POLITECNICO
MILANO 1863

Mathematical Engineering A.Y. 2022-23
Bayesian Statistics(052499) - Prof. Alessandra Guglielmi
Tutoring team: Raffaele Argiento, Sirio Legramanti, Lucia Paci

# BAYESIAN MIXTURE MODEL FOR ENVIRONMENTAL APPLICATION

Final Report

P. Bogani, P. Botta, S. Caresana, R. Carrara, G. Corbo, L. Mainini

## Contents

# 1 Introduction

The World Health Organization considers air pollution a major global environmental risk to human health. Pollutants have shown to be responsible for respiratory and cardiovascular diseases. Despite improvements over the past two decades, Europe's air quality remains poor in many places.
Our objective is to develop Bayesian-mixture-model-based clustering algorithms for environmental applications. Specifically, we focus our attention on PM10. Only in the EU in 2020, a total of 238,000 premature deaths were linked to exposure to particulate matter. Clustering environmental data may be useful for further studies, for instance aimed at analyzing their relations with specific diseases or human health problems.
In our study, we use two hierarchical linear regression mixed models that take into account level, trend, seasonal, and time-dependent components. A first-order auto-regressive process is used, in both, to model the temporal effect.
In the first model, a non-parametric prior is assumed for the joint distribution of the random effects and the coefficients related to trend and seasonality. The Poisson-Dirichlet process' discreteness will be exploited to cluster the data.
To provide additional insights, we introduced a second model. This higher-hierarchical-level clustering approach is based on the persistence of the time series and assumes a Dirichlet prior directly to the distribution of the auto-regressive parameters.

# 2 Data Exploration

With the term Particulate Matter (or PM) one refers to a collection of solid and liquid particles, with a wide variety of characteristics, dispersed in the atmosphere for sufficiently long times to undergo diffusion and transport phenomena. The sources of these particles may be natural (like soil erosion, volcanoes, pollen dispersal etc.) or anthropogenic (for example from industry, heating or vehicular traffic). It is therefore a very different pollutant from all others, presenting itself not as a specific chemical entity but as a mixture of particles with the most varied properties.

PM10 is the fraction of particles collected by a sorting system with an efficiency established by the standard (UNI EN12341/2001) and equal to 50% for the aerodynamic diameter of $10\mu m$. Atmospheric particulate matter has a major environmental impact on climate, water and soil contamination and, above all, on the health of living beings. For this reason, it is important to constantly monitor PM10 levels by means of control units located throughout the territory, so that critical concentration levels for health are not exceeded.

The data under consideration are collected from 162 monitoring stations between 2013 and 2019 in Northern Italy (European Environmental Agency) and for each station we have a daily measure of PM10 concentration. In Figure 1 is analyzed the PM10 concentration of 10 stations just to have an idea of the level and seasonality of the time series.

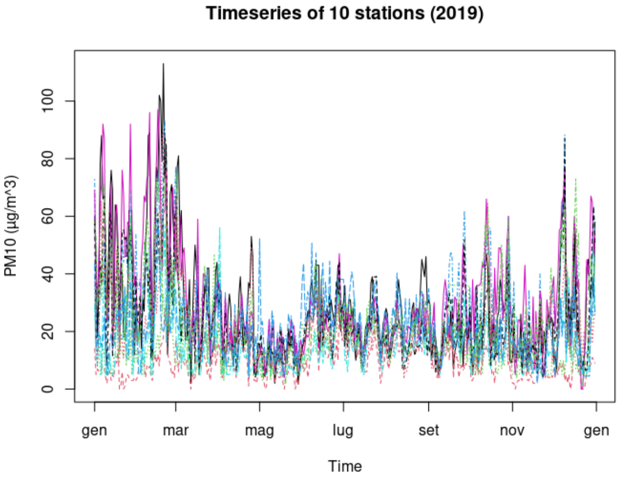


Figure 1: Concentration of 10 random stations from 2019

The European Environment Agency (EEA) has set two limit values for PM10: the PM10 daily mean value may not exceed 50 micrograms per cubic metre ($\mu m/m^3$) more than 35 times in a year and the PM10 annual mean value may not exceed 40 micrograms per cubic metre ($\mu m/m^3$).

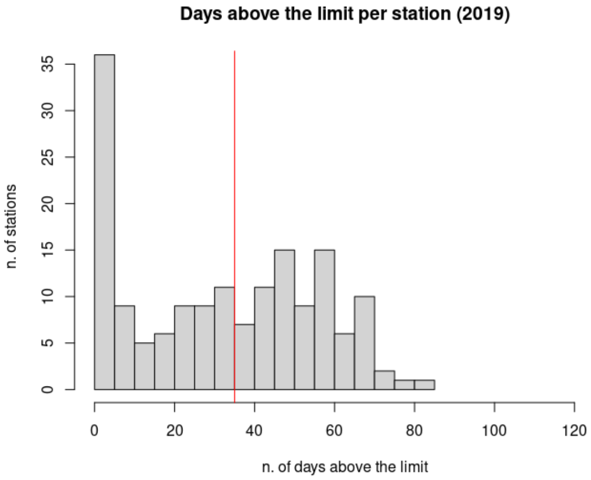


Figure 2: Number of stations above the daily limit in 2019

From the previous plot, we can say that 77 stations out of 162 overcame the limit, which is fixed at 35 days per year. In particular, some of them reached the maximum concentration per day for 80 days.

