared the empirical distributions of $PM_{10}$ concentrations recorded during weekdays with those collected during weekends, but no peculiar pattern came out. Same result was obtained from the comparison of empirical distributions of $PM_{10}$ concentrations recorded in the seven days of the week, providing no relevant information to include into the model.

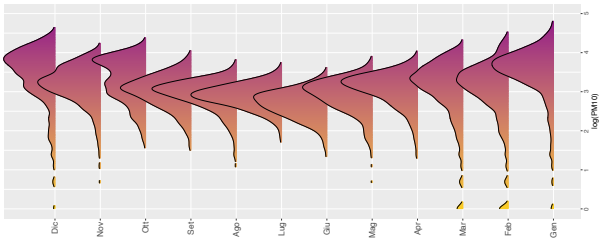



Figure 6: Kernel density estimate of $PM_{10}$ log-values recorded in each month of 2018.

However, the comparison of data distributions recorded in each month highlights an interesting behaviour, as shown in Figure 6. Indeed, the data variability seems to increase in the colder months, while it decreases during the summer season. This peculiar behaviour can be exploited

in the formulation of the model, as better explained in Section 3. Together with the above features, which can be generalized to all regions, also the zoning of each station is provided. However, the zoning characterization is a region-specific partitioning of the territory, thus it will not be included in the model. In fact, we aim at defining a model that could be generalized to any set of stations over any time horizon, independently of regional specifications, in order to adopt it in future work.

## 3. Hierarchical Modeling

Our goal is to build a Bayesian spatio-temporal model, able to explain the trend of $PM_{10}$ log-concentrations over 2018, and to include territorial information. We first provide a quickly review of basic theory about spatial models and the Bayesian approach. Then, this general framework is fitted to the $PM_{10}$ log-concentrations data collected in Emilia-Romagna during 2018. Finally, we propose three different models which were fitted to the data.

### 3.1. Modeling spatial data

Here we will report material from [1]. We reduce the large field of spatial data type to the more specific framework of Gaussian stationary spatial processes. In this section, $Y(\mathbf{s})$ represents a random vector evaluated at location $\mathbf{s} \in \mathbb{R}^r$, where $\mathbf{s}$ varies continuously over $D \subset \mathbb{R}^r$, its $r$-dimensional domain. Specifically, we will assume $r = 2$, since our sites will be characterized by only two spatial coordinates: latitude and longitude. Stationary spatial processes present a constant mean and their covariance matrix can be defined as $Cov[Y(\mathbf{s}) \ , \ Y(\mathbf{s}+\mathbf{h})] = C(\mathbf{h})$ for any $\mathbf{h} \in \mathbb{R}^r$. Hence, the covariance relationship between process' values at any two locations will be expressed by the function $C(\mathbf{h})$, depending only on the value of the separation vector $\mathbf{h} \in \mathbb{R}^r$. Specifically, we will deal with intrinsically stationary processes, i.e. processes for which $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})] = 0$ and the value of $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2$ depends only on $\mathbf{h}$. For these spatial processes we can define the *variogram function* as:

$$2\gamma(\mathbf{h}) = Var(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) \qquad (1)$$

We restrict our analysis to spatial processes presenting an *isotropic* variogram. In this scenario, the value assumed by the semivariogram function $\gamma(\mathbf{h})$ depends only on the length of the separation vector $||\mathbf{h}||$. Hence, the variogram becomes a real-valued function with an univariate input argument, and we write $\gamma(||\mathbf{h}||)$. Moreover, the relationship between the variogram and the covariance function can be summarized as follows:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \qquad (2)$$

where $C(\mathbf{0})$ denotes the variance of the stationary process. If a process is both intrinsically stationary and isotropic, then it is called *homogeneous*. Finally, assuming to have a homogeneous Gaussian process, simple parametric formulations can be proposed for the semivariogram. One of the most popular covariance kernels is the following exponentiated quadratic function:

$$C(d) = \begin{cases} \alpha^2 + \sigma^2 & \text{if } d = 0 \\ \alpha^2 \exp\left(-\frac{d}{2\rho^2}\right) & \text{if } d > 0 \end{cases} \qquad (3)$$

where $d = ||\mathbf{h}||$. The hyperparameters $\alpha, \rho, \sigma$ specify the behaviour of the covariance function, which can be seen as the convolution of two independent Gaussian processes with kernels $C_1(d) = \alpha^2 exp\left(\frac{-1}{2\rho^2}d\right)$ and $C_2(d) = \sigma^2 \mathbb{1}_{d=0}$. Let then $Y(\mathbf{s})$ be the homogeneous Gaussian process evaluated in a generic site $\mathbf{s}$, and assume:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \qquad (4)$$

where $\mu(\mathbf{s})$ is the process' average value. The residual of this model is divided in two pieces. The spatial residual term $w(\mathbf{s})$ is assumed to be a realization from a zero mean Gaussian process, capturing the residual spatial association of data. The non-spatial residuals $\epsilon(\mathbf{s})$ are instead uncorrelated pure error terms. We will better specify this last formulation in Section 3.3, including spatial residuals in the model definition.

### 3.2. The Bayesian approach

In this section we summarize material from [1, 10]. The Bayesian approach combines complex data models and external prior knowledge. In the regression context, classic statistical methods such as ordinary least squares (OLS) aim at estimating unknown but fixed parameters value. On the contrary, in the Bayesian framework, both the target variable and any unknown

parameters are modeled as random quantities. Following this approach, the joint distribution for the target variable $\mathbf{y} = [y_1, \dots, y_n]$ and the vector of unknown parameters $\boldsymbol{\theta} = [\theta_1, \dots, \theta_k]$ has to be specified. One way to specify the joint density is to assign the conditional law $f(\mathbf{y}|\boldsymbol{\theta})$ of $\mathbf{y}$ given $\boldsymbol{\theta}$. Then, the vector of parameters $\boldsymbol{\theta}$ is modeled as a random quantity from the prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where the vector $\boldsymbol{\lambda}$ contains the hyperparameters specifying the prior. If hyperparameters in $\boldsymbol{\lambda}$ are known, inference about the parameters in $\boldsymbol{\theta}$ is based on the following posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\mathbf{y}|\boldsymbol{\lambda})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\theta} \tag{5}$$

where $\Theta$ is the $k$-dimensional parametric space. The posterior distribution, i.e., the updated distribution of parameters after having observed the data, is computed exploiting the well known *Bayes' theorem* (5). Note that data contribution (given by the likelihood $f$) and external knowledge (expressed by the prior $\pi$) are both involved in computing the posterior distribution. Very often $\boldsymbol{\lambda}$ is not known, hence we will need a second stage distribution $h(\boldsymbol{\lambda})$, called *hyperprior*, in order to specify the model. Note that an implicit hierarchical structure is adopted in the distributional sense. Typically, when using this hierarchical approach, the primary interest is focused on the parameters $\boldsymbol{\theta}$ level. A very popular method to sample from the target posterior is the Markov Chain Monte Carlo (MCMC) method. This method relies on the construction of a Markov chain having the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ as its invariant distribution. Then, sampling path averages of this MC, we estimate the characteristics of the target posterior.

### 3.3. Spatio-temporal models for $PM_{10}$ in Emilia-Romagna

We propose here three models to be fitted to the data. All the models include a function of time $f(t)$ expressing the trend of the $PM_{10}$ log-concentrations, as specified in (6). We defined this formulation using Fourier basis, assuming annual and quarterly frequencies for the trigonometric basis functions, coherently with the observed seasonality of the phenomenon.

$$\begin{aligned} f(t) = {} & a_1 sin(\omega t) + b_1 cos(\omega t) \\ & + a_2 sin(4\omega t) + b_2 cos(4\omega t) + c \end{aligned} \tag{6}$$

where $\omega = 2\pi/365$. The $f(t)$ coefficients are assumed to be different in the cases of rural and non-rural recording stations, coherently with the trend shown in Figure 3. The corresponding functions will be indicated by $f_R(t)$ and $f_{NR}(t)$. In addition to this time-dependent component, we included three regressors in the models: two dummy variables indicating traffic or industrial locations and the continuous variable specifying the altitude. A first model (M1) is defined without including the spatial correlation. Hence, the spatial residual term $w(\mathbf{s})$ in (4) will be omitted, assuming that the nugget parameter $\sigma^2$ explains the whole data variability, i.e. $C(0) = \sigma^2$ and $C(d) = 0 \quad \forall d > 0$ instead of (3). We propose, then, a second model (M2) including the spatial correlation term as in (4). Here $w(\mathbf{s})$ is modeled as a Gaussian process having covariance matrix defined as in (3). Finally, the third and most complex model (M3) is a further development of M2, accounting also for the comments about Figure 6. In this last case the spatial correlation is included as before, together with a month-specific variance parameter $\sigma^2_{m(t)}$. The function $m(t)$ indicates the month corresponding to every day of the year $t = 1, \dots, 365$, taking values in $1, \dots, 12$. We define in (7)-(18) the formulation of M3, while formulations of M1 and M2 can be easily obtained by properly removing terms. In what follows the residual term $w(\mathbf{s}_i)$ will be indicated by $w_i$, for the sake of a more concise notation.

$$Y_i(t) \overset{ind}{\sim} \mathcal{N}(\mu_i(t)\ , \sigma^2_{m(t)}) \tag{7}$$

$$i = 1, \dots, 49\ , \quad t = 0, \dots, 364$$

$$\begin{aligned} \mu_i(t) = {} & f_R(t)\mathbb{1}_{\{area(i)=rural\}} \\ & + f_{NR}(t)\mathbb{1}_{\{area(i)\neq rural\}} \\ & + \mathbf{x}_i^T\boldsymbol{\beta} + w_i \end{aligned} \tag{8}$$

$$\begin{aligned} f_R(t) = {} & a_{R1} sin(\omega t) + b_{R1} cos(\omega t) \\ & + a_{R2} sin(4\omega t) + b_{R2} cos(4\omega t) \\ & + c_R \end{aligned} \tag{9}$$

$$f_{NR}(t) = a_{NR1}sin(\omega t) + b_{NR1}cos(\omega t) + a_{NR2}sin(4\omega t) + b_{NR2}cos(4\omega t) + c_{NR} \quad (10)$$

$$\sigma_1^2, \ldots, \sigma_{12}^2 \overset{iid}{\sim} InvGamma(3,2) \quad (11)$$

$$a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R \overset{iid}{\sim} \mathcal{N}(0,1) \quad (12)$$

$$a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR} \overset{iid}{\sim} \mathcal{N}(0,1) \quad (13)$$

$$\beta_0, \beta_1, \beta_2 \overset{iid}{\sim} \mathcal{N}(0,1) \quad (14)$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0},\ \Sigma) \quad (15)$$

$$\Sigma_{i,j} = \alpha^2 exp\left(-\frac{1}{2\rho^2}||s_i - s_j||^2\right) \quad (16)$$

$$\alpha \sim \mathcal{N}(0.3,\ 0.1) \quad (17)$$

$$\rho \sim Beta(3,10) \quad (18)$$

Posterior inference for model M3 is briefly summarized here below. Note that all the posterior estimates have been obtained using *Stan* ([11]) for the MCMC simulations.

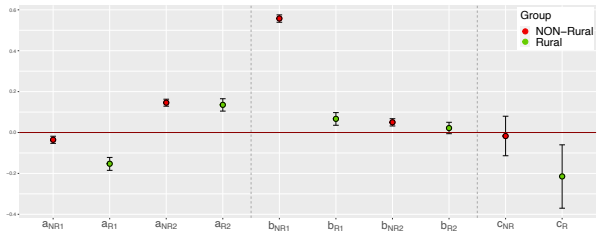



Figure 7: 95% marginal posterior Credibility Intervals (CI) of parameters $a_{NR1}, a_{R1}, a_{NR2}, a_{R2}, b_{NR1}, b_{R1}, b_{NR2}, b_{R2}, c_{NR}, c_R$ for M3.

Figures 7 and 8 summarize the posterior inference for $f_R(t)$ and $f_{NR}(t)$. The CIs of rural and non-rural coefficients might assume very different values (see $b_{NR1}$ vs $b_{R1}$), supporting the choice of distinguish these two scenarios.

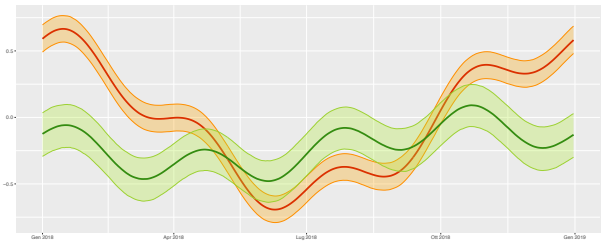


Figure 8: 95% posterior credibility bands for the time-dependent functions $f_R(t)$ (in green) and $f_{NR}(t)$ (in red) for Model 3.

Notice that also the posterior credibility bands of the two functions are disjoint almost everywhere, proving that $f_R(t)$ and $f_{NR}(t)$ can not be assumed to be equal (see Figure 8). Figure 9 reports posterior inference for $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$, the regression parameter corresponding to covariates. As expected from the EDA in Section 2.2, the altitude coefficient ranges over negative values (lower pollution at higher altitudes), while coefficients specifying industrial and traffic stations are associated to positive values (higher pollution in traffic/industrial locations).

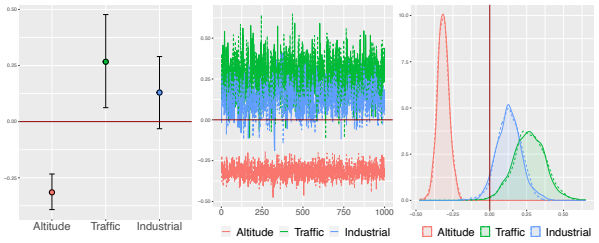



Figure 9: Posterior inference for the regression parameter $\boldsymbol{\beta}$: 95% marginal posterior credibility intervals (left), traceplots (center) and posterior marginal densities (right).

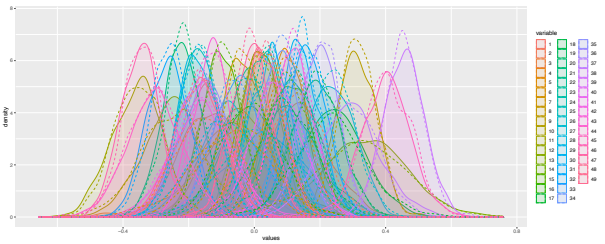


Figure 10: Marginal posterior densities of $w_1, \ldots, w_{49}$ (spatial residual terms) in Model 3.

Figure 10 shows the marginal posterior densities of the spatial residual term $\mathbf{w}$ for each of the 49 stations under study. This distributional pattern enlighten the presence of relevant spatial correlation among recording sites.

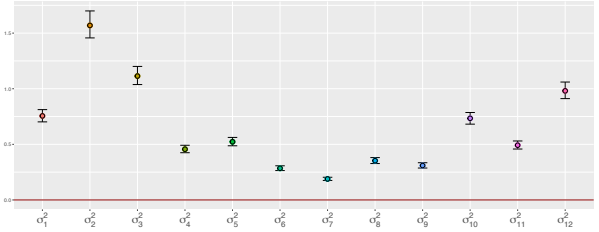


Figure 11: 95% posterior marginal credibility intervals of the 12 month-specific parameters $\sigma_1^2, \ldots, \sigma_{12}^2$ expressing the data variability.

Finally, Figure 11 reports the CIs for the twelve month-specific variance parameters. The posterior CIs appear to be really different from one another, thus there is evidence supporting the definition of month-specific variances.

### 3.4. Model Selection

The models will be evaluated from a predictive point of view using the Widely Applicable Information Criterion (WAIC) and the Leave One Out Cross Validation (LOO-CV), which are Bayesian model selection methods exploiting the whole posterior distribution. These methods consist in estimating pointwise out-of-sample prediction accuracy from the fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values. Moreover, the values assumed by WAIC and LOO are asymptotically equal for a sufficiently large number of data.

**Models Comparison**

| | WAIC | LOO |
|---|---|---|
| **M1** | 42129.5 | 42129.5 |
| **M2** | 41063.6 | 41063.8 |
| **M3** | 38284.5 | 38284.8 |

Table 1: WAIC and LOO computed for the three proposed models.

We computed the values reported in Table 1 using the `loo()` package in R (see [12]). The best model, according to these criteria, is the one providing the lowest value of WAIC and LOO. Table 1 shows that the third model (M3) is to be preferred, even if it is the most complex one (i.e. with more parameters).

## 4. Model-based clustering

The model-based clustering will rely on the best model we just provided (i.e. M3). Since we aim at grouping together the stations showing a similar trend, this section will deal with f(t) and its coefficients. A suitable formulation for this clustering procedure is the Dirichlet Process Mixture Model (DPMM) approach, which is described in general as follows:

$$p(y|\gamma, P) = \sum_{k=1}^{\infty} \pi_k p(y|\theta_k) \tag{19}$$

$$\pi_1 = V_1, \ \pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j) \quad \forall k \geq 2 \tag{20}$$

$$V_1, V_2, \ldots|\gamma \overset{iid}{\sim} Beta(1, \gamma) \tag{21}$$

This is an infinite mixture of randomly weighted parametric densities, with parameters randomly selected from a distribution $P$ centered on the base fixed distribution $P_0$. Model (19)-(21) is equivalent to the general DPMM involving an infinite series of terms. In real-life applications we use finite mixtures to achieve an approximation, considering a truncation of the above series.

$$Y_i|\theta_i \overset{ind}{\sim} p(y_i|\theta_i), \quad i = 1, \ldots, n \tag{22}$$

$$\theta_1, \ldots, \theta_n|P \overset{iid}{\sim} P \tag{23}$$

$$P \sim DP(\gamma, P_0) \tag{24}$$

see, for instance, [3]. In our case, the likelihood $p(y|\theta)$ is the Gaussian kernel, and $P$ is the Dirichlet Process centered in $P_0$ with dispersion parameter $\gamma$. If $\theta_1, \ldots, \theta_n$ are a sample from a DP as in (23), then $\theta_i = \theta_j$ with positive probability. In this case a random partition of the data label set $\rho = \{S_1, \ldots, S_G\}$ will be used for the actual clusterization of data. Each subset $S_g$ contains the data indexes allocated together, according to the relation $Y_i \sim Y_j \iff \theta_i = \theta_j$ from (22)-(24). Hence the data are grouped according to their parameters distribution. Note that in general $G$, the number of clusters, is assumed to be random. The final grouping of data will be given by the value of $\rho$ obtained by minimizing the Binder's loss function, as in [3]. Applying this approach to the parameters of individual mean $f_i(t)$, a first cluster estimate is obtained looking only at the annual cosine coefficient $b_i$. Then, a second multivariate cluster estimate will be obtained considering all the five coefficients at the same time. In Sections 4.1 and 4.2 we provide the mixture formulations for these two cases, together with the estimated posterior clusterization.

### 4.1. Univariate clustering

Building on M3 (7)-(18), only the following terms have been modified:

$$\mu_i(t) = f_i(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \tag{25}$$

$$f_i(t) = a_1 sin(\omega t) + b_i cos(\omega t) + a_2 sin(4\omega t) + b_2 cos(4\omega t) + c \tag{26}$$

$$b_i \overset{iid}{\sim} \sum_{g=1}^{G} \eta_g \mathcal{N}(m_g, s_g^2) \quad i = 1, \dots, 49 \tag{27}$$

$$\eta_1 = v_1 \ , \quad \eta_g = v_g \prod_{j=1}^{g-1} (1 - v_j) \tag{28}$$

$$\sum_{g=1}^{G} \eta_g = 1 \tag{29}$$

$$v_g \overset{iid}{\sim} Beta(1, 2) \quad g = 1, \dots, G \tag{30}$$

where, considering some a-priori frequentist analysis, a possible feasible definition for the hyperparameters prior is:

$$m_g \overset{iid}{\sim} \mathcal{N}(0, 150 s_g^2) \tag{31}$$

$$s_g^2 \overset{iid}{\sim} InvGamma(4.5, 0.015) \quad g = 1, \dots, G \tag{32}$$

Note that, instead of assuming station-specific parameters $b_i$ as iid from (19)-(21), we truncate the infinite sum to $G = 20$ terms. The associated cluster estimate is reported in Figure 12. Note that while two groups (blue and green) show negative values of the annual cosine coefficient, the other two (pink and purple) are related to positive values of $b_i$, following a different annual trend.

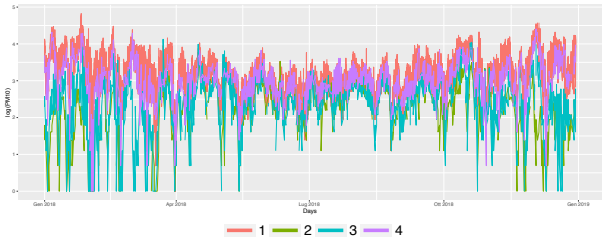



Figure 12: $PM_{10}$ log-concentrations time series of the stations, coloured according to cluster estimates in 4 groups.

### 4.2. Multivariate clustering

Starting from M3 (7)-(18), only the following terms have been modified:

$$\mu_i(t) = f_i(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \tag{33}$$

$$f_i(t) = a_{1i} sin(\omega t) + b_{1i} cos(\omega t) + a_{2i} sin(4\omega t) + b_{2i} cos(4\omega t) + c_i \tag{34}$$

$$[a_{1i}, b_{1i}, a_{2i}, b_{2i}, c_i] \overset{iid}{\sim} \sum_{g=1}^{G} \eta_g \mathcal{N}_5(\mathbf{m}_g, \Sigma_g) \tag{35}$$

$$i = 1, \dots, 49$$

$$\Sigma_g = diag(s_{g1}^2, \dots, s_{g5}^2) \tag{36}$$

where, the weights $\eta_g$ are defined as before in (28)-(30). Using an empirical Bayes approach, we have assumed as hyperparameter prior:

$$\mathbf{m}_g \overset{iid}{\sim} \mathcal{N}_5(\mathbf{0}, 150 \Sigma_g) \tag{37}$$

$$s_{g1}^2, \dots, s_{g5}^2 \overset{iid}{\sim} InvGamma(4.5, 0.015) \quad g = 1, \dots, G \tag{38}$$

The associated cluster estimate is reported in Figure 13.


Figure 13: $PM_{10}$ log-concentrations time series of the stations, coloured according to associated multivariate cluster estimate in 2 groups.

This second cluster estimate provided a neat division among the two identified clusters. As can be seen from Figure 14, this cluster estimate divides the stations located across the Apennines (in blue) from the other stations (in pink). This division is coherent with all our comments about the influence of altitude, urbanization and anthropic factors over the pollution level.

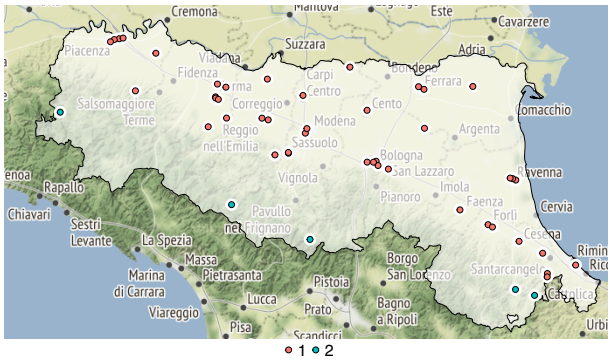



Figure 14: Geographical map of Emilia-Romagna displaying the 49 recording sites coloured according to BNP grouping.

## 5. Conclusions and future developments

Applying the Bayesian approach to the $PM_{10}$ log-concentrations data, we improved our knowledge about this pollution phenomenon from a spatio-temporal point of view. Moreover, the model-based cluster estimates provided additional information about the sites, grouping stations that would not be put together by looking only at territorial specifications. Moreover, our work allows for many further extensions. First of all, the models we have proposed could be extended to the whole dataset, including all the stations in the Po valley and data from multiple years. Then, Bayesian spatial prediction methods could be implemented to estimate the spatial residual term in unknown locations (kriging techniques). Another interesting and necessary generalization of the model concerns the inclusion of meteorological factors as regressors. Indeed, the weather conditions heavily affect the air pollution level, being responsible for both air masses exchange and pollutants built-up. However, meteorological data and pollution data are collected by different ARPA stations, having different locations and a different data timing. Hence pollution data and weather information were not comparable. A more complex procedure of data processing should be implemented to properly combine the two datasets in a unique extended database, requiring a huge amount of time given the data dimensionality. Thus we were not able to include the meteorological data in our analysis, but it could be interesting to develop this extension in future work.

## 6. Acknowledgements

I would like to thank Matteo Gianella, PhD researcher at Politecnico di Milano, for his help and suggestions during *Stan* code implementation.

## References

[1] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2015.

[2] M. Cameletti. The Effect of Corona Virus Lockdown on Air Pollution: Evidence from the City of Brescia in Lombardia Region (Italy). *Atmospheric Environment*, 239:117794, 07 2020.

[3] D. B. Dahl. *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*, page 201–218. Cambridge University Press, 2006.

[4] EEA. Air quality in Europe - 2019 Report: Technical Report. Technical report, European Environmental Agency (EEA), 2019. `https://www.eea.europa.eu/publications/air-quality-in-europe-2019`.

[5] B. Larsen, S. Gilardoni, K. Stenström, J. Niedzialek, J. Jimenez, and C. Belis. Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmospheric Environment*, 50:203–213, 04 2012.

[6] G. Lonati and F. Riva. Effetti degli interventi di contrasto alla diffusione del COVID19 sulla qualitÀ dell'aria in Pianura Padana. *Ingegneria dell'Ambiente*, 8(1/2021):24–39, 2021.

[7] G. Lonati and F. Riva. Regional Scale Impact of the COVID-19 Lockdown on Air Quality: Gaseous Pollutants in the Po Valley, Northern Italy. *Atmosphere*, 12(2), 2021.

[8] M. Masiol, S. Squizzato, G. Formenton, R. Harrison, and C. Agostinelli. Air quality across a European hotspot: Spatial gradients, seasonality, diurnal cycles and trends

in the Veneto region, NE Italy. *Science of The Total Environment*, 576:210–224, 01 2017.

[9] A. Pozzer, S. Bacer, S. D. Z. Sappadina, F. Predicatori, and A. Caleffi. Long-term concentrations of fine particulate matter and impact on human health in Verona, Italy. *Atmospheric Pollution Research*, 10(3):731–738, 2019.

[10] G. L. Rosner, P. W. Laud, and W. O. Johnson. *Bayesian Thinking in Biostatistics*. CRC Press, 2021.

[11] Stan Development Team. Stan modeling language users guide and reference manual, version 2.30. `https://mc-stan.org`, 2022. 2022-08-10.

[12] A. Vehtari, J. Gabry, M. Magnusson, Y. Yao, P.-C. Bürkner, T. Paananen, and A. Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2022. R package version 2.5.1.

POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# Spatio-Temporal Models for Particulate Matter in the Po Valley

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Michela Frigeri**

Student ID: 953285
Advisor: Prof. Alessandra Guglielmi
Co-advisors: Giovanni Lonati
Academic Year: 2021-2022

# Abstract

In this thesis the major theme of air pollution is discussed. Specifically, this work focus on the time series of particulate matter (PM) concentrations registered in the Po valley. Once provided a general overview of the phenomenon in the whole area of interest, we deepen the analysis for data collected in Emilia-Romagna during 2018. This choice was driven by the huge dimensionality of the original database received from the ARPA associations of multiple regions. At first, we propose Bayesian spatio-temporal hierarchical models in order to explain the $PM_{10}$ concentrations trend. Then, well-known posterior predictive goodness-of-fit criteria will be involved to select the "best" model among the proposed ones. Subsequently, we generalized this "best" model to provide model-based clustering for the recording stations in Emilia-Romagna. In general, model-based clustering gives hints on the clustering structure of the individuals that goes beyond the characteristics of the individuals themselves. Applied to our data, this Bayesian non-parametric clustering procedure has provided interesting results, separating the stations located on the Apennines mountains from the other recording sites. Furthermore, the models we have proposed can provide spatio-temporal prediction and allow for the inclusion of data in the whole Po valley over multiple years. We also plan to include new regressors, such as meteorological factors, into the models.

**Keywords:** air pollution, spatio-temporal hierarchical models, Bayesian inference, Gaussian processes, particular matter

# Abstract in lingua italiana

In questa tesi viene affrontato il tema di rilievo dell'inquinamento atmosferico. In particolare, questo lavoro si concentrerà sulle time-series delle concentrazioni di particulate matter (PM) registrate nella Pianura Padana. Una volta fornita una panoramica generale del fenomeno in tutta l'area di interesse, sono state implementate tecniche più complesse considerando solo i dati raccolti in Emilia-Romagna nel corso del 2018. Questa scelta è stata guidata dalle grandi dimensioni del database originale ricevuto dalle associazioni ARPA di più regioni. Inizialmente verranno proposti dei modelli gerarchici spazio-temporali seguendo l'approccio bayesiano, per spiegare l'andamento delle concentrazioni del $PM_{10}$ nel tempo. Verranno poi utilizzati dei noti criteri di predictive goodness-of-fit per selezionare il modello "migliore" tra quelli proposti. Il modello scelto verrà poi usato come punto di partenza per modelli mistura per il clustering, che verranno applicati ai dati per ottenere una partizione delle stazioni in Emilia-Romagna. I gruppi ottenuti grazie a questi modelli, detti modelli bayesiani non parametrici per il clustering, hanno fornito interessanti scoperte, separando le stazioni situate sugli Appennini dagli altri siti di registrazione. In generale, la clusterizzazione basata su modelli fornisce informazioni aggiuntive sui dati, essendo i gruppi trovati diversi da quelli definiti dalle caratteristiche delle stazioni. Inoltre, i modelli proposti forniscono un buon punto di partenza adatto per lo sviluppo di modelli ancora più complessi, che permetterebbero di ottenere nuovi risultati come la previsione spazio-temporale, l'estensione dei modelli a più regioni in più anni e l'inclusione di nuovi regressori (per esempio i fattori metereologici).

**Parole chiave:** modelli gerarchici spazio-temporali, inferenza bayesiana, processi gaussiani, particolato, inquinamento atmosferico

# Contents

# Introduction

In this master thesis we will discuss the important topic of air pollution. In the Po valley the air quality problem has became of great significance, especially in the last years, as explained in Cameletti (2020). Moreover, as can be seen from EEA (2019), this area is classified as one of the most polluted in Europe, due to both geographical and anthropic factors. The massive presence of air pollutants has a negative effect over the population life expectancy, increasing the natural mortality rate, as reported in Wan Mahiyuddin et al. (2012); Pozzer et al. (2019). From here, it is clear the need for deeper analysis and appropriate modeling of the phenomenon. In this paper the attention will be focused over a single atmospheric pollutant, namely the particulate matter(PM) having diameter less or equal to $10\mu m$, which is referred as $PM_{10}$. Indeed, the particulate matter is one of the most diffused air pollutants in the Po valley, because of many sources such as vehicle exhaust, industries, residential heating etc. (see Larsen et al. (2012)). This project aims at defining appropriate models to describe the $PM_{10}$ concentrations trend, together with more complex models which will provide a cluster estimate of the recording stations. The models will be defined and implemented using the Bayesian approach. The $PM_{10}$ pollution problem will then be approached as follows.

The dataset concerning $PM_{10}$ air pollution in the Po valley will be presented in detail in Section 1.1. The data have been collected from Lombardia, Emilia-Romagna, Veneto and Piemonte ARPA associations, thanks to the help of professor G. Lonati. Each of these datasets provided the daily average concentration of $PM_{10}$, recorded by multiple sensing stations over a time horizon going from 2014 to 2020, together with the territorial characterization of these stations. Before proceeding with the statistical modeling of the problem, an accurate data polishing was performed over the four datasets. Once we cleaned the data, a logarithmic transformation was applied to the $PM_{10}$ concentrations, trying to achieve log-normal distributed values. Then, preliminary explorative data analysis has been performed over the four regional dataset, and the resulting findings have been reported in Section 1.2. The information emerged from this first analysis will then be used to define which factors affect the $PM_{10}$ log-concentrations and how to include them in a regression model. Due to the huge dimensionality of the whole dataset, the case

study was reduced to the recordings taken in Emilia-Romagna during 2018. In Section 2.4 can be found some comments about the relevant factors that we decided to include in the $PM_{10}$ regression model.

After this first graphical exploration of the data, in Chapter 2 we will report the theoretical bases for the construction of the model. In this thesis the data will be modeled using Bayesian hierarchical models, which will be better defined in Section 2.2. Moreover, dealing with spatio-temporal data, we provided also some basic theory about spatial processes in Section 2.1. Indeed, since each station is associated with its geographical coordinates, specific techniques will be used to properly define the spatial correlation between recording sites. Then, we will propose three different models in Sections 2.4.1, 2.4.2 and 2.4.3. The "best" model will be defined via predictive goodness-of-fit (GOF) criteria as explained in Section 2.4.4. Finally, two possible formulations for the model-based clustering of stations will be provided in Sections 2.5.1 and 2.5.2, just after the necessary theoretical results reported in Section 2.5. We will then report the posterior inference for each model in Chapter 3.

Looking at the posterior inference reported in Chapter 3, some interesting findings can be pointed out. First of all, according to the GOF criteria, the most complex model (i.e. the one with more parameters) appears to be the best one to explain $PM_{10}$ data. This result validates the choice of including in the model both spatial residual terms and month-specific variability, which actually contributed at explaining the phenomenon. Secondly, when looking at the model-based clustering posterior inference in Sections 3.5 and 3.6, the posterior cluster estimates provide an interesting partition of the recording stations. Indeed, the estimated clusters could not be found by simply looking at the categorical variable characterizing each station, and our resulting partition provides additional knowledge about the pollutant behaviour. Notice that, according to our model-based clustering, stations located along the Apennines detach from other stations, showing a very different behaviour.

The model provided in Section 2.4.3 allows for many extensions beside clustering. Indeed, as better specified in Chapter 4, the model could be expanded to the whole Po valley dataset and to multiple years registrations. Moreover, it would be interesting to include as regressors the meteorological factors affecting the $PM_{10}$ concentrations. However, this was not possible since meteorological data and pollution data are recorded by ARPA using different stations in different locations and with a different timing. Consequently, obtaining and combining the two database would require an unreasonable amount of time, and hence the meteorological data were not included in this work. A further possible extension for this type of models is the Bayesian spatial prediction (kriging), which can be

performed exploiting the marginal posterior distribution of the Gaussian process defining spatial residual terms. During the thesis work, some preliminary analysis have been conducted on this topic, but the resulting models needed further refinement and thus they have not been included in this paper. Finally, also temporal prediction could be implemented, in order to estimate the pollution level in a "new" day.

# 1 | Air Pollution in the Po Valley

The problem of air pollution is of paramount importance, especially nowadays, to understand and analyze the quality of living in a certain area. By air pollution, we mean the release by human activities of gases and particulates into the atmosphere. The most common air pollutants are: carbon monoxide, sulfur dioxide, chlorofluorocarbons (CFCs), nitrogen oxides, photo–chemical ozone and smog, particulate matter, or fine dust, characterized by their diameter size ($PM_{10}$ and $PM_{2.5}$). In the following work, the analysis will focus only on the phenomenon of $PM_{10}$ air-pollution. This choice was due to the massive presence of this type of pollutants in the area of interest, which was not improved even in the COVID19 lockdown period (see Lonati & Riva (2021b,a); Cameletti (2020)). Indeed, the Po Valley is considered an hotspot for PM pollution in Europe (see Masiol et al. (2017); EEA (2019)). When investigating this phenomenon, several factors must be considered, for instance, the high population density, the massive presence of urban and industrial areas, the area geographic shape and meteorological factors. The population density and the preponderance of industries are the main responsible for PM emissions in this area, while its shape and climate hampers the exchange of polluted air masses, favouring pollutants built-up. Indeed, as reported in Larsen et al. (2012)), the origin of *Particular Matter* (PM) air pollution is to be found mainly in the anthropic scenario. As debated in the scientific reports by Larsen et al. (2012); Lenschow et al. (2001), among the first sources of air pollution there are factors such as long range transport, vehicle exhaust emissions, tyre abrasion and resuspended soil particles. All this pollution factors are strictly correlated to the human footprint on the environment, which has been irreparably altered by the construction of massive transport infrastructures and highly urbanized cores, leading unavoidably to busy areas characterized by intensive traffic and industrial waste. Unfortunately the presence of all this air pollutants has irreparably serious effects over the quality of human life. Indeed, there are strong basis to state that all air pollutants have a significant association with respiratory mortality (see Pozzer et al. (2019)). The most negative contribution appear to be given by $O_3$, but also $PM_{10}$ has been proved to be responsible in shortening the life expectancy, as reported in the study Wan Mahiyuddin et al. (2012). All the above mentioned risk factors contribute to make

the air pollution field an hot topic, from here the urge to define adequate models explaining the pollutant behaviour. As mentioned in the introduction, it could be very useful to identify some specific factors that contribute to this phenomenon, allowing the experts to bring together areas having the same "pollution pattern". Given this environmental framework, the need for a model explaining air pollution phenomenon arises naturally, in order to correctly describe the pollutants behavior over a specific time horizon and in a certain geographical area.

