POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Clustering PM_{10} and other cute stuff

Project Report of Bayesian Statistics - Mathematical Engineering

Giulia Mezzadri, Ettore Modina, Oswaldo Jesus Morales Lopez, Federico Angelo Mor, Abylaikhan Orynbassar, Federica Rena

Professor:
Alessandra Guglielmi

Tutors:
Michela Frigeri
Alessandro Carminati

Academic year:
2023-2024

Abstract[a] In this project we undertake a comprehensive clustering analysis of PM_{10} levels in the Lombardy region (Italy), employing four different Bayesian models to account for the complex nature of our data, which comprise spatio-temporal measurements of PM_{10}, together with many other environmental variables, collected from various monitoring stations displaced across the entire region.
The main objective was to leverage on covariates, station locations and time trends to cluster weekly PM_{10} data over a one-year period. Our analysis revealed distinct clusters for each time step, with a noteworthy influence of morphological terrain characteristics (e.g. altitude, wind speed) and anthropological factors (e.g. agricultural activities, vehicles and road transports, etc.).
The analysis was executed concurrently across a set of four models, to study the different interactions and combinations of spatio-temporal aspects and covariates information. Despite some variations among the models, that however highlighted peculiar patterns and characteristics which each model independently dwelt on, a unanimous consensus emerged regarding the overall division between the stations. This study contributes valuable insights into the delicate interaction of spatial, temporal, and covariate variables in shaping PM_{10} levels, providing a robust foundation for understanding the clustering dynamics in the Lombardy region.

[a] See `https://github.com/federicomor/progetto-bayesian` for all the project codes and `https://federicomor.github.io/assets/figures/visualize.html` for the visualization page.

1. Introduction

Particulate matter with a diameter of 10 micrometers or less, known as PM_{10}, comprises small airborne particles which pose serious health risks upon inhalation due to their ability to deeply penetrate the respiratory system. Therefore, a meticulous monitoring of their levels is crucial for a comprehensive air quality assessment and the subsequent safeguarding of public health.

The PM_{10} particles arise from different sources, mainly combustion processes (like waste burning, house heating, energy generation and consumption), industrial and agricultural activities, vehicles exhaust, construction and demolition operations [otEU24]. This wide context poses under great attention the highly industrialized countries, such as Italy, which indeed in 2017 faced a significant pressure from the European Commission to address the persistently high levels of PM_{10} particles [Com17]. Both annual and daily limits were too often exceeded by several regions, with the most critical one being Lombardy.

This paper therefore embarks on the project with the aim of identifying both natural and anthropogenic factors contributing to elevated PM_{10} levels in the Lombardy region, employing a Bayesian clustering analysis.

Drawing upon data from the Agrimonia project, our focus relies on weekly averages values over a one-year timeframe. Our analytical approach involves the implementation of four models, including spatial and temporal ones alongside with covariate informed ones.

Now, in Section 2, we will delve into the dataset exploration and cleaning process, followed by Section 3 with a detailed presentation of each individual model, and to conclude with the models comparison and subsequent interpretation of the results in Section 4. Visualization played a pivotal role in our exploration, having to deal with a spatial and temporal framework as well as many available covariates, from which in the end we drew the interpretation insights. Arguably, our lack of technical knowledge in the phenomenon of the air pollutants required a fully data-driven approach, but in the end we assessed the validity of the our results and interpretation through a more specific research on the field under analysis.

2. Dataset inspection and processing

The Agrimonia dataset, developed in [FRFM+23], spans from January 1, 2016, to December 31, 2021, recording observations from a network of 141 stations in the surroundings of Lombardy region. The dataset gathers measurements from five different covariate groups: air quality (AQ), weather and climate (WE), pollutants' emissions (EM), livestock (LI), land and soil characteristics (LA). In total there are 38 covariates, with our target variable lying in the AQ group, namely `AQ_pm10`. Each row of the dataset is distinctively recognized by the combination of the station code and the date, making a total of 309072 rows, consisting of daily recorded values. The geographical coordinates, comprising longitude and latitude, are also available for each station in the dataset.

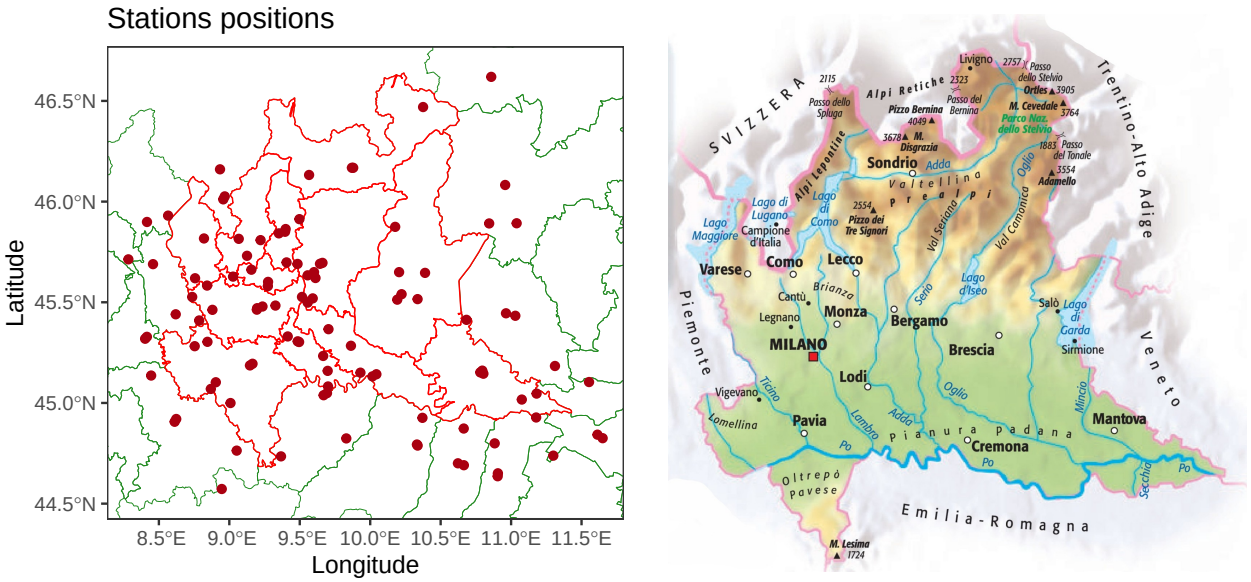

Figure 1: Map of the 105 selected stations after the data preprocessing (on the left), together with the physical map of the region under analysis (on the right).

The goal proposed for the project was "clustering weekly data of one year of PM_{10}", and as such we started by selecting the year and then to divide the dataset by weeks. One main concern for the year selection was the presence of missing data (NA) both in the target variable and in many other ones, as we can see in Figure 2.

About the covariates' selection, we noticed a considerable scarcity in the AQ group, so we were forced to remove them and save only our target variable PM_{10}. This may look like a relevant information loss, since the other pollutants like $PM_{2.5}$, SO_2 or NH_3 could have related well to the PM_{10} concentrations. Actually the EM group of variables, related to emissions, stored information about those pollutants, so the information they carried was preserved. We also removed variable `LA_soil_use`, which was considerably empty for most of the stations. After this procedure we remained with 36 variables, of which 31 covariates and 5 context-related (the temporal and spatial coordinates, the target variable and the station IDs), from the original 43.

Figure 2: Heatmap of the missing values of all the variables in the available dataset. The percentage is computed considering all the six years data (that is, before the year selection). On the rows there are the variables, on the columns the original 141 stations.

Regarding the PM_{10} levels, instead, there were many stations which were totally lacking of any recorded value, as revealed by Figure 3; therefore we removed them and we were left with 105 stations out of the original 141, which is still a very representative set. After this cleaning procedure, a natural choice for the year would have been to select the most recent one, but due to 2021 showing an increase in the missing values and being it still close to the covid-affected period after 2019, we decided in the end to choose 2018, hoping that the present years, somehow recovered from the pandemic anomalous levels, would be similar to that one.

Then we moved to the task of the weekly averaging. For the covariates, the selected year showed to be almost full of values except for three stations, which were lacking of all values in three variables of the LI and LA groups. Since this was a small problematic case and concerned stations outside the Lombardy area, the one

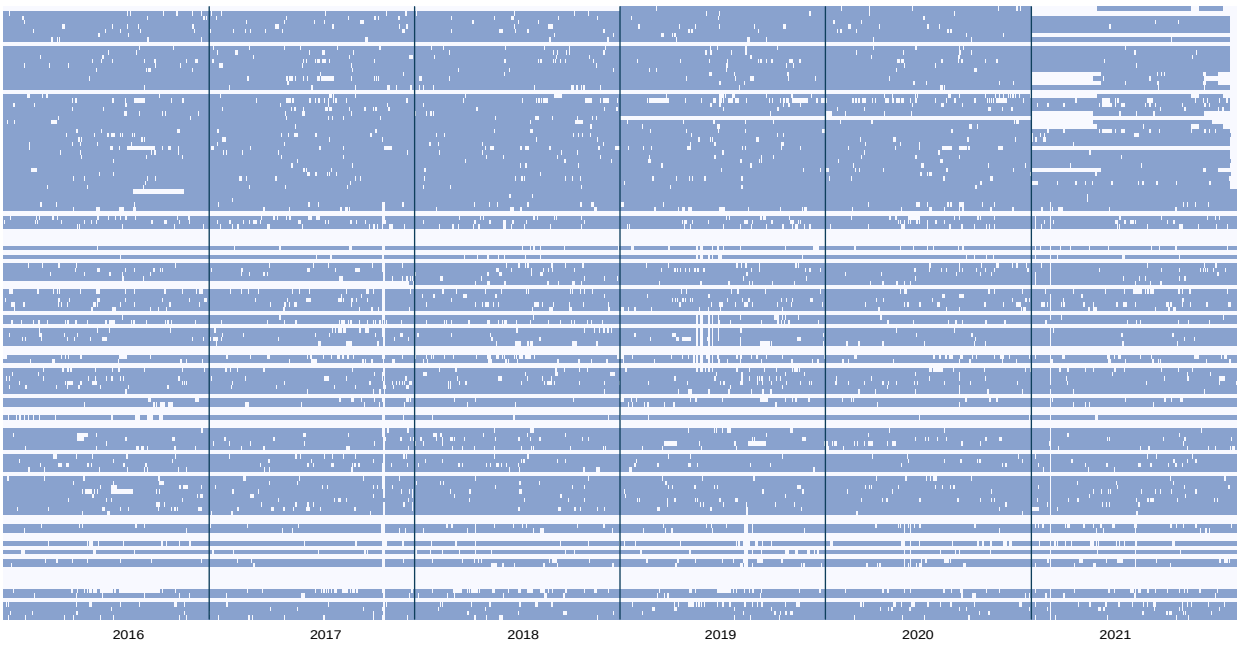

Figure 3: Heatmap of the missing data (in white) of the PM_{10} values. On the rows there are all the original 141 stations, on the columns all the 2192 days composing the six years.

of primary interest, we didn't deem necessary to remove them completely, but instead we filled in the missing data with an average of the values of the three closest stations on the map.

