Detecting clusters in spatially correlated waveforms

Francesca Di Salvo* Renata Rotondi**
Giovanni Lanzano***

*University of Palermo, **CNR, Milan, ***INGV, Milan

November 2017

Introduction

- goal:
  - Investigating the effectiveness of some approaches relied on depth measures in constructing basic tools for clustering of waveforms.
  - Combining clustering of waveforms with clustering of metadata.
- motivation:
  - the analysis of waveforms
  - Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. National grant MIUR, PRIN-2015 program, Prot.20157PRZC4
- basic points:
  - Working on collections of seismic data, dealing with high-dimensionality.
  - Avoiding strict parametric assumptions, clustering and aligning functional data

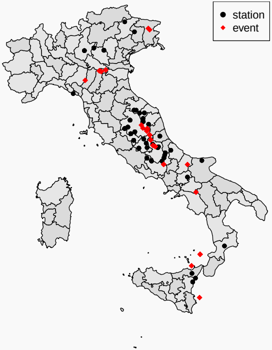
Selected collection of seismic data

- Source: Engineering Strong Motion database (http://esm.mi.ingv.it/)
- Engineering Strong Motion database, ESM allows users to query earthquake and station information and download waveforms for events (M>4.0) recorded in the European-Mediterranean and the middle-East regions. ESM is fully compatible with the European Integrated Data Archive (EIDA).
- A sample of 21 Italy earthquakes with magnitudo > 5.5.
- Recordings refer to a set of 41 station of class $EC8 - A$. The distances from epicenter are in $50 - 100 Km$
- For each recording, waveform data and some related metadata are considered.

Selected collection of seismic data

- Source: Engineering Strong Motion database (`http://esm.mi.ingv.it/`)

Figure: Geographic Coordinates of the events and stations



Selected collection of seismic data

- A sample of 21 Italy earthquakes from 1976 to 2017. Recordings refer to a set of 41 stations.

Table: Number of recordings for 4 main events

Event	Latitude	Longitude	Recordings
EMSC-20161030-0000029	42.8322	13.1107	8
EMSC-20161026-0000095	42.9087	13.1288	10
EMSC-20170118-0000034	42.5293	13.2823	12
EMSC-20160824-0000006	42.6983	13.2335	14

- For the other events, from 1 to 4 recordings are in the sample.

Selected collection of seismic data

Data collected can be arranged in:

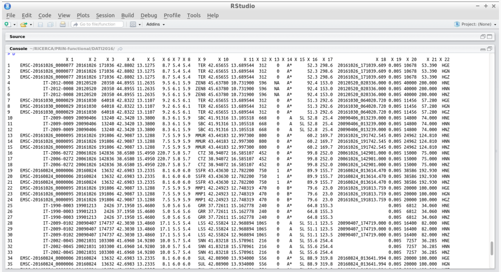
Event		
TIME	LATITUDE DEGREE	LONGITUDE DEGREE
EVENT DEPTH Km	MAGNITUDE W	

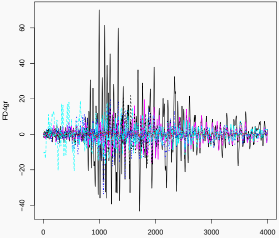
Station (EC8-A)		
LATITUDE DEGREE	LONGITUDE DEGREE	ELEVATION m
SITE CLASS	MORPHOLOGIC CLASS	

Waves		
DIMENSION (E-N-Z)	PGA cms^2	TIME PGA
DURATION	FREQUENCY HZ	ACCELERATION

Framework of the Data

Figure: Metadata for multivariate statistical analysis (right) and functional data for waveforms analysis (left)





A vector of 46 data is available for each recordings. Seismograms record in three cartesian axes (x, y, and z), representing the horizontal directions (E and N) and vertical direction Z.

The methodology

The proposed approach links different methodologies so as to combine information from metadata with waveform data.

Steps:

1. A hierarchical clustering is applied to obtain homogeneous clusters of recordings (**Multivariate Statistical Tecnique**)
2. A waveform analysis is implemented inside the clusters, aiming to the **characterization of the seismic waves**.
    - This second step, is handled in a **functional data** setting. The functional nature of the data are exploited in order to highlight the temporal dynamics of the signals.
    - The key contribution is to detect clusters of similar waveforms by mean of **Depth measures**.
    - A crucial point is represented by the **alignment of waves** with different lengths.