Now, we look at the mean of the PM10 concentration along all the stations, dividing the data for each years. We obtain the following plot:

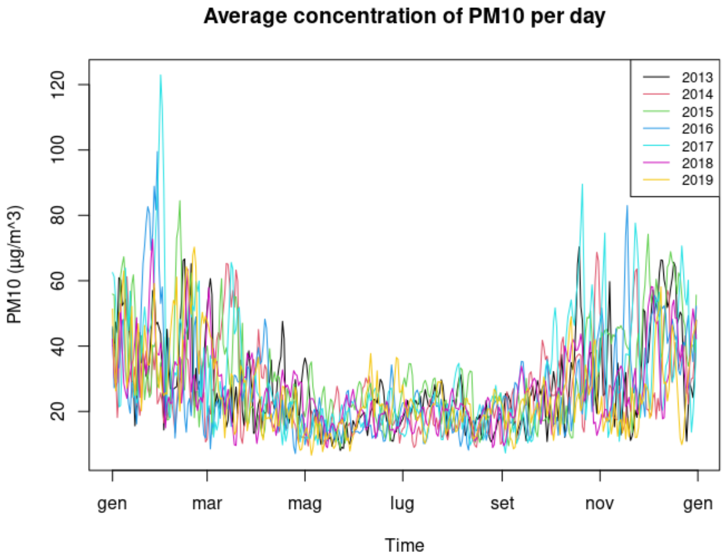


Figure 3: Average concentration of PM10 per day

We can say that all the analysed years behave similarly, and the average concentration of PM10 is visibly lower in the summertime rather than in the winter months. For this reason, we take into account the seasonality effect in both our models.

# 3 Models

## 3.1 Starting model

The first model we adopted was inspired by the one introduced by Nieto-Barajas and Contreras-Cristan[1] in the article *"Bayesian Non-parametric clustering for time series"*.

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, n$$
$$\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \ldots, \epsilon_{iT}) \sim \mathrm{N}_T\left(\mathbf{0}, \sigma^2_{\epsilon_i}\mathbf{I}\right),$$
$$\theta_{it} = \rho\theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{it} \sim \mathrm{N}\left(0, \sigma^2_\theta\right),$$

where $\mathbf{Z}$ and $\mathbf{X}$ are two design matrices of dimension $T \times p$ and $T \times d$ respectively. The $p \times 1$ dimensional vector $\boldsymbol{\alpha}_i$, the $d \times 1$ dimensional vector $\boldsymbol{\beta}_i$ and the $T \times 1$ dimensional vector $\boldsymbol{\theta}_i$ are parameters of the model such that $\boldsymbol{\eta}_i = (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$, but only $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$ will be considered for clustering. In our case the clustering is based on everything else rather than the level $\mu_i$ then we would take $\boldsymbol{\alpha}_i = \mu_i$ and $\boldsymbol{\beta}_i = (\boldsymbol{\omega}_i, \boldsymbol{v}_i)$, where $\boldsymbol{\omega}_i$ denotes a polynomial trend of the series and $\boldsymbol{v}_i$ denotes the seasonal component.
Finally, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})^T \sim \mathrm{N}_T\left(\mathbf{0}, \sigma^2_{\epsilon_i}\mathbf{I}\right)$ is the vector of measurement errors such that I is the identity matrix of dimension $T \times T$.

### 3.1.1 Prior distributions

We adopted a generalization of the Dirichlet process as prior associated to the distribution of the coefficients used for clustering $\boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$

$$\boldsymbol{\gamma}_i \mid G \overset{\text{iid}}{\sim} G, \text{ for } i = 1, \ldots, n \quad \text{ with } G \sim \mathcal{PD}\left(a, b, G_0\right),$$

The almost-certain discretization of the Poisson-Dirichlet Process induces a clustering among the data, grouping the observations with the same latent variables $\boldsymbol{\gamma}_i$ sampled from the DP, the so-called **ties**.

$$G_0(\boldsymbol{\gamma}) = G_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathrm{N}_d\left(\boldsymbol{\beta} \mid \mathbf{0}, \boldsymbol{\Sigma}_\beta\right) \times \mathrm{N}_T(\boldsymbol{\theta} \mid \mathbf{0}, \mathbf{R}),$$

$$R_{jk} = \sigma^2_\theta \rho^{|j-k|},$$
$$\boldsymbol{\Sigma}_\beta = \operatorname{diag}\left(\sigma^2_{\beta 1}, \ \ldots \ , \sigma^2_{\beta d}\right).$$

The number of clusters m (unique values in $\boldsymbol{\gamma}_1, \ \ldots \ , \boldsymbol{\gamma}_n)$) is determined by the parameters $(a, b)$. Larger values of either $a$ or $b$, within the valid ranges, produce a larger $m$.
Fixing $i$, we recall that $\boldsymbol{\gamma}_{-i}$ denotes the set of all $\gamma$ 's excluding the $i^{th}$ element. $\gamma^*_{j,i}$'s denote the unique values in $\boldsymbol{\gamma}_{-i}$, each occurring with frequency $n^*_{j,i}$, with $j = 1, \ldots, m_i$, where $m_i$ represents the number of unique values in $\boldsymbol{\gamma}_{-i}$.
The joint distribution of the $\gamma_i$ 's is characterized by a generalized Polya urn

mechanism with conditional distribution that depends on the density $g_0$ associated to $G_0$ and given by

$$f\left(\gamma_i \mid \gamma_{-i}\right)=\frac{b+am_i}{b+n-1}g_0\left(\gamma_i\right)+\sum_{j=1}^{m_i}\frac{n_{j,i}^*-a}{b+n-1}\delta_{\gamma_{j,i}^*}\left(\gamma_i\right),$$

$$\sigma_{\epsilon_i}^2\sim\mathrm{IGa}\left(c_0^\epsilon,c_1^\epsilon\right),\quad\sigma_{\beta_j}^2\sim\mathrm{IGa}\left(c_0^\beta,c_1^\beta\right),\quad\sigma_{\alpha_k}^2\sim\mathrm{IGa}\left(c_0^\alpha,c_1^\alpha\right),$$

$$f\left(\rho,\sigma_\theta^2\right)\propto\frac{\sqrt{1+\rho^2}}{1-\rho^2}\left(\sigma_\theta^2\right)^{-1},$$

$$f(a)=\pi I_{\{0\}}(a)+(1-\pi)\,\mathrm{Be}\left(a\mid q_0^a,q_1^a\right),$$

$$f(b\mid a)=\mathrm{Ga}\left(b+a\mid q_0^b,q_1^b\right).$$

### 3.1.2 First results

Applying the first model to our data we obtained the following clusters:

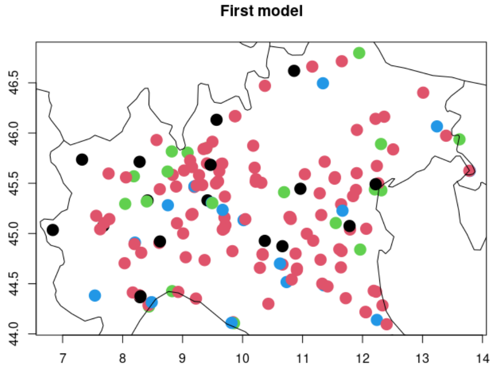


Figure 4: Results of the first model

We could have expected a poor outcome because there isn't much difference in level or trend between the stations, as can be seen in the matplot of the PM10 concentration in Northern Italy.

We also tested the first model on more heterogeneous data considering stations throughout different European regions, and the algorithm was able to divide them into groups. However, considering only data from the Po Valley, this model provides limited information, as you can see above. This induced us to proceed with a second model based on a new approach, clustering on a higher hierarchical level, the parameters of the auto-regressive process $\rho$ and $\sigma_\theta^2$.

## 3.2 Proposed model

The idea was to develop a clustering technique based on the **persistence** of the pollutant, investigating the correlation between consecutive observations in the same station. In order to better understand the meaning of persistence, we can say that a persistent series is one in which the variable's value at a given time has a strong correlation with its previous value. To implement this idea, we decided to cluster our data adopting a Dirichlet process as prior for the distribution of the parameters of the auto-regressive process of $\boldsymbol{\theta}_i$ ($\rho_i$ and $\sigma_i$).
In Figure 5, you can see the distribution of the $\boldsymbol{\rho}$'s, obtained by fitting an auto-regressive model of order 1 using the function `arima` in R.

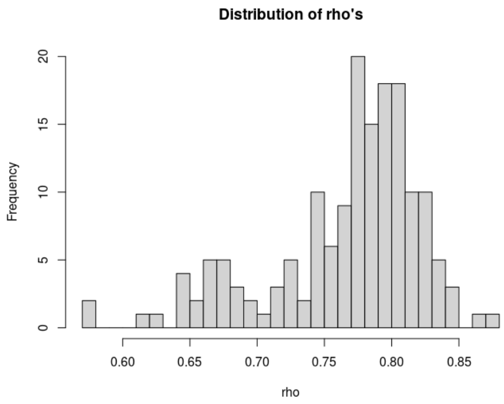


Figure 5: $\boldsymbol{\rho}$'s obtained by fitting an AR(1) on PM10 concentrations

Given this distribution, we may be able to obtain a reasonable clustering.
In Figure 6, you can see the distribution of the $\rho_i$ for each station considered. The green points are the stations with a low value of $\rho$, so a low persistence, while the red points are the ones with persistent time series. In particular, we can see that in the Po valley the persistence is high while in the zones near the sea or in a more mountaineer city the value of the persistence is smaller.

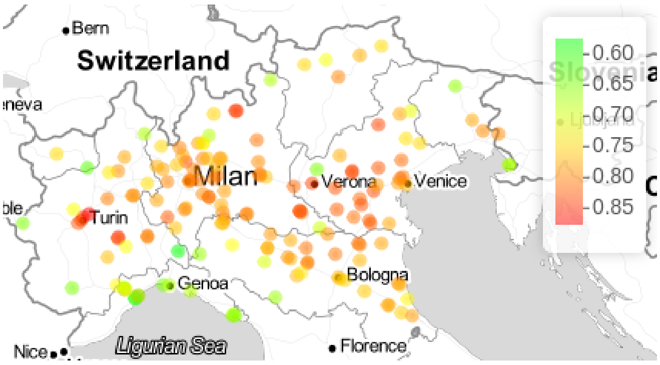


Figure 6: $\boldsymbol{\rho}$'s distribution in Northern Italy

Considering this approach, we decided to implement from scratch a second model, using some specifications from the previous one, but most importantly setting as prior for the distribution of the parameters $\rho_i$ and $\sigma_i$, a Dirichlet process. The definition of $\mathbf{y}_i$ is similar to the first one, but the $\boldsymbol{\beta}$'s are not present since they represented the covariates on which the cluster is made. Since the cluster is now not made on any covariate, but on a higher hierarchical level, they are all represented by the $\boldsymbol{\alpha}$'s, the variables on which we don't cluster. Moreover, from now on $\sigma_i^2$ represents the $\sigma_\theta^2$ of the $i$-th station.

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \quad \text{for } i = 1, 2, \dots n,$$

$$\begin{gathered}
\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \dots, \epsilon_{iT}) \sim \mathrm{N}_T\left(\mathbf{0}, \sigma_{\epsilon_i}^2 \mathbf{I}\right), \\
\theta_{it} = \rho_i \theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{it} \sim \mathrm{N}\left(0, \sigma_i^2\right), \\
\gamma_i = (\rho_i \sigma_i^2), \\
\boldsymbol{\theta}_i \sim \mathrm{N}_T(\mathbf{0}, \mathbf{R}(\gamma_i)) \quad \text{with } R_{ls}(\gamma_i) = \sigma_i^2 \rho_i^{|l-s|} \text{ and } l, s = 1, \dots, T.
\end{gathered}$$