## 1.1. The $PM_{10}$ Dataset

The data considered for the analysis in this thesis, come from the air quality monitoring network of the Po Valley. This network is managed by Regional Agencies of Environmental Protection (ARPA) in the four regions of interest: Emilia Romagna, Lombardia, Piemonte and Veneto. The pollutants concentrations are collected by fixed monitoring stations, distributed across the whole Po valley area. Each station collects data regarding many pollutants, in particular: benzene (C6H6), nitrogen dioxide (NO2), particulate matter (PM10, PM2.5) and ammonia (NH3). Since each regional agency is autonomous, it may occur that the same data is registered in different time scales. For example, NO2 concentrations are usually available as hourly averages, but in Piemonte NO2 data are registered as daily average. Luckily, particulate matter (PM) concentrations are collected as daily averages in all regions, avoiding the downscaling of data. Exploratory data analysis of the whole dataset is provided in Lonati & Riva (2021b), where the authors investigate the effects of COVID–19 lockdown policies on the air pollution level in the target area. As mentioned above, the data concerning $PM_{10}$ pollution level over time are expressed through a single daily average value, i.e. the pollutant atmospheric concentration, and are intended with the proper measurement unit $\mu g/m^3$. The unit of measurement will be generally omitted, unless it is necessary.

The data have been collected by the stations accordingly to specific regional directives, but by law all these data are comparable to each other since they all must follow the European regulation Karagulian et al. (2019), which states the characteristics and benchmarks that air quality sensors must comply. The $PM_{10}$ concentration values are traditionally collected by applying the *gravimetric method*, which is well explained in the document ARPA Valle d'Aosta (2022) provided by ARPA Valle d'Aosta association. However, in the latest recordings, an automatic method has been preferred by almost all the registering stations. Indeed the two methods are deemed to provide equivalent value registrations, but in practice the so called *Beta attenuation method* is preferred to the gravimetric one,

since it is able to provide real-time measurement data. Following this approach, the procedure below is applied: the pollutant particles are suitably selected from a sample head that determines the dimensional cut of the sample, allowing the measurement of PM10 or PM2.5. After passing the sample head, the particles are deposited on a filter. The measurement of the attenuation of ß particles produced by a radioactive source by the sample (generally 14C) makes it possible to calculate, by difference from the attenuation of the white filter, the atmospheric particulate concentration of the selected dimensional class. This procedure is well explained in the portal of ARPA Lombardia ARPA Lombardia (2022), where all the adopted measurements techniques are properly described.

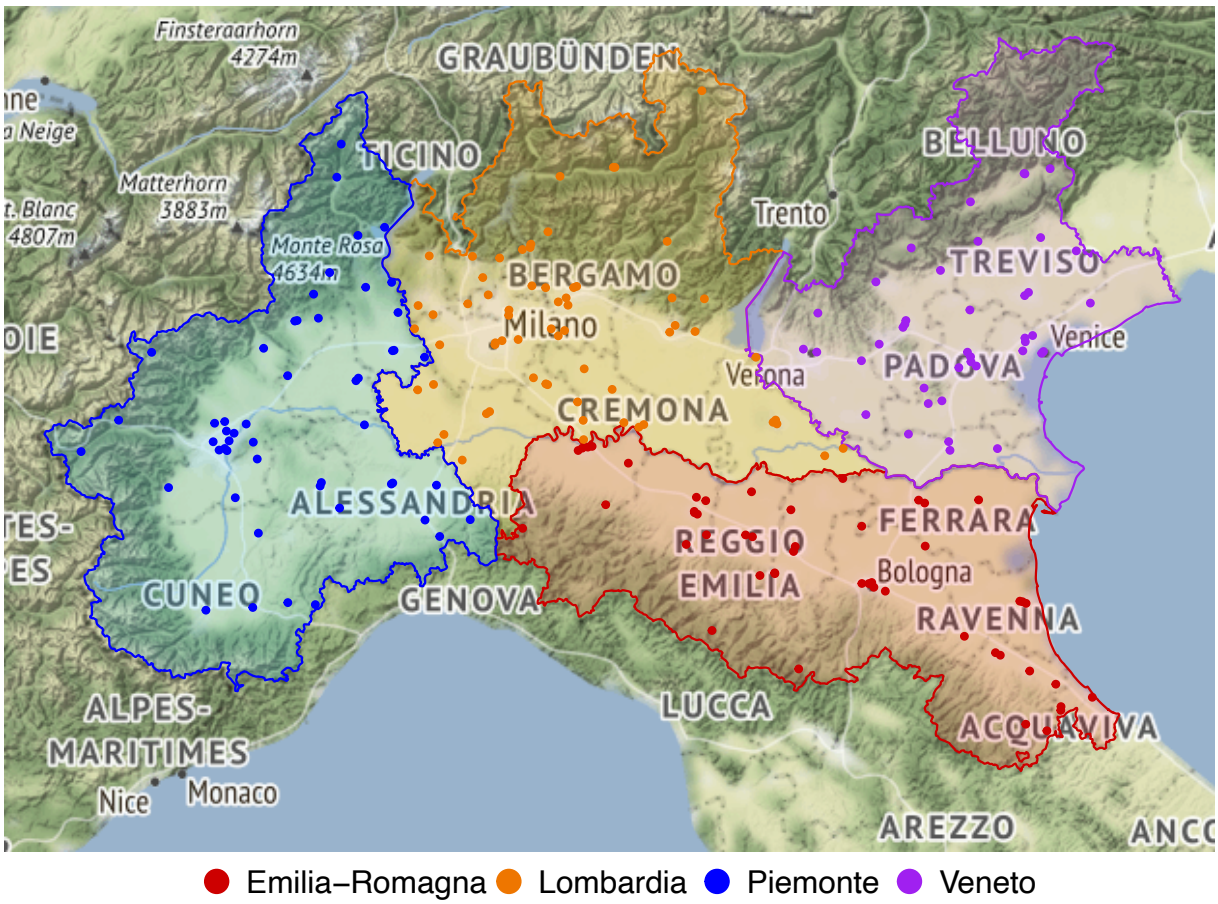



Figure 1.1: Location of the 201 recording stations divided by colour across the four regions of interest : Emilia-Romagna, Lombardia, Piemonte, Veneto.

### 1.1.1. Data Structure

The daily values of $PM_{10}$ concentration have been collected from 2014 to 2020 by a total of 201 stations located in northern Italy. The stations are partitioned across the four regions as follows:

- 49 stations in Emilia-Romagna
- 64 stations in Lombardia
- 51 stations in Piemonte
- 37 stations in Veneto

See Figure 1.1 for the map of recording stations.

Each recording station is uniquely identified through an ID and a specific station-name. Moreover extra information on the location of every station are recorded:

- **Area**: The station can be located in a *Rural, Suburban* or *Urban* area, according to its neighbourhood characteristics. For instance, the monitoring station "Milano - Pascal Città Studi" is labeled as urban, while "Casirate d'Adda" is located in a rural area, being a fraction of Bergamo which is little urbanized.
- **Type**: The station can be of *Traffic, Industrial* or *Background* type. This characterization describes the specific anthropological footprint in the recording site, that must account for traffic pollution or industrial smoke when in presence of higher $PM_{10}$ concentrations.
- **Zoning**: Each Region has its unique zoning definition, grouping the stations according to some region-specific geographical cores. This category divides stations according to a merely geopolitical classification, such as urban agglomerates around big cities or specific land properties like mountain range. For instance, the zoning of Piemonte is given by: Torino agglomerate, plain, hills and mountains.
- **Altitude**: The Altitude at which every station is located, which is crucial when dealing with air pollution phenomena.
- **Geographical Coordinates**: the exact location of each station is provided by its specific latitude and longitude or by UTM coordinates, depending on the region that registered the data.

In Figure 1.2 is reported the temporal series of $PM_{10}$ concentrations going from 2014 to 2020 recorded in Veneto. Also the other three regions show a similar behaviour, revealing an interesting annual trend of $PM_{10}$ data. Indeed the time-series shows a peculiar U-shaped behaviour repeating along the whole time horizon. As already mentioned, this seasonal pattern could be due to the specific nature of data under study, which are expected to be strongly influenced by seasonal phenomena such as domestic heating or massive vehicle use.

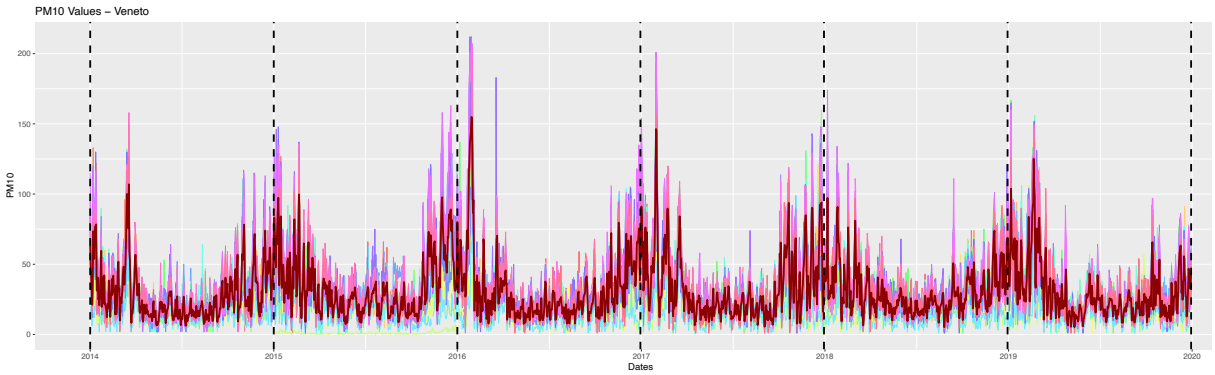



Figure 1.2: Time series of $PM_{10}$ values registered in Veneto from 2014 to 2020, illustrative of the data annual repeating pattern.

### 1.1.2. Issues about the Dataset

Analyzing the data supplied by the four ARPA organizations some common problems have been detected. First of all some observations were missing, probably due to some malfunctioning or sensor breakdown. Dealing with a real dataset this kind of problem is pretty common and can be solved in different ways. For instance the *missing values* could be approximated using a suitable interpolation of available data, or set equal to the mean of all data or to a more restricted time-horizon mean value (such as monthly or weekly mean). In the specific case under study, missing observations were not substituted with any numerical value. In fact, for each region, missing data represented less than 10% of the total observations. Instead of trying to guess a realistic value covering absent information, these problematic observation will be treated as sort of 'holes' in the $PM_{10}$ time series and then the missing concentrations will be predicted by the chosen model, coherently with the model structure and the other available observed data, thanks to the Bayesian approach. This is one of the advantages of the Bayesian approach, where missing data are simulated from the proper full conditional in the Bayesian model.

A further issue was the presence of some unrealistic observations that were registered by

the ARPA of Lombardia and Piemonte. In both of these dataset indeed *multiple registrations* from the same station in the same day were recorded. Dealing with Lombardy, this problem was simply an issue of doubled observations, that is multiple identical registration of the same value. It was sufficient to remove one of the two redundant observations to solve the problem. The problematic case of Piemonte was more difficult to solve, indeed in this scenario the multiple registrations reported different values of $PM_{10}$ concentration in the same day from the same station. This was possibly due to the use of different detection techniques, since the field reporting the measurement method was the only difference between these double observations. The three possible options for $PM_{10}$ sensing over all the available data are identified as follows: "Basso Volume - $\mu g/m^3$", "Beta(media giornaliera) - $\mu g/m^3$", "Beta - $\mu g/m^3$". These three methods should be equivalent and hence the concentration of $PM_{10}$ detected in the same place at the same time should be nearly the same using two different sensing methods. Unfortunately, for a not negligible portion of data, this was not the case and two different values were registered by the same station following two different approaches. After having investigated the phenomenon with professor G. Lonati, the problem was solved choosing to keep the values registered through the Beta detection method, that denotes the automatic sensing of $PM_{10}$, which at the moment it is considered the most popular and reliable method for air-pollutant detection.

Another common and obstructive issue was the presence of some very low or even *zero-valued registrations*. In general, very low pollution levels are expected after heavy rain storms or downpours. However, it is rather unrealistic to register no pollutant at all. It is well-known that meteorological factors heavily affects the PM concentration in the air (see for example Dung et al. (2019); Onuorah et al. (2019); Tian et al. (2014)), however the presence of a null mean value for air pollution over a whole day still appear nearly impossible. This is an important problem since this data values are unrealistic and wrong, but also because, when these values are equal to 0, the log transformation, so common for this type of data, is not possible. The zero valued observations could not be just considered as missing values or 'NA' due to some malfunctioning, as suggested in the study Taylor et al. (2018). Indeed, when some failure occurred in the sensing mechanism, the corresponding daily registration was simply missing or actually labeled with a 'NA' value. On the contrary these anomalous zero-valued observations were actually been recognised somehow by the detecting station. To understand the nature of the problem and propose a solution, some clarifications about the data sensing have been asked to one of the institutions dealing with this kind of data. A very interesting discovery was to learn that, obviously, each specific sensor (and hence each specific recording station) has

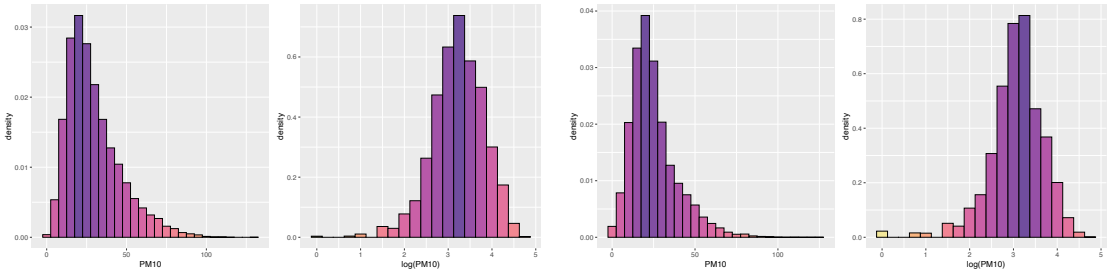
a proper detection threshold, i.e. all the tiny values lying below this threshold can not be 'seen' by the sensor. Consequently, it has been assumed that the zero-valued registrations correspond to a very low pollution level, that is probably not zero but still below the detection threshold and hence is not registered by the sensing station. Moreover this assumption is supported by the common practice of replacing this problematic abnormally low values with an automatic value equal to half of the detection threshold or to the detection threshold itself. Hence, being the collected values all positive integers, if the detection threshold is for instance equal to 1, its half will be approximated with 0 and thus one of the former problematic values is registered. Unfortunately to make sure that this assumption is correct, every single recording station of every single region should be contacted to ask for the exact detection threshold of its sensors, which is fairly unfeasible. Summing up, according to the suggestion by prof. Lonati and accounting also for the second lowest values registered in the various dataset (which probably represented the detection threshold), these zero-valued observations have been set equal to 1 (which is indeed the second smallest observed value). This data adjustment will also prevent issue arise from the logarithmic transformation of $PM_{10}$ concentrations, which will be explained in details in the next section. It could be interesting to notice that the only dataset being untouched by this problem is the one provided from Piemonte ARPA, for which the lowest concentration recorded was $1\mu g/m^3$. However, as can be seen from Figure 1.4d, a sort of censoring process seems to have been applied to the data, truncating the registrations at unit value and excluding all the concentrations lying below that threshold. This is important detail to notice, since it will require a specific processing when included in the model, surely different from the one applied to other regions data.

*Minor issues*, such as typos or missing specifications (for instance the altitude of some locations), were easily solved through a cross-check with other information sources.

Finally, in the analysis that follows, only data recorded in 2018 have been considered for different reasons. First of all, looking at the peculiar outline of $PM_{10}$ time series, it is straightforward to notice that there is a strong annual seasonality providing the U-shaped repeating pattern that can be seen in Figure 1.2. Considering one year at a time within the available time horizon (2014-2020), and excluding 2020 due to possible COVID-19 influences over the data trend (see Lonati & Riva (2021b,a); Cameletti (2020)), 2018 was the year presenting the lowest number of missing values, considering obviously the whole data observations recorded by all the four regions. Given the just mentioned specifications, in the rest of the paper only data in 2018 were considered, for a *restricted time horizon* of 365 days going from 01-01-2018 to 31-12-2018 (since 2018 was not a leap year). We plan to extend the descriptive analysis and the Bayesian models to more subsequent years.

### 1.1.3. Data Transformation

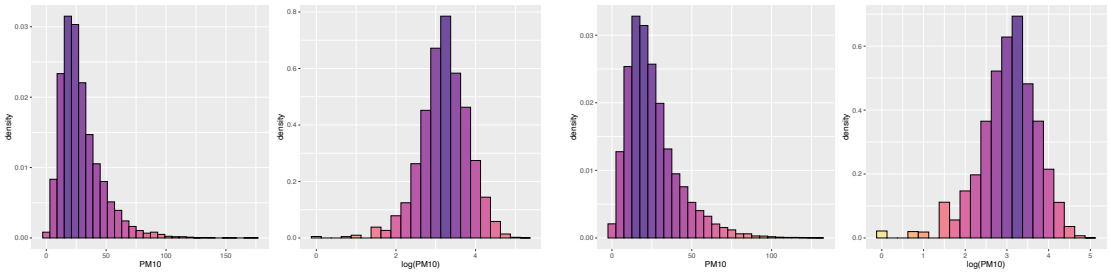
After the data "polishing" described in Section 1.1.2, the logarithmic transformation was applied to all $PM_{10}$ concentration values. The choice of transforming data was mainly leaded by two reasons: first of all it is common use to log-transform the PM concentrations in the atmospherical research field (see Taylor et al. (2018)) since the data are strictly positive. Hence, for a matter of coherence with the already published studies, it is convenient to apply the same transformation to the $PM_{10}$ data. The second advantage of log-transforming $PM_{10}$ concentrations is that this transformation will make the data distribution more symmetric, which is a very useful property to exploit in model definition. In fact, the PM concentrations are usually assumed to be log-Gaussian distributed, since the logarithmic transformation help in fitting the data to a Gaussian framework. To give an idea of this effect, the empirical distributions of the $PM_{10}$ values before and after the log-transformation are reported below in Figure 1.3. Observe that the values of $PM_{10}$ concentration were initially ranging between 1 $\mu g/m^3$ and 120 $\mu g/m^3$, while after the log-transformation the data variability was significantly reduced, ranging between 0 $\mu g/m^3$ and 5 $\mu g/m^3$. Hence the log-transformed data show a smaller empirical variance, which is preferable since it will be easier to handle during the MCMC procedure. Finally, the $PM_{10}$ time series before and after log-transformation are reported in Figure 1.4.




(a) Lombardia




(b) Emilia-Romagna

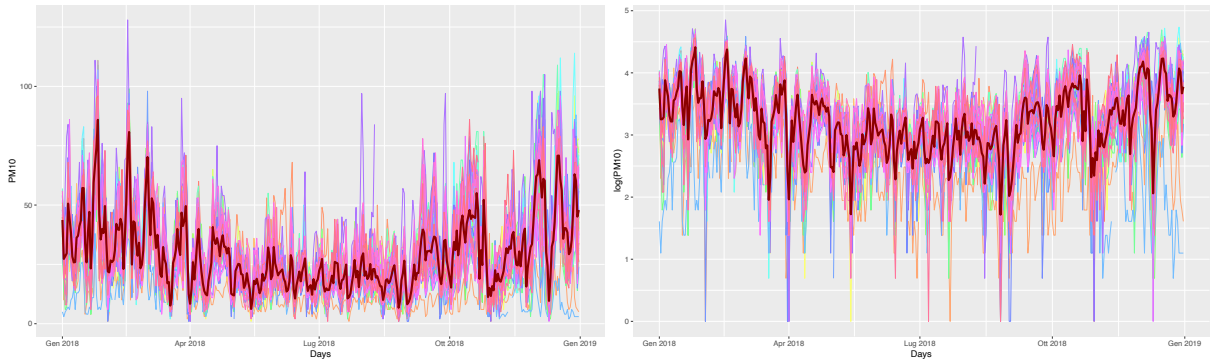



(c) Veneto

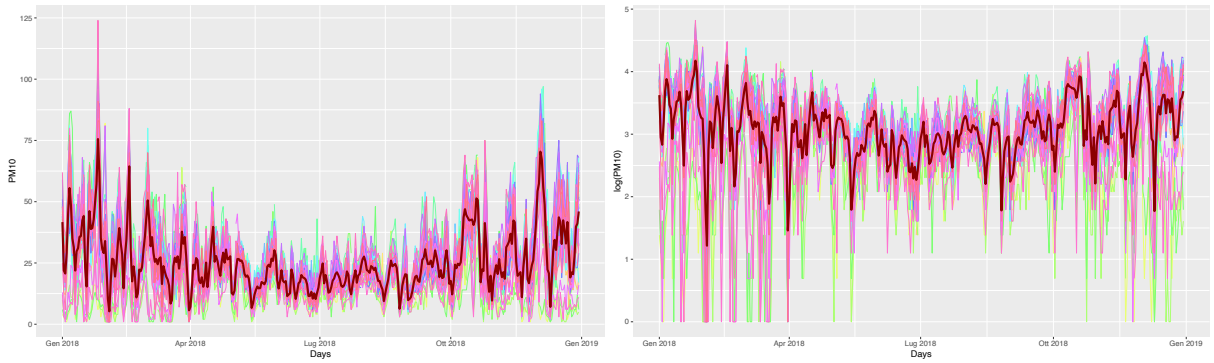



(d) Piemonte

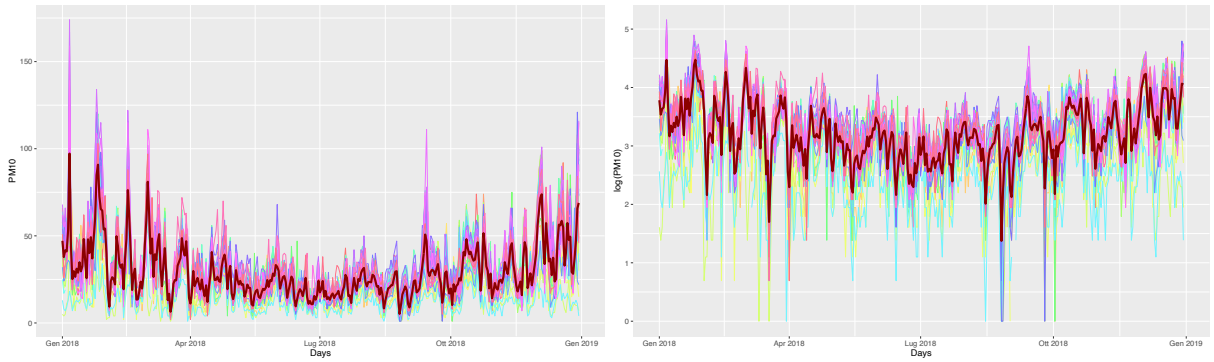
Figure 1.3: Empirical distribution of $PM_{10}$ concentrations before (on the left) and after (on the right) logarithmic transformation in the four regions.



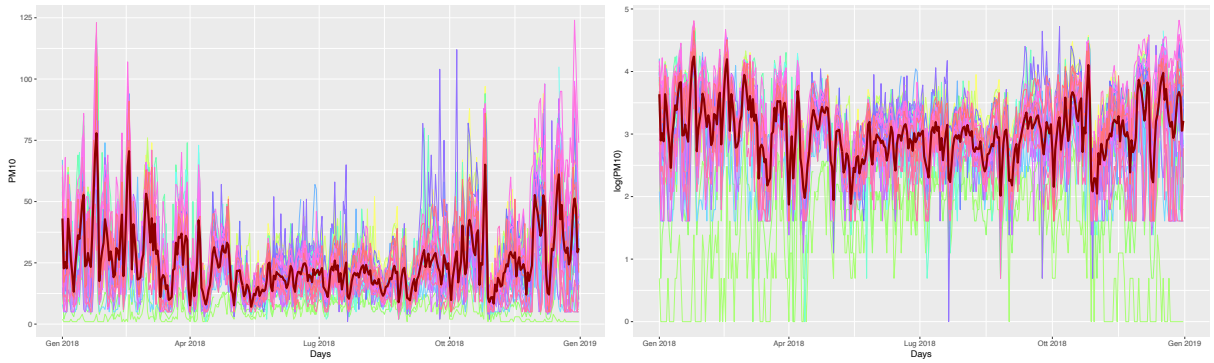

(a) Lombardia




(b) Emilia-Romagna




(c) Veneto




(d) Piemonte

Figure 1.4: Data trend in the four regions before and after the logarithmic transformation of $PM_{10}$ concentrations.

## 1.2. Explorative Data Analysis

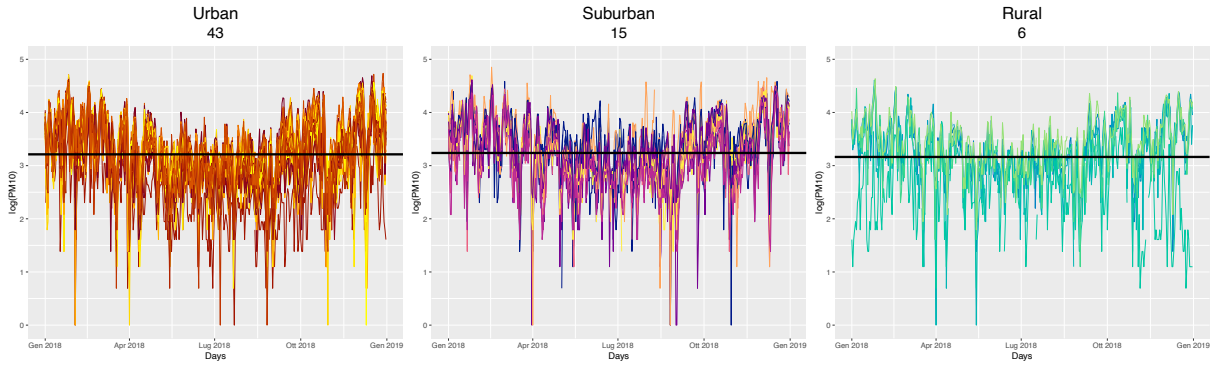
Here is reported the exploratory data analysis of the 2018 data which will be considered in the rest of the Chapter. A graphical analysis of the more relevant categories is presented as follows, aiming at enlightening data behavior in the four regions and defining some peculiar patterns that may be useful for the model definition. The data will be treated as time-series, assuming the $PM_{10}$ log-concentrations as a function of time, having one value for each day over a time horizon of one year (2018). Each station will be represented by a single time series reporting the daily registrations of that station over the 365 days under study. Notice that, even reducing the analysis to a single year, the the number of data still be considerable. Indeed there will be used $365 \cdot 49 = 17885$ observations in Emilia Romagna, $365 \cdot 64 = 23360$ in Lombardia, $365 \cdot 51 = 18615$ in Piemonte and $365 \cdot 37 = 13505$ in Veneto.

### 1.2.1. Area

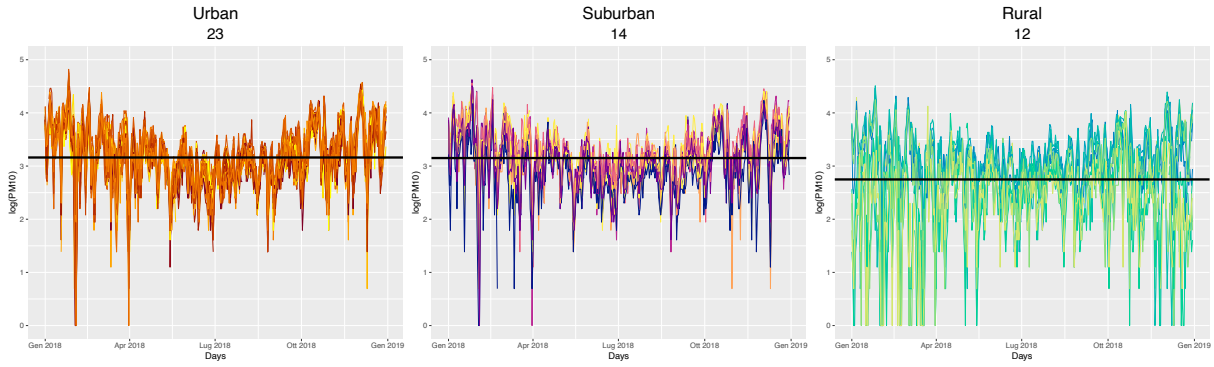
Each detection station should be located in a suitable position, in order to be as reliable as possible in representing the air quality status of the agglomerate or zone in which it is located. The specific area characterizing the zone in which each station is located can be identified with one of these three possible labels :

- Urban: station located in a completely, or at least predominantly built area.
- Suburban: station located in a zone which is mostly built, but in which both built-up areas and not urbanized areas are present.
- Rural: station located in a zone that can not be labeled neither as urban or suburban. Moreover if the station is located at a distance higher than 50 km from pollution sources (f.i. industrial exhales or busy streets), the site is labeled as remote rural.

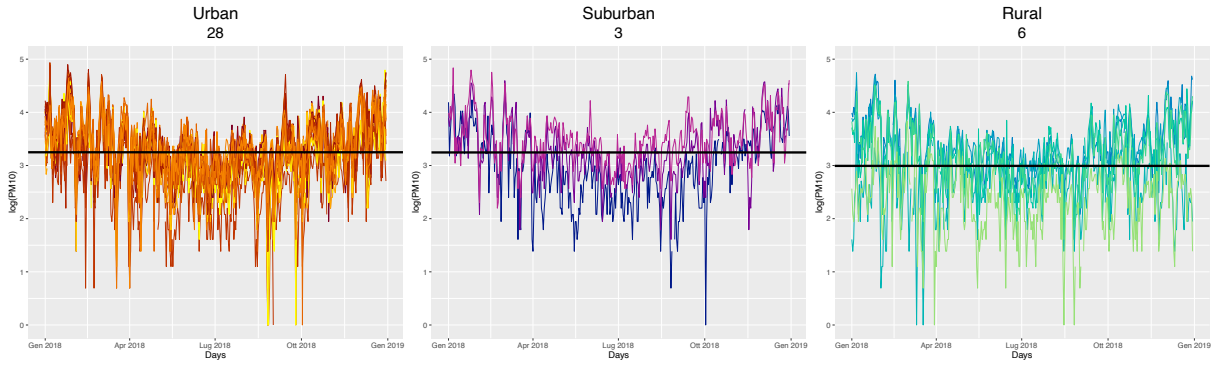
The $PM_{10}$ log-concentrations registered in these three different type of zones are reported in Figure 1.5. It can be noticed that the U-shaped behaviour characterizing the PM pollution level is emphasized in the urban and suburban areas, while the rural stations seem to follow a more linear trend, or in some cases even an inverse U-shaped trend.



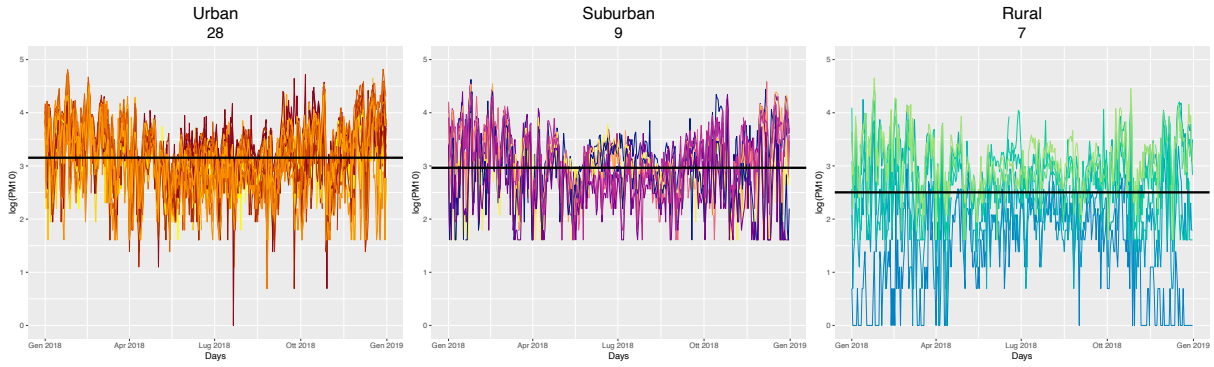

(a) Lombardia




(b) Emilia-Romagna




(c) Veneto




(d) Piemonte

Figure 1.5: Time series of $PM_{10}$ data recorded by each station, divided according to their area characterization.

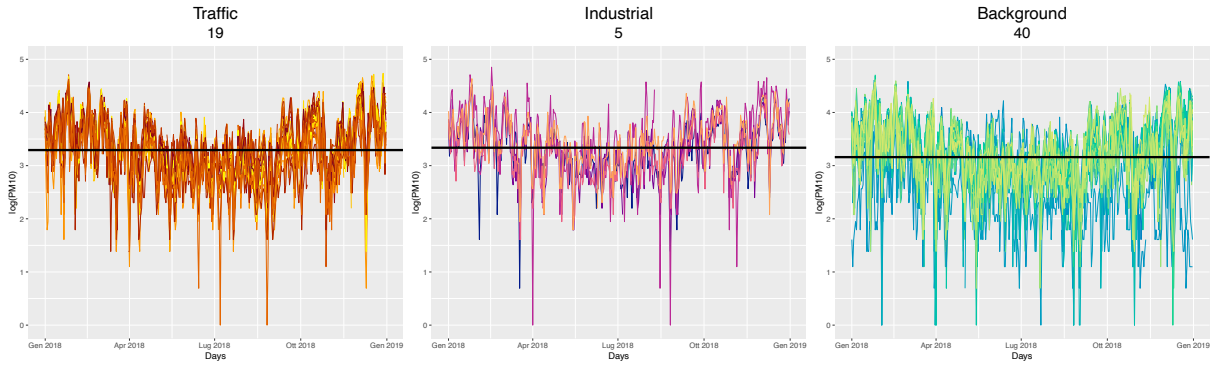
### 1.2.2. Type

Independently from the area in which the station is located, three possible types of stations can be identified. Indeed, as specified in ARPA Lombardia (1999), a good pollutant detection network should be provided with both background stations, measuring the diffuse pollution that is universally spread in the territory, and stations located in peak positions, for instance nearby streets or factories, in order to be able to measure the air quality also in this more critical scenarios. The categories describing the specific station type are the following:

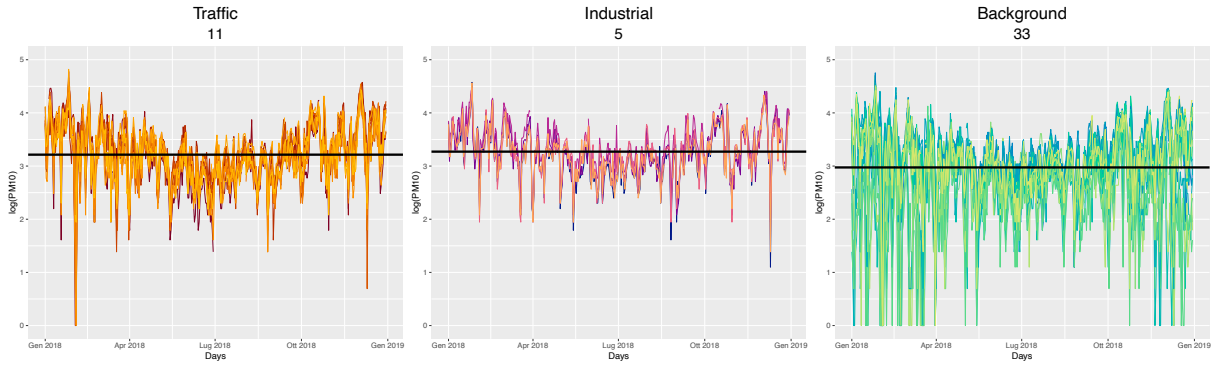
- Traffic : stations located in specific positions in which the air pollution is mainly due to vehicle traffic emissions, coming from highly or medium travelled neighboring streets.
- Industrial : stations that are located close to industrial areas, specifically the stations having a position such that it can be assumed that the recorded pollutants mainly derive from a unique industrial source or from a neighbouring industrial site.
- Background : Stations located in positions for which it cannot be assumed that the air pollution level is mainly due to specific sources (industries, traffic, domestic heating etc.). For this kind of stations the recorded values of $PM_{10}$ are given by the cumulative contribution of all the pollution sources located windward with respect to the station.