Selected collection of seismic data

Table: Summary of Metadata

	range of variability
EVENT LATITUDE DEGREE	37.195 - 46.300
EVENT LONGITUDE DEGREE	10.345 - 15.495
EVENT DEPTH KM	4.300 - 220.700
MAGNITUDE W	5.400 - 6.900
EPICENTRAL DISTANCE KM	51.100 - 99.800
EARTHQUAKE BACKAZIMUTH DEGREE	0.800 - 357.900
PGA $cm.s^2$	-57.109 - 102.517
TIME PGA s	0.825 - 59.820
DURATION s	12.125 - 230.015

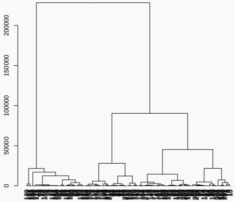
Selected collection of seismic data

Table: Variables used in hierarchical clustering

	range of variability
MAGNITUDE W	5.400 - 6.900
EPICENTRAL DISTANCE KM	51.100 - 99.800
EARTHQUAKE BACKAZIMUTH DEGREE	0.800 - 357.900
PGA $cm.s^2$	-57.109 - 102.517
TIME PGA s	0.825 - 59.820
DURATION s	12.125 - 230.015

Agglomerative hierarchical clustering

1. Method: Minimization of total within-cluster variance (WARD)



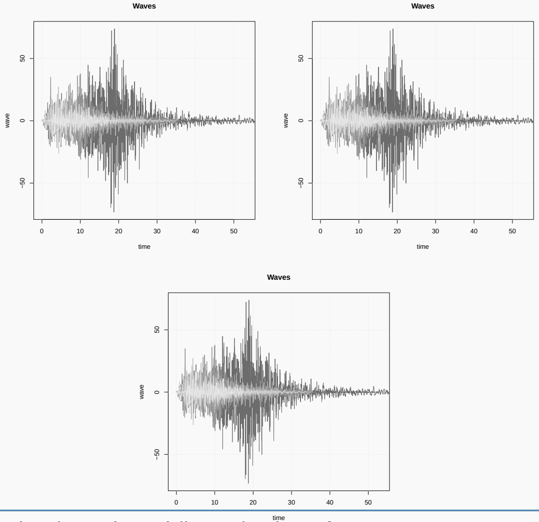
2. Choice of the number of clusters:
   Averaging the distances between each cluster and its centroid,
   K^* is the value that maximize over $K = 2, \ldots, N$:

$$max_K \Delta(K) = \frac{(N-K)}{K} \sum_k \left(\frac{d^{M_k M}}{\sum_{k=1}^{K} d^{x_{ik} M_k}} \right)$$

3. K^* =3 clusters are identified

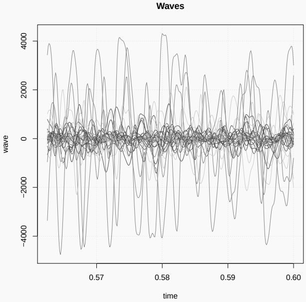
Waves as functional data - FDA, Ramsay et al.(2005)

Figure: E-component of the waves clustered in 3 initial groups



Functional data show peaks and other features at different time points. The underlying variability can be ascribed to two sources:

1. Amplitude (variability along $y - axis$)
2. Phase (variability along $x - axis$)

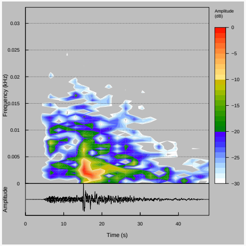


- Alignment Procedure:
  Short Time Fourier Transform (Shumway, 2003) and *elastic shape analysis of functional data* (Tucker et al.,2013).

Short Time Fourier Transform

Selection of time intervals $->STFT(y^p{}_i(t)) = Z^p{}_i(\omega;t)$

Figure: STFT performed on the E-component of a signal



1. Time intervals are splitted into frames and the variability of the time-frequency content of the waveforms is computed. The partition with the minimum number of frames retaining at least an α% of the whole variabilility is selected.
2. This allows to cut the signals obtainig informative sequences of the same length.