### 3.2.1 Prior distributions

$$\begin{gathered}
\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathrm{N}_p\left(\mathbf{0}, \boldsymbol{\Sigma}_\alpha\right) \quad \text{with } \boldsymbol{\Sigma}_\alpha = \mathrm{diag}(\sigma_{\alpha_1}, \dots, \sigma_{\alpha_p}), \\
\sigma_{\epsilon_i}^2 \sim \mathrm{IGa}\left(c_0^\epsilon, c_1^\epsilon\right), \quad \sigma_{\alpha_k}^2 \sim \mathrm{IGa}\left(c_0^\alpha, c_1^\alpha\right), \\
\gamma_i \mid P \overset{\text{iid}}{\sim} P, \text{ for } i = 1, \dots, n \text{ with } P \sim \mathrm{Dir}\left(a^P, P_0\right), \\
p_0\left(\gamma_i\right) = p_0(\sigma_i^2) \times p_0(\rho_i) \\
p_0(\sigma_i^2) = \mathrm{IGa}(a, b) \qquad p_0(\rho_i) = \mathrm{Beta}(c, d),
\end{gathered}$$

for $i = 1, \dots, n$.

### 3.2.2 Full conditionals

In the following computations, we will make use of some properties related to multivariate Gaussian distribution that we report here.

**Proposition A.** Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right), \\
p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right),
\end{aligned}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right), \tag{1}$$

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right), \tag{2}$$

where

$$\mathbf{\Sigma} = (\Lambda + A^T LA)^{-1}.$$

Proof of the previous properties can be found at Pag. 91-92 of Christopher Bishop *"Pattern recognition and machine learning"* [2].

If we let $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}_1', \dots, \boldsymbol{\alpha}_n'), \boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_n')$, and $\boldsymbol{\sigma}_\epsilon' = \left(\sigma_{\epsilon_1}^2, \dots, \sigma_{\epsilon_n}^2\right)$ then the likelihood function is given by

$$f\left(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\sigma}_\epsilon\right) = \prod_{i=1}^{n} \mathrm{N}_T\left(\mathbf{y}_i \mid \mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i, \sigma_{\epsilon_i}^2 \mathbf{I}\right).$$

With this notation, we denote that $\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\theta}_i, \boldsymbol{\sigma}_{\epsilon_i}$ are distributed as an $T$-variate multivariate normal distribution with mean vector $\mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i$ and variance-covariance matrix $\sigma_{\epsilon_i}^2 \mathbf{I}$. We will continue using this notation from now on.

Using Property (1) of Proposition A.:

$$f\left(\mathbf{y}_i \mid \boldsymbol{\alpha_i}, \sigma_{\epsilon_i}^2, \mathbf{\Sigma}_\alpha, \gamma_i\right) = \mathrm{N}_p\left(\mathbf{y}_i \mid Z\boldsymbol{\alpha}_i, \mathbf{W}_i\right), \tag{3}$$

with matrices $\mathbf{W}_i = \sigma_{\epsilon_i}^2 \mathbf{I} + \mathbf{R}(\gamma_i)$.
The conditional posterior distribution of $\boldsymbol{\alpha}_i$ becomes

$$f\left(\boldsymbol{\alpha}_i \mid \mathbf{y}, \sigma_{\epsilon_i}^2, \mathbf{\Sigma}_\alpha\right) = \mathrm{N}_p\left(\boldsymbol{\alpha}_i \mid \boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha\right), \tag{4}$$

for $i = 1, \dots, n$, where $\boldsymbol{\mu}_\alpha = \mathbf{V}_\alpha \mathbf{Z}' \mathbf{W}_i^{-1} \mathbf{y}_i$ and $\mathbf{V}_\alpha = \left(\mathbf{Z}' \mathbf{W}_i^{-1} \mathbf{Z} + \mathbf{\Sigma}_\alpha^{-1}\right)^{-1}$ with matrices

$$\mathbf{W}_i = \sigma_{\epsilon_i}^2 \mathbf{I} + \mathbf{R}(\gamma_i),$$

of dimensions $T \times T$.
The conditional posterior distribution for the variances $\sigma_{\epsilon_i}^2, i = 1, \dots n$, and $\sigma_{\alpha_k}^2, k = 1, \dots, p$ given the data and the rest of the parameters are all conditionally conjugate. The conditional posterior distribution for $\sigma_{\epsilon_i}^2$ has the form

$$f\left(\sigma_{\epsilon_i}^2 \mid \mathbf{y}, -\right) = \mathrm{IGa}\left(\sigma_{\epsilon_i}^2 \mid c_0^\epsilon + \frac{T}{2}, c_1^\epsilon + \frac{1}{2}\mathbf{M}_i'\mathbf{M}_i\right), \tag{5}$$

where $\mathbf{M}_i = \mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \boldsymbol{\theta}_i$, for $i = 1, \dots, n$.
The conditional posterior distribution for $\sigma_{\alpha_j}^2$ has the form

$$f\left(\sigma_{\alpha_j}^2 \mid \mathbf{y}, -\right) = \mathrm{IGa}\left(\sigma_{\alpha_j}^2 \mid c_0^\alpha + \frac{n}{2}, c_1^\alpha + \frac{1}{2}\sum_{i=1}^{n} \alpha_{ij}^2\right), \tag{6}$$

for $j = 1, 2, \dots, p$.
The conditional posterior distribution for $\boldsymbol{\theta_i}$ has the form

$$f\left(\boldsymbol{\theta}_i \mid \mathbf{y}, \sigma_{\epsilon_i}^2, \mathbf{R}(\gamma_i)\right) = \mathrm{N}_T\left(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_\theta, \mathbf{S}_\theta\right), \tag{7}$$

for $i = 1, \ldots, n$, where $\mathbf{S}_\theta = \left( \left( \sigma^2_{\epsilon_i} \mathbf{I} \right)^{-1} + \mathbf{R}(\gamma_i)^{-1} \right)^{-1}$ and $\boldsymbol{\mu}_\theta = \mathbf{S}_\theta \left( \sigma^2_{\epsilon_i} \mathbf{I} \right)^{-1} (\mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i)$.

