Clustering spatial functional data

Vincent Vandewalle, Cristian Preda, Sophie Dabo-Niang

▶ **To cite this version:**

Vincent Vandewalle, Cristian Preda, Sophie Dabo-Niang. Clustering spatial functional data. J. Mateu and R. Giraldo. Geostatistical Functional Data Analysis : Theory and Methods. Editors: Jorge Mateu, Ramon Giraldo, 1, Wiley, 2021, Geostatistical Functional Data Analysis : Theory and Methods, 10.1002/9781119387916.ch7 . hal-01948934

HAL Id: hal-01948934
https://inria.hal.science/hal-01948934

Submitted on 26 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Clustering spatial functional data

Abstract

In this chapter we present two approaches for clustering spatial functional data. The first one is the model-based clustering that uses the concept of density for functional random variables. The second one is the hierarchical clustering based on univariate statistics for functional data such as the functional mode or the functional mean. These two approaches take into account the spatial features of the data: two observations that are spatially close share a common distribution of the associated random variables. The two methodologies are illustrated by an application to air quality data.
Keywords: Model-based clustering, hierarchical clustering, functional data analysis, spatial data.

Contents

0.1 Introduction

The purpose of this chapter is to present two techniques for clustering spatial functional data. Generally, in any clustering framework, data inside each cluster should be as similar as possible but different from those in other clusters. Recent researches on the clustering of independent functional data are available in the literature devoted to functional data analysis (FDA). In particular, k-means techniques are adjusted to functional data, hierarchical algorithm and some of its variants are proposed as well, mainly for independent data (e.g Abraham *et al.* (2006), Dabo-Niang *et al.* (2007), Auder and Fischer (2012), Abraham *et al.* (2003), Chiou and Li (2007), Cuevas *et al.* (2001), Dabo-Niang *et al.* (2007), García-Escudero and Gordaliza (2005), Romano *et al.* (2015), Tarpey and Kinateder (2003), Jacques and Preda (2014b)). A revue of clustering methods for functional data under the independent model is provided in Jacques and Preda (2014a). Other model-based approaches for clustering functional data are given in Floriello and Vitelli (2017) and James and Sugar (2003). In several domains, data are of spatio-functional nature, observations may be dependent curves at some spatial locations and clustering these data taking into account the spatial dependency can be more accurate. The independence hypothesis does not hold in this case. Few works exist on such dependent data: Dabo-Niang *et al.* (2010), Giraldo *et al.* (2012) extended some approaches on hierarchical clustering to the context of spatially correlated functional. Giraldo *et al.* (2012) measured the similarity between two curves by the trace-variogram (Giraldo *et al.* (2011)) while the spatial variation is taken into account by using kernel mode and density estimation in Dabo-Niang *et al.* (2010). Other approaches for clustering spatial functional data are given recently in Romano *et al.* (2015) and Romano *et al.* (2017).
An appropriate clustering approach should lead to homogeneous clusters and heterogeneity between them. Consequently, the number of clusters is an important issue (e.g. Milligan and Cooper (1985)). Some clustering methods do not automatically determine the number of clusters. Techniques are developed in the literature to overcome this difficulty. Most of them propose to estimate or to select the number of clusters by solving an optimization problem involving some cluster homogeneity index (e.g. Milligan and Cooper (1985), Krzanowski and Lai (1988), Cuevas *et al.* (2000)).

We deal with a measurable spatial process $X = \left(X_s,\ s \in \mathbb{R}^N\right),\ N \geq 1$, defined on some probability space $\left(\Omega,\ \mathcal{A}, \mathbf{P}\right)$. Assume that the process X is observed on some spatial region $\mathcal{I} \subseteq \mathbb{R}^N$ of cardinal n, $\mathcal{I} = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^N$, $i = 1 \ldots n$. We assume also that for each location $s \in \mathcal{I}$, the random variables X_s are valued in a metric space $\left(\mathcal{E},\ d\right)$ of eventually infinite dimension and are locally identically distributed (see for instance Klemelä (2008)). Here $d(.,.)$ is some measure of proximity, for instance a metric or a semi-metric. This means that when a site u is close enough to site v, the variables X_u and X_v have same or similar distributions. This assumption is less restrictive that strict stationarity. It is motivated by the fact that one can imagine that, variables located at neighbors sites may be similar

and have the same local distribution that may be different to the local distribution of another set of variables at other locations. In the classical framework of FDA, the space $\mathcal{E}$ is a space of functions, typically the space of squared integrable functions defined on some finite interval $\mathcal{T} = [0, T], T > 0$.
Let denote with S the set of the n curves, $S = \{X_s, s \in \mathcal{I}\}$ (renamed sometime in an arbitrary way, $S = \{X_1, ..., X_n\}$).
First, we present the problem of clustering spatial functional data generated by a mixture of Gaussian processes with logistic prior weights depending on the location. Second, we present an extension to spatial data of the method studied by Dabo-Niang *et al.* (2007) which is a descendant HC procedure based on distances between the modal and mean curves of a set of curves. The two approaches are illustrated with pollution data.