Also for the PM_{10} data there were some missing spots, sparse but affecting almost all stations. Initially we thought of filling them by using, for each station, a linear interpolation between the closest-in-time present data around a set of missing ones. This would have allowed the build of the weekly division by simply averaging over those (now all complete) values. But we thought that this method would have induced a double approximation: the first one in the NA filling and the second one in the weekly averaging. So in the end we decided to directly build the weekly division by averaging not necessarily on the complete set of seven days, but just on the available values in a given week. We applied this procedure on all the numeric variables, as well as on the categorical ones (e.g. the wind direction) but using the mode instead of the mean.

This way we got the final dataset, on which we then performed a logarithmic transformation to the PM_{10} variable, to achieve a normal distribution, followed by a shift to bring them into having zero mean. The obtained trend is depicted in Figure 4. We also standardized the numerical covariates, including the spatial coordinates. This allowed us to enhance the suitability of the data for the subsequent statistical models, which for example assumed a normal distribution of the target data, and in general worked better using centered data, to accommodate the prior distribution support of the parameters. This comprehensive processing dataset formed the foundation for our investigation into the factors influencing PM_{10} levels in the Lombardy region.

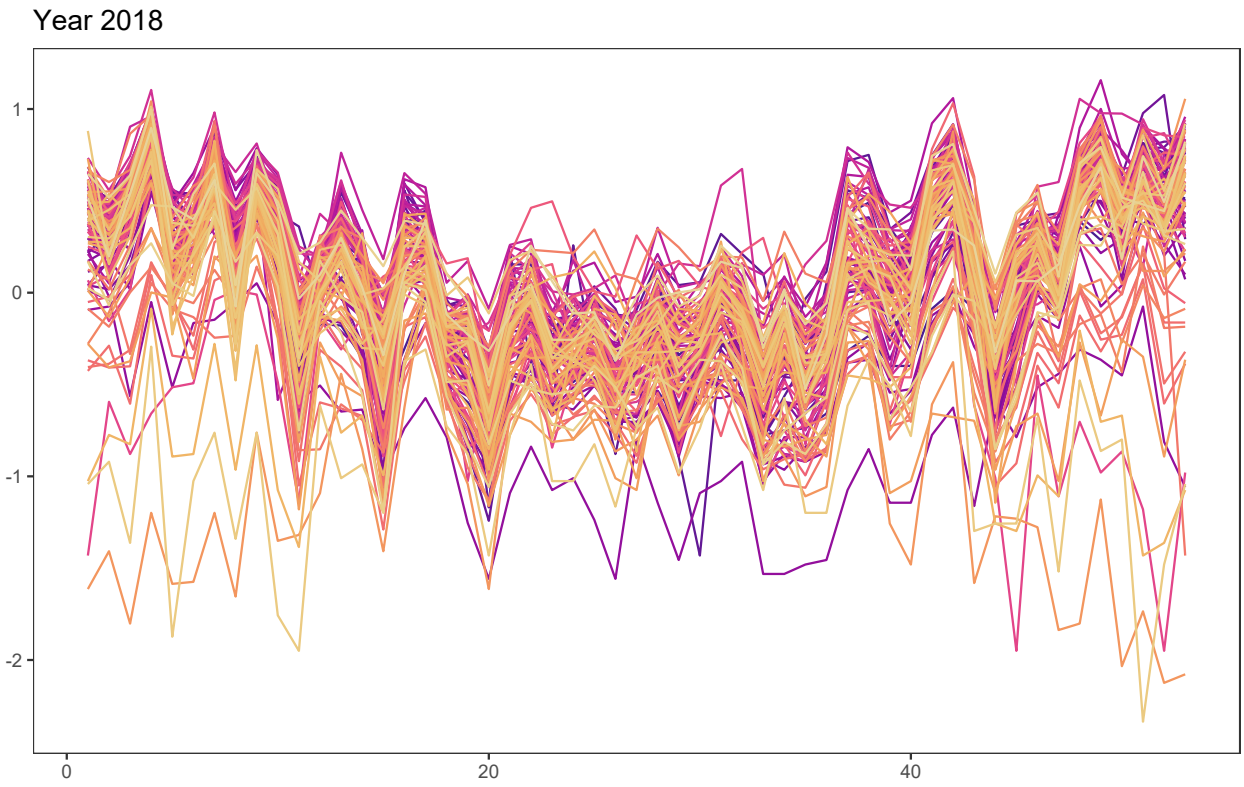

Figure 4: Trend of the PM_{10} values in the 2018 year, for the 105 selected stations, after the data processing step (log-transformation and centering).

3. Models

For our analysis we looked into models which could tailor the complex nature of our data, exploiting spatial and temporal information, together with covariates, with a clustering target in mind. Unfortunately, there was no "holy grail" which could manage to harness all those levels of information, but nonetheless we found four models which in the end worked well for our task.

We will now see them in details, but for a clear preview of their characteristics refer to Table 1.

model name	Time	Space	Covariates
sPPM	✗	✓	✗
DRPM	✓	✓	✗
Gaussian PPMx	✗	✗	✓
Curve PPMx	✗	✗	✓

Table 1: Summary of the functional characteristics of the models at hand. All of them were able to perform clustering natively.

3.1. sPPM model

We started with a purely spatial model, developed in [PQ16], which consisted in a Gaussian likelihood and a spatial PPM prior on partitions. This design aims to cluster together observations (the stations, in our case) which exhibit a common trend in the PM_{10} values but keeping into account their spatial locations, assuming that stations closer in space tend to show a similar behaviour, a priori. That is, in addition to the measurements of the target variable, the proximity of observations influences their partitioning. The full model is the following:

$$\begin{aligned} Y_i|\mu^\star, \sigma^\star, c_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mu^\star_{c_i}, \sigma^{\star 2}_{c_i}) \quad i = 1, \dots, n \\ \mu^\star_j|\mu_0, \sigma_0^2 &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \sigma^\star_j|A &\sim \mathcal{U}(0, m_s) \\ \mu_0|m, s^2 &\sim \mathcal{N}(\mu_0, s_0^2) \\ \sigma_0|B &\sim \mathcal{U}(0, m_{s0}) \\ \rho_m|M, \xi &\sim \text{sPPM} \end{aligned} \tag{1}$$

where $\mathcal{N}$ and $\mathcal{U}$ denote the Normal and Uniform laws. We can see how response variable is modelled trough a Normal distribution, of cluster-specific mean and variance, while the partition variable ρ_m on the m locations has a spatially-informed prior, defined by

$$\Pr(\rho_m|vecs) \propto \prod_{h=1}^{k_n} C(S_h, \boldsymbol{s}^\star_h)$$

where $C(S_h, \boldsymbol{s}^\star_h)$ is a cohesion function that measures how likely elements of S_h are clustered a priori, while $\boldsymbol{s}^\star_h = \{\boldsymbol{s}_i : i \in S_h\}$ is the vector carrying the spatial information of the cluster h. In our dataset we had $\boldsymbol{s}_i = (\text{latitude}_i, \text{longitude}_i)$ for all units $i = 1, \dots, 105$, i.e. the stations.

M	method	MSE	MSPE	LPML	WAIC
$M = 0.01$	$C_{1_{\alpha=1}}$	0.11982303	0.05766370	5.2563275	-25.203210
	$C_{1_{\alpha=2}}$	0.11312414	0.05604355	6.4879950	-22.511087
	C_2	0.12443479	0.05083555	1.1840116	-11.885645
	C_3	0.06498192	0.06338080	-0.6759250	-11.933451
	C_4	0.10825500	0.05733057	3.9262961	-28.335136
$M = 0.1$	$C_{1_{\alpha=1}}$	0.11524981	0.06309311	1.1684726	-17.622761
	$C_{1_{\alpha=2}}$	0.11843205	0.06461063	-2.2174683	-7.156824
	C_2	0.13093623	0.05086688	1.0103033	-14.315375
	C_3	0.07020742	0.06177079	0.8533671	-14.122342
	C_4	**0.10878478**	**0.04986580**	**6.8232228**	**-30.527901**
$M = 1$	$C_{1_{\alpha=1}}$	0.11678516	0.07271717	-12.0532115	7.392855
	$C_{1_{\alpha=2}}$	0.11821033	0.08026768	-26.2609990	30.812101
	C_2	0.11386666	0.05615466	-5.0202824	-8.652450
	C_3	0.09224991	0.06171361	2.9784456	-19.611827
	C_4	0.12542675	0.04993301	6.1906711	-28.117866

Table 2: Model fit comparison of the different choices for the cohesion functions the parameter M, using the data of 2017.

For the model fit, implemented on `R` through the `PPMSuite` package, we tested the four different cohesion functions which were there available, alongside with different values of a parameter M on which the functions depend, in order to look for the best arguments which could suit our dataset. For the first two cohesion functions a further tuning was available through a parameter α. We ran this selection procedure using the 2017 year dataset, as we split the stations into train and test sets to build the fit on the former and evaluate the performance on the latter. The performance scores were both about the statistical fit (LPML and WAIC) but also on accuracy (MSE and MSPE, which are mean squared errors associated to train and test data, respectively), as we used the generated clusters to extract predictions for the left out stations of the test sets. The results are summarized in Table 2. Since the variability of the metrics was similar, and to be robust of outliers, we compared the median of them.

According to those tests, the best combination for the model parameters was choosing the cohesion function C_4 with $M = 0.1$. Therefore we selected them to fit the final sPPM model on the 2018 data of our interest, collecting 20000 iterations, discarding the first 10000, and thinning them by 10. For this model, and for the other ones which do not include time, we just ran 53 different fits, one for each week.

3.2. DRPM model

The second model we focused on, and noticeably the only one including time, is the Dependent Modeling of Temporal Sequences of Random Partitions (DRPM), developed in [PQD22]. The main objective of the authors was to define a spatio-temporal model capable of performing "smooth" clusterings, i.e. a framework which would favour a gentle evolution in time of the units allocations, rather than abrupt (and therefore less interpretable) changes in them. This result was clearly reached also in our analysis, as we will describe more precisely in section 4.3, where we witnessed a more regular trend in the clusters definition for the DRPM model with respect to the other ones.

