On the using of modal curves for radar waveforms classification

Sophie Dabo-Niang\textsuperscript{a,}\textsuperscript{*}, Frédéric Ferraty\textsuperscript{b}, Philippe Vieu\textsuperscript{b}

\textsuperscript{a}Université Charles De Gaulle, Lille 3, Maison de la Recherche, domaine du pont de bois, BP 60149, 59653 Villeneuve d’Ascq cedex, France
\textsuperscript{b}Université Paul Sabatier, Laboratoire de Statistique et Probabilités, UMR C5583 - 118 route de Narbonne, 31062 Toulouse Cedex 04

Available online 31 July 2006

Abstract

Recent advances in nonparametric functional data analysis allow to define the notion of mode for a sample of curves. A kernel-type estimator is proposed for estimating this modal curve. In addition, other centrality curves can be easily extended to the functional case, namely mean and median curves. A nonparametric unsupervised classification method which is also a cluster analysis is investigated. This is based on the comparison between the modal curve with another centrality curve for measuring some heterogeneity index. The main point is to show the good practical behaviour of this hierarchical classification on a sample of radar waveforms registered from the satellite Topex/Poseidon upon the Amazonian basin. In addition, a few theoretical advances are provided.

Keywords: Curves classification; Cluster analysis; Curves data; Functional statistics; Hierarchical method; Modal curve; Nonparametric statistics; Radar waveforms; Satellite data

1. Introduction

The using of satellite measurements is known to be of great interest for many purposes. This is particularly true for the hydrological setting that we wish to investigate here. Basically, the data which are recorded by a satellite are so finely discretized that it makes reasonable to consider them as curves data. So, these kinds of data are natural candidates to be treated by some of the recent statistical methods developed for functional data (see, for instance, Ramsay and Silverman, 1997; Ferraty and Vieu, 2003a, 2006 for large discussions). Here, we wish to deal with an unsupervised classification statistical problem involving curves. As far as we know, functional classification methods exist in the much simpler setting of supervised classification (known also as discrimination problem), corresponding to situations when the different groups of curves are identified before the beginning of the study (see Ferraty and Vieu, 2003a, 2006 for recent advances and references). Here, we will present a method for dealing with unsupervised curves classification, but let us before fix the ideas by presenting the satellite curves data set which motivated our approach.
Note this satellite curves data set has been previously studied in Dabo-Niang et al. (2004) and because our knowledge is progressing both on the nonparametric classification mechanism and on the data set itself, we are able to propose now a deeper study.

Topex/Poseidon is a satellite which emits a radar signal towards an impact point on the Earth and registers its echo. The original mission of the satellite Topex/Poseidon was the study of ocean level and its apparatus have been calibrated for processing echos from such kind of surface. This means that Topex/Poseidon is not calibrated

\* Corresponding author. Tel.: +33 3 20416510; fax: +33 3 20416171.
E-mail addresses: sophie.dabo@univ-lille3.fr (S. Dabo-Niang), ferraty@cict.fr (F. Ferraty), vieu@cict.fr (P. Vieu).

0167-9473/ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.07.012
for computing the altitude of any other kind of ground than seas or oceans. However, geophysicists would like to use also this satellite for a different goal, that is, for hydrological purposes. The idea is to use the waveform curves for deducing the altitude of the impact point, when the satellite is running upon areas of earth other than seas or oceans.

The data set which is studied in this paper consists of 472 waveforms obtained when the satellite is running upon the Amazonian basin, and more precisely around an area of 25 km upon the Amazon River. The satellite is registering 10 curves each second along a straight trajectory. Each observed curve is a waveform on the range $[0, 70]$. Because we know from experts on these radar-type curves that any value of the echo registration which is smaller than 50 can be considered as noise, we have subtracted 50 from all the observed curves (and then we set any negative value to be equal to 0). Fig. 1 contains a set of 20 randomly selected among these 472 curves which display a wide variety of waveforms. It can already be noted that these curves show different shapes regarding the number (and sharpness) of
peaks. Compare, for example, curves 1, 3, 4 and 21. More comments about these radar curves and their possible links between the various kinds of ground can be found in Frappart (2003).