Finally, to obtain the full conditional distribution of $\gamma$ we adopted an algorithm to sample from Dirichlet process mixture models, adapting the Algorithm 8 proposed by Neal [3]. Let $\mathbf{c} = \{c_1, \ldots, c_n\}$ be the vector of the "latent classes" to which each observation $y_i$ is associated. We denote the unique values in $\boldsymbol{\gamma}$ with $\gamma^*$'s. For each class $c_i$, the parameters $\gamma^*_{c_i}$ determine the distribution of observations from that class.
We can now update $c_i$ by sampling from its conditional distribution given $\theta_i$ and the parameters of all existing and empty clusters.
Specifically, for $i = 1, \ldots, n$, let $m_i$ the number of distinct $c_i$ for $j \neq i$ and $h = m_i + n_{\text{aux}}$. If $c_i = c_j$ for some $j \neq i$, draw values independently from $P_0$ for those $\gamma^*_c$ for which $m_i < c \leq h$ . If $c_i \neq c_j$ for all $j \neq i$, let $c_i$ have the label $m_i + 1$, and draw values independently from $P_0$ for those $\gamma^*_c$ for which $m_i + 1 < c \leq h$. Draw a new value for $c_i$ from $\{1, \ldots, h\}$ using the following probabilities:

$$\mathbb{P}\left[c_i = c \mid \boldsymbol{c}_{-i}, \boldsymbol{\theta}_i, \gamma^*_1, \ldots, \gamma^*_h\right] = \begin{cases} b \cdot \frac{n^*_{c,i}}{n-1+a^p} \cdot f\left(\boldsymbol{\theta}_i \mid \gamma^*_{c_i}, -\right), & \text{for } 1 \leq c \leq m_i \\ b \cdot \frac{a^p/n_{\text{aux}}}{n-1+a^p} \cdot f\left(\boldsymbol{\theta}_i \mid \gamma^*_{c_i}, -\right), & \text{for } m_i < c \leq h \end{cases}$$

where $n^*_{c,i}$ is the number of $c_j$ for $j \neq i$ that are equal to $c$, and $b$ is the appropriate normalizing constant. The values of $\gamma^*_{c_i}$ where $m_i < c \leq h$ are drawn independently from $P_0$.

## 3.3 Bayesian model based clustering analysis

At each iteration, the Gibbs sampler produces an implicit clustering of the parameters $\gamma = (\gamma_1, \gamma_2, \ldots \gamma_n)$, for which each $\gamma_i$ (partially) characterizes the time series $\mathbf{y}_i$, thus inducing a clustering of the time series $\mathbf{y}_i$ 's.

MCMC clustering information can be summarized by counting the number of times (iterations) that two parameters, say $\gamma_i$ and $\gamma_j$, belong to the same cluster. This method is better than only registering the cluster membership because it allows to avoid the label-switching problem. With this information, it is possible to build a similarity matrix containing the relative frequencies of pairwise clustering corresponding to the event that $\mathbf{y}_i$ and $\mathbf{y}_j$ share the same $\gamma$ parameter values, that is $\gamma_i = \gamma_j$.

Under the Bayesian paradigm, the canonical approach is to introduce a loss function $L(\mathbf{c}, \hat{\mathbf{c}})$ and then choose the clustering $\hat{\mathbf{c}}$ that minimizes the posterior expectation of the chosen loss function. The expectation of this function is approximated using posterior samples. For instance, the binder loss function is defined as:

$$L_{\text{Binder}}\left(\mathbf{c}, \hat{\mathbf{c}}\right) = \sum_{i<j} \left(a \cdot \mathbb{I}\left\{c_i = c_j\right\} \mathbb{I}\left\{\hat{c}_i \neq \hat{c}_j\right\} + b \cdot \mathbb{I}\left\{c_i \neq c_j\right\} \mathbb{I}\left\{\hat{c}_i = \hat{c}_j\right\}\right).$$

As proposed by Wade [4] and Dahl [5], we choose the partition minimizing the posterior of the variation of information loss function developed by Meila. It can be shown that under the VI, the optimal partition $\hat{\mathbf{c}}^*$ is the following:

$$\begin{aligned}\hat{\mathbf{c}}^* =& \underset{\hat{\mathbf{c}}}{\operatorname{argmin}} \mathbb{E}\left(L_{VI}(\mathbf{c}, \hat{\mathbf{c}}) \mid \mathcal{D}\right) \\ =& \underset{\hat{\mathbf{c}}}{\operatorname{argmin}} \sum_{i=1}^{n} \log_2 \left( \sum_{j=1}^{n} \mathbb{I}\left(\hat{c}_j = \hat{c}_i\right) \right) \\ & - 2 \sum_{i=1}^{n} \mathbb{E}\left( \log_2 \left( \sum_{j=1}^{n} \mathbb{I}\left(c_j = c_i\right) \mathbb{I}\left(\hat{c}_j = \hat{c}_i\right) \right) \mid \mathcal{D} \right),\end{aligned}$$

where $\mathcal{D}$ represents data and $\mathbf{c}$ is the true cluster partition. In particular, we use the SALSO algorithm developed by Dahl which implements a greedy, stochastic search given the number of desired clusters [5].