The model starts by assuming a first order dependence relation between clusters, meaning that the conditional distribution of ρ_t given $\rho_{t-1}, \dots, \rho_1$ just depends on ρ_{t-1}. This idea is implemented using a temporal dependence parameter $\alpha \in [0, 1]$ which controls the level of flexibility in the cluster allocation variables: the higher is α, the higher is the tendency of units to remain in their current cluster, meaning that clusters ρ_{t+1} will be similar to ρ_t. Conversely, when α approaches 0, we would get more independent clusters. In this way the clusters allocations variables $\boldsymbol{c}_t$ will follow a temporal Random Partition Model (the entry tRPM in the model formulation) driven by the sequence of α_t and the Dirichlet dispersion parameter M. The full definition of the model is:

$$
\begin{aligned}
Y_{it}|Y_{it-1}, \boldsymbol{\mu}_t^\star, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{\eta}, \boldsymbol{c}_t &\overset{\text{ind}}{\sim} \mathcal{N}(\mu^\star_{c_{it}t} + \eta_{1i} Y_{it-1}, \sigma^{2\star}_{c_{it}t}(1-\eta_{1i}^2)) \quad i = 1, \dots, n \quad \text{and} \quad t = 2, \dots, T \\
Y_{i1} &\overset{\text{ind}}{\sim} \mathcal{N}(\mu^\star_{c_{i1}1}, \sigma^{2\star}_{c_{i1}1}) \\
\xi_i = \text{Logit}(\tfrac{1}{2}(\eta_{1i}+1)) &\overset{\text{ind}}{\sim} \text{Laplace}(a, b) \\
(\mu^\star_{jt}, \sigma^\star_{jt}) &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \\
\theta_t|\theta_{t-1} &\overset{\text{ind}}{\sim} \mathcal{N}((1-\phi_1)\phi_0 + \phi_1\theta_{t-1}, \lambda^2(1-\phi_1^2)) \\
(\theta_1, \tau_t) &\sim \mathcal{N}(\phi_0, \lambda^2) \times \mathcal{U}(0, A_\tau) \\
(\phi_0, \phi_1, \lambda) &\sim \mathcal{N}(m_0, s_0^2) \times \mathcal{U}(-1, 1) \times \mathcal{U}(0, A_\lambda) \\
\{\boldsymbol{c}_t, \dots, \boldsymbol{c}_T\} &\sim \text{tRPM}(\boldsymbol{\alpha}, M) \quad \text{with} \quad \alpha_t \overset{\text{iid}}{\sim} \text{Beta}(a_\alpha, b_\alpha)
\end{aligned}
\tag{2}
$$

About the target variable Y_{it}, the authors modelled it with a Normal law with mean $\boldsymbol{\mu}_t^\star$ and variance $\boldsymbol{\sigma}_t^{2\star}$. The mean of that distribution actually incorporates a more sophisticated modelling introducing an autoregressive part both at the observations and at the parameters (or "atoms") level. Indeed, the Y_{it} depend on Y_{it-1} through the parameter η_{i1}, while for the $\mu^\star_{jt}$ the autoregressive structure is inside the parameter θ_t which enters in his prior distribution definition.

This deepening level allowed us to test different subsets of models and to select the best one which would suit our data. Through their package `drpm` on `R`, we fitted 8 different models based on the binary choices available for those three key parameters: the α could be set constant or varying in time, while the η_{1i} and ϕ_1 could be present (therefore introducing the autoregressive design) or not.

According to those tests, the best model for our scenario turned out to be the one using a time specific α and with an autoregressive component just at the atoms level, while not for the observations. Surprisingly, the model at his full complexity scored last in the ranking. We then we ran another fit on the best model, using some further refined parameters in terms of samples collection, to get the definitive results. Each fit of the 8 models tested above took around one hour, while the final fit took a little more than two hours and we ran

model at test				LPML	WAIC
model	η:No	ϕ:Yes	α_t:Yes	**1077.64**	**-2366.48**
model	η:No	ϕ:No	α_t:Yes	950.17	-2117.36
model	η:Yes	ϕ:No	α_t:No	724.34	-1474.02
model	η:No	ϕ:Yes	α_t:No	693.04	-1458.70
model	η:Yes	ϕ:No	α_t:Yes	605.32	-1287.13
model	η:No	ϕ:No	α_t:No	504.41	-1129.83
model	η:Yes	ϕ:Yes	α_t:No	445.16	-913.62
model	η:Yes	ϕ:Yes	α_t:Yes	403.05	-1264.03

Table 3: Metrics values computed for the DRPM model selection, sorted by best to worst. Higher LPML and lower WAIC values denote a better fit.

100000 iterations, discarding the first 60000, and thinning by 40; thus getting 1000 iterates. The high value of burn in was deemed necessary after seeing some significant oscillations even after a lot of iterations, while the thinning value was suggested by the authors and confirmed by the good trend of almost all our ACF plots (see Appendix B for them).

3.3. Gaussian PPMx model

After sPPM and DRPM, which dealt with space and time, we started looking for models which could incorporate the covariates, and as third choice we implemented a Gaussian PPMx model, still from the `PPMSuite` package (the same of sPPM and of the later discussed Curve PPMx).

The approach of this model, presented in [PQ18], was to handle the covariates influence on the clustering only through a calibration of the partition prior distribution. In this way the cohesion function now gets updated with a term defined as a nonnegative function $g(\boldsymbol{x}_h^\star)$ of the cluster-specific vectors of covariates $\boldsymbol{x}_h^\star$, i.e. $\boldsymbol{x}_h^\star = \{\boldsymbol{x}_i : i \in S_h\}$ with $\boldsymbol{x}_i$ the covariates of unit i. The function $g(\boldsymbol{x}_h^\star)$ is called similarity function and measures the homogeneity of those covariates, leading to higher values when there is more likeness among them. The full model can be defined as follows:

$$
\begin{aligned}
Y_i|\mu_j^\star, \sigma_j^\star, S_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_{S_i}^\star, \sigma_{S_i}^{\star 2}) \quad i = 1, \dots, m \\
\sigma_j^\star &\sim \mathcal{U}(0, A), \\
\mu_j^\star|\mu_0, \sigma_0^2 &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2) \quad j = 1, \dots, k_m \\
\sigma_0 &\sim \mathcal{U}(0, A_0), \\
\mu_0 &\sim \mathcal{N}(m, s^2), \\
\Pr(\rho_m|\boldsymbol{x}) &\propto \prod_{h=1}^{K_m} c(S_h) g(\boldsymbol{x}_h^\star)
\end{aligned}
\tag{3}
$$

where $m = 105$ is the number of subjects under analysis (the stations), while K_m the number of clusters. In our fit we decided to use the the auxiliary similarity function, defines as

$$
g(\boldsymbol{x}_h^\star) = \int \left(\prod_{i \in S_h} q(\boldsymbol{x}_i|\boldsymbol{\xi}_h^\star) \right) q(\boldsymbol{\xi}_h^\star)\, d\boldsymbol{\xi}_h^\star
$$

where $q(\boldsymbol{x}_i|\boldsymbol{\xi}_h^\star)$ and $q(\boldsymbol{\xi}_h^\star)$ are density functions whose selection depend on the covariate types.

Having decided the structure of the model we dived into the selection of the covariates. In order to make the results as interpretable as possible, we considered to restrict us to a subset of five variables, selected by a maximization of goodness of fit metrics such as LPML. More precisely, before obtaining the final model, we considered intermediate models, in an forward selection style, aimed solely at choosing the most informative variables.

Following this idea we firstly compared 31 different models, each characterized by a single variable, and we selected as the first covariate the one corresponding to the model with the highest LPML. This turned out to be `Altitude`, and indeed from the plots available in the visualization page or showed in Section 4, altitude information was a distinctive key of the clusterings. After this first iteration we considered 30 different models,

each characterized by `Altitude` and another variable, among the present ones in the dataset, and comparing the derived models we selected the second covariate. Repeating this procedure we obtained the best subset of five covariates, in terms of LPML, which consisted of `Altitude`, `EM_nox_sum`, `WE_mode_wind_direction_100m` (a categorical variable), `WE_wind_speed_100m_max` and `LA_lvi`.

Once we finalized the set of covariates to use in the model, we further refined the choice of the priors and fitted the final model. Not seeing any particular trouble about the convergence diagnostics, we limited to run 4300 MCMC iterates were collected, with a burn-in of 300.

3.4. Curve PPMx model

The last model that we considered was the functional version of Gaussian PPMx model, called Curve PPMx. This method applies a hierarchical functional data model, in which B-spline coefficients undergo clustering employing either a PPM or a PPMx prior approach on partitions, but using the same approach of the Gaussian PPMx. The PPM and PPMx priors are used to group similar elements based on their characteristics, where the PPMx prior integrates the concept that individuals sharing similar covariate values are more likely to be grouped together [PQ15]. The difference is that now the "elements" are no more point realizations but become functional curves.

The use of B-splines allows for the flexible clustering of functional realizations based on the characteristics of the B-spline coefficients and covariates [MQR11]. More precisely, the model can be represented as follows:

$$Y_i(t) = \sum_{j=1}^{p} \beta_{ij} B_j(t) + \varepsilon_i(t) \quad i = 1, \ldots, n \tag{4}$$

where $Y_i(t)$ is the set of functional realizations, $B_j(t)$ is the j-th B-spline basis function, while β_{ij} are the B-spline coefficients to be clustered. The term $\varepsilon_i(t)$ represents the error component, and t the time variable. The PPMx prior is then used to model the distribution of the B-spline coefficients β_{ij} and their clustering based on covariates [ZNS11].

Since the model uses the same approach of Gaussain PPMx, we ran it with the same set of covariates mentioned in Section 3.3. We fitted the model for each week of 53 weeks of the year so that we could observe how the clusterings changed over time, as otherwise the use of the full set of weeks as timeframe would have produced time informed clusterings but fixed in all instants. In this way Curve PPMx model is in principle able to account for time, as DRPM, but rather than the latter, which allows change points in the clusters, the former would produce just a summary, a global clustering of the units over the selected timespan.

The fit function had several hyperparameters to be tuned, and this allowed us to test different models and choose the one with the best fitting metrics, meaning highest LPML and lowest WAIC values. We tested three values of the scale parameter M, related to the dispersion parameter of a Dirichlet process, and different number knots used for the splines basis. The summary of the tests are reported in Table 4.

M	#knots	LPML	WAIC
$M = 0.1$	20	124.1328	-495.2763
	40	**2293.467**	**-4796.269**
$M = 1$	20	78.86386	-398.1941
	40	2072.848	-4392.669
$M = 10$	20	126.533	-507.5256
	40	2064.878	-4365.364

Table 4: Model fit comparisons for the different scale parameters and number of knots. Higher LPML and lower WAIC values denote a better fit.