Elastic Shape Analysis , (Tucker et al.,2013)

- **Warping functions** are tranformations of time $\gamma(t)$:

$$\Gamma = \{\gamma : [0, T] \rightarrow [0, T] \mid \gamma(0) = 0,\ \gamma(T) = T\}$$

- Let define the *Square Root Slope Functions* SRSFs:

$$q(t) = sign\ f(t)\sqrt{\left|\frac{df(t)}{dt}\right|}$$

- for any f_1, f_2 , let define the distance D_γ :

$$D_\gamma(f_1, f_2) = inf_{\gamma\in\Gamma}||q_1 - (q_2 \circ \gamma)||$$

- The optimal warping function γ is the solution of the minimization of $D_\gamma(f_1, f_2)$ over Γ. (by dynamic programming algorythm)
- The **warped (ALIGNED) functions** are the compositions:

$$f \circ \gamma : [0, T] \rightarrow R$$

Choice of a depth

- Robust nonparametric tools, based on the concept of data depth can be applied for clustering purposes in the functional data setting
- the underlying idea is to determine the clusters providing an order within a sample of curves.
- Several depth notions generalizes unidimensional concepts of robust statistics to multivariate data
- Not all the depths are able to be generalized to functional data, due to the high dimensionality
- We focus on Modified Band Depth (MBD, López-Pintado and Romo, 2009)

Band Depth

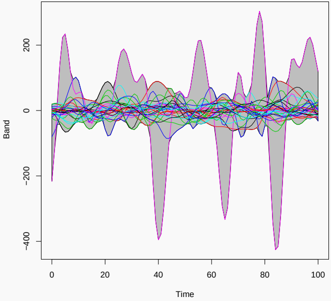
Basic concepts:

1. The *Band* in R^2 delimited by a set of n observed curves:

$$(y_1(t_{1_i}), \ldots, y_n(t_{n_i})) \tag{1}$$

with $t_{j_i} \in [0,1]$ is defined as:

$$B(y,t) = (t,y) : min_{j=1,\ldots,n}(y_j(t_{j_i})) \leq y(t) \leq max_{j=1,\ldots,n}\left(y_j(t_{1_j})\right) \tag{2}$$



Band Depth and Modified Band Depth

For any of the n observed waves,

1. The *Band Depth* is defined as:

$$BD_n(y) = \sum_{j=1}^{J} \binom{n}{j}^{-1} \sum_{1<\ldots i_k \ldots<n} I\left(G(x) \in B(y_{i1}, \ldots, y_{ij})\right) \quad (3)$$

   BD_n is the proportion of bands, made up of $2, 3, \ldots, J$ curves containing the graph of y. (Lopez-Pintado & Romo, 2009)

2. The *Modified Band Depth*, given λ, a Lebesgue measure in $[0,1]$ is:

$$MBD_n(y) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1<\ldots i_k \ldots<n} \frac{\lambda(A(x; x_{i1} \ldots x_{ij}))}{\lambda(T)} \quad (4)$$

   MBD_n is the portion of time that $y(t)$ is in the bands, made up of $2, 3, \ldots, J$ curves containing the graph of y.

Clustering Modified Band Depth

- The *center* $\rightarrow$ *outward* ordering provided by MBD_n is exploited in the proposed clustering procedure:
- selecting $\alpha : 0 \leq \alpha \leq 1$ and considering a partition of n curves in K clusters, the $\alpha-$ trimmed median function and only the $(1-\alpha)100\%$ of deepest curves are retained in the cluster (kernel).
- each of the $100\alpha\%$ most external curves is allocated to the cluster w.r.t. its MBD is highest.
- the kernels of the clusters are computed again, after the memership is changed, and the new $100\alpha\%$ of most external curves is assigned to on of the clusters, maximizing the MBD.
- After some steps the clusters achieve the optimal configuration
- The area of the Kernels of the clusters give a measure of their cohesiveness.

