Bayesian Clustering Analysis of PM_{10} in Lombardy

Exploring Spatio-Temporal Trends and Covariates Effects

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

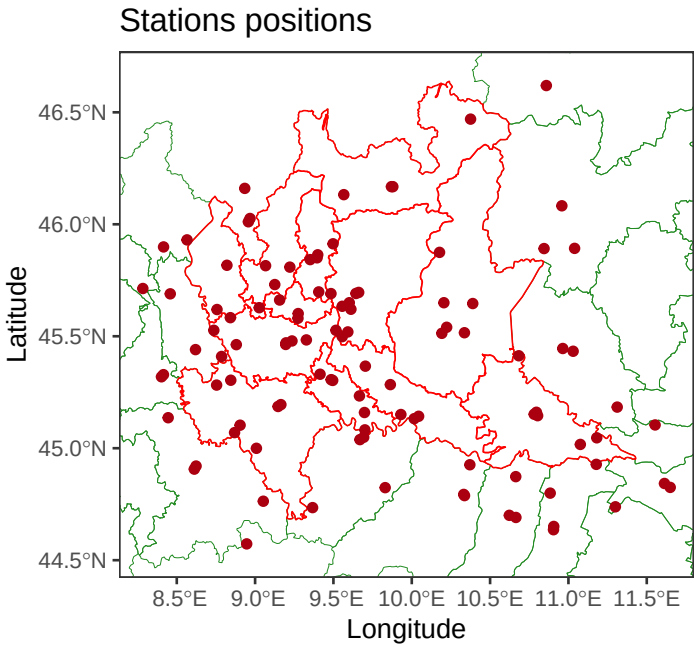
Politecnico of Milano
Bayesian Statistics course

February 15, 2024

github project repository: `https://github.com/federicomor/progetto-bayesian`
visualization page: `https://federicomor.github.io/assets/figures/visualize.html`

Presentation Flow

Dataset and Goal



Dataset: AGRIMONIA project
Developed by Agriculture Impact On Italian Air (AGRIMONIA) project to assess the impact of livestock on air quality
Five groups of data: air quality (AQ), weather and climate (WE), pollutants' emissions (EM), livestock (LI) and land and soil characteristics (LA)

Our goal: clustering weekly data of one year of PM_{10} using different Bayesian models.

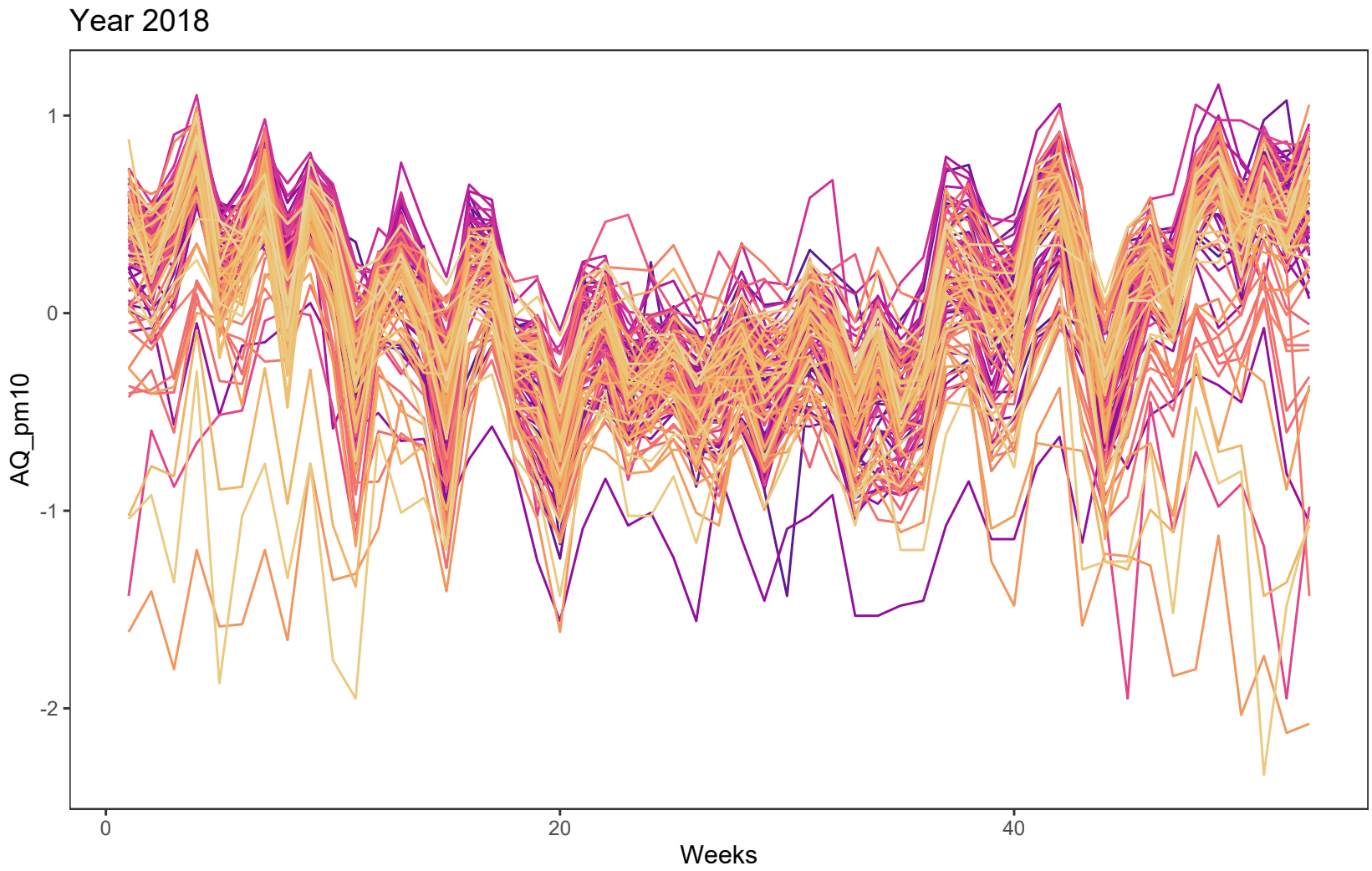
Missing data analysis

We removed some stations that were not measuring PM_{10} values and the almost-empty covariates of the AQ group and `LA_soil_use`.
We selected the year 2018, as it had few missing values with respect to the others and was not part of the pandemic period.

Some sparse missing data we removed during the weekly averaging process, averaging over the present values to avoid a double approximation.

Heatmap of the missing values of all the variables in the available dataset

PM_{10} trend after the data processing



Models summary

We looked into models which could tailor the complex nature of our data: spatial and temporal information plus covariates, with a clustering target in mind.
Unfortunately, there was no "holy grail" which could manage to tame all those levels together, but our results are still satisfactory.

We will now see them in more details, but for now this is a clear preview:

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

sPPM

Purely spatial model.

The starting point.

$$
\begin{aligned}
Y(\boldsymbol{s}_i)|c_i, \mu^{\star}_{c_i}(\boldsymbol{s}_i), \sigma^2 &\overset{\text{ind}}{\sim} \mathcal{N}(\mu^{\star}_{c_i}(\boldsymbol{s}_i), \sigma^2) \quad i = 1, \ldots, n \\
\mu^{\star}_h(\boldsymbol{s}_i) &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2) \quad h = 1, \ldots, k_n \\
\sigma &\sim \mathcal{U}(0, A) \\
\mu_0 &\sim \mathcal{N}(m, s_0^2) \\
\sigma_0 &\sim \mathcal{U}(0, B) \\
\{c_i\}_{i=1}^n &\sim \text{sPPM} \\
\Pr(\rho_n|\boldsymbol{s}) &\propto \prod_{h=1}^{k_n} C(S_h, \boldsymbol{s}^{\star}_h)
\end{aligned}
$$

sPPM

There were 4 possible cohesion functions and different choices of the parameters; we compared them using data from the previous year to select the best combination of parameters for the fit.

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

DRPM

Spatial and temporal model.
The only model which includes time.

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

DRPM

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.

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

Gaussian PPMX

Model with covariates.

Five covariates chosen with a forward selection method: `Altitude`, `EM_nox_sum`, `WE_mode_wind_direction_100m` (categorical), `WE_wind_speed_100m_max` and `LA_lvi`.

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

Curve PPMx

Functional version of Gaussian PPMx.

Now the data are no more point realizations but functional curves, with the clustering method (still including covariates) applying to the B-spline coefficients of the observations.

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

For the rest it maintains the same architecture of Gaussian PPMx. And the same covariates were chosen.

Linear Model

A baseline simple model for comparison.