According to those tests we chose $M = 0.1$ and a number of knots equal to 40. The run of all the fits with these best hyperparameters for the entire timeframe took around 2 hours. We set the number of iterations equal to 10000, using a burn-in of 5000, and thinning by 5. Those parameters allowed us to have the good trend of almost all our ACF plots, as reported in Appendix B.

3.5. Linear Model

Beyond the described models, available in the cited the packages on `R`, we thought that it could have proved useful to implement also a baseline model to be used for comparison, variable selection and a better insight on

the data.

With respect to the more complex models described in the previous sections, we implemented a simpler model which would grant us faster fits and the possibility of trying out a vast range of methods. With the target in mind of not being too complex but allowing the inclusion of covariates, we chose the linear approach and, since it wouldn't have been able to capture spatial variability anyway, we actually implemented a model for each station, leaving the coordinates and altitude out. The general structure was the following:

$$\begin{aligned} Y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2) \quad i = 1, \dots, n \\ \boldsymbol{\beta} &\sim \mathcal{N}_p(\boldsymbol{b}_0, \sigma^2 B_0) \end{aligned} \tag{5}$$

The models considered the numerical covariates linearly but tried to enable more variability in time by considering also their sine, cosine and square transformations.

The first idea was also to develop a clustering procedure on the linear model, maybe switching to linear mixed effects models and grouping together the stations according to the similarity among the betas, but it was soon discarded as we deemed it redundant, since we already had different models more suited for a precise and explainable clustering. Therefore we decided to just focus on variable selection instead. Since at our disposal we had plenty of covariates it would have been extremely long and computationally heavy to try methods based on partitions of combinations of covariates or spike and slab; and consequently we firstly implemented the model through JAGS to try and use another kind of selection method as seen in [KM98]. In particular, this procedure returned a matrix with confidence quantiles on the columns and the covariates on the rows, with value 1 if the covariate was considered relevant at that confidence and 0 otherwise. Our idea was then to select a quantile, keeping only the corresponding column as a vector, and sum element-wise the values for all stations, expecting an higher value in correspondence of a useful covariate to select while a significantly lower one for covariates discarded by most stations. Unfortunately, this attempt did not lead to any solid conclusion, since even changing the threshold and hyperparameters the final vector presented very similar values on all covariates, usually between 50 and 60, suggesting that all variables had been selected only for about half the stations, and none was significantly more important than the others.

For this reason we moved to Bayesian lasso and horseshoe methods, using the corresponding `R` packages and the same count-based approach as before, since both methods returned a binary vector indicating whether or not to keep a variable. Horseshoe analysis was inconclusive, discarding all covariates, while lasso showed a great weight on the precipitation variable (`WE_tot_precipitation`), and a more moderate but still interesting on the percentage of total green area of low and high vegetation type (`LA_lvi` and `LA_hvi`). Indeed these last two variables showed also a good separation in the mode clustering plots, so they actually resulted to be relevant, as we will see in the plots in Section 4.

4. Analysis of the results

As we can see in the summary clusterings of Figure 5, almost all models exhibited a stratification pattern. Regions with flat terrain generally displayed higher PM_{10} concentrations and were frequently clustered together, such as the Milan area or Milan and Mantua area. As elevation increased towards the Alps, fewer polluted clusters were observed. In the southwest area, some noticeable station emerged consistently across all models, clustering a small group of stations together but separately from the surrounding Milan cluster.

Now we dive into the detailed interpretation of the clusters, where we focused our attention on a few set of covariates which had the most distinct trend among clusters, were easy to interpret, and had a relevant and meaningful influence on the PM_{10} concentrations. We report them in separate two sections, with plots (not of all the four models, just for clarity and order) depicting what explained in the paragraphs. All the "bands" appearing in the next plots refer to regions delimited by the 25% and 75% quantiles of the observations in the clustering corresponding to that color, while the solid line refers to the median.

4.1. Morphological factors interpretation

The first class of relevant covariates includes altitude (`Altitude`) and vegetation characteristics (`LA_hvi` and `LA_lvi`), together with total precipitations (`WE_tot_precipitation`) and maximum wind speed at 100m above the ground (`WE_wind_speed_100m_max`).

Altitude and vegetation Elevation and green areas are complementary: at higher altitudes we find denser vegetation, while at lower altitudes vegetation become sparser. The growth pattern of both types of vegetation is consistent throughout the year, with high levels in summer and low levels in winter. We can interpret this

effect in two ways:

1. It's the altitude that brings the PM_{10} down; or
2. It's the more vegetation which contributes to reduce the PM_{10} levels.

The reason behind that could be that plants use their special micro-morphological structure to retain particulate matter, while altitude helps simply because it implies distance from industrial areas. So in general, high vegetation and high altitude are generally correlated with lower PM_{10} levels, as we can see in Figure 6.

Total precipitations The total precipitation appears to influence the reduction of PM_{10} levels, as it is well known how rainwater can take away air pollutants and carry away some of the particulate matter, to some extent, by delivering and depositing the contaminants to the ground. This process is also known as precipitation scavenging or washout. This phenomenon is evidenced also in our analysis by the events observed in November, characterized by peaks in rainfall which corresponded to a drop in PM_{10} levels, clearly visible in Figure 7.

Max intensity of the wind speed at a height of 100m High intensity of the wind increases PM_{10} levels, but this effect is mitigated by the presence of rain. For instance, in October, despite high levels of both wind and rain, PM_{10} levels decrease. In December, when precipitation levels were low and the influence of rain could

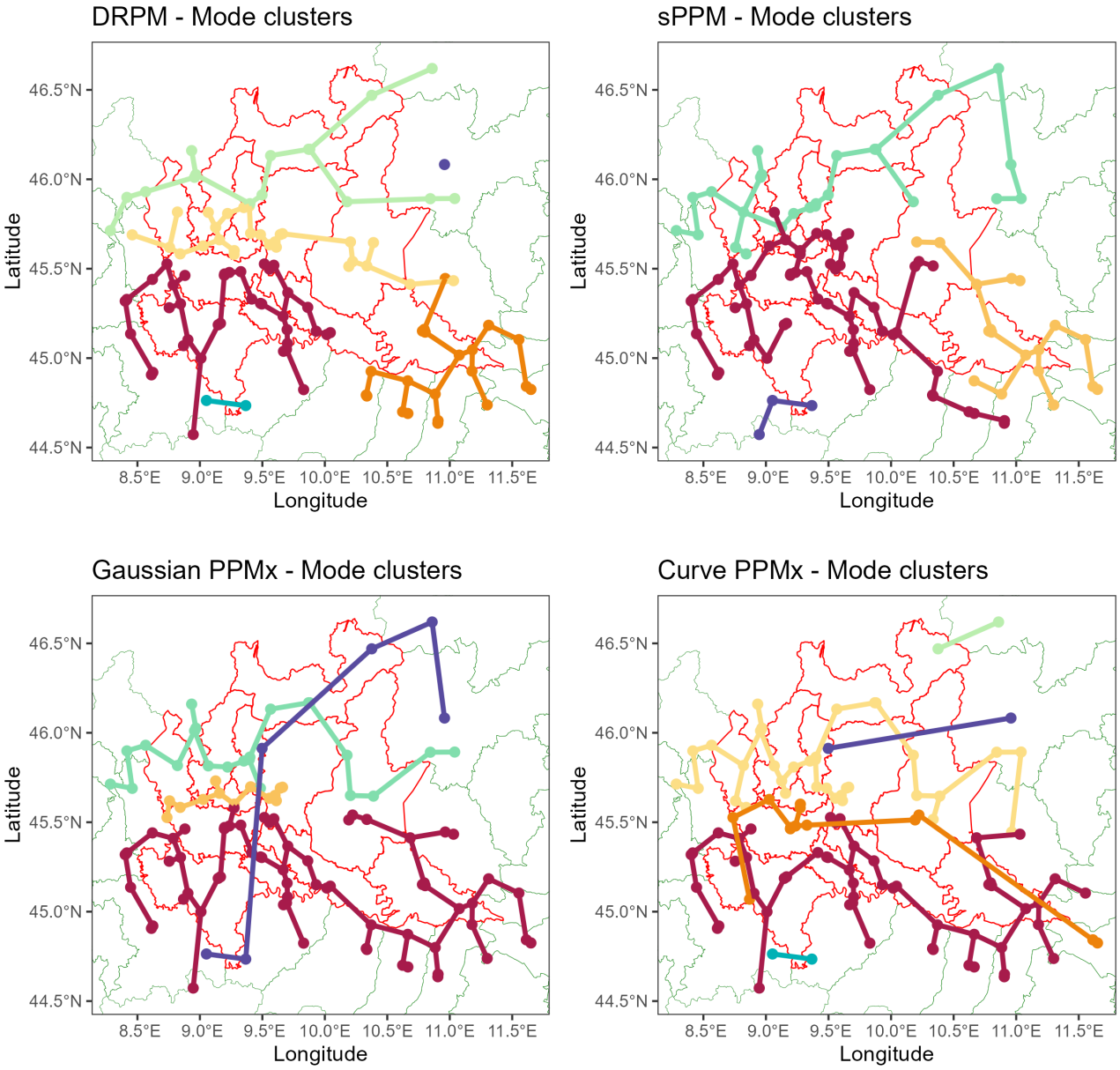

Figure 5: Maps of the most frequent clusters, throughout the 53 weeks of 2018, for all the models. The color palette is also informative, as clusters are colored from lower on average values of PM_{10} (in blue) towards higher on average values (in red).

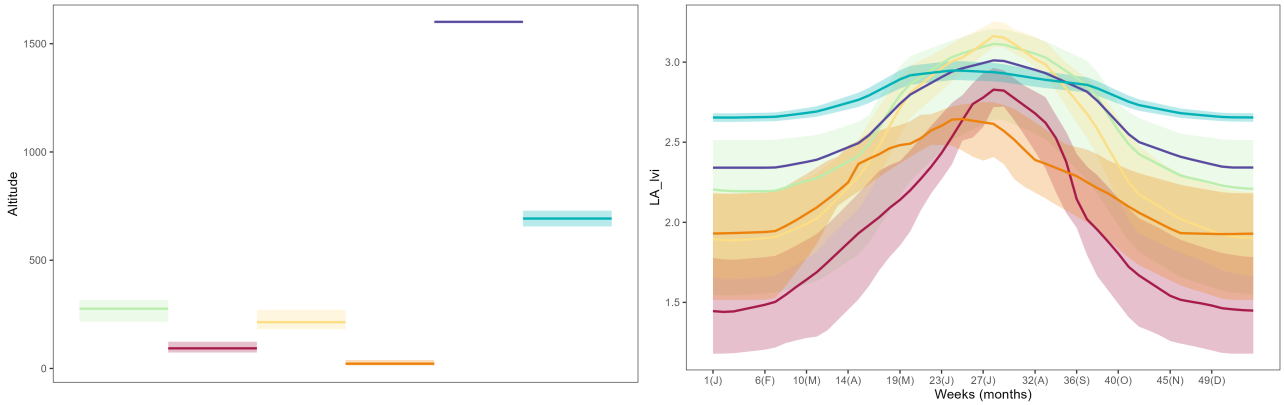

Figure 6: (Model DRPM) Plots of the trend in `Altitude` and `LA_lvi` variables of the mode clusters, showing how the high altitude and high vegetation are associated to lower PM_{10} values (in blue and light blue).