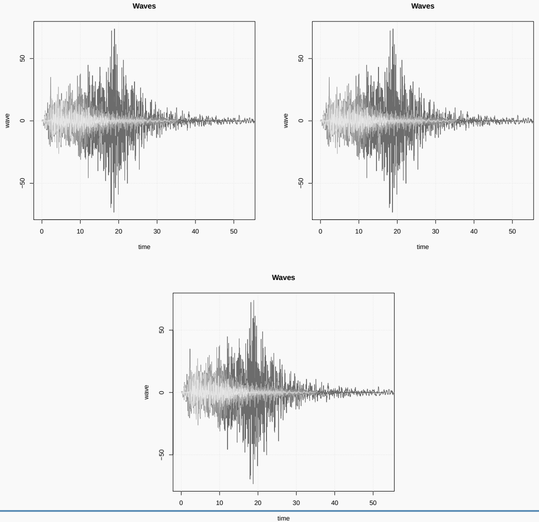
Clustering Algorithm

Steps of the Algorithm

1. Determining clusters on the space of the Metadata.
2. Aligning waves inside the clusters:
   2.1 STFT
   2.2 ESA
3. Depth-based clustering of waveforms:
   3.1 compute the depth (MBD) for the initial partition
   3.2 find the kernel made up by the $\alpha\%$ of deepest curves
   3.3 ri-allocate the $(1-\alpha)\%$ of most external curves in the cluster w. r. t. the MBD is highest.
   3.4 repeat steps $3.2-3.3$ until the allocation of the curves improves in terms of increasing MBD.
   3.5 stop when all the external curves have the highest MBD with its cluster

