# BAYESIAN MIXTURE MODEL FOR ENVIRONMENTAL APPLICATION

-Lasya Sri Nidumukkala

# CONTENTS

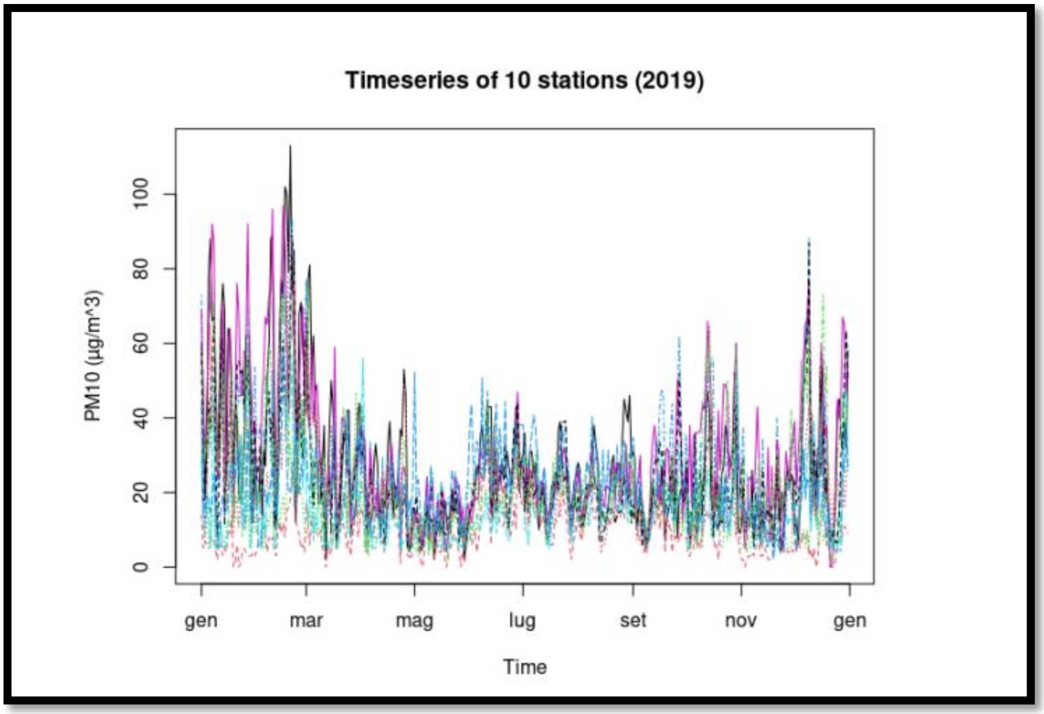
## Introduction:

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.

## Data Set , Source and Data Exploration:

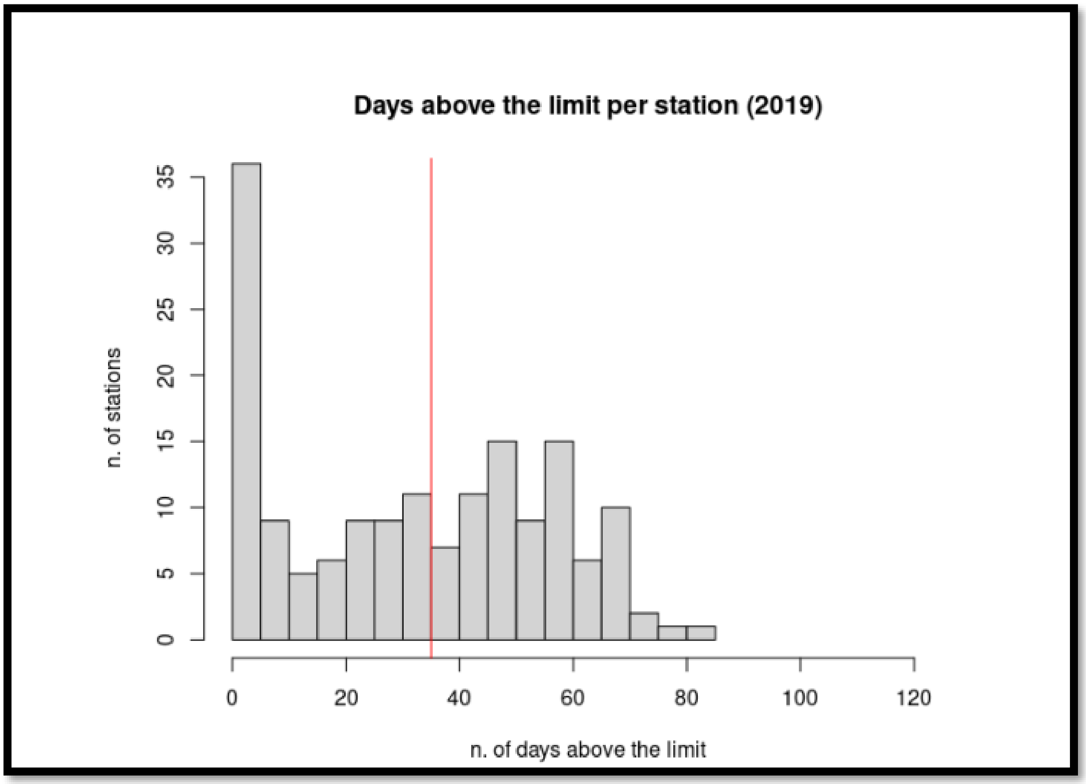
With the term Particulate Matter (or PM) refers to a collection of solid and liquid particles, with a wide variety of characteristics, dispersed in the atmosphere for sufficiently long times to undergo diffusion and transport phenomena. The sources of these particles may be natural (like soil erosion, 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 10µm. 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. Analyzed the PM10 concentration of 10 stations just to have an idea of the level and seasonality of the time series.