### 3.4 Algorithm scheme

The implemented algorithm employs the following scheme:

- Construction of the design matrices
- Initialization of the parameters
- Beginning of **Gibbs sampling**
    1. Sample $\alpha$ from its full conditional
    2. Sample $\theta$ from its full conditional
    3. Sample $\gamma' s = (\rho, \sigma^2)$
    4. Sample $\sigma_\epsilon^2$ from its full conditional
    5. Sample $\sigma_\alpha^2$ from its full conditional
- End of Gibbs sampling
- Determining which cluster configuration minimizes a specific posterior loss

## 4 Simulations

### 4.1 Synthetic data generation

Recalling the structure of the model (letter $i$ denotes the $i$-th station):

$$\begin{gathered}\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \ldots, \epsilon_{iT}) \sim \mathrm{N}_T\left(\mathbf{0}, \sigma_{\epsilon_i}^2 \mathbf{I}\right), \\ \theta_{it} = \rho_i \theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{i1} \sim \mathrm{N}\left(0, \sigma_i^2\right),\end{gathered}$$

The unknown parameters are: $\boldsymbol{\alpha}_i$, $\sigma^2_{\epsilon_i}$, $\rho_i$ and $\sigma_i$

We suppose that stations belonging to the same cluster are characterized by the same value of $\rho_i$ and $\sigma_i$. For example, if stations 1 and 2 come from the same cluster $\rho_1 = \rho_2$ and $\sigma_1 = \sigma_2$.

To verify the correctness of the algorithm we set some random values for the unknown parameters and we generated $\theta_{it}$ and $\boldsymbol{y}_i$ accordingly.

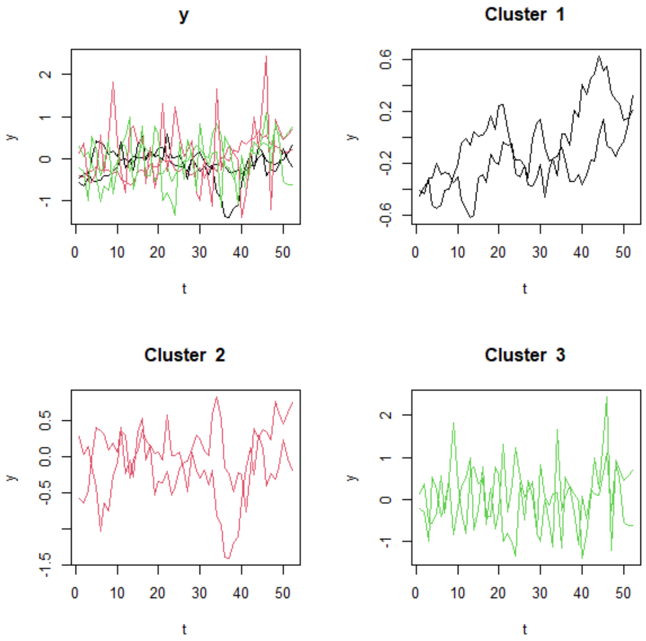


Figure 7: Example of synthetic data

The above data are obtained simulating 6 stations equally divided in three clusters. $\alpha_i = 0$ for each i, and $\sigma_{\epsilon_i} = 0.0001$ for each i (i.e. low noise)

1. The first cluster has $\rho = 0.9$ and $\sigma = 0.1$
2. The second cluster has $\rho = 0.9$ and $\sigma = 0.5$
3. The third cluster has $\rho = 0.1$ and $\sigma = 0.5$

The generated data in this small example follow our expectations: stations from the first cluster have a strong persistence, while stations from the third cluster are practically white noises.

## 4.2 Posterior sampling

To verify the correctness of the posteriors we used a larger synthetic dataset with 45 stations and three clusters (with the same $\rho_i$ and $\sigma_i$ as before). This time, $\sigma_{\epsilon_i} = 0.1$ and $(\boldsymbol{\alpha}_i)_4 = 1$ for each $i$. This induces a seasonality and a higher noise.

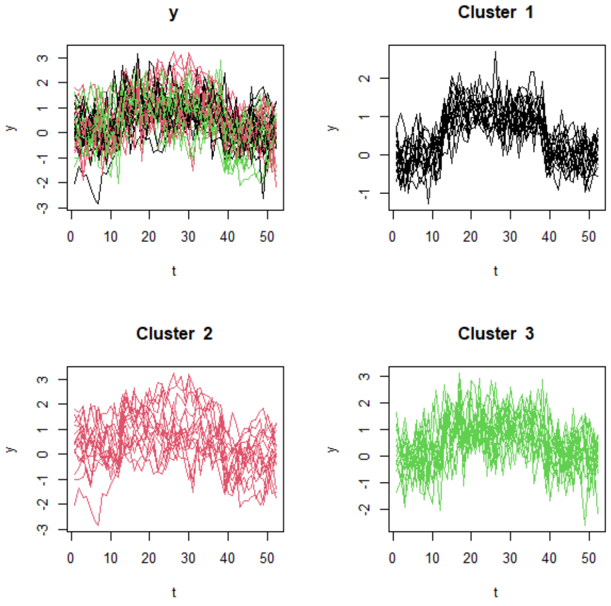
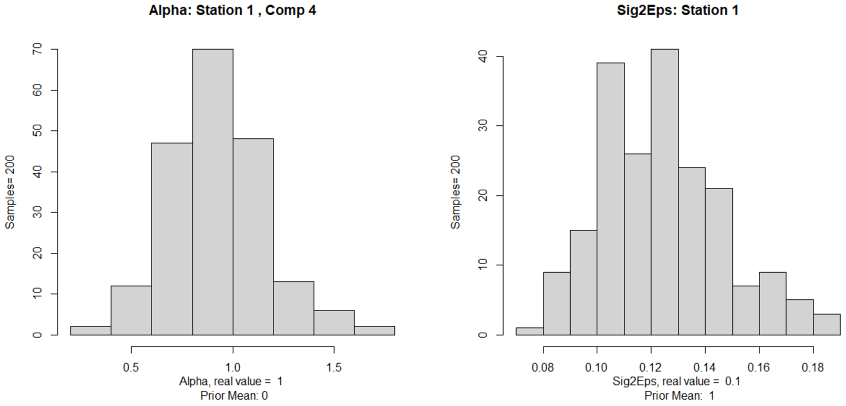