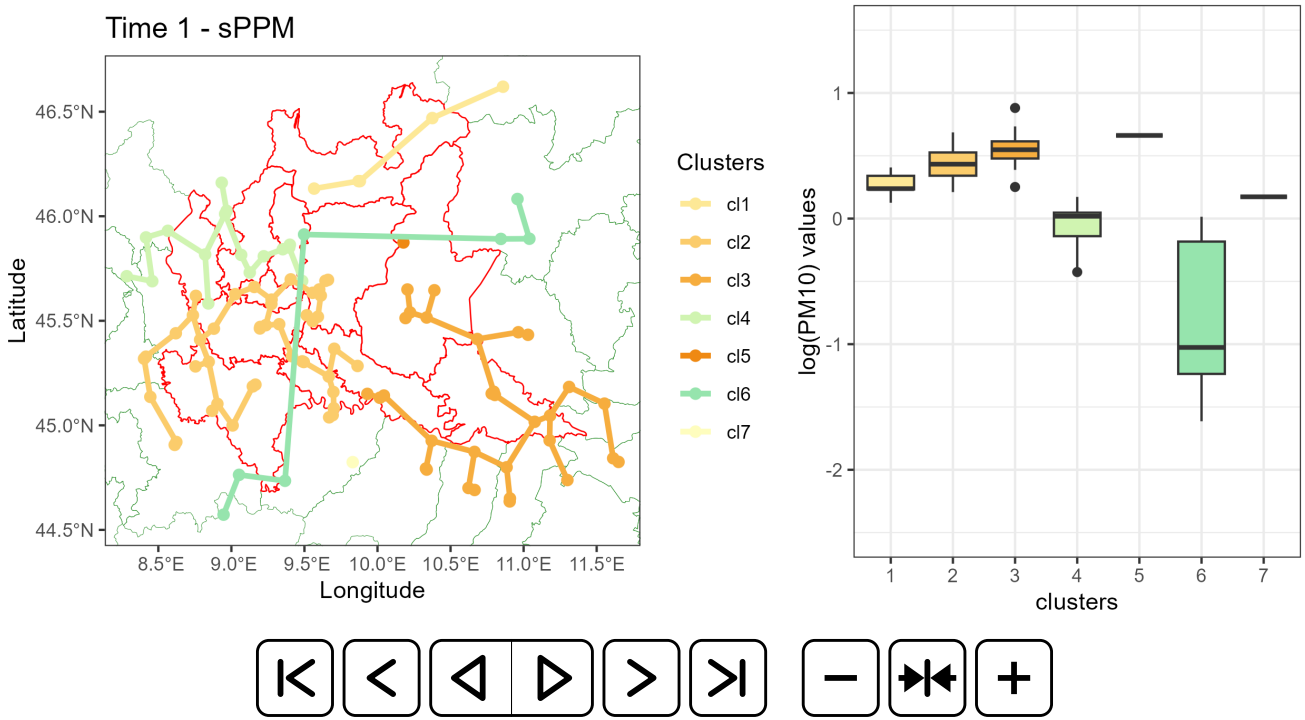
$$Y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad i = 1, \ldots, n$$
$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{b}_0, \sigma^2 B_0)$$

Can't capture spatial variability, so one model for each station to try and achieve precision on a smaller scale.
Covariates and functions of time (sine, cosine, square).

Useful for variable selection (through Kuo-Mallick, horseshoe and lasso) and interpretation of the results.

Analysis of the results - weekly view

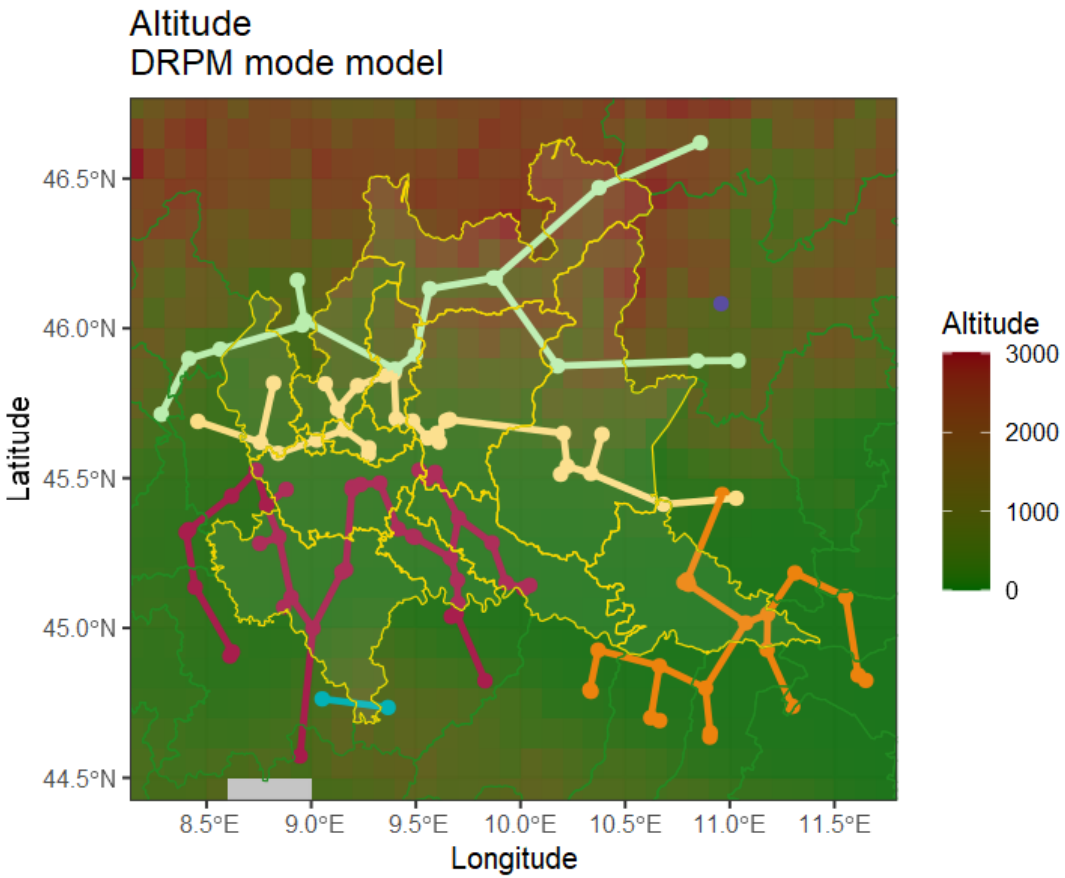
The task was weekly clustering, which for example generated this result:



But to have a better overview of the various aspects influencing PM_{10} levels we now dive more into the global trend and interpretation.