We are interested in building some procedure for splitting automatically these 472 radar waveforms into different groups. After then, geophysicists expect to get a correspondence between the shape of the waveform and the kind of ground (the location of the satellite is known at each time), which would allow to achieve a preprocessing to obtain accurate altitudes. From a statistical point of view, this is an unsupervised curves classification problem because we do not have at hand any categorical response variable and the data sets are clearly of functional nature. This problem is also named cluster analysis and has been widely studied in the statistical literature for finite-dimensional data (see, for instance, Anderberg, 1973; Hartigan, 1975; Everitt, 1993 for a general overview and large discussion on this field).

In recent years, there is an increasing interest in the development of statistical methods for treating functional data. This field of statistics usually known as functional statistics covers different approaches. Two popular directions are including longitudinal data studies (see, for instance, Zimmerman and Nunez-Anton, 2001 for a recent survey) and linear models for functional data (see, for instance, Ramsay and Silverman, 1997, 2002; Bosq, 2000 for recent monographies). More recently, nonparametric methods for functional variables have emerged (see, Ferraty and Vieu, 2003a, 2006), and these advances on nonparametric functional data analysis have allowed in particular to develop new methods for supervised curves classification (see, Ferraty and Vieu, 2003b). The unsupervised curves classification problem that we treat here is much more difficult to deal with, because the groups are not known at the beginning of the study and there are just a few contributions available in the literature. These contributions are mainly restricted to the works by Abraham et al. (2003b); Tarpey and Kinateder (2003) and Cuesta-Albertos and Fraiman (2006) in which \( k \)-means techniques for cluster analysis are extended to curves data, and to the hierarchical algorithm proposed by Dabo-Niang et al. (2006) and also studied by Abraham et al. (2005). This hierarchical algorithm will be at the heart of our paper.

In finite-dimensional settings, nonparametric density estimates and/or modal considerations have been often used for cluster analysis purposes (see, for instance, Hartigan and Hartigan, 1985; Cuevas et al., 2000, 2001). On the other hand, recent advances on nonparametric estimation of the density function for random variables valued in abstract infinite-dimensional spaces have been performed, making possible the nonparametric estimation of some features of a distribution of random curves like density operator or modal curves and to develop hierarchical curves classification algorithms. Whereas these first works in this direction (see Dabo-Niang et al., 2006; Abraham et al., 2005) were essentially theoretical, the main aim of this contribution is to study this recent unsupervised hierarchical method for the radar waveforms classification problem described before. More precisely, an heterogeneity index based on the notion of modal curve is introduced in order to provide an efficient tool for classifying curves. Our paper is organized as follows. Section 2 describes the nonparametric functional methodology. As we will see, a heuristic definition of modal curve will play a major role for defining an heterogeneity index. Combined with small ball probabilities considerations (see Section 2.3), a hierarchical descending classification algorithm is presented in Section 3.2. The final classification of the radar waveforms and practical aspects is detailed in Section 3 which illustrates the good behaviour from a practical point of view. Finally, Section 4 develops a short theoretical discussion in order to highlight the theoretical pertinence of our heuristic modal curve.

2. The nonparametric functional methodology

2.1. Centrality notions for a sample of curves

The statistical modelling for treating curves data consists in looking at them as a sample of independent realizations \( \{X_1, \ldots, X_n\} \) of some functional variables distributed like \( X \) and taking values in some abstract infinite dimensional space \( (E, d) \), \( d \) being some measure of proximity between curves (for instance a metric or a semi-metric). In the radar data set discussed before, each \( X_i \) is a whole wave curve:

\[ \forall i \in \{1, \ldots n\}, \quad X_i = \{X_i(t), t \in (0, 70)\}. \]
The notion of mean curve is easily deduced from
\[ ∀t ∈ (0, 70), \quad X_{\text{mean}}(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t), \]
and the mean curve is defined by
\[ X_{\text{mean}} = \{X_{\text{mean}}(t), t ∈ (0, 70)\}. \]
Similarly, the notion of median curve can be extended to the functional framework (Avérous and Meste, 1997 for general definition and Cadre, 2001 for recent advances). Here, we will define the median curve to be a curve such that
\[ X_{\text{median}} = \min_{m ∈ \{X_1, ..., X_n\}} \sum_{i=1}^{n} d(X_i, m). \]
However, the extension of the notion of mode for a sample of curves is less trivial. From a heuristic point of view, one can see a modal curve as the one whose probability is “locally maximum”. A first naive way for translating mathematically this concept consists in defining the modal curve as follows:
\[ \tilde{X}_{\text{modal}} = \max_{m ∈ \{X_1, ..., X_n\}} \sum_{i=1}^{n} 1_{B(m, h)}(X_i), \]
where \( B(m, h) = \{\chi ∈ E, \quad d(m, \chi) < h\} \) is the ball of radius \( h \) and centered at \( m \). This means that among all the balls of radius \( h \), \( B(\tilde{X}_{\text{modal}}, h) \) is the one containing the largest number of curves. In other words, \( B(\tilde{X}_{\text{modal}}, h) \) defines an area of \( E \) where the sample of curves is the most dense. This is the reason why one can say that \( \tilde{X}_{\text{modal}} \) overperforms locally the set of curves, the local aspect coming from that it depends on \( h \) (the radius of the balls). Using an indicator function means that all the \( X_i \)'s lying to \( B(m, h) \) play the same role for computing the modal curve. Consequently, there is a strong border effect. In order to reduce this border effect, one can follow kernel smoothing ideas. To do that, it suffices to replace the indicator with an asymmetrical kernel function \( K \) (generally a positive decreasing function defined on the positive real line \((0, +∞)\)). So, from now on, a modal curve of a sample of curves \( S \) will be defined as
\[ X_{\text{modal}} = \max_{m ∈ \{X_1, ..., X_n\}} \sum_{i=1}^{n} K\left(\frac{d(m, X_i)}{h}\right). \]
In fact, \( K \) acts as a weight function: the larger is \( d(m, X_i) \) and the smaller is \( K\left(\frac{d(m, X_i)}{h}\right) \). We will not enter here on the discussion about the kernel function \( K \), but to fix the ideas, let us just tell that in the altimetric satellite application presented below in Section 3.3 we will use the following weight:
\[ K(u) = \frac{3}{2} \left(1 - u^2\right) 1_{(0,1)}(u). \]
As explained in Section 2.2, the modal curve will play a major role in our unsupervised classification procedure. The pertinence of this heuristic definition of a modal curve will be seen along Section 4 where it will be viewed as a consistent estimator of the curve which maximizes the density of the theoretical distribution drawing the sample of curves. Before that, let us extend the previous definitions to any sample of curves \( S \) in the following way:
\[ X_{\text{mean}, S}(.) = \frac{1}{\text{Card}(S)} \sum_{\chi ∈ S} \chi(.), \]
\[ X_{\text{median}, S} = \arg \min_{m ∈ S} \sum_{\chi ∈ S} d(m, \chi), \]
\[ X_{\text{modal}, S} = \arg \max_{m ∈ S} \sum_{\chi ∈ S} K\left(\frac{d(m, \chi)}{h}\right). \]
2.2. Modal curve and unsupervised classification

As in the finite-dimensional case, both the mean and the median curves are interesting only in the situation of homogeneous data, while the modal curve would be more useful for detecting any structural differences between data. This will be seen clearly later in Section 3.4 along the satellite data application. This is exactly the idea that we are going to use for our classification purpose, and we will look for some difference (in the sense of the measure \(d\)) between the modal curve \(X_{\text{modal},S}\) and one among \(X_{\text{mean},S}\) or \(X_{\text{median},S}\). Because of the horizontal shift in the curves (see deeper discussion in Section 3.1) it is much more reasonable for these waveforms-type data to use \(X_{\text{median},S}\) rather than \(X_{\text{mean},S}\) (but the same methodology could be used with \(X_{\text{mean},S}\) and could be efficient for other functional data sets). One way for measuring the heterogeneity of a sample \(S\) of curves consists in defining the heterogeneity index:

\[
HI(S) = \frac{d \left( X_{\text{modal},S}, X_{\text{median},S} \right)}{d \left( X_{\text{median},S}, 0 \right) + d \left( X_{\text{modal},S}, 0 \right)}.
\]