Figure 8: Synthetic data used for posterior sampling

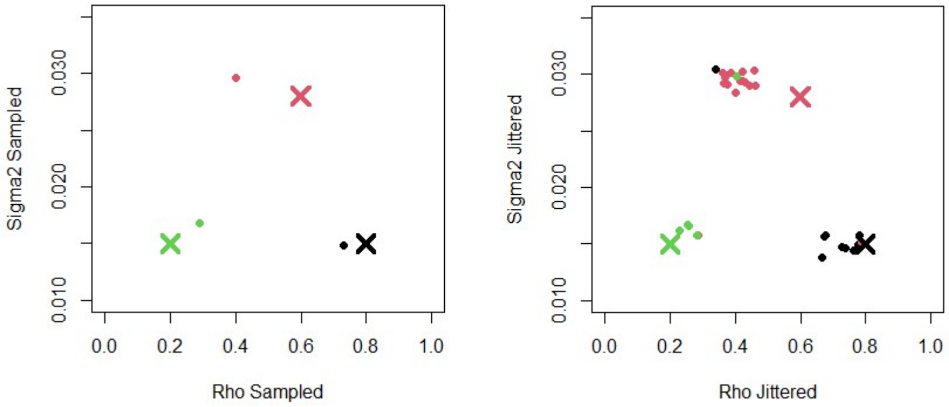
We executed the algorithm on this data and we plotted some posterior samples from $(\boldsymbol{\alpha}_1)_4$ and $\sigma_{\epsilon_1}$. Of course, we could also choose another station different from the first one.



As we said, the real value of $(\boldsymbol{\alpha}_1)_4$ which generated the data is 1. So, we expect a posterior mean approximately equal to 1. As you can see, the majority of our posterior samples vary between 0.5 and 1.5. The mass is not centered in 1 but it is slightly shifted to the left. That's correct considering that the posterior mean is also affected by the prior (mean), which is equal to 0. Similarly, the real value of $\sigma_{\epsilon_1}$ is 0.1 but the posterior samples are slightly shifted to the right due to the prior (mean) equals to 1. Thus, our posteriors are perfectly able to estimate the real parameters. A similar check can be executed for $\theta_{it}$, $\rho_i$, $\sigma_i$.

## 4.3 Testing

Finally, we applied the algorithm on a synthetic dataset with 3 clusters with dimensions 9,15,5 and $\rho$ respectively equal to 0.8, 0.6 and 0.2 and $\sigma^2 = 0.015, 0.028, 0.015$. To generate the data, we impose a small $\sigma^2_\epsilon$. In the figure below, you can see the results. We have decided here to select the cluster structure, choosing the iteration minimizing the Binder loss function with respect to the known real cluster we created.



The points in the figures represent the sampled values of $\rho$ and $\sigma$, and as expected we obtained 3 clusters with values acceptably similar to the real and original value of the $\rho$ and $\sigma$ of the 3 clusters (plotted as $\times$). In the second figure are represented each observation with a jittered value of $\rho$ and $\sigma^2$, just for visual purposes.

At last, we plotted the similarity matrix of the cluster obtained, and we can indeed denote from the dendrogram the formation of the three clusters.

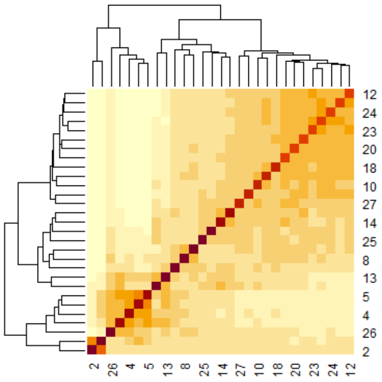


Figure 9: Similarity matrix

# 5 Case study

The last step was to apply the proposed model to the PM10 data described above. Due to the high computational workload, the results shown are obtained on a reduced version of the dataset.
In this case we adopted the greedy, stochastic search approach, selecting the cluster structure minimizing the VI loss function and setting the desired number of clusters to 4 using the R package `salso`.

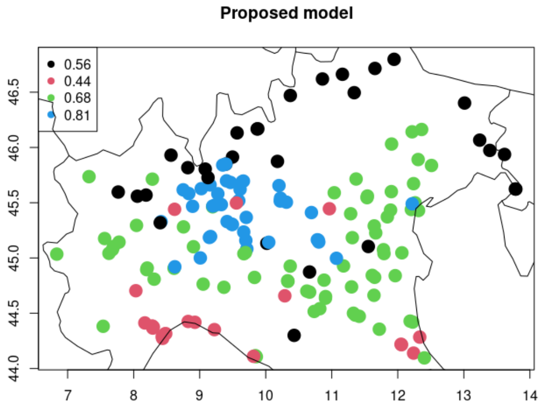
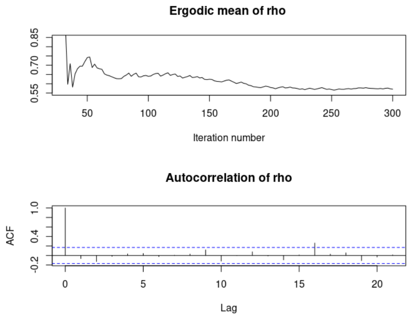
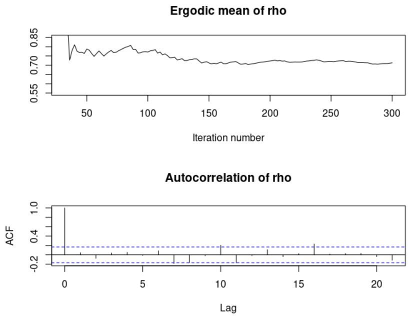


Figure 10: Results of the proposed model

We can note that the cluster obtained is comparable to Figure 6, and we can see that the four clusters obtained can be divided into 4 distinct natural regions. The Milan area has the highest persistence cluster, and the cities of the Po valley are home to another significant cluster. The other two clusters, with less persistence, are in the Genoa or marine cities zone and in the stations of the Alps. In fact, a search of the literature [6] [7] revealed studies analysing PM10 concentration that confirmed that the persistence is often higher in the proximity of urban areas and dry zones while it is lower in greener areas and more breezy regions.
We've then decided to compare the convergence of the ergodic mean of $\rho$ of two stations belonging to different clusters.