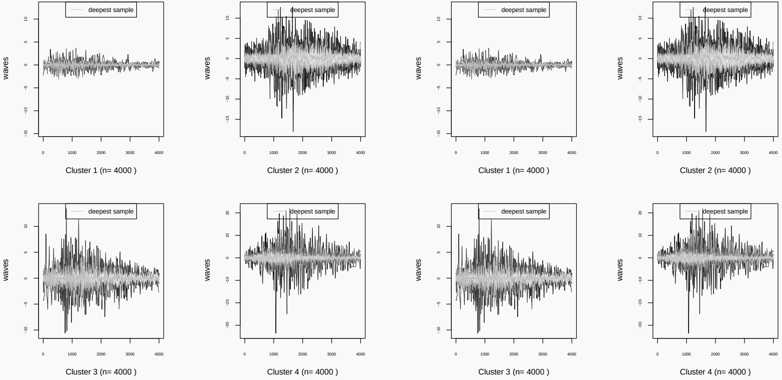
Initial clusters

Figure: E-component of the waves clustered in 3 initial clusters



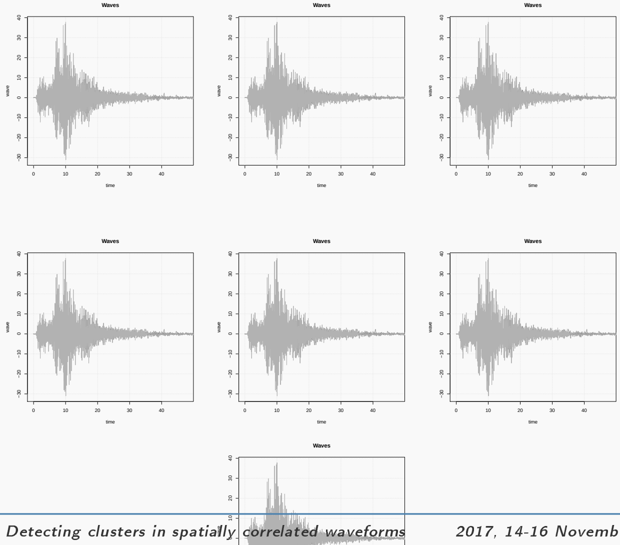
Clusters for STFT and ESA results

Figure: final clusters



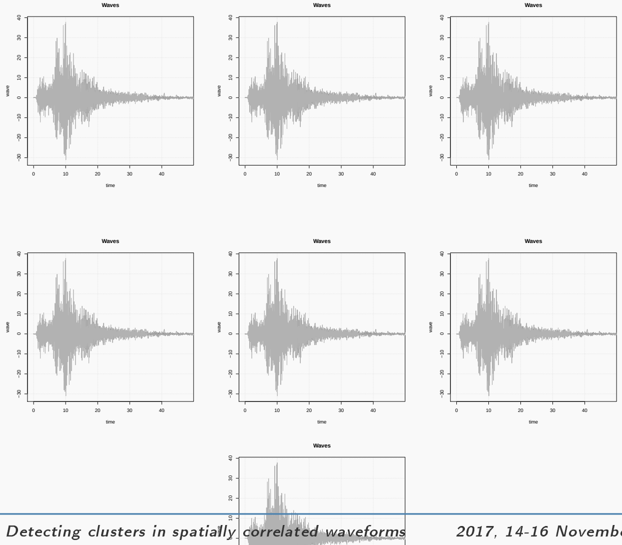
Clusters based on MBD

Figure: E-component of the waves clustered in 7 final clusters



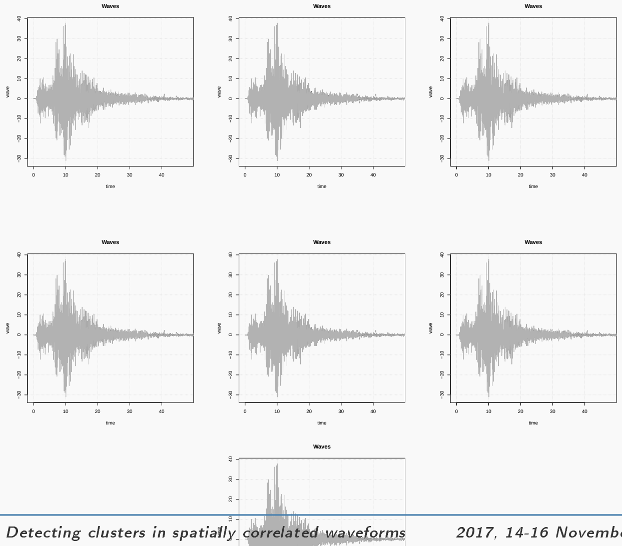
Clusters based on MBD

Figure: N-component of the waves clustered in 7 final clusters

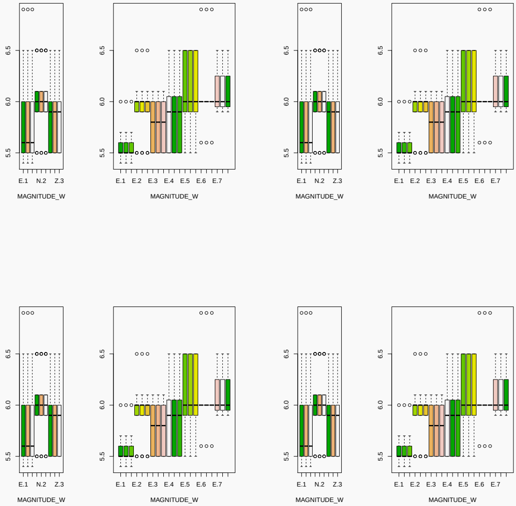


Clusters based on MBD

Figure: Z-component of the waves clustered in 7 final clusters



Metadata in final clusters

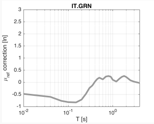


Residuals from attenuation models

$$\mu_{es} = \mu_{es_ref} + S_k + \delta S2S_s + \delta P2P_{sr} + \delta L2L_r$$

where $k = 1, 2, \ldots, 7$
(Al Atik et al.2010 ; Lanzano et al. 2017)

Figure: Residuals from cluster 1: Duration 36.28 -57.37 s; $PGA1 - 3cm/s^2$

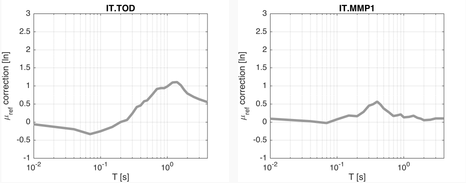


Residuals from attenuation models

$$\mu_{es} = \mu_{es_r ef} + S_k + \delta S2S_s + \delta P2P_{sr} + \delta L2L_r$$

where $k = 1, 2, \ldots, 7$
(Al Atik et al.2010 ; Lanzano et al. 2017)

Figure: Residuals from cluster 2: Duration 160 - 230 s; $PGA 4 - 10 cm/s^2$

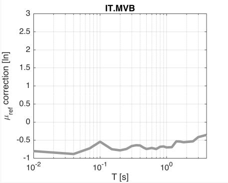


Residuals from attenuation models

$$\mu_{es} = \mu_{es_ref} + S_k + \delta S2S_s + \delta P2P_{sr} + \delta L2L_r$$

where $k = 1, 2, \ldots, 7$
(Al Atik et al.2010 ; Lanzano et al. 2017)