In order to get a more stable criterion, one can use \(L\) (randomly generated and not necessarily separated) subsamples \(S(l) \subset S\) (each being of the same fixed size \(M_S\)), and introduce the following subsampling heterogeneity index (\(SHI\)) for the set of curves \(S\):

\[
SHI(S) = \frac{1}{L} \sum_{l=1}^{L} \frac{d \left( X_{\text{modal},S(l)}, X_{\text{median},S(l)} \right)}{d \left( X_{\text{median},S(l)}, 0 \right) + d \left( X_{\text{modal},S(l)}, 0 \right)}.
\]

An unsupervised classification procedure should be able to answer both the following questions. How can we construct a possible splitting of \(S\) into \(G\) fixed subgroups \(S_1, \ldots, S_G\)? Is this splitting accurate or not? While the first one will be attacked later in Section 2.3, the indice \(SHI\) provides a nice way for replying the second question in some automatic way. This automatic procedure is based on the comparison of the heterogeneity indice \(SHI(S)\) obtained from the initial sample \(S\) with the one obtained from the split version of \(S\):

\[
SHI(S; S_1, \ldots, S_G) = \frac{1}{\text{Card}(S)} \sum_{k=1}^{G} \text{Card}(S_k)SHI(S_k).
\]

More precisely, we will use as a statistics to decide for splitting or not the following splitting score:

\[
SC = SC(S; S_1, \ldots, S_G) = \frac{SHI(S) - SHI(S; S_1, \ldots, S_G)}{SHI(S)}.
\]

We will do (respectively undo) the splitting of \(S\) into \(S_1, \ldots, S_G\) if \(SC\) is (respectively is not) greater than some fixed threshold \(\tau\).

It just remains now to propose a way for deciding what should be \(G\) and for building the \(G\) subgroups \(S_1, \ldots, S_G\) from \(S\). This the aim of the next section.

2.3. Small balls probabilities, bandwidth choice and building of subgroups

As usual in nonparametric setting, as well for functional as for finite-dimensional purposes, the choice of the smoothing factor is a crucial point to insure a good behaviour of the procedure. To choose the bandwidth parameter \(h\), the procedure has to take into account the objective of the study. Concretely here, that means that this parameter has to be chosen via some criterion which is adapted to the classification problem. Let \(S\) be a sample of curves. We propose to build our method by considering, for various values of \(h\), the small ball probabilities

\[
P(X \in B(x, h)),
\]

which play a key role in the theoretical properties of our mode estimate (see Dabo-Niang et al., 2006). We estimate these probabilities in a classical fashion by using

\[
\hat{p}_x(h) = \frac{1}{\text{Card}(S)} \sum_{\chi \in S} \mathbf{1}_{B(x, h)}(\chi),
\]
which gives us $\text{Card}(S)$ small ball concentration curves $\{\hat{p}_S(.)\}_{\chi \in S}$. For each fixed value of $h$, it is natural to consider the quantities $\{\hat{p}_{S,h}(\chi), \chi \in S\}$ as a set (of size $\text{Card}(S)$) of independent realizations of the same real random variable, whose density $\hat{d}_{S,h}$ can be estimated by mean of an usual one-dimensional smoothing technique. For instance, each of these density estimates $\hat{d}_{S,h}$ can be obtained by mean of the \textit{sm.density} function of the S-PLUS, 2005 library provided in \cite{BowmanAzzalini1997} or by the \textit{density} function if the R language (\cite{R2004}) is used.