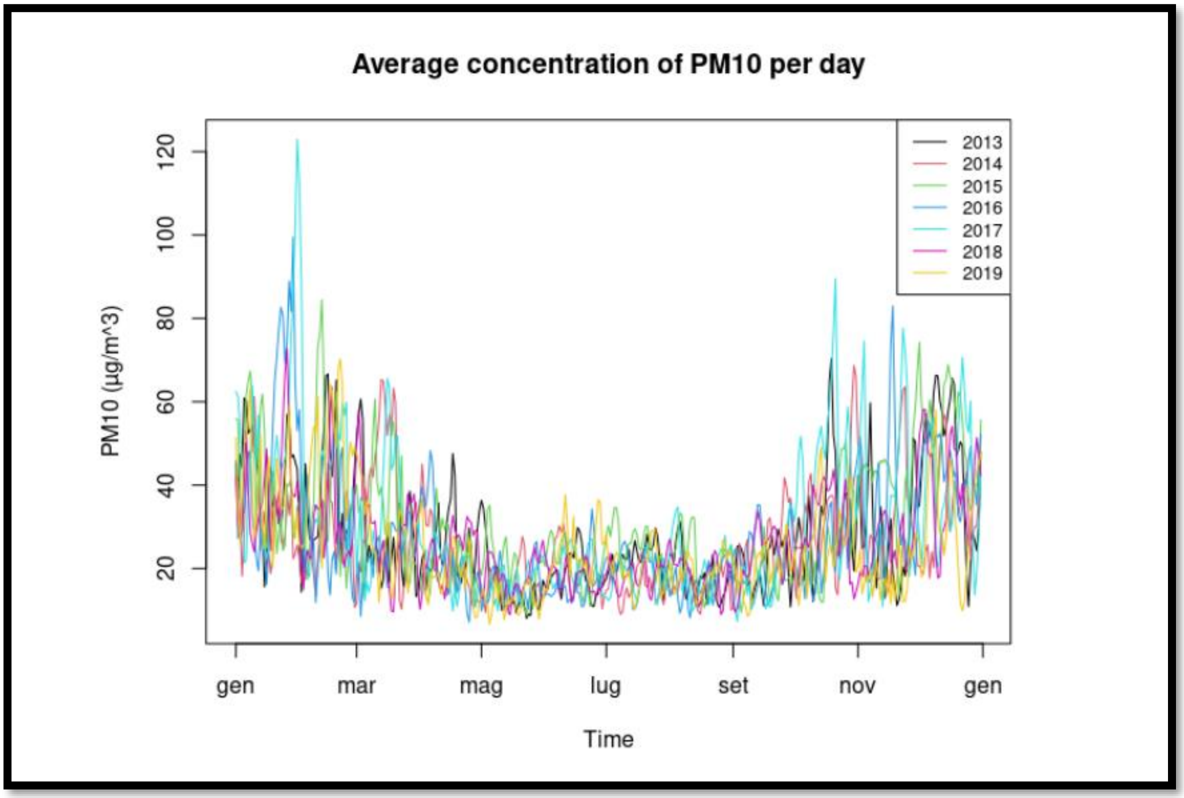
***Concentration of 10 random stations from 2019***

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 (µm/m3 ) more than 35 times in a year and the PM10 annual mean value may not exceed 40 micrograms per cubic meter (µm/m3)



***Number of stations above the daily limit in 2019***

From above plot we can infer 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. Looking at the mean of the PM10 concentration along all the stations, dividing the data for each years. We obtain the following plot:



The PM10 concentration along all the stations, dividing the data for each years. 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 took into account the seasonality effect in both our models.