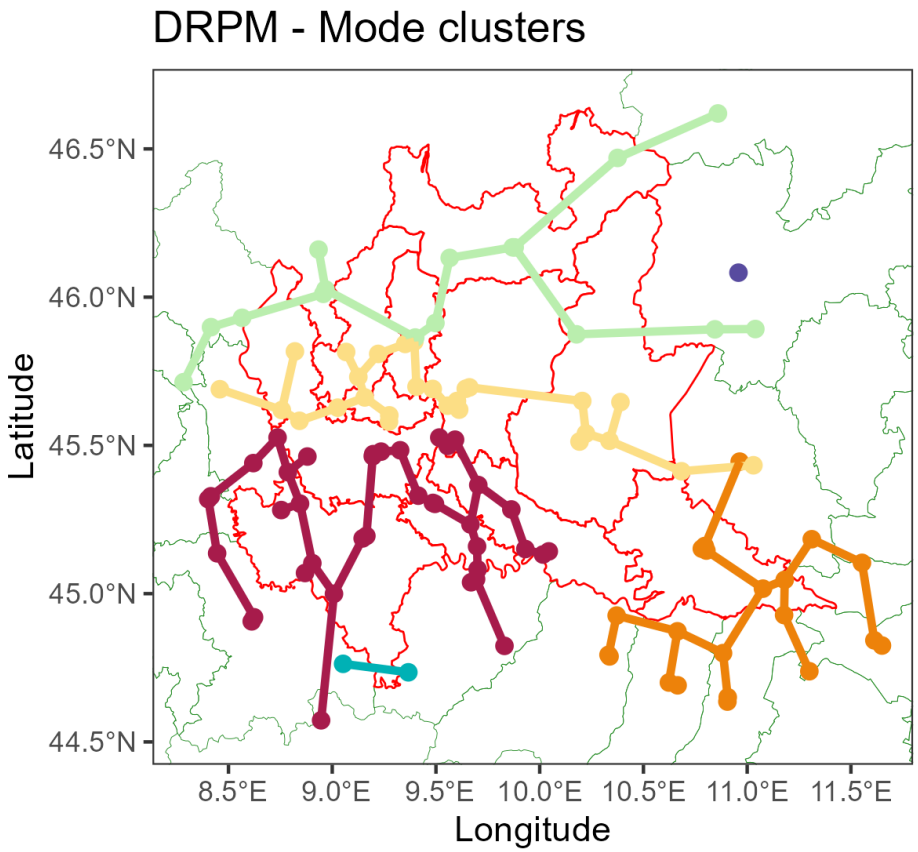
Stratification effect of Altitude

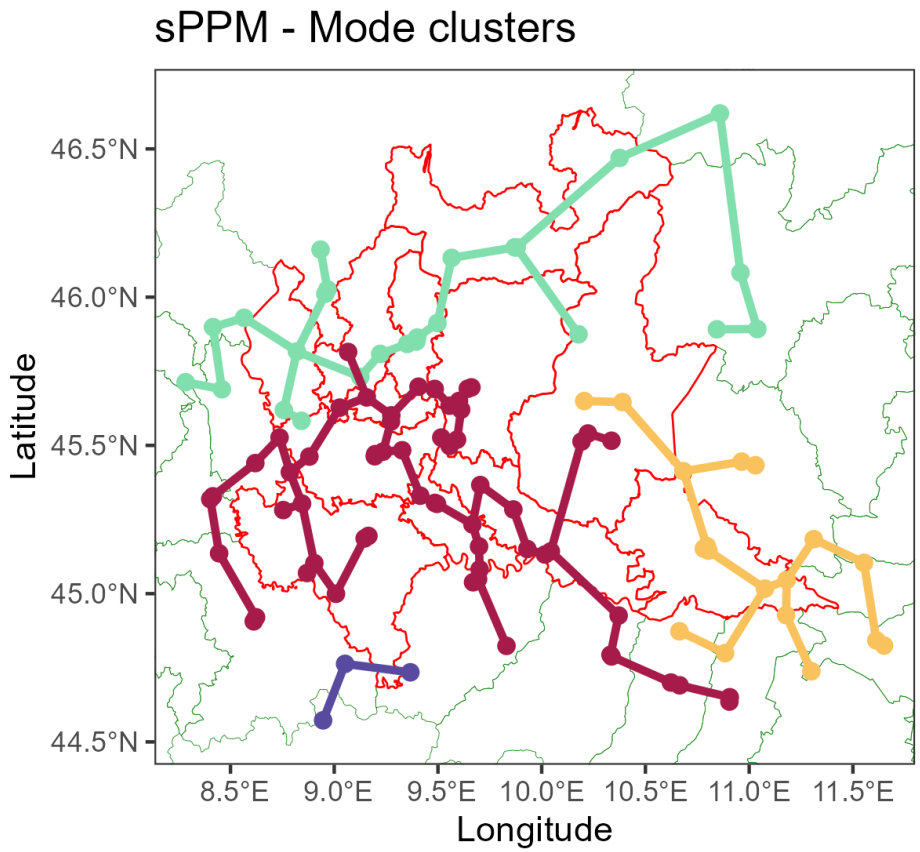
Almost all models exhibited a stratification pattern.
Regions with flat terrain generally displayed higher PM_{10} concentrations and were frequently clustered together.



Mode clusters 1/2

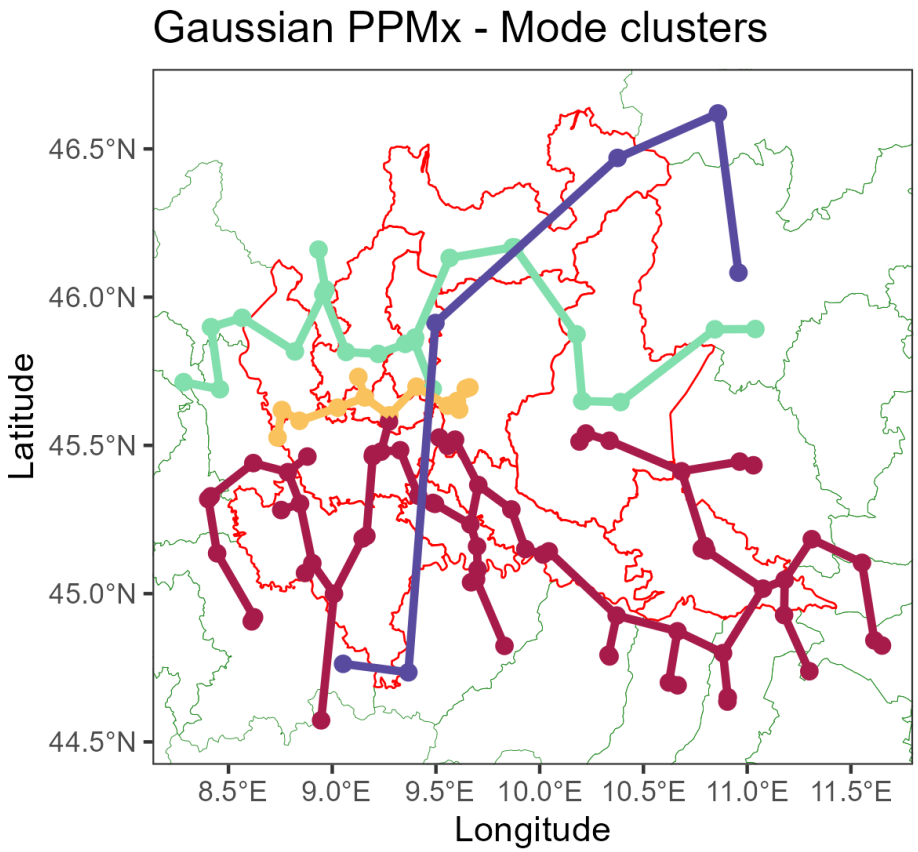
Almost all models exhibited a stratification pattern.
Regions with flat terrain generally displayed higher PM_{10} concentrations and were frequently clustered together.

