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 PM10, 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 PM10 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

we still need to finsh the introduction

[Com17] Particulate matter with a diameter of 10 micrometers or less, known as PM10, comprises small airborne particles sourced from various origins, posing potential health risks upon inhalation due to their ability to deeply penetrate the respiratory system. The meticulous monitoring of PM10 levels is imperative for comprehensive air quality assessment and the safeguarding of public health.

This paper embarks on a project with the overarching aim of identifying both natural and anthropogenic factors contributing to elevated PM10 levels. Employing a clustering analysis, our objective is to delineate distinct regions within Lombardy, unveiling discernible patterns influencing particulate levels.

Drawing upon data from the Agrimonia project, which encompasses diverse measurements, our focus centers on weekly averages across a one-year timeframe. Our analytical approach involves the utilization of various models, including DRPM and SPPM, alongside additional models for covariate selection.

In subsequent sections, we delve into the dataset cleaning process, present individual analyses for each model, and expound on our interpretation of results. Visualization plays a pivotal role in our exploration, with a particular emphasis on manual (visual) interpretation to extract nuanced insights.

It is essential to acknowledge the inherent limitations of our approach. Notably, our lack of technical expertise in the phenomenon necessitated a wholly data-driven analysis, underscoring the importance of contextualizing our findings within this parameter.

2. The dataset

this part maybe is too short? but dont know what else to add

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 total of 309,072 rows. The geographical coordinates, comprising longitude and latitude, are available for each station in the dataset.

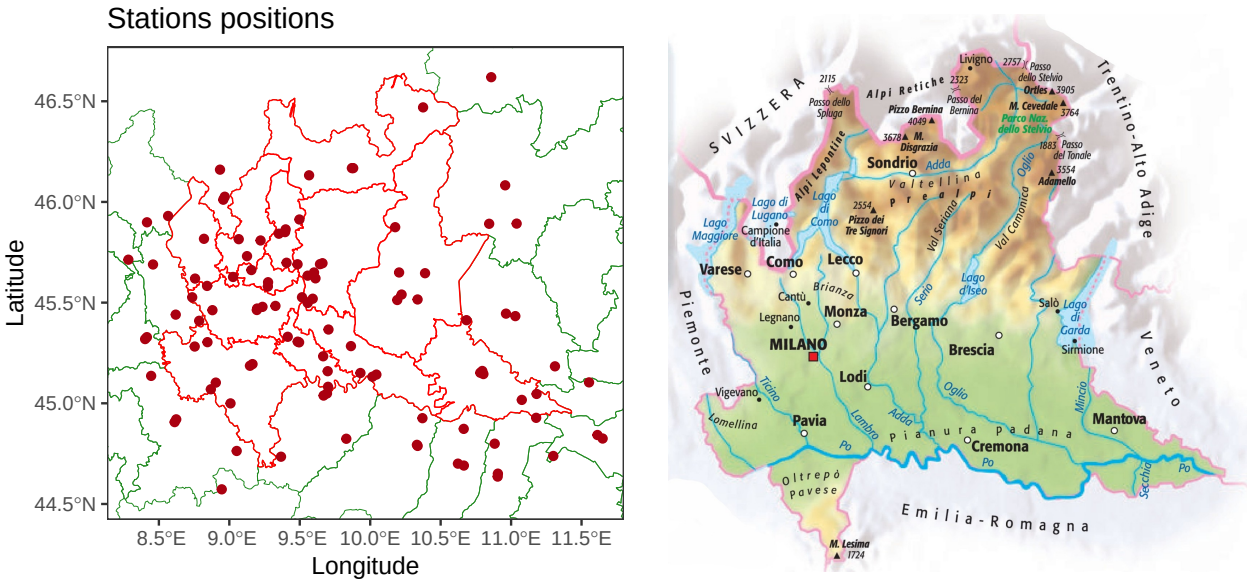

Figure 1: Map of the 105 selected stations after the data preprocessing (on the left), together with the physical map of the region under analysis (on the right).

2.1. Data inspection and processing

The goal proposed for our project was "clustering weekly data of one year of PM_{10}" and as such we started by selecting the year and then dividing the dataset by weeks, since originally it consisted of daily recordings. 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 were covariates and 5 were related to temporal and spatial coordinates, target and station id, from the original 43.

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

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.

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 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

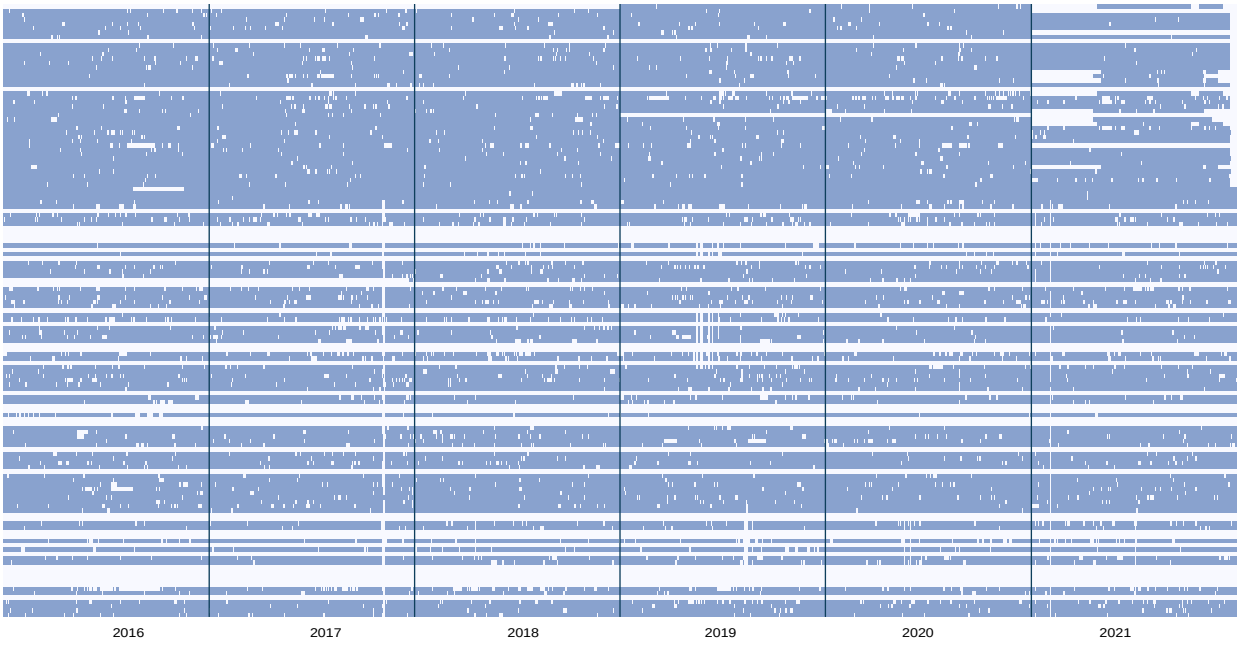

Figure 3: Heatmap of the missing data (in white) of the PM_{10} values recorded in the available dataset. On the rows there are all the original 141 stations, on the columns all the 2192 days composing the six years.

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. 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 PM10 levels in the Lombardy region.