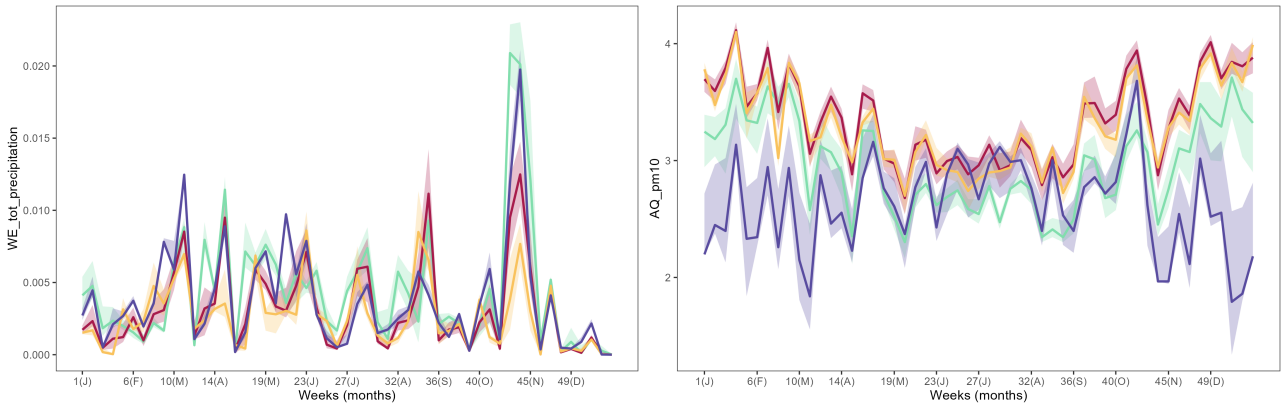

Figure 7: (Model sPPM) Plots of the trend in `AQ_pm10` and `WE_tot_precipitation` variables of the mode clusters, showing, around November, a peak in rainfalls and a drop in PM_{10} concentrations.

be filtered out, showing how high wind speeds indeed elevated the PM_{10} levels, as we can see in Figure 8. This is probably due to wind generating and moving dust from the ground (where pollutants may have been deposited), a problem more highlighted in dry and rural areas such as the Lombardy valley.

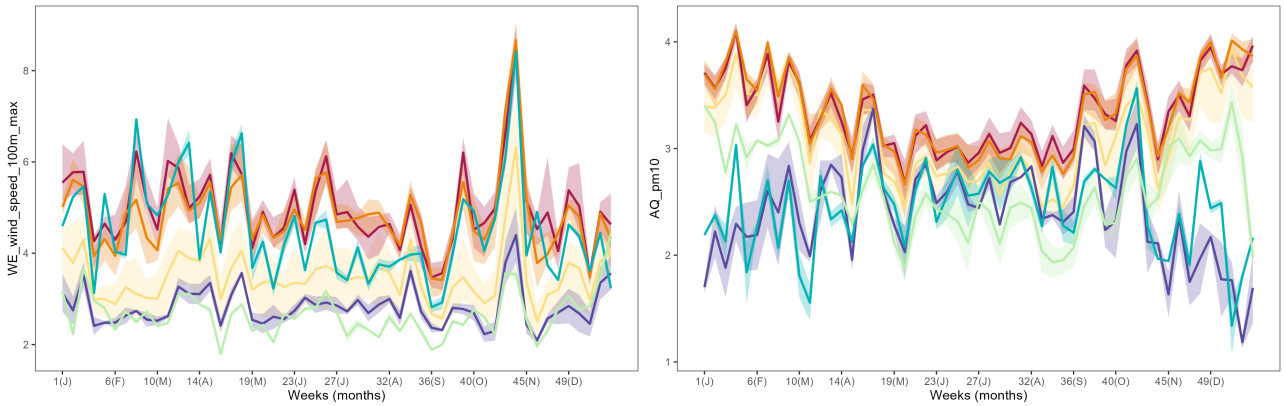

Figure 8: (Model Curve PPMx) Plots of the trend in `WE_wind_speed_100m_max` and `AQ_pm10` variables of the mode clusters, showing, especially around December (where the rain levels dropped), the correlation among high intensities of wind and high concentrations of PM_{10}.

4.2. Anthropological factors interpretation

The second class of relevant covariates includes the effects of human activities on the other pollutants generation such as NO_x and NH_3, which can then affect the concentrations of PM_{10} by chemical transformations. These variables are morally all the ones from the EM group (`EM_nh3_agr_soils`, `EM_nh3_agr_waste_burn`, `EM_nh3_livestock_mm`, `EM_nh3_sum` and `EM_nox_sum`, `EM_nox_traffic`).

Emissions of NO_x This other pollutant is positively correlated with PM_{10}, with higher levels observed during winters when PM_{10} values are also elevated. It's particularly higher for clusters in regions characterized by high levels of industrialization and transportation. This distinction is more pronounced in models incorporating covariates, as we can see in Figure **??** in the case of the Gaussian PPMx model.

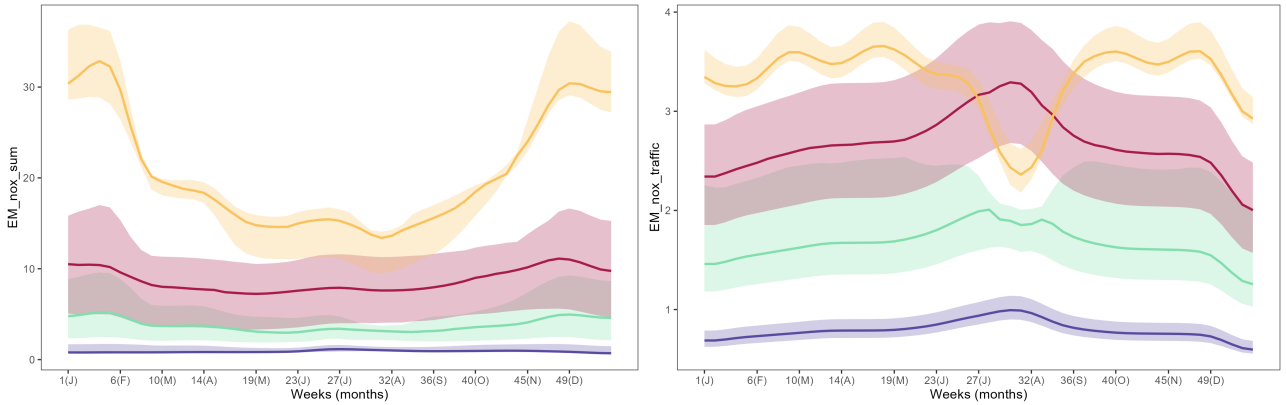

Figure 9: (Model Gaussian PPMx) Plots of the trend in `EM_nox_sum` and `EM_nox_traffic` variables of the mode clusters, showing the clear separation reached in the clustering. Interestingly we see an inverse trend in the yellow and red curves (corresponding to the Monza province and the south region of Lombardy, respectively) in the traffic emission variable, but it's not clear why there is such a trend.

Emissions of NH_3 and livestock In the Milan (and Mantua) area we noticed a strong presence of NH_3 especially imputable to livestock farming, particularly during the central part of the year. Presumably, the presence of NH_3 from agricultural and breeding activities (the pretty much the only source of that pollutant) also contributes to the elevation of PM_{10} levels. Naturally, those areas corresponds to a high number of animals or crops. An example of the trend of these covariates is displayed in Figure **??**.

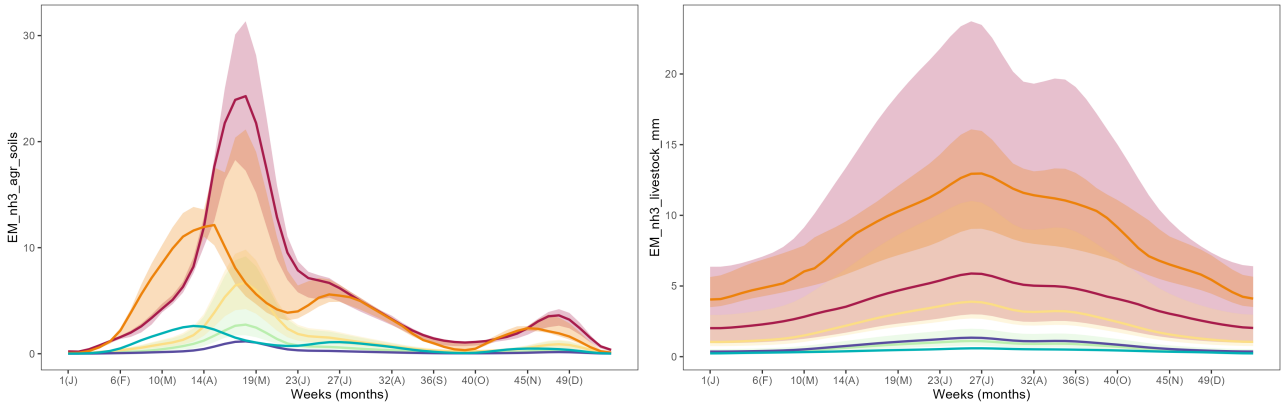

Figure 10: (Model DRPM) Plots of the trend in `EM_nh3_agr_soils` and `EM_nh3_livestock_mm` variables of the mode clusters, showing the interesting trend among the clusters and also a noticeable distinction, relevant since DRPM model did not comprise covariates in its fit.

To conclude our analysis, we deemed it important to analyze the presence of certain small clusters in all models. We can consider them as a sort of outlier sets, or just stations exhibiting a remarkable and distinctive behavior. In fact, their PM_{10} values differ significantly from the other stations, displaying an opposite trend, with low levels in winter and high levels in summer. These stations are located in high-altitude regions, and are the blue

or light blue spots that we can see in the north-east and south-west part of the map of Figure 5. The Gaussian PPMx model tends to group those stations together, in one single cluster, while the others keep them separated into smaller and isolated chunks. Additionally, these outliers exhibit the lowest values of PM_{10} and `EM_nox`. The causes behind this inverse trend surely include their position in elevated regions, suggesting maybe the role of snow in lowering levels during the winter months, by capturing the pollutants particles instead of releasing them in the air.

