

First Meeting Presentation

Project A2

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

Politecnico of Milano
Bayesian Statistics course

Project revision of
November 23, 2023

Presentation Flow

- ① Project Overview
 - Goal and Definition
 - Data Exploration
- ② Models
 - General model construction
 - Models from literature
- ③ Expected workflow
- ④ References

Goals

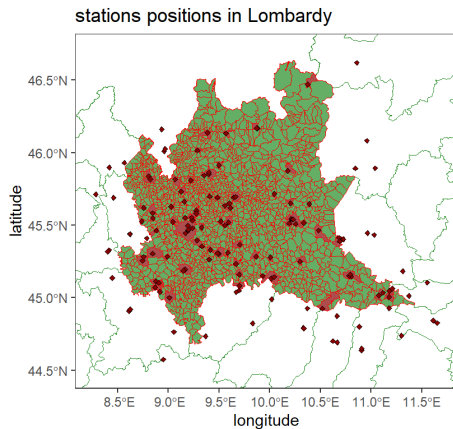
Clustering weekly data of one year of PM10 (plus covariates) using different Bayesian models.

Understand and apply model from a paper by **Garritt et al.** and **ppmSuite** package in R and compare the clusterings of different models.

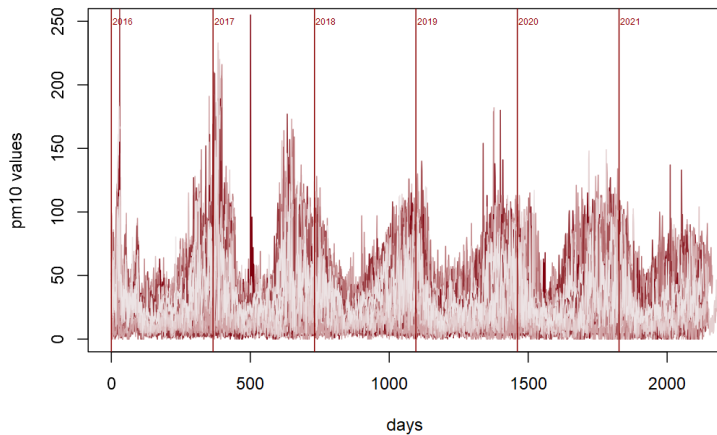
About the dataset

- **Dataset:** AGRIMONIA project, at <https://zenodo.org/records/7563265>
- It was 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)
- In total 38 covariates
- 141 stations
- Timeframe: 01/01/2016 to 31/12/2021

Spatial Exploration

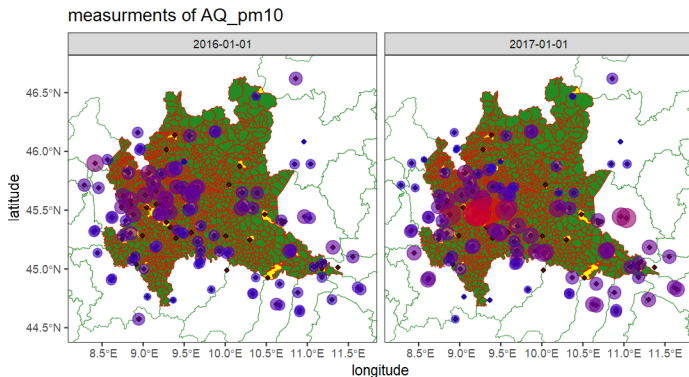


Temporal Exploration

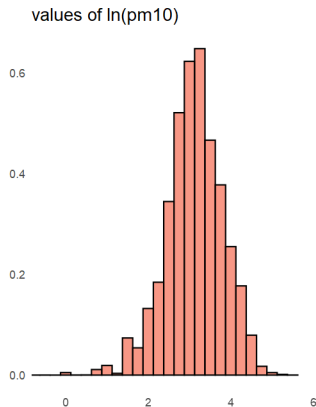
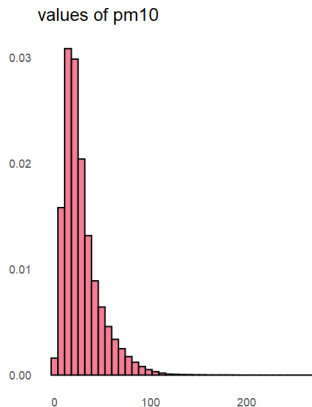


PM10 concentration

We have created a library to analyse the variables in our dataset.