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 by fitting a model with Gaussian likelihood and spatial PPM as prior on partitions:

$$\begin{aligned} Y_i|\mu^*, \sigma^*, c_i &\overset{\text{iid}}{\sim} N(\mu^*_{c_i}, \sigma^{*2}_{c_i}), \quad i = 1, \dots, n \\ \mu^*_j|\mu_0, \sigma_0^2 &\sim N(\mu_0, \sigma_0^2) \\ \sigma^*_j|A &\sim U(0, ms) \\ \mu_0|m, s^2 &\sim N(\mu_0, s0^2) \\ \sigma_0|B &\sim U(0, ms0) \\ \rho_m|M, \xi &\sim sPPM \end{aligned}$$

ρ_n denotes a partitioning on the n locations and its prior is defined by

$$\Pr(\rho_m|x) \propto \prod_{j=1}^{k_n} C(S_j)$$

where $C(S_j)$ for $S_j \subset \{1, ..., n\}$ is a cohesion function that measures how likely elements of S_j are clustered a priori. We analysed the 4 available cohesion functions available in the `PPMSuite` package:

We examined the 4 cohesion functions with different values for M (= 0.01, 0.1 and 1). For cohesion C_1 we considered both $\alpha = 1$ and $\alpha = 2$, for C_2 we set a equal to the median distance among all pairwise distances, for C_3 and C_4 we set the parameters of the Normal/Normal-Inverse-Wishart with $(0, 1, 2, I_2)$ as in the sppm paper. The analysis and selection of the cohesion functions were done using the dataset of the previous year (2017) and the best model was used to our data of interest (year 2018).

To make the method invariant to the scale of location, we standardized the coordinates to have zero mean and unit variance. Furthermore, to assess predictions, we divided the stations into 70% for the training 30% for the testing. We used the following metrics to compare the different models:

- MSE: the squared prediction error associated to the training data;
- MSPE: the squared prediction error associated to the testing data;
- LPML: the logarithm of the pseudo marginal likelihood, it is a goodness-of-fit metric that takes into account model complexity;
- WAIC: the Watanabe-Akaike information criterion.

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 comparisons for the different cohesion functions and different choice of M, data of 2017

Table 2 provides the results of the metrics for the different models. Since the variability of the metrics were similar and to be robust of outliers, we compare the median of the metrics

In terms of MSPE, LPML, WAIC, the best one was the cohesion function C4 with M = 0.1. So, finally, we fitted the sppm model using cohesion C_4 with M = 0.1 using our data of 2018.

3.2. DRPM model

The second model we focused on, outside of the `PPMSuite` package, 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 4.2, where we witnessed a more regular trend in the clusters definition for the DRPM model with respect to the other ones.

The model that we used, fully detailed in (1), 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.

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

About the target variable Y_{it}, they 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.

method				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.

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 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 therefore as third choice we implemented a Gaussian PPMx model, from the `PPMSuite` package like sPPM, developed in [PQ18].

The main idea is to calibrate the covariate-dependent partition model by capping the influence that covariates have on partition probabilities. The approach presented in [PQ18] was to temper covariate influence on clustering only through the partition prior distribution. The authors carried this out by calibrating the similarity functions.

A product partition distribution is a discrete probability distribution for ρ_m that is comprised of a so-called cohesion function $C(S) \geq 0$ for $S \subset \{1, \dots, m\}$. The cohesion measures the compactness of the elements in S and is used to produce the following unnormalized partition probabilities.

When covariates are available, they are incorporated in the model by introducing a nonnegative function $g(x_j^*)$.

The function $g(x_j^*)$ is called similarity function and measures the homogeneity of the $x_i \in \mathbf{x}_j$ by producing larger values for x's that are more similar.

The model can be defined as follows:

$$\begin{aligned} Y_i|\mu_j^*, \sigma_j^*, S_i &\overset{\text{iid}}{\sim} N(\mu_{S_i}^*, \sigma_{S_i}^{*2}) \quad i = 1, \ldots, m \\ \sigma_j^* &\sim U(0, A), \\ \mu_j^*|\mu_0, \sigma_0^2 &\overset{\text{iid}}{\sim} N(\mu_0, \sigma_0^2) \quad j = 1, \ldots, k_m \\ \sigma_0 &\sim U(0, A_0), \\ \mu_0 &\sim N(m, s^2), \\ \Pr(\rho_m|x) &\propto \prod_{j=1}^{K_m} c(S_j) g(x_j^*) \end{aligned}$$

In our model, we decided to use the the Auxilliary *N-NIG* similarity function. It is defined as follows:

$$g(x_j) = \int \left(\prod_{i \in S_j} q(x_i|\xi_{\mathbf{j}}^*) \right) q(\xi_{\mathbf{j}}^*) d\xi_{\mathbf{j}}^*$$

where q$(\cdot|\cdot)$ and q$(\cdot)$ are density functions, $\xi_{\mathbf{j}}^* = (m_j^*, v_j^*)$ and $q(\xi_{\mathbf{j}}^*) = q(m_j^*, v_j^*) =$ N-IG$(m_j^*, v_j^*|m_0, k_0, v_0, n_0)$, the Normal–Inverse-Gamma density function which is parametrized so that m_0, v_0 are a priori "guesses" for m_j^* and v_j^* and k_0, n_0 the corresponding a priori "sample sizes."

The initial step we took to implement the model was to select the covariates. In order to make the results as interpretable as possible, we considered a subset of five variables that maximized the LPML. Specifically, before obtaining the final model, we considered preliminary models aimed solely at variable selection.

Firstly, we compared 31 different models, each characterized by a single variable. The first variable selected, was the one corresponding to the model with the highest LPML: `Altitude`. Secondly, we considered 30 different models, each characterized by the first chosen variable and another variable present in the dataset. The second variable chosen was the one corresponding to the model with two covariates with the highest LPML: `EM_nox_sum`.

By repeating this procedure three more times, we obtained the best subset of five variables in terms of LPML, consisting of: `Altitude`, `EM_nox_sum`, `WE_mode_wind_direction_100m`, `WE_wind_speed_100m_max` and `LA_lvi`.

Once we finalized the set of covariates to use in the model, we focused on the choice of priors and other required parameters.

At the end, 4300 MCMC iterates were collected, with a burn-in of 300.

In this manner, we achieved favorable trends in nearly all of our ACF plots (refer to Appendix B for them), thus reinforcing the value of our efforts, the decisions made, and the outcomes obtained.

3.4. Curve PPMx model

The last model that we considered was the Functional Gaussian PPMx model which was implemented in `PPMSuite` package using the `curve_ppmx` function. This function applies a hierarchical functional data model, wherein B-spline coefficients undergo clustering utilizing either a PPM or a PPMx prior approach on partitions using Gaussian PPMx approach. The details of latter model is provided in section 3.3. The PPM and PPMx priors are used to group similar functional curves together based on their characteristics. The PPMx prior integrates the concept that individuals sharing similar covariate values are more likely to be grouped together [PQ15]. The use of B-splines allows for the flexible clustering of functional realizations based on the characteristics of the B-spline coefficients and covariates.

Since the model uses Gaussain PPMx model approach, we ran the model with the same covariates mentioned in sections 3.3. We fitted the model for each week of 53 weeks of the year so that we can observe how the clusterings change over time. The resulting clusters are provided in section 4.1.

Having different hyperparameters of the function allowed us to test different models and choose the model with the highest LPML and lowest WAIC values. We have tested 3 values of Scale Parameter M, which is related to the dispersion parameter of a Dirichlet process, and different number knots used for the splines. The summary of the tests are reported in Table 4

M	number of 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.

