First Meeting Presentation

Project A2

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

Politecnico of Milano
Bayesian Statistics course

Project revision of
November 23, 2023

Presentation Flow

Goals

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

we are unsure year was missing

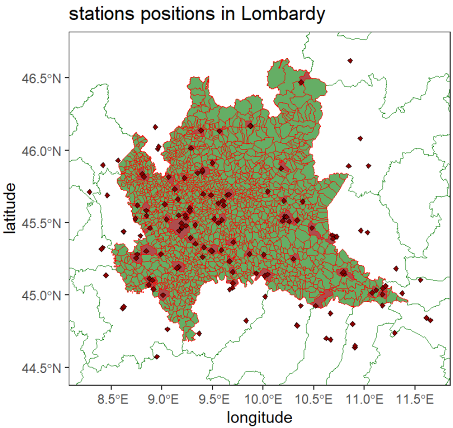
Page [2022]

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

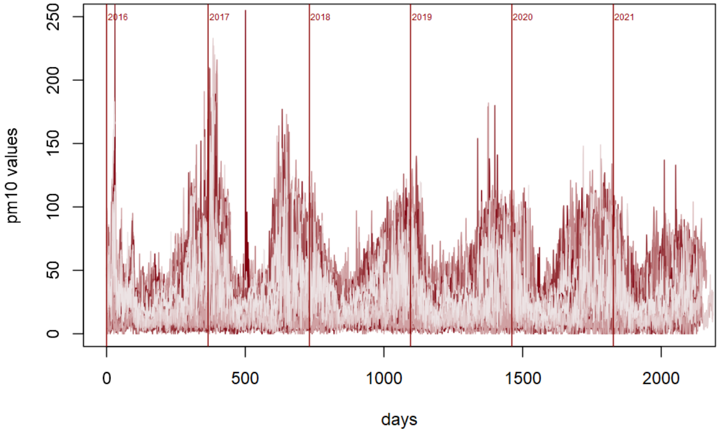
About the dataset

- **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

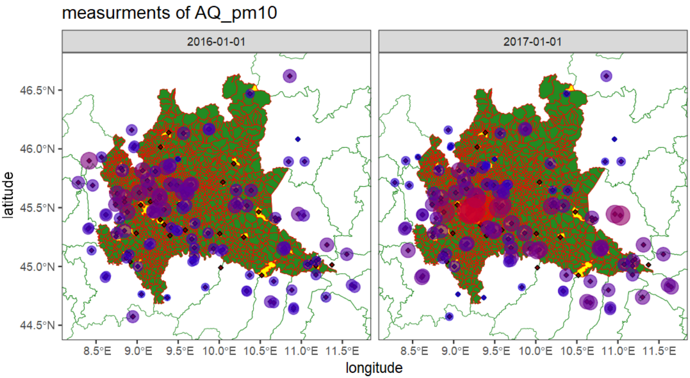


Temporal Exploration

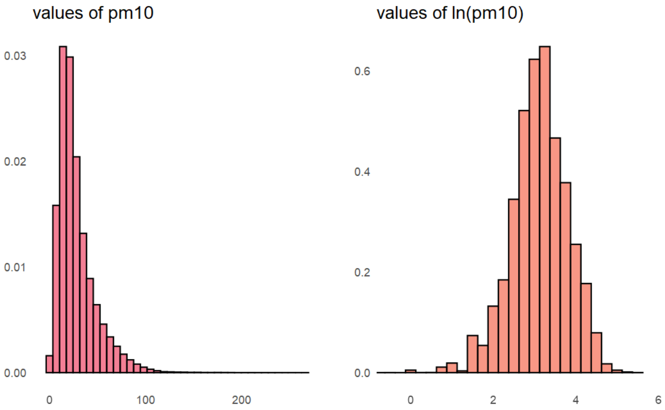


PM10 concentration

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



PM10 distribution