4.3. ARI metric comparison

A more numerical way to compare the clustering results is through the Adjusted Random Index (ARI), developed by [HA85], which is a sort of correlation index which measures the similarity between clusterings. Given two partitions ρ_1 and ρ_2, the $\text{ARI}(\rho_1, \rho_2)$ describes the amount of accordance between them, i.e. the level of agreement that they show in clustering the data. The ARI values are bounded above by one, which refers to a perfect alignment, and have zero expected value, which refers to the case of comparing two random generated partitions.

This metric was developed as a correction of the Random Index (RI) to take into account the fact the some concordance can happen by chance. While this correction is outside the scope of this analysis, the idea behind the definition of the original index lies in computing the frequency of agreements for any possible pair of units. Therefore we get $\text{RI}(\rho_1, \rho_2) = (a + b)/\binom{n}{2}$, where a is the number of pairs allocated in the same subset (i.e. when the two units are clustered together in both ρ_1 and ρ_2 allocations) and b the number of pairs allocated in different clusters (i.e. when the two units do not belong to the same cluster in both ρ_1 and ρ_2 allocations).

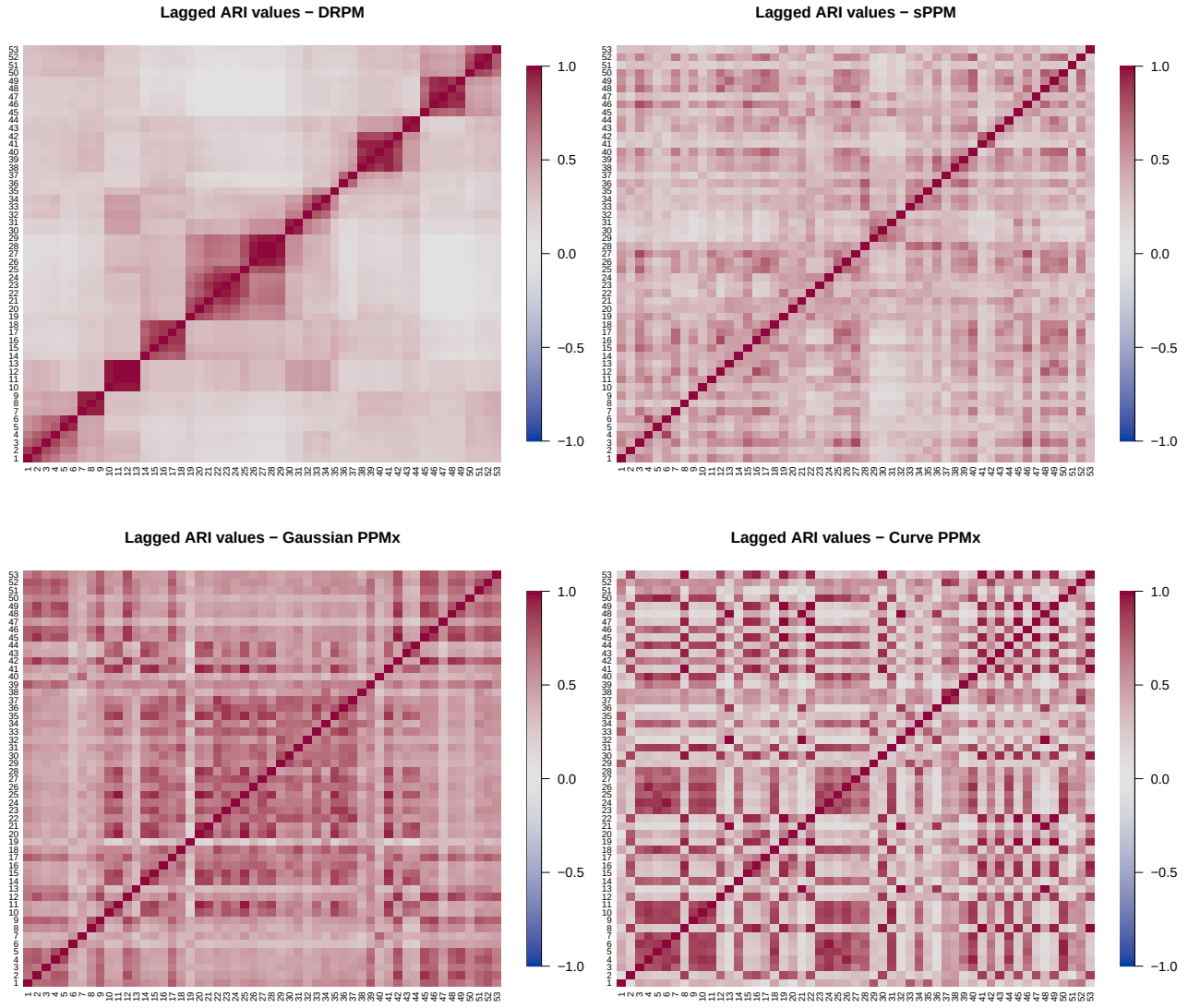

Figure 11: Lagged ARI values of the four models.

This metric allows for example to compare a proposal clustering with the real one, if available, to see how good is the matching; but in our case, where there was no correct answer, we used it to study the time evolution of the clusterings and to check the agreement level among the different models.

Regarding the time evolution we studied the Lagged ARI values, meaning that for each model we computed $\text{ARI}(\rho_t, \rho_{t+k})$ for $t \in \{1, \dots, 53\}$ and for all the valid values of k. In this way we obtained an information about the relation among the clusters throughout the year. As depicted in Figure 11, we can see how the DRPM exhibits a gentle evolution of the clusters, in a sort of time persistency, where almost every ρ_t tends to be similar to the subsequent clusterings, while losing connections with the ones of further away in time. Those regions of similar colored squares, highlighting the correlated clusterings, also appear in sPPM and Gaussian PPMx, but in a milder, less distinct way. Moreover, the latter displays a general dominance of high values, possibly denoting a more rigid and recurrent structure in the definition of the clusters. About Curve PPMx there is an interesting but quite erratic pattern, with an appearance of clear and intense spots where clusters tend to remain similar to each other, like it happened in the DRPM case, but now with less regularity.

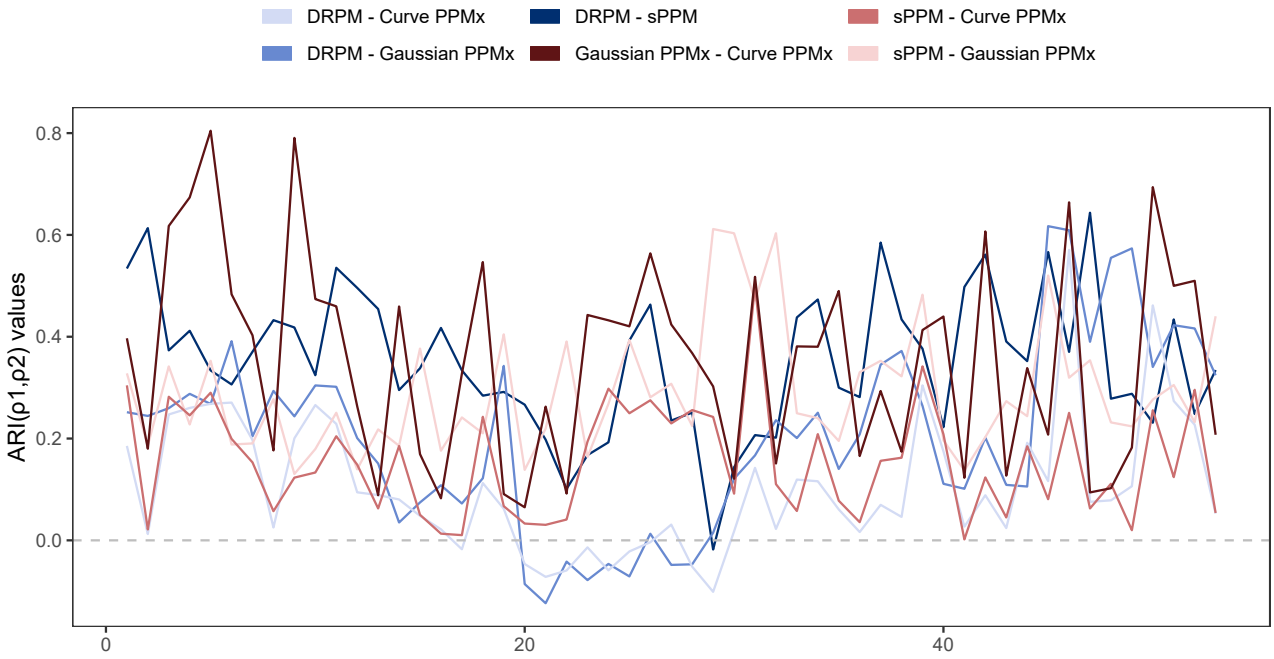

Figure 12: Plot of the ARI values for each pairwise comparison of the models, for all the weeks.

The other aspect that the ARI metric allowed to study was the agreement among the clusterings generated by the models, to see if a sort of general consensus and division appeared clearly in all models. To study this we computed $\text{ARI}(\rho_t^{M_1}, \rho_t^{M_2})$ for $t \in \{1, \dots, 53\}$ and for each pair of models M_1 and M_2 in the four that we fitted. From this computation, which produced Figure 12, we can see how the level of agreement among the models remained relatively consistent over time, indicating robustness in the modeling approach and coherence in the clusters generated. More precisely, during the non-summer months, all models seemed to agree, since the ARI values are always positive and quite high; and this is especially true for the classes of similar models. Indeed, the pairs which showed on average the greatest similarity were Gaussian PPMx vs Curve PPMx, which share the same underlying structure, and DRPM vs sPPM, which both incorporate the spatial information. This significant level of similarity, even in models with substantially different architectures, is probably due to the values of PM_{10} being more spread out in those fall-to-spring months, as we can see from Figure 4. Consequently, this led the models to an "easier" task, as they could rely just on the dominant influence of the more distinctive PM_{10} concentrations to generate clusters.

During the summer months, instead, we see an inverse trend, where there is some indecision and ambiguity in the models agreement. This is probably due to the trend of PM_{10} which becomes more condensed and uniform in that period, resulting in a lack of distinct patterns across stations and hindering the clustering task. In this way models needed to rely on other aspects of their architectures, which being different may have led to the different responses we see. Indeed, in that period we see two relevant disagreements, among DRPM vs the Gaussian and Curve PPMx. This contrast can be explained by guessing that the PPMx models tried to exploit and put more trust on the information encoded in the covariates, while DRPM, in the absence of that, tried to

take advantage of the spatial and temporal characteristics that it only owned.