## Starting model

The first model we adopted was inspired by the one introduced by Nieto-Barajas and Contreras-Cristanin the article ”Bayesian Non-parametric clustering for time series”.

In this model, non-parametric priors and the Dirichlet-Poisson process are used which are interconnected. The non-parametric priors allowed for flexible modeling of the distributions of trend and seasonality coefficients without assuming a specific parametric form. The Dirichlet-

Poisson process, a specific kind of non-parametric prior, is used because of its discrete and clustering properties. It effectively groups similar observations, aiding in the identification of latent structures within the time series data. This combination is key in analyzing complex environmental data like PM10 concentrations, where standard parametric models might be too restrictive or inaccurate.

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

where Z and X are two design matrices of dimension T ×p and T ×d respectively. The p × 1 dimensional vector αi , the d × 1 dimensional vector βi and the T × 1 dimensional vector θi are parameters of the model such that ηi = (αi , βi , θi), but only βi and θi will be considered for clustering. In our case the clustering is based on everything else rather than the level µi then we would take αi = µi and βi = (ωi , vi), where ωi denotes a polynomial trend of the series and vi denotes the seasonal component. Finally, ϵi = (ϵi1, . . . , ϵiT ) T ∼ NT 0, σ2 ϵi I  is the vector of measurement errors such that I is the identity matrix of dimension T ×T.

We adopted a generalization of the Dirichlet process as prior associated to the distribution of the coefficients used for clustering γi = (βi , θi)

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

The almost-certain discretization of the Poisson-Dirichlet Process induces a clustering among the data, grouping the observations with the same latent variables γi sampled from the DP, the so-called tie

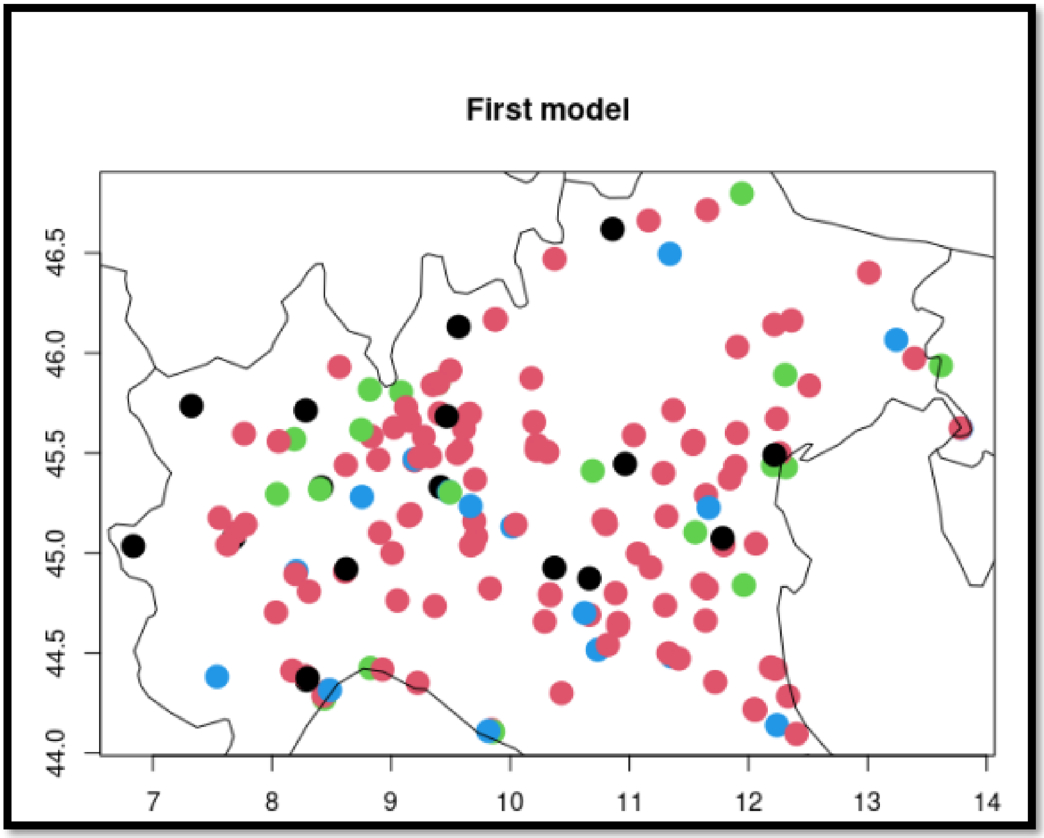
$$G_0(\boldsymbol{\gamma}) = G_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathrm{N}_d\left(\boldsymbol{\beta} \mid \mathbf{0}, \boldsymbol{\Sigma}_\beta\right) \times \mathrm{N}_T(\boldsymbol{\theta} \mid \mathbf{0}, \mathbf{R}),$$