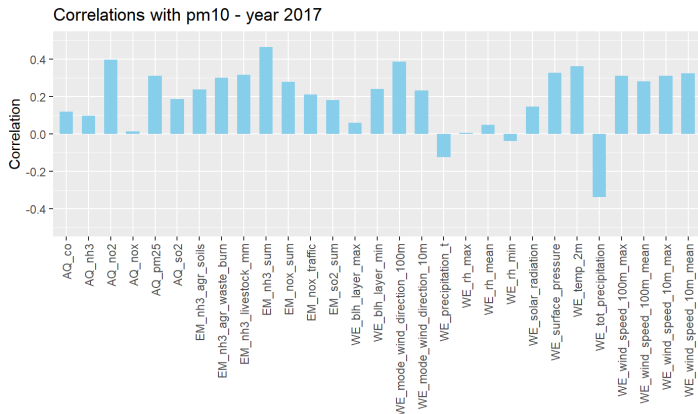


PM10 distribution



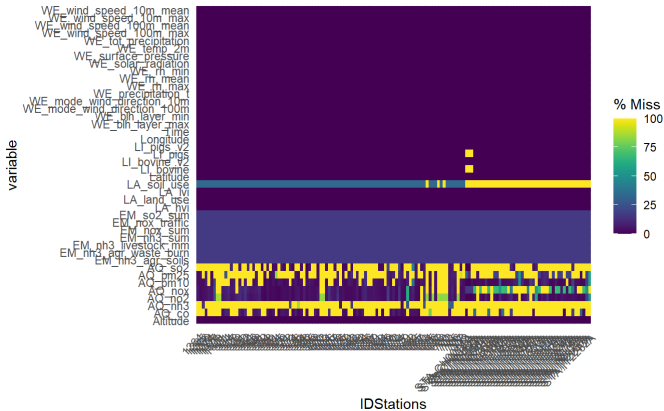
Correlation: PM10 and others

We compute the spearman's correlation index between PM10 and the other covariates, which in the functional framework quantifies with a value in $[-1, 1]$ the tendency of 2 random variables X_t and Y_t to be perfect monotone functions one of each other.



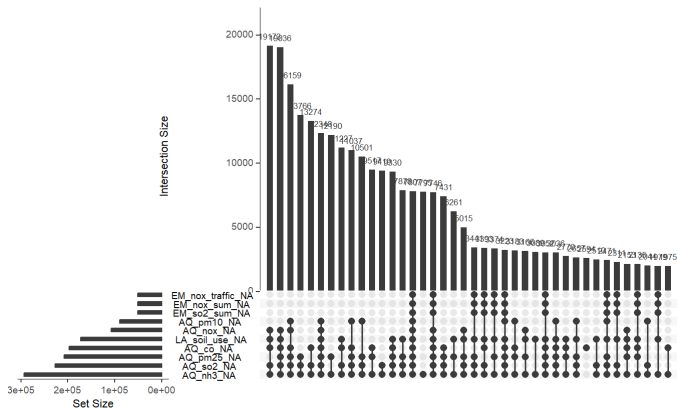
Missing value exploration

General patter on missing values.



Missing value exploration

Combinations of variables with the most missing values.



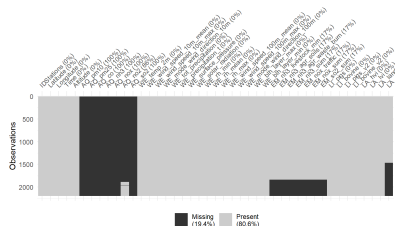
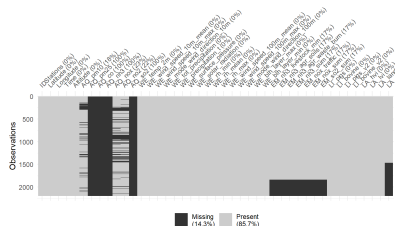
Missing value exploration

Plot of missing data divided by station to see particular patterns to help select variables.

Possible to remove some stations that are not measuring PM10 values.

Some columns present missing values in most stations, so we can remove the corresponding covariates.