According to those tests we have chosem $M = 0.1$ with number of knots of splines being 40. The time needed to run all the fits with the best hyperparameters for the entire timeframe was around 2 hours, with number of iterations being 10000, number of burn-in being 5000 and thinning parameter being equal 5. Those parameters allowed us to have the good trend of almost all our ACF plots (see Appendix B for them).

3.5. Linear Model

We thought it could prove useful to implement also a simpler baseline model to use for comparison, variable selection and to better understand the data, that could allow us to try a vast range of methods faster than more complex models.

Being simpler and allowing the inclusion of covariates, we chose the linear approach, actually implementing a linear model for each station. The models considered the numerical covariates linearly but tried to allow more variability on the time considering also its sine, cosine and square.

The first idea was to try a clustering on the linear model, maybe grouping together the stations according to the betas they assigned to each variable but it was soon discarded as we deemed it redundant, since we already had different models more fitted for a more precise and explainable clustering, and we decided to focus on variable selection instead. Since with so many covariates it would have been extremely long and computationally heavy to try methods based on partitions of combinations of covariates or spike and slab, the model was firstly implemented through jags to try and use another kind of selection method as seen in [KM98]. In particular, it 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, 0 otherwise. Our idea was to select a quantile, keep only the corresponding column as a vector, and sum element-wise for all stations, expecting an higher value in correspondence of a useful covariate to select, and a significantly lower for covariates discarded by most stations. Unfortunately this first 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 all variables had been selected only for about half the stations, and none was significantly more important than the others.

For this reason we tried 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 total precipitations, and a more moderate but still interesting on the livestock, lvi and hvi (related to total green area per unit horizontal ground surface area for low and high vegetation type) variables.

4. Models comparison

4.1. Cluster trends

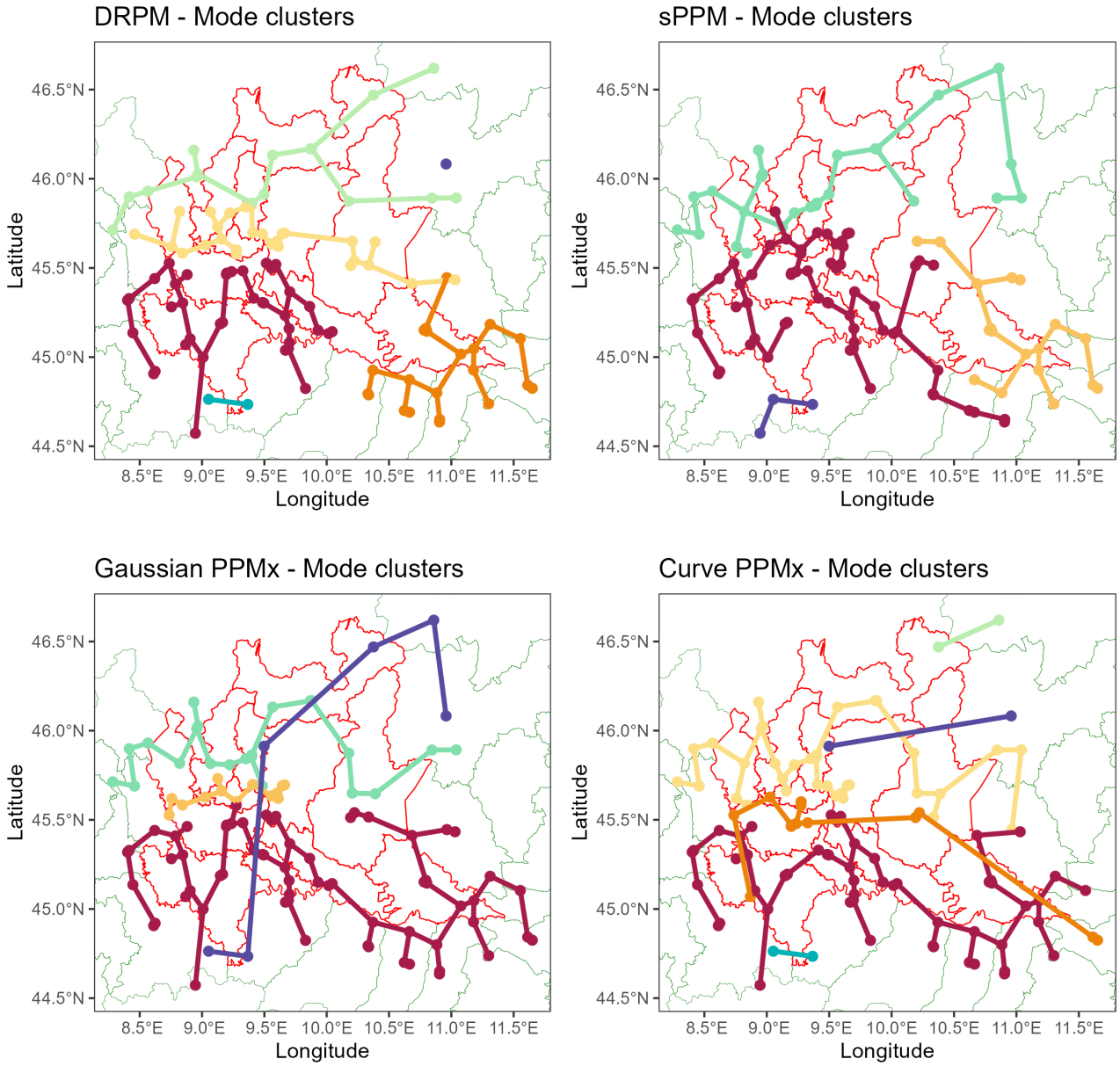

Figure 4: Maps of the most frequent clusters, throughout the 53 weeks of 2018, for all the models. See `https://federicomor.github.io/assets/figures/visualize.html` for a more detailed analysis of the plots (e.g. for all the week by week clusterings).

4.2. ARI metric DONE

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.