0.2 Model-based clustering for spatial functional data

In the framework of clustering, the model-based techniques assume that there exists a latent categorical random variable Z defining G clusters of data such that probability distribution of data is a mixture of cluster distributions. Let denote by f the probability distribution of X and by f_g the probability distribution of X given $Z = g$. Then, the mixture model is written as

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

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

In the particular case of spatial dependency we extend the model given in Equation (1) by involving the location s ($s \in \mathcal{I}$) into the priors probabilities of clusters. The mixture model becomes:

$$f(x|s) = \sum_{g=1}^{G} \pi_g(s; \beta) f_g(x), \tag{2}$$

where β is some parametrization of the spatial prior. Thus, conditionally to the cluster $Z = g$, the distribution of observations within the cluster is independent of the location, all spatial dependency being captured by the priors $\pi_g(s; \beta)$. This idea is used in Cheam *et al.* (2017) for clustering spatio-temporal data. The authors propose the multinomial logistic regression as a model for the $\pi_g(s; \beta)$,

$$\ln \frac{\pi_g(s; \beta)}{\pi_G(s; \beta)} = \beta_{0g} + \langle \beta_g, s \rangle_{\mathbb{R}^N}. \tag{3}$$

In a parametric framework, the conditional distribution f_g is depending on parameters θ_g. For example, in the Gaussian model θ_g is the mean and the covariance

matrix of cluster g. Let denote by θ the set of all parameters including also those defining the $\pi_g(s;\beta)$. Thus the model becomes

$$f(x|s;\theta) = \sum_{g=1}^{G} \pi_g(s;\beta) f_g(x;\theta_g). \tag{4}$$

In the finite dimensional setting (see for instance Celeux and Govaert (1995)) the multivariate probability density function is the main tool for estimating such a model using the EM algorithm. For functional random variables the notion of probability density in not well defined because of the infinite dimension of data. To overcome this difficulty, James and Sugar (2003) and Bouveyron and Jacques (2011) use the expansion coefficients of X into some finite basis of functions. This approach allows them to get a well defined probability density function on the coefficients. In Delaigle and Hall (2010) the functional principal component analysis is used to define a surrogate of the probability density for functional data. This approach is used in the context of model based clustering in Jacques and Preda (2013) and Jacques and Preda (2014b). In a spatial setting Ruiz-Medina *et al.* (2014) have proposed a mixed-effect model, in which the fix effect can take into account the spatial dependencies. Moreover, assuming a spatial autoregressive dynamic for the random effect, they propose a functional classification criterion to detect local spatially homogeneous regions. In what follows we assume that given $Z = g$, X is a Gaussian process. Then, within the cluster g, we consider a modified version of the pseudo-density defined in Delaigle and Hall (2010):

$$f_g^{(q_g)}(x;\theta_k) = \prod_{j=1}^{q_g} f_{gj}(c_{gj}(x);\lambda_{gj}) \prod_{j'=q_g+1}^{d} f_{gj'}(c_{gj'}(x);\bar{\lambda}_g), \tag{5}$$

where f_{gj} is the probability density of the j-th principal component C_{gj} of X within the cluster g. The random variables C_{gj} ($j = 1,\ldots,q_g$) are independent Gaussian zero-mean with variance equal to the eigen values λ_{gj} of the covariance operator of X, and the random variables $C_{gj'}$ ($j' = q_g+1,\ldots,d$) are independent Gaussian zero-mean with variance equal to the mean $\bar{\lambda}_g$ of the eigen values $\lambda_{gj'}$ ($j' = q_g+1,\ldots,d$) of the covariance operator of X. Thus the parameters $\theta_g = (\lambda_{g1},\ldots,\lambda_{gq_g},\bar{\lambda}_g)$, q_g and d need to be defined. Notice that compared to the definition of Delaigle and Hall (2010), we have added the term $\prod_{j'=q_g+1}^{d} f_{gj'}(c_{gj'}(x);\bar{\lambda}_g)$.

In fact the proposed surrogate density car be interpreted as a true density if the functional data belong to a finite dimensional space of functions spanned by some basis $\{\phi_1,\ldots,\phi_d\}$, $d \geq 1$, i.e.

$$X(t) = \sum_{j=1}^{d} \alpha_j \phi_j(t), \quad t \in [0,T], T > 0.$$

Thus we will take d as the dimension of the basis which has been used to perform the smoothing of the data. In this case the principal components C_{kj} of the functional

PCA 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. Thus, if the learning data considered are now the expansion coefficients multiplied by $M^{1/2}$ then the proposed approach can simply be re-interpreted as learning a parsimonious high dimensional model (see Bouveyron *et al.* (2007)) on these new data.

Let notice that it is also possible to consider sparse versions of the mixture model such as for instance to consider the homoscedastic setting (equal covariance process by cluster).

0.2.1
The Expectation-Maximization (EM) algorithm

We are now ready to describe the EM algorithm for estimating θ and therefore the clustering.