We propose to compute the entropy which is known to be a useful index for measuring the homogeneity of each density estimate $\hat{d}_{S,h}$. So, because we look at this stage for the bandwidth which is the most accurate for discovering heterogeneity in the set of curves $S$, the bandwidth will be selected as follows:

$$\hat{h}_S = \arg \min_h \int \hat{d}_{S,h}(t) \log \hat{d}_{S,h}(t) \, dt.$$  

Finally, the small ball probability density estimate obtained with this selected bandwidth will be considered as the most accurate for detecting different structures in the set $S$. To reduce the notations, let us replace the notation $\hat{d}_{S,h}$ with $\hat{d}_{S,\hat{h}}$. At this stage we still do not know how to split the set of curves $S$, but we will use this “optimal” concentration curve to propose a division of the set $S$ into $G$ subsets. The number of groups will be determined by the number $G$ of peaks of $\hat{d}_S$. This means obviously that if $G = 1$ then we will not split the set $S$. Otherwise for $G > 2$, let $\hat{d}_S(p_{1,S}), \ldots, \hat{d}_S(p_{G-1,S})$ be the $G-1$ local minima of $\hat{d}_S$ in such a way that $p_{1,S} < \cdots < p_{G-1,S}$. The following splitting of $S$ can be considered:

$$S = \bigcup_{k=1}^{G} S_k \quad \text{where} \quad S_k = \{\chi \in S, \ p_{k-1,S} < \hat{p}_S(\hat{h}_S) \leq p_{k,S}\},$$

with $p_{0,S} = 0$ and $p_{G,S} = 1$.

It should be noted that $\hat{d}_{S,h}$ is not an estimate of the density of the variable $\hat{p}_S(h)$. Therefore, even if under standard assumptions on $h$ the random variable $\hat{p}_S(h)$ could be expected (for each fixed $\chi$) to have an asymptotic (and hence, unimodal) distribution, the estimate $\hat{d}_{S,h}$ can exhibit several modes (according to the value of $h$).

Recall that, in practice, the accuracy of this splitting should be checked (either by some splitting criterion or by a testing-type procedure as in \cite{Ferratyetal2006}). Here, we will consider the splitting score $SC$ defined in Section 2.2.

3. Classification of radar waveforms

3.1. Notion of proximity between waveforms

The only thing to be chosen now is the function $d$ that has to be used for measuring proximity between curves. This function plays a great role for ensuring good properties of the modal curve (and therefore of our classification procedure). The choice of $d$ can be driven from the practical problem which is investigated, that is from our altimetric context. First of all look again at Fig. 1, and note that there is clearly some horizontal shift between curves. Indeed, the shapes of curves numbered 21 and 22 are very similar, but they are affected by this horizontal shift. However, according to the experience of geophysicists, this kind of waveforms reflects the same water area (deep rivers); we would like to assign them to a same group. As a first consequence, taking the mean curve as a centrality curve does not make sense. This is why our heterogeneity index is based on the comparison between modal and median curves. As a second consequence, a good function of proximity between such curves has to be invariant under translation (because standard distances between curves 21 and 22 would produce large values whereas one expects small ones). For these reasons, we decided to use as a proximity measure between two radar curves $\chi$ and $\chi'$, the following function $d$:

$$d(\chi, \chi') = \inf_{z \in (-\varepsilon_0, +\varepsilon_0)} \frac{1}{b-a-2|z|} \int_{a+|z|}^{b-|z|} (\chi(t+z) - \chi'(t-z))^2 \, dt,$$

where $(a, b) = (0, 70)$ is the range of each wave curve, and where the rescaling parameter $\varepsilon_0$ has been chosen such that $\varepsilon_0 = (b-a)/5 = 14$. 

3.2. The algorithm for radar waveforms classification

The following unsupervised classification algorithm has been implemented in R and S-PLUS languages. The programs will be available on line through the *npda* package that will be go with the monography Ferraty and Vieu, 2006.

**Preliminary step**

→ Put $S$ to be the whole data set $S = \{X_1, \ldots, X_n\}$;
→ Compute for $\chi \in S$ and $\chi' \in S$ the proximity measures $d(\chi, \chi')$;
→ Compute for $\chi \in S$ and for a range of $h$, the concentration curves $\hat{p}_\chi(h)$.

**Step 1: Heterogeneity of $S$**

→ Compute directly $X_{\text{median}, S}$;
→ Compute $X_{\text{modal}, S}$ from the following steps:
  → Compute the density estimates $\hat{d}_{S,h}(\cdot)$ from the values of $\hat{p}_\chi(h)$;
  → Compute the optimal bandwidth $\hat{h}_S$ by mean of the entropy criterion;
  → Compute $X_{\text{modal}, S}$ by using the optimal bandwidth $\hat{h}_S$;
→ Compute the subsampling heterogeneity index $SHI(S)$.