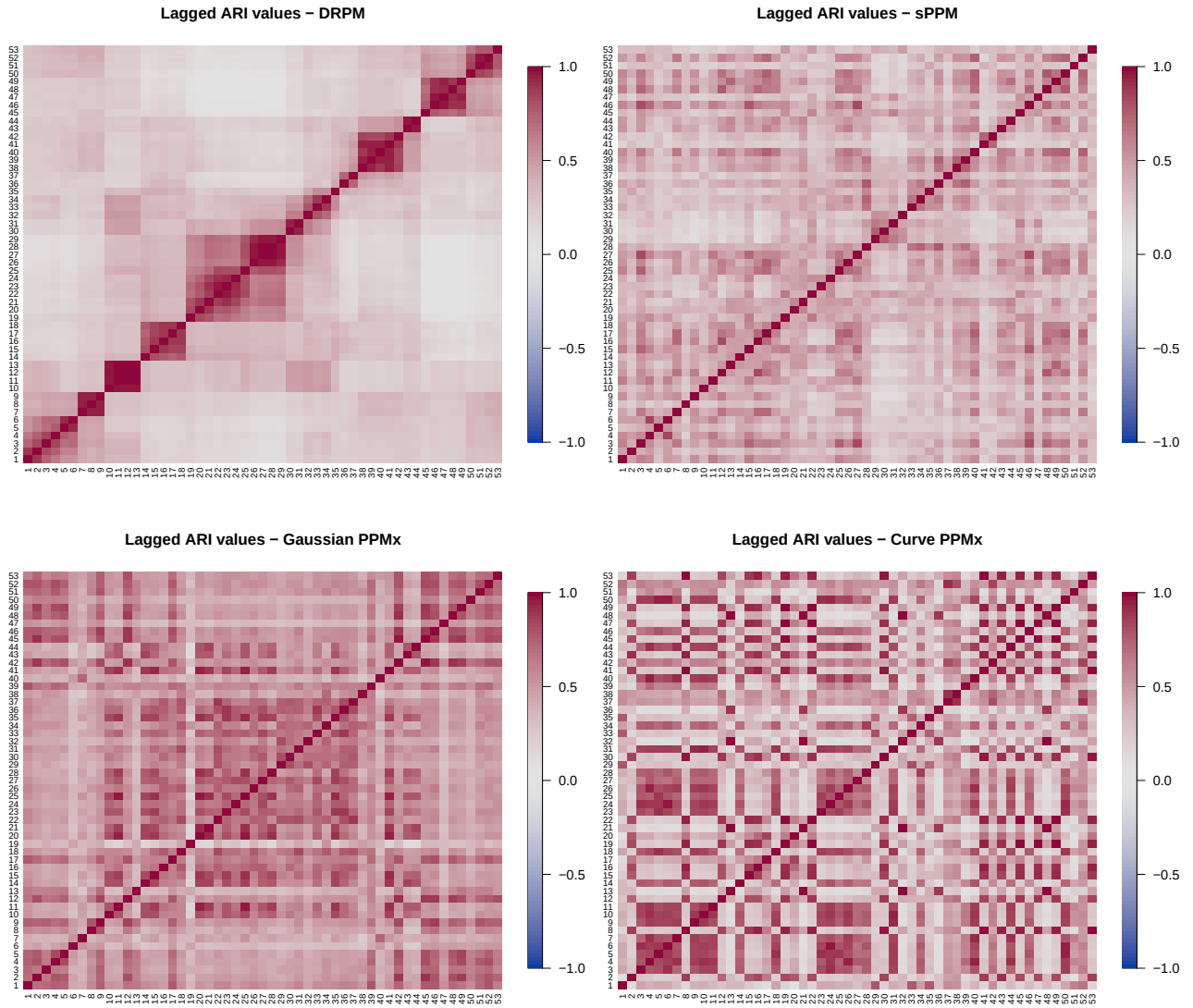

Figure 5: Lagged ARI values of the four models.

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).

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 5, 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.

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 6, 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

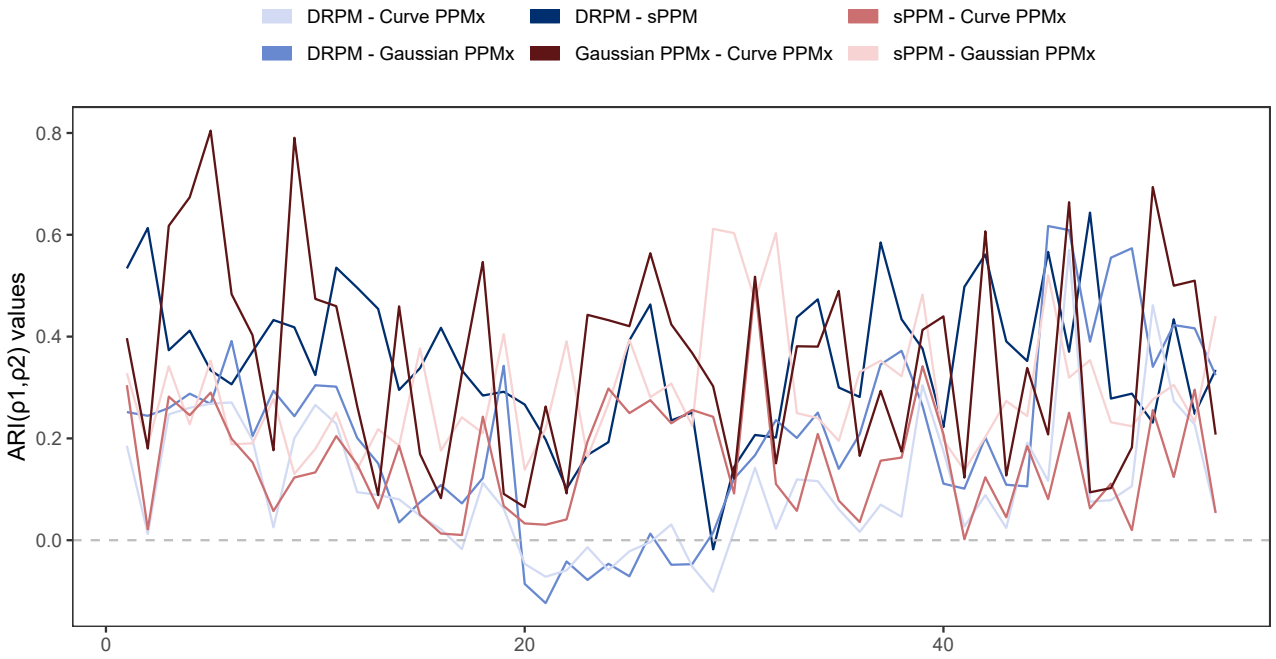

Figure 6: Plot of the ARI values for each pairwise comparison of the models, for all the weeks.

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 7. 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. Analysis of the results

5.1. Differences among the different clusterings

5.2. Interpretation of the obtained results

Almost all models exhibited a stratification pattern. Regions with flat terrain generally displayed high PM10 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 outliers emerged consistently across all models, clustering together but separately from the surrounding Milan cluster. These stations exhibited lower PM10 concentrations and high levels of `LA_lvi` (indicating significant vegetation presence).

For a detailed interpretation of the clusters, we focused our attention on a few variables that are easy to interpret and have dense significance.