5. Conclusions

6. Further developements

Several avenues for a further development remain, presenting opportunities to enhance the depth and precision of our analysis:

- **Use previous years data for model priors** In our approach we decided to standardize the covariates to make them suit the given prior distribution parameters in the model. But of course another idea could have been to keep the original covariates and consider incorporating data from the preceding years to establish fine tuned priors for the models. While our approach surely eased the models convergence, integrating historical data could offer additional insights and refine the robustness of our findings.
- **Distinguish between weekends and weekdays** Rather than gather data by weeks, this alternative differentiation could uncover different patterns, associated to more specific and human-related factors and contributing to a more nuanced understanding of particulate matter dynamics.
- **Ensemble modeling** Meaning that in principle we could get potential benefits by combining the outputs of the different models we obtained. This approach could enhance the overall accuracy and reliability of the clustering analysis, since by leveraging the strengths of individual models we can obtain a more comprehensive and robust estimation of the identified clusters.
- **Complete the DRPM model** Being remarkably the only model including space and time, a natural idea could be working on an update which include also covariates, to make a really final and "definitive" model.

These proposed extensions aim to further refine our methodology, enriching the interpretability of results and providing a more extensive understanding of the intricate factors influencing PM_{10} levels in the Lombardy region.

References

[Com17] European Commission. Infringement actions for excessive levels of PM10 in Italy, 2017. `https://ec.europa.eu/commission/presscorner/detail/ET/IP_17_1046`.

[FRFM+23] A. Fassò, J. Rodeschini, A. Fusta Moro, Q. Shaboviq, P. Maranzano, M. Cameletti, F. Finazzi, N. Golini, R. Ignaccolo, and P. Otto. AgrImOnIA: Open Access dataset correlating livestock and air quality in the Lombardy region, Italy (3.0.0), 2023.

[HA85] Lawrence J. Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[KM98] Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics*, 60:65–81, 1998.

[MQR11] Peter Müller, Fernando Quintana, and Gary L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, Jan 2011.

[otEU24] Council of the European Union. Infographic - Air pollution in the EU: facts and figures, 2024. `https://www.consilium.europa.eu/en/infographics/air-pollution-in-the-eu/`.

[PQ15] Garritt L. Page and Fernando A. Quintana. Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. *Bayesian Analysis*, 10(2), Jun 2015.

[PQ16] Garritt L. Page and Fernando A. Quintana. Spatial Product Partition Models. *Bayesian Analysis*, 11(1):265 – 298, 2016.

[PQ18] Garritt L. Page and Fernando A. Quintana. Calibrating covariate informed product partition models. *Statistics and Computing*, 28:1–23, 09 2018.

[PQD22] Garrit L. Page, Fernando A. Quintana, and David B. Dahl. Dependent modeling of temporal sequences of random partitions. *Journal of Computational and Graphical Statistics*, 31(2):614–627, 2022.

[ZNS11] Hui Zheng, David M. Nathan, and David A. Schoenfeld. Using a multi-level b-spline model to analyze and compare patient glucose profiles based on continuous monitoring data. *Diabetes Technology and Therapeutics*, 13(6):675–682, Jun 2011.

A. Visualiation methods

A pivotal component of a robust statistical analysis lies in the effective interpretation of results. To address this crucial aspect, we developed a sort of library of auxiliary functions, enabling us to visually investigate the various aspects of our research.

Given the intrinsic temporal and spatial dimensions of our dataset, we opted for a dynamic approach by creating videos, animations, and an interactive html page alongside the classical images. For the visualisation of spatial variables, we devised two principal tools: a grid map and an expanding circles plot, available at ??.

- **Grid Map** This tool exploits a distinct dataset inside the Agrimonia ones, which includes measurements across the entire region organized on a grid of evenly spaced points, rather than just the records we got through the station measurements. This finer dataset offers a panoramic overview of key variables, such as altitude and weather measurements, providing a comprehensive understanding of spatial patterns and how they evolve in time.
- **Expanding Circles Plot** Focused on station-level measurements, this tool illustrates the magnitude of variables by properly setting the radius and color intensity of the circles centered around each station. This approach granted us insights into more localized patterns, enhancing our comprehension of variable distribution across the region.

Finally, to enhance the clarity of our cluster representations, we devised a function that establishes connections between stations within the same cluster. These connections are formed by solving a minimum spanning tree problem, to avoid redundant or confusing or even random edges among clustered stations; a strategic approach chosen to yield a more organized and visually coherent representation of the clusters.

B. MCMC diagnostics

Here we present the plots that we used to check the convergence of the MCMC values generated by the models fit functions. They are, for summary purposes, just on two of the most relevant parameters of each model, and on weeks running from 1 to 50 with jumps of length 7.

We used trace plots to ensure that the chosen values for the burn-in were high enough to remove any unstable behaviour in the iterates. However, even after really high burn-in periods (e.g. 60000 in DRPM), occasionally there were still some oscillations, but by looking at the y axis we can see that there is no significant variation after all. Also sometimes there are divergent iterations, especially in the σ^2 trace plots. We think that these small issues were due to the complexity of the models, spanning on lots of subjects (the 105 stations) and several time instants (the 53 weeks), together with implementing a deep hierarchical structure.

We also looked at ACF plots to tune the thinning parameter, by seeing the trend of the auto correlation on subsequent iterates.

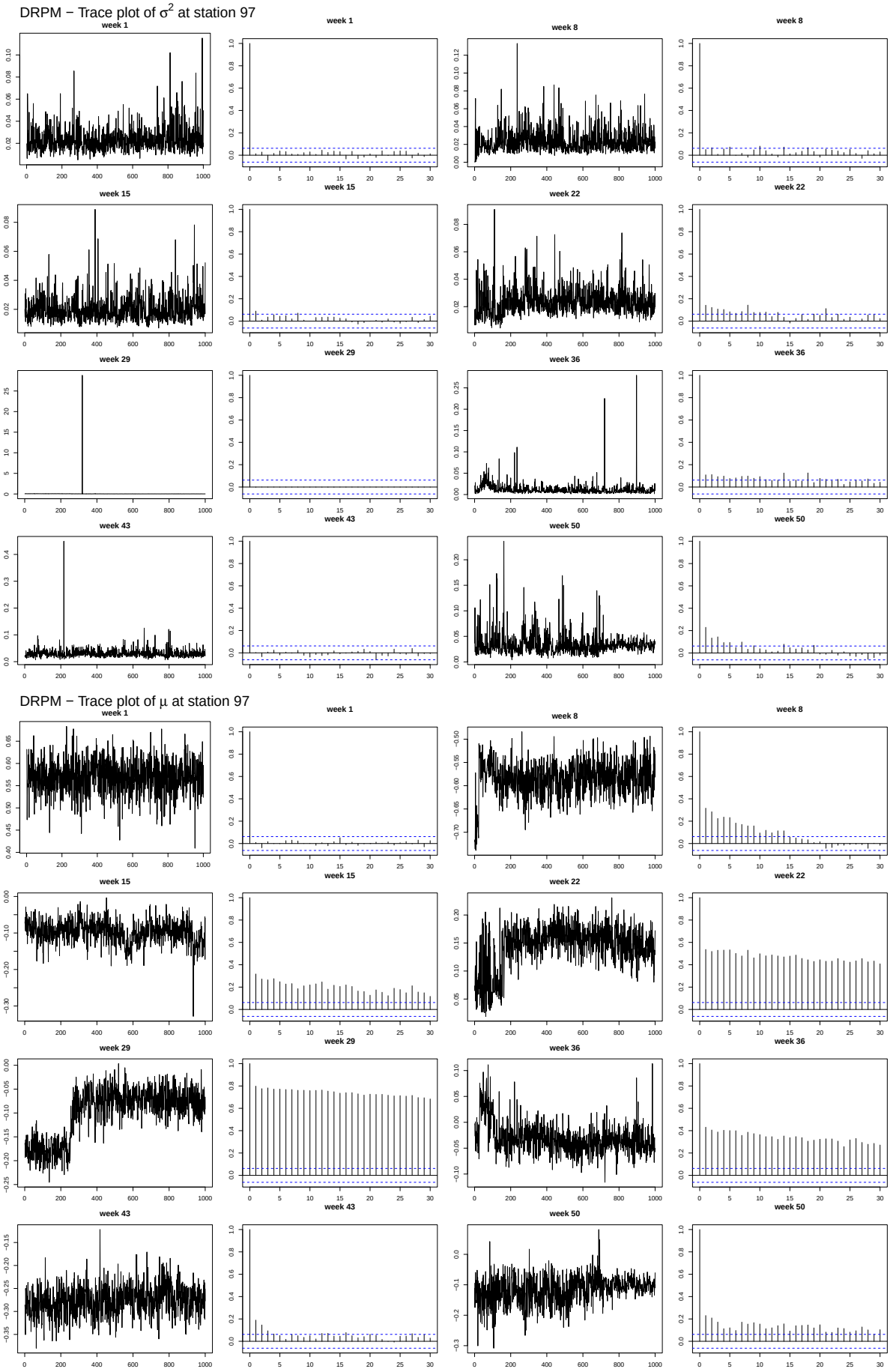

Figure 13: Trace and ACF plots of parameters σ^2 and μ of the DRPM model.