**Step 2: Splitting of $S$**

→ Identify the location of the local minima $p_{1,S} < \cdots < p_{G-1,S}$ of $\hat{d}_{S,h_S}(\cdot)$;
→ If $G = 1$, Stop the procedure;
  If $G > 1$, Build the subsets $S_1, \ldots, S_G$;
→ Compute the splitting score $SC(S; S_1, \ldots, S_G)$.

**Step 3: Iteration of the procedure**

→ If $SC(S; S_1, \ldots, S_G) < \tau$:
  → Reject the splitting of $S$ into $S_1, \ldots, S_G$;
  → Stop the procedure.
Table 1

Sizes of the final groups

<table>
<thead>
<tr>
<th>GROUP 11</th>
<th>GROUP 12</th>
<th>GROUP 21</th>
<th>GROUP 22</th>
<th>GROUP 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>144</td>
<td>100</td>
<td>84</td>
<td>18</td>
<td>126</td>
</tr>
</tbody>
</table>

Fig. 3. Samples of waveforms for each group.

> If $SC(S; S_1, ..., S_G) \geq \tau$:
> - Split $S$ into $S_1, ..., S_G$;
> - For each $k = 1, ..., G$, go back to Step 1 with $S = S_k$.

3.3. The results

The previous algorithm has been used on our set of 472 radar-type curves with the threshold parameter ($\tau$) fixed at 5%, and $L = 0$ for reducing the computational cost (i.e. $HI$ is used instead of $SHI$). Let us denote by $S_0$ the whole sample of curves. In Fig. 2, the results of the different iterations of our procedure are presented. The sizes of the final groups are given in the Table 1.

At the first iteration the data set is split into two groups (GROUP 1 and GROUP 2), since the corresponding concentration density $\hat{d}_{S_0}$ had two modes. Because the SC in terms of homogeneity index is large enough (25%), this splitting is accepted. After the second iteration, GROUP 1 is split into GROUP 11 and GROUP 12, and this splitting is again accepted since the SC is around 13%. The GROUP 2 is split into three subgroups (that means that the corresponding density $\hat{d}_{GROUP 2}$ had three modes) with a splitting score close to 22%. At the third iteration the procedure has been stopped because all the splitting scores were smaller than the fixed threshold (5%).

Fig. 3 displays a sample of 16 waveforms for each group. GROUP 11 and GROUP 12 concern essentially “unimodal” waves (only one main peak), the difference between these two groups coming from the width of the peak base (small for the first group and large for the second one). GROUP 21, GROUP 22 and GROUP 23 contain “multimodal” curves. The main difference from one group to another one comes from the number of peaks and the width of their bases.

Why are these groups pertinent? In fact, geophysicists have additional important informations, namely the location of the area corresponding to each waveform. Thus, for each curve, the kind of ground is known (rivers, lakes, vegetation,
etc.). According to geophysicians, these groups of radar waves reflect various categories of ground (see Frappart, 2003); more the surface is reflective, more heavy is the peak (with high top). Therefore, deep rivers which act as good mirrors have a high reflective level. In the opposite, ground, rocks, etc. are not reflective surfaces and lead to low-level curves. In particular, “unimodal” waves with an heavy peak (for instance, curves 21 and 22 in Fig. 1) indicate a deep water area like large rivers. Waveforms having several peaks with larger base correspond to successive transitions between vegetation and water areas (for instance curves 44, 7, 132, 4, etc.).

Thus, one can say that a curve of GROUP 11 corresponds to a deep river whereas a member of GROUP 12 is in relation with a less deep river or a small lake (large width for the peak base). Concerning GROUP 21, it concerns essentially transition between ground and water areas. GROUP 23 displays globally waveforms coming from a mixing of water areas and vegetations. GROUP 22 is an intermediate set. Finally, one can summarize the classification procedure in the following way: the first splitting separates waveforms coming from either deep water areas or transitions between various land types. The second step splits the previous groups into more homogeneous ones.