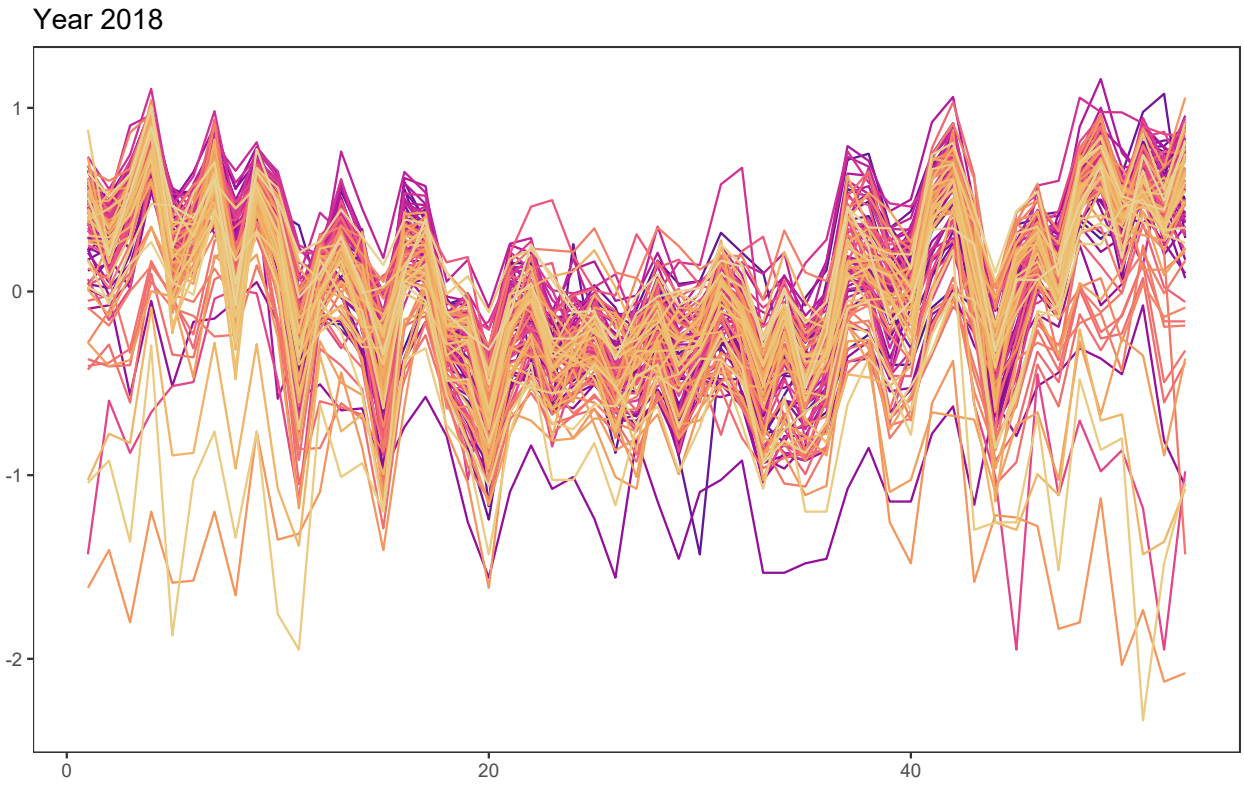

Figure 7: Trend of the PM_{10} values in the 2018 year, for the 105 selected stations, after the data processing step (log-transformed and centered).

Altitude and vegetation:

High and low vegetation are complementary: at higher altitudes, we find dense vegetation, while at lower altitudes vegetation is sparse. The growth patterns of both types of vegetation are consistent throughout the year, with high levels in summer and low levels in winter. We can interpret this effect in two ways:

- It's the altitude that brings the PM10 down.
- In high altitudes, we have higher vegetation, which contributes to the reduction of PM10 levels.

In general, high vegetation and high altitude are generally correlated with lower PM10 levels.

Total precipitations:

The total precipitation appears to influence the reduction of PM10 levels, as evidenced by the events observed in October, characterized by peaks in rainfall followed by a decrease in PM10 levels.

Max intensity of the wind speed at a height of 10 m:

High intensity of the wind increases PM10 levels, but this effect is mitigated by the presence of rain. For instance, in October, despite high levels of both wind and rain, PM10 levels decrease. In December, when precipitation levels were low and the influence of rain could be filtered out, high wind speeds were observed to elevate PM10 levels.

Emissions of NO_x:

It's positively correlated with PM10, with higher levels observed during winters when PM10 levels 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.

Emissions of NH_3 and livestock:

In the Milan (and Mantua) area we noticed a strong presence of NH_3 especially attributable to livestock farming,

particularly during the central part of the year. Presumably, the presence of NH_3 from livestock contributes to the elevation of PM10 levels. Naturally, this area corresponds to a high number of animals.

To conclude our analysis, we deemed it important to analyze the presence of certain small clusters in all models. We can consider them as outlier stations. In fact, their PM10 behavior differs significantly from the others: the trend is opposite, with low levels in winter and high levels in summer. These stations are situated in high-altitude regions. The Gaussian PPMx model tends to group the stations together in one single cluster, while the others keep them separated into smaller and isolated clusters. Additionally, these outliers exhibit the lowest values of PM10 and `EM_nox`. Possible causes of the inverse trend include their location in higher regions, suggesting the role of snow in lowering levels during the winter months.

6. Conclusions

7. Further developements

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

- *Utilize Previous Year Data for Model Priors:* consider incorporating data from the preceding year to establish priors for the models. While our current dataset facilitated model convergence, integrating historical data could offer additional insights and refine the robustness of our findings.
- *Distinguish Between Weekends and Weekdays:* exploring the impact of human activities on PM10 levels by stratifying the analysis between weekends and weekdays. This differentiation may uncover patterns associated with specific human-related factors, contributing to a more nuanced understanding of particulate matter dynamics.
- *Ensemble Modeling:* exploring the potential benefits of ensemble modeling by combining the outputs of different models. This approach can enhance the overall accuracy and reliability of our clustering analysis. By leveraging the strengths of individual models, we can obtain a more comprehensive and robust estimation of the identified clusters.

These proposed extensions aim to further refine our methodology, enrich the interpretability of results, and provide a more comprehensive understanding of the intricate factors influencing PM10 levels in the Lombardy region.

Also, we can mentioned the development code for the drpm model with covariates:) Since we already have a theory

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`.

[CTAa] CTAN. BiBTeX documentation.

[CTAb] CTAN. pgf – create PostScript and PDF graphics in TEX.

[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.

[Knu74] Donald E. Knuth. Computer programming as an art. *Commun. ACM*, pages 667–673, 1974.

[Knu92] Donald E. Knuth. Two notes on notation. *Amer. Math. Monthly*, 99:403–422, 1992.

[Kot15] Stefan Kottwitz. *LaTeX Cookbook*. Packt Publishing Ltd, 2015.

[Lam94] Leslie Lamport. *LaTeX: A Document Preparation System*. Pearson Education India, 1994.

[OPHS95] Tobias Oetiker, Hubert Partl, Irene Hyna, and Elisabeth Schlegl. The not so short introduction to latex2ε. *Electronic document available at http://www. tex. ac. uk/tex-archive/info/lshort*, 1995.

[PQ15] Garritt L. Page and Fernando A. Quintana. Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. *Bayesian Analysis*, 10:379–410, Jun 2015.

[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.

A. Appendix A

A pivotal component of a robust statistical analysis lies in the effective interpretation of results. To address this crucial aspect, we meticulously constructed a library of auxiliary functions, empowering us to visually scrutinize various facets of our research.

Given the inherent temporal and spatial dimensions of our dataset, we opted for a dynamic approach, creating videos instead of static images to seamlessly navigate the temporal component.

For the visualisation of spatial variables, we devised two principal tools: a grid map and an expanding circles plot.

1. **Grid Map:** - This tool harnesses a distinct dataset featuring measurements across the entire region, organized on a grid of evenly spaced points. It offers a panoramic overview of key variables, such as Altitude and Weather measurements, providing a comprehensive understanding of spatial patterns. 2. **Expanding Circles Plot:** - Focused on station-level measurements, this tool illustrates the magnitude of variables by employing radius and color intensity of circles centered around each station. This approach grants us insights into localized patterns, enhancing our comprehension of variable distribution across the region.

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, a strategic approach chosen to yield a more organized and visually coherent representation of the clusters.

B. MCMC diagnostics DONE

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.