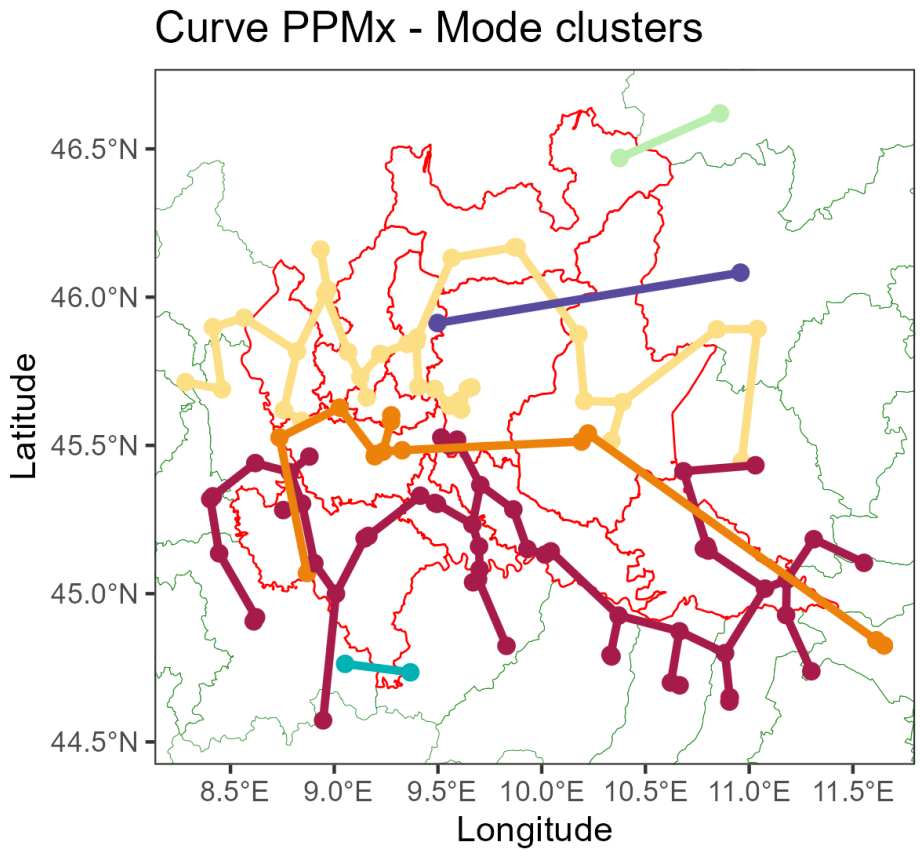




Mode clusters 2/2

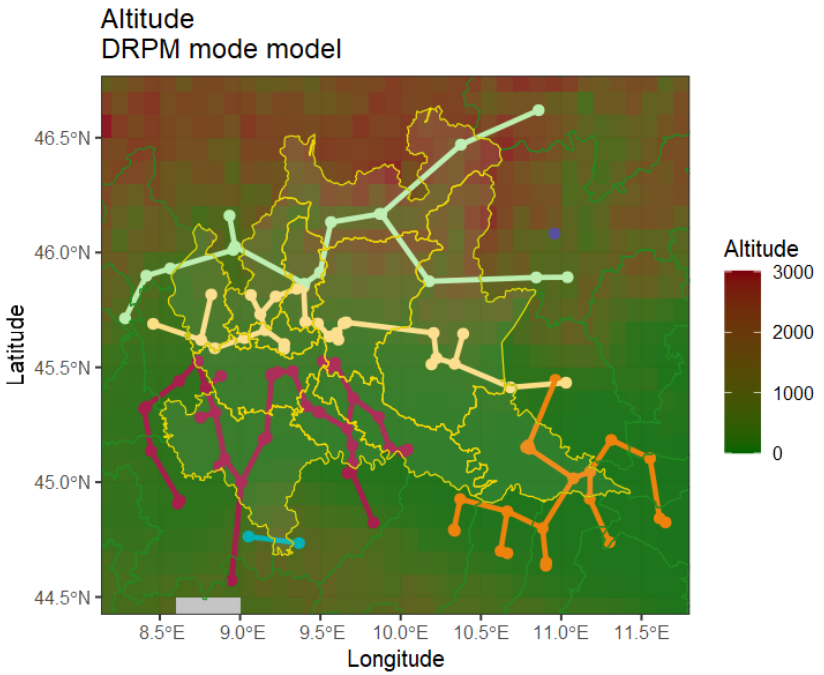
As elevation increased towards the Alps, fewer polluted clusters were observed. In the southwest area, some noticeable station emerged consistently across all models.



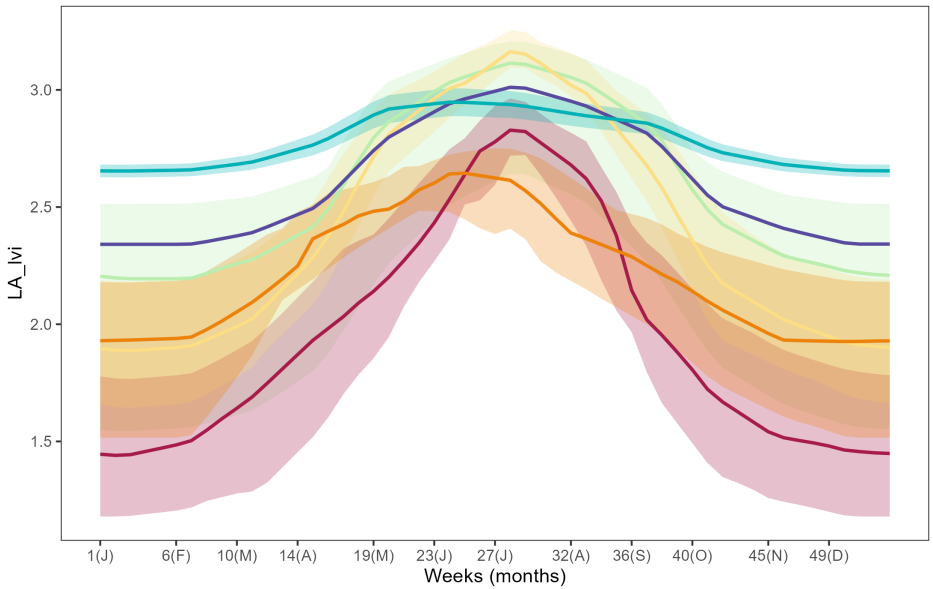


Altitude and Vegetation

In general, vegetation and altitude are correlated with lower PM_{10} levels.
Interpretation Altitude could help because it implies distance from industrial areas, while plants could use their special micro-morphological structure to retain particulate matter.



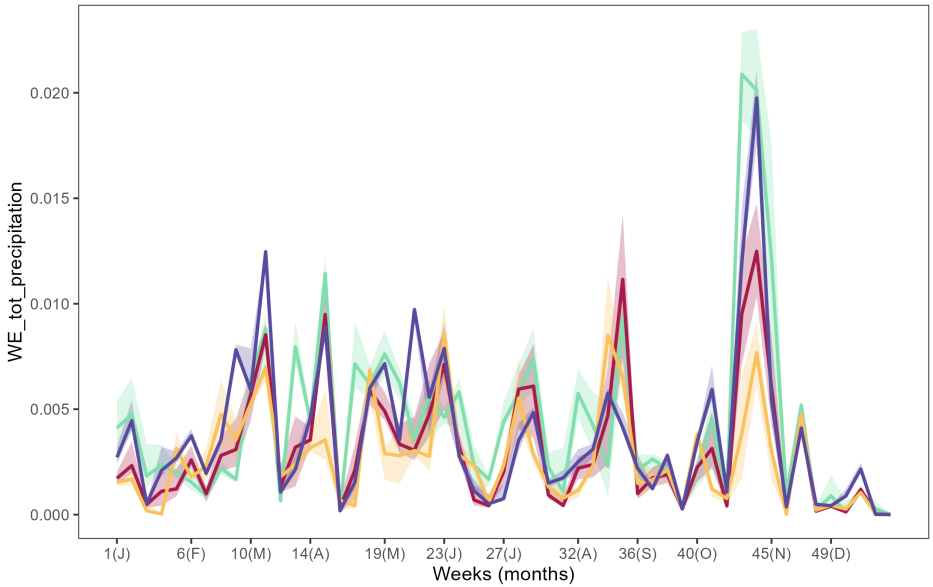
Altitude



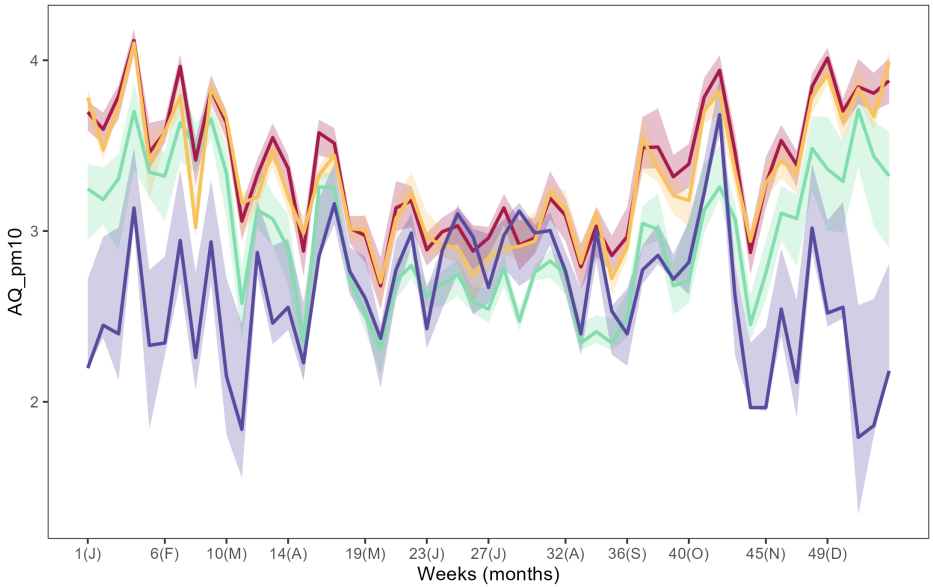
Vegetation

Total precipitations

Also rainfalls appears to influence the reduction of PM_{10}.
Interpretation Rainwater can take away air pollutants, to some extent, by delivering and depositing the contaminants to the ground (a process known as precipitation scavenging or washout).