As in the finite setting, based on Equation (5) we define a likelihood of the sample of curves $S = \{x_s,\ s \in \mathcal{I}\}$ by:

$$l(\theta; S) = \prod_{s \in \mathcal{I}} \left(\sum_{g=1}^{G} \pi_g(s; \beta) f_g^{(q_g)}(x_s; \theta_g) \right). \tag{6}$$

A classical way to maximize the likelihood when data are missing (here the variable Z) is to use the iterative EM algorithm. We use this algorithm to maximize the likelihood (6), and adapt it for updating the principal components scores of each group as well as the parameters β defining the $\pi_g(s)$ in (3).
The algorithm consists in maximizing the approximated completed log-likelihood. Let $Z_g(s)$ be the indicator random variable for the cluster g at location s. Then the completed log-likelihood is given by:

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

which is known to be easier to maximize than its incomplete version. Let $\theta^{(h)}$ be the estimated parameter value at iteration $h \geq 0$ of the algorithm.

E step.

As the groups belonging $Z_g(s)$'s are unknown, the **E** step consists in computing the conditional expectation of the approximated completed log-likelihood:

$$\begin{aligned} \mathcal{Q}(\theta; \theta^{(h)}) &= E_{\theta^{(h)}}[L_c(\theta; S, Z)|S] \\ &= \sum_{s \in \mathcal{I}} \sum_{g=1}^{G} t_g^{(h+1)}(s) \left(\log \pi_g(s; \beta) + \log f_g^{(q_g)}(x_s; \theta_g) \right) \end{aligned}$$

where $t_g^{(h+1)}(s)$ is the probability for the curve X_s to belong to the group g conditionally to $C_{gj} = c_{gj}(x_s), j = 1, \ldots, q_g$:

$$t_g^{(h+1)}(s) = E_{\theta^{(h)}}[Z_g(s)|s] = \frac{\pi_g(s;\beta^{(h)}) f_g^{(q_g)}(x_s;\theta_g^{(h)})}{\sum_{\ell=1}^{G} \pi_\ell(s;\beta^{(h)}) f_\ell^{(q_\ell)}(x_s;\theta_\ell^{(h)})}. \quad (7)$$

M step.

The M step consists in maximizing the conditional expectation of the completed log-likelihood with respect to θ:

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

and

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

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

The EM algorithm starts with an initial random partition of data S into G clusters.

Notice that for homoscedastic models one has a modification of the update θ_g^{h+1}. See also Bouveyron *et al.* (2007) for more details.

0.2.2 Model selection

In order to select the number of cluster G when q_g $(g = 1, \ldots, G)$ are known, we propose to maximize the Bayesian Information Criterion (BIC) criterion defined below:

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

where $\nu_G = (N+1)(G-1)+Gd+\sum_{g=1}^{G}(q_g(d-(q_g-1)/2)+1)$ is the number of parameters of the model (spatial mixing proportions, center means, principal scores and variances) and $n = |\mathcal{I}|$.

When the values q_g $(g = 1, \ldots, G)$ are unknown they can be selected in order to maximize the BIC criterion by considering the following modified M step which tries to maximize the conditional expectation of the BIC criterion:

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

where $\nu_{q_g} = q_g(d - (q_g - 1)/2)$ is the additional number of parameters required for the model with q_g principal components.

Let notice that if we consider the homoscedastic setting, the value of the BIC criterion can be easily computed at each step of the EM algorithm for each possible value of q which does not depends on g since in this case this parameter is the same for each cluster. In this case the expression of ν_G would be $\nu_G = (N + 1)(G - 1) + Gd + (q(d - (q - 1)/2) + 1)$.

In section 0.4 we present the results of the application of this technique to air quality data.

0.3 Descendant Hierachical Classification (HC) based on centrality methods

Recent advances in nonparametric FDA allow to define centrality features for a sample of curves (see e.g. Ferraty and Vieu (2006)). Dabo-Niang *et al.* (2007) indicated that both the mean and the median curves are interesting when dealing with homogeneous data, while the modal curve would be more useful for detecting possible different structures in the data. Consequently, Dabo-Niang *et al.* (2007) used a descendant HC method based on comparing the modal curve either with the mean or the median. Location measures (mean, mode and median) summarize the data and aim to provide a representative element of the sample. The spatial mean used is the same as in the *i.i.d.* setting compare to the mode and median.
In our context, for the set of curves S, we define the mean curve as

$$X_{mean,S} = \frac{1}{n} \sum_{s \in \mathcal{I}} X_s.$$

The notion of median curve for i.i.d functional data can be extended to the spatial framework, see Cadre (2001) and Dabo-Niang *et al.* (2007), for general definition in i.i.d data and Dabo-Niang *et al.* (2010), for a heterogeneity spatial index. Here, let the median curve be:

$$X_{median,S} = \arg \min_{X_t \in S,\, t \in \mathcal{I}} \sum_{s \in \mathcal{I}} d_m(X_t, X_s),$$