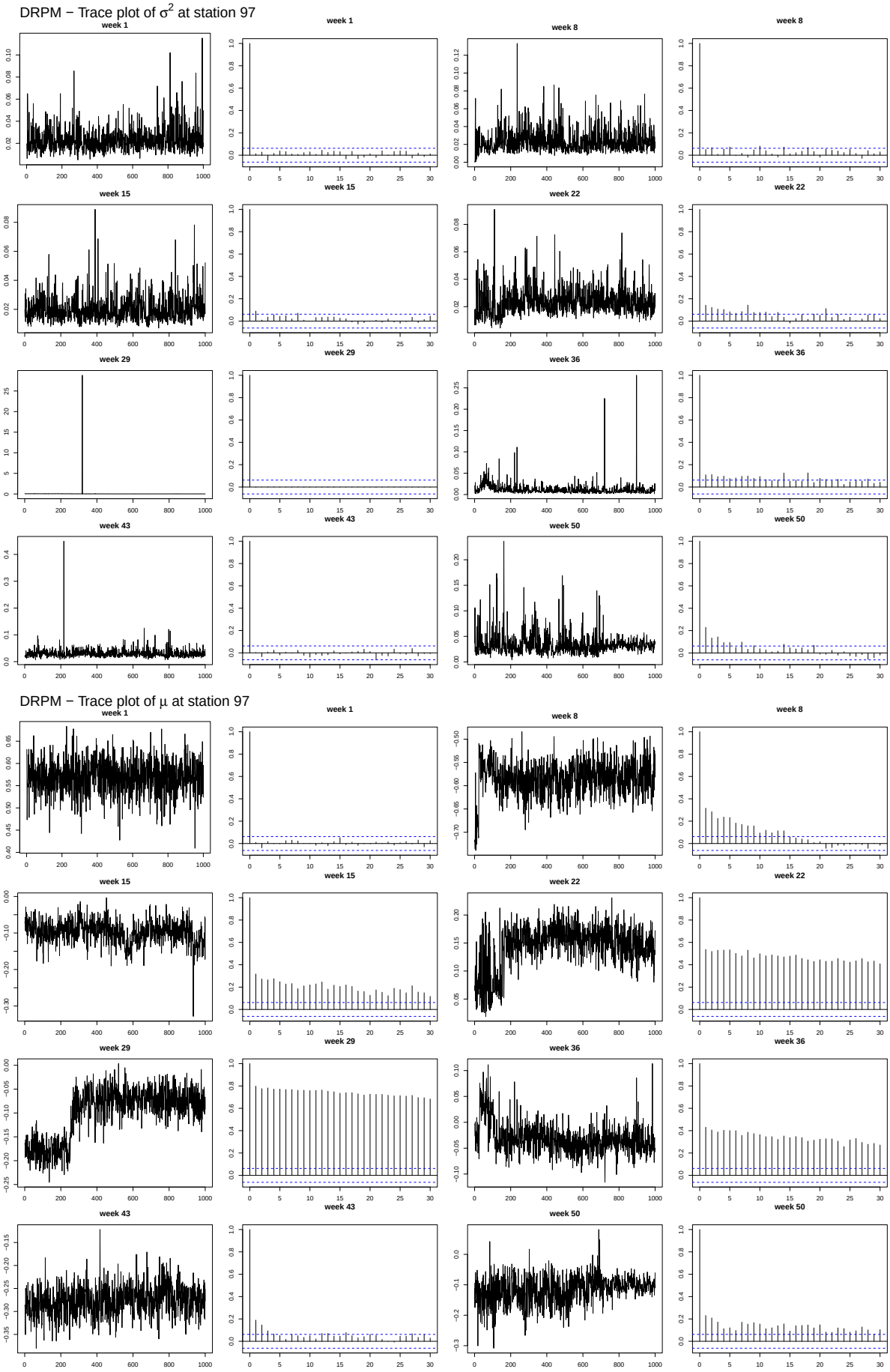

Figure 8: Trace and ACF plots of parameters σ^2 and μ of the DRPM model.

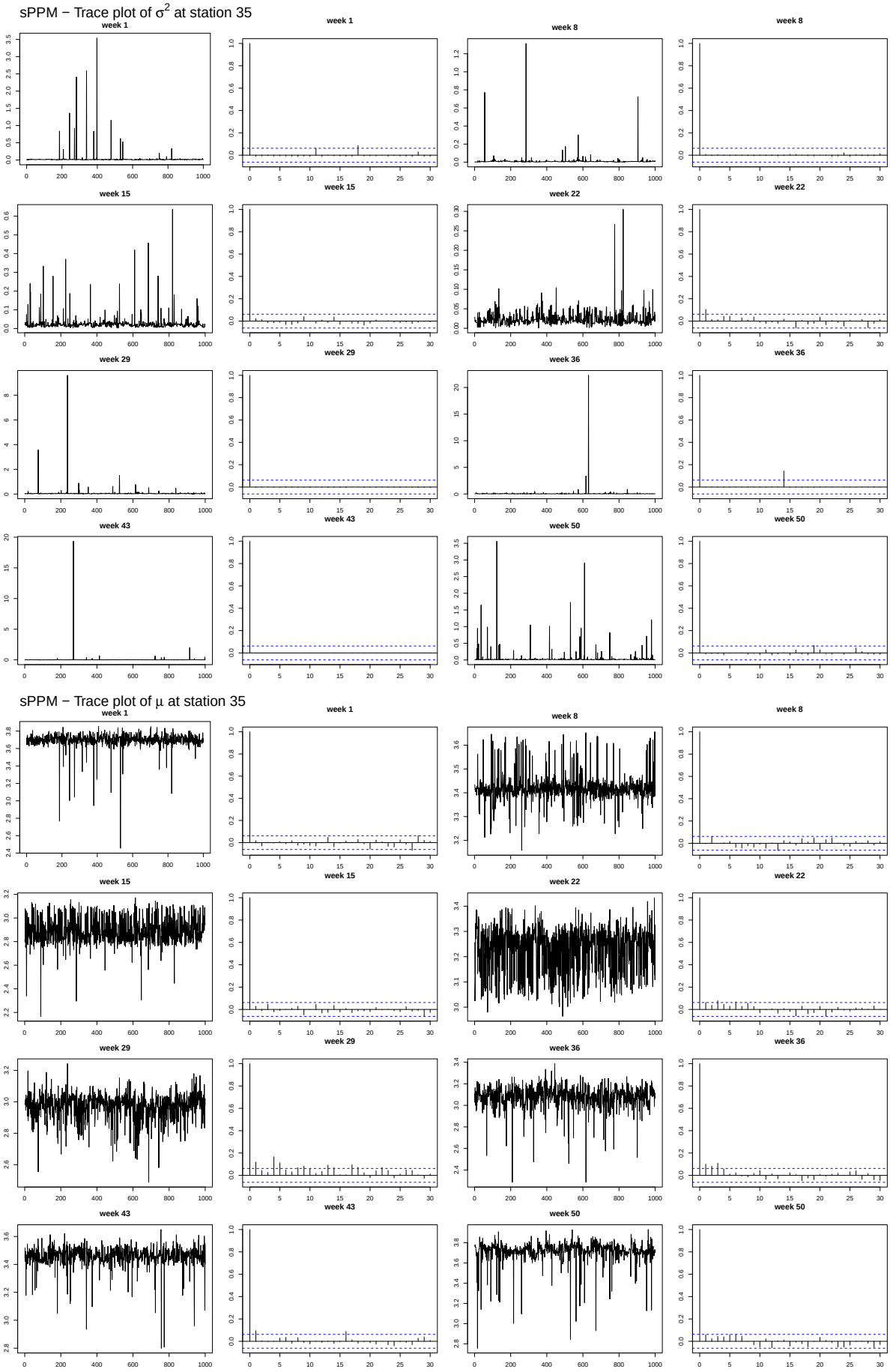

Figure 9: Trace and ACF plots of parameters σ^2 and μ of the sPPM model.