Some missing observations are concentrated in specific periods such as during last years.



Missing value exploration

```
[1] "1274"      "1800"      "1801"      "501"       "504"
[6] "514"       "529"       "539"       "544"       "545"
[11] "551"       "552"       "573"       "588"       "602"
[16] "603"       "606"       "607"       "626"       "652"
[21] "656"       "657"       "665"       "672"       "682"
[26] "694"       "696"       "698"       "699"       "700"
[31] "702"       "707"       "STA.IT1518A" "STA.IT1751A" "STA.IT1924A"
[36] "STA.IT2282A"
```

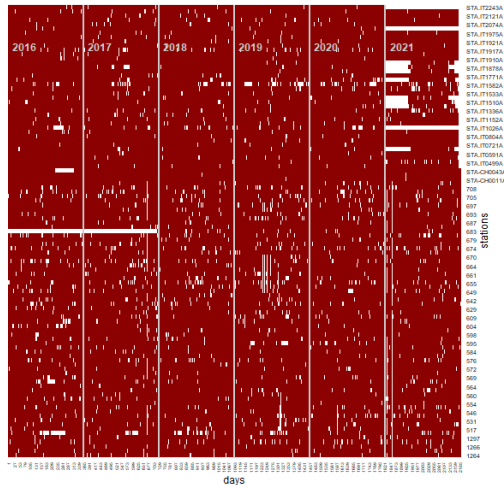
First selection to remove non informative stations, then count of missing values for column, removing those above a chosen threshold.

```
IDStations      Latitude
0               0
Longitude        Time
0               0
Altitude         AQ_pm10
0               12719
AQ_pm25         AQ_co
136792          136560
AQ_rh3          AQ_nox
217540          72088
AQ_no2          AQ_so2
21266          156205
WE_temp_2m      WE_wind_speed_10m_mean
0               0
WE_wind_speed_10m_max WE_node_wind_direction_10m
0               0
WE_tot_precipitation WE_precipitation_t
0               0
WE_surface_pressure WE_solar_radiation
0               0
WE_rh_mean      WE_rh_max
0               0
WE_rh_min       WE_wind_speed_100m_mean
0               0
WE_wind_speed_100m_max WE_node_wind_direction_100m
0               0
WE_b1h_layer_max WE_b1h_layer_min
0               0
EH_rh3_livestock_nm EH_rh3_agr_soils
38325              38325
EH_rh3_agr_waste_burn EH_rh3_sum
38325              38325
EH_nox_traffic       EH_nox_sum
38325              38325
EH_so2_sum           LT_pigs
38325              6576
LT_bovine            LT_pigs_v2
6576                0
LT_bovine_v2        LA_hvi
0                   0
LA_lv1              LA_land_use
0                   0
LA_soil_use         236656
```

Missing value exploration

Missing values over time by stations.

NA time series (white are NAs) all data



Models: a complex task

Considering the nature of the data, our models should account for different levels of information:

- spatial context
- temporal context
- covariates

which is a not-so-trivial task.

Now we see the general incremental idea to build such models.

Purely spatial model

- We have n distinct locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, where $\mathbf{s}_i = (\text{lat}_i, \text{long}_i)$.
- There we record data y_i and (possibly) covariates \mathbf{x}_i , for $i = 1, \dots, n$. The goal is to define a model for partitioning them into k groups.
- So we define $\rho = \{S_1, \dots, S_k\}$ the cluster set variable (with $S_h \subseteq \{1, \dots, n\}$ for $h = 1, \dots, k$).
An equivalent formulation is possible through some cluster indicator variables c_1, \dots, c_n ; where $c_i = h \iff i \in S_h$ for $i = 1, \dots, n$.
- In general, the law for ρ follows a spatial Product Partition Model (sPPM):

$$\mathbb{P}(\rho = \{S_1, \dots, S_k\}) \propto \prod_{h=1}^k C(S_h, \mathbf{s}_h^*)$$

where $\mathbf{s}_h^* = \{\mathbf{s}_i : i \in S_h\}$ and $C(S_h, \mathbf{s}_h^*)$ is a cohesion function; for $h = 1, \dots, k$.