with $d_m(X_t, X_s) = d(X_t, X_s) W_{s,t}$, where $W_{s,t}$ is a spatial weight. Indeed, the spatial dependency structure between the n spatial units is described by a non-stochastic spatial weights $n \times n$ matrix W_n that depends on n. The elements $W_{s,t} = W_{s,t,n}$ of this matrix are usually considered as inversely proportional to distances between spatial units s and t with respect to some metric (physical distance, social networks or economic distance, see for instance Pinkse and Slade (1998)) and Chapter 13 of this book. Here, the spatial weighted matrix W_n is constructed by taking k-neighbors of each spatial unit using K-Nearest Neighbor (KNN) method (k Nearest Neighbors Algorithm). This k-neighbors matrix can be computed by for instance, the function *knn2nb* of the R package *spdep* (Bivand *et al.*, 2015) of the software R (R Core Team, 2017).
From a theoretical point of view, the mode, when it exists, is an observation whose probability is locally maximum. So, the modal curve of the sample S can be estimated as:

$$X_{modal,S} = \arg \max_{X_t \in \mathcal{S},\, t \in \mathcal{I}} \sum_{s \in \mathcal{I}} K\left(\frac{d_m(X_t, X_s)}{h}\right)$$

where $K(\cdot)$ is a kernel function, $h = h_n$ is a sequence of positive numbers called bandwidth, considered as a smoothing parameter. The kernel K acts as a weight function: the larger is $d_m(x, X_s)$ and the smaller is $K\left(d_m(x, X_s)/h\right)$. This means that among all the curves X in S, the modal curve defines a spatial neighboring area where the sample of curves is the most dense and dependent. The pertinence of this estimate of a modal curve and asymptotic properties are similar to that given in Dabo-Niang *et al.* (2010). This last assumed a mixing condition on the spatial process and used it to measure the spatial heterogeneity of the data.

The elements K, $d(.,.)$ and h are essential in nonparametric estimation. In the functional context, a semi-metric $d(\cdot,\cdot)$ is often used as a proximity measure. In particular, a semi-metric based on the first q scores of a functional principal components analysis, used in the application section, is defined by

$$m_q^{pca}(X_i, X_j)^2 \quad = \quad \int \left(X_i^{(q)}(t) - X_j^{(q)}(t)\right)^2 dt,$$

where $X^{(q)}$ denotes the vector of the first qth scores components of X (see Ferraty and Vieu (2006) for more details). A kernel K is a weighting function used in nonparametric estimation techniques. There exists a large variety of kernels in the FDA context, the most classical ones are the positive and symmetrical kernels such as, among others, box, triangle, quadratic and Gaussian (see Ferraty and Vieu (2006)). In Section 0.4 we use the following kernel:

$$K(u) = \frac{3}{2}(1 - u^2)1_{(0,1)}(u).$$

We choose this kernel because it gives more relevant results (among several kernel functions investigated) from the classification point of view of the air quality data considered.

Methodology

The proposed methodology performs iteratively by splitting S into increasingly homogeneous classes. To measure the heterogeneity of a given sample S of curves, Dabo-Niang *et al.* (2007) compared modal and mean curves by computing the Subsampling Heterogeneity Index (SHI). The median curve can also be used instead of the mean, e.g. when one wants to assign to the same group all curves that have the same shape but which are affected by some clearly horizontal shift (see Dabo-Niang *et al.* (2007)). The SHI is computed by using a large number L of randomly generated subsamples $S^{(l)} \subset S$ of the same size

$$SHI_{mean}(S) = \frac{1}{L}\sum_{l=1}^{L} \frac{m(X_{modal,S^{(l)}}, X_{mean,S^{(l)}})}{m(X_{mean,S^{(l)}}, 0) + m(X_{modal,S^{(l)}}, 0)}, \quad (8)$$

where $m(X, 0)$ denotes the proximity measure between a function X and the constant null function 0. A large value of $m(X_{modal,S^{(l)}}, X_{mean,S^{(l)}})$ indicates that $X_{modal,S^{(l)}}$ and $X_{mean,S^{(l)}}$ have different behaviors according to $m(\cdot,\cdot)$. The larger $SHI_{mean}(S)$ is, the more heterogeneous the sample S will be. However, since the goal is to decide if the set S should be splitted into G classes $S_1, \ldots, S_G$ another index is required. The splitting will be accepted if the heterogeneity in each class is smaller than before splitting. To this end, the Partitioning Heterogeneity Index (PHI) is considered. It is defined as a weighted average of the SHI over classes

$$PHI_{mean}(S; S_1, \ldots, S_G) = \frac{1}{\text{Card}(S)}\sum_{g=1}^{G} \text{Card}(S_g) SHI_{mean}(S_g). \quad (9)$$

The larger PHI is, the more heterogeneous each class $S_1, \ldots, S_G$ is. Both $SHI_{mean}(S)$ and $PHI_{mean}(S; S_1, \ldots, S_G)$ are employed to define a score SC given by

$$\begin{aligned} SC &= SC_{mean}(S; S_1, \ldots, S_G) \quad (10)\\ &= \frac{SHI_{mean}(S) - PHI_{mean}(S; S_1, \ldots, S_G)}{SHI_{mean}(S)} \end{aligned}$$

A positive score SC indicates a gain of homogeneity inside classes. The splitting is accepted if SC is greater than a fixed threshold τ. For instance, $\tau = 5\%$ indicates that a splitting is accepted if it brings more than 5% of homogeneity within classes. If the score is negative, then S does not require this splitting. A value of τ that is too small indicates that the considered splitting is not required and the gain in terms of homogeneity is not significant. The value of τ is chosen according to the type of the data and the purpose of the classification. It is analogous to the choice of the first kind error in hypotheses testing. All the details concerning this methodology are given in Ferraty and Vieu (2006).