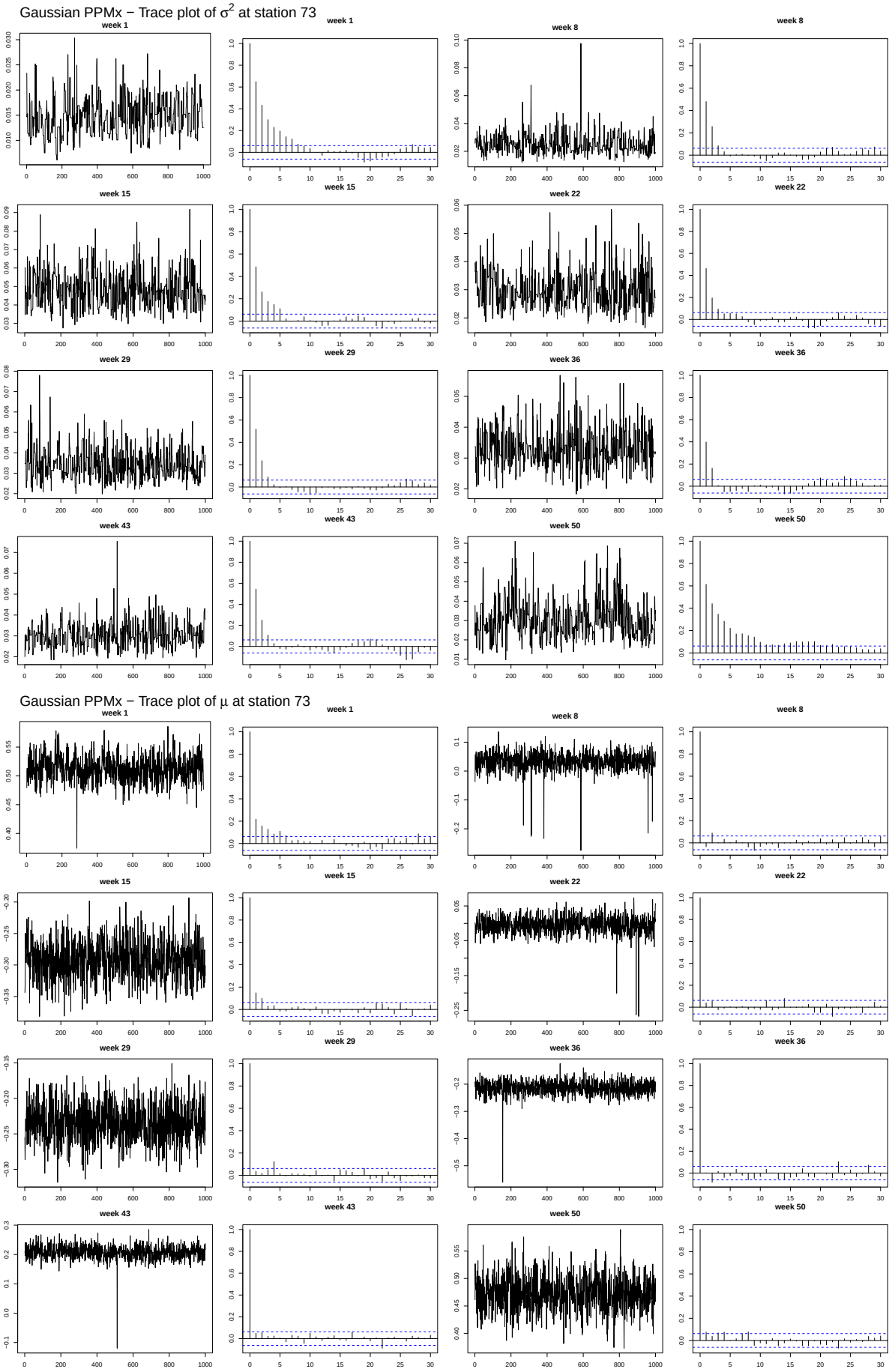

Figure 10: Trace and ACF plots of parameters σ^2 and μ of the Gaussian PPMx model.

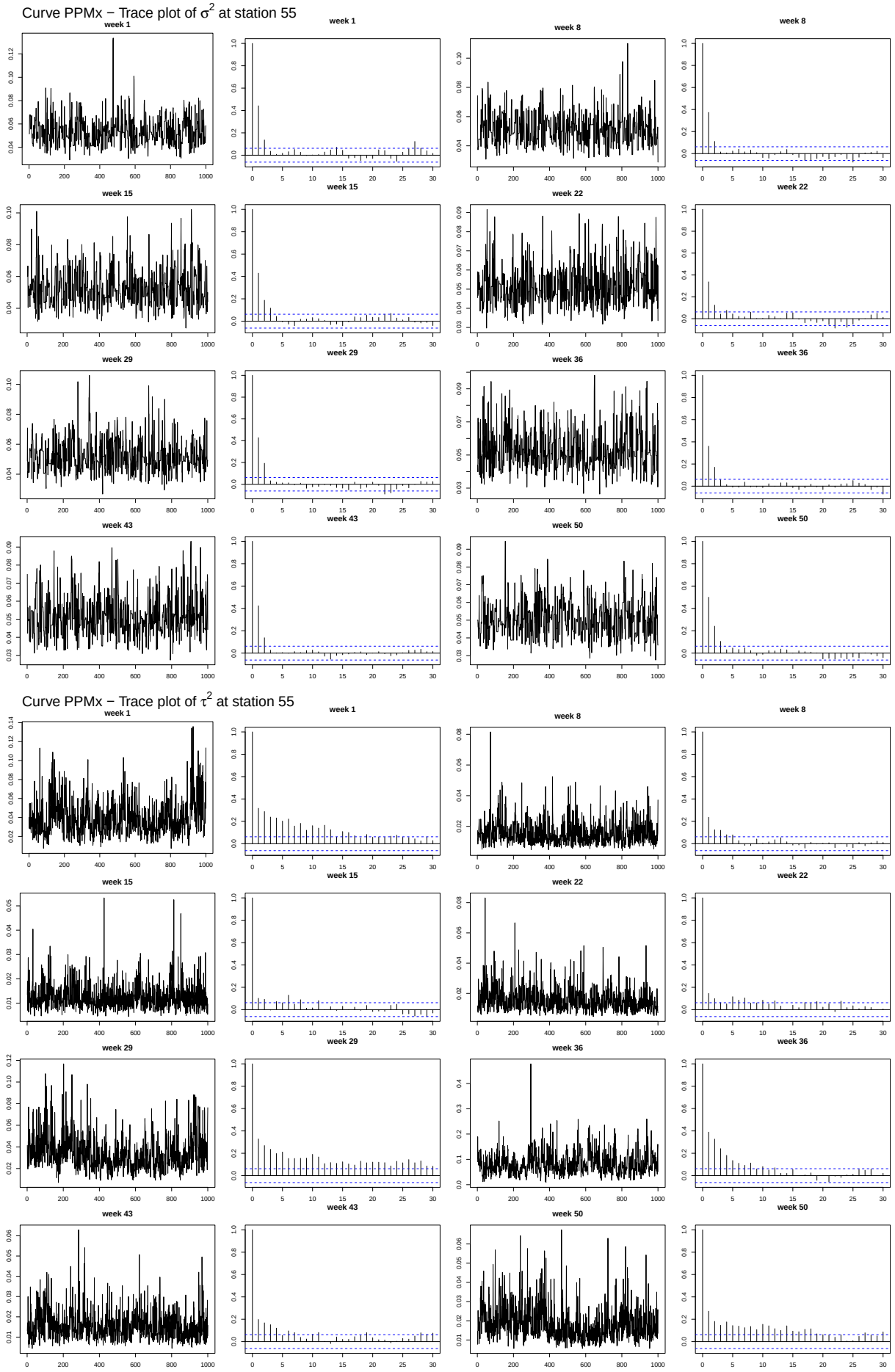

Figure 11: Trace and ACF plots of parameters σ^2 and τ^2 of the Curve PPMx model.