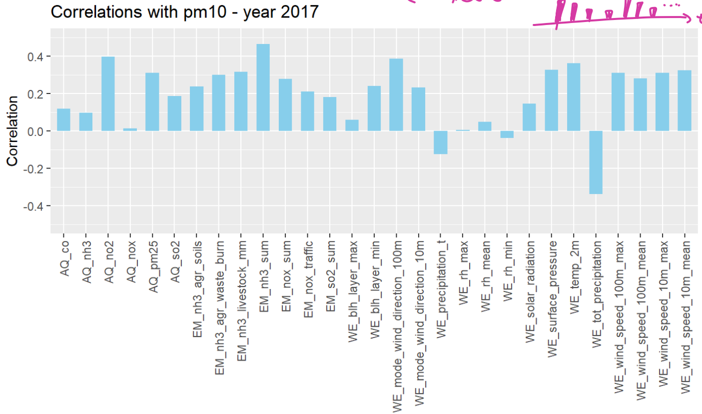


Correlation: PM10 and others

we should also check the corr in time. we avoid if we have it, for our models

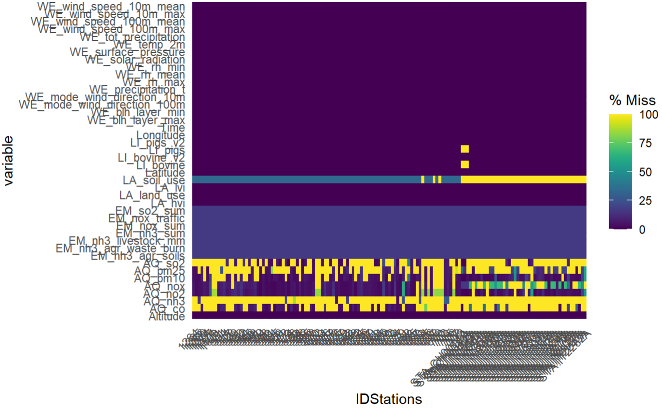
We compute the spearman's correlation index between PM10 and the other covariates, which in the functional framework quantifies with a value in $[-1, 1]$ the tendency of 2 random variables X_t and Y_t to be perfect monotone functions one of each other.

should be the AUTO-CORR plot



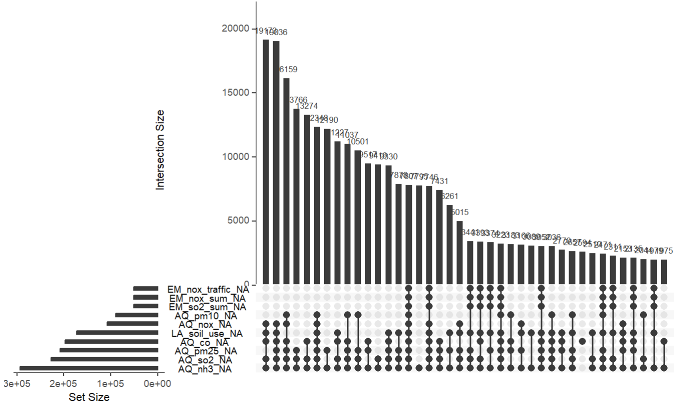
Missing value exploration

General patter on missing values.



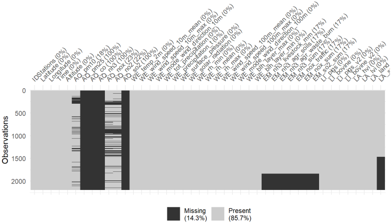
Missing value exploration

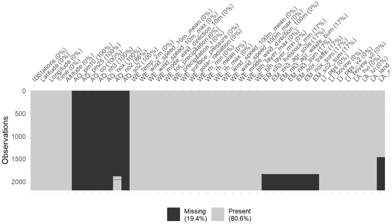
Combinations of variables with the most missing values.