Purely spatial model

- We have n distinct locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, where $\mathbf{s}_i = (\text{lat}_i, \text{long}_i)$.
- There we record data y_i and (possibly) covariates \mathbf{x}_i , for $i = 1, \dots, n$. The goal is to define a model for partitioning them into k groups.
- So we define $\rho = \{S_1, \dots, S_k\}$ the cluster set variable (with $S_h \subseteq \{1, \dots, n\}$ for $h = 1, \dots, k$).
An equivalent formulation is possible through some cluster indicator variables c_1, \dots, c_n ; where $c_i = h \iff i \in S_h$ for $i = 1, \dots, n$.
- In general, the law for ρ follows a spatial Product Partition Model (sPPM):

$$\mathbb{P}(\rho = \{S_1, \dots, S_k\}) \propto \prod_{h=1}^k C(S_h, \mathbf{s}_h^*)$$

where $\mathbf{s}_h^* = \{\mathbf{s}_i : i \in S_h\}$ and $C(S_h, \mathbf{s}_h^*)$ is a cohesion function; for $h = 1, \dots, k$.

Purely spatial model

- We have n distinct locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, where $\mathbf{s}_i = (\text{lat}_i, \text{long}_i)$.
- There we record data y_i and (possibly) covariates \mathbf{x}_i , for $i = 1, \dots, n$. The goal is to define a model for partitioning them into k groups.
- So we define $\rho = \{S_1, \dots, S_k\}$ the cluster set variable (with $S_h \subseteq \{1, \dots, n\}$ for $h = 1, \dots, k$).

An equivalent formulation is possible through some cluster indicator variables c_1, \dots, c_n ; where $c_i = h \iff i \in S_h$ for $i = 1, \dots, n$.

- In general, the law for ρ follows a spatial Product Partition Model (sPPM):

$$\mathbb{P}(\rho = \{S_1, \dots, S_k\}) \propto \prod_{h=1}^k C(S_h, \mathbf{s}_h^*)$$

where $\mathbf{s}_h^* = \{\mathbf{s}_i : i \in S_h\}$ and $C(S_h, \mathbf{s}_h^*)$ is a cohesion function; for $h = 1, \dots, k$.

Purely spatial model

- We have n distinct locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, where $\mathbf{s}_i = (\text{lat}_i, \text{long}_i)$.
- There we record data y_i and (possibly) covariates \mathbf{x}_i , for $i = 1, \dots, n$. The goal is to define a model for partitioning them into k groups.
- So we define $\rho = \{S_1, \dots, S_k\}$ the cluster set variable (with $S_h \subseteq \{1, \dots, n\}$ for $h = 1, \dots, k$).

An equivalent formulation is possible through some cluster indicator variables c_1, \dots, c_n ; where $c_i = h \iff i \in S_h$ for $i = 1, \dots, n$.

- In general, the law for ρ follows a spatial Product Partition Model (sPPM):

$$\mathbb{P}(\rho = \{S_1, \dots, S_k\}) \propto \prod_{h=1}^k C(S_h, \mathbf{s}_h^*)$$

where $\mathbf{s}_h^* = \{\mathbf{s}_i : i \in S_h\}$ and $C(S_h, \mathbf{s}_h^*)$ is a cohesion function; for $h = 1, \dots, k$.

Spatial and temporal model

We have n distinct locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, where $\mathbf{s}_i = (\text{lat}_i, \text{long}_i)$.

There we record data y_i and (possibly) covariates \mathbf{x}_i , for $i = 1, \dots, n$.

The goal is to define a model for partitioning them into k_t groups, with t spanning over $1, \dots, T$.

So we define $\rho_t = \{S_{1,t}, \dots, S_{k_t,t}\}$ the cluster set variable (with $S_{h,t} \subseteq \{1, \dots, n\}$ for $h = 1, \dots, k_t$).

In general, the law for ρ_t follows a spatial Product Partition Model (sPPM) updated to account for the time relation (stPPM); meaning that we need a formulation of a joint probability model for ρ_1, \dots, ρ_T .

This update can be explicated for example by

- supposing a Markov Chain structure, letting ρ_t depend just on ρ_{t-1} ;
- introducing some cluster reallocation variable $\gamma_{i,t} \in \{0, 1\}$.