### 3.4. What about homogeneity of the split groups?

In this section, we will highlight the interest of the notion of modal curve, both as a tool for characterizing a homogeneous set of curves and for detecting different functional structures in heterogeneous situations. In particular, we will see that the five final groups obtained by our algorithm are rather homogeneous. For each among the terminal leaves (groups) of our classification tree, the corresponding modal curve has been evaluated. The plot of modal curves are presented in Fig. 4. Each modal curve has to be compared with its corresponding group of curves. For each group, the corresponding modal curve gives clearly a good idea of centrality. In particular, the modal curves summarize very well the splitting procedure explained in the previous section. In the opposite, mean curves displayed in Fig. 5 crush the waveforms and do not reflect really their structure. To confirm that, we tried to perform the classification procedure with an heterogeneity index based on the comparison between modal curves and mean curves, and obviously none was performed.

To emphasize this point, Fig. 6 focuses on the first stage of our procedure and displays, for each group, the modal, the median and the mean curves. It is clear again that modal and median curves reflect similar structure whereas mean curves are not good centrality curves.
4. Theoretical aspects of the notion of mode in infinite-dimensional spaces

The radar curves can be viewed as independent observations of a functional variable which is valued in some functional (i.e. infinite-dimensional) space. The mode plays a crucial role in our classification procedure, and the aim of this section is to introduce the reader to a mathematical support for defining rigorously what is a mode in infinite-dimensional space and why the heuristic definition of the modal curve proposed in Section 2.1 is pertinent. A first approach for estimating modal curve can be found in Gasser et al. (1998) when the initial curves are replaced with a truncated basis expansion. Here, we discuss another approach which takes into account the whole functional feature and allows to see our heuristic modal curve as an empirical version of a consistent estimator of a theoretical mode. Note that, more generally, a functional variable is any random variable which takes its values in some infinite-dimensional space, and this covers the cases when the statistical unit is a surface, an image, etc. The theoretical background that we state below is presented in a general mathematical framework, in order to be useful to various kinds of functional data and not only for curves. Note that, despite obvious strong connexions with the beginning of our paper, this theoretical
section is self-contained, and its reading can be made independently of the waveforms classification problem studied before.

Let $X$ denote a generic functional random variable valued in the infinite-dimensional semi-metric space $(E, \, d)$. The key point is to assume that the probability distribution of the functional random variable $X$ admits some density function $f$ which is a non-linear operator defined on $E$ and valued in $(0, +\infty]$ with respect to some abstract measure $\mu$ on $E$. Note that the asymptotic results are given without specifying the measure of reference; we just require that $\mu$ is $\sigma$-finite, diffuse and is such that $0 < \mu(A) < \infty$, for any open ball $A \subset E$. Formally, a mode $\theta$ of the probability distribution of $X$ is a value that locally maximizes its density $f$. A simple way for defining the mode in the functional setting consists in introducing the mode as the global maximizer of $f$ over some fixed subset of $E$. To this end, we assume all along this section that there is some subset $C$ of $E$ such that $\theta$ is the unique solution of the problem

$$\sup_{\chi \in C} f(\chi).$$

In order to make computable the mode, $f$ has to be estimated. A nonparametric estimate using functional kernel smoothing techniques (see Ferraty and Vieu, 2006) is proposed. Let $S = \{X_1, \ldots, X_n\}$ be independent random variables which are identically distributed as $X$, given a kernel $K$ and a sequence of positive numbers $h_n$, a pseudo kernel density estimator of $f$ is defined as

$$f_n(\chi) = \frac{1}{n Q(K, h_n)} \sum_{i=1}^{n} K \left( \frac{d(\chi, X_i)}{h_n} \right),$$

where $Q(K, h_n)$ is a quantity which does not depend on $\chi$. Now, one can define an estimator of the functional mode by

$$\hat{\theta} = \arg \sup_{\chi \in C} f_n(\chi)$$

or by giving the following equivalent definition:

$$\hat{\theta} = \arg \sup_{\chi \in C} \sum_{i=1}^{n} K \left( \frac{d(\chi, X_i)}{h_n} \right),$$

since $Q(K, h_n)$ does not depend on $\chi$. Let us just remark now that it is the theoretical version of the definition of modal curve introduced in Section 2.1.