According to this specific categorization of the stations, it is common to observe higher values of pollutant in the time series recorded by traffic and industrial stations.

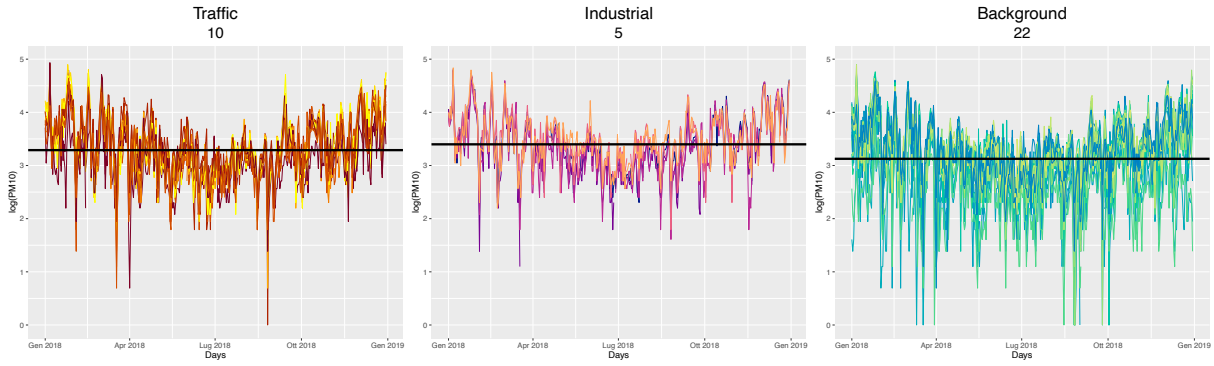
In Figure 1.6 is reported the graphical overview of the $PM_{10}$ log-concentrations registered during 2018 by each station, divided according to the just mentioned type categories. Also in this case the analysis has been made separately for the four regions under study.



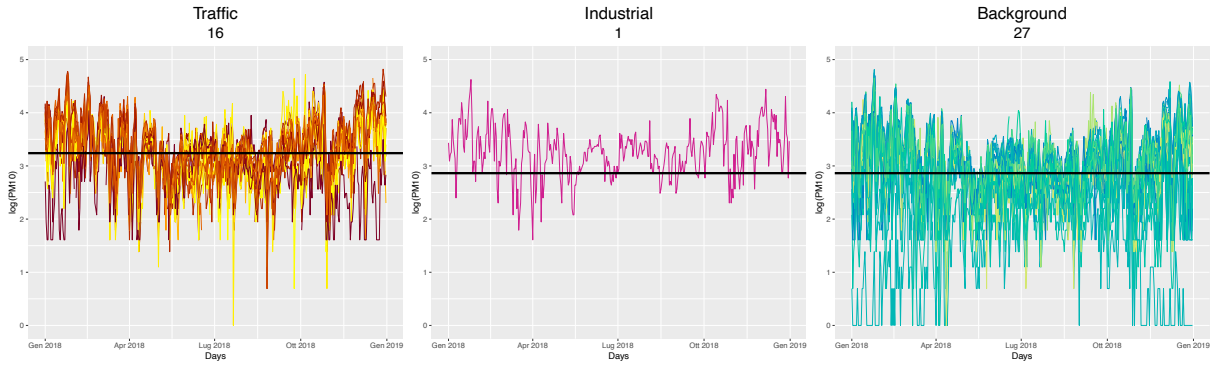

(a) Lombardia




(b) Emilia-Romagna




(c) Veneto

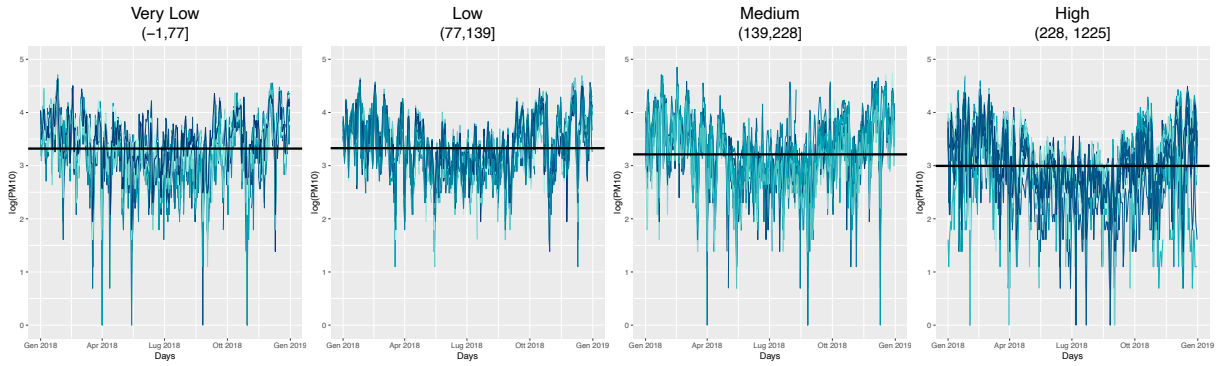



(d) Piemonte

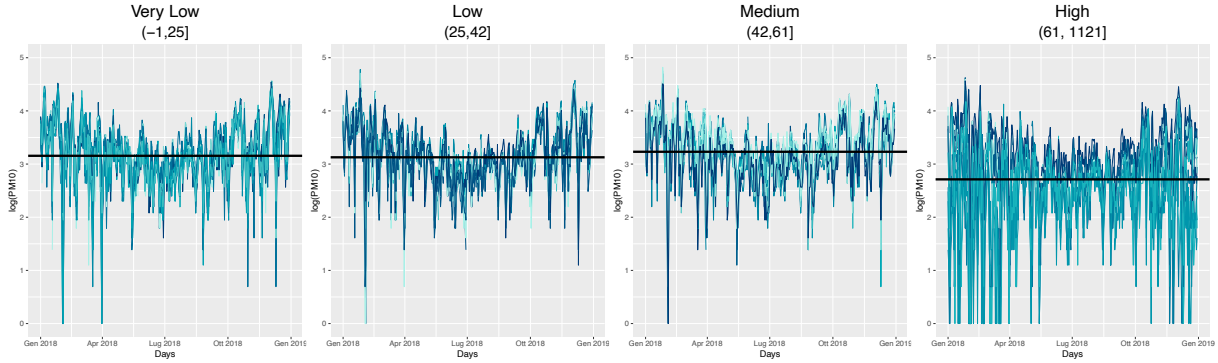
Figure 1.6: Time series of $PM_{10}$ data recorded by each station, divided according to their type characterization.

### 1.2.3. Altitude

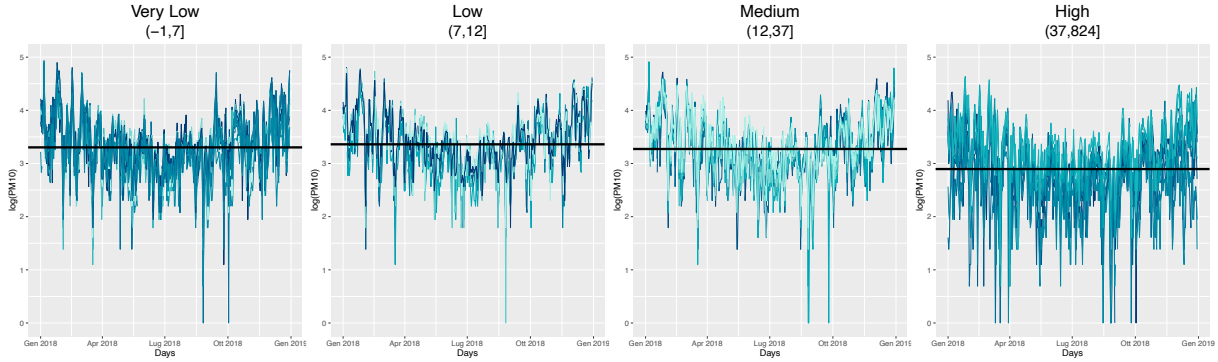
The altitude at which the $PM_{10}$ measurements are taken plays a fundamental role in the phenomenon explanation. Indeed while it has already been proven that at higher altitudes the general outgoing of air pollution seems to get worse, as widely discussed in U.S. EPA (1978), in the case of particular matters such as $PM_{10}$ or $PM_{2.5}$ the converse happens. The concentration of these specific pollutant seem to decrease when the altitude is increasing, delineating an inverse relation between the recorded level of air pollution and the distance of the station from the sea level. This peculiar effect is probably due to the extremely specific nature of this kind of air pollutant, indeed as displayed in Larsen et al. (2012) the main sources of $PM_{10}$ pollution are linked to anthropological issues such as transport emissions, brake or tire wear and re-suspension of soil dust. In the territory under study the altitude increases only in strongly mountainous areas, which are characterized by a weaker urbanization and present less dense populated districts, leading to a milder anthropization of the environment and hence to a minor presence of roads and transport infrastructures. Moreover also the minor sources of PM pollution, such as domestic heating or soil re-suspension, can be assumed to be less present in the mountain setting since also the habitation are usually dislocated and it is pretty hard to find big urbanized cores. In conclusion analyzing the continuous datum characterizing the altitude of the stations in the four regions, higher values of pollutant are expected to be found at lower height and vice versa, coherently with the just described $PM_{10}$ behaviour. Figure 1.7 reports the data outgoing in the four qualitative altitude-driven groups in which the data have been divided. The time-series recorded by every station have been partitioned among recording sites located at *Very Low, Low, Medium* or *High* heights. This discretization of the altitude factor (which is in fact a continuous variable) corresponds to a discretization in classes fixed by quantiles, i.e. accordingly to their belonging to the first, second, third or fourth quartile, computed with respect to the empirical distribution of stations heights.



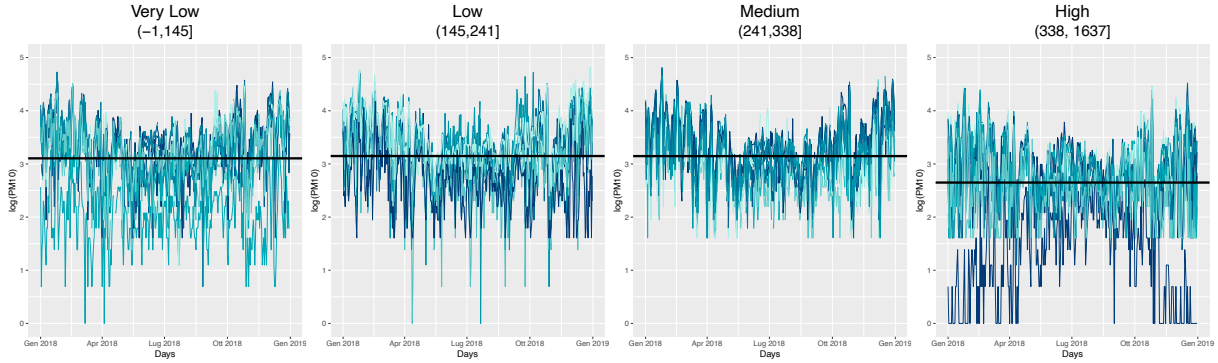

(a) Lombardia




(b) Emilia-Romagna




(c) Veneto




(d) Piemonte

Figure 1.7: Time series of $PM_{10}$ data recorded by each station, divided according to their altitude class.

### 1.2.4. Zoning

The zoning consist in a region-specific partitioning of the whole territory, independently of municipalities or provinces in the region. Since the measurement of air quality is useful to guarantee the protection and the maintenance of the population health, a categorizing of the territory is of crucial importance for the regional organization. Being region-specific, this kind of categorical information is not suitable to be included in a hierarchical model, but still pretty useful to understand the data behavior and distribution across the areas of the regional domain. Coherently with the nature of this characterization of recording sites, the graphical analysis reporting the time series related to the zoning will be exhibited for one region at a time.

Investigating the zoning characterization of Lombardia, all the possible categories can be summarized as follow :

- Urban agglomerations in the most extended and populated cities of the region (Bergamo, Brescia, Milano).
- Zone A : plains with high levels of urbanization.
- Zone B : zone characterized by prevalence of lowlands.
- Zones C and D : mountainous areas (Prealpi and Appennino) including also the valley bottom.

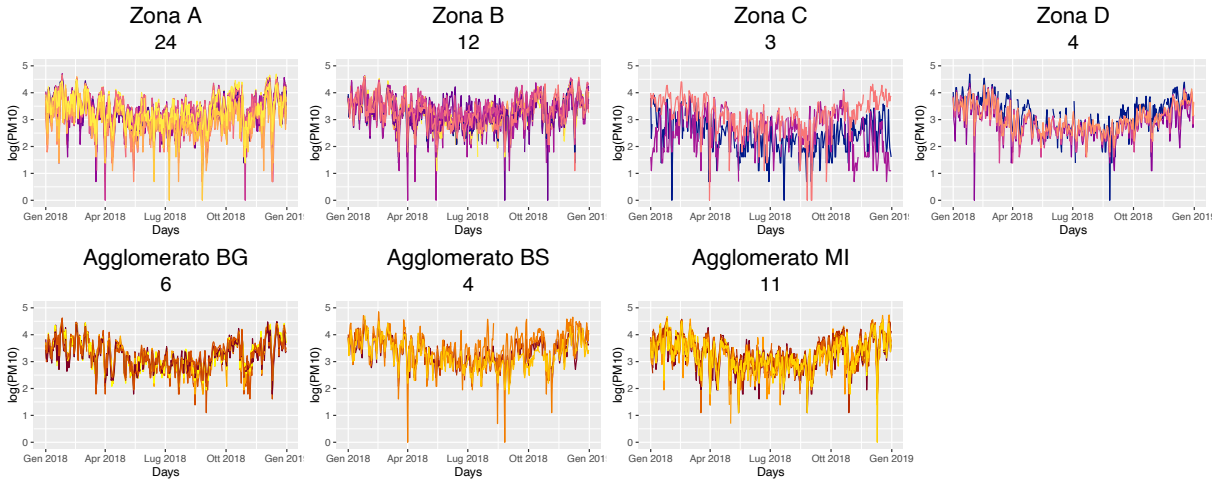



Figure 1.8: Region specific zone characterization for the recording stations in Lombardia.

Figure 1.8 shows that lower concentrations of $PM_{10}$ were registered in zone C and zone B, corresponding to plains and mountains devoid of significantly urbanized cores, while in the more urbanized areas the pollutant concentrations tend to assume higher values.

Moving on to the next region, Emilia-Romagna zoning can be summed up as:

- Urban agglomerations, including all the stations located in highly built-up areas.
- Mountainous area (Appennino).
- Zones characterized by a plain setting (Pianura Est, Pianura Ovest) that however could include some urbanized areas.

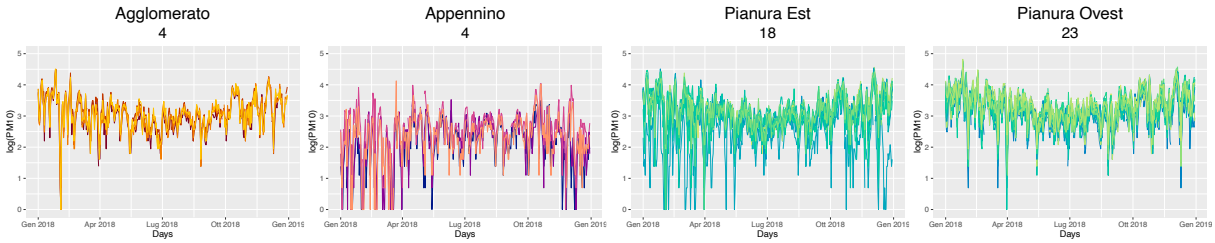



Figure 1.9: Region specific zone characterization for the recording stations in Emilia-Romagna.

From Figure 1.9 it can be noticed that in the mountainous and east plain areas lower levels of pollution were registered, especially during the winter period, while in the urban agglomerate and in the west plain the pollutant concentration takes larger values.

Looking then at the zoning describing Veneto territory, three main macro-categories can be highlighted :

- Urban agglomerations in the most extended and populated cities of the region (Padova, Treviso, Venezia, Vicenza, Verona).
- Portions of territory characterized by the preponderant presence of plains and lower hills (Bassa pianura e colli, Pianura e capoluogo della bassa pianura).
- Mountainous areas (Alpi e Prealpi).

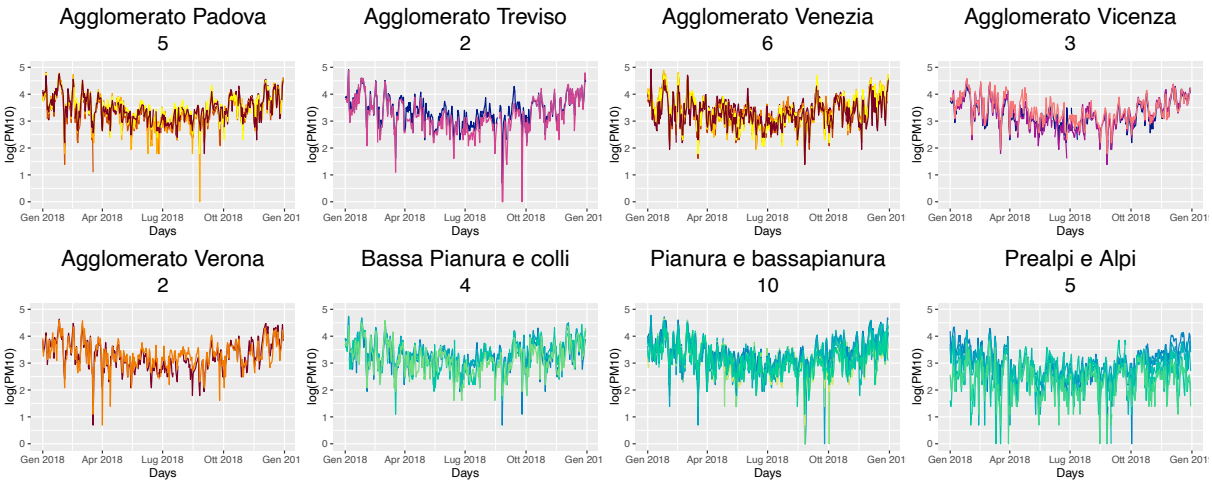



Figure 1.10: Region specific zone characterization for the recording stations in Veneto.

Also looking at Figure 1.10, as expected, the concentrations of $PM_{10}$ registered in the urban agglomerates appear to be higher with respect to the measurements performed in less urbanized areas, such as plains, hills and mountains.

Finally, considering the data regarding the Piemonte region, the following zoning classification is presented:

- Urban agglomeration of the most populated city Torino (TO).
- Montainous areas (Montagna)
- Territory characterized by the preponderant presence of hills (Collina).
- Territory characterized by the preponderant presence of plains (Pianura).

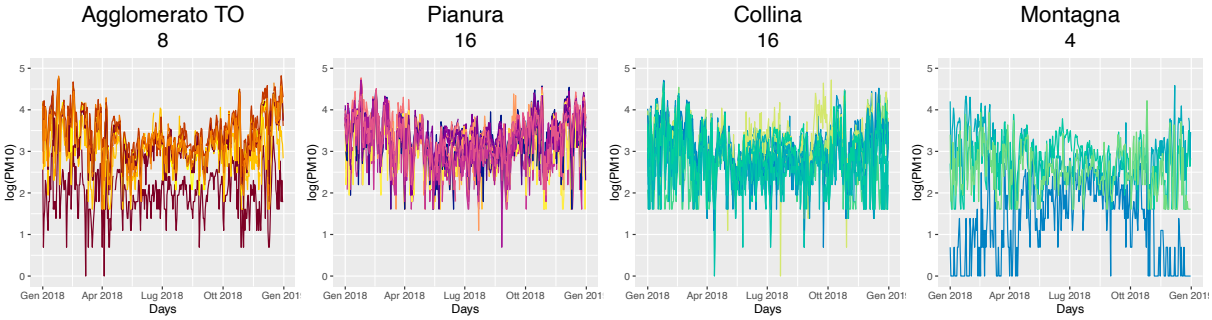
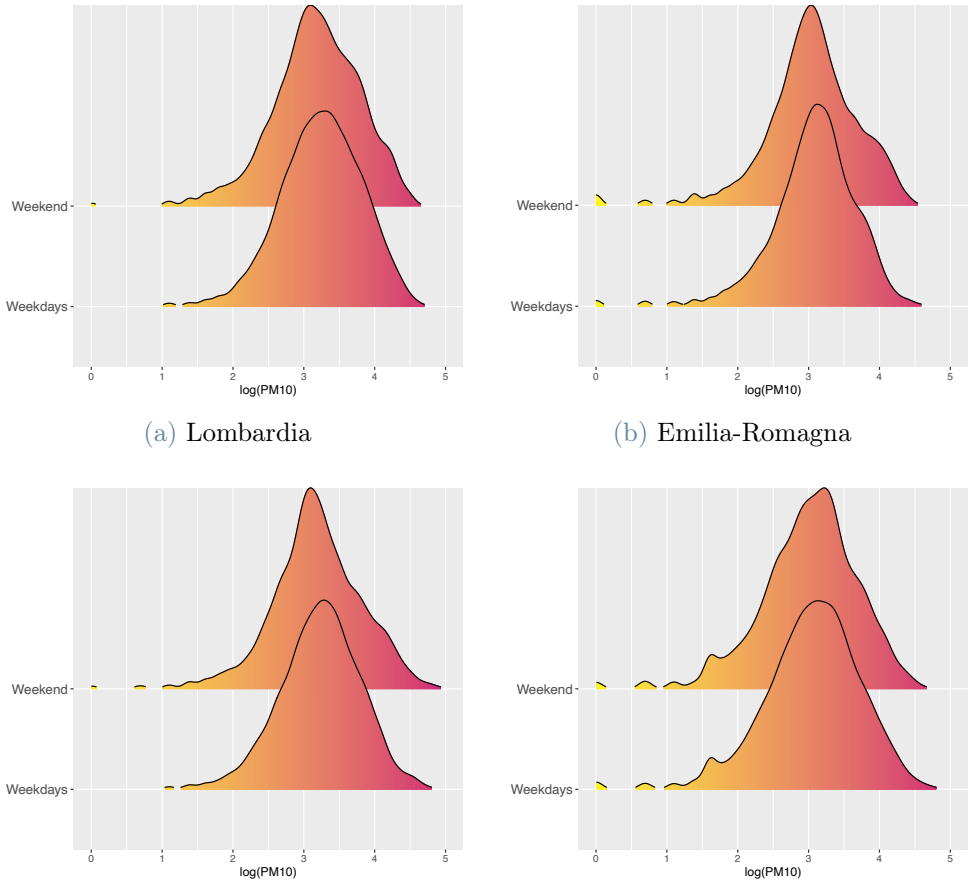



Figure 1.11: Region specific zone characterization for the recording stations in Piemonte.

Figure 1.11 shows that the lower concentrations of $PM_{10}$ seem to be registered in the zones including mountains and hills, while the Turin urban core and the land zone are characterized by higher levels of air pollution.

### 1.2.5. Seasonality

Dealing with time series such as the daily recordings of $PM_{10}$ concentration, entails very often the presence of some periodical patterns. A first good practice is to investigate the presence of differences between the daily values registered during weekdays versus the concentrations observed in the weekends. This first inspection aims at identifying the influence of some anthropic phenomena, such as the work leaded traffic or the depopulation of urbanized cores during weekends, that may affect the preponderance of air-pollution. In order to verify if the recorded values actually behave differently in weekend or weekdays, the data distributions are visually compared below, considering as usual the data partitioned in the four different regions of interest. As can be seen in Figure 1.12 no relevant difference can be identified among the two distributions, which means that the data tend to assume the same range of values in the two scenarios and hence no distinction is needed when dealing with weekend or weekdays data.

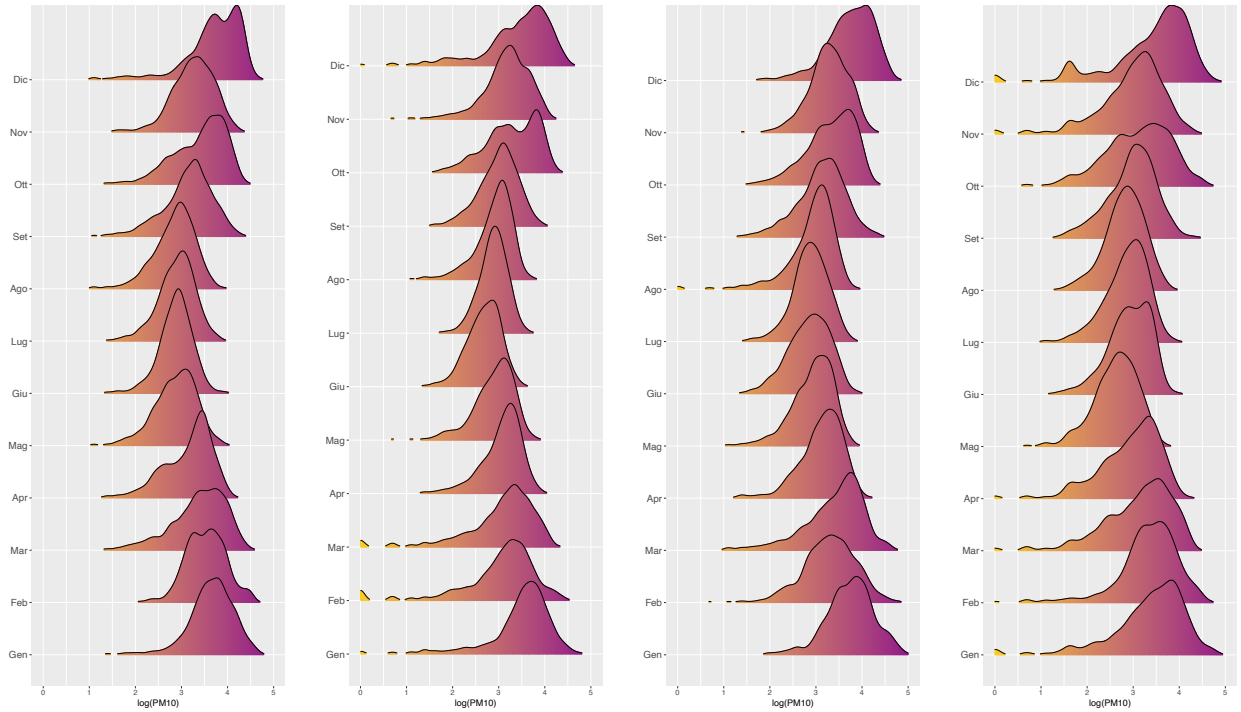



(a) Lombardia (b) Emilia-Romagna

(c) Veneto (d) Piemonte

Figure 1.12: Estimated kernel densities of $PM_{10}$ log-concentrations registered during weekend or weekdays.

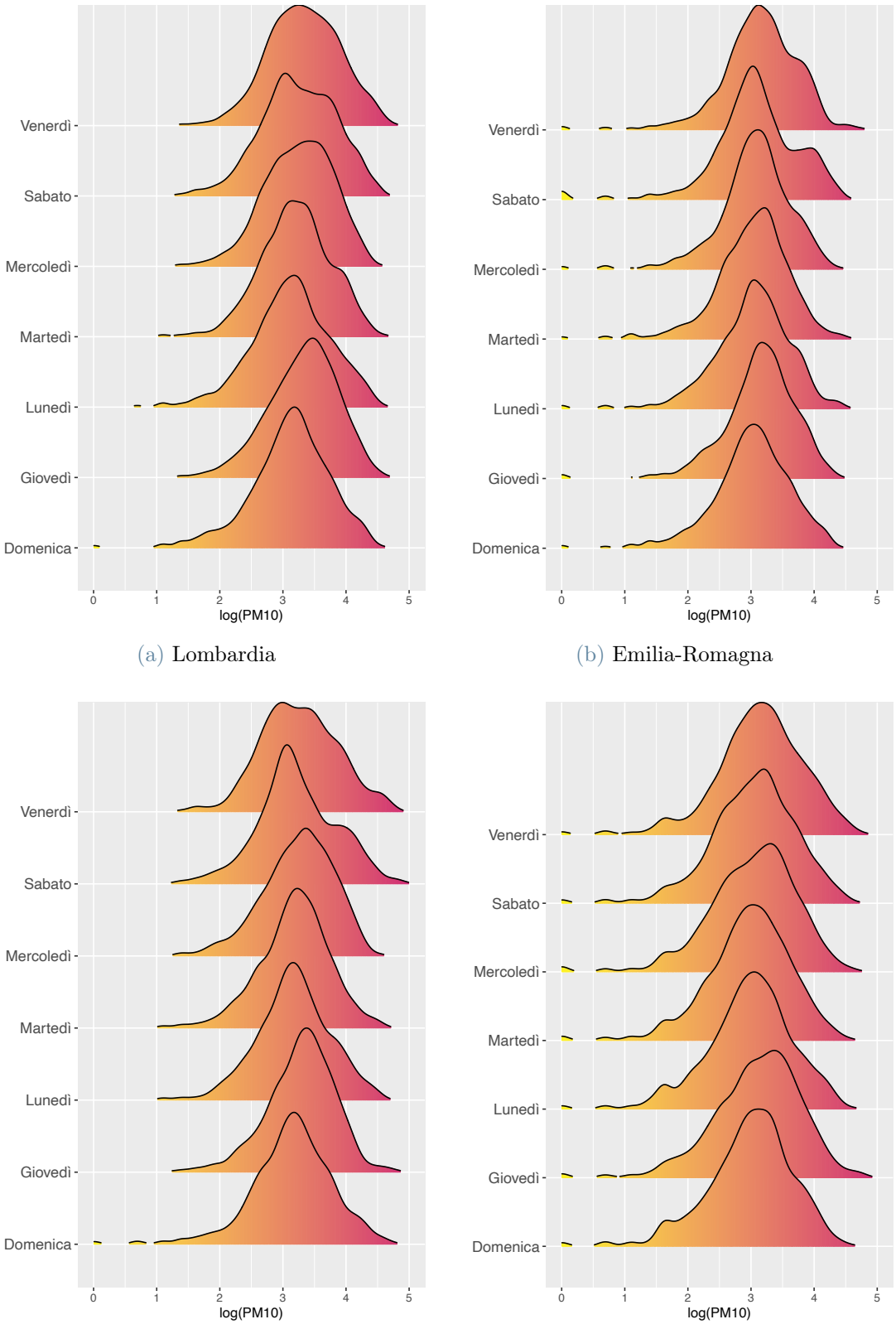
Looking for a peculiar pattern due to the month in which the registrations took place, the distributions in Figure 1.13 are outlined. It is straightforward to notice that in the cold season the data variability is higher and the pollutant log-concentration reaches bigger values, while in the hottest summer months the variability seems to be reduced and the air pollution level appear to be lowered and concentrated around smaller values. The discovering of this pretty specific behaviour may come in handy later on when modeling the data variability.




(a) Lombardia (b) Emilia-Romagna (c) Veneto (d) Piemonte

Figure 1.13: Estimated kernel densities of $PM_{10}$ log-concentrations registered during each month of the year (2018).

Another good practice is to investigate the presence of some weekly periodic event. Indeed it could happen that the data behave differently according to the specific day of the week in which they are recorded. This phenomenon could be due to the presence of some event or recurrence that take place always (or very often) the same day of the week. As before, the idea to check for the presence of these anomalies is to compare the values distribution in each day of the week ad look for some odds between the densities. By looking at Figure 1.14 it can be noticed that also in this case no peculiar pattern has been pointed out, meaning that the day of the week in which the recording takes place does not influence the $PM_{10}$ observed value.




(a) Lombardia (b) Emilia-Romagna

(c) Veneto (d) Piemonte

Figure 1.14: Estimated kernel densities of $PM_{10}$ log-concentrations registered during the seven different days of the week.

## 1.3. Building the Model

For the model-based analysis we consider only data from Emilia-Romagna, since it was the region presenting less missing values and hence the most suitable one for defining a model following as precisely as possible the real outline of data.

Figure 1.15 shows the time series recorded by the 49 stations in Emilia-Romagna, which have been divided into two groups according to their area characterization. It is straightforward to distinguish two completely different annual trends when grouping the rural and non-rural stations (where non-rural means urban or suburban). This partition was driven by the very similar behaviour of suburban and urban stations, in contrast with the rural ones, as can be seen in Figure 1.5b. Indeed while the non-rural stations follow a U-shaped outline, the trend of rural stations seems instead to define a sort of M-shaped silhouette. According to this clearly different performances, two specific functions of time will be defined, in order to better follow the trend of these two different class of recording stations. Hence, following a functional approach which will be better specified in Section 2.4, the average annual trend will be expressed by two different shaped functions $f_R(t)$ or $f_{NR}(t)$ depending on the station area classification, where $t$ denotes the day of the year, i.e. $t \in \{0, \dots, 364\}$.

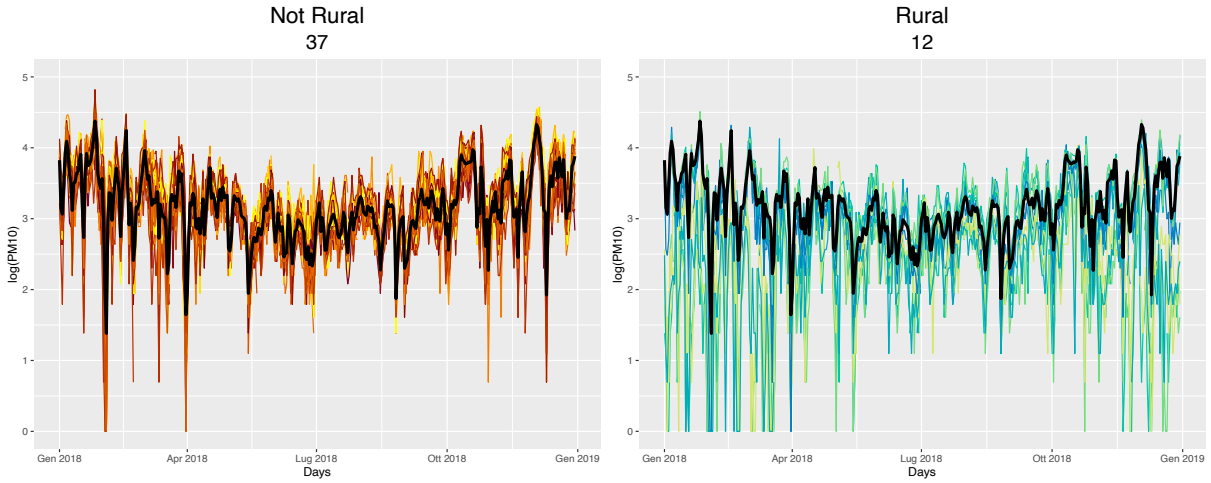
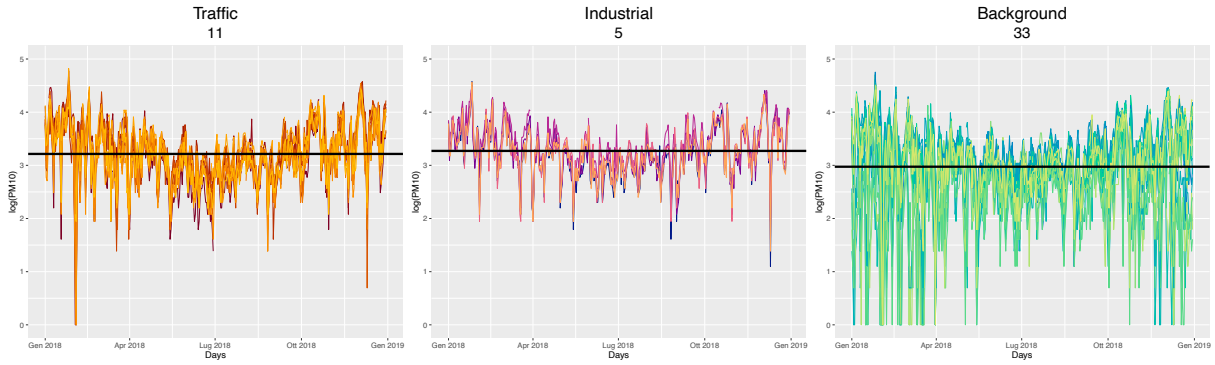


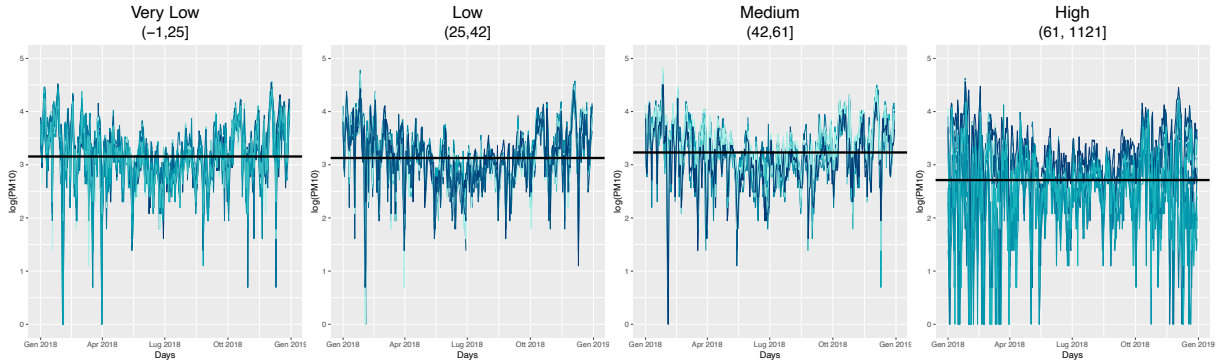

Figure 1.15: Time series of $PM_{10}$ log-concentrations recorded by stations located in non-rural (left) and rural (right) areas. The bold line delineates the daily average pollution level, which was computed among all the stations in the corresponding group.