Model 1



Garritt L. Page, Fernando A. Quintana, David B. Dahl (2022)

Dependent Modeling of Temporal Sequences of Random Partitions. [Journal of Computational and Graphical Statistics](#), 31:2, 614-627.

$$\begin{aligned}
 Y_{it} | \boldsymbol{\mu}_t^*, \boldsymbol{\sigma}_t^{2*}, \mathbf{c}_t &\stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_{c_{it}t}^*, \sigma_{c_{it}t}^{2*}) \quad i = 1, \dots, n \quad \text{and} \quad t = 1, \dots, T \\
 (\mu_{jt}, \sigma_{jt}) | \theta_t, \tau_t^2 &\stackrel{\text{ind}}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \quad j = 1, \dots, k_t \\
 (\theta_t, \tau_t) &\stackrel{\text{iid}}{\sim} \mathcal{N}(\phi_0, \lambda^2) \times \mathcal{U}(0, A_\tau) \quad t = 1, \dots, T \\
 (\phi_0, \lambda) &\sim \mathcal{N}(m_0, s_0^2) \times \mathcal{U}(0, A_\lambda) \\
 \{\mathbf{c}_t, \dots, \mathbf{c}_T\} &\sim \text{tRPM}(\boldsymbol{\alpha}, M) \quad \text{with} \quad \alpha_t \stackrel{\text{iid}}{\sim} \text{Beta}(a_\alpha, b_\alpha)
 \end{aligned}$$

Adding covariates

Given a partition, now we can easily design models which also account for covariates. For example we can update the previous model into

$$\begin{aligned}
 Y_{it} | \beta_t^*, \sigma_t^{2*}, \mathbf{c}_t &\stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_{it}^T \beta_{c_{it}t}^*, \sigma_{c_{it}t}^{2*}) \\
 (\beta_{jt}, \sigma_{jt}) | \theta_t, \tau_t^2 &\stackrel{\text{ind}}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \\
 &\vdots
 \end{aligned}$$

or we can further characterize the time dependance with some AR(.) model

$$\begin{aligned}
 Y_{it} | Y_{it-1}, \beta_t^*, \sigma_t^{2*}, \mathbf{c}_t &\stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_{it}^T \beta_{c_{it}t}^* + Y_{it-1} \eta_i, \sigma_{c_{it}t}^{2*} (1 - \eta_i)^2) \\
 Y_{i1} | \beta_1^*, \sigma_1^{2*}, \mathbf{c}_1 &\sim \mathcal{N}(\mathbf{x}_{i1}^T \beta_{c_{i1}1}^{2*}, \sigma_{c_{i1}1}^{2*}) \\
 &\vdots
 \end{aligned}$$

Model 2



Mozdzen A., Cremaschi A., Cadonna A., Guglielmi A., Kastner G. (2022)

Bayesian modeling and clustering for spatio-temporal areal data: An application to Italian unemployment. *Spatial Statistics* 52, 100715.

$$Y_{it} | \mathbf{x}_{it}, \beta_{s_i}^*, w_{it}, \sigma^2, s_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_{it}^T \beta_{s_i}^* + w_{it}, \sigma^2)$$

$$\mathbf{w}_t | \mathbf{w}_{t-1}, \xi_s^*, \mathbf{s}, \tau^2, \rho, W \sim \mathcal{N}_I(\text{diag}(\xi_s^*) \mathbf{w}_{t-1}, \tau^2 Q(\rho, W)^{-1})$$

$$\mathbf{w}_1 | \tau^2, \rho, W \sim \mathcal{N}_I(\mathbf{0}, \tau^2 Q(\rho, W)^{-1})$$

$$\sigma^2 \sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$\tau^2 \sim \text{Inv-Gamma}(a_{\tau^2}, b_{\tau^2})$$

$$\rho \sim \text{Beta}(\alpha_\rho, \beta_\rho)$$

$$\mathbf{s} | \alpha \sim \text{PólyaUrn}(\mathbf{s} | \alpha)$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$$

$$\phi_1^*, \dots, \phi_{K_I}^* | \mu_\beta, \Sigma_\beta, \alpha_\xi, \beta_\xi \stackrel{\text{iid}}{\sim} P_0, \quad \phi_j^* = (\beta_j^*, \xi_j^*) \quad j = 1, \dots, K_I$$

$$P_0(d\phi^*) = \mathcal{N}_{p+1}(d\beta^* | \mu_0, \Sigma_0) \text{Beta}_{(-1,1)}(d\xi^* | \alpha_\xi, \beta_\xi)$$