$$R_{jk} = \sigma_\theta^2 \rho^{|j-k|},$$

$$\boldsymbol{\Sigma}_\beta = \mathrm{diag}\left(\sigma_{\beta 1}^2, \ \ldots \ , \sigma_{\beta d}^2\right).$$

The number of clusters m (unique values in γ1, . . . , γn)) is determined by the parameters (a, b). Larger values of either a or b, within the valid ranges, produce a large

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

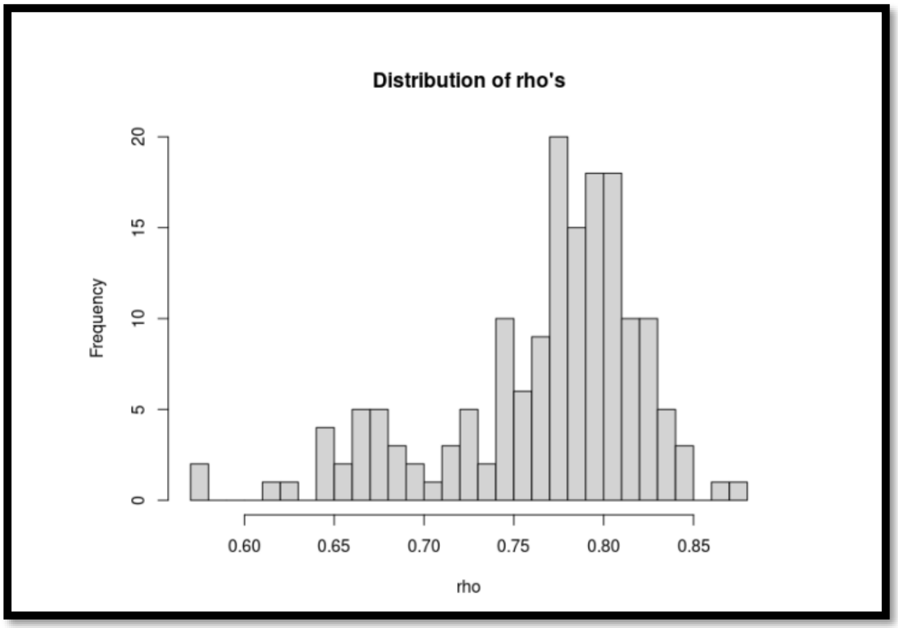


We could have expected a poor outcome because there isn't much difference in level or trend between the stations, as can be seen in the matplot of the PM10 concentration in Northern Italy. This model provides limited information, as you can see above. This induced us to proceed with a second model based on a new approach, clustering on a higher hierarchical level, the parameters of the auto-regressive process ρ and $\sigma_\theta^2$.

## Second Model:

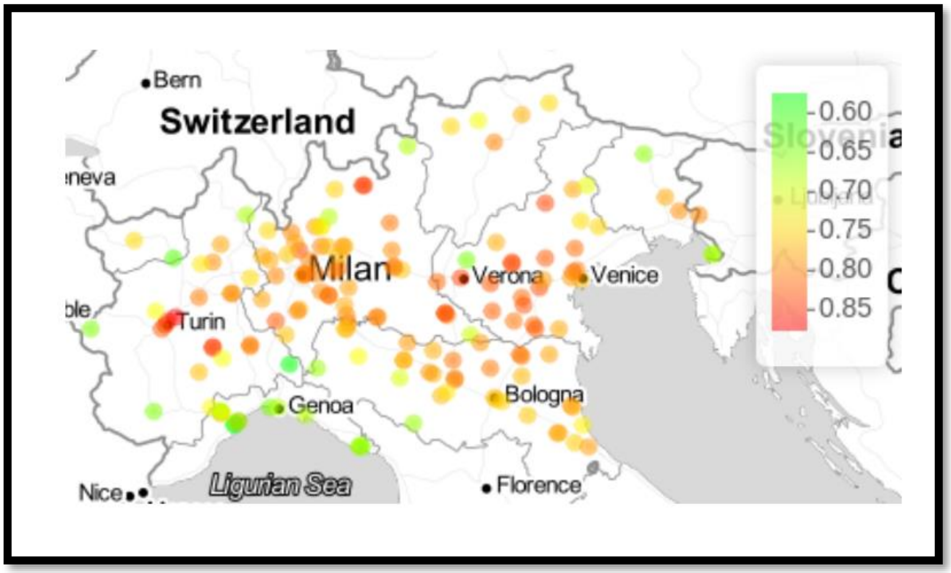
The idea was to develop a clustering technique based on the persistence of the pollutant, investigating the correlation between consecutive observations in the same station. In order to better understand the meaning of persistence, we can say that a persistent series is one in which the variable's value 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 $\sigma_i^2$).