Figure: Residuals from cluster 3: Duration 65.72 - 91-23 s; $PGA2 - 10cm/s^2$

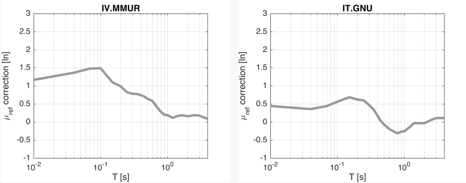


Residuals from attenuation models

$$\mu_{es} = \mu_{es_r ef} + S_k + \delta S2S_s + \delta P2P_{sr} + \delta L2L_r$$

where $k = 1, 2, \ldots, 7$
(Al Atik et al.2010 ; Lanzano et al. 2017)

Figure: Residuals from cluster 4: Duration 97 - 140 s; $PGA5 - 30cm/s^2$



Analysis of the event Accumoli, $24-08-2016$

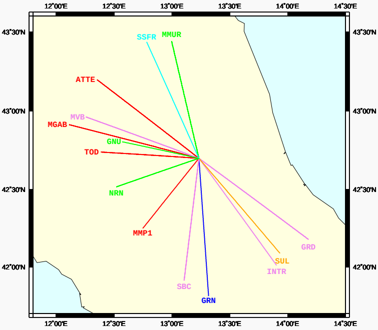
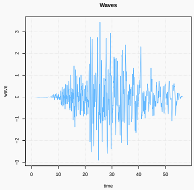
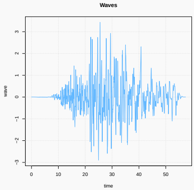
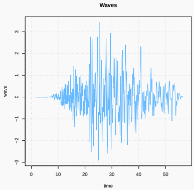


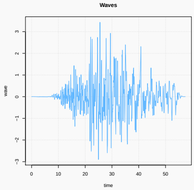
Figure: Map of the signals

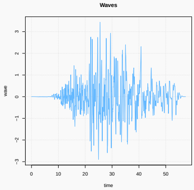
Analysis of the event: Accumoli, 24-08-2016

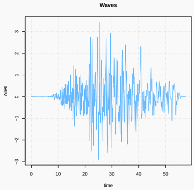












Adelfio, G., Chiodi, M., D'Alessandro, A. and Luzio, D., D'Anna, G., Mangano, G. (2012) Simultaneous seismic wave clustering and registration. Computers & Geosciences, 8(44), 60–69.

Antoniadis, A., Paparoditis, E. and Sapatinas, T. (2006) A functional waveletkernel approach for time series prediction. Journal Royal Statisticial Society Series B Statistical Methodology, 68(5):837

Bindi, D., Castro, R. R., Franceschina, G., Luzi, L., Pacor, F. (2004). The 1997 - 1998 Umbria - Marche sequence (central Italy): Source, Path and Site effects estimated from strong motion data recorded in the epicentral area. J. Geophys. Res., 109, B04312, doi:10.1029 2003JB002857.

Di Salvo F., Adelfio G., Sottile G. (2017) Depth-based methods for clustering of functional data TIES 2017 Conference, Bergamo.

Everitt, B. (1993) Cluster analysis. Wiley, New York

Garcìa-Escudero, L. A. and Gordaliza, A. (2005). A proposal for robust curve clustering, Journal of classification, 22, 185-201.

- Giraldo, R., Delicado, P., Comas, C., Mateu, J.(2011) Hierarchical clustering of spatially correlated functional data. Stat. Neerl. 66, 403–421.
- Jacques J. and Preda C. (2014). Functional data clustering: a survey. Advances in Data Analysis and Classification, Springer Verlag, 2014, 8 (3) 231-255
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Ramsay, J.O., Li, X. (1998). Curve registration. Journal of the Royal Statistical Society, Section B 60, 351–363.
- Ramsay, J. O. and Silverman, B. W. (2005). Functional Data Analysis. Springer, New York.
- Romano E., Mateu J. Giraldo R. (2015) On the performance of two clustering methods for spatial functional data. Advances in Data Analysis and Classification, Springer Verlag, 467-492