Station of the "red" cluster



Station of the "blue" cluster

We can note how indeed the two ergodic mean converge to two different values, with the Milan area station persistence higher than the one in Genoa zone. Finally, we plotted the similarity matrix of the cluster obtained:

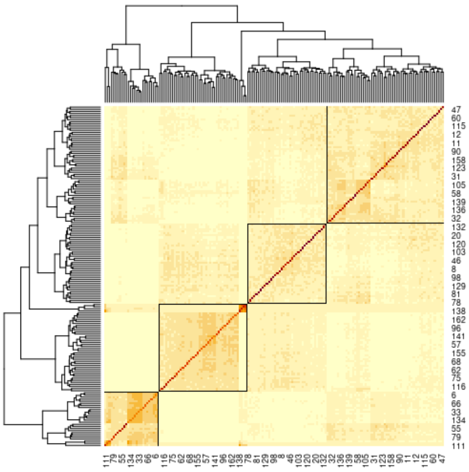


Figure 11: Similarity matrix of the proposed model

We can indeed denote the 4 clusters in the left-bottom corner, center and right-up corner. Figure 12 shows the distribution of the number of clusters obtained by the algorithm after 300 iterations.

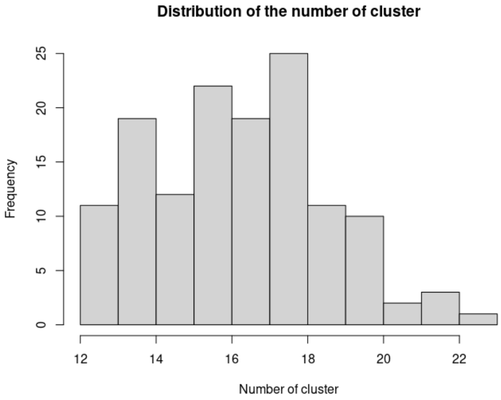


Figure 12: Distribution of the number of cluster

We can see that even if the number of stations is high (162), the minimum number of clusters obtained by the algorithm is still high (12). Deeper analysis and tuning of the hyper-parameters, number of iterations or the introduction of an acceleration step (in case we assume different priors) could be useful to have a histogram of the number of clusters more centered in a small value than the one obtained.

# 6 Conclusions

This project aimed to develop a Bayesian-model-based clustering algorithm for environmental data. We tested two multi-hierarchical linear mixture models which incorporate a first-order auto-regressive process to model the temporal effect. Clusters are obtained thanks to the non-parametric prior's discreetness. In the first model, a Poisson-Dirichlet prior is assumed for the joint distribution of the random effects and the coefficients related to trend and seasonality. This algorithm is especially effective if there is variability between group averages and thus allows to detect particularly heterogeneous groups as environmental zones with particularly distant average pollution levels. However, Northern Italy does not have pollution levels that far apart, and the first model does not provide particular insights. For this reason, we developed a second model based on persistence. In this case, we will adopt a Dirichlet Process for the prior of the distribution of the auto-regressive parameters. Cluster's results show that the persistence is often higher in the proximity of urban areas and dry zones while it is lower in areas with more vegetation or breezy regions (for instance closer to sea).

To conclude, the possible improvements for this project are multiple, for instance, the introduction of another non-parametric prior or a spatial component to better cluster stations close to each other. An idea could be to introduce a spatial product partition model [8] as prior for the parameters of the auto-regressive process $\gamma$ to spatially cluster on a higher hierarchical level. Another option could be to consider the data as areal data of a zone and propose to model the density of each area through a finite mixture of Gaussian distributions[9].

Furthermore, deeper analysis and tuning of the hyper-parameters or the introduction of an acceleration step (in case we assume different priors) could be useful to have a histogram of the number of clusters more centred in a small value than the one obtained (Figure 12).

# References

[1] Luis E Nieto-Barajas and Alberto Contreras-Cristán. 'A Bayesian nonparametric approach for time series clustering'. In: *Bayesian Analysis* 9.1 (2014), pp. 147–170.

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[3] Radford M. Neal. 'Markov Chain Sampling Methods for Dirichlet Process Mixture Models'. In: *Journal of Computational and Graphical Statistics* 9.2 (2000), pp. 249–265.

[4] Sara Wade and Zoubin Ghahramani. 'Bayesian cluster analysis: Point estimation and credible balls (with discussion)'. In: (2018).

[5] David B. Dahl, Devin J. Johnson and Peter Müller. 'Search Algorithms and Loss Functions for Bayesian Clustering'. In: *Journal of Computational and Graphical Statistics* 31.4 (2022), pp. 1189–1201.

[6] M. Meraz et al. 'Statistical persistence of air pollutants (O3,SO2,NO2 and PM10) in Mexico City'. In: *Physica A: Statistical Mechanics and its Applications* 427 (2015), pp. 202–217. ISSN: 0378-4371. DOI: `https://doi.org/10.1016/j.physa.2015.02.009`. URL: `https://www.sciencedirect.com/science/article/pii/S0378437115001065`.

[7] Carlos Zafra, Yenifer Ángel and Eliana Torres. 'ARIMA analysis of the effect of land surface coverage on PM10 concentrations in a high-altitude megacity'. In: *Atmospheric Pollution Research* 8.4 (2017), pp. 660–668.

[8] Garritt L Page and Fernando A Quintana. 'Spatial product partition models'. In: (2016).

[9] Mario Beraha et al. 'Spatially dependent mixture models via the Logistic Multivariate CAR prior'. In: *Spatial Statistics* 46 (2021), p. 100548.