After having defined the general outline characterizing the time-series of $PM_{10}$ log-concentrations, some relevant features of the sensing stations will be included in the model as covariates. First of all, considering the type class to which the station belongs (Traffic, Industrial

or Background), it is easy to see that higher values of pollutant have been registered by traffic stations, while the lowest reached values were collected by background stations. Hence this feature will be included into the model as a categorical regressor, raising or lowering the expected pollution level depending on the type of station which is recording the $PM_{10}$ concentrations. The categorical classification of stations will be included in the form of two *dummy variables*, assuming as default value the one related to background stations ([0,0]) and adding alternatively two different terms if the station is in a traffic zone ([1,0]) or in an industrial zone ([0,1]). In Figure 1.16a are reported the time series recorded by 49 stations in Emilia Romagna, grouped according to their type categories. By taking a look at the just mentioned graphics, it is not wrong to suppose that the coefficients related to these two dummies will intuitively take value in a strictly positive numeric range. Indeed observing the overall mean of values in the three categories (represented by the black horizontal line), the background type class of stations fluctuates around a reportedly lower mean with respect to the industrial and traffic ones.




(a) Station Type




(b) Altitude

Figure 1.16: Time series of $PM_{10}$ log-concentrations grouped by stations type (top row) and altitude (bottom row), together with their overall mean value.

An analogous reasoning can be made for the stations altitude, but in this case the regressor would be continuous. As can be seen in Figure 1.16b the pollution level is expected to decrease when the altitude increases, thus the coefficient associated to the altitude covariate is expected to take value in a strictly negative numeric range. Moreover, since the altitude specification covers a fairly wide range of values, going from $0m$ to $1121m$ over sea level, it is a good practice to standardize this continuous variable before including it in the model as a covariate in order to obtain better performances and higher accuracy.

Other peculiar features of the $PM_{10}$ dataset are enlightened by the monthly empirical distributions reported in Figure 1.17. At first it can be noticed the tendency of pollution level to be lower in summer and higher in winter, as expected by taking a look to previous studies (see Dung et al. (2019); Onuorah et al. (2019); Tian et al. (2014)). But an even more interesting observation can be made by looking at Figure 1.17, that is the tendency of log-concentrations to be distributed over a wider range of values during the cold season, while they seem to be shrinked over a restricted numerical range during the hot season. This phenomenon provides important information about the data variability, which appears to be different according to the time period of the year. Hence, when actually modelling the data in Section 2.4, a relevant role will be assumed by this peculiar feature, providing the basis to make important assumptions about the proposed model (specifically in Subsection 2.4.3).

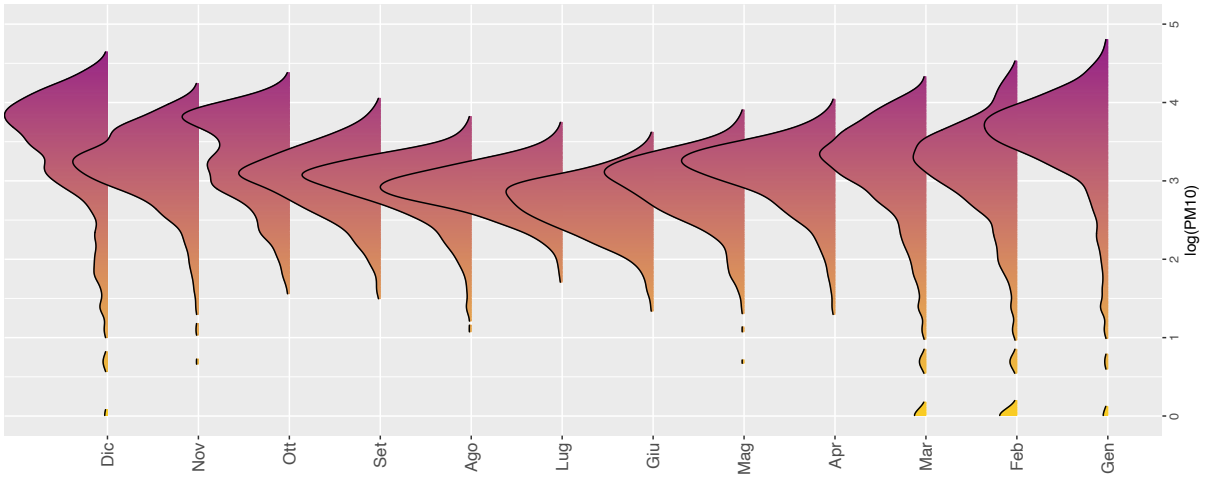



Figure 1.17: Empirical distributions of $PM_{10}$ log-values recorded in each month of 2018.

Before applying the model, we have standardized the responses ($PM_{10}$ log-concentrations), i.e. we have subtracted the overall mean and divided by the overall standard deviation. All the model proposed in Chapter 2 and the related posterior inference reported in Chapter 3 will refer to the standardized data.

# 2 | Hierarchical Modeling for $PM_{10}$ Data

The aim of this study is to build a spatio-temporal model, able to explain the trend of $PM_{10}$ log-concentrations over 2018, and to include territorial information. The chosen approach is Bayesian. First of all, the likelihood of this method is defined, and the basic theory of spatial models is quickly reviewed. Then, this general framework will be fitted to the specific case of the data under study, that is the time-series of $PM_{10}$ log-concentrations collected in Emilia-Romagna during 2018. Different suitable models, which could fit the data, will be discussed. The best performing model will then be chosen among the proposed ones, exploiting well known predictive information criteria, which will be computed in Chapter 3. Finally the Bayesian approach for this kind of problems will be briefly summarized and the Monte Carlo Markov Chain (MCMC) method will be explained.

## 2.1. Hierarchical Modeling for Spatial Data

In this section will be provided a review of basic theory of stationary spatial process models, which are the building blocks for more flexible spatio-temporal models. First of all, *spatial point-referenced* data need to be introduced, reporting definitions from Chapter 1 of Banerjee et al. (2015). $Y(\mathbf{s})$ will represent a random vector evaluated at a location $\mathbf{s} \in \mathbb{R}^r$, where $\mathbf{s}$ varies continuously over $D$, a fixed subset of $\mathbb{R}^r$ that contains an $r$-dimensional rectangle of positive volume. In the study case $r = 2$ will be fixed, since each site $\mathbf{s}$ will be characterized by two geographical coordinates (Latitude and Longitude). As follows material from chapter 2 of Banerjee et al. (2015) is presented.

A spatial process is *Gaussian* if for any $n > 1$ and any set of sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, the multivariate response variable $\mathbf{Y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^T$ follows a multivariate normal distribution. A process is then labeled as *strictly stationary* if for any $n > 1$ , any set of sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and any vector $\mathbf{h} \in \mathbb{R}^r$ the vector $[Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]$ has the same distribution as the vector $[Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h})]$ . An interesting subcategory of these

processes is given by spatial processes having a constant mean, i.e. $\mu(\mathbf{s}) = \mu$ for any $\mathbf{s}$, and being such that their covariance matrix can be defined as $Cov[Y(\mathbf{s}) \ , \ Y(\mathbf{s}+\mathbf{h})] = C(\mathbf{h})$ for any $\mathbf{h} \in \mathbb{R}^r$, i.e. the covariance of the response variable measured into two different sites only depends upon distance among the sites. If these assumptions are verified, the spatial process is said to be *weakly stationary* (or second-order stationary). The following discussion will be restricted to *Gaussian stationary spatial process*. Hence the covariance relationship between process' values at any two locations will be expressed by a covariance function $C(\mathbf{h})$, depending only on the value of the separation vector $\mathbf{h} \in \mathbb{R}^r$.

Another fundamental element to be defined when dealing with spatial processes is the *variogram*. The variogram is a function of $\mathbf{h} \in \mathbb{R}^r$ which can be defined for *intrinsically stationary processes*, i.e. processes for which $E[Y(\mathbf{s}+\mathbf{h}) \ - \ Y(\mathbf{s})] = 0$ and such that the value of $E[Y(\mathbf{s}+\mathbf{h}) \ - \ Y(\mathbf{s})]^2$ does not depend on the site $\mathbf{s}$, but only on the value assumed by $\mathbf{h}$. For these spatial processes the variogram function can then be defined as:

$$2\gamma(\mathbf{h}) = Var(Y(\mathbf{s}+\mathbf{h}) \ - \ Y(\mathbf{s})) \tag{2.1}$$

If the variogram of the spatial process is well defined, then also the function $\gamma(\mathbf{h})$ is well defined and it is called *semivariogram*. The relationship between the variogram and the covariance function can be summarized as follows:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \tag{2.2}$$

where $C(\mathbf{0})$ denotes the variance of the stationary process. Another interesting group of processes is given by spatial processes presenting an *isotropic* variogram. For these processes the value assumed by the semivariogram function $\gamma(\mathbf{h})$ depends only on the length of the separation vector $||\mathbf{h}||$; hence the variogram is real-valued function having a univariate input argument and can be written as $\gamma(||\mathbf{h}||)$. On the contrary, if the previous does not hold, the variogram is said to be *anisotropic*. Concluding the list of definitions, if a process is both intrinsically stationary and isotropic then in the literature it is said to be *homogeneous*. Specifically, if a homogeneous Gaussian process is assumed, simple parametric formulations can be proposed as feasible candidates for the semivariogram. Here will be used $d$ to denote the separation vector length $||\mathbf{h}||$. Again, for ease of reading, the following material is reported from Banerjee et al. (2015). The *exponential kernel* is characterized by semivariogram and covariance functions, as specified below:

$$\gamma(d) = \begin{cases} \sigma^2 + \alpha^2(1 - e^{-\phi d}) & \text{if } d > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

$$C(d) = \begin{cases} \alpha^2 + \sigma^2 & \text{if } d = 0 \\ \alpha^2 e^{-\phi d} & \text{if } d > 0 \end{cases} \tag{2.4}$$

Another interesting example of homogeneous Gaussian processes is when the variogram is the *Matérn Kernel*:

$$\gamma(d) = \begin{cases} \sigma^2 + \alpha^2 \left[1 - \frac{(2d\phi\sqrt{\nu})^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2d\phi\sqrt{\nu})\right] & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases} \tag{2.5}$$

$$C(d) = \begin{cases} \alpha^2 + \sigma^2 & \text{if } d = 0 \\ \frac{\alpha^2}{2^{\nu-1}\Gamma(\nu)} (\phi d)^\nu K_\nu(\phi d) & \text{if } d > 0 \end{cases} \tag{2.6}$$

where $\nu > 0$ is a parameter controlling the smoothness and $\phi$ is a spatial decay parameter. The function $\Gamma()$ is the well-known gamma function, while the function $K_\nu()$ is the modified Bessel function of order $\nu$. Note that assuming $\nu = 1/2$ the Matèrn kernel and the exponential kernel coincide, while considering instead $\nu \to \infty$ the Gaussian kernel is obtained (see Banerjee et al. (2015), Section 2.1.3). A very popular formulation, which is also the one that will be assumed for the $PM_{10}$ data, is a combination of the two kernels above. Indeed, as suggested in Stan Development Team (2022a) relying also on Rasmussen et al. (2006), a widely used covariance function is the following *exponentiated quadratic function*:

$$C(d) = \begin{cases} \alpha^2 + \sigma^2 & \text{if } d = 0 \\ \alpha^2 \exp\left(-\frac{d}{2\rho^2}\right) & \text{if } d > 0 \end{cases} \tag{2.7}$$

where the hyperparameters $\alpha, \rho$ and $\sigma$ define the behaviour of the covariance function, which is actually obtained through the convolution of two independent Gaussian processes having kernels $C_1(d) = \alpha^2 exp\left(\frac{-1}{2\rho^2} d\right)$ and $C_2(d) = \sigma^2 \mathbb{1}_{d=0}$. Note that the addition of the noise term variability $\sigma^2$ ensures the positive definiteness of the covariance matrix in the case of two identical sites (i.e. in the case in which $d = 0$). Hyperparameter $\rho$ is the *length-scale* parameter, regulating the impact of spatial correlation in the Gaussian process. Small values of the length-scale parameter ($\rho \to 0$) lead the Gaussian process to have nearly null covariance between different sites, while higher values of $\rho$ lead to higher spatial correlation. Hyperparameter $\alpha$ represents the *marginal standard deviation*. It controls the magnitude of the variability assumed to characterize the Gaussian process.

As follow is illustrated further basic material on spatial models, as presented in Chapter

6 of Banerjee et al. (2015). Let $Y(\mathbf{s})$ be the unidimensional process evaluated in a generic site $\mathbf{s}$ and assume:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{2.8}$$

where $\mu(\mathbf{s})$ is the average value. In a generalized linear model (GLM) framework, typically is assumed $\mu(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s})$ is the vector containing the covariates characterizing site $\mathbf{s}$ while $\boldsymbol{\beta}$ is the vector of the corresponding coefficients. The residual of this model is divided in two pieces, one spatial and one non-spatial. The spatial residual term $w(\mathbf{s})$ is assumed to be a realization from a Gaussian spatial process having zero mean and capturing the residual spatial association of data. The non-spatial residuals $\epsilon(\mathbf{s})$ are instead assumed to be uncorrelated pure error terms. Considering the above characterization of residual terms and assuming the data covariance functions formulated as reported in (2.4) and (2.6), the partial sill $\alpha^2$ and range $\phi$ parameters are given by the spatial residual term $w(\mathbf{s})$, while the non-spatial residual term $\epsilon(\mathbf{s})$ brings the additional nugget effect $\sigma^2$. Whereas, seeking for a greater clarity, referring to the notation adopted in the covariance function (2.7) the range parameter is proportional to $\rho$. In the field of spatial prediction, also known as *kriging*, the classical approach (i.e. minimum mean-squared error) can be summarized as follows. In order to predict the value of the response variable $Y$ in a new site $\mathbf{s}_0$ is crucial to identify the best performing model fitting the observed available values $\mathbf{Y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^T$. Reducing the modeling problem only to the Gaussian processes environment, and assuming at first to have no covariates affecting the process outcome, the simplest possible model for observed data is given by :

$$\begin{aligned} \mathbf{Y} &= \mu\mathbf{1} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \Sigma) \end{aligned} \tag{2.9}$$

If the the spatial covariance structure of the process has no nugget effect, the covariance matrix defining the residual term can be specified as:

$$\Sigma = \sigma^2 H(\phi) \quad \text{where} \quad H_{ij}(\phi) = g(\phi;\ d_{ij}) \tag{2.10}$$

where the term $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ represents the distance between the two sites (usually computed using the Euclidean distance, i.e. L2 distance), while the function $g()$ is assumed to be a valid covariance function among those previously defined in this section. If the aim is to build a model including a nugget effect, the residual covariance matrix is instead defined as:

$$\Sigma = \sigma^2 H(\phi) + \tau^2 \mathbf{I} \tag{2.11}$$

where $\tau^2$ is indeed the nugget effect variance.

If covariates $\mathbf{x} = [x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n)]^T$ are also incorporated into the analysis, given that also the covariates in the new location (i.e. $x(\mathbf{s}_0)$) are known, the model assumes the more general form:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \Sigma) \end{aligned} \tag{2.12}$$

where $\Sigma$ is specified as above, depending on the presence or the absence of the nugget effect.

This section is concluded by noting that, so far, the time-series nature of the process under analysis has not be considered. In this section only the spatial side of the phenomenon has been considered for the model proposals, while the temporal aspect did not appear in any formulation. Section 2.4 incorporates both the temporal and the spatial features of $PM_{10}$ data into the final formulation of the model. However, as mentioned at the beginning of this chapter, we will follow the Bayesian approach. In the next section the Bayesian approach to hierarchical models will be briefly described as the natural evolution of the above mentioned classical kriging methods. Indeed assuming to have a Gaussian process following the general formulation (2.12) the response variable will be modeled as:

$$\mathbf{Y}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\ \sigma^2 H(\phi) + \tau^2 \mathbf{I}) \tag{2.13}$$

where $\boldsymbol{\theta}$ represents the parameters vector, which will be specified by a proper prior as better explained in Section 2.4. In this Bayesian setting, parameter estimates will then be obtained from their posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ which accounts for the available observed realizations of the response variable $\mathbf{Y}$.

## 2.2. Bayesian Approach for Hierarchical Models

For ease of reading, as follow is presented material from Chapter 5 of Banerjee et al. (2015). The Bayesian approach combines complex data models and external prior knowledge or expert opinion. Contrary to the classic statistical methods in the regression context, such as Ordinary Least Squares (OLS) aiming at estimating the unknown but fixed parameters value, in the Bayesian framework both the observed data and any unknown parameters are modeled as random variables. Following this approach, the joint distribution for both the observed data $\mathbf{y} = [y_1, \ldots, y_n]$ and the vector of unknown parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]$ has to be specified. One way to specify the joint density is to assign the conditional law $f(\mathbf{y}|\boldsymbol{\theta})$ of $\mathbf{y}$ given $\boldsymbol{\theta}$ first. Then the vector of parameters $\boldsymbol{\theta}$ is assumed to be a random quantity from the *prior distribution* $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ where the vector $\boldsymbol{\lambda}$ contains the

hyperparameters characterizing the prior distribution. If the vector of hyperparameters $\boldsymbol{\lambda}$ is known, inference about the parameters in $\boldsymbol{\theta}$ is based on the following *posterior distribution* :

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\mathbf{y}|\boldsymbol{\lambda})} = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{\int_\Theta p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})\, d\theta} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\theta} \tag{2.14}$$

where $\Theta$ is the multidimensional ($k$-dim) parametric space. The *posterior distribution*, which is the updated distribution of parameters after having observed the available data, is computed exploiting the well known *Bayes theorem* (2.14). Note that in the formula above the contribution of data (given by the likelihood $f$) and the external knowledge (expressed by the prior $\pi$) are both considered for the definition of the posterior distribution. However in real life applications the hyperparameters vector $\boldsymbol{\lambda}$ will not be known, and hence a second stage distribution $h(\boldsymbol{\lambda})$, also known as *hyperprior*, will be needed to provide the specification of the model. In this scenario the model formulation expressed in (2.14) becomes:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int_H f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\lambda}{\iint_{H,\Theta} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\lambda\, d\theta} \tag{2.15}$$

where $H$ is the space in which the hyperparameters $\boldsymbol{\lambda}$ take values. Note that in (2.15) an implicit *hierarchical structure* is adopted, indeed three different levels of distributional specifications can be identified. Typically, when using this specific approach, the primary interest is focused on the parameters $\boldsymbol{\theta}$ level. The name hierarchical models refers exactly to this specific structure which deals with multiple levels of "nested" probabilistic distributions. Another popular approach is the *empirical Bayes analysis* which assumes, as a value for $\boldsymbol{\lambda}$, an estimate $\hat{\boldsymbol{\lambda}}$ typically obtained by maximizing the marginal distribution $p(\mathbf{y}|\boldsymbol{\theta}) = \int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(d\boldsymbol{\theta}|\boldsymbol{\lambda})$ which is treated as a function of $\boldsymbol{\lambda}$. Posterior inference is then carried on working with the *estimated posterior distribution* $p(\boldsymbol{\theta}|\mathbf{y}, \hat{\boldsymbol{\lambda}})$ obtained by replacing $\boldsymbol{\lambda}$ with $\hat{\boldsymbol{\lambda}}$ into the equation (2.14). Of course, the Bayesian approach has negative aspects that must be mentioned. One of the most problematic aspects is indeed that the integration required in (2.14) (and hence also in (2.15)) are generally not available in an analytical closed form, and thus the posterior must be numerically approximated. The most popular approach, so far, is to build a Markov chain with state space $\Theta$, whose invariant distribution is the posterior $\pi(\theta|\mathbf{y})$. There exist specific software probabilistic programming languages, that is able to build such Markov chain, using few line of code (see Beraha et al. (2021)). This software allows for the application of Markov Chain Monte Carlo (MCMC) integration methods that will be explained in Section 2.3.

## 2.3. MCMC Sampling

As mentioned before, a very popular and efficient method to sample from the target posterior distribution is the Markov Chain Monte Carlo (MCMC) method. This is a method for exploring a distribution $p$ relying on the construction of a Markov chain having $p$ as its invariant distribution, sampling path averages of this MC to estimate the characteristics specifying $p$. In Bayesian statistics very often the posterior distribution is continuous, and hence MCMC samplers having a continuous state-space must be implemented. This topic has been widely explored in the literature (see for instance Gilks et al. (1995), chapter 4) and still be a vivid evolving field in statistics, with experts always seeking for better performances in the practical numerical implementation of this class of methods. Here is reviewed basic material from Section 4.4 in Rosner et al. (2021). A Markov chain is defined assigning $\Theta$, the state space ($\theta \in \Theta$), and the transition mechanism, defining how to proceed sampling the next iterate given the current one. This mechanism is technically called a *transition kernel* and it is represented by the conditional law $r(\theta^{(t+1)}|\theta^{(t)})$ which determines the stochastic path of the Markov chain. The goal of all this procedure is to be able to sample from the parameter posterior distribution $p(\theta|\mathbf{y})$, allowing to make inference about the unknown parameter. In this section the notation $|\mathbf{y}$ will be omitted, and the posterior or target distribution will be simply denoted by $p(\theta)$. According to well-known theory of general state-space Markov chains, under specific conditions, the distribution of the samples $\theta^{(t)}$ will converge to the posterior distribution of $\theta$ as $t$ grows ($t = 0, 1, 2, \dots$), regardless of the starting point of the chain (i.e. the choice of $\theta^{(0)}$ does not affect the result). The main condition to be fulfilled, ensuring the chain correctness, concerns the transition kernel which must satisfy the following assumption :

$$p(\theta) = \int_{\Theta} r(\theta|\theta^*)p(\theta^*)\, d\theta^* \tag{2.16}$$

When the conditional law $r(\cdot|\cdot)$ satisfies the former condition, then it is called *stationary transition kernel*. This is a necessary condition for the proper functioning of MCMC method, i.e. to obtain that the iterates will be eventually sampled from the posterior. In this scenario, the transition distributions satisfying (2.16) are referred to as *stationary transition kernels of the chain* while the target distribution $p(\cdot)$ is called *stationary distribution of the chain*. It is interesting to notice that there could exist multiple kernels $r(\cdot|\cdot)$ satisfying (2.16), which means that there may be different chains, generated by different kernels, leading to samples from the same target distribution $p(\cdot)$. Notice that if the first iterate is sampled from the stationary distribution, i.e. $\theta^{(0)} \sim p(\theta)$, then using the law of total probability it is easy to verify that also $\theta^{(1)} \sim p(\theta)$ and thus $\theta^{(t)} \sim p(\theta)$ for any

$t$. This is precisely the meaning of the name *stationary distribution* given to the target distribution $p(.)$. However, since it is unlikely to sample the initial iterate from the target distribution, this result is practically useless. In real applications the idea is to start from a reasonable initial iterate $\theta^{(0)}$ and then assuming that sooner or later a sample from $p(\theta)$ will be obtained. In this case also all the subsequent iterates will be sampled from $p(\theta)$. In practice a value for $\theta^{(0)}$ is selected, possibly by sampling from a prior distribution or by just selecting a reasonable value according to some prior knowledge about $\theta$. Then, starting from $\theta^{(0)}$, the Markov chain theory asserts that we will eventually sample from the target posterior distribution. The former results guarantees that, once fixed a sufficiently large number denoted as the *"burn-in"* $(BI)$ , if $t > BI$, then the corresponding iterates are such that $\theta^{(t)} \sim p(\theta)$ approximately. Moreover the same theory states also that, given any integrable function $g(.)$, the following holds:

$$\frac{\sum_{t=1}^{M} g(\theta^{(t)})}{M} \quad \rightarrow \int_{\Theta} g(\theta)p(\theta)\, d\theta \text{ as } M \rightarrow \infty \tag{2.17}$$

This means that the Markov chain Monte Carlo average above (left-hand term) will become almost surely arbitrarily close to the posterior mean of $g(\theta)$ as $M$ grows. This result is called the *Ergodic Theorem* and it is analogous to the law of large numbers (LLN) for independent identically distributed (iid) sequences. Note that the variance of the estimator in the left hand-side of (2.17) is given by a first term as in the case of iid sequence, plus a second term involving correlation between successive iterates of the Markov chain.

There are two more issues when dealing with this kind of methods. The first issue concerns the *convergence* of the chain, which can be trivially overcome by selecting a suitable large value for the burn-in $(BI)$. The second issue is about the total number of MCMC samples after the burn-in, i.e., to harvest $M - BI$, which indeed may need to be very large, because of the correlation between iterates that comes along with the Markov property of the chain.

The Markov chain theory allows the definition of many well-known sampling techniques, all relying on the MCMC method, such as Gibbs Sampler, Slice Sampling, Metropolis-Hastings Algorithm etc. For the purpose of this project, where the software STAN (see Stan Development Team (2022b)) is used, the theoretical method of major interest is the *Hamiltonian Markov Chain Monte Carlo Sampling*. As briefly summarized in Rosner et al. (2021) (Section 4.4.4), this method makes use of Hamiltonian dynamics from physics to generate new feasible candidate iterates that will then either be accepted or rejected according to the Metropolis acceptance probability. Indeed, following the *Metropolis-Hastings Algorithm* approach, a new candidate $\theta$ can be accepted or rejected as the next iterate value in the Markov chain stochastic path. The acceptance probability with which

$\theta^{(t+1)} = \theta$ is accepted is given by:

$$\alpha(\theta, \theta^{(t)}) = min\left\{1,\ \frac{p(\theta)q(\theta^{(t)};\theta)}{p(\theta^{(t)})q(\theta;\theta^{(t)})}\right\} \tag{2.18}$$

otherwise, if the candidate value is rejected, the chain does not advance and the path remains in the current state, i.e. $\theta^{(t+1)} = \theta^{(t)}$. Moreover, since in the Metropolis- Hastings framework $q(\cdot;\cdot)$ is required to be symmetric, the acceptance probability formula in (2.18) is simplified as :

$$\alpha(\theta, \theta^{(t)}) = min\left\{1,\ \frac{p(\theta)}{p(\theta^{(t)})}\right\} \tag{2.19}$$

Compared with the classical Metropolis-Hastings Algorithm, the Hamiltonian MCMC Sampling (HMCMC) enables a more efficient exploration of parameters by expanding the parameter space $\Theta$. This procedure is involved in many applications, resulting very helpful when dealing with high-dimensional spaces. Then, in this expanded space, the HMCMC method exploits Hamiltonian dynamics to "generate" a deterministic candidate for the next iteration, instead of relying on the usual random-walk Metropolis-Hastings method. As reported in Rosner et al. (2021), it has been argued in the literature that the Hamiltonian MCMC method, if properly optimized, can be much more efficient than its more classical predecessors. The Hamiltonian MCMC is of massive importance for this thesis, since all the numerical simulations have been carried on using STAN, which is a state-of-the-art platform for statistical modeling and high-performance statistical computation. In the STAN environment, full Bayesian statistical inference can be performed with MCMC sampling relying on Hamiltonian Markov Chains. As reported in the *Stan Reference Manual* (see Stan Development Team (2022c)), the MCMC sampling is based on simulating the Hamiltonian of a particle having initial position equal to the current parameter values and an initial momentum (kinetic energy) generated randomly. In the usual approach to implement Hamiltonian Markov chains, the Hamiltonian dynamics of the particle is simulated using the leapfrog integrator, which discretizes the smooth path of the particle into a number of small time steps denoted as leapfrog steps. A standard Metropolis accept/reject step is then required to retain the detailed balance condition, implying the posterior distribution invariant property, and to ensure that draws are marginally distributed according to the desired target distribution. This Metropolis adjustment is based on comparing log probabilities, here defined by the Hamiltonian, which is the sum of the potential (negative log probability) and kinetic (squared momentum) energies. In theory, the Hamiltonian is invariant over the path of the particle and rejection should never occur. In practice, the probability of rejection is determined by the accuracy of the leapfrog approximation to the true trajectory of the parameters. A

proper balance between effort and rejection rate is required in order to obtain a well performing algorithm. If the proposal is accepted, the parameters are updated to their new values. Otherwise, the sample is the current set of parameter values. All the techniques explained in details in this section will be needed to properly understand what will come next, becoming the theoretical foundation supporting all the findings reported in chapter 3. For more detailed insights about Hamiltonian MCMC tuning, see Hoffman & Gelman (2011).

## 2.4. The Spatio-temporal Model for $PM_{10}$ in Emilia-Romagna

Going back to the main goal of this study, the idea is to apply the Bayesian approach to the general spatial model defined in (2.13), assigning the prior of parameters vector $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma^2, \tau^2, \phi]$. In this way the hierarchical spatio-temporal model is defined. The temporal aspect of the phenomenon, which has been neglected till now, will be included in the model in the form of a time-depending function defining the general trend of data, i.e. the mean of $Y(s,t)$. From Figure 1.2 the general data trend shows a sort of U-shaped annual pattern and this seasonal behaviour can be expressed through a suitable function which will surely include an annual periodicity. This peculiar function of time $f(t)$ will be better specified in Section 2.4 and will take as argument the time $t$, i.e. the day of the year in which the pollutant registration took place. Summing up, the final model will follow the general Gaussian spatial process formulation proposed in (2.13), along with its spatial correction defined as (2.8). Moreover the expected value of the Gaussian process (GP) will include both the time-dependent trend of data, i.e. $f(t)$, and the station-specific characterizations, embedded into the model as numerical and categorical covariates. Referring to classical statistical matters, this model could be fitted into the *Linear Model* (LM) framework, which is well defined in Section 7.1 of Rosner et al. (2021). The mean of the process will be the sum of the time-dependent component with a linear combination of the covariates (linear w.r.t. the covariates coefficients), which corresponding parameters will then be modeled following the Bayesian approach. Also the time-dependent function will be specifically formulated in order to be linear in the coefficients, while the seasonality trend will be expressed by proper functions of time that will be used as Fourier basis for the definition of $f(t)$.

Going into the specifics of the $PM_{10}$ spatio-temporal model and considering only observations collected in Emilia-Romagna during 2018 (see Section 1.3), all the theoretical basis developed in the previous sections will be exploited to formulate a proper specific model.

Let $Y_i(t)$ be the response of the model, denoting the log-concentration of $PM_{10}$ at day $t$ recorded by station $i$ at location $\mathbf{s}_i$. For the dataset concerning Emilia Romagna, we will have $i = 1, \dots, 49$, i.e. a total of 49 recording stations, and $t = 0, \dots, T$, where $T = 364$. So the vector $\mathbf{Y} = [Y_1, \dots, Y_n]$ of all responses will be $n = 49 \cdot 365 = 17885$ dimensional. Each element of this vector will be indicated with the notation $Y_i(t)$ where $i \in \{1, \dots, 49\}$ and $t \in \{0, \dots, 364\}$, specifying the station and the day of the log-concentration sensing. Each observed log-concentration of pollutant can then be seen as the product of the combination of multiple factors, such as the zone characterization, the time of the year etc. as widely explained in Section 1.3. The following model is proposed:

$$Y_i(t) = \begin{cases} f_R(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w(\mathbf{s}_i) + \epsilon(t, \mathbf{s}_i) & \text{if } area(i) = rural \\ f_{NR}(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w(\mathbf{s}_i) + \epsilon(t, \mathbf{s}_i) & \text{if } area(i) \neq rural \end{cases} \quad \text{for } t = 0, 1, \dots, 364. \tag{2.20}$$

where $area(i)$ returns the area of station $i$, taking values in $\{rural,\ suburban,\ urban\}$, while $w(\mathbf{s}_i)$ indicates the *spatial residual* concerning station $i$, which depends upon the spatial coordinates of the recording site and it is incorporated into the model as specified in (2.8). The vector of covariates $\mathbf{x}_i$ has dimensionality equal to three, including two dummy variables expressing whether the station is in a traffic or industrial zone, and a continuous variable containing the standardized value of the station altitude. The vector of coefficients $\boldsymbol{\beta}$ has dimension three as well, i.e $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$. The non-spatial residual $\epsilon(t, \mathbf{s}_i)$ represents the classical residual term of the linear regression model and it contains different information with respect to $w(\mathbf{s}_i)$, as better explained in Section 2.1. Finally the two time-dependent functions $f_R(t)$ and $f_{NR}(t)$ will be properly defined in order to follow as close as possible the trend of the data. As can be seen in Figure 1.15 there is a graphical evidence to assume that two different function should be defined in order to successfully describe the real data $y(\mathbf{s}_i, t)$.

### Time-Dependent Component: Specifying $f_R(t)$ and $f_{NR}(t)$

Multiple options have been considered for the definition of $f_R(t)$ and $f_{NR}(t)$. The $PM_{10}$ data are time-series i.e. , time is discrete. However, language and tools from functional data analysis will be borrowed, expanding the two functions in the Fourier basis. Indeed when dealing with functional data, the first step is to choose a proper smoothing technique providing a set of smooth functional data to work with. In order to smooth the data, a basis needs to be specified. The chosen basis is a set of functions defining the functional object, being such that each functional object can be expressed as a linear weighted

combination of the elements belonging to the basis. In this scenario the functional data can be approximated through a multitude of different methods, such as polynomial smoothing, splines or natural splines approaches, Fourier basis smoothing and so on and so forth. In the specific case of the $PM_{10}$ time-series, since the data outline shows a strong seasonal component, a Fourier basis smoothing approach seems to be the most suitable one (see Essomba (2017)). In the Fourier basis setting, the functions composing the basis are all trigonometric functions, which indeed are intrinsically periodic depending on the specified *angular frequency* $\omega$ describing the phenomenon. In the specific case study of $PM_{10}$ concentrations data, a suitable value for the angular frequency is $\omega = \frac{2\pi}{365}$ expressing an annual periodic behaviour (period $T = 365$). Given this frequency and adopting a Fourier basis smoothing approach, a general formulation for the time dependent functions is reported as follows:

$$f(t) = \sum_{k=1}^{K} (a_k sin(k\omega t) + b_k cos(k\omega t)) + c, \tag{2.21}$$

where $K$ is a positive fixed integer. This formulation of the problem approximates the functional data via a Fourier basis composed of $2K$ sine and cosine functional elements, plus an additional constant term $c$. Finally the $2K+1$ coefficients controlling the function trend will be assumed to differ in the two cases of rural and non-rural areas, while the Fourier basis will remain the same for both functions of time.

At this point one last important issue remains to be solved, that is the choice of a proper number of basis for a good approximation. This topic is quite delicate, since assuming a large number of basis the resulting function will follow too precisely the data trend and the model will suffer from overfitting problems, while assuming a small number of basis functions, the function $f(t)$ will lead to a really poor fitting model, which will be useless for prediction. The final form of (2.21) has been chosen after many attempts. In the decision making process multiple factors have been considered; first of all, in order to obtain a well-defined function, the number of bases must be strictly smaller than the number of time instant, hence $2K < 365$ necessarily. Secondly, $K$ was fixed to some specific values in order to obtain precisely a weekly, bi-weekly, monthly, bi-monthly or semi-annual periodic behaviour. The combination of all these multiple frequencies, or maybe only a restricted subset of them, will probably lead to a well performing outcome since they are constructed to catch a real-life observed phenomenon that is suitable to follow such periodicity.