Aside from the above splitting criteria, it is required to define classes of S. Ferraty and Vieu (2006) proposed for independent data a procedure to establish the subgroups $S_1, \ldots, S_G$ as well as their number G. The procedure is related to the choice of the bandwidth parameters h and b. The choice of h is done by using the small ball probabilities. They play a key role in the theoretical properties of mode estimate (see Dabo-Niang *et al.* (2007)). A small ball probability is defined as $\mathbb{P}[X_i \in B(X,h)]$ for $X, X_i \in S$ which is the probability that a curve $X_i \in S$ belongs to the ball $B(X,h)$ with center X and radius h. For a given bandwidth h, one has at hand n probability points $\mathbb{P}[X_i \in B(X,h)], i = 1, \ldots, n$ for which the corresponding density $d_{S,h}$ can be estimated by a kernel estimate $\hat{d}_{S,h}$. The estimated density can be computed using, for instance, the package *np* (Hayfield *et al.*, 2008) of the R language (R Core Team, 2017). The number of groups G will be determined by the number of peaks of $\widehat{d}_{S,\widehat{h}_S}$. In practice, the bandwidth is selected using the entropy such that $\widehat{h}_S = \arg\min_h \int \widehat{d}_{S,h}(t) \log \widehat{d}_{S,h}(t)\, dt$.
The main algorithm of this classification approach is illustrated in Figure 6. The reader is referred to Ferraty and Vieu (2006), Dabo-Niang *et al.* (2007) and Dabo-Niang *et al.* (2010) for more details. A *R* (R Core Team, 2017) code, in the context of i.i.d data, is available at
https://www.math.univ-toulouse.fr/ ferraty/SOFTWARES/NPFDA/index.html.

0.4 Application

We illustrate the methodologies by using data of Ozone Concentration (OC) (units of measurement) collected in 106 monitoring stations of United States (see Figure 1) in 2015. The dataset is available at *https://www.epa.gov/outdoor-air-quality-data*.

Specifically for each one of the 106 stations we have data of OC recorded hourly from July 19 at 12 am to July 20 at 11 pm (Figure 2). We use linear interpolation to estimate some missing values. We denote the OC at time t, $t \in [1, 48]$ as $X(t)$. In order to apply the methodologies before described, were obtained OC functions (Figure 2) by expanding the discrete data (48 data at each station) in terms of 25 Fourier basis functions. The number of basis was chosen by using cross-validation.

0.4.1 Model-based clustering

We apply the EM algorithm for clustering the spatial functional data above described. A homocedastic model has been applied since it gives more relevant results from the classification point of view, and the value of q was selected during the EM algorithm by maximizing the BIC computed at each step for each possible value of q.

Under this setting, the BIC indicates that two or three clusters could be appropriated (Figure 3).

In Figure 4 we show the classification of the monitoring stations in two and three

Figure 1 Location of 106 monitoring stations (in the same number of cities) of ozone concentration in United States. Source: *https://www.epa.gov/outdoor-air-quality-data*

groups respectively. For the clustering in two clusters, $q = 18$ principal components have been retained. We see on the map that the obtained clustering well separates the East cities from the West cities. Moreover, we see on the curves that the clusters are also well separated from the curves point of view. In average we see on Figure 5 that West cities have higher pollution than Est cities. For the clustering in three clusters, $q = 17$ principal components have been retained. We see on the map that the obtained clustering still well separates the East curves from the West curves, but also the North from the South for the West side. When looking at these curves on Figure 5, we see that it is the six first hours that well separate cluster 1 (North) from cluster 3 (South).

As a conclusion of this application, for the clusterings (with two or three clusters) let observe that the method makes an trade-off between the geographical proximity and the common features of the curves, which allows to take into account spatial dependency while performing clustering. We see on the application that the obtained results are easily interpretable, and give a relevant spatial segmentation.

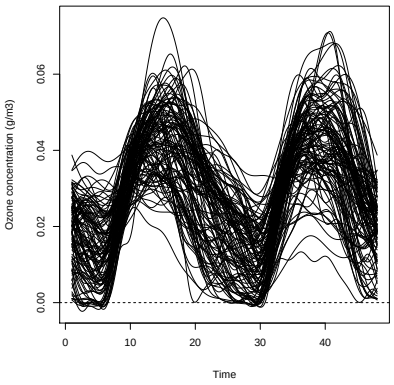


Figure 2 Ozone concentration curves (obtained after smoothing the data by using a Fourier basis) at 106 monitoring stations of the United States

0.4.2 Hierarchical classification

The previous algorithm has been used on our set of 106 curves with the threshold parameter (τ) fixed at 25%, and $L = 0$ for reducing the computational cost (i.e. HI is used instead of SHI). Let us denote by S the whole sample of curves, given in Figure 2.
Recall that the spatial weighted matrix W_n is constructed by taking k neighbors of each unit using KNN method.

