Clustering spatial functional data

Vincent Vandewalle, Cristian Preda, Sophie Dabo

Clustering spatial functional data

Vincent VANDEWALLE[1,2]
Joint work with Cristian PREDA[2,3] and Sophie DABO[2,4]

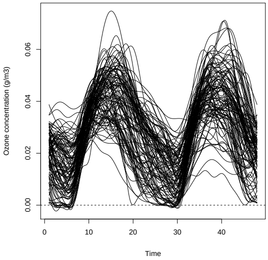
[1] Univ. Lille, EA2694 Santé publique: épidémiologie et qualité des soins
[2] Inria
[3] UMR 8524, Laboratoire Paul Painlevé
[4] UMR 9221, Laboratoire Lille Economie et Management

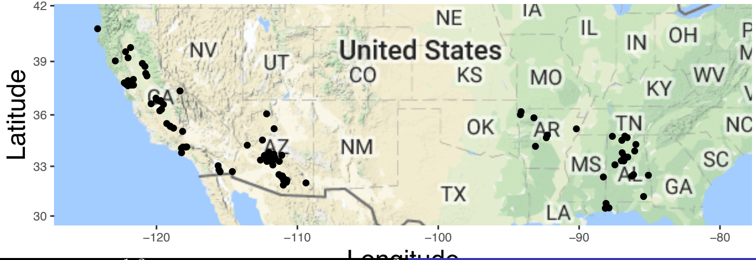
ERCIM 2018
Sunday 16th December 2018
Pisa

Outline

Running example: US ozone data (*https://www.epa.gov/outdoor-air-quality-data*)

- 106 monitoring stations of United States in 2015
- Ozone recorded hourly from July 19 at 12 am to July 20 at 11 pm
- Expansion of the discrete data (48 data at each station) in terms of 25 Fourier basis functions





State of the art

Clustering of independent functional data

- k-means techniques adjusted to functional data, hierarchical algorithm mainly for independent data
- Revue of clustering methods for functional data under the independent model provided in Jacques and Preda (2014).

State of the art

Clustering of independent functional data

- k-means techniques adjusted to functional data, hierarchical algorithm mainly for independent data
- Revue of clustering methods for functional data under the independent model provided in Jacques and Preda (2014).

Clustering of spatio-functional data: observation of dependent curves at some spatial locations

- Few works: Dabo-Niang et al. (2010), Giraldo et al. (2012) extended some approaches on hierarchical clustering to the context of spatially correlated functional
- Dabo-Niang et al. (2010) : Spatial variation taken into account by using kernel mode and density estimation
- Giraldo et al. (2011): Similarity between two curves by the trace-variogram
- Other approaches: see Romano et al. (2015) and Romano et al. (2017).

Basic notations for functional data

- $X = (X_s,\ s \in \mathbb{R}^N)$, a measurable spatial process $N \geq 1$, defined on some probability space $(\Omega,\ \mathcal{A}, \mathbf{P})$
- X observed on some spatial region $\mathcal{I} \subseteq \mathbb{R}^N$ of cardinal n, $\mathcal{I} = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^N$, $i = 1 \ldots n$.
- In each location $s \in \mathcal{I}$, the random variables X_s are valued in a metric space $(\mathcal{E},\ d)$ of eventually infinite dimension and are locally identically distributed.
- $d(.,.)$ is some measure of proximity, for instance a metric or a semi-metric
- u and v close $\Rightarrow$ X_u and X_v have same or similar distributions, less restrictive that strict stationarity
- In Functional Data Analysis (FDA): $\mathcal{E}$ is a space of functions, typically the space of squared integrable functions defined on some finite interval $\mathcal{T} = [0, T]$, $T > 0$.
- S the set of the n curves, $S = \{X_s, s \in \mathcal{I}\}$.

Clustering of spatial function data: what focus and how?

Focus

- Clustering of the times
- Clustering of the spatial regions
- Clustering of both times and spatial regions
- Global explicative model of the process without particular clustering structure

Clustering of spatial function data: what focus and how?

Focus

- Clustering of the times
- **Clustering of the spatial regions**
- Clustering of both times and spatial regions
- Global explicative model of the process without particular clustering structure

Clustering of spatial function data: what focus and how?

Focus

- Clustering of the times
- **Clustering of the spatial regions**
- Clustering of both times and spatial regions
- Global explicative model of the process without particular clustering structure

Method

- Geometric based approach: distance between curves weighted by the spatial proximity
- Model based approach: integrate the spatial aspect in the model

Clustering of spatial function data: what focus and how?