Total precipitations

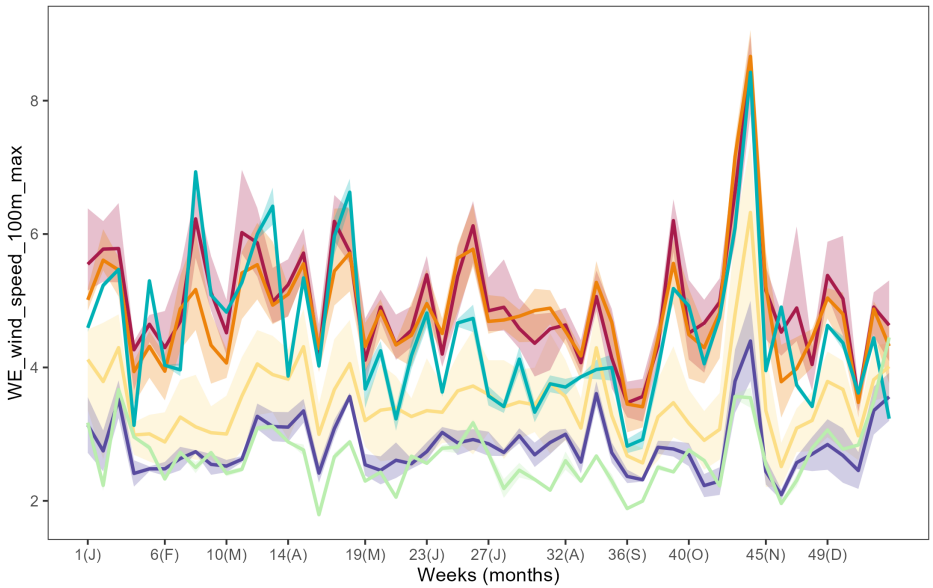


PM_{10} concentrations

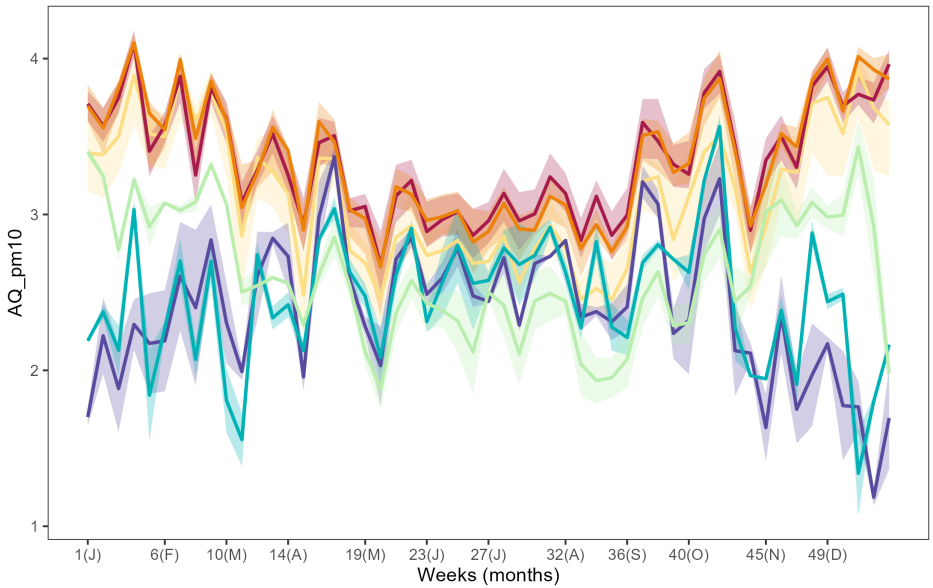
Max wind intensity

High intensity of the wind increases PM_{10} levels; but this effect is mitigated by the presence of rain.

Interpretation Wind generates and moves dust from the ground, where pollutants may have been deposited, and bring them in the air.



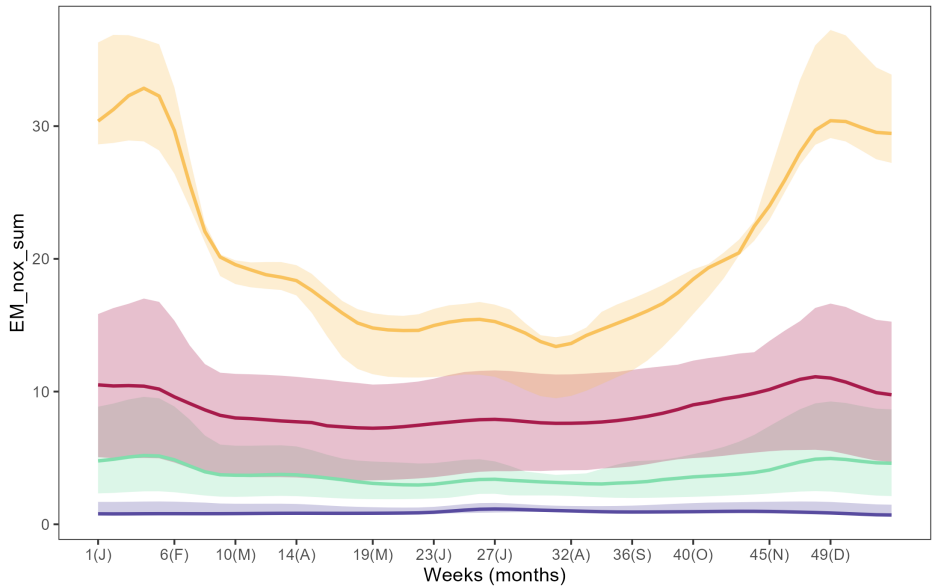
Max wind speed at 100m



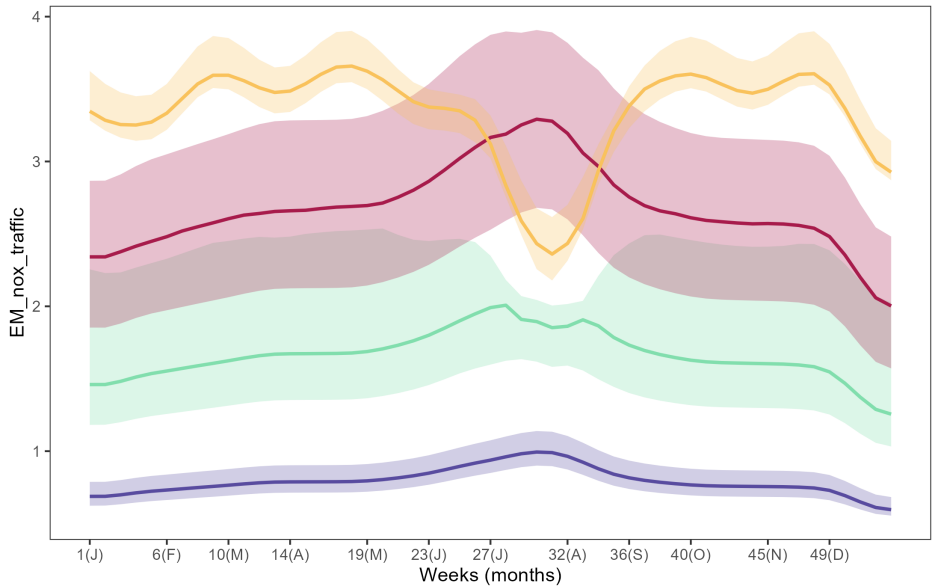
PM_{10} concentrations

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.



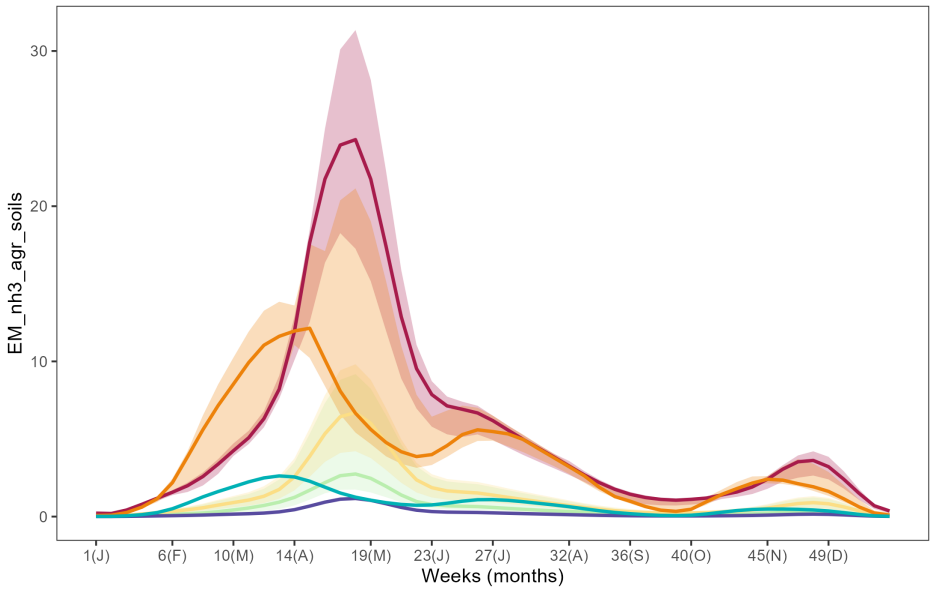
NO_x across all sectors



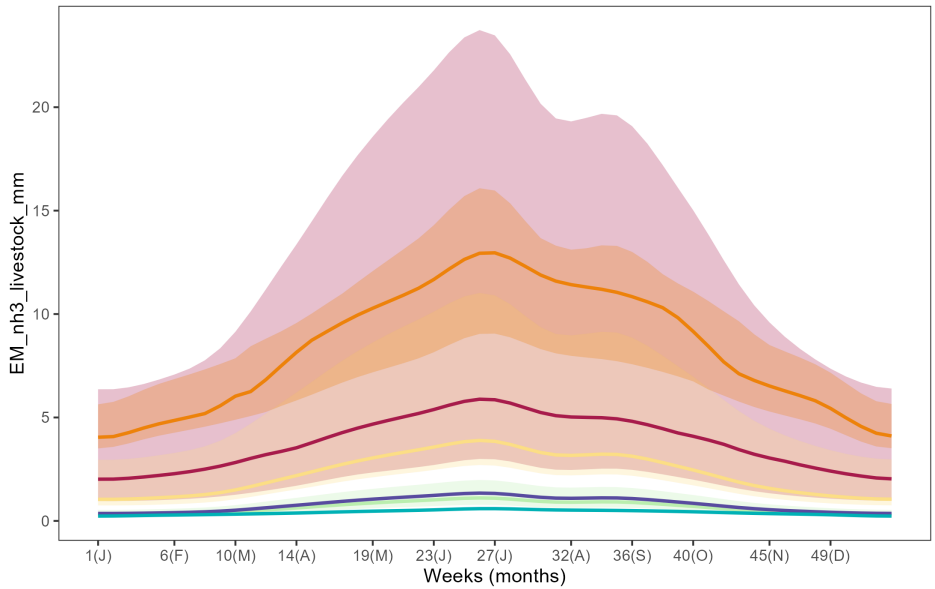
NO_x from traffic

Emissions of NH_3

The presence of NH_3 from agricultural and breeding activities also seems to correlate to the elevation of PM_{10} levels.
Naturally, those areas corresponds to a high number of animals or crops.