After having considered many feasible combinations of the above mentioned time frame periods, the most suitable option appeared to be the one considering only the *annual* and

*quarterly* frequency for the trigonometric bases, besides the additive constant term. The final formulation of the time-depending function becomes then:

$$f(t) = a_1 sin(\omega t) + b_1 cos(\omega t) + a_2 sin(4\omega t) + b_2 cos(4\omega t) + c \tag{2.22}$$

where the seasonal frequency is expressed by the term $4\omega = \frac{2\pi}{365/4}$ which contribute to the formula adding a term with a nearly 3-months periodicity, corresponding to the four season of the year (winter, spring, summer and autumn). The outcome obtained by fitting this model to the rural and non-rural data, where coefficients have been estimated through the Ordinary Least Squared (OLS) classical method, is provided below in Figure 2.1, while Figure 2.2 displays the corresponding basis functions. As already mentioned, the Fourier basis will remain the same for both the rural and non-rural cases, while the coefficients will be all assumed to be different in the two subsets of data, providing the following formulation for the time-dependent trend component :

$$f_R(t) = a_{R1} sin(\omega t) + b_{R1} cos(\omega t) + a_{R2} sin(4\omega t) + b_{R2} cos(4\omega t) + c_R \tag{2.23}$$

$$f_{NR}(t) = a_{NR1} sin(\omega t) + b_{NR1} cos(\omega t) + a_{NR2} sin(4\omega t) + b_{NR2} cos(4\omega t) + c_{NR} \tag{2.24}$$

which will lead to two different type of time trend, as shown in Figure 2.1.

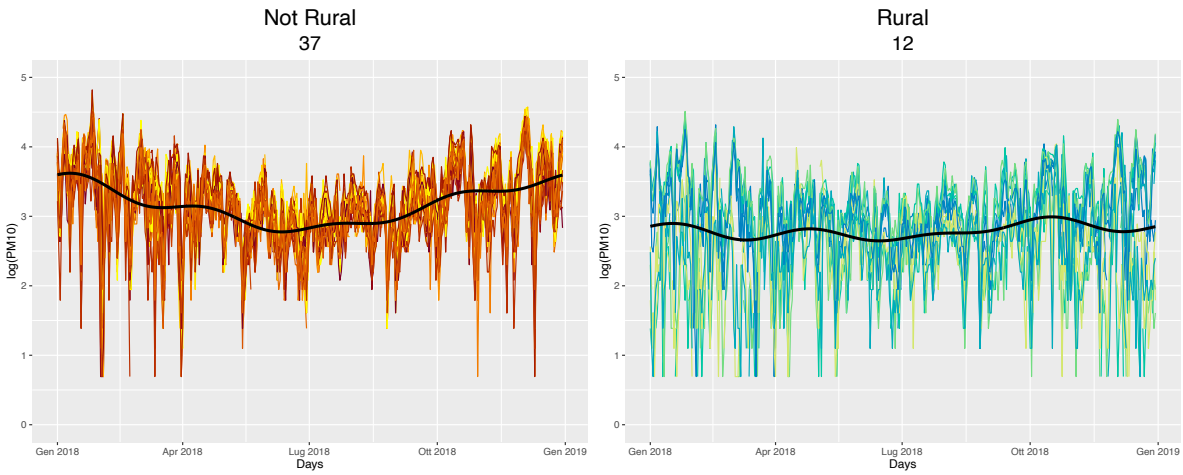



Figure 2.1: $PM_{10}$ log-concentrations average trend in rural (on the right) and non rural (on the left) areas, estimated by implying selected Fourier basis.

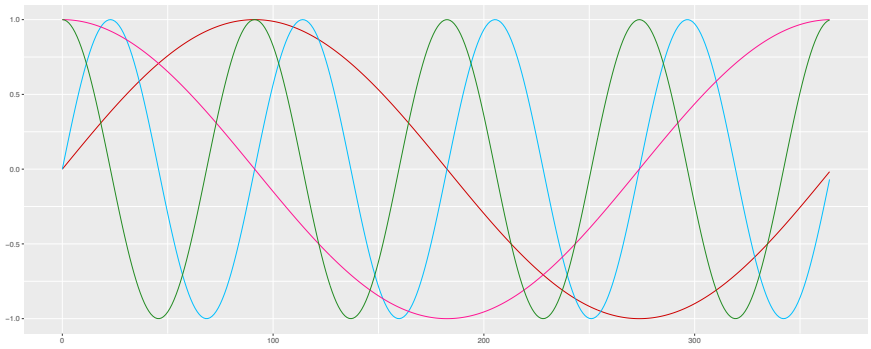



Figure 2.2: Annual and quarterly basis functions outline.

## The Bayesian Approach

Adopting a Bayesian approach, not only the coefficients of the covariates (i.e. $\boldsymbol{\beta}$) and the parameters defining the spatial process $w(\mathbf{s})$ will be treated as unknowns to be modeled, but also the coefficients associated to the Fourier basis elements will be assumed to be unknown parameters. Recalling Section 2.1, the spatial correlation terms will be modeled as a Gaussian process having exponentiated quadratic function kernel (2.7), while the non-spatial residual term will behave as a zero average Gaussian process, providing only the nugget term $\sigma^2$ in the definition (2.7). Hence the unknown parameters to be modeled can be summarized as $\boldsymbol{\theta} = [\sigma^2, \mathbf{w}, \alpha, \rho, \boldsymbol{\beta}, a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R, a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR}]$. In this framework, as better explained in Section 2.2, by exploiting prior knowledge and experts opinion, the marginal prior distributions of parameters must be defined. The description of three different models for the $PM_{10}$ log-concentrations data will be reported here. The corresponding inference is reported in Chapter 3.

### 2.4.1. Model 1: No Spatial Correlation

A first model proposal is the one considering the time-dependent function $f(t)$, expressing the seasonal data trend, the continuous and categorical covariates and the general residual term $\epsilon$. Here the spatial correction $w(\mathbf{s})$ residual term is not included. This model can be seen as a sort of base model to be compared with the more complex ones that will follow, in order to verify if the increased complexity and the additional terms are actually providing an improvement in the model performances. Hence the $PM_{10}$ log-concentrations are modeled here as follows:

$$Y_{i,t} = \begin{cases} f_R(t) + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon(t, \mathbf{s}_i) & \text{if } area(i) = rural \\ f_{NR}(t) + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon(t, \mathbf{s}_i) & \text{if } area(i) \neq rural \end{cases} \tag{2.25}$$

where each of the two time-depending functions is defined through the five different coefficients reported in (2.23) and (2.24). Summing up, the unknown parameters are $\boldsymbol{\theta} = [\sigma^2, \boldsymbol{\beta}, a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R, a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR}]$, where $\sigma^2$ is the variance of the non-spatial residual term $\epsilon(t, \mathbf{s}_i)$, which in this scenario explains alone the whole data variability. The priors characterizing all the unknown parameters involved in this first model are represented by the stochastic distributions reported below.

$$Y_i(t) \overset{ind}{\sim} \mathcal{N}(\mu_i(t), \sigma^2) \quad i = 1, \dots, 49 \quad , \quad t = 0, \dots, 364 \tag{2.26}$$

$$\mu_i(t) = f_R(t)\mathbb{1}_{\{area(i)=rural\}} + f_{NR}(t)\mathbb{1}_{\{area(i)\neq rural\}} + \mathbf{x}_i^T\boldsymbol{\beta} \tag{2.27}$$

$$f_R(t) = a_{R1}sin(\omega t) + b_{R1}cos(\omega t) + a_{R2}sin(4\omega t) + b_{R2}cos(4\omega t) + c_R \tag{2.28}$$

$$f_{NR}(t) = a_{NR1}sin(\omega t) + b_{NR1}cos(\omega t) + a_{NR2}sin(4\omega t) + b_{NR2}cos(4\omega t) + c_{NR} \tag{2.29}$$

$$\sigma^2 \sim InvGamma(3, 2) \tag{2.30}$$

$$a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.31}$$

$$a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR} \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.32}$$

$$\beta_0, \beta_1, \beta_2 \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.33}$$

where all the initial numerical hyperparameters have been decided by looking at the standardized $PM_{10}$ log-concentrations. Indeed the $f(t)$ coefficients and the betas have as prior a zero-centered Gaussian distribution, while the prior of the variance $\sigma^2$ has unitary mean.

### 2.4.2. Model 2: Including Spatial Correlation

Relying on the base model presented in the previous Section 2.4.1, and coherently with the Gaussian spatial process formulation presented in details in Section 2.1, when including the spatial correction $w(\mathbf{s}_i)$ depending on *latitude* and *longitude* coordinates of the sensing stations (here referred as the recording sites $\mathbf{s}_i$ ) the updated model becomes :

$$Y_i(t) = \begin{cases} f_R(t) + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon(t, \mathbf{s}_i) + w(\mathbf{s}_i) & \text{if } area(i) = rural \\ f_{NR}(t) + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon(t, \mathbf{s}_i) + w(\mathbf{s}_i) & \text{if } area(i) \neq rural \end{cases} \tag{2.34}$$

In this scenario, for notation simplicity, the spatial residual $w(\mathbf{s}_i)$ associated to station $i$ is denoted by $w_i$. The updated vector of unknown parameters is then $\boldsymbol{\theta} = [\sigma^2, \mathbf{w}, \alpha, \rho, \boldsymbol{\beta}, a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R, a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR}]$. All the parameters that were already present in the base model are still be modeled as before, leading to the following prior distributions of the involved parameters:

$$Y_i(t) \overset{ind}{\sim} \mathcal{N}(\mu_i(t)\ , \sigma^2) \quad i = 1, \dots, 49\ , \quad t = 0, \dots, 364 \tag{2.35}$$

$$\mu_i(t) = f_R(t)\mathbb{1}_{\{area(i)=rural\}} + f_{NR}(t)\mathbb{1}_{\{area(i)\neq rural\}} + \mathbf{x}_i^T\boldsymbol{\beta} + w_i \tag{2.36}$$

$$f_R(t) = a_{R1}sin(\omega t) + b_{R1}cos(\omega t) + a_{R2}sin(4\omega t) + b_{R2}cos(4\omega t) + c_R \tag{2.37}$$

$$f_{NR}(t) = a_{NR1}sin(\omega t) + b_{NR1}cos(\omega t) + a_{NR2}sin(4\omega t) + b_{NR2}cos(4\omega t) + c_{NR} \tag{2.38}$$

$$\sigma^2 \sim InvGamma(3, 2) \tag{2.39}$$

$$a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.40}$$

$$a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR} \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.41}$$

$$\beta_0, \beta_1, \beta_2 \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.42}$$

In this case the mean $\mu_i(t)$ of the process also includes the spatial correction $w_i$, which obviously is uniquely defined accordingly to the sensing station $i$ under study. For this candidate model some further assumptions must be made, since also the spatial residual term needs a proper modelling. In order to provide a suitable formulation for this last term, a Gaussian spatial process similar to the one reported in equation (2.13) was chosen and further specified assuming as covariance function the exponential quadratic form stated in (2.7). This choice was leaded by the already implemented computational methods available in the software STAN (Stan Development Team (2022b)). Following the just mentioned theoretical guideline, the spatial corrective term incorporated into the $PM_{10}$ data modelling can be characterized by the following hyperparameters, along with

their associated marginal priors:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0},\ \Sigma) \tag{2.43}$$

$$\Sigma_{i,j} = C(\mathbf{s}|\alpha,\rho)_{i,j} = \alpha^2 exp\left(-\frac{1}{2\rho^2}||s_i - s_j||^2\right) \tag{2.44}$$

$$\alpha \sim \mathcal{N}(0.3,\ 0.1) \tag{2.45}$$

$$\rho \sim Beta(3, 10) \tag{2.46}$$

where $||s_i - s_j||$ means the *Euclidean distance* between the two spatial points, i.e., $||s_i - s_j||^2 = (|lat(s_i) - lat(s_j)|^2 + |long(s_i) - long(s_j)|^2)$ . Here the vector of spatial residual terms is composed of 49 unknown variables, one for each recording station, and has been modeled as a Gaussian process following what explained in Section 2.1. The priors characterizing the hyperparameters behaviour, i.e. the marginal distributions of $\alpha$ and $\rho$, have been defined with very specific distributions. Looking at inference reported in Chapter 3, this choice will lead the MCMC mechanism to have less freedom in the exploration of the parameters space. However, the spatial hyperparameter priors must be defined as precisely as possible, in order to avoid unidentifiability problems. Indeed, when dealing with spatial processes, some precautions must be taken to avoid the occurrence of spatial-specific problems, such as correlation issues between the two hyperparameters $(\alpha, \rho)$ and the unidentifiability problem. This last issue is related to the presence of multiple peaks in the posterior estimated distribution of one (or eventually both) hyperparameter, which of course represent a major problem for the posterior inference analysis of the model. Hence, aiming to avoid these unwanted situations, a common practice is to assume one of the two hyperparameters to be fixed to a suitable value, which would have been previously estimated using spatial geostatistical methods. Alternatively, as was made in the proposed model, it is possible to impose very specific prior distributions, providing a more controllable behavior. All the just mentioned considerations lean on the theoretical issues briefly debated in Section 6.1.1.1 of the book by Banerjee et al. (2015). Finally, as before, in the model formulation all the fixed numerical values have been chosen coherently to the data structure.

### 2.4.3. Model 3: Including Spatial Correlation and Month-specific Variance

From Section 1.3 it is clear that the common variance assumption is not true. In this last model the variance $\sigma^2$ related to the non-spatial residual term $\epsilon$, which actually represents the so called nugget effect (see Section 2.1), is assumed to be time-specific. More precisely, relying on the comments made in Section 1.3, a month-specific variability will be included into the model. A specific variance parameter will be introduced for each month of the year, leading to the inclusion of twelve different parameters $\sigma^2_1, \dots, \sigma^2_{12}$. These parameters will be assumed to be independent and identically distributed (i.i.d.), and they will be given an Inverse Gamma prior distribution. The main difference from the previous model is the obvious expansion of the parameters space dimensionality and the chance of (possibly) having a different range of variability concerning each month, as could be expected by looking at the monthly data distribution in Figure 1.17. All the marginal priors related to the time-dependent function, covariate coefficients and spatial residual term will remain unchanged. The month-specific variable defined above will be referred as dependent on a function of time $m(t)$ specifying the month of the year in which the $PM_{10}$ sensing took place, and thus taking values in $\{1, \dots, 12\}$. Hence in the following model specification, the variability of the Gaussian prior distribution of log-concentration data will be expressed by $\sigma^2_{m(t)}$. The latter model formulation can then be summarized as:

$$Y_i(t) \overset{ind}{\sim} \mathcal{N}(\mu_i(t)\ , \sigma^2_{m(t)}) \quad i = 1, \dots, 49\ , \quad t = 0, \dots, 364 \tag{2.47}$$

$$\mu_i(t) = f_R(t)\mathbb{1}_{\{area(i)=rural\}} + f_{NR}(t)\mathbb{1}_{\{area(i)\neq rural\}} + \mathbf{x}_i^T\boldsymbol{\beta} + w_i \tag{2.48}$$

$$f_R(t) = a_{R1}sin(\omega t) + b_{R1}cos(\omega t) + a_{R2}sin(4\omega t) + b_{R2}cos(4\omega t) + c_R \tag{2.49}$$

$$f_{NR}(t) = a_{NR1}sin(\omega t) + b_{NR1}cos(\omega t) + a_{NR2}sin(4\omega t) + b_{NR2}cos(4\omega t) + c_{NR} \tag{2.50}$$

$$\sigma^2_1, \dots, \sigma^2_{12} \overset{iid}{\sim} InvGamma(3, 2) \tag{2.51}$$

$$a_{R1}, b_{R1}, a_{R2}, b_{R2}, c_R \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.52}$$

$$a_{NR1}, b_{NR1}, a_{NR2}, b_{NR2}, c_{NR} \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.53}$$

$$\beta_0, \beta_1, \beta_2 \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.54}$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0},\ \Sigma) \tag{2.55}$$

$$\Sigma_{i,j} = \alpha^2 exp\left(-\frac{1}{2\rho^2}||s_i - s_j||^2\right) \tag{2.56}$$

$$\alpha \sim \mathcal{N}(0.3,\ 0.1) \tag{2.57}$$

$$\rho \sim Beta(3, 10) \tag{2.58}$$

### 2.4.4. Model Selection Criteria

The models will then be compared via predictive goodness of fit criteria in Section 3.4, after having described the posterior inference provided by each candidate model. Once found the best model from the predictive point of view, it will be used for the stations clustering, building on it to include a cluster mechanism as explained in Section 2.5.
The analysis carried out in Section 3.4 is based on two well established predictive model selection criteria, i.e. the *Widely Applicable Information Criterion* (WAIC) and the *Leave One Out Cross Validation* (LOO-CV) method. As specified in Vehtari et al. (2016), these two methods consist in estimating pointwise out-of-sample prediction accuracy from the fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values. The values assumed by WAIC and LOO are asymptotically equal, as explained by Watanabe (2010). In the Bayesian framework these tools are of massive importance, since they define Bayesian model selection methods exploiting the whole posterior distribution other than the point estimates, as properly defined in Yong (2018). However, these tools require to be computationally evaluated. Vehtari et al. (2016) defined an optimized method implemented in R (see `loo()` package, Vehtari et al. (2022)) which provides an efficient computation of LOO using Pareto-smoothed importance sampling (PSIS), a new procedure for regularizing importance weights. The former procedure has been adopted in the computation of the values reported in Table 3.1 in Chapter 3. It has been proven in literature (see Yong (2018)) that WAIC and LOO-CV criteria appear to be the best performing ones, compared other four popular goodness-of-fit (GOF) criteria, which are the *Likelihood Ratio Test* (LRT), *Akaike Information Criterion* (AIC), *Bayesian Information Criterion* (BIC), and *Deviance Information Criterion* (DIC). The WAIC for model $M_j$ is formally defined as:

$$WAIC_j = -2(LPPD_j) + 2\sum_{i=1}^{n} Var_{\theta_j|\mathbf{y}}[log(p_i(y_i|\theta_j, M_j))] \tag{2.59}$$

where $LPPD_j$ is called the *Log-Pointwise Predictive Density* of the model indexed by $j$ (i.e. $M_j$) and its theoretical definition is provided below, together with its approximation in the MCMC framework:

$$LPPD_j = \sum_{i=1}^{n} log(m(y_i|\mathbf{y}, M_j)) \tag{2.60}$$

$$\text{MCMC computed } LPPD_j = \sum_{i=1}^{n} log\left(\frac{1}{M}\sum_{k=1}^{M} p_i(y_i|\theta_j^{(k)}, M_j)\right) \tag{2.61}$$

where $m(.|., M_j)$ represents the posterior marginal distribution under model $M_j$, while $\theta_j^{(k)}$ is the $k$-th MCMC sample from the parameter posterior distribution. The LOO-CV method relies instead on the *Estimated Log-pointwise Predictive Density* (elpd), providing the following Bayesian LOO estimate of out-of-sample predictive fit:

$$elpd_{loo} = \sum_{i=1}^{n} log(p(y_i|\mathbf{y}_{-i})) \tag{2.62}$$

$$\text{where} \quad p(y_i|\mathbf{y}_{-i}) = \int_{\Theta} p(y_i|\theta)p(\theta|\mathbf{y}_{-i}). \tag{2.63}$$

The latter quantity $p(y_i|\mathbf{y}_{-i})$ represents the leave-one-out predictive density given the observed data without the $i$-th data point. The above quantity is computed in the MCMC framework through the computationally advantageous technique provided by Vehtari et al. (2016), and the corresponding selection criterion, which will be referred as PSIS-LOO, is easily computed by converting $elpd_{loo}$ to deviance scale, i.e :

$$\text{PSIS-LOO} = -2\ elpd_{loo} \tag{2.64}$$

Assuming the WAIC as discriminant criterion for the model selection, given that for $M$ large enough it is equivalent to LOO-CV, the best performing model will be the one providing the **lowest** value of WAIC (and hence also the **lowest** value of PSIS-LOO).

## 2.5. Mixture Models for Stations Clusterization

The goal of this section is to introduce a method for clustering the recording stations of Emilia-Romagna (or whichever region). The general idea is to group together the stations which show a similar behaviour in a functional sense, relying upon the model that will be selected in Section 3.4, among the three models described in Section 2.4. In order to achieve this clusterization, a suitable strategy is to assume different coefficients in $f(t)$ for each of the 49 station, and then group together the stations having coefficients that most likely follow the same posterior distribution. Indeed, as widely explained in Section 2.4, the general outline of $PM_{10}$ functional data is given by a time-dependent behaviour represented in the model by $f(t)$, while the other regressors and the spatial correction term only provide an upward or downward shift in the pollution level. Hence, looking for a clusterization based on the functional shape of each station, the main characters involved in this section will be the function of time $f(t)$ and its parameters. A suitable formulation for the above mentioned clustering procedure is the *Dirichlet Process Mixture Model* (DPMM) approach, which will be properly explained as follows.

First of all, as described in Rosner et al. (2021), Section 3.3.1, the idea of mixture models comes from the need of more flexible models, allowing to better describe the data through a proper mixture of parametric densities over a common domain. In general, data could be modeled as a finite mixture of $m$ parametric probability density functions (pdf), providing the following notation :

$$p(y|\pi_1,\dots,\pi_m,\theta_1,\dots,\theta_m) = \sum_{i=1}^{m} \pi_i p(y|\theta_i) \tag{2.65}$$

$$\text{where} \quad \pi_i > 0 \quad \forall i = 1,\dots,m \ , \text{ and } \quad \sum_{i=1}^{m} \pi_i = 1$$

The above formulation will require, of course, the definition of proper prior distributions for the parameters $\theta_i$ and the mixture weights $\pi_i$. Since the latter must sum to one, a natural choice for their distribution would be an *m-dimensional Dirichlet distribution* $\mathcal{D}_m(\alpha_1,\dots,\alpha_m)$, which is a multidimensional generalization of the Beta distribution (easily obtained by considering $m = 2$ in the above). Considering then the parameters $\theta_i$, a possible choice for their prior distribution could be $\theta_i \overset{iid}{\sim} P$ for a fixed known distribution $P$ defined on the parameters space $\Theta$. In the case under study, since the $f(t)$ coefficients have been modeled as Gaussian distributed in model (2.47)-(2.58), it can be considered a generalization of (2.47)-(2.58) and assume (2.65) with $p(y|\theta_i)$ the Gaussian density. From Rosner et al. (2021), Section 3.3.2, for $\theta = [m, \sigma^2]$ denoting the mean and variance of the

Gaussian distribution, the following model could be considered:

$$y_i|\theta_i \sim p(y|\theta_i), \;\; i = 1, 2, ... \tag{2.66}$$

$$\theta_i|P \overset{iid}{\sim} P(\cdot), \;\; i = 1, 2, ... \tag{2.67}$$

If the hyperparameters of $P$ are unknown, a prior need to be fixed for them. The specification of $P$ will determine how flexible the mixture density will be, making the prior distribution choice of massive importance.

Let's introduce the *Dirichlet Process*(DP) [Ferguson, 1973].

$$\theta_1, \theta_2, \dots |P_0 \overset{iid}{\sim} P_0 \tag{2.68}$$

$$V_1, V_2, \dots |\alpha \overset{iid}{\sim} Beta(1, \gamma) \tag{2.69}$$

where $P_0$ is referred to as the *base distribution*. Both sequences are made of iid random variables independent of each others, and the second sequence defines the random variates $V_1, V_2, \dots$ that will be used in the definition of the weights $\pi_1, \pi_2, \dots$. For any $A$, (measurable) subset of $\Theta$, the Dirichlet process is defined as follow:

$$P(A) \overset{d}{=} \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(A) \tag{2.70}$$

where $\overset{d}{=}$ represents the equality in distribution, while $\delta_{\theta_k}(A)$ is equal to 1 if $\theta_k \in A$ and 0 otherwise. The weights in the expression above are defined as:

$$\pi_1 = V_1, \;\; \pi_2 = V_2(1 - V_1), \dots, \pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j), \dots \tag{2.71}$$

It can be proven that $\sum_k \pi_k = 1$, as desired for the mixture definition. Moreover, since $V_k$ and $\theta_k$ are independent of each other, also the weights $\pi_k$ will be independent of $\theta_k$, providing $E[P(A)] = P_0(A)$. In this case the random distribution $P$ is said to be "centered" on the fixed distribution $P_0$ and it is actually modeled as a Dirichlet process, i.e $P \sim DP(\gamma, P_0)$. It is important to notice that the value of $\gamma$ determines how shrinked the distribution $P$ will be around the base distribution $P_0$; if $\gamma$ is very large, then $P$ will be very close to $P_0$. The formulation above is called the *Stick-Breaking* (SB) formulation of the Dirichlet process (see Sethuraman (1994)). Practically speaking, it consists of a randomly weighted average of random point masses, obtained by hypothetically 'breaking' a stick of length 1 into an infinite number of bits with random lengths $\pi_i$'s. Note that

$\pi_1 \geq \pi_2 \geq \pi_3 \geq \dots$ a.s. The target variable $Y$, that in the study case will be represented by the $f(t)$ coefficients, can be then modeled as:

$$p(y|\gamma, P) = \sum_{k=1}^{\infty} \pi_k p(y|\theta_k) \tag{2.72}$$

$$\pi_1 = V_1, \ \pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j) \quad \forall k \geq 2 \tag{2.73}$$

$$V_1, V_2, \dots | \gamma \overset{iid}{\sim} Beta(1, \gamma) \tag{2.74}$$

which is an infinite mixture of randomly weighted parametric densities, having parameters that are randomly selected from a random distribution $P$ centered on the base fixed distribution $P_0$. This model is called *Dirichlet Process Mixture Model* (DPMM). The general formulation provided in equation (2.72) involves an infinite series of terms, but in real-life applications finite mixtures are often used to achieve an approximation. In practice a truncation of the above sum is provided in order to define a finite mixture, and hence it will be necessary to select a proper number of terms to include in the mixture. In the literature can be found many valid proposals to properly define the cardinality of the finite mixture, see for example the approximation method by Ishwaran & Zarepour (2002) or the a priori truncation method proposed in Argiento et al. (2015). Mixture models provide a straightforward mathematical framework for model-based clustering. For instance, if data $Y_i$ are assumed to be i.i.d. from the DPMM in (2.72), an equivalent formulation is describes as follows.

Once provided the necessary theoretical background concerning DPMM, the clustering technique relying on this kind of model needs to be defined. This *Bayesian Non-Parametric* (BNP) clustering approach is well introduced and specified in the paper by Dahl (2006) and is based on the following general formulation:

$$Y_i|\theta_i \overset{ind}{\sim} p(y_i|\theta_i), \quad i = 1, \dots, n \tag{2.75}$$

$$\theta_1, \dots, \theta_n | P \overset{iid}{\sim} P \tag{2.76}$$

$$P \sim DP(\gamma, P_0) \tag{2.77}$$

where the likelihood $p(y|\theta)$ is, for instance, the Gaussian kernel, and $P$ is the Dirichlet Process centered in $P_0$ with dispersion parameter $\gamma$. Because of (2.70), if $\theta_1, \dots, \theta_n$ are a sample from a Dirichlet process as in (2.76), then $\theta_i = \theta_j$ with positive probability. In this case the parameter $\rho$ can be introduced, being the *random partition of the data label set*, i.e. a partition of $\{1, \dots, n\}$ , and will be used for the actual clusterization of

data. More explicitly $\rho = \{S_1, \dots, S_G\}$ where each subset $S_g$ contains all the data indexes allocated in the same group, according to the relation $Y_i \sim Y_j \iff \theta_i = \theta_j$ from (2.75)-(2.77). Hence the data are grouped according to their parameters distribution. Note that in general $G$, the number of clusters, is assumed to be random. The prior distribution on the partition $\rho$ is induced by (2.76)-(2.77), while its posterior given the observed data $y_1, \dots, y_n$ will determine the final desired clusterization. Indeed the data clusterization is defined by exploiting the posterior distribution of $\rho$. Taking for instance the *Binder's Loss Function* (BLF) the final grouping of data will be given by the value of $\rho$ obtained by minimizing the chosen loss function, which for every couple of $\rho$ and $x$ will assign the cost of estimating the "true" $\rho$ by the feasible partition $x$. Thus the estimated data labels partition will be given by:

$$\hat{\rho} = \arg\min_x E[B(\rho, x)|data] \tag{2.78}$$

where the Binder's loss function $B(.,.)$ returns a fixed cost $b$ when two objects are wrongly classified in the same cluster, and a fixed cost $a$ when instead they are wrongly classified in different clusters. The value of the above mentioned costs is fixed such that $\frac{b}{a+b} \in [0, 1]$, and typically this fraction is assumed to be equal to $\frac{1}{2}$. Another possible choice for the loss function is to consider the *Variation of Information* (VI) loss function, which is a more conservative method, usually providing a smaller number of non-empty clusters with respect to the ones provided by the BLF, see Wade & Ghahramani (2018) and Dahl et al. (2022). Finally, defining the *vector of allocation variables* as $\mathbf{q} = [q_1, \dots, q_n]$, where each $q_i$ takes value in $\{1, \dots, G\}$ denoting the cluster associated to each observation $i = 1, \dots, n$, is possible to obtain a posterior estimate of the desired data clusterization. Indeed the posterior of the random partition $\mathbf{q}$ can be approximated as:

$$\mathbb{P}(q_i = g|data) \simeq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{q_i^{(m)} = g\}}, \quad g = 1, \dots, G \quad i = 1, \dots, n \tag{2.79}$$

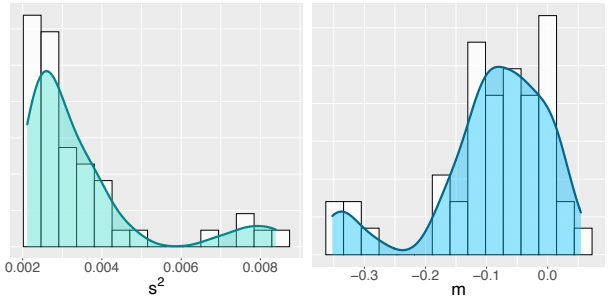
And the related posterior probability of two observations to be allocated together is coherently defined as:

$$\pi_{ij} = \mathbb{P}(c_i = c_j|data) \simeq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{c_i^{(m)} = c_j^{(m)}\}}, \quad i, j = 1, \dots, n, \quad i \neq j \tag{2.80}$$

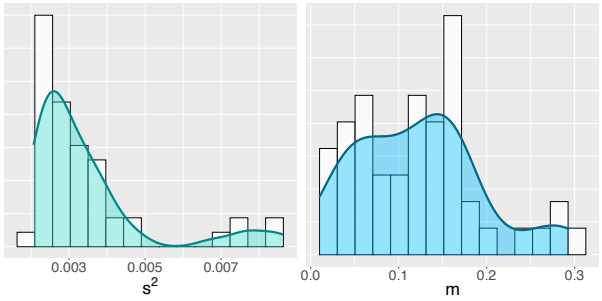
In both the formulations (2.79) and (2.80) the estimated quantities are obtained through the MCMC sampling techniques explained in Section 2.3, and hence the used notation is coherent with the one adopted in the previous section.

Provided the necessary theoretical background, the specific model-based clustering for $PM_{10}$ data will be discussed as follow. In the just presented BNP clustering scenario the base distribution $P_0$ needs to be specified. Since the clustering will be performed over $f(t)$ coefficients (modeled as Gaussian distributed in Sections 2.4.1, 2.4.2 and 2.4.3), the mixture will have a Gaussian kernel, providing as unknown parameter the vector $\theta = [m, s^2]$. Hence, for each one of the coefficients $(a_1, b_1, a_2, b_2, c)$ and for each recording station $i = 1, \dots, 49$, a different $\theta_i = [m_i, s_i^2]$ will be considered. In the further sections will be presented two different models for the clustering problem: at first a *Univariate* clusterization will be performed, grouping stations with respect to only one coefficient, then a multidimensional generalization of the model will be provided by a *Multivariate* clusterization, performed over all the coefficients. A prior for the base distribution $P_0$ will be needed, which in this specific case results in providing a prior for $m, s^2$. Shi et al. (2019) suggest that the prior should be specified with care, being vague enough to allow the chain correct functioning but avoiding no-information priors. In order to have an indication about the hyperprior to assign to the hyperparameters, frequentist estimates of the empirical distributions of $m$ and $\sigma^2$ have been computed. For each one of the 49 stations of interest, a point estimate of the $f(t)$ coefficients has been obtained through the classical statistical method of *Ordinary Least Squares* (OLS), together with its standard error. Figure 2.3 shows the empirical distributions of the estimates of the hyperparameters for every coefficient. For each coefficient, the 49 OLS estimated values have been used to define the empirical distribution of the mean, while the estimated squared standard errors have been used for outlining the variance empirical distribution.

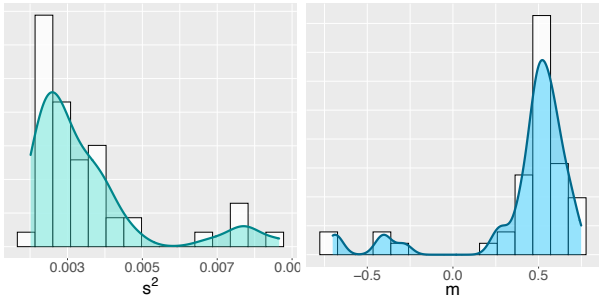
From Figure 2.3 it is clear that every coefficient present at least two peaks in the distribution of the mean term, moreover also in the estimated $s^2$ distributions multiple peaks can be identified. This peculiar distributional behaviour suggests the effective presence of at least two different groups of data (i.e. at least a two-element partition of the stations exists), and thus the ongoing clustering analysis actually make sense. Other important remarks concern the distributional domain; indeed looking at the variance empirical distribution, the curve seems to be shrinked around very low values tending to zero. This suggests that the estimated coefficients present a reduced variability among the 49 stations considered, and this information must be accounted in the definition of $s^2$ prior distribution. Looking then at the mean empirical distribution, it can be noticed that the domain clearly includes both positive and negative values, suggesting a zero-centered prior distribution for the hyperparameter $m$. Assuming a *Normal-InverseGamma* framework for $m, s^2$, suitable specifications of the distributions must be made in order to account for the former considerations.



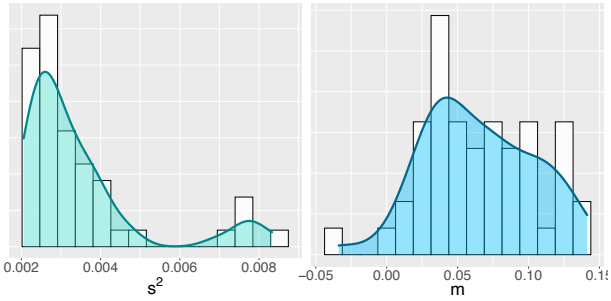

(a) Annual sine coefficient $a_1$



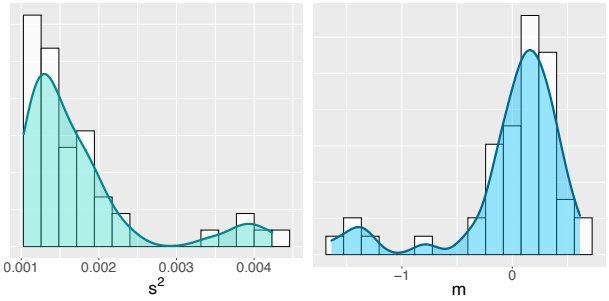

(b) Seasonal sine coefficient $a_2$




(c) Annual cosine coefficient $b_1$




(d) Seasonal cosine coefficient $b_2$




(e) Constant additive term $c$

Figure 2.3: Empirical distributions of the hyperparameters OLS estimates, provided for each one of the five $f(t)$ coefficients.