At the first iteration, the number of neighbors is equal to $k = 4$ and a number $q = 8$ of eigenvalues used in the semi-metric $d(.,.)$ (different values have been taken, this last gives better results in term of homogeneity) and the data is splitted into two groups (CLASS 1 and CLASS 2 of respective sizes 5 and 101), since the corresponding concentration density $\hat{d}_S$ had two modes. This splitting is accepted since the gain of homogeneity (score SC) is larger (26%) than the threshold. Then, for the second iteration, the second (CLASS 2) group is splitted into two subgroups (CLASS 21, CLASS 22) with $k = 1$ and $q = 7$ and splitting score equal to 43%. At the third iteration, only CLASS 21 is splitted into two subgroups (CLASS 211 and CLASS 212) with $k = 1$ and $q = 6$ and splitting score equal to 55%. The procedure has been stopped with all the splitting scores smaller than 25% at the fourth iteration.

In Figure 7, the results of the different iterations of our procedure are presented. The sizes of the final groups CLASS 1, CLASS 22, CLASS 211 and CLASS 212 are respectively 5, 51, 23 and 27. The mean and mode curves are given in Figure 8.

$CLASS$ 1 (resp. CLASS 22) concerns essentially stations with higher main peak

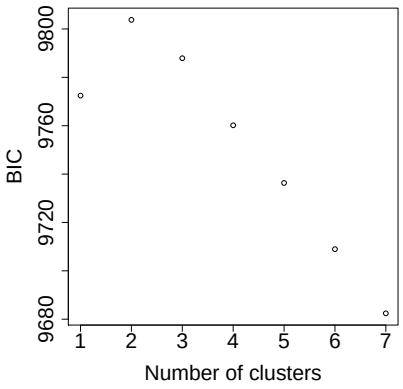


Figure 3 Value of BIC criterion according to the number of clusters

(around the middle of a day) ozone concentration (resp. smaller) than in other groups, particularly for the first day. The main difference between CLASS 211 and CLASS 212 comes also from the importance of the ozone concentration peak (smaller for the second group) and the width of their bases (larger for CLASS 211). It seems that at each iteration, the algorithm splits according to the maximum of the ozone concentration.

Regarding Figures 8 and 9, we may say that as in the first method, the clusters are separated from the curves point of view and geographical proximity. We see on the map of Figure 8 that there are mainly two groups of curves from the west (CLASS 1, CLASS 211 and CLASS 212) cities with higher pollution, while CLASS 22 (with the largest number of curves) is distributed in the west and east parts and has mainly smaller ozone concentration than the other groups. The CLASS 211 is mainly located in the west part.

0.5 Conclusion

The purpose of the present chapter is the classification of functional spatial curves using the functional framework. Two functional classification methods are considered, namely, descendant hierarchical classification based on modal curve, and the model-based clustering using a mixture model. These functional classification methods are presented and applied to ozone concentration data. We see on the application that the obtained results are interpretable, and give a relevant spatial segmentation.

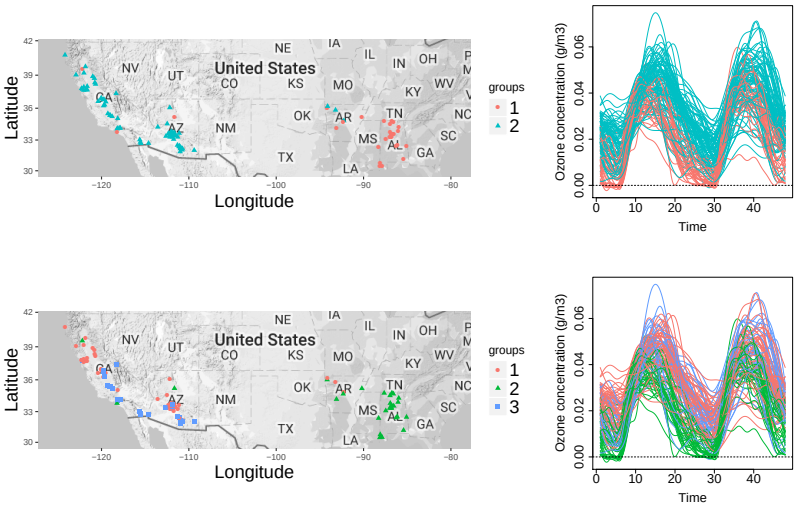


Figure 4 Locations of the stations coloured according to the cluster (left) and curves coloured according to the cluster (right) for two clusters (above) and three clusters (bellow)

Although this work is mainly practical, consistency results may be easily obtained, see the references therein. An advantage of spatial functional approaches is that they allow the clustering to take in account some spatial dependency. For the two different methods considered, the different classification results allow to identify two main ozone concentration area. The first located in the west is characterized by large peak, and the second is characterized by high volume, it is located in the east and west parts. Two separate clustering could be tried in the west and east parts of the considered region. This could be adapted in order to account for regional spatial variability. Future efforts can focus on adapting the distance measure used in the hierarchical method to the classification objective (risk analysis, etc.) and to the data record length and quality.