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"
```

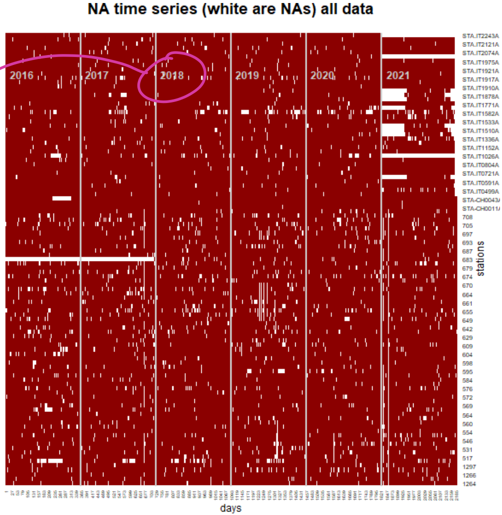
```
            IDStations                Latitude
                     0                       0
             Longitude                    Time
                     0                       0
              Altitude                 AQ_pm10
                     0                   12719
               AQ_pm25                   AQ_co
                136792                  136560
                AQ_nh3                  AQ_nox
                217540                   72988
                AQ_no2                  AQ_so2
                 21266                  156205
           WE_temp_2m    WE_wind_speed_10m_mean
                     0                       0
   WE_wind_speed_10m_max WE_mode_wind_direction_10m
                     0                       0
   WE_tot_precipitation      WE_precipitation_t
                     0                       0
    WE_surface_pressure      WE_solar_radiation
                     0                       0
             WE_rh_min              WE_rh_mean
                     0                       0
             WE_rh_max   WE_wind_speed_100m_mean
                     0                       0
  WE_wind_speed_100m_max WE_mode_wind_direction_100m
                     0                       0
      WE_blh_layer_max        WE_blh_layer_min
                     0                       0
    EM_nh3_livestock_mm       EM_nh3_agr_soils
                 38325                   38325
  EM_nh3_agr_waste_burn             EM_nh3_sum
                 38325                   38325
        EM_nox_traffic              EM_nox_sum
                 38325                   38325
            EM_so2_sum                 LI_pigs
                 38325                    6576
             LI_bovine              LI_pigs_v2
                  6576                       0
          LI_bovine_v2                  LA_hvi
                     0                       0
                LA_lvi             LA_land_use
                     0                       0
           LA_soil_use
                136656
```

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

Missing value exploration

Missing values over time by stations.

- professor suggests to use 2018
- our models to use also two years



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 $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$, where $\boldsymbol{s}_i = (\text{lat}_i, \text{long}_i)$.
- There we record data y_i and (possibly) covariates $\boldsymbol{x}_i$, 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):

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

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

Purely spatial model

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

$$\mathbb{P}(\rho = \{S_1, \ldots, S_k\}) \propto \prod_{h=1}^{k} C(S_h, \boldsymbol{s}_h^\star)$$

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

Purely spatial model

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

$$\mathbb{P}(\rho = \{S_1, \ldots, S_k\}) \propto \prod_{h=1}^{k} C(S_h, \boldsymbol{s}_h^\star)$$

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

Purely spatial model

- We have n distinct locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$, where $\boldsymbol{s}_i = (\text{lat}_i, \text{long}_i)$.
- There we record data y_i and (possibly) covariates $\boldsymbol{x}_i$, 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):

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

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

Spatial and temporal model

We have n distinct locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, where $\boldsymbol{s}_i = (\text{lat}_i, \text{long}_i)$.
There we record data y_i and (possibly) covariates $\boldsymbol{x}_i$, for $i = 1, \ldots, n$.
The goal is to define a model for partitioning them into k_t groups, with t spanning over $1, \ldots, T$.
So we define $\rho_t = \{S_{1,t}, \ldots, S_{k_t,t}\}$ the cluster set variable (with $S_{h,t} \subseteq \{1, \ldots, n\}$ for $h = 1, \ldots, k_t$).
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 $\rho_1, \ldots, \rho_T$.
This update can be explicited 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.