NH_3 from agricultural activities



NH_3 from breeding activities

ARI metric

A more numerical way to compare the clustering results is through the Adjusted Random Index (ARI): a sort of correlation index which measures the similarity between clusterings.
Given two partitions ρ_1 and ρ_2, the ARI(ρ_1, ρ_2) describes the level of agreement that they show in clustering the data.
It is defined as a correction of the Random Index (RI), which is RI$(\rho_1, \rho_2) = (a + b)/\binom{n}{2}$ where

- $a =$ #pairs (of units) clustered together cluster according to ρ_1 and ρ_2
- $b =$ #pairs that do not belong to the same cluster according to ρ_1 and ρ_2

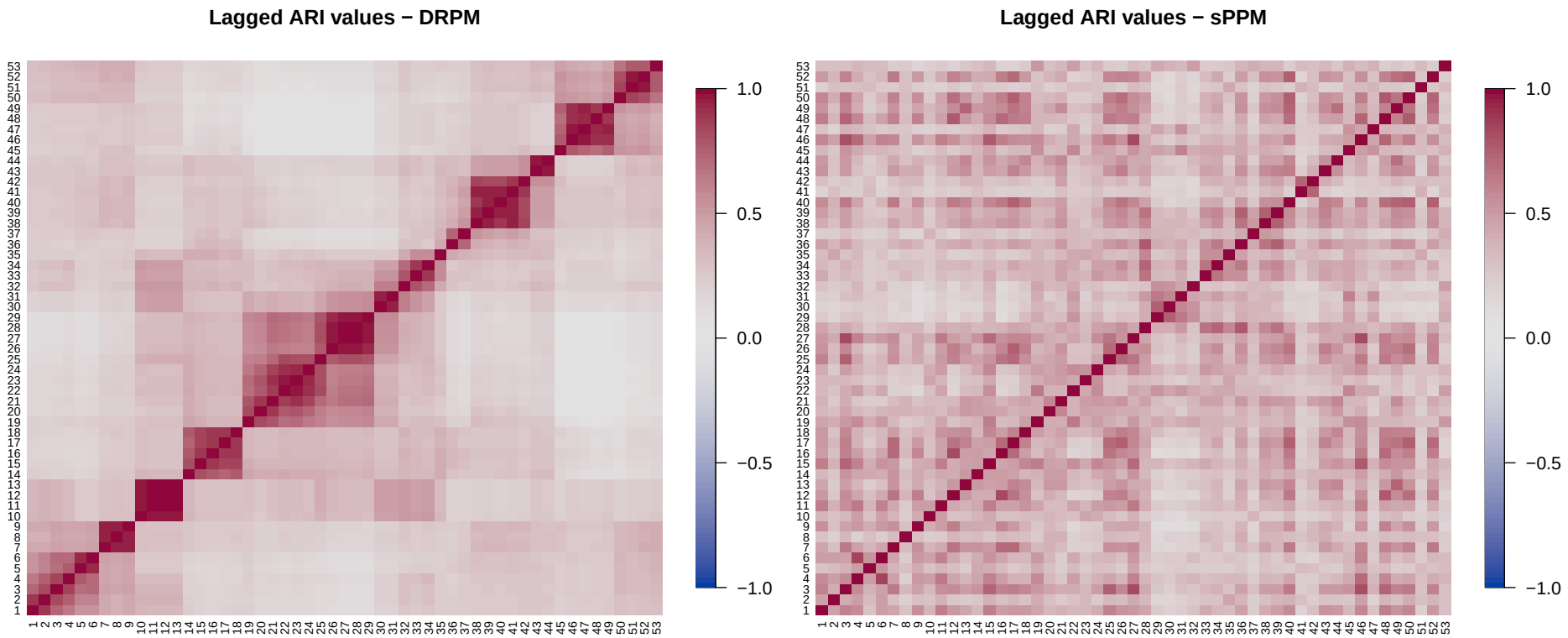
How to use it?

This metric allows for example to compare a proposal clustering with the real exact 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.

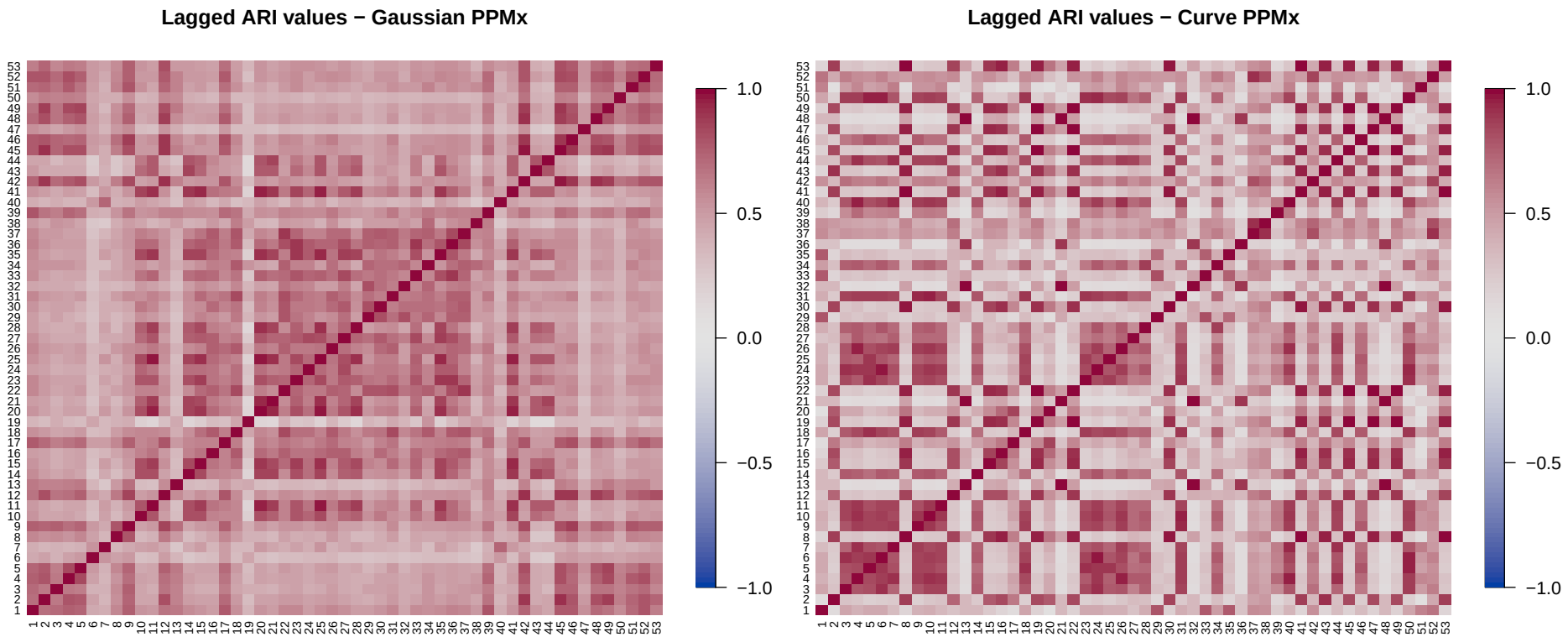
Time evolution of the clusters 1/2

For each model we computed ARI(ρ_t, ρ_{t+k}), for $t \in \{1, \ldots, 53\}$ and for all valid values of the lag (or time-shift) k.