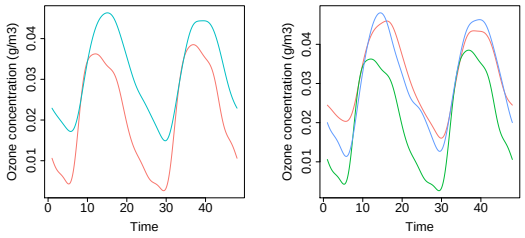


Figure 5 Average curves by cluster, respectively for two clusters (left) and three clusters (right)

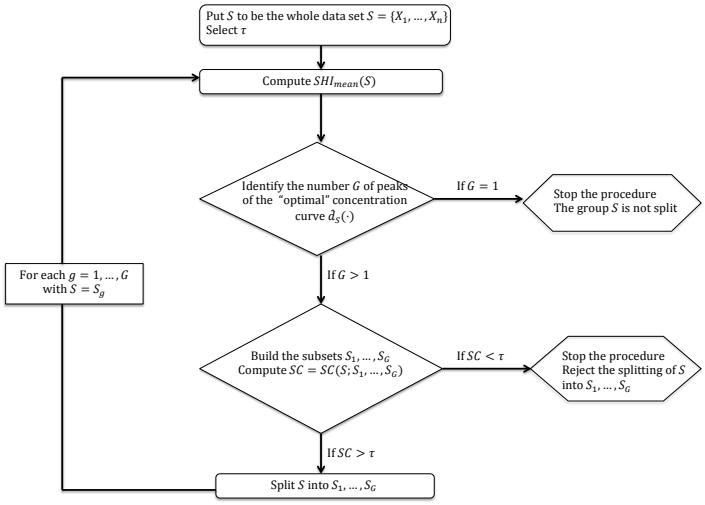


Figure 6 Algorithm of the descendant HC

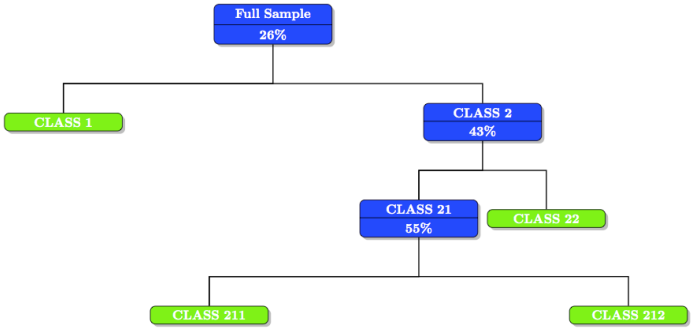
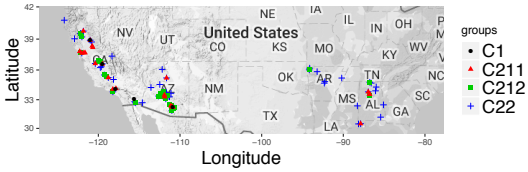
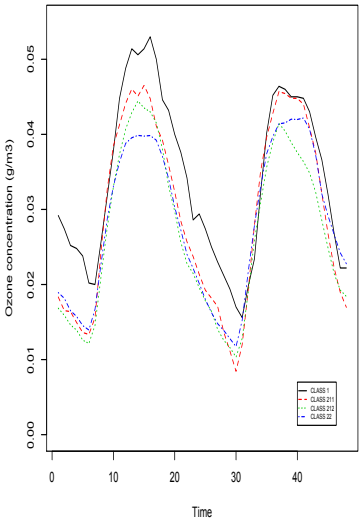


Figure 7 The classification results of the descendant HC





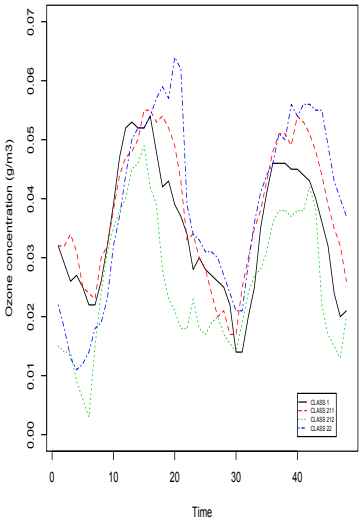


Figure 8 Locations of the stations colored according to the cluster (above), mean (left) and mode (right) curves colored according to the cluster (bellow)