### 2.5.1. Univariate Clustering

A first cluster estimate of the stations has been obtained by considering only the *annual cosine coefficient* $b_1$ as a discriminating term. Hence the general framework of the model still be defined as in equations (2.47) - (2.58), with the same marginal priors for the covariates coefficients and the spatial correlation term. The only change regards the time-depending function. Indeed in this scenario there will be no distinction between rural and non-rural stations, assuming to have the same coefficients $a_1, a_2, b_2, c$ defining $f(t)$ for all the recording sites, exception made for the annual cosine term $b_1$ that instead will be different for each station and labeled as $b_i$; this last term will be modeled using the DPMM approach. Thus each recording station will be identified by a specific unique function $f_i(t)$ defined as:

$$f_i(t) = a_1 sin(\omega t) + b_i cos(\omega t) + a_2 sin(4\omega t) + b_2 cos(4\omega t) + c \quad i = 1, \dots, 49 \tag{2.81}$$

Summing up, the model proposed in Section 2.4.3 assumes the following form:

$$Y_i(t) \overset{ind}{\sim} \mathcal{N}(\mu_i(t)\ , \sigma^2_{m(t)}) \quad i = 1, \dots, 49\ , \quad t = 0, \dots, 364 \tag{2.82}$$

$$\mu_i(t) = f_i(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \tag{2.83}$$

$$f_i(t) = a_1 sin(\omega t) + b_i cos(\omega t) + a_2 sin(4\omega t) + b_2 cos(4\omega t) + c \tag{2.84}$$

$$\sigma^2_1, \dots, \sigma^2_{12} \overset{iid}{\sim} InvGamma(3, 2) \tag{2.85}$$

$$a_1, a_2, b_2, c \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.86}$$

$$\beta_0, \beta_1, \beta_2 \overset{iid}{\sim} \mathcal{N}(0, 1) \tag{2.87}$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0},\ \Sigma) \tag{2.88}$$

$$\Sigma_{i,j} = \alpha^2 exp\left(-\frac{1}{2\rho^2}||s_i - s_j||^2\right) \tag{2.89}$$

$$\alpha \sim \mathcal{N}(0.3,\ 0.1) \tag{2.90}$$

$$\rho \sim Beta(3, 10) \tag{2.91}$$

In addition the DPMM specifying the annual cosine term must be defined as a finite mixture, allowing to further perform the BNP clusterization over the stations. Having a total of 49 stations, and given the visual observations taken on Figure 2.3, a reasonable choice could be to assume a total of $G = 20$ elements to be included in the mixture. The station-specific coefficient is then modeled as:

$$b_i \overset{iid}{\sim} \sum_{g=1}^{G} \eta_g \mathcal{N}(m_g, s_g^2) \quad i = 1, \dots, 49 \ , \tag{2.92}$$

where the weights are defined through the stick-breaking technique as:

$$\eta_1 = v_1 \ , \quad \eta_g = v_g \prod_{j=1}^{g-1} (1 - v_j) \tag{2.93}$$

$$\sum_{g=1}^{G} \eta_g = 1 \tag{2.94}$$

$$v_g \overset{iid}{\sim} Beta(1, 2) \quad g = 1, \dots, G \tag{2.95}$$

where the Dirichlet process concentration parameter has been fixed to $\gamma = 2$. The last step is to properly specify the prior distributions of the mixture hyperparameters $m_g, s_g^2$. Accordingly to the solutions proposed by Shi et al. (2019) and considering also the observations regarding Figure 2.3, a possible feasible definition for the hyperparameters prior is:

$$m_g \overset{iid}{\sim} \mathcal{N}(0, 150 s_g^2) \quad g = 1, \dots, G \tag{2.96}$$

$$s_g^2 \overset{iid}{\sim} InvGamma(4.5, 0.015) \quad g = 1, \dots, G \tag{2.97}$$

For a better understanding of the former distributions outline, in Figure 2.4 is reported the overlapping of the chosen prior distributions (2.96), (2.97) over the frequentist estimate of the hyperparameters empirical distributions (see Figure 2.3c). The posterior inference of this univariate clustering model is reported in Section 3.5.

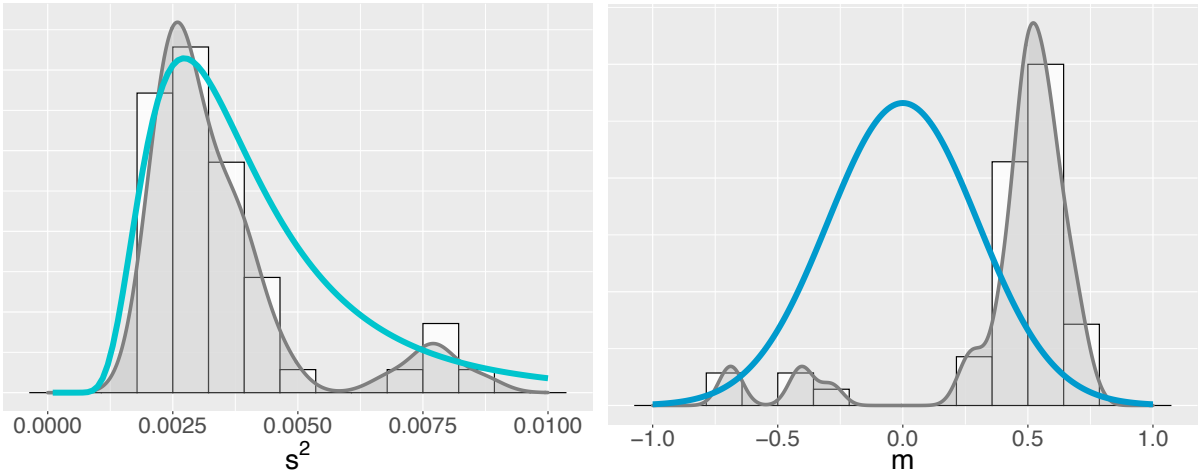



Figure 2.4: Comparison between the estimated hyperparameters empirical distributions and their prior distribution in the model-based univariate clustering.

### 2.5.2. Multivariate Clustering

In the multivariate clustering scenario, the clusterization is performed considering all the five coefficients defining $f(t)$ to be station-specific and hence modeled following a DPMM approach. In this setting, the function of time is defined for each station as:

$$f_i(t) = a_{1i}sin(\omega t) + b_{1i}cos(\omega t) + a_{2i}sin(4\omega t) + b_{2i}cos(4\omega t) + c_i \ , \quad i = 1, \dots, 49 \quad (2.98)$$

In this multivariate model-based clustering scenario, the model proposed in Section 2.4.3 assumes the following form:

$$Y_i(t) \overset{ind}{\sim} \mathcal{N}(\mu_i(t) \ , \sigma^2_{m(t)}) \quad i = 1, \dots, 49 \ , \quad t = 0, \dots, 364 \quad (2.99)$$

$$\mu_i(t) = f_i(t) + \mathbf{x}_i^T \boldsymbol{\beta} + w_i \quad (2.100)$$

$$f_i(t) = a_{1i}sin(\omega t) + b_{1i}cos(\omega t) + a_{2i}sin(4\omega t) + b_{2i}cos(4\omega t) + c_i \quad (2.101)$$

$$\sigma^2_1, \dots, \sigma^2_{12} \overset{iid}{\sim} InvGamma(3, 2) \quad (2.102)$$

$$\beta_0, \beta_1, \beta_2 \overset{iid}{\sim} \mathcal{N}(0, 1) \quad (2.103)$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \ \Sigma) \quad (2.104)$$

$$\Sigma_{i,j} = \alpha^2 exp\left(-\frac{1}{2\rho^2}||s_i - s_j||^2\right) \quad (2.105)$$

$$\alpha \sim \mathcal{N}(0.3, \ 0.1) \quad (2.106)$$

$$\rho \sim Beta(3, 10) \quad (2.107)$$

where the coefficients characterizing each of the 49 functions $f_i(t)$ are modeled as a 5-dimensional unknown vector through a multidimensional Gaussian DPMM approach. Also in this case the mixture is finite and truncated to a total of $G = 20$ terms, leading to the following definition:

$$[a_{1i}, b_{1i}, a_{2i}, b_{2i}, c_i] \overset{iid}{\sim} \sum_{g=1}^{G} \eta_g \mathcal{N}_5(\mathbf{m}_g, \Sigma_g) \quad i = 1, \dots, 49 \ , \quad (2.108)$$

where the weights $\eta_g$ are defined through the stick-breaking technique as in 2.93. As suggested in Shi et al. (2019), the covariance matrix $\Sigma_g$ has been assumed to be diagonal.

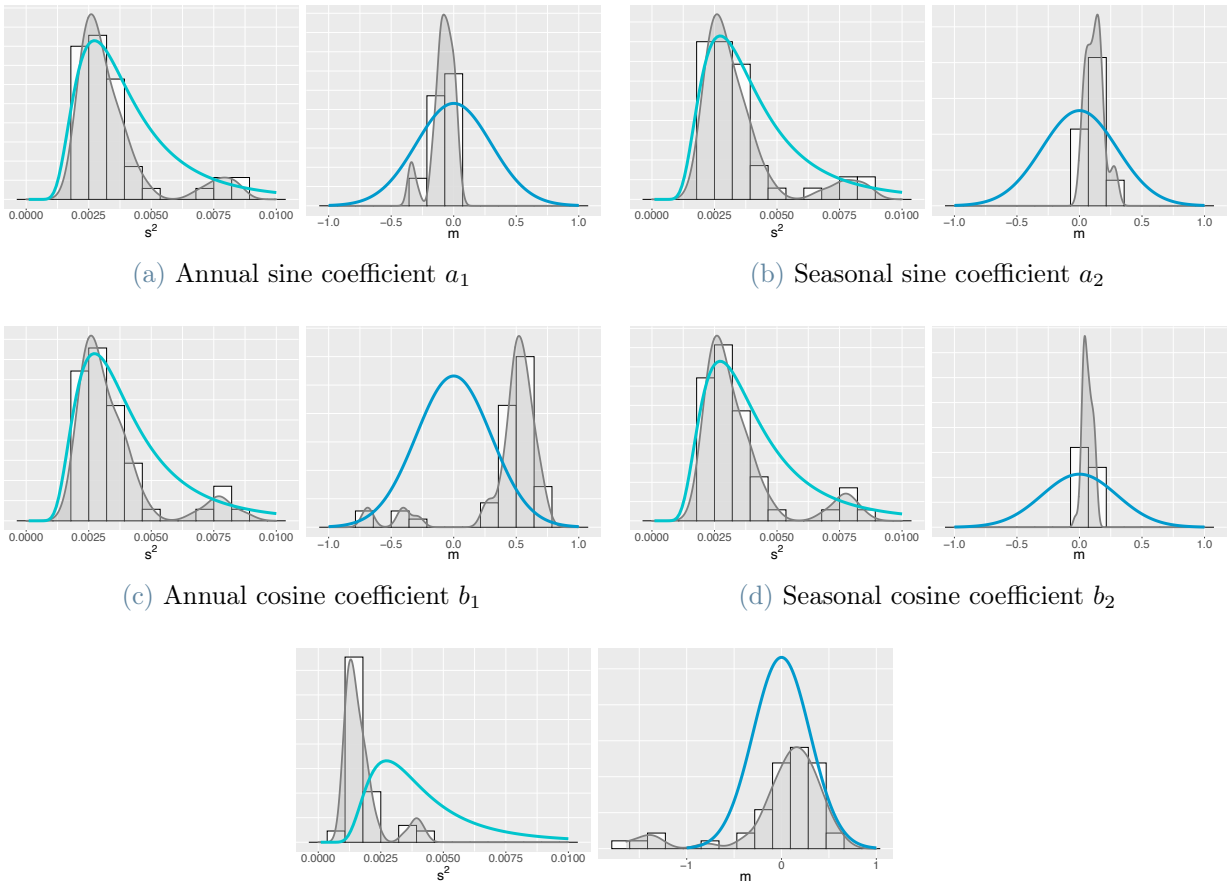
More specifically: $\Sigma_g = \begin{bmatrix} s^2_{g1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s^2_{g5} \end{bmatrix}$

Looking at the estimated empirical distributions reported in Figure 2.3, both hyperparameters show similar behaviours and a common domain for all the five coefficients. Thus the elements defining each of the two multidimensional hyperparameters, i.e. $\mathbf{m}$ and $\Sigma$, will be modeled as independent and identically distributed(iid). The hyperparameters are then modeled as:

$$\mathbf{m}_g \overset{iid}{\sim} \mathcal{N}_5(\mathbf{0}, 150\Sigma_g) \qquad g = 1, \dots, G \tag{2.109}$$

$$s^2_{g1}, \dots, s^2_{g5} \overset{iid}{\sim} InvGamma(4.5, 0.015) \qquad g = 1, \dots, G \tag{2.110}$$

As before, for the sake of clarity, in Figure 2.5 are reported five plots, one for each coefficient of $f(t)$. In each plot, for the two hyperparameters $m$ and $s^2$, the prior distributions outlines have been overlapped to the estimated empirical distributions.




(a) Annual sine coefficient $a_1$

(b) Seasonal sine coefficient $a_2$

(c) Annual cosine coefficient $b_1$

(d) Seasonal cosine coefficient $b_2$

(e) Constant additive term $c$

Figure 2.5: Comparison between the estimated hyperparameters empirical distributions and their prior distribution in the model-based multivariate clustering.

The posterior inference of this multivariate clustering model can be found in Section 3.6.

# 3 | Posterior Inference on $PM_{10}$ in Emilia Romagna

In this chapter we report numerical and graphical overview of the posterior inference of the three models presented in Section 2.4, as well as the clustering estimates associated to models in Section 2.5. Posterior analysis will be reported in Sections 3.1 , 3.2 and 3.3 for the three models defined in Sections 2.4.1 , 2.4.2 and 2.4.3. Section 3.4 reports goodness-of-fit (GOF) of the three models. The "best" model (according to WAIC and LOO), as generalized in Section 2.5, and the associated inference are presented in Sections 3.5 and 3.6 as far as clustering is concerned. All the estimates we report here are based on MCMC simulations with the software STAN, running two parallel chain and assuming 1000 iterations as burn-in ($BI = 1000$) and 1000 sampling iterations ($M = 1000$), unless otherwise specified.

## 3.1. Model 1 : No Spatial Correlation

In this section is presented the posterior inference for model (2.26)-(2.33).

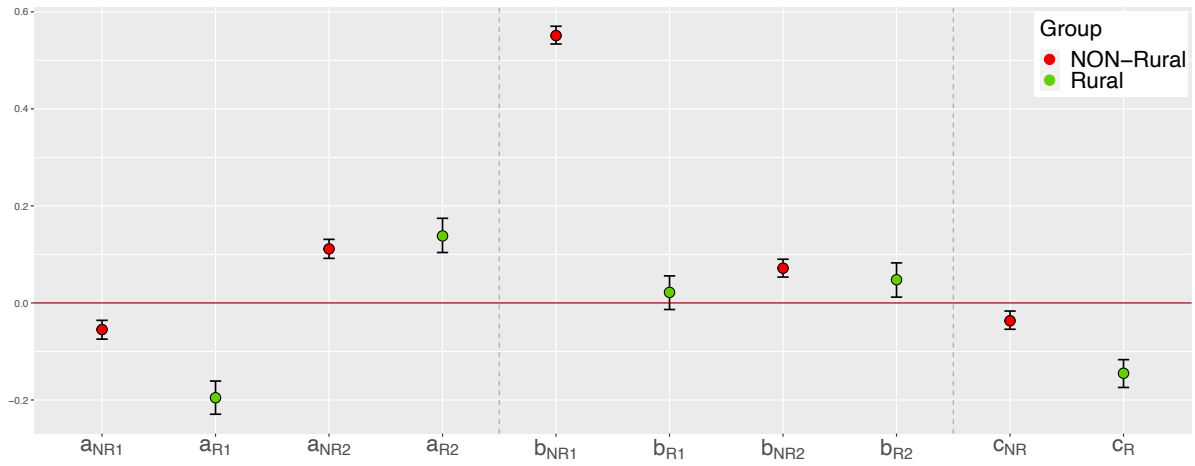



Figure 3.1: 95% marginal posterior Credibility Intervals (CI) of parameters $a_{NR1}, a_{R1}, a_{NR2}, a_{R2}, b_{NR1}, b_{R1}, b_{NR2}, b_{R2}, c_{NR}, c_R$ for Model 1.

Figure 3.1 shows the 95% marginal posterior credibility intervals for the coefficients of $f_R(t)$ and $f_{NR}(t)$. Note that all the coefficients vary in a very tight range of values.

Coherently, also the 95% posterior credibility bands for $f_R(t)$ and $f_{NR}(t)$ reported in Figure 3.2 are quite narrow. Indeed, both rural and non-rural cases show an estimated functional range shrinked around the median value (represented by the thicker lines in Figure 3.2).

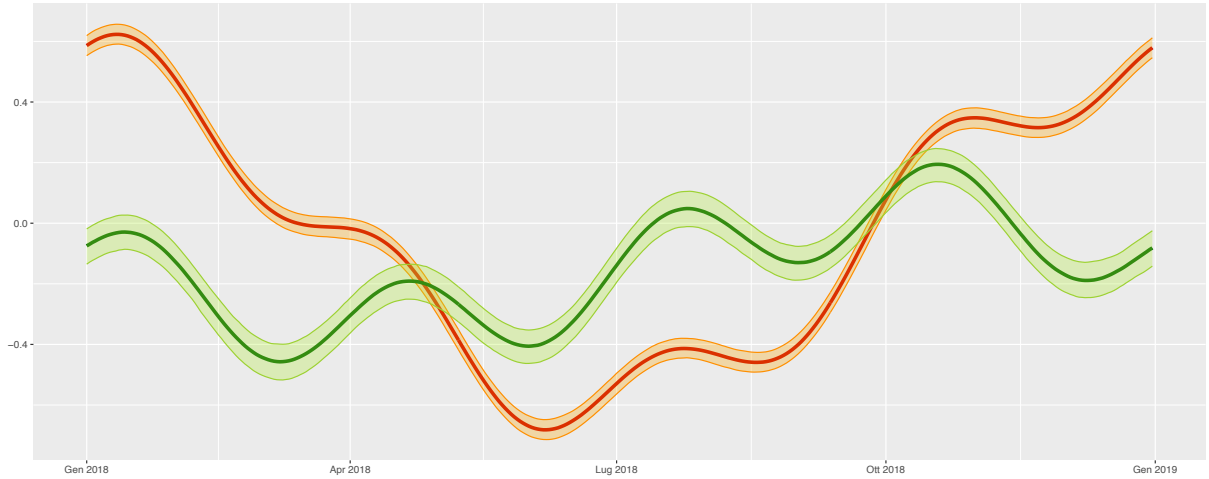



Figure 3.2: 95% posterior credibility bands for the time-dependent functions $f_R(t)$ (in green) and $f_{NR}(t)$ (in red) for Model 1.

The same argument applies also to the 95% marginal posterior credibility intervals of $\beta_0, \beta_1, \beta_2$, which are reported in Figure 3.3. Even in this case the credibility intervals are quite tight and shrinked around the posterior median (colored points in Figure 3.3 on the left). Notice that, as expected from Section 1.3, the altitude coefficient varies in a range of strictly negative values, denoting an inverse relation w.r.t. the air pollution level, while the coefficients indicating traffic and industrial zones assume strictly positive values, representing an increase in the pollutant concentrations.

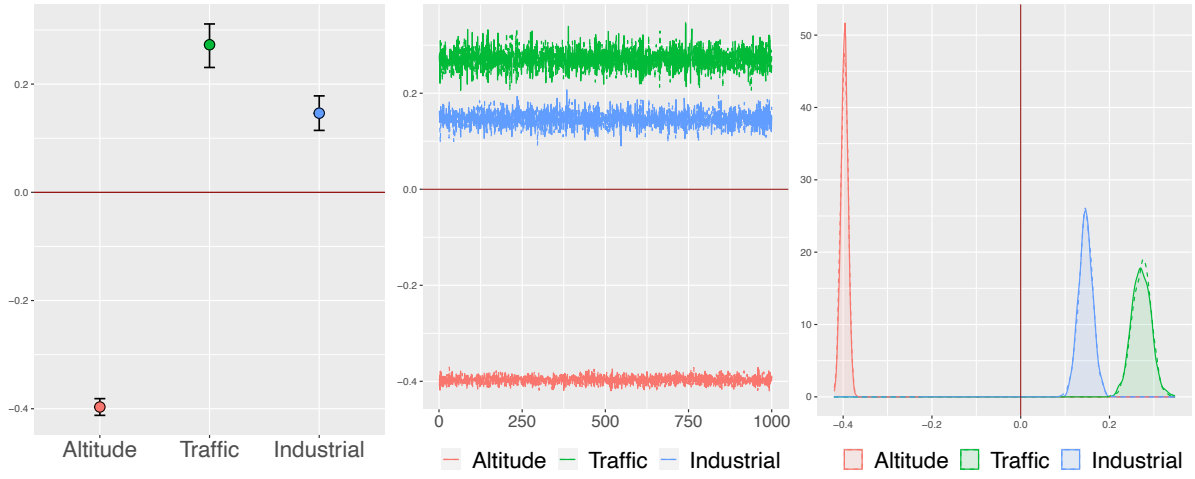



Figure 3.3: Posterior inference for the regression parameter $\boldsymbol{\beta}$: 95% marginal posterior credibility intervals (left), traceplots (center) and estimated posterior marginal densities (right).

## 3.2. Model 2 : Including Spatial Correlation

In this section is reported the posterior inference on the parameters of model (2.35)-(2.42).

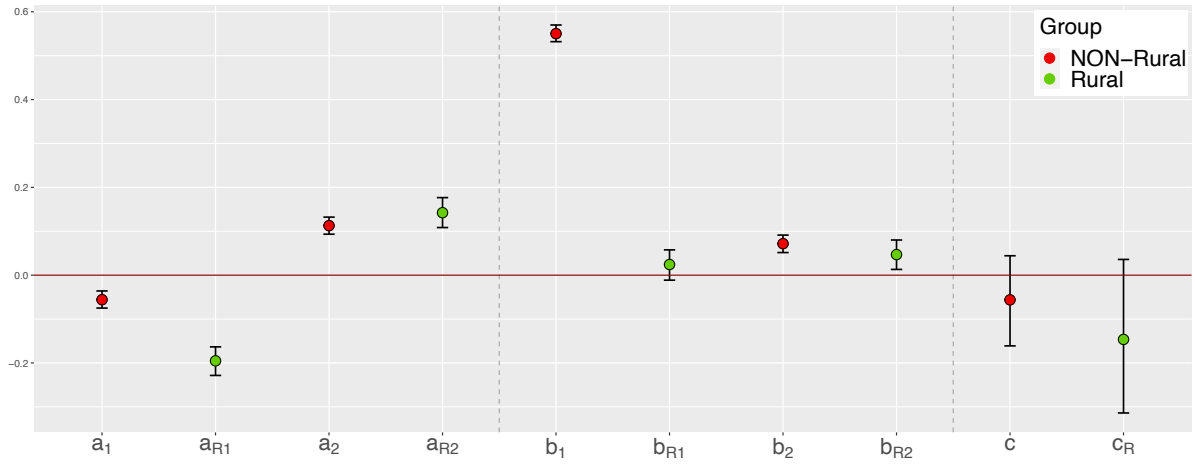



Figure 3.4: 95% marginal posterior Credibility Intervals (CI) of parameters $a_{NR1}, a_{R1}, a_{NR2}, a_{R2}, b_{NR1}, b_{R1}, b_{NR2}, b_{R2}, c_{NR}, c_R$ for Model 2.

We report here both the 95% marginal posterior credibility intervals for the time depending function coefficients in Figure 3.4 and the 95% posterior credibility bands for $f_R(t)$ and $f_{NR}(t)$. First of all, it can be noticed that, as expected, including the 49 unknown parameters for the spatial residual term $\mathbf{w}$ increased the estimates uncertainty. Indeed

both the posterior credibility bands in Figure 3.5 and the marginal posterior CIs of $f(t)$ coefficients in Figure 3.4 appear to be wider with respect to the ones reported in Figures 3.2 and 3.1 respectively. The increased variability is particularly evident when looking at the 95% marginal posterior CIs of the constant terms $c_{NR}$ and $c_R$.

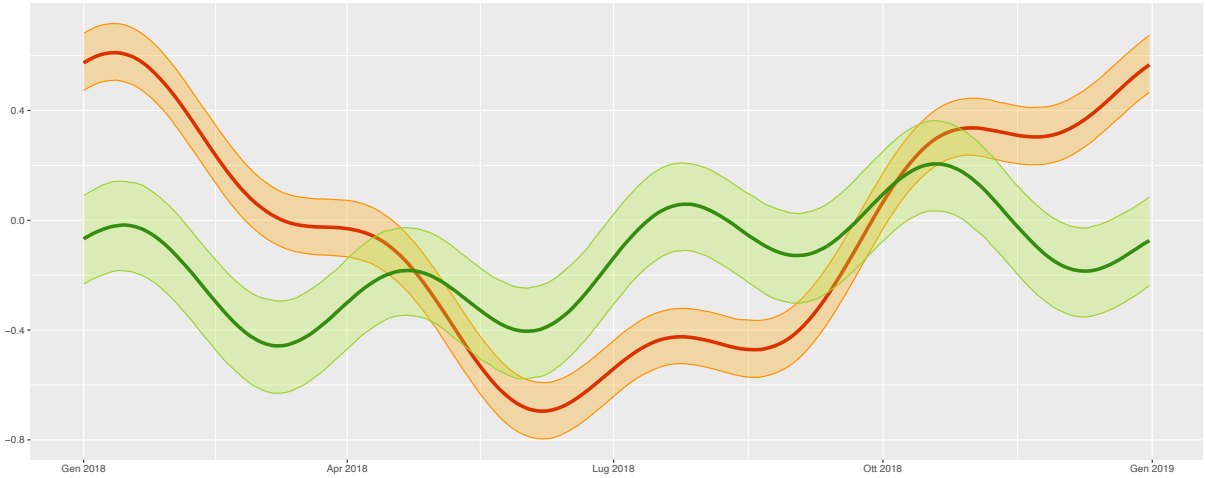



Figure 3.5: 95% posterior credibility bands for the time-dependent functions $f_R(t)$ (in green) and $f_{NR}(t)$ (in red) for Model 2.

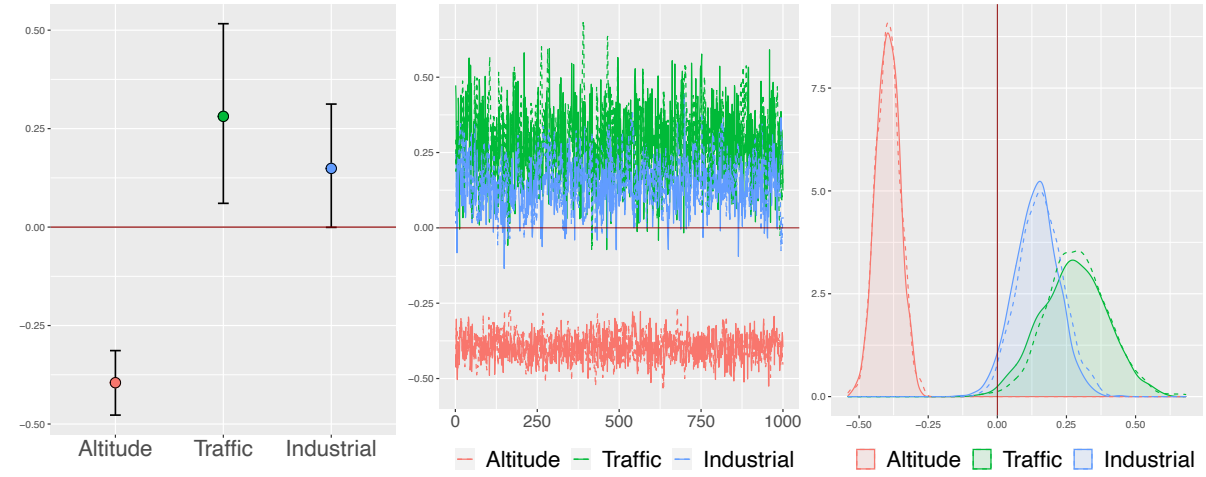



Figure 3.6: Posterior inference for the regression parameter $\boldsymbol{\beta}$: 95% marginal posterior credibility intervals (left), traceplots (center) and posterior marginal densities (right).

Figure 3.6 reports the posterior inference for the regression parameter $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$, providing the 95% marginal posterior CIs, together with their traceplots and posterior marginal densities. Also in this case, comparing the intervals width with the ones reported in Figure 3.3, it is possible to notice an increased variability. The comments taken about

the positive or negative values assumed by the three coefficients remain unchanged with respect to the previous section (Section 3.1).

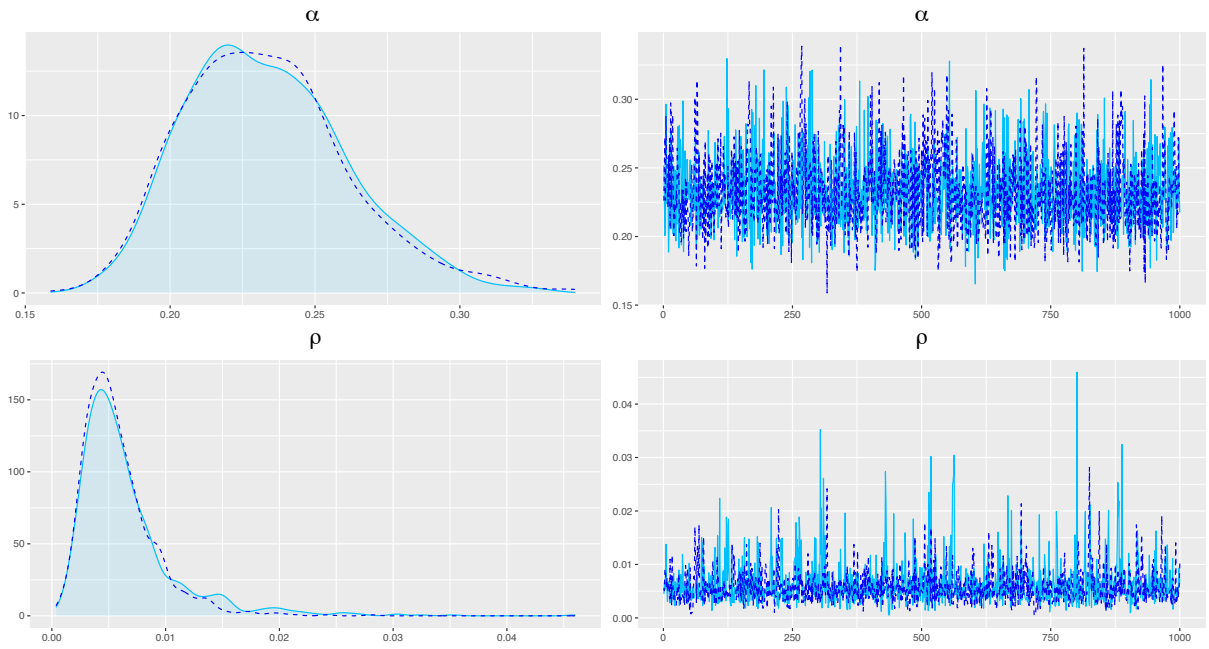
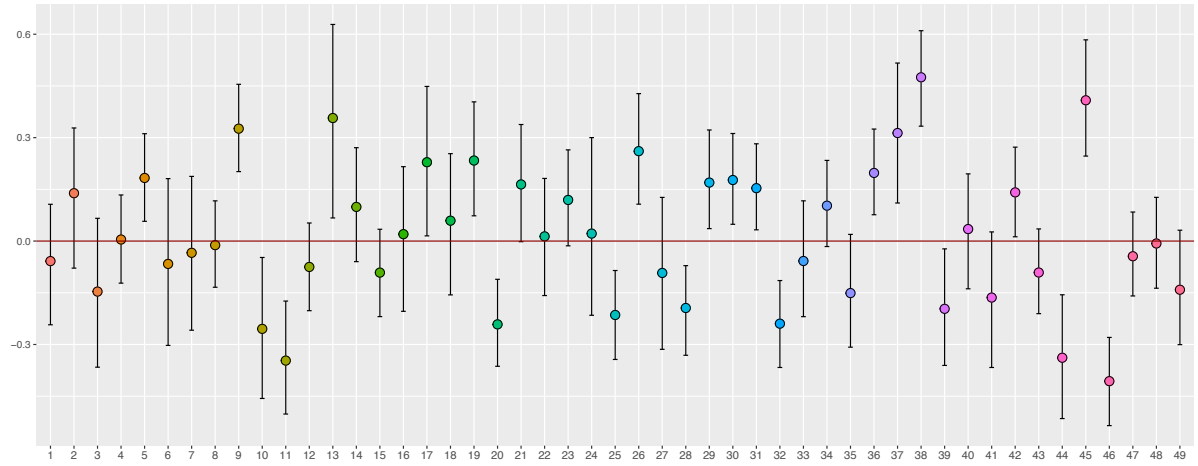
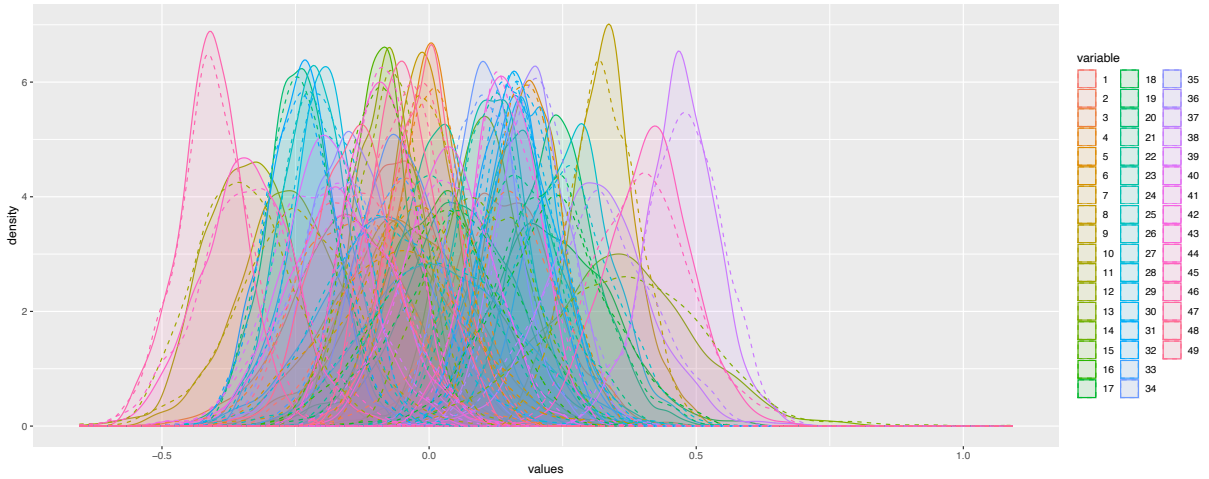



Figure 3.7: Posterior inference for the parameters characterizing the spatial residuals $\mathbf{w}$. Marginal posterior densities and traceplots of $\alpha$ (top row) and $\rho$ (bottom row).

Figure 3.7 reports the posterior inference for parameters $\alpha$ and $\rho$, which characterize the Gaussian process defining spatial residuals $\mathbf{w}$. Looking at $\alpha$ and $\rho$ marginal posterior densities, which are outlined on the left side of Figure 3.7, the choice of very specific prior distributions prevented the chain from non-identifiability problems. Indeed, neither of the two distributions shows the presence of multiple peaks. However, looking at the traceplots in Figure 3.7 on the right, it can be noticed that $\rho$ tends to take values really close to zero, even if it should be a strictly positive parameter. This tendency of $\rho$ to be distributed over very small values may represent a problem, since if it reaches the null value it would make the chosen kernel formulation (2.7) undefined. The former issue must be taken into account when dealing with $\rho$, being very careful in the choice of its prior distribution in order to avoid undesirable behaviours of the MCMC chain.




(a) 95% marginal posterior CIs of **w**




(b) Marginal posterior density distribution of **w**

Figure 3.8: Marginal posterior credibility intervals and densities of $w_1, \ldots, w_{49}$ (spatial residual terms) in model 2.

Dealing instead with the posterior inference for the spatial residual term **w** reported in Figure 3.8, some comments can be made. First of all, it can be noticed that even if the marginal posterior densities in Figure 3.8b appear to be distributed over small values, there is evidence to assume a spatial differentiation among the considered recording sites. Indeed, by looking also to the 95% marginal posterior CIs of **w** in Figure 3.8a, many credibility intervals are disjoint and the majority of them present different median values as well as different variability (enlightened by the different widths of the intervals).