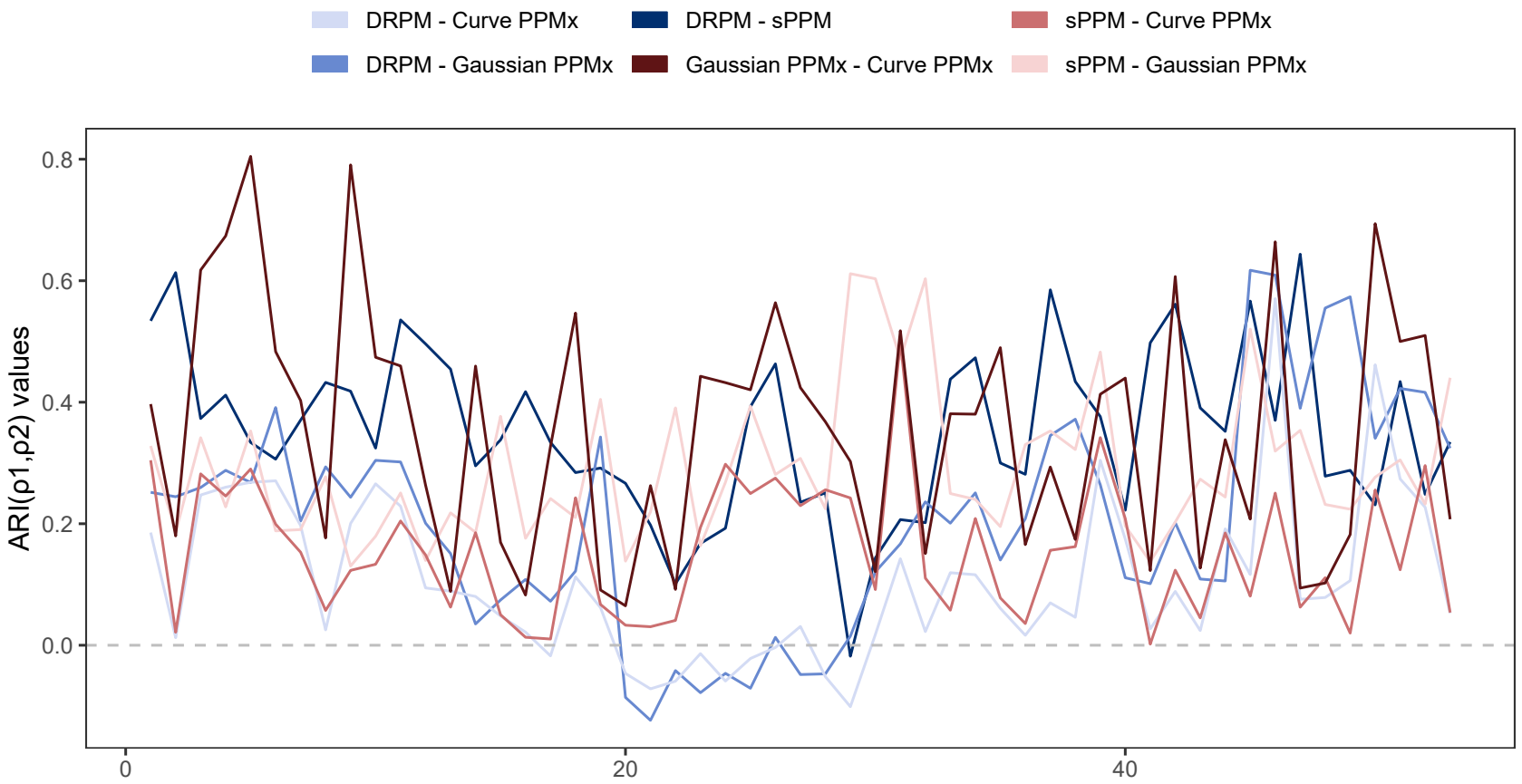
Time evolution of the clusters 2/2

For each model we computed ARI(ρ_t, ρ_{t+k}), for $t \in \{1, \ldots, 53\}$ and for all valid values of the lag (or time-shift) k.