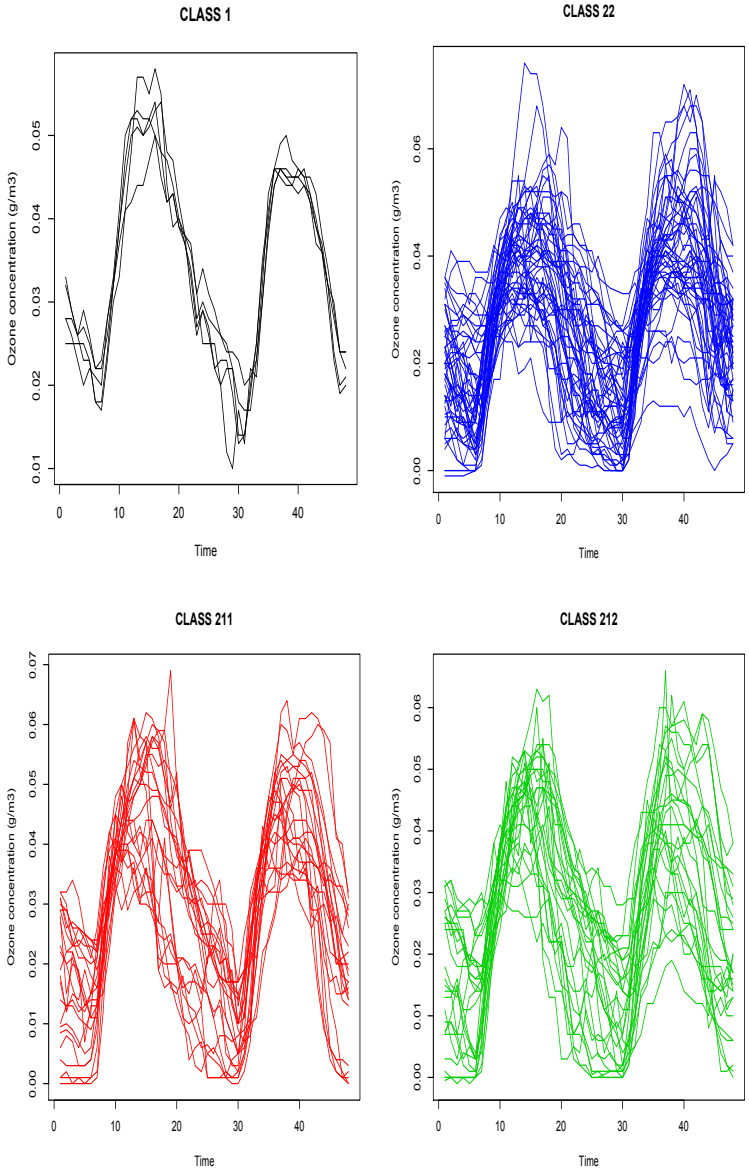


Figure 9 The curves of the different groups by the descendant HC.

References

Abraham, C., Biau, G., and Cadre, B. (2006) On the kernel rule for function classification. *Ann. Inst. Statist. Math.*, **58** (3), 619–633.

Abraham, C., Cornillon, P.A., Matzner-Løber, E., and Molinari, N. (2003) Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, **30** (3), 581–595.

Auder, B. and Fischer, A. (2012) Projection-based curve clustering. *J. Stat. Comput. Simul.*, **82** (8), 1145–1168.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., and Blanchet, G. (2015) Package ?spdep? *See ftp://garr.tucows.com/mirrors/CRAN/web/packages/spdep/spdep.pdf (accessed 9 December 2015)*.

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

Bouveyron, C. and Jacques, J. (2011) Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, **5** (4), 281–300.

Cadre, B. (2001) Convergent estimators for the l1-median of banach valued random variable. *Statistics: A Journal of Theoretical and Applied Statistics*, **35** (4), 509–521.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern recognition*, **28** (5), 781–793.

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

Chiou, J.M. and Li, P.L. (2007) Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69** (4), 679–699.

Cuevas, A., Febrero, M., and Fraiman, R. (2000) Estimating the number of clusters. *Canad. J. Statist.*, **28** (2), 367–382.

Cuevas, A., Febrero, M., and Fraiman, R. (2001) Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, **36** (4), 441–459.

Dabo-Niang, S., Ferraty, F., and Vieu, P. (2007) On the using of modal curves for radar waveforms classification. *Comput. Statist. Data Anal.*, **51** (10), 4878–4890.

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

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

Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

Floriello, D. and Vitelli, V. (2017) Sparse clustering of functional data. *Journal of Multivariate Analysis*, **154**, 1–18.

García-Escudero, L.A. and Gordaliza, A. (2005) A proposal for robust curve clustering. *Journal of Classification*, **22** (2), 185–201.

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

Giraldo, R., Delicado, P., and Mateu, J. (2012) Hierarchical clustering of spatially

correlated functional data. *Statistica Neerlandica*, **66** (4), 403–421.

Hayfield, T., Racine, J.S. *et al.* (2008) Nonparametric econometrics: The np package. *Journal of statistical software*, **27** (5), 1–32.

Jacques, J. and Preda, C. (2013) Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, **112**, 164–171.

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

Jacques, J. and Preda, C. (2014b) Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, **71**, 92–106.

James, G.M. and Sugar, C.A. (2003) Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98** (462), 397–408.

Klemelä, J. (2008) Density estimation with locally identically distributed data and with locally stationary data. *J. Time Ser. Anal.*, **29** (1), 125–141.

Krzanowski, W.J. and Lai, Y. (1988) A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pp. 23–34.

Milligan, G.W. and Cooper, M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50** (2), 159–179.

Pinkse, J. and Slade, M.E. (1998) Contracting in space: an application of spatial statistics to discrete-choice models. *J. Econometrics*, **85** (1), 125–154.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

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

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

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

Tarpey, T. and Kinateder, K.K.J. (2003) Clustering functional data. *J. Classification*, **20** (1), 93–114.