Below is the distribution of the ρ's, obtained by fitting an autoregressive model of order 1 using the function arima in R.



***ρ's obtained by fitting an AR(1) on PM10 concentrations.***

Below is the distribution of the ρi for each station considered. The green points are the stations with a low value of ρ, so a low persistence, while the red points are the ones with persistent time series.



***ρ's distribution in Northern Italy***

Using some specifications from the previous one, most importantly setting as prior for the distribution of the parameters ρi and σ^2i , a Dirichlet process. The definition of yi is similar to the first one, but the β'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 α's, the variables on which we don't cluster. Moreover, from now on σ^2i represents the σ^2θ of the i-th station.

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

## Prior distributions:

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

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

## Bayesian model based clustering analysis:

In Bayesian clustering analysis, the Gibbs sampler from the MCMC method produces clusters of the time series parameters γ. These parameters facilitate clustering by characterizing each time series. To assess the clustering, the approach evaluates how frequently two parameters are in the same cluster over multiple iterations, which circumvents the label-switching problem. A similarity matrix is then constructed from these assessments, reflecting the relative clustering frequencies of the time series.

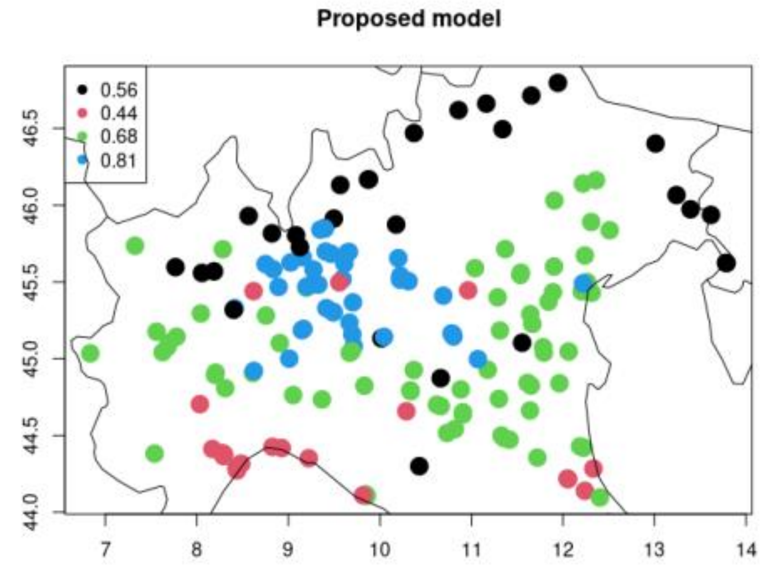
To determine the optimal clustering, a loss function is introduced within the Bayesian framework. The expected loss is minimized using the posterior samples. The variation of information (VI) loss function is typically used for this purpose. The optimal clustering c^∗ is obtained by minimizing the posterior expectation of the VI loss function given the data D, as shown:

This optimization is executed using the SALSO algorithm, which performs a greedy, stochastic search for the best partitioning of the data into the desired number of clusters.

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 conditionals
    2. Sample $\theta$ from its full conditional
    3. Sample $\gamma ' s = (\rho, \sigma 2 )$
    4. Sample $\sigma 2 \epsilon$ from its full conditional
    5. Sample $\sigma 2 \alpha$ from its full conditional
- End of Gibbs sampling
- Determining which cluster configuration minimizes a specific posterior loss

The last step was to apply the proposed model to the PM10 data described above. 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.



Proposed model

We can note that the cluster obtained is comparable to result of first model, 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 revealed studies analyzing 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.

## 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  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 . 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 centered in a small value than the one obtained

**References :**

[1]Luis E Nieto-Barajas and Alberto Contreras-Crist´an. ‘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 M¨uller. ‘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 Angel 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.