Focus

- Clustering of the times
- **Clustering of the spatial regions**
- Clustering of both times and spatial regions
- Global explicative model of the process without particular clustering structure

Method

- Geometric based approach: distance between curves weighted by the spatial proximity
- **Model based approach: integrate the spatial aspect in the model**

The proposed model: main idea

Standard mixture model

- Let Z a latent categorical random variable defining G clusters
- Let f the probability distribution of X and f_g the probability distribution of X given $Z = g$.
- The mixture model pdf:

$$f(x) = \sum_{g=1}^{G} \pi_g f_g(x),$$

where $\pi_g = P(Z = g)$ is the prior probability of cluster g.

Extension to the spatial setting: involving the location s into the priors

$$f(x|s) = \sum_{g=1}^{G} \pi_g(s) f_g(x),$$

the distribution of observations within the cluster g is independent of the location $\Rightarrow$ spatial dependency captured by the priors $\pi_g(s)$.

The proposed model: parametrisation

Parametric framework

- f_g is depending on parameters θ_g
- π_g depending on parameters β
- $\theta = (\theta_1, \ldots, \theta_G, \beta)$

$$f(x|s;\theta) = \sum_{g=1}^{G} \pi_g(s;\beta) f_g(x;\theta_g).$$

Remarks

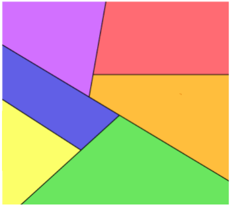
- Allows to perform model choice
- Need to define accurate parametric model on f_g et π_g

The proposed model: parametrisation of the priors

Prior model

- $\pi_g(s) = \pi_g(s; \beta)$
- Cheam et al. (2017) for clustering spatio-temporal data.
- Multinomial logistic regression as a model for the $\pi_g(s; \beta)$,

$$\ln \frac{\pi_g(s; \beta)}{\pi_G(s; \beta)} = \beta_{0g} + \langle \beta_g, s \rangle_{\mathbb{R}^N}.$$



Pros

- Defines simple and interpretable boundaries
- Tractable computation

Cons

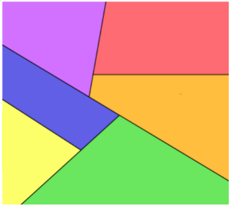
- Low flexibility
- May enforce some predefined structure on clustering

The proposed model: parametrisation of the priors

Prior model

- $\pi_g(s) = \pi_g(s;\beta)$
- Cheam et al. (2017) for clustering spatio-temporal data.
- Multinomial logistic regression as a model for the $\pi_g(s;\beta)$,

$$\ln \frac{\pi_g(s;\beta)}{\pi_G(s;\beta)} = \beta_{0g} + \langle \beta_g, s\rangle_{\mathbb{R}^N}.$$



Pros

- Defines simple and interpretable boundaries
- Tractable computation

Cons

- Low flexibility: **possibility to use non-linear logistic regression**
- May enforce some predefined structure on clustering: **possible model selection**

The proposed model: parametrization of the functional part

Density for functional data?

- $X|Z = g \sim$ Gaussian process
- The pseudo-density defined in Delaigle and Hall (2010) is considered:

$$f_g^{(q_g)}(x;\theta_k) = \prod_{j=1}^{q_g} f_{gj}(c_{gj}(x);\lambda_{gj}),$$

  - f_{gj} is the p.d.f of j-th principal component C_{gj} of X within the cluster g
  - C_{gj} $(j = 1, \ldots, q_g)$ independent Gaussian zero-mean with variance equal to the eigen values λ_{gj} of the covariance operator of X
- q_g and d need to be defined.

The proposed model: parametrization of the functional part

Density for functional data?

- $X|Z = g \sim$ Gaussian process
- The pseudo-density defined in Delaigle and Hall (2010) is considered:

$$f_g^{(q_g)}(x;\theta_k) = \prod_{j=1}^{q_g} f_{gj}(c_{gj}(x);\lambda_{gj}),$$

  - f_{gj} is the p.d.f of j-th principal component C_{gj} of X within the cluster g
  - C_{gj} $(j = 1, \ldots, q_g)$ independent Gaussian zero-mean with variance equal to the eigen values λ_{gj} of the covariance operator of X
- q_g and d need to be defined.