Now let us consider for all $\epsilon > 0$, the sets

$$\mathcal{A}_\epsilon = \{\chi \in C, \, f(\theta) - f(\chi) < \epsilon\}.$$

Following the ideas described in finite-dimensional setting by Abraham et al. (2003a), the flatness of the density function $f$ around its mode $\theta$ can be controlled by mean of the diameter of these sets:

$$T(\epsilon) = \sup_{\chi \in \mathcal{A}_\epsilon} \sup_{\chi' \in \mathcal{A}_\epsilon} d(\chi, \chi').$$

Under some additional assumptions acting on the small ball probabilities for the functional variable $X$, on $f$ and on its pseudo-estimator $f_n$, the following almost complete convergence properties for $f_n$ and $\hat{\theta}$ have been stated in Ferraty and Vieu (2006):

$$\lim_{n \to 0} \sup_{\chi \in C} |f_n(\chi) - f(\chi)| = 0 \quad \text{a.co.}$$

and

$$\lim_{n \to 0} d(\hat{\theta}, \theta) = 0 \quad \text{a.co.}$$

The goal of the next theorem is to precise the rates of convergence.
Theorem 1. If \( \sup_{\chi \in \mathcal{C}} |f_n(\chi) - f(\chi)| = O(v_n) \), a.c.o., where \( v_n \to 0 \), and if \( \lim_{\varepsilon \to 0} T(\varepsilon) = 0 \), then

\[
d(\hat{\theta}_n, \theta) = O(T(v_n)) \quad \text{a.c.o.}
\]

Proof. We deduce from the condition \( \lim_{\varepsilon \to 0} T(\varepsilon) = 0 \) that

\[
\forall \varepsilon > 0, \ \exists \eta > 0, \ \forall \chi \in \mathcal{C},\ d(\hat{\theta}, \chi) > \varepsilon \Rightarrow |f(\hat{\theta}) - f(\chi)| > \eta.
\]

The result holds as soon as one takes \( \chi = \hat{\theta} \) and \( \varepsilon = T(v_n) \). □

Remark 1. A technical step consists in stating the rate of convergence \( v_n \) for the kernel density estimate \( f_n \). We have decided to present our result in such a very synthetic way for avoiding to present this long technical step. This rate of convergence could be reached by imposing more restrictive conditions such as a Lipschitz condition on the density \( f \) (for more details and a deep discussion, see Ferraty and Vieu, 2006). When \( d \) is a metric, one can request a differentiability condition for the density \( f \) which is detailed in Dabo-Niang et al. (2006). Finally, as explained in detail in Ferraty and Vieu (2006), the rate \( v_n \) is strongly linked with the semi-metric \( d \) through the notion of small ball probabilities.

5. Concluding remarks

The results of this study show that the proposed hierarchical classification methodology for a sample of curves works well on the waveforms. The algorithm produces automatically homogeneous subgroups that can be easily interpreted in terms of differences of grounds (and in the Amazonian basin in terms of river, lake, vegetation, etc.). This method is really a functional way for classifying curves because the functional feature of the data set is taken into account. Firstly, a functional proximity is introduced for comparing curves and secondly, functional centrality curves as modal, median and mean have been used. In addition, recent kernel smoothing ideas for functional data have been adapted to this classification setting for defining what is a modal curve.

As a by-product, this work highlights the practical interest of the notion of “modal curve” whose definition makes sense with the recent theoretical advances on nonparametric functional data analysis discussed in Section 4. This is a good example of the interaction between practical and theoretical developments, which is the spirit of the nonparametric functional data analysis. Finally, a website going with the monograph Ferraty and Vieu (2006) is available (http://www.lsp.ups-tlse.fr/staph/npfda) where the proposed classification procedure (implemented with R or S-PLUS) will be online (with many other nonparametric functional data analysis methodologies).

Acknowledgements

The authors would like to thank all the participants in the working group STAPH on functional and operatorial statistics in Toulouse, for their continuous comments and encouragements, as well as Frédéric Frappard from the LEGOS-CNES who kindly provided us the radar data set. The pertinent comments of two referees have also been particularly useful for improving the quality of this paper.

References