## 3.3. Model 3 : Including Spatial Correlation and Month-Specific Variance

Here is reported posterior inference for parameters in model (2.47)-(2.58).

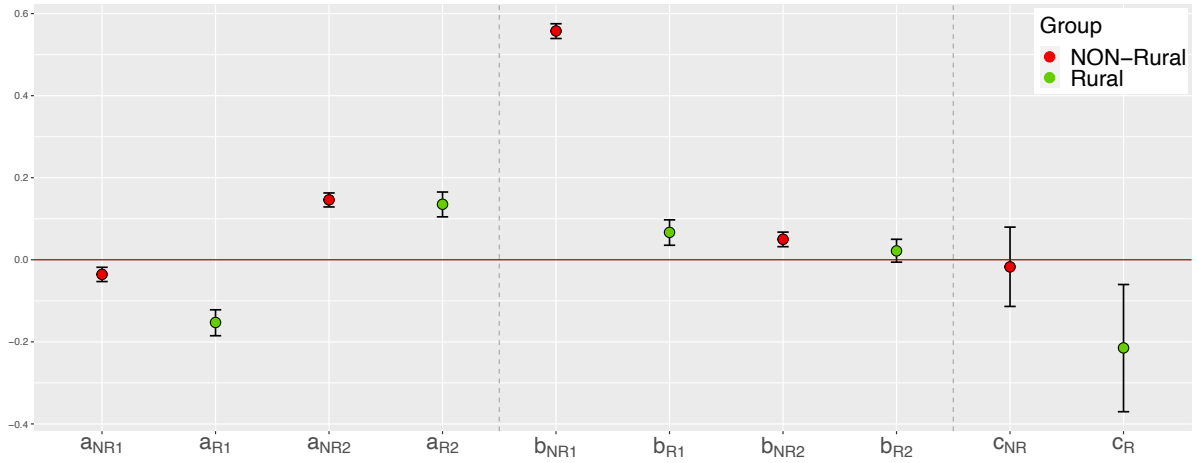



Figure 3.9: 95% marginal posterior Credibility Intervals (CI) of parameters $a_{NR1}, a_{R1}, a_{NR2}, a_{R2}, b_{NR1}, b_{R1}, b_{NR2}, b_{R2}, c_{NR}, c_R$ for Model 3.

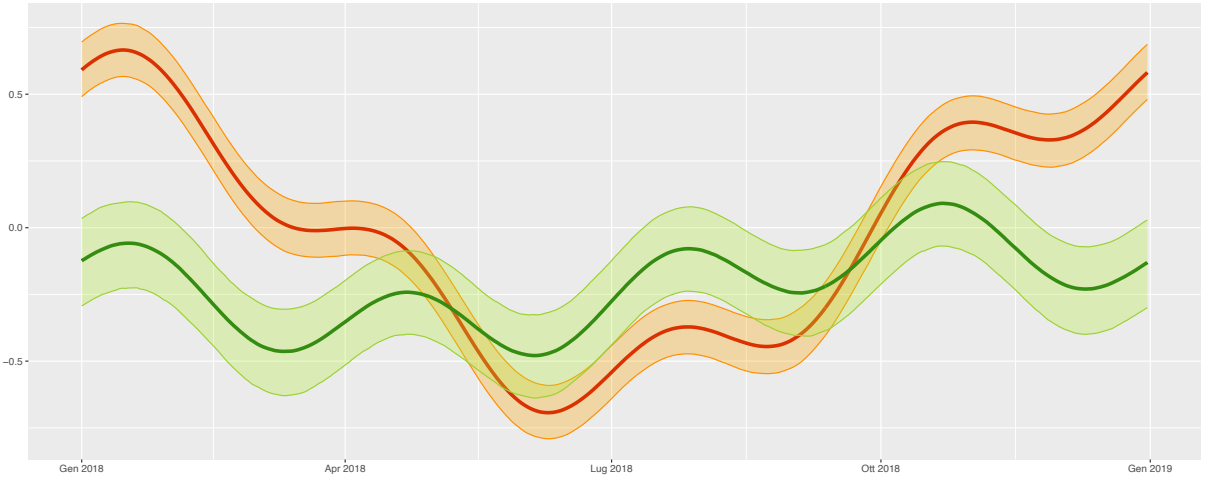



Figure 3.10: 95% posterior credibility bands for the time-dependent functions $f_R(t)$ (in green) and $f_{NR}(t)$ (in red) for Model 3.

Figure 3.9 shows 95% marginal posterior Credibility intervals of $f_R(t)$ and $f_{NR}(t)$ coefficients. It is clear that corresponding coefficients in the two scenarios are ranging over different values. This behavior is particularly evident when looking at $a_1$ and $b_1$, i.e. the annual sine and cosine coefficients, which present even disjoint marginal posterior CIs in

the rural and non-rural cases. These observations support the choice of assuming two different functions to describe the time-dependent trend in the two different areas. Also the 95% posterior credibility bands for the two functions, which are reported in Figure 3.10, highlight different behaviours for the two considered cases.

Figure 3.11 shows the posterior inference for parameters $\beta_0, \beta_1, \beta_2$. Looking at 95% marginal posterior CIs, traceplots and densities, comments similar to the ones in Section 3.2 can be made, since no particular changes can be identified.

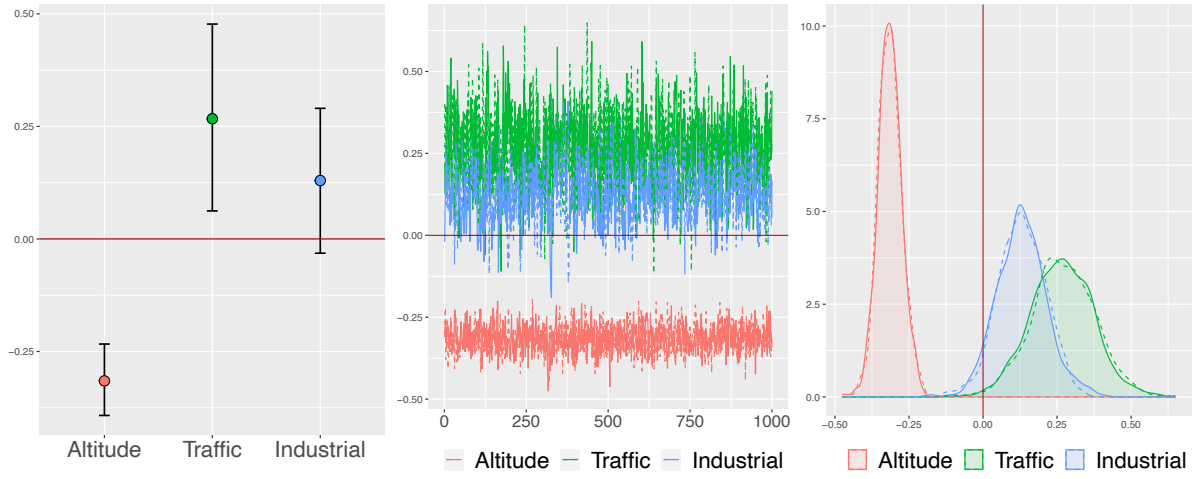



Figure 3.11: Posterior inference for the regression parameter $\boldsymbol{\beta}$: 95% marginal posterior credibility intervals (left), traceplots (center) and posterior marginal densities (right).

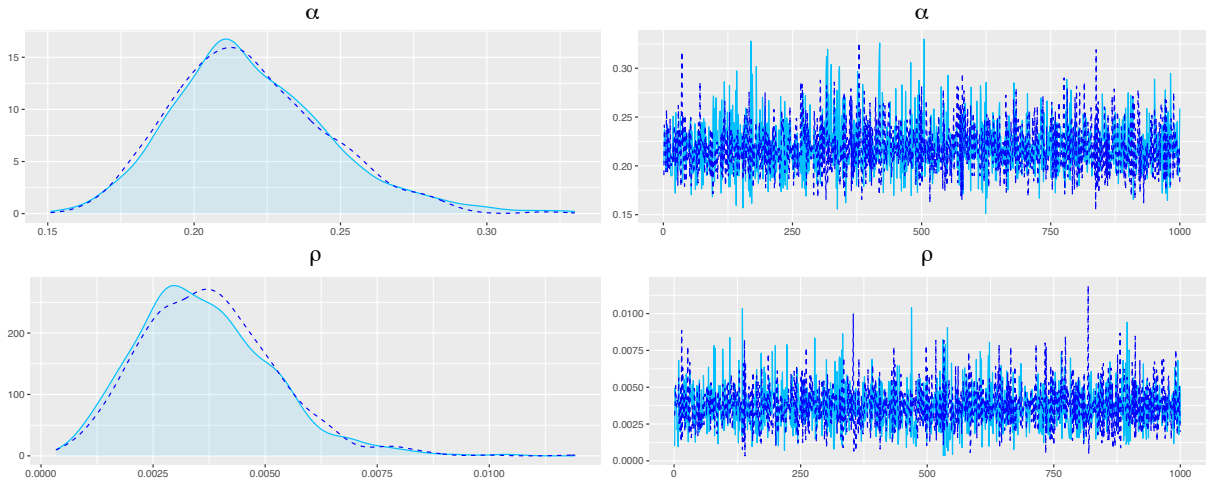
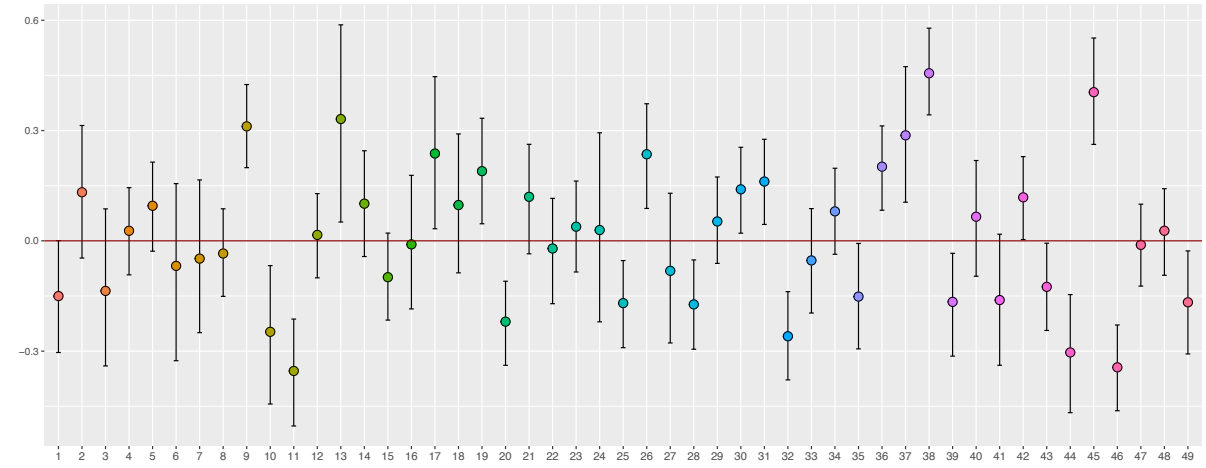


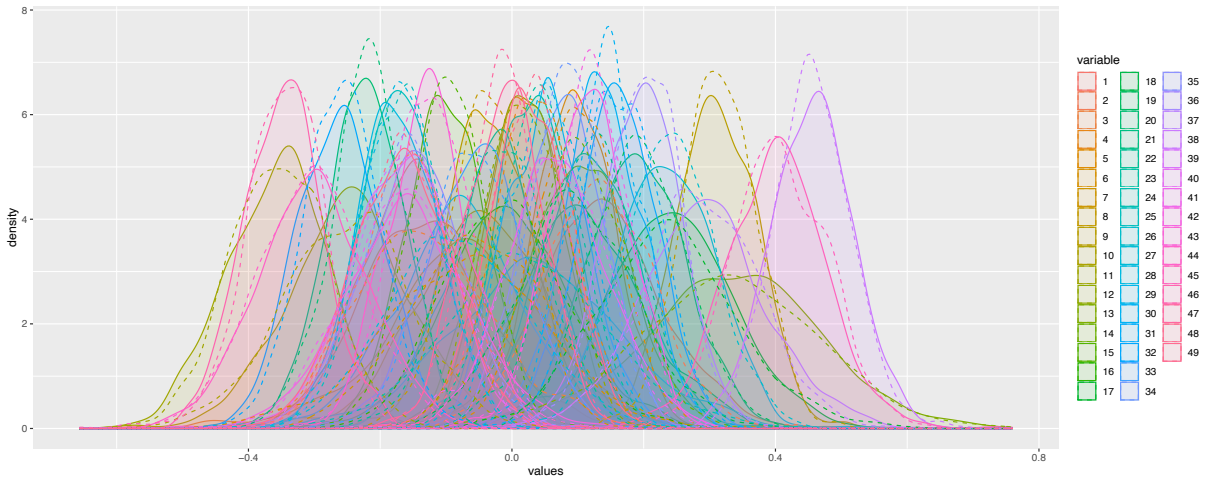

Figure 3.12: Posterior inference for the parameters characterizing the spatial residuals $\mathbf{w}$. Marginal posterior densities and traceplots of $\alpha$ (top row) and $\rho$ (bottom row).

While posterior inference for parameters $a_{NR1}, a_{R1}, a_{NR2}, a_{R2}, b_{NR1}, b_{R1}, b_{NR2}, b_{R2}, c_{NR}, c_R$

and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$ led to comments similar to the ones reported in Section 3.2, some observations can be made on the spatial parameter $\rho$. Looking at Figure 3.12, which shows marginal posterior densities and traceplots of $\alpha$ and $\rho$, it can be noticed that the parameter $\rho$ seems to be distributed over values far from zero, contrary to what has been observed in Figure 3.7. This change derives from the inclusion of the month-specific variance into the model, which now expresses information about the data that was improperly carried by $\rho$ before.




(a) 95% marginal posterior CIs of $\mathbf{w}$




(b) Marginal posterior density distribution of $\mathbf{w}$

Figure 3.13: Marginal posterior credibility intervals and densities of $w_1, \ldots, w_{49}$ (spatial residual terms) in model 3.

Figure 3.13 shows the posterior inference for the spatial residual term $\mathbf{w}$ in this last model. Both 95% marginal posterior CIs and densities seem to behave much like the ones reported in Figure 3.8, leading to the same comments.

Looking at the 95% posterior marginal CIs for the new month-specific parameter $\boldsymbol{\sigma^2}$, which are reported in Figure 3.14, more interesting comments can be made. The 12 parameters appear to be distributed over disjoint range of values, supporting the choice of assuming a different parameter $\sigma^2$ for each month. Moreover, the shape outlined by the twelve median values (coloured points in Figure 3.14) reminds the one defined in Figure 1.17, as could be expected.

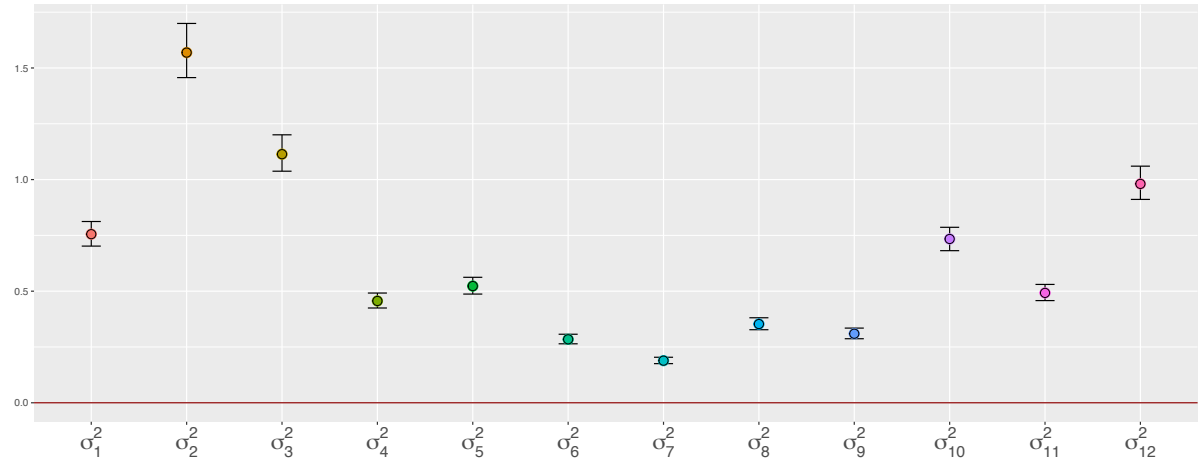



Figure 3.14: 95% posterior marginal credibility intervals of the 12 month-specific parameters $\sigma^2_1, \ldots, \sigma^2_{12}$ expressing the data variability.

All these observations suggest that the third model should be the best performing one. In the following section, that is Section 3.4, reliable goodness-of-fit criteria for the three model will be provided in order to choose the "best" model, which will then be used for the clustering mechanism.

## 3.4. Model Selection

Table 3.1 shows WAIC and LOO-CV (as described in Section 2.4.4) for each of the three models. As expected, since theoretically the WAIC and LOO-CV are assumed to be equal asymptotically, the values provided by the two methods for each model are nearly the same. From the table it is clear that the lowest values achieved for both criteria are

| MODEL | WAIC | PSIS-LOO |
|---|---|---|
| Model 1 | 42129.5 | 42129.5 |
| Model 2 | 41063.6 | 41063.8 |
| Model 3 | 38284.5 | 38284.8 |

Table 3.1: WAIC and LOO computed for the three above mentioned models

those of Model 3, which is hence the "best" model. Thus, even if it is the most expensive one in terms of complexity and computational costs, Model 3 is preferable to Model 1 and Model 2.

## 3.5. Univariate Clustering

In this section are reported the posterior cluster estimates for model (2.82)-(2.91), obtained simulating one MCMC chain with $BI = 1000$ and $M = 1000$. Figure 3.15 provides a graphical overview of the number of non-empty clusters at each iteration of the MCMC chain. Noticed that at each iteration at least two different clusters can be identified, since the traceplot and histogram reported in Figure 3.15 do not cover values lower than two.

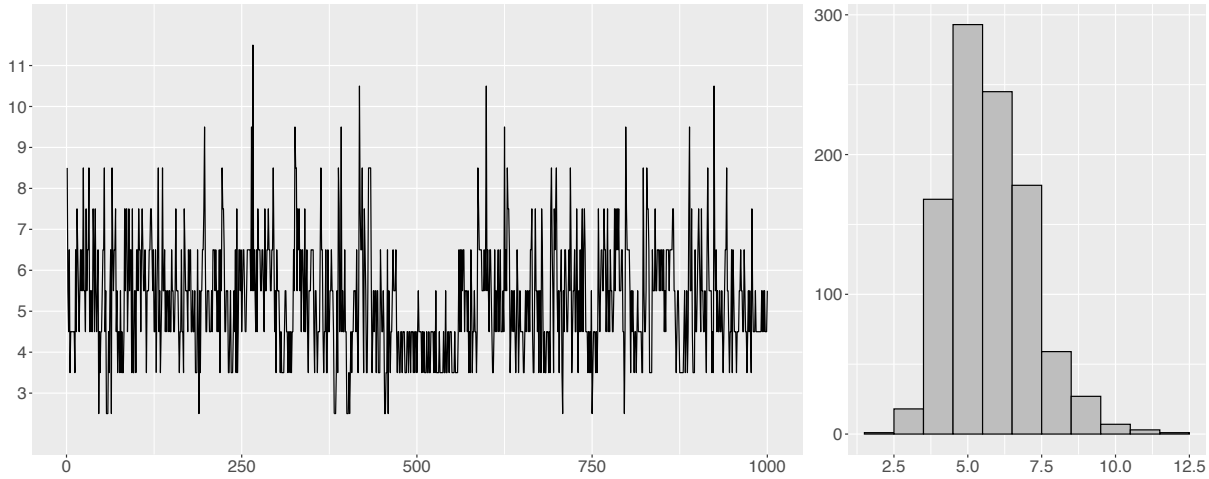



Figure 3.15: Chain (left) and histogram (right) specifying the number of non-empty clusters allocated by the MCMC chain at each iteration.

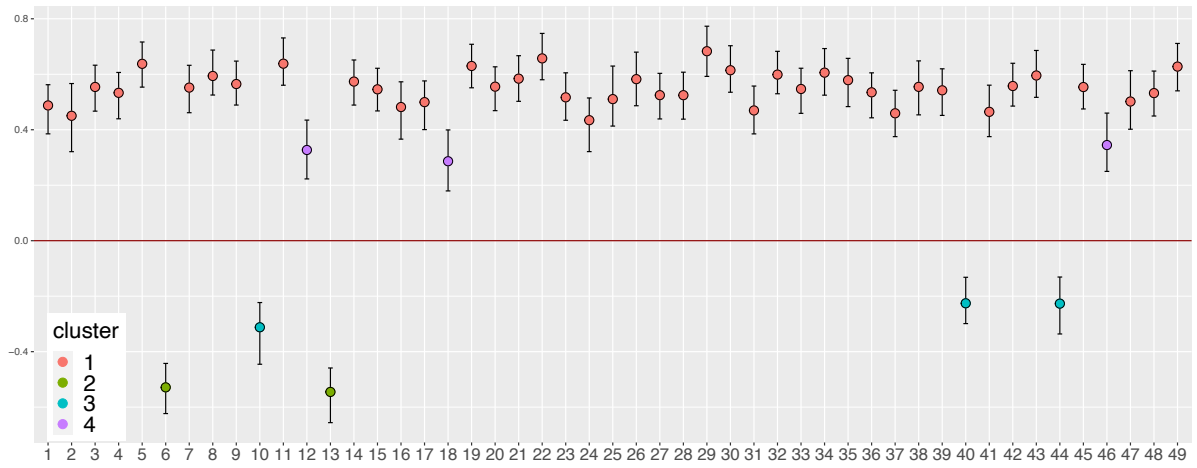



Figure 3.16: Marginal posterior 95% CIs for the annual cosine parameter $b_1$ in each of the 49 stations. The colour of each median point is assigned accordingly to the clusterization.

Figure 3.16 shows the the 95% marginal posterior CIs of the 49 parameters $b_i$, each of them characterizing the annual cosine coefficient associated to the corresponding station. The median value has been coloured according to the resulting clusterization, which has been obtained minimizing the Binder's loss function, following what explained in Section 2.5. The number of estimated clusters via Binder's loss function is **four**. They will be labeled as group $1, 2, 3$ and 4 and will be associated to the same four colours in all the figures reported in this section. The clusterization has been performed by looking only at parameter $b_i, \quad i = 1, \ldots, 49$, i.e. the annual cosine coefficient for each station, defining a sort of "stratification" given by the four groups (see Figure 3.16). All the stations in group 1 present an annual sine coefficient $b_i$ distributed over the highest values available, while stations belonging to group 4 show CIs centered around slightly lower positive values. On the contrary, groups 2 and 3 contain stations associated to a negative valued annual cosine coefficient, enlightening an inverse trend with respect to the stations in groups 1 and 4. This "stratification" can be identified also in Figure 3.17, which reports the time series of $PM_{10}$ log-concentrations recorded by each station. Each of the 49 time series has been coloured according to the cluster-allocation of its corresponding station. From the outline of the 49 station recordings coloured by group, two main trends can be identified: while time series associated to groups 1 and 4 delineate an upward concavity (given by positive-valued $b_i$), the ones in groups 2 and 3 seem to follow a down-facing concave behaviour (due to negative-valued $b_i$).

By looking at the map reported in Figure 3.18, some comments can be made about the relation between estimated clusters and geographical locations of the stations. Notice

that stations assigned to groups 3 and 4 are actually detached from the others, being located along the Apennines mountain range. This distinction among spatial recording sites seems to validate the estimated clusterization of the 49 stations.

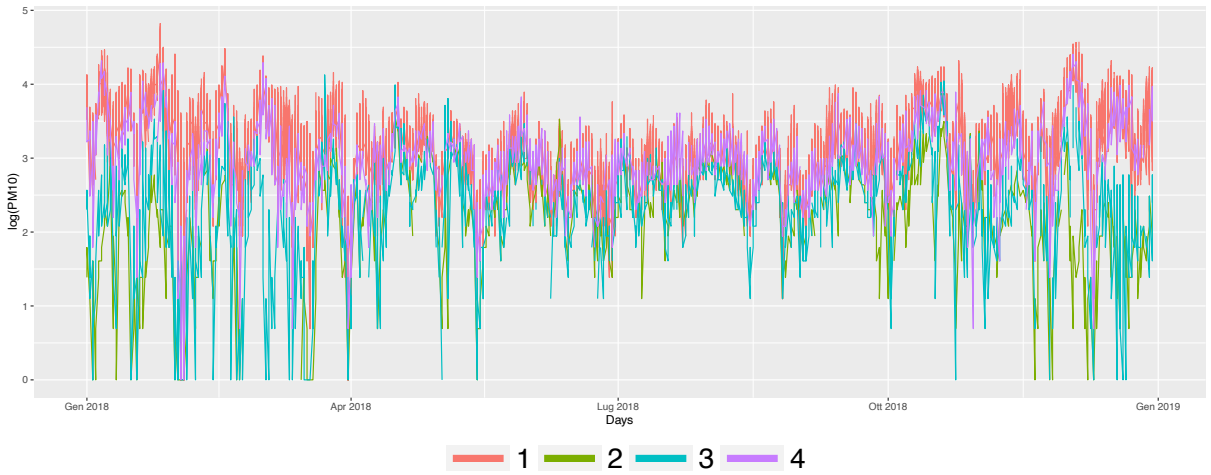



Figure 3.17: Functional outline of the 49 stations recordings of $PM_{10}$ log-concentrations, coloured according to the BNP univariate clusterization.

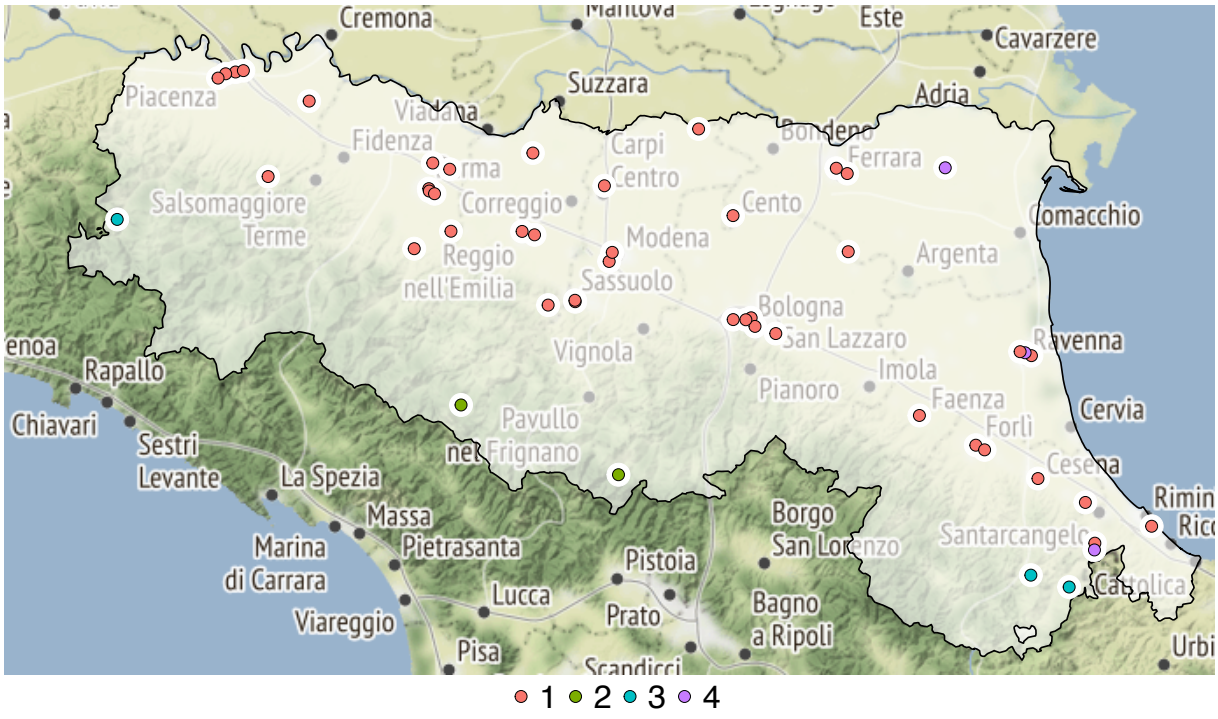
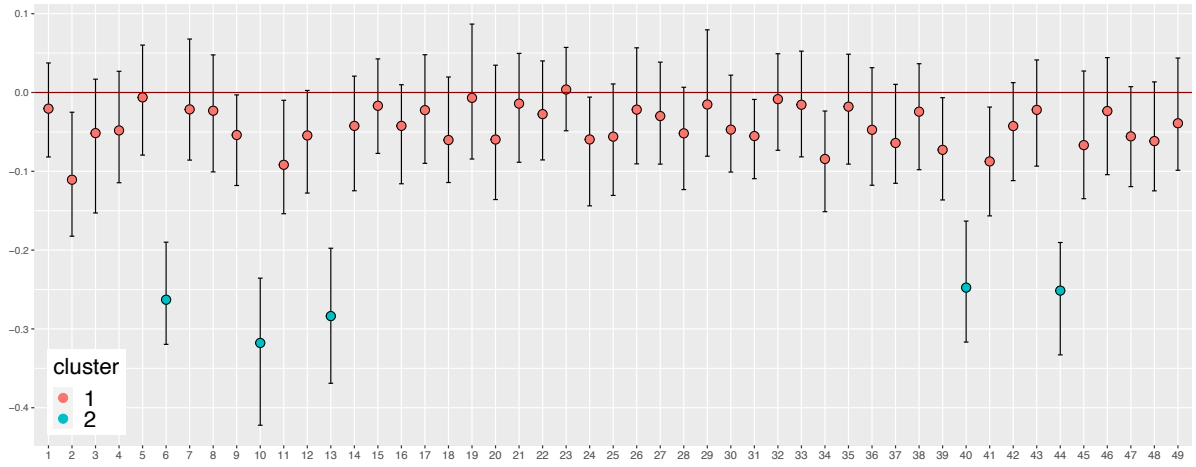


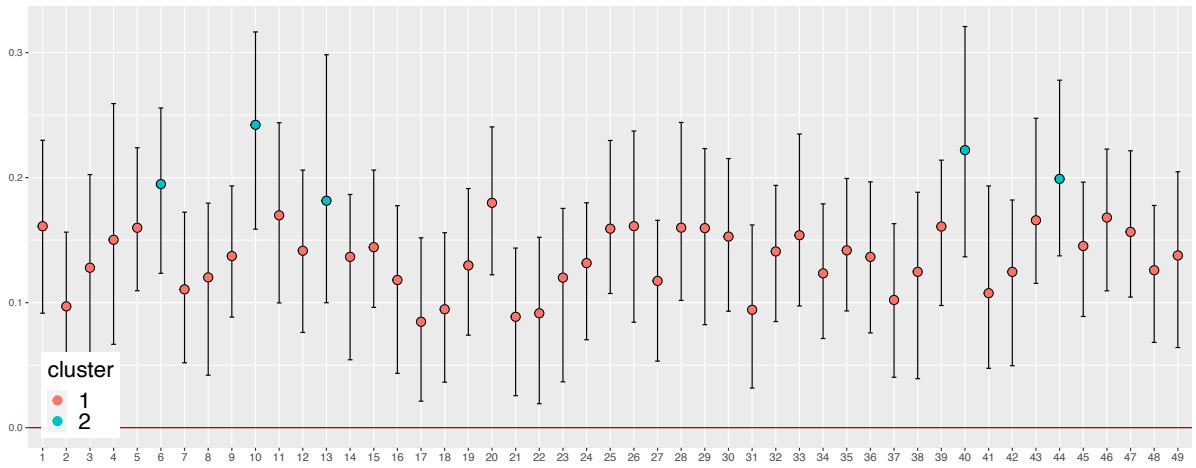

Figure 3.18: Geographical map of Emilia-Romagna displaying the 49 recording sites coloured according to the group they have been assigned.

## 3.6. Multivariate Clustering

Here are reported the posterior cluster estimates for model (2.99)-(2.107), obtained simulating one MCMC chain with $BI = 500$ and $M = 500$ due to the high computational cost. The procedure is analogue to the one carried out for the univariate case. Also in this framework the Binder's loss function has been used for defining the optimal stations partitioning and the obtained clusters will be identified by the same colours in all the figures. In this second case only **two clusters** have been identified among the 49 stations. Figure 3.19 and 3.20 show the 95% marginal posterior CIs for each of the five parameters in each station. As before, the median are coloured according to the corresponding stations group-allocation. The visual arrangement of the CIs actually suggests a two-cluster partition when looking at the "stratification" of the median values.

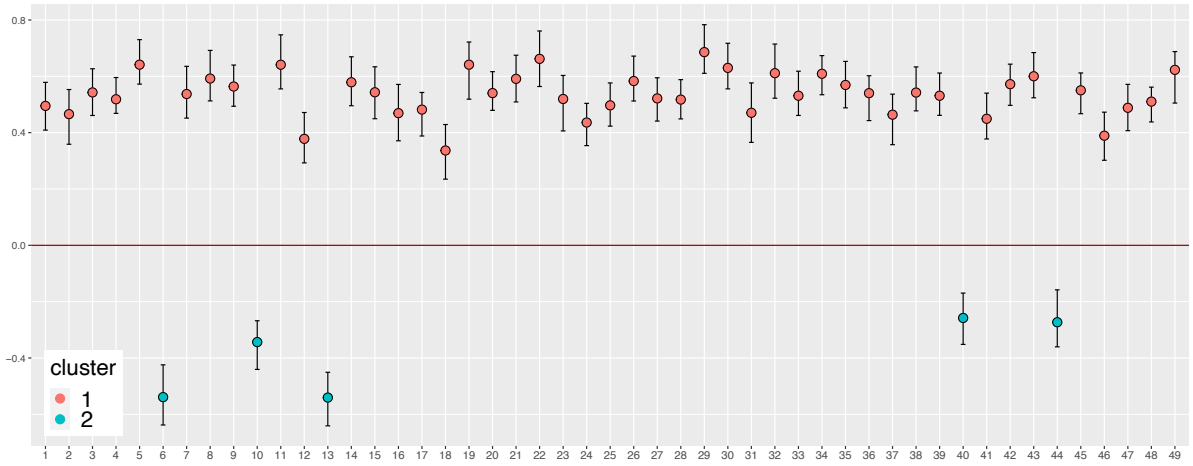



(a) Annual sine coefficient $a_1$

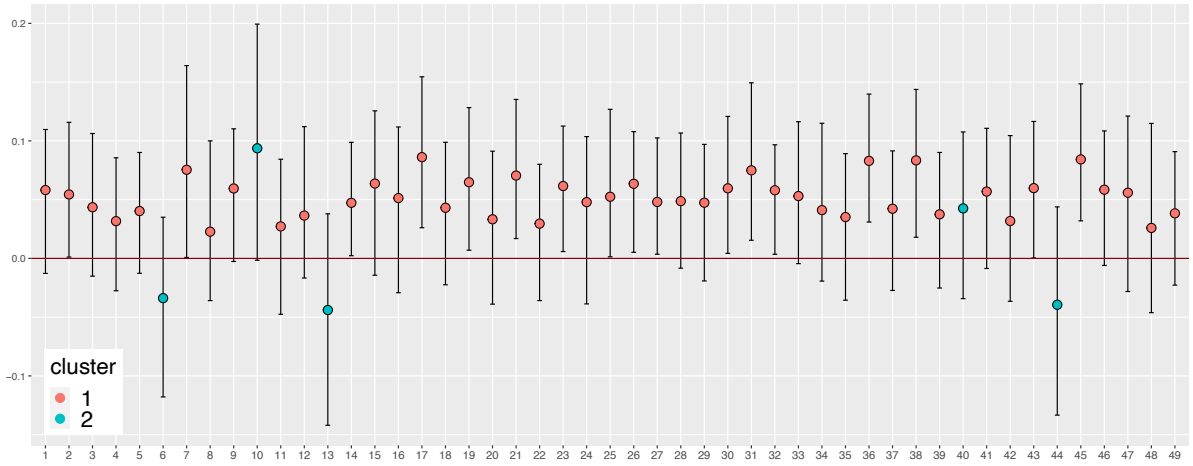



(b) Quarterly sine coefficient $a_2$

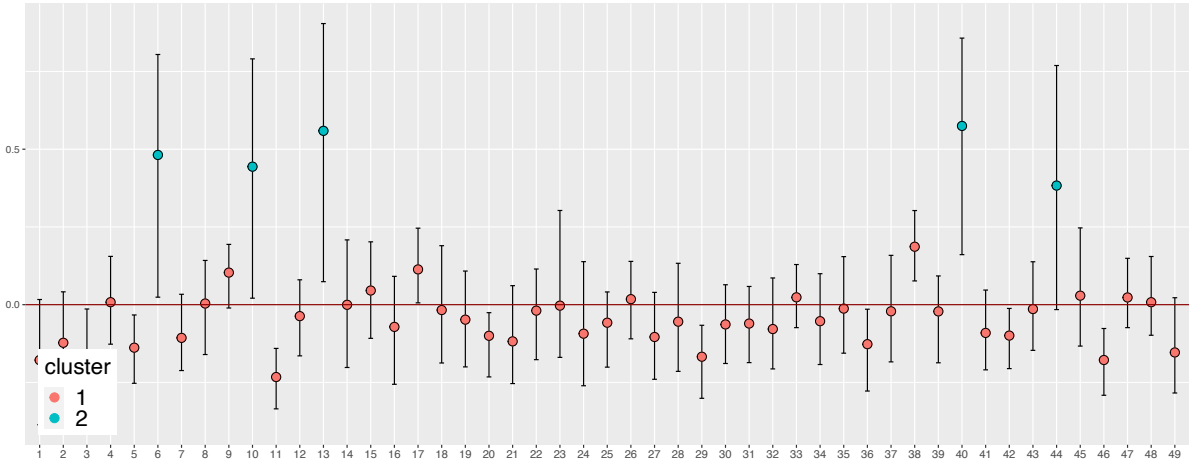
Figure 3.19: Marginal posterior 95% CIs for parameters $a_1, a_2$ in each of the 49 stations.