Proposal if $X(t) = \sum_{j=1}^{d} \alpha_j \phi_j(t)$

$$f_g^{(q_g)}(x;\theta_k) = \prod_{j=1}^{q_g} f_{gj}(c_{gj}(x);\lambda_{gj}) \prod_{j'=q_g+1}^{d} f_{gj'}(c_{gj'}(x),\bar{\lambda}_g),$$

- $\bar{\lambda}_g$ mean of the eigen values $\lambda_{gj'}$ $(j' = q_g + 1, \ldots, d)$ of the covariance operator of X
- $f_g^{(q_g)}$: true density if $X \in \text{span}(\{\phi_1, \ldots, \phi_d\})$

Remarks

Equivalence with parsimonious high dimensional model (Bouveyron et al., 2007)

- d is chosen as the dimension of the basis which has been used to perform the smoothing of the data
- C_{kj} can be obtained by performing PCA on the expansion coefficients of X in the metric M given by the inner product of the basis functions
- learning data: expansion coefficients multiplied by $M^{1/2}$ $\Leftrightarrow$ learning a parsimonious high dimensional model (see Bouveyron et al. (2007))

An other possible model

Ruiz-Medina et al. (2014) propose a mixed-effect model, in which the fix effect can take into account the spatial dependencies, assuming a spatial autoregressive dynamic for the random effect, they propose a functional classification criterion to detect local spatially homogeneous regions

Parameters estimation (1/2)

log-likelihood of the sample of curves $S = \{x_s,\ s \in \mathcal{I}\}$

$$\ell(\theta; S) = \sum_{s\in\mathcal{I}} \log\left(\sum_{g=1}^{G} \pi_g(s;\beta) f_g^{(q_g)}(x_s;\theta_g)\right).$$

Completed log-likelihood

$Z_g(s)$ the indicator random variable for the cluster g at location s.

$$\ell_c(\theta; S, Z) = \sum_{s\in\mathcal{I}} \sum_{g=1}^{G} Z_g(s)\left(\log \pi_g(s;\beta) + \log f_g^{(q_g)}(x_s;\theta_g)\right),$$

Parameters estimation (2/2)

The EM algorithm

Start with an initial random partition of data S into G clusters. Let $\theta^{(h)}$ be the estimated parameter value at iteration $h \geq 0$ of the algorithm

- **E step**

$$t_g^{(h+1)}(s) = E_{\theta^{(h)}}[Z_g(s)|s] = \frac{\pi_g(s;\beta^{(h)}) f_g^{(q_g)}(x_s;\theta_g^{(h)})}{\sum_{\ell=1}^{G} \pi_\ell(s;\beta^{(h)}) f_\ell^{(q_\ell)}(x_s;\theta_\ell^{(h)})}.$$

- **M step**

$$\theta_g^{(h+1)} = \arg\max_{\theta_g} \sum_{s\in\mathcal{I}} t_g^{(h+1)}(s) \log f_g^{(q_g)}(x_s;\theta_g),$$

$$\beta^{(h+1)} = \arg\max_{\beta} \sum_{s\in\mathcal{I}} \sum_{g=1}^{G} t_g^{(h+1)}(s) \log \pi_g(s;\beta).$$

Notice that $\beta^{(h+1)}$ is obtained as solution of a weighted logistic regression.

For homoscedastic models just a simple modification of the update $\theta_g^{(h+1)}$. See also Bouveyron et al. (2007) for more details.

Model selection

Selection of G when q_g $(g = 1, \ldots, G)$ are known

Maximize the Bayesian Information Criterion (BIC) criterion:

$$BIC(G) = \log \ell(G) - \frac{\nu_G}{2} \log(n),$$

ν_G is the number of parameters of the model: spatial mixing proportions, center means, principal scores and variances and $n = |\mathcal{I}|$.

Model selection

Selection of G when q_g $(g = 1, \ldots, G)$ are known

Maximize the Bayesian Information Criterion (BIC) criterion:

$$BIC(G) = \log \ell(G) - \frac{\nu_G}{2} \log(n),$$

ν_G is the number of parameters of the model: spatial mixing proportions, center means, principal scores and variances and $n = |\mathcal{I}|$.

Selection of G when q_g $(g = 1, \ldots, G)$ are **unknown**

Modified M step which tries to maximize the conditional expectation of the BIC criterion:

$$(q_g^{(h+1)}, \theta_g^{(h+1)}) \quad = \arg\max_{(q_g,\theta_g)} \sum_{s\in\mathcal{I}} t_g^{(h+1)}(s) \log f_g^{(q_g)}(x_s; \theta_g) - \frac{\nu_{q_g}}{2} \log n,$$

where ν_{q_g} is the number of parameters required for the model with q_g principal components.

Model selection

Selection of G when q_g $(g = 1, \dots, G)$ are known

Maximize the Bayesian Information Criterion (BIC) criterion:

$$BIC(G) = \log \ell(G) - \frac{\nu_G}{2} \log(n),$$

ν_G is the number of parameters of the model: spatial mixing proportions, center means, principal scores and variances and $n = |\mathcal{I}|$.