==========

POLIMI TEMPLATE EXAMPLE

C. Introduction

This document is intended to be both an example of the Polimi LaTeX template for Master Theses in article format, as well as a short introduction to its use. It is not intended to be a general introduction to LaTeX itself, and the reader is assumed to be familiar with the basics of creating and compiling LaTeX documents (see [OPHS95, Kot15]).
The cover page of the thesis in article format must contain all the relevant information: title of the thesis, name of the Study Programme, name(s) of the author(s), student ID number, name of the supervisor, name(s) of the co-supervisor(s) (if any), academic year.
Be sure to select a title that is meaningful. It should contain important keywords to be identified by indexer. Keep the title as concise as possible and comprehensible even to people who are not experts in your field. The title has to be chosen at the end of your work so that it accurately captures the main subject of the manuscript.

It is convenient to break the article format of your thesis (in article format) into sections and subsections. If necessary, subsubsections, paragraphs and subparagraphs can be used. A new section is created by the command

```
\section{Title of the section}
```

The numbering can be turned off by using `\section*{}`. A new subsection is created by the command

```
\subsection{Title of the subsection}
```

and, similarly, the numbering can be turned off by adding an asterisk as follows

```
\subsection*{}
```

It is recommended to give a label to each section by using the command

```
\label{sec:section_name}%
```

where the argument is just a text string that you'll use to reference that part as follows: *Section C contains* INTRODUCTION ....

D. Equations

This section gives some examples of writing mathematical equations in your thesis.

Maxwell's equations read:

$$\begin{cases} \nabla \cdot \boldsymbol{D} = \rho, & \text{(2a)} \\ \nabla \times \boldsymbol{E} + \dfrac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}, & \text{(2b)} \\ \nabla \cdot \boldsymbol{B} = 0, & \text{(2c)} \\ \nabla \times \boldsymbol{H} - \dfrac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}. & \text{(2d)} \end{cases}$$

Equation (2) is automatically labeled by `cleveref`, as well as Equation (2a) and Equation (2c). Thanks to the `cleveref` package, there is no need to use `\eqref`. Equations have to be numbered only if they are referenced in the text.

Equations (3), (4), (5), and (6) show again Maxwell's equations without brace:

$$\nabla \cdot \boldsymbol{D} = \rho, \tag{3}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}, \tag{4}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{5}$$

$$\nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}. \tag{6}$$

Equation (7) is the same as before, but with just one label:

$$\begin{cases} \nabla \cdot \boldsymbol{D} = \rho, \\ \nabla \times \boldsymbol{E} + \dfrac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}, \\ \nabla \cdot \boldsymbol{B} = 0, \\ \nabla \times \boldsymbol{H} - \dfrac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}. \end{cases} \tag{7}$$

E. Figures, Tables and Algorithms

Figures, Tables and Algorithms have to contain a Caption that describes their content, and have to be properly referred in the text.

E.1. Figures

For including pictures in your text you can use `TikZ` for high-quality hand-made figures [CTAb], or just include them with the command

```
\includegraphics[options]{filename.xxx}
```

Here xxx is the correct format, e.g. `.png`, `.jpg`, `.eps`, ...

Figure 12: Caption of the Figure.

Thanks to the `\subfloat` command, a single figure, such as Figure 12, can contain multiple sub-figures with their own caption and label, e.g. Figure 13a and Figure 13b.

(a) One PoliMi logo.

(b) Another one PoliMi logo.

Figure 13: Caption of the Figure.

E.2. Tables

Within the environments `table` and `tabular` you can create very fancy tables as the one shown in Table 5.

Example of Table (optional)

	column1	column2	column3
row1	1	2	3
row2	α	β	γ
row3	alpha	beta	gamma

Table 5: Caption of the Table.

You can also consider to highlight selected columns or rows in order to make tables more readable. Moreover, with the use of `table*` and the option `bp` it is possible to align them at the bottom of the page. One example is presented in Table 6.

E.3. Algorithms

Pseudo-algorithms can be written in LaTeX with the `algorithm` and `algorithmic` packages. An example is shown in Algorithm 1.

Algorithm 1 Name of the Algorithm

```
Initial instructions
for for − condition do
    Some instructions
    if if − condition then
        Some other instructions
    end if
end for
while while − condition do
    Some further instructions
end while
Final instructions
```

F. Some further useful suggestions

Theorems have to be formatted as follows:

Theorem F.1. *Write here your theorem.*

Proof. If useful you can report here the proof.

Propositions have to be formatted as follows:

Proposition F.1. *Write here your proposition.*

How to insert itemized lists:

	column1	column2	column3	column4	column5	column6
row1	1	2	3	4	5	6
row2	a	b	c	d	e	f
row3	α	β	γ	δ	ϕ	ω
row4	alpha	beta	gamma	delta	phi	omega

Table 6: Highlighting the columns

- first item;
- second item.

How to write numbered lists:

1. first item;
2. second item.

G. Use of copyrighted material

Each student is responsible for obtaining copyright permissions, if necessary, to include published material in the thesis. This applies typically to third-party material published by someone else.

H. Plagiarism

You have to be sure to respect the rules on Copyright and avoid an involuntary plagiarism. It is allowed to take other persons' ideas only if the author and his original work are clearly mentioned. As stated in the Code of Ethics and Conduct, Politecnico di Milano *promotes the integrity of research, condemns manipulation and the infringement of intellectual property*, and gives opportunity to all those who carry out research activities to have an adequate training on ethical conduct and integrity while doing research. To be sure to respect the copyright rules, read the guides on Copyright legislation and citation styles available at:

`https://www.biblio.polimi.it/en/tools/courses-and-tutorials`

You can also attend the courses which are periodically organized on "Bibliographic citations and bibliography management".

I. Conclusions

A final section containing the main conclusions of your research/study and possible future developments of your work have to be inserted in the section "Conclusions".

J. Bibliography and citations

Your thesis must contain a suitable Bibliography which lists all the sources consulted on developing the work. The list of references is placed at the end of the manuscript after the chapter containing the conclusions. It is suggested to use the BibTeX package and save the bibliographic references in the file `bibliography.bib`. This is indeed a database containing all the information about the references. To cite in your manuscript, use the `\cite{}` command as follows:
Here is how you cite bibliography entries: [Knu74], or multiple ones at once: [Knu92, Lam94].
The bibliography and list of references are generated automatically by running BibTeX [CTAa].