(a) Annual cosine coefficient $b_1$




(b) Quarterly cosine coefficient $b_2$




(c) Constant additive term $c$

Figure 3.20: Marginal posterior 95% CIs for parameters $b_1, b_2, c$ in each of the 49 stations.

In Figure 3.21 are reported the 49 time-series of $PM_{10}$ log-concentrations, coloured according to the group allocation of their corresponding stations. Also in this case it is easy to distinguish two well-defined trends, provided by the two estimated clusters.

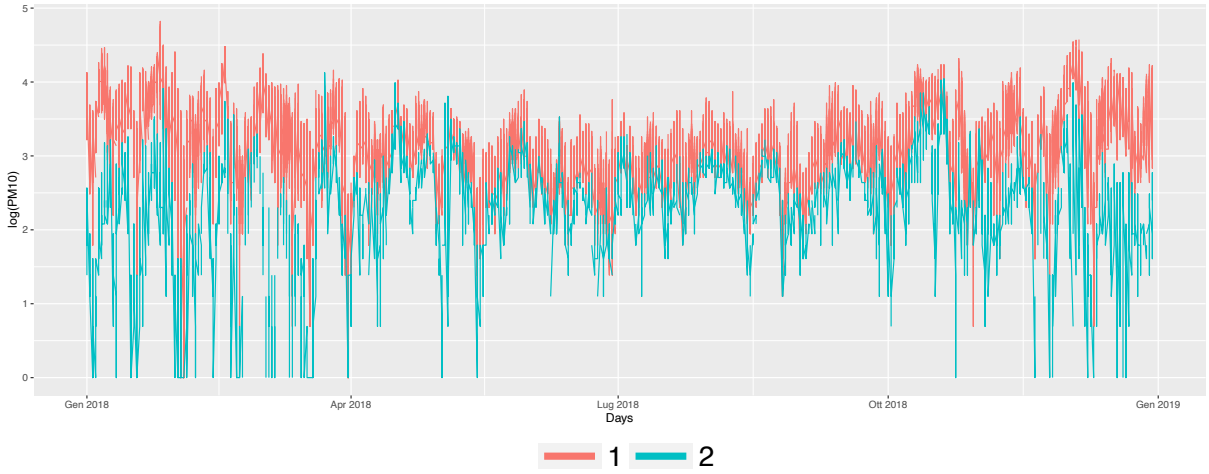



Figure 3.21: Functional outline of the 49 stations recordings of $PM_{10}$ log-concentrations, coloured according to the BNP multivariate clusterization.

Figure 3.22 shows Emilia-Romagna geographical map, where the 49 recording sites have been coloured according to their cluster allocation. In this second case, the estimated clusters define a sharp distinction between the stations located on the Apennines mountain range and the others. This partition is coherent with the lower expected pollution level in mountainous areas. Indeed, the recording stations over the Apennines are probably associated with higher altitudes and limited urbanization areas, which are factors linked to lower pollutant concentrations.

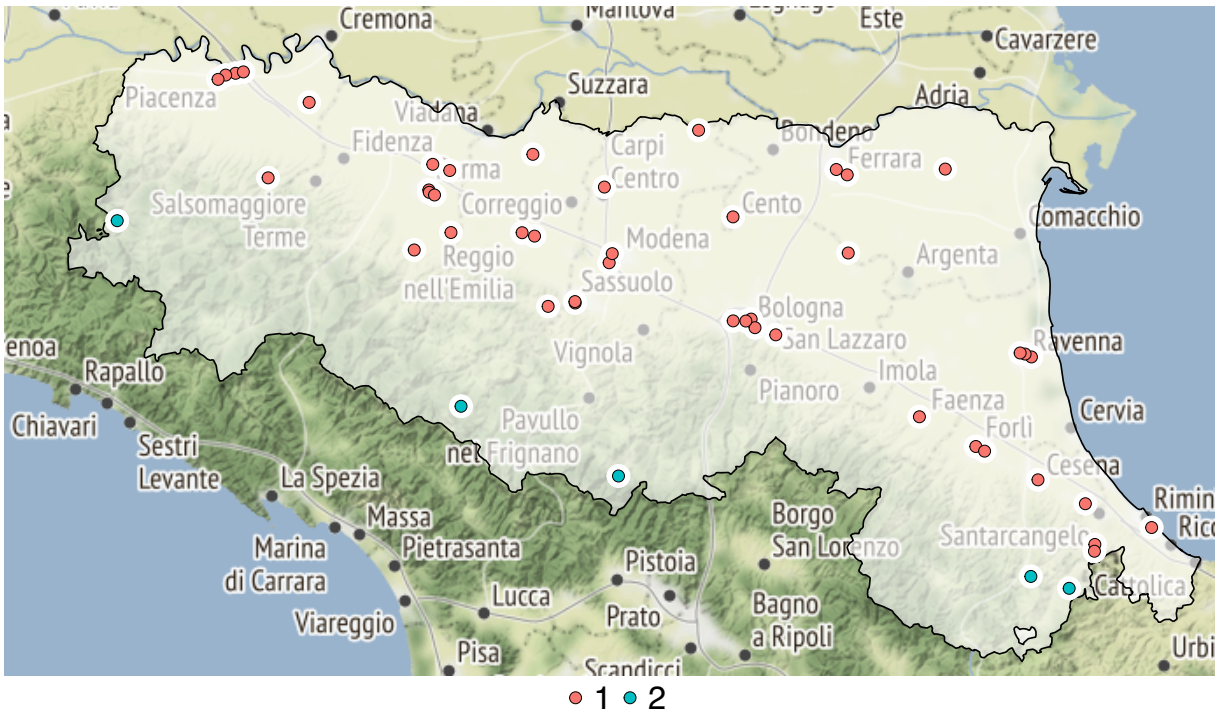



Figure 3.22: Geographical map of Emilia-Romagna displaying the 49 recording sites coloured according to the group they have been assigned.

# 4 | Conclusions and Future Developments

The analysis presented in this thesis allows for many extensions. In this work we have restricted the data to recording stations in Emilia-Romagna during 2018. However, the proposed models can easily be extended to multiple years and different regions. Of course, when extending the time horizon and considering a higher number of recording stations, the computational effort required to obtain MCMC posterior inference will be significantly increased. This issue can be solved by using more powerful devices to perform the MCMC simulations, which however would require a proper reformulation of the models (see for instance Rue et al. (2009); Martins et al. (2013); Lindgren & Rue (2015)). Because of lack of time, in this thesis we have discarded the computation of spatio-temporal prediction. Indeed, expected log-concentrations of $PM_{10}$ in a "new" day and in a specific location could be computed under our framework, exploiting the marginal posterior distributions obtained with the Bayesian approach. Other possible developments are spatial prediction (kriging) and the inclusion of weather conditions as regressors, as better specified in the sections below.

## 4.1. Spatial prediction

We refer to model (2.47)-(2.58), which from Table 3.1 appears to be the "best" one among the candidates. Having modeled the spatial residual term $\mathbf{w}$ as a Gaussian process (GP) depending on stations coordinates, it is possible to update the model and provide estimates of $w$ in new locations, exploiting the GP posterior estimate. In this work, the value of $w$ has been estimated only in the recording sites, but with this improvement we could estimate $w$ also for locations which are not "covered" by a sensing station. This procedure could be very useful, providing estimates of the spatial correlation allover the area of interest. By including this calculation in our model, and considering data from all the available ARPA stations, we could provide estimates for $w$ in the whole Po valley area. Theoretical background for these calculations can be found in Banerjee et al. (2015).

## 4.2. Meteorological factors

As it is well known from available literature (see Dung et al. (2019); Onuorah et al. (2019); Tian et al. (2014)), some meteorological factors heavily affect the particular matter (PM) concentrations in the air. While rainfall, snowfall and strong winds contribute to lowering the pollutant level, phenomena such as fog, humidity and high atmospheric pressure seem to promote the air pollutant stagnation. Hence, we would like to include such influential factors as regressors in the $PM_{10}$ modeling. The main problem in doing so, is that the ARPA monitoring network is made of separate stations for pollution and meteorological measurements, which are located in different places. Moreover, considering the pollutant concentrations data and the meteorological data, a time downscaling problem can be identified. Indeed, while the $PM_{10}$ registrations are collected daily, the meteorological factors are registered with a different timing (for instance every hour). The former issues represent a huge problem when trying to combine these two databases, which would require a proper polishing in order to be comparable. This procedure, together with the implementation of suitable ad-hoc methods to overcome spatial and temporal issues, would require a huge amount of time and thus it has not been developed in this master thesis.

# References

Argiento, R., Bianchini, I., & Guglielmi, A. (2015). A blocked Gibbs sampler for NGG-mixture models via a priori truncation. *Statistics and Computing*, *26*. doi: 10.1007/s11222-015-9549-6

ARPA Lombardia. (1999). *Tipologia delle Stazioni.* Retrieved 2022-08-04, from `https://www.arpalombardia.it/Pages/Aria/Rete-di-rilevamento/Criteri-di-rilevamento/Tipologia-delle-stazioni`

ARPA Lombardia. (2022). *Criteri di rilevamento - particolato atmosferico.* Retrieved 2022-08-14, from `https://www.arpalombardia.it/Pages/Aria/Rete-di-rilevamento/Criteri-di-rilevamento.aspx`

ARPA Valle d'Aosta. (2022). *Principi e metodi di misura dei principali inquinanti atmosferici.* Retrieved 2022-08-15, from `https://www.arpa.vda.it/images/stories/ARPA/aria/lavalutazioneQA/retemonitoraggio/principi_di_misura_inquinanti.pdf`

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*. CRC Press.

Beraha, M., Falco, D., & Guglielmi, A. (2021). *JAGS, NIMBLE, STAN: a detailed comparison among Bayesian MCMC software.* arXiv. Retrieved from `https://arxiv.org/abs/2107.09357` doi: 10.48550/ARXIV.2107.09357

Cameletti, M. (2020). The Effect of Corona Virus Lockdown on Air Pollution: Evidence from the City of Brescia in Lombardia Region (Italy). *Atmospheric Environment*, *239*, 117794. doi: 10.1016/j.atmosenv.2020.117794

Dahl, D. B. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In K.-A. Do, P. Müller, & M. Vannucci (Eds.), *Bayesian inference for gene expression and proteomics* (p. 201–218). Cambridge University Press. doi: 10.1017/CBO9780511584589.011

Dahl, D. B., Johnson, D. J., & Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 1-13.

Dung, N., Hong Son, D., Nguyen The Duc, H., & Tri, D. (2019). Effect of Meteorological Factors on PM10 Concentration in Hanoi, Vietnam. *Journal of Geoscience and Environment Protection*, *07*, 138-150. doi: 10.4236/gep.2019.711010.

EEA. (2019). *Air quality in Europe - 2019 Report: Technical Report* (Tech. Rep.). European Environmental Agency (EEA). Retrieved from `https://www.eea.europa.eu/publications/air-quality-in-europe-2019`.

Essomba, R. (2017). *Smoothing techniques using basis functions: Fourier basis.* Retrieved 2022-08-11, from `https://datascienceplus.com/smoothing-techniques-using-basis-functions-fourier-basis/`

Gilks, W., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. CRC Press.

Hoffman, M. D., & Gelman, A. (2011). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.* arXiv. Retrieved from `https://arxiv.org/abs/1111.4246` doi: 10.48550/ARXIV.1111.4246

Ishwaran, H., & Zarepour, M. (2002). Exact and Approximate Sum Representations for the Dirichlet process. *Canadian Journal of Statistics*, *30*, 269 - 283. doi: 10.2307/3315951

Karagulian, F., Gerboles, M., Barbiere, M., Kotsev, A., Lagler, F., & Borowiak, A. (2019). *Review of sensors for air quality monitoring* (Tech. Rep. No. EUR 29826 EN). Publications Office of the European Union, Luxembourg: Joint Research Centre (JRC). (ISBN 978-92-76-09255-1, doi:10.2760/568261, JRC116534)

Larsen, B., Gilardoni, S., Stenström, K., Niedzialek, J., Jimenez, J., & Belis, C. (2012). Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmospheric Environment*, *50*, 203-213. doi: 10.1016/j.atmosenv.2011.12.038

Lenschow, P., Abraham, H.-J., Kutzner, K., Lutz, M., Preuß, J.-D., & Reichenbächer, W. (2001). Some ideas about the sources of PM10. *Atmospheric Environment*, *35*. doi: 10.1016/S1352-2310(01)00122-4

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1-25.

Lonati, G., & Riva, F. (2021a). effetti degli interventi di contrasto alla diffusione del COVID19 sulla qualità dell'aria in Pianura Padana. *Ingegneria dell'Ambiente*, *8*(1/2021), 24-39. doi: https://doi.org/10.32024/ida.v8i1.327

Lonati, G., & Riva, F. (2021b). Regional Scale Impact of the COVID-19 Lockdown on Air Quality: Gaseous Pollutants in the Po Valley, Northern Italy. *Atmosphere*, *12*(2). Retrieved from `https://www.mdpi.com/2073-4433/12/2/264` doi: 10.3390/atmos12020264

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, *67*, 68-83.

Masiol, M., Squizzato, S., Formenton, G., Harrison, R., & Agostinelli, C. (2017). Air quality across a European hotspot: Spatial gradients, seasonality, diurnal cycles and trends in the Veneto region, NE Italy. *Science of The Total Environment*, *576*, 210-224. doi: 10.1016/j.scitotenv.2016.10.042

Onuorah, C., Leton, T., & Momoh, Y. (2019). Influence of Meteorological Parameters on Particle Pollution (PM2.5 and PM10) in the Tropical Climate of Port Harcourt, Nigeria. *Archives of Current Research International*, 1-12. doi: 10.9734/acri/2019/v19i130149

Pozzer, A., Bacer, S., Sappadina, S. D. Z., Predicatori, F., & Caleffi, A. (2019). Long-term concentrations of fine particulate matter and impact on human health in Verona, Italy. *Atmospheric Pollution Research*, *10*(3), 731-738. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1309104218303465` doi: https://doi.org/10.1016/j.apr.2018.11.012

Rasmussen, Edward, C., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Rosner, G. L., Laud, P. W., & Johnson, W. O. (2021). *Bayesian thinking in biostatistics*. CRC Press.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, *71*, 319-392.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*(2), 639–650. Retrieved 2022-09-04, from `http://www.jstor.org/stable/24305538`

Shi, Y., Martens, M., Banerjee, A., & Laud, P. (2019). Low information omnibus (LIO) priors for Dirichlet process mixture models. *Bayesian Analysis*, *14*(3), 677–702.

Stan Development Team. (2022a). *Gaussian Process Regression.* Retrieved 2022-08-10, from `https://mc-stan.org/docs/stan-users-guide/gaussian-process-regression.html`

Stan Development Team. (2022b). *Stan modeling language users guide and reference manual, version 2.30.* Retrieved 2022-08-10, from `https://mc-stan.org`

Stan Development Team. (2022c). *Stan modeling language users guide and reference manual, version 2.30.* Retrieved 2022-08-10, from `https://mc-stan.org/docs/reference-manual/sampling.html`

Taylor, C., Yousif, A., & Mwitondi, K. (2018). Statistical analysis of particulate matter data in Doha, Qatar. In (Vol. 230, p. 107-118). doi: 10.2495/AIR180101

Tian, G., Qiao, Z., & Xu, X. (2014). Characteristics of particulate matter (PM10) and its relationship with meteorological factors during 2001–2012 in Beijing. *Environmental Pollution*, *192*, 266-274. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0269749114001857` doi: https://doi.org/10.1016/j.envpol.2014.04.036

U.S. EPA. (1978). *Altitude as a factor in air pollution* (Tech. Rep.). Washington, D.C.: U.S. Environmental Protection Agency. (EPA/600/9-78/015 (NTIS PB285645))

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and waic for bayesian models.* Retrieved from `https://mc-stan.org/loo/` (R package version 2.5.1)

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. Retrieved from `https://doi.org/10.1007/s11222-016-9696-4` doi: 10.1007/s11222-016-9696-4

Wade, S., & Ghahramani, Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Analysis*, *13*(2), 559 – 626. Retrieved from `https://doi.org/10.1214/17-BA1073` doi: 10.1214/17-BA1073

Wan Mahiyuddin, W. R., Sahani, M., Aripin, R., Latif, M. T., e, T.-Q., & e, C.-M. (2012). Short-term effects of daily air pollution on mortality. *Atmospheric Environment*, *65*, 69. doi: 10.1016/j.atmosenv.2012.10.019

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine*

*Learning Research*, *11*(116), 3571–3594. Retrieved from `http://jmlr.org/papers/v11/watanabe10a.html`

Yong, L. (2018). *LOO and WAIC as Model Selection Methods for Polytomous Items.* arXiv. Retrieved from `https://arxiv.org/abs/1806.09996` doi: 10.48550/ARXIV.1806.09996

# A | Appendix: *Stan* Code

## A.1. Model 1

```
data {
    int<lower=0> N_obs;
    int<lower=0> N_miss;

    int<lower=0> p;
    int<lower=0> N_staz;
    int<lower=0> N_days;

    vector[N_obs] Y_obs;
    array[N_obs] int t_obs;
    array[N_obs] int ID_obs;

    array[N_miss] int t_miss;
    array[N_miss] int ID_miss;

    matrix[N_staz, p] X;
    vector[N_staz] r;

    real omega;
    vector[N_days] day;
}



parameters {
    vector[N_miss] Y_miss;
```

```
    real<lower=0> sigma_sq;

    vector[p] betas;

    vector[2] a;
    vector[2] b;
    real c;

    vector[2] a_r;
    vector[2] b_r;
    real c_r;
}



transformed parameters {

    real<lower=0> sigma = sqrt(sigma_sq);

    vector[N_days] ft;
    ft = rep_vector(c, N_days);
    ft += a[1]*sin(omega*day) + b[1]*cos(omega*day);
    ft += a[2]*sin(4*omega*day) + b[2]*cos(4*omega*day);

    vector[N_days] ft_r;
    ft_r = rep_vector(c_r, N_days);
    ft_r += a_r[1]*sin(omega*day) + b_r[1]*cos(omega*day);
    ft_r += a_r[2]*sin(4*omega*day) + b_r[2]*cos(4*omega*day);

    vector[N_obs] mu_obs;
    mu_obs[1:N_obs] = r[ID_obs[1:N_obs]] .* ft_r[t_obs[1:N_obs]]
                    + (1 - r[ID_obs[1:N_obs]]).* ft[t_obs[1:N_obs]];
    mu_obs[1:N_obs] += X[ID_obs[1:N_obs]] * betas;



    vector[N_miss] mu_miss;
    mu_miss[1:N_miss] = r[ID_miss[1:N_miss]] .* ft_r[t_miss[1:N_miss]]
```

```
                        + (1 - r[ID_miss[1:N_miss]]).* ft[t_miss[1:N_miss]];
    mu_miss[1:N_miss] += X[ID_miss[1:N_miss]] * betas;
}



model {
    sigma_sq ~ inv_gamma(3, 2);

    betas ~ normal(0, 2);

    a ~ normal(0, 1);
    b ~ normal(0, 1);
    c ~ normal(0, 1);

    a_r ~ normal(0, 1);
    b_r ~ normal(0, 1);
    c_r ~ normal(0, 1);

    Y_obs ~ normal(mu_obs, sigma);
    Y_miss ~ normal(mu_miss, sigma);
}



generated quantities {
  vector[N_obs] log_lik;
  for (j in 1:N_obs) {
    log_lik[j] = normal_lpdf(Y_obs[j] | mu_obs[j], sigma);
  }
}
```

## A.2. Model 2

```
data {
    int<lower=0> N_obs;
    int<lower=0> N_miss;

    int<lower=0> p;
    int<lower=0> N_staz;
    int<lower=0> N_days;

    vector[N_obs] Y_obs;
    array[N_obs] int t_obs;
    array[N_obs] int ID_obs;

    array[N_miss] int t_miss;
    array[N_miss] int ID_miss;

    matrix[N_staz, p] X;
    vector[N_staz] r;
    vector[2] coord[N_staz];

    real omega;
    vector[N_days] day;
}



parameters {
    vector[N_miss] Y_miss;

    real<lower=0> sigma_sq;

    vector[p] betas;

    vector[2] a;
    vector[2] b;
    real c;
```

```
    vector[2] a_r;
    vector[2] b_r;
    real c_r;

    vector[N_staz] w;
    real<lower=0> rho;
    real<lower=0> alpha;
}



transformed parameters {

    real<lower=0> sigma = sqrt(sigma_sq);

    vector[N_days] ft;
    ft = rep_vector(c, N_days);
    ft += a[1]*sin(omega*day) + b[1]*cos(omega*day);
    ft += a[2]*sin(4*omega*day) + b[2]*cos(4*omega*day);

    vector[N_days] ft_r;
    ft_r = rep_vector(c_r, N_days);
    ft_r += a_r[1]*sin(omega*day) + b_r[1]*cos(omega*day);
    ft_r += a_r[2]*sin(4*omega*day) + b_r[2]*cos(4*omega*day);

    cov_matrix[N_staz] H = cov_exp_quad(coord, alpha, rho);

    vector[N_obs] mu_obs;
    mu_obs[1:N_obs] = r[ID_obs[1:N_obs]] .* ft_r[t_obs[1:N_obs]]
                    + (1 - r[ID_obs[1:N_obs]]).* ft[t_obs[1:N_obs]];
    mu_obs[1:N_obs] += X[ID_obs[1:N_obs]] * betas;
    mu_obs[1:N_obs] += w[ID_obs[1:N_obs]];



    vector[N_miss] mu_miss;
    mu_miss[1:N_miss] = r[ID_miss[1:N_miss]] .* ft_r[t_miss[1:N_miss]]
```

```
                        + (1 - r[ID_miss[1:N_miss]]).* ft[t_miss[1:N_miss]];
    mu_miss[1:N_miss] += X[ID_miss[1:N_miss]] * betas;
    mu_miss[1:N_miss] += w[ID_miss[1:N_miss]];
}



model {

    sigma_sq ~ inv_gamma(3, 2);

    betas ~ normal(0, 2);

    a ~ normal(0, 1);
    b ~ normal(0, 1);
    c ~ normal(0, 1);

    a_r ~ normal(0, 1);
    b_r ~ normal(0, 1);
    c_r ~ normal(0, 1);

    rho ~ beta(3, 10);
    alpha ~ normal(0.3, 0.1);

    w ~ multi_normal(rep_vector(0, N_staz), H);

    Y_obs ~ normal(mu_obs, sigma);
    Y_miss ~ normal(mu_miss, sigma);
}



generated quantities {
  vector[N_obs] log_lik;
  for (j in 1:N_obs) {
    log_lik[j] = normal_lpdf(Y_obs[j] | mu_obs[j], sigma);
  }
```

```
}
```

## A.3. Model 3

```
data {
    int<lower=0> N_obs;
    int<lower=0> N_miss;

    int<lower=0> p;
    int<lower=0> N_staz;
    int<lower=0> N_days;

    vector[N_obs] Y_obs;
    array[N_obs] int t_obs;
    array[N_obs] int m_obs;
    array[N_obs] int ID_obs;

    array[N_miss] int t_miss;
    array[N_miss] int m_miss;
    array[N_miss] int ID_miss;

    matrix[N_staz, p] X;
    vector[N_staz] r;
    vector[2] coord[N_staz];

    real omega;
    vector[N_days] day;
}



parameters {
    vector[N_miss] Y_miss;

    vector<lower=0>[12] sigma_sq;

    vector[p] betas;
```

```
    vector[2] a;
    vector[2] b;
    real c;

    vector[2] a_r;
    vector[2] b_r;
    real c_r;

    vector[N_staz] w;
    real<lower=0> rho;
    real<lower=0> alpha;
}



transformed parameters {

    vector<lower=0>[12] sigma = sqrt(sigma_sq);

    vector[N_days] ft;
    ft = rep_vector(c, N_days);
    ft += a[1]*sin(omega*day) + b[1]*cos(omega*day);
    ft += a[2]*sin(4*omega*day) + b[2]*cos(4*omega*day);

    vector[N_days] ft_r;
    ft_r = rep_vector(c_r, N_days);
    ft_r += a_r[1]*sin(omega*day) + b_r[1]*cos(omega*day);
    ft_r += a_r[2]*sin(4*omega*day) + b_r[2]*cos(4*omega*day);

    cov_matrix[N_staz] H = gp_exp_quad_cov(coord, alpha, rho);

    vector[N_obs] mu_obs;
    mu_obs[1:N_obs] = r[ID_obs[1:N_obs]] .* ft_r[t_obs[1:N_obs]]
                    + (1 - r[ID_obs[1:N_obs]]).* ft[t_obs[1:N_obs]];
    mu_obs[1:N_obs] += X[ID_obs[1:N_obs]] * betas;
    mu_obs[1:N_obs] += w[ID_obs[1:N_obs]];
```

```
    vector[N_miss] mu_miss;
    mu_miss[1:N_miss] = r[ID_miss[1:N_miss]] .* ft_r[t_miss[1:N_miss]]
                      + (1 - r[ID_miss[1:N_miss]]).* ft[t_miss[1:N_miss]];
    mu_miss[1:N_miss] += X[ID_miss[1:N_miss]] * betas;
    mu_miss[1:N_miss] += w[ID_miss[1:N_miss]];
}



model {

    sigma_sq ~ inv_gamma(3, 2);

    betas ~ normal(0, 2);

    a ~ normal(0, 1);
    b ~ normal(0, 1);
    c ~ normal(0, 1);

    a_r ~ normal(0, 1);
    b_r ~ normal(0, 1);
    c_r ~ normal(0, 1);

    rho ~ beta(3, 10);
    alpha ~ normal(0.3, 0.1);

    w ~ multi_normal(rep_vector(0, N_staz), H);

    Y_obs[1:N_obs] ~ normal(mu_obs[1:N_obs], sigma[m_obs[1:N_obs]]);

    Y_miss[1:N_miss] ~ normal(mu_miss[1:N_miss], sigma[m_miss[1:N_miss]])
}



generated quantities {
```

```
  vector[N_obs] log_lik;
  for (j in 1:N_obs) {
    log_lik[j] = normal_lpdf(Y_obs[j] | mu_obs[j], sigma[m_obs[j]]);
  }
}
```

## A.4. Univariate Clustering

```
data {
    int<lower=0> N_obs;
    int<lower=0> N_miss;

    int<lower=0> p;
    int<lower=0> N_staz;
    int<lower=0> N_days;

    vector[N_obs] Y_obs;
    array[N_obs] int t_obs;
    array[N_obs] int m_obs;
    array[N_obs] int ID_obs;

    array[N_miss] int t_miss;
    array[N_miss] int m_miss;
    array[N_miss] int ID_miss;

    matrix[N_staz, p] X;
    vector[2] coord[N_staz];

    real omega;
    vector[N_days] day;

    int<lower=0> C;
}
```

```
parameters {
    vector[N_miss] Y_miss;

    vector<lower=0>[12] sigma_sq;

    vector[p] betas;

    real a2, b2;
    real a;
    real c;
    vector[N_staz] b;

    vector[N_staz] w;
    real<lower=0> rho;
    real<lower=0> alpha;

    vector<lower=0>[C] s_sq;
    vector[C] m;
    vector<lower=0, upper=1>[C-1] v;
}



transformed parameters {

    vector<lower=0>[12] sigma = sqrt(sigma_sq);

    matrix[N_staz, N_days] ft;
    for(i in 1:N_staz){
      ft[i, ] = (rep_vector(c, N_days))';
      ft[i, ] += (a*sin(omega*day) + b[i]*cos(omega*day)
                  + a2*sin(4*omega*day) + b2*cos(4*omega*day)
                    )';
      }

    cov_matrix[N_staz] H = gp_exp_quad_cov(coord, alpha, rho);
```

```
    vector[N_obs] mu_obs;
    for(i in 1:N_obs){
      mu_obs[i] =  ft[ID_obs[i], t_obs[i]];
      }
    mu_obs[1:N_obs] += X[ID_obs[1:N_obs], ]*betas;
    mu_obs[1:N_obs] += w[ID_obs[1:N_obs]];

    vector[N_miss] mu_miss;
    for(i in 1:N_miss){
      mu_miss[i] = ft[ID_miss[i], t_miss[i]];
      }
    mu_miss[1:N_miss] += X[ID_miss[1:N_miss], ]*betas;
    mu_miss[1:N_miss] += w[ID_miss[1:N_miss]];

    vector<lower=0>[C] s = sqrt(s_sq);


    // Stick Breaking
    vector<lower=0, upper=1> [C-1] cumprod_one_minus_v;
    cumprod_one_minus_v = exp(cumulative_sum(log1m(v)));

    simplex[C] eta;
    eta[1] = v[1];
    eta[2:(C-1)] = v[2:(C-1)] .* cumprod_one_minus_v[1:(C-2)];
    eta[C] = cumprod_one_minus_v[C - 1];

    real param = 2.0;
}



model {

    sigma_sq ~ inv_gamma(3, 2);

    betas ~ normal(0, 1);
```

```
    a2 ~ normal(0, 1);
    b2 ~ normal(0, 1);
    a ~ normal(0, 1);
    c ~ normal(0, 1);

    rho ~ beta(3, 10);
    alpha ~ normal(0.3, 0.1);

    w ~ multi_normal(rep_vector(0, N_staz), H);

    Y_obs[1:N_obs] ~ normal(mu_obs[1:N_obs], sigma[m_obs[1:N_obs]]);

    Y_miss[1:N_miss] ~ normal(mu_miss[1:N_miss], sigma[m_miss[1:N_miss]])


    // Clustering
    m[1:C] ~ normal(0, 10*s[1:C]);
    s_sq ~ inv_gamma(4.5, 0.015);    //(3, 0.006)
    v ~ beta(1, param);

    // Finite Mixture Model
    for (i in 1:N_staz){
        vector[C] lps = log(eta);
        for (j in 1:C){
            lps[j] += normal_lpdf(b[i] | m[j], s[j]);;
            }
        target += log_sum_exp(lps);
        }
}



generated quantities  {
  vector[N_obs] log_lik;
  for (j in 1:N_obs) {
    log_lik[j] = normal_lpdf(Y_obs[j] | mu_obs[j], sigma[m_obs[j]]);
  }
```

```
  int cluster_allocs[N_staz];
  for(i in 1:N_staz){
    vector[C] log_probs = log(eta);
    for (j in 1:C){
      log_probs[j] += normal_lpdf(b[i] | m[j], s[j]);;
      }

    //Sampling from Discrete Distribution
    cluster_allocs[i] = categorical_rng(softmax(log_probs));
  }
}
```

## A.5. Multivariate Clustering

```
data {
    int<lower=0> N_obs;
    int<lower=0> N_miss;

    int<lower=0> p;
    int<lower=0> N_staz;
    int<lower=0> N_days;

    vector[N_obs] Y_obs;
    array[N_obs] int t_obs;
    array[N_obs] int m_obs;
    array[N_obs] int ID_obs;

    array[N_miss] int t_miss;
    array[N_miss] int m_miss;
    array[N_miss] int ID_miss;

    matrix[N_staz, p] X;
    vector[2] coord[N_staz];

    real omega;
```

```
    vector[N_days] day;

    int<lower=0> C;
}



parameters {
    vector[N_miss] Y_miss;

    vector<lower=0>[12] sigma_sq;

    vector[p] betas;

    matrix[N_staz, 5] abc;

    vector[N_staz] w;
    real<lower=0> rho;
    real<lower=0> alpha;

    matrix<lower=0>[C, 5] S;
    matrix[C, 5] M;
    vector<lower=0, upper=1>[C-1] v;
}



transformed parameters {
    vector<lower=0>[12] sigma = sqrt(sigma_sq);

    matrix[N_staz, N_days] ft;
    for(i in 1:N_staz){
      ft[i, ] = (rep_vector(abc[i, 5], N_days))';
      ft[i, ] += (abc[i, 1]*sin(omega*day) + abc[i, 2]*cos(omega*day) +
                  abc[i, 3]*sin(4*omega*day) + abc[i,4]*cos(4*omega*day))
      }
```

```
    cov_matrix[N_staz] H = gp_exp_quad_cov(coord, alpha, rho);

    vector[N_obs] mu_obs;
    for(i in 1:N_obs){
      mu_obs[i] =  ft[ID_obs[i], t_obs[i]];
      }
    mu_obs[1:N_obs] += X[ID_obs[1:N_obs]] * betas;
    mu_obs[1:N_obs] += w[ID_obs[1:N_obs]];

    vector[N_miss] mu_miss;
    for(i in 1:N_miss){
      mu_miss[i] =  ft[ID_miss[i], t_miss[i]];
      }
    mu_miss[1:N_miss] += X[ID_miss[1:N_miss]] * betas;
    mu_miss[1:N_miss] += w[ID_miss[1:N_miss]];



    // Stick Breaking
    vector<lower=0, upper=1> [C-1] cumprod_one_minus_v;
    cumprod_one_minus_v = exp(cumulative_sum(log1m(v)));

    simplex[C] eta;
    eta[1] = v[1];
    eta[2:(C-1)] = v[2:(C-1)] .* cumprod_one_minus_v[1:(C-2)];
    eta[C] = cumprod_one_minus_v[C-1];

    real param = 2.0;
}



model {
    sigma_sq ~ inv_gamma(3, 2);

    betas ~ normal(0, 1);

    rho ~ beta(3, 10);
```

```
    alpha ~ normal(0.3, 0.1);

    w ~ multi_normal(rep_vector(0, N_staz), H);

    Y_obs[1:N_obs] ~ normal(mu_obs[1:N_obs], sigma[m_obs[1:N_obs]]);

    Y_miss[1:N_miss] ~ normal(mu_miss[1:N_miss], sigma[m_miss[1:N_miss]])


    // Clustering
    for (i in 1:C){
       M[i, ] ~ normal(0, 10*sqrt(S[i, ])');
       S[i, ] ~ inv_gamma(4.5, 0.015);
    }
    v ~ beta(1, param);


    // Finite Mixture Model
    for (i in 1:N_staz){
        vector[C] lps = log(eta);
        for (j in 1:C){
            lps[j] += normal_lpdf(abc[i, :] | M[j, :], sqrt(S[j, :])');
            }
        target += log_sum_exp(lps);
        }
}



generated quantities  {
  vector[N_obs] log_lik;
  for (j in 1:N_obs) {
    log_lik[j] = normal_lpdf(Y_obs[j] | mu_obs[j], sigma[m_obs[j]]);
  }

  array[N_staz] int cluster_allocs;
  for(i in 1:N_staz){
```

```
    vector[C] log_probs = log(eta);
    for (j in 1:C){
      log_probs[j] += normal_lpdf(abc[i, :] | M[j, :], sqrt(S[j, :])'); }

    // Sampling from Discrete Distribution
    cluster_allocs[i] = categorical_rng(softmax(log_probs));
  }
}
```

# List of Figures

# List of Tables