Selection of G when q_g $(g = 1, \dots, G)$ are **unknown**

Modified M step which tries to maximize the conditional expectation of the BIC criterion:

$$(q_g^{(h+1)}, \theta_g^{(h+1)}) \quad = \arg\max_{(q_g, \theta_g)} \sum_{s \in \mathcal{I}} t_g^{(h+1)}(s) \log f_g^{(q_g)}(x_s; \theta_g) - \frac{\nu_{q_g}}{2} \log n,$$

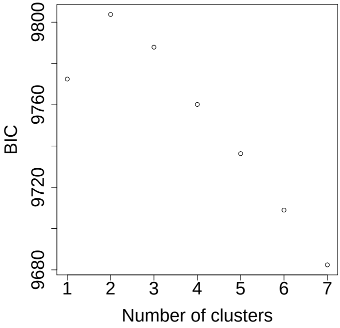
where ν_{q_g} is the number of parameters required for the model with q_g principal components.

Remark

Also possible to consider the homoscedastic setting.

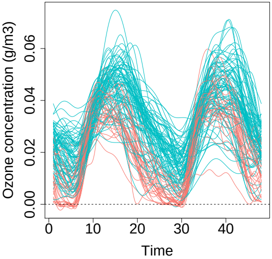
Model choice

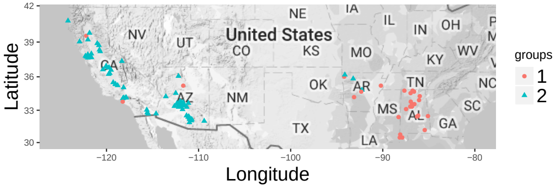
- homocedastic model applied ⇒ more relevant results from the classification point of view
- value of q selected during the EM algorithm by maximizing the BIC computed at each step
- BIC indicates that two or three clusters could be appropriated.

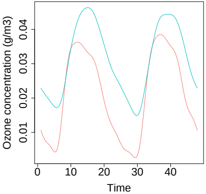


Solution with $G = 2$ clusters

- $q = 18$ principal components retained
- good separation between East cities from the West cities
- good separation between red and blue curves
- in average West cities have higher pollution than Est cities

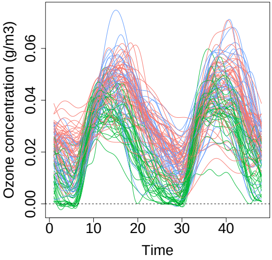


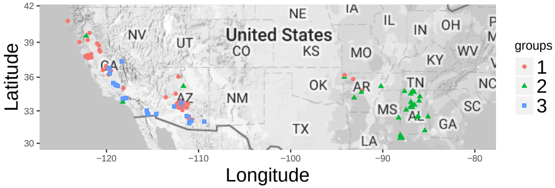


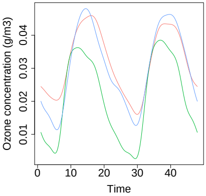


Solution with $G = 3$ clusters

- $q = 17$ principal components retained
- well separation of the East curves from the West curves, but also the North from the South for the West side.
- we see that it is the six first hours that well separate cluster 1 (North) from cluster 3 (South).







Conclusion

- Sparse model to take into account spatial dependency
- Smoothing of the clustering based on spatial location

Conclusion

- Sparse model to take into account spatial dependency
- Smoothing of the clustering based on spatial location

Perspectives

- Avoid prior smoothing of the data $\Rightarrow$ perform the selection of the basis and other smoothing parameters based on the model
- Propose more flexible model for $\pi_k(s)$

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.

A. Cheam, M. Marbac, and P. McNicholas. Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics*, 28(3), 2017.

S. Dabo-Niang, A.-F. Yao, L. Pischedda, P. Cuny, and F. Gilbert. Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, 24(4): 487–497, 2010.

A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, pages 1171–1193, 2010.

R. Giraldo, P. Delicado, and J. Mateu. Ordinary kriging for function-valued spatial data. *Environ. Ecol. Stat.*, 18(3):411–426, 2011. ISSN 1352-8505.

R. Giraldo, P. Delicado, and J. Mateu. Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, 66(4):403–421, 2012.

J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

E. Romano, J. Mateu, and R. Giraldo. On the performance of two clustering methods for spatial functional data. *AStA Adv. Stat. Anal.*, 99(4):467–492, 2015. ISSN 1863-8171.

E. Romano, A. Balzanella, and R. Verde. Spatial variability clustering for spatially dependent functional data. *Stat. Comput.*, 27(3):645–658, 2017. ISSN 0960-3174.

M. D. Ruiz-Medina, R. M. Espejo, and E. Romano. Spatial functional normal mixed effect approach for curve classification. *Advances in Data Analysis and Classification*, 8(3):257–285, 2014.