$$Y_{it} \mid \boldsymbol{\mu}_t^{\star}, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{c}_t \overset{\text{ind}}{\sim} \mathcal{N}(\mu^{\star}_{c_{it}t}, \sigma^{2\star}_{c_{it}t}) \quad i = 1, \dots, n \quad \text{and} \quad t = 1, \dots, T$$

$$(\mu_{jt}, \sigma_{jt}) \mid \theta_t, \tau_t^2 \overset{\text{ind}}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \quad j = 1, \dots, k_t$$

$$(\theta_t, \tau_t) \overset{\text{iid}}{\sim} \mathcal{N}(\phi_0, \lambda^2) \times \mathcal{U}(0, A_\tau) \quad t = 1, \dots, T$$

$$(\phi_0, \lambda) \sim \mathcal{N}(m_0, s_0^2) \times \mathcal{U}(0, A_\lambda)$$

$$\{\boldsymbol{c_t}, \dots, \boldsymbol{c_T}\} \sim \text{tRPM}(\boldsymbol{\alpha}, M) \quad \text{with } \alpha_t \overset{\text{iid}}{\sim} \text{Beta}(a_\alpha, b_\alpha)$$

When explaining / writing models, the "prior is one", there are no priors. All the others are marginal laws. The prior is over on the joint params law in class.

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

$$
\begin{gathered}
Y_{it} \mid \beta_t^\star, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{c}_t \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_{it}^T \beta^\star_{c_{it}t}, \sigma^{2\star}_{c_{it}t}) \\
(\beta_{jt}, \sigma_{jt}) \mid \theta_t, \tau_t^2 \overset{\text{ind}}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \\
\vdots
\end{gathered}
$$

or we can further characterize the time dependance with some AR($\cdot$) model

$$
\begin{gathered}
Y_{it} \mid Y_{it-1}, \beta_t^\star, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{c}_t \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_{it}^T \beta^\star_{c_{it}t} + Y_{it-1}\eta_i, \sigma^{2\star}_{c_{it}t}(1-\eta_i)^2) \\
Y_{i1} \mid \beta_1^\star, \boldsymbol{\sigma}_1^{2\star}, \boldsymbol{c}_1 \sim \mathcal{N}(\boldsymbol{x}_{i1}^T \beta^{2\star}_{c_{i1}1}, \sigma^{2\star}_{c_{i1}1}) \\
\vdots
\end{gathered}
$$

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.

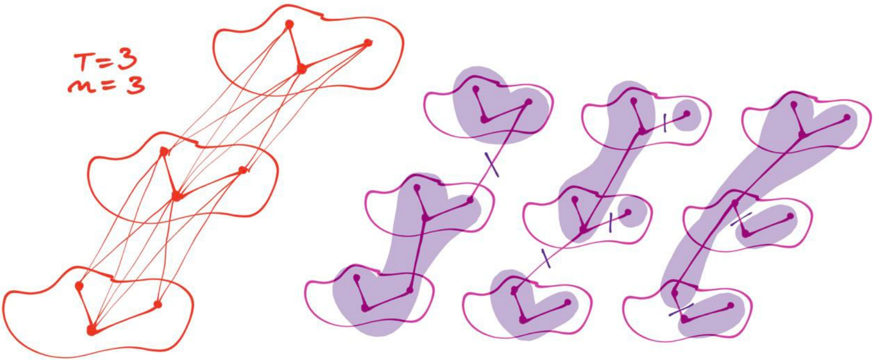
$$
\begin{aligned}
Y_{it} | \boldsymbol{x}_{it}, \beta^\star_{s_i}, w_{it}, \sigma^2, s_i &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_{it}^T \beta^\star_{s_i} + w_{it}, \sigma^2) \\
\boldsymbol{w}_t | \boldsymbol{w}_{t-1}, \boldsymbol{\xi}^\star_{\boldsymbol{s}}, \boldsymbol{s}, \tau^2, \rho, W &\sim \mathcal{N}_I(\text{diag}(\boldsymbol{\xi}^\star_{\boldsymbol{s}}) \boldsymbol{w}_{t-1}, \tau^2 Q(\rho, W)^{-1}) \\
\boldsymbol{w}_1 | \tau^2, \rho, W &\sim \mathcal{N}_I(\mathbf{0}, \tau^2 Q(\rho, W)^{-1}) \\
\sigma^2 &\sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
\tau^2 &\sim \text{Inv-Gamma}(a_{\tau^2}, b_{\tau^2}) \\
\rho &\sim \text{Beta}(\alpha_\rho, \beta_\rho) \\
\boldsymbol{s} | \alpha &\sim \text{PólyaUrn}(\boldsymbol{s} | \alpha) \\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
\boldsymbol{\phi}^\star_1, \ldots, \boldsymbol{\phi}^\star_{K_I} | \boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}, \alpha_\xi, \beta_\xi &\overset{\text{iid}}{\sim} \mathrm{P}_0, \qquad \boldsymbol{\phi}^\star_j = (\boldsymbol{\beta}^\star_j, \boldsymbol{\xi}^\star_j) \qquad \mathrm{j} = 1, \ldots, \mathrm{K}_I \\
\mathrm{P}_0(d\boldsymbol{\phi}^\star) &= \mathcal{N}_{p+1}(d\boldsymbol{\beta}^\star | \boldsymbol{\mu}_{\mathbf{0}}, \Sigma_0) \text{Beta}_{(-1,1)}(d\xi^\star | \alpha_\xi, \beta_\xi)
\end{aligned}
$$

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 $j = 1, \ldots, n$ and $t = 1, \ldots, T$, we now move to Y_i for $i \in I = \{1, \ldots, nT\}$ 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 $\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\}$ 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 $\beta_{\mathcal{G}_1}, \ldots, \beta_{\mathcal{G}_c}$.

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.

$$
\begin{aligned}
Y_i \mid \mathcal{T}, \boldsymbol{\pi}, \beta_{\mathcal{G}_k} &\overset{\text{iid}}{\sim} f(Y_i \mid \beta_{\mathcal{G}_k}; \boldsymbol{x}_i) \quad i \in \mathcal{G}_k \\
\beta_{\mathcal{G}_1}, \ldots, \beta_{\mathcal{G}_c} \mid \mathcal{T}, \boldsymbol{\pi} &\sim \prod_{k=1}^{c} f(\beta_{\mathcal{G}_k}) \\
\mathbb{P}(\boldsymbol{\pi} = \{\mathcal{G}_1, \ldots, \mathcal{G}_c\} \mid \mathcal{T}) &\propto \prod_{k=1}^{c} \kappa(\mathcal{G}_k) \\
\mathcal{T} &\sim \mathcal{U}(\mathrm{St}(\mathcal{G}))
\end{aligned}
$$

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`

- meeting in January with the professor

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.