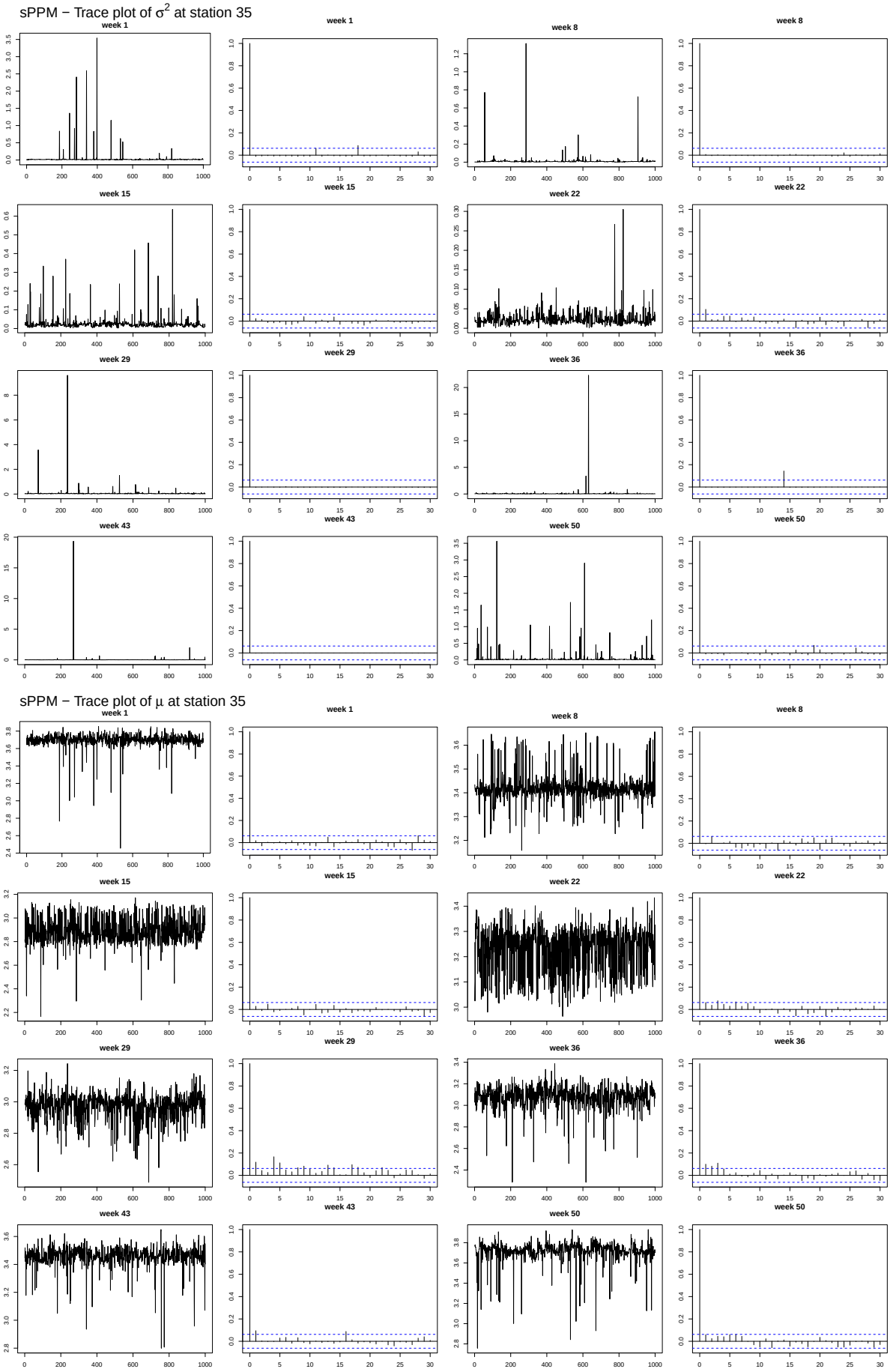

Figure 14: Trace and ACF plots of parameters σ^2 and μ of the sPPM model.

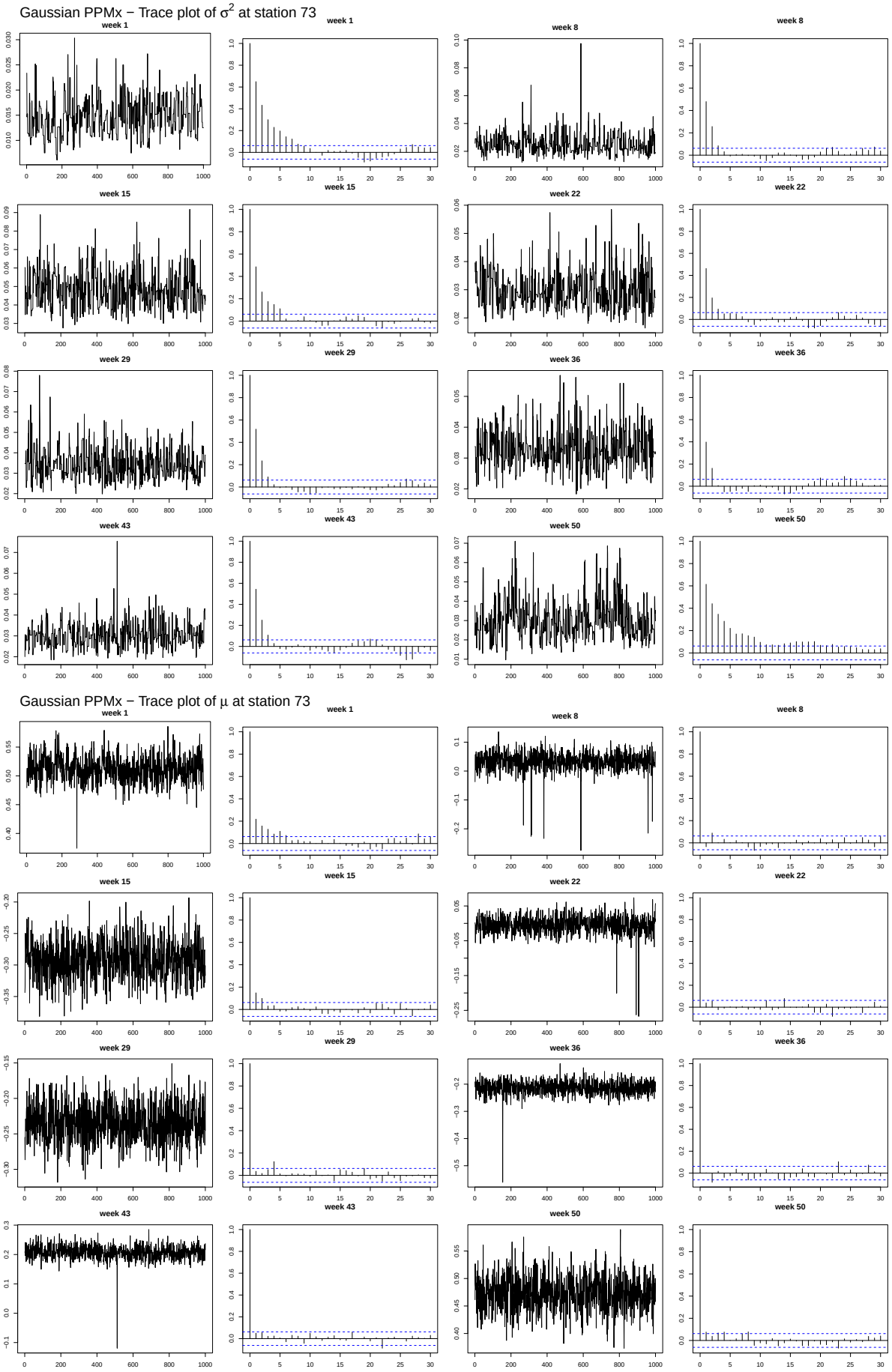

Figure 15: Trace and ACF plots of parameters σ^2 and μ of the Gaussian PPMx model.

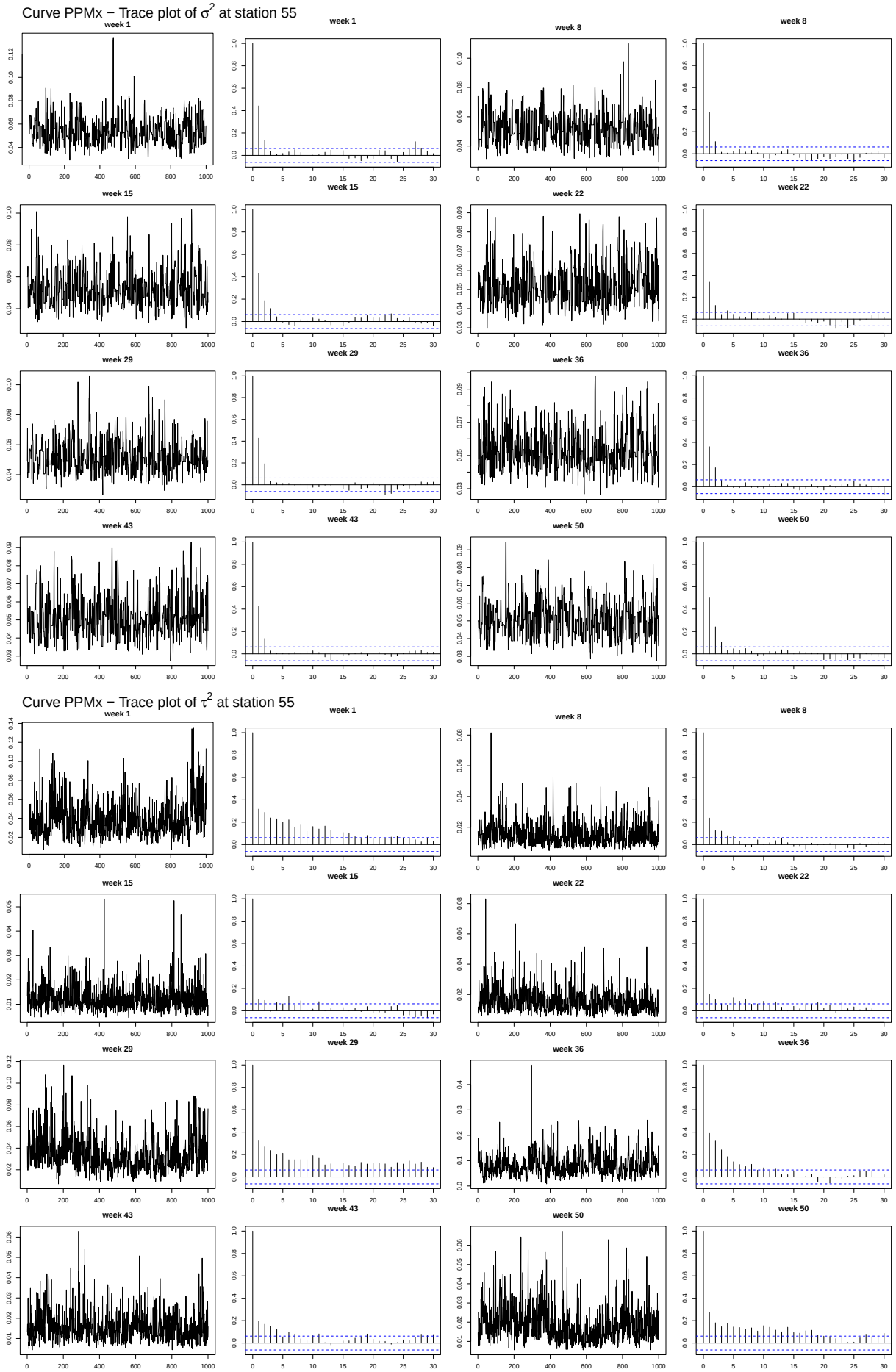

Figure 16: Trace and ACF plots of parameters σ^2 and τ^2 of the Curve PPMx model.