Model 3

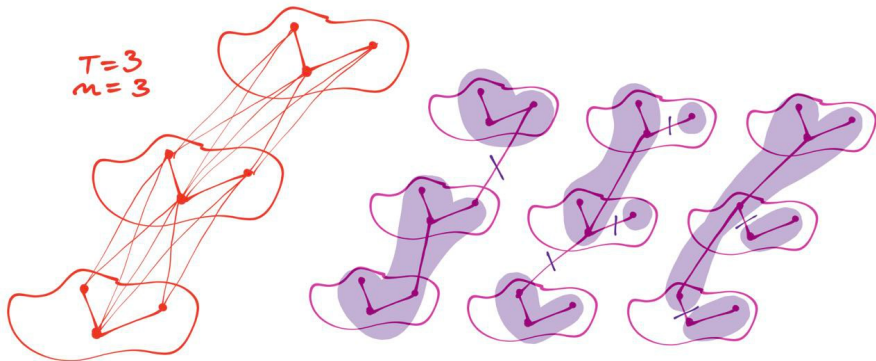


Leonardo V. Teixeira, Renato M. Assunção, Rosangela H. Loschi (2019)
Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees.
Journal of Machine Learning Research 20, 85, 1–35.

This model works on a graph structure, which incorporates together space and time. That is, from data Y_{jt} for $j = 1, \dots, n$ and $t = 1, \dots, T$, we now move to Y_i for $i \in I = \{1, \dots, nT\}$ by stacking T times the spatial map. So we have a graph $\mathcal{G} = (V, E)$ of nT nodes and edges built according to time and space connections.

The idea is to search a partition $\pi = \{\mathcal{G}_1, \dots, \mathcal{G}_c\}$ of I (with \mathcal{G}_k subgraphs for \mathcal{G}), on randomly selected spanning trees \mathcal{T} of \mathcal{G} , on which we set cluster-specific parameters $\beta_{\mathcal{G}_1}, \dots, \beta_{\mathcal{G}_c}$.

Model 3



Model 3



Leonardo V. Teixeira, Renato M. Assunção, Rosangela H. Loschi (2019)
 Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees.
[Journal of Machine Learning Research](#) 20, 85, 1–35.

$$Y_i | \mathcal{T}, \boldsymbol{\pi}, \beta_{\mathcal{G}_k} \stackrel{\text{iid}}{\sim} f(Y_i | \beta_{\mathcal{G}_k}; \mathbf{x}_i) \quad i \in \mathcal{G}_k$$

$$\beta_{\mathcal{G}_1}, \dots, \beta_{\mathcal{G}_c} | \mathcal{T}, \boldsymbol{\pi} \sim \prod_{k=1}^c f(\beta_{\mathcal{G}_k})$$

$$\mathbb{P}(\boldsymbol{\pi} = \{\mathcal{G}_1, \dots, \mathcal{G}_c\} | \mathcal{T}) \propto \prod_{k=1}^c \kappa(\mathcal{G}_k)$$

$$\mathcal{T} \sim \mathcal{U}(\text{St}(\mathcal{G}))$$

Expected workflow

- PM10 covariates analysis (physical/chemical relation)
- First models implementation and comparison
- Implementation of variations of those simple models, or more complex models from literature
- Gif/Video interactive plots for displaying results.

Here you can find some trials: <https://drive.google.com/drive/folders/1CbcHMSkEIXhIL8-yNpRa0raeXJ2nQIH6>

References



Garritt L. Page, Fernando A. Quintana, David B. Dahl (2022)
Dependent Modeling of Temporal Sequences of Random Partitions. *Journal of Computational and Graphical Statistics*, 31:2, 614-627.



Mozdzen A., Cremaschi A., Cadonna A., Guglielmi A., Kastner G. (2022)
Bayesian modeling and clustering for spatio-temporal areal data: An application to Italian unemployment. *Spatial Statistics* 52, 100715.



Leonardo V. Teixeira, Renato M. Assunção, Rosangela H. Loschi (2019)
Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees. *Journal of Machine Learning Research* 20, 85, 1-35.