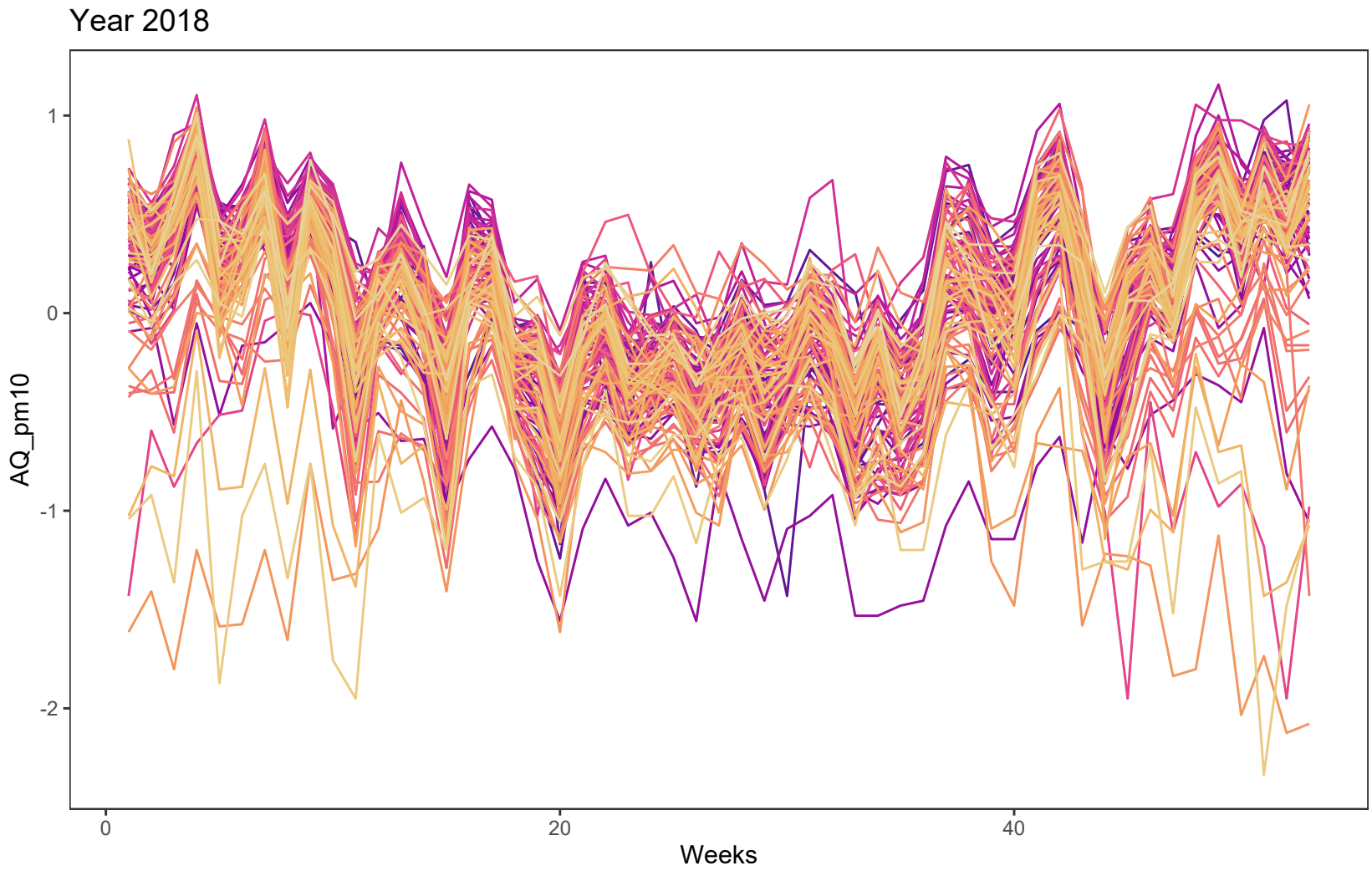


Agreement level of the models

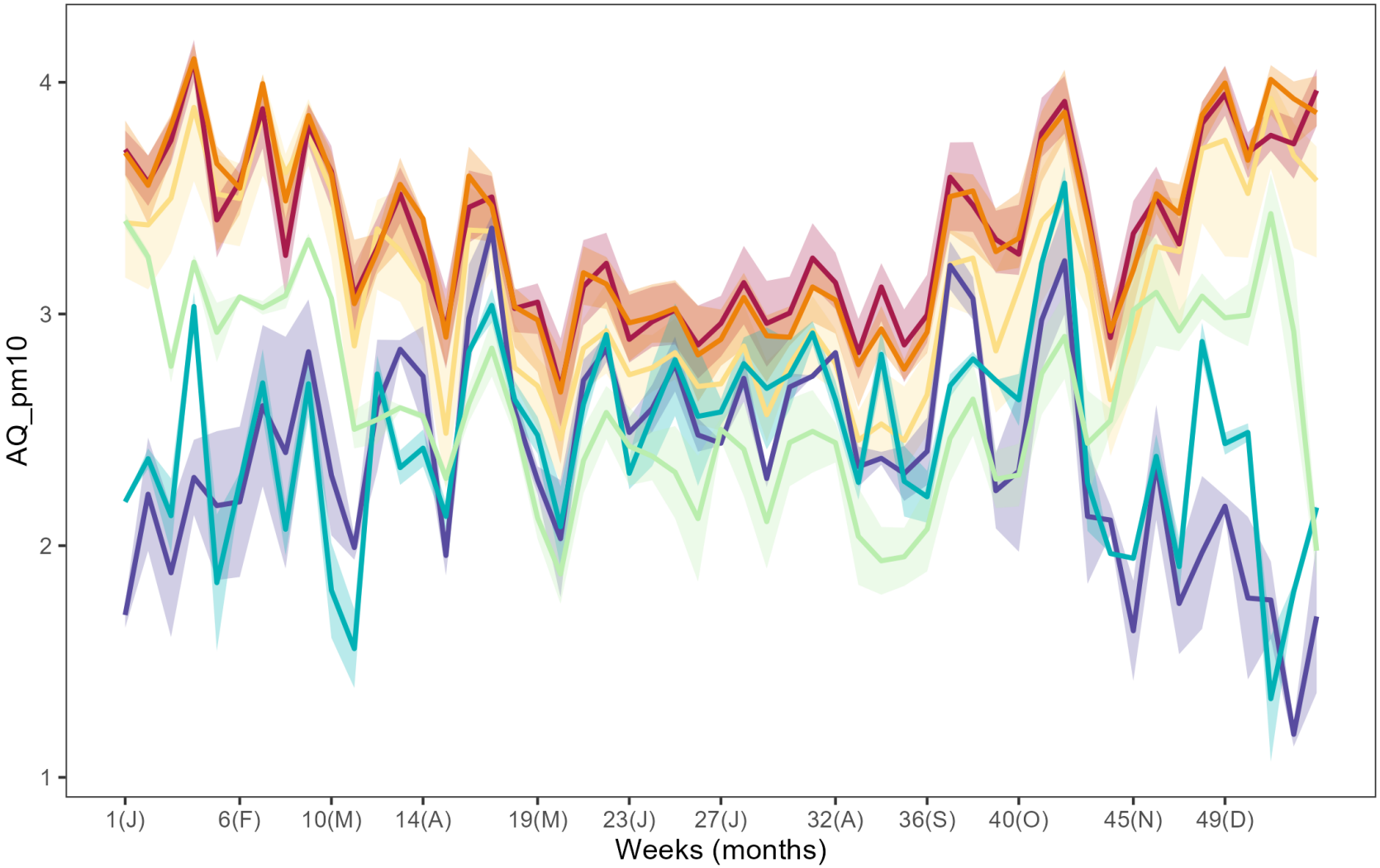
We computed $\text{ARI}(\rho_t^{M_1}, \rho_t^{M_2})$ for $t \in \{1, \ldots, 53\}$ and for each pair of models M_1 and M_2 in the four that we fitted.



Remember this?

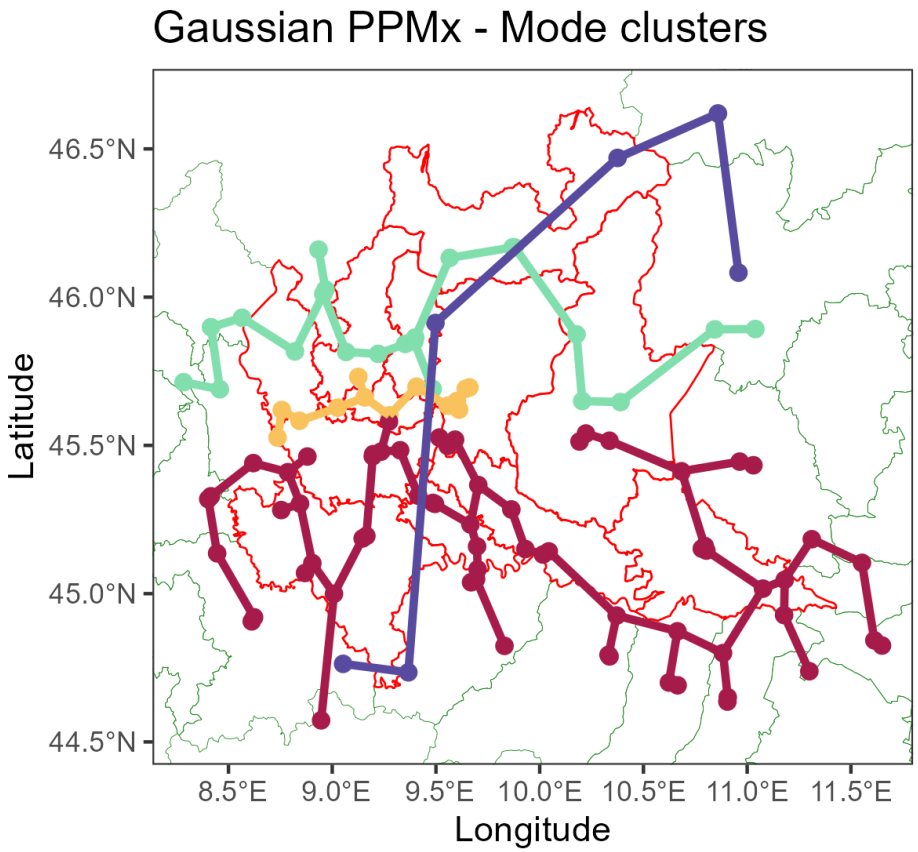


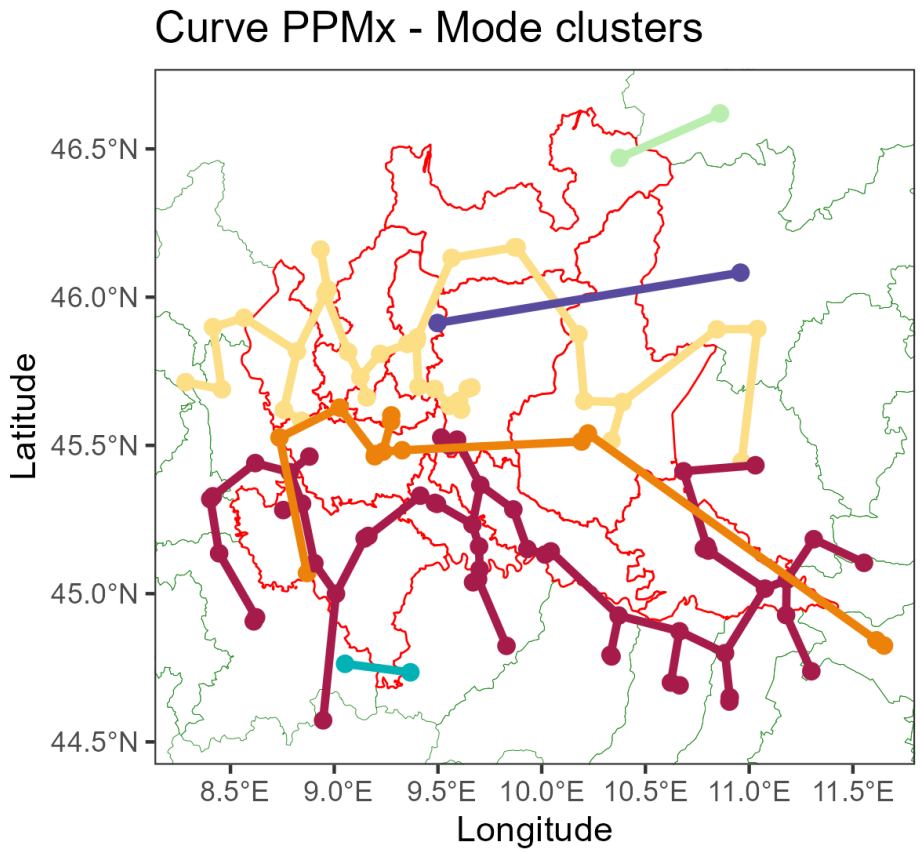
Now after clustering



Opposite trend clusters

Some stations had an inverted trend: low levels in winter and high in summer. The causes could be high elevation and the presence of snow, which captures the pollutants particles





Conclusion

Our findings underline the influential role of both

- morphological characteristics
- anthropological factors

Both are to be considered for a comprehensive analysis.
Despite variations, a general trend was successfully spotted.

Visualization methods

To address the interpretation of the results, we developed a sort of library of auxiliary functions which enabled 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.

All those results are available *to play with* at the page `https://federicomor.github.io/assets/figures/visualize.html`.

References I

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

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.

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

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

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.

References II

Yevgen Nazarenko, Sébastien Fournier, Uday Kurien, Rodrigo Benjamin Rangel-Alvarado, Oleg Nepotchatykh, Patrice Seers, and Parisa A. Ariya.
Role of snow in the fate of gaseous and particulate exhaust pollutants from gasoline-powered vehicles.
Environmental Pollution, 223:665–675, 2017.

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/.

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.

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

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

References III

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.

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.