DETECTING CLUSTERS IN SPATIALLY CORRELATED WAVEFORMS

F. Di Salvo¹, R. Rotondi², G. Lanzano³

¹ Dipartimento di Scienze Agrarie Alimentari e Forestali, Università di Palermo, Italy
² Istituto di Matematica Applicata e Tecnologie Informatiche Enrico Magenes, CNR, Milano, Italy
³ Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Italy

Introduction. Seismic networks often record signals characterized by similar shapes that can also be considered according to their geographic positions.

Many techniques have been proposed for the analysis of similarity between seismic waves and in our proposal the functional nature of the data are exploited in order to highlight the temporal dynamics of the signals and investigate their spatial dependence.

In the situation under consideration, the observed data $y_s(t)$ can be assumed as a realization of a spatio-temporal process, where $s$ is the geographic position and $t$ indicates the observed time in a continuous interval. An overview of statistical methods for analyzing functional data is shown in Ramsay and Silverman (2005).

The data $y_s(t)$ are functional in the time dimension and can also be multidimensional: in particular we consider three components of the signal, recorded in three different directions, N-S, E-W and vertical.

Therefore, we assume multidimensional curves, as functions of time, $y_s(t)$, with $d=3$, recorded, as a set of discrete measurements of spatially interdependent curves.

The analysis of functional data provides new perspectives for deriving models and tools for the analysis of high-dimensional data in presence of spatio-temporal structures. Among the proposed methods, functional clustering has been adapted to the case of geographically referenced samples in order to delineate relatively contiguous zones with similar facies. Clustering of spatially correlated curves is a recent field of research; more details are given in Giraldo et al. (2010).

Here the key contribution is to incorporate a spatial structure in clusters of waveforms relying on depth measures. A crucial point is represented by the alignment of the curves: functional data not perfectly aligned show peaks and other features at different locations and as a consequence, any pointwise synthesis of curve values become meaningless. A preprocessing step consists in aligning observed curves in order to discard nuisance effects. An application of the proposal is given in the section 3, where the methodology is applied to the set of recordings of the first mainshock of 2012 Emilia sequence (20 May; Mw 6.1). The analysis is performed by using the R software package (R Development Core Team, http://www.R-project.org)

The methodology. In applied sciences there is an increasing interest for modeling correlated functional data: it is the case when samples of functions are observed over an interval at discrete time points (temporally correlated functional data) and when these functions are observed in different sites of a region (spatially correlated functional data).

The considered methodology is based on waveform clustering dividing the spatial domain into clusters and extracting information from the shapes of the underlying functions. In a sample data the given functions may not be aligned and the mechanism for alignment is an important topic of the analysis. Warping (or aligning) procedures have also great importance in the waveform clustering process, because they reduce a negligible component of amplitude variability, identify nuisance effects in phase variation, that, if ignored, may result in a possible loss of information; the immediate consequence is that the underlying pattern could not be retained.

The adopted warping procedure is based on a technique called elastic shape analysis of curves, proposed by Tucker et al. (2013); the technique aims to separate the phase (x-axis) and amplitude (y-axis) of the functional data.

An important advantage of this procedure is a relative reasonable computational time, that is a crucial point when data come from dense seismic networks with three components broadband sensors.
Hierarchical clustering methods are recently extended to both geographically referenced data and functional data in order to handle with the problem of classifying spatially correlated curves in groups of curves which are spatially homogeneous.

In the more traditional setting, the spatial structure is taken into account by mean of the variogram, that is the variance of the increments of the observed data, defined as function of the geographic distance; an efficient way to weight dissimilarities between functions is the use of the trace-variogram (Giraldo et al., 2007) that is the variogram of the coefficients accounting for the temporal dynamics of the observed data.

Waveforms clustering, based on cross-correlation measures between signals, may presents some limitations (see Adelfio et al., 2012), so we refer to more recent contributes relating data-depth based clustering analysis.

Statistical data depth is a robust technique providing an alternative way to find the “center” of multivariate data sets and is robust for clustering. Statistical depth functions provide an ordering of all points from the center outward in a multivariate setting, extending the concept of linear order induced in one-dimensional observations. But with the complication that for dimension greater than 2, there is no natural order.
The statistical depth identifies the deepest function (a sort of multivariate median function) and induces a center-out-ward ordering. Several depth notions generalize unidimensional robust statistics to multivariate data but 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). The choice of MBD in our clustering of waves is motivated by the following reasons: due to high oscillations of the curves, many curves leave their cluster for short intervals and the use of other depth measures does not allow a correct classification. Moreover when there is not a characterization of the clusters in terms of a representative shape, other alternative measures are not able to recognize the different clusters (see López-Pintado and Romo, 2009).

Application. The data consist of a set of recordings of the first mainshock of 2012 Emilia sequence (20 May; Mw 6.1). The waveforms are relative to the seismic network RAN (Rete Accelerometrica Nazionale) managed by the Italian Department of Civil Protection, are obtained from ESM/ITACA database (esm.mi.ing.it; itaca.mi.ingv.it; Luzi et al., 2016).

The signals were recorded by the stations, whose distances from the epicenter are in the range R=(16.1, 343.7) km. The goal is to divide the spatial domain into clusters by extracting information from the shapes of the underlying curves.
As the signals start at different times, in a preprocessing step the time is standardized and the waves are synchronized; then the alignment is applied in order to retain the underlying pattern. The algorithm provides the classification of the stations into two groups, characterized by different amplitude variability: starting with randomly selected clusters, the algorithm provides further allocations in order to optimize the classification; in Fig. 1 and Fig. 2, the aligned waves are represented in a small interval of time, for a better understanding of the dynamics of the waves and of their differences: Fig. 1 shows the initial allocation: the grouped signals are represented for the three dimensions respectively in the column on the left, cluster 1 and in the column on the right for the cluster 2. Fig. 2 shows the results for the optimal allocation. Fig. 3 shows the clusters for the entire sequences of the three dimensions. The distances of the stations from epicenter are reported in Tab. 1, where their spatial separation is confirmed.

Tab. 1 - Distances of the stations from the epicenter in km, according to their membership to the two groups.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>75.1</th>
<th>98.6</th>
<th>64.7</th>
<th>79.8</th>
<th>16.1</th>
<th>42.4</th>
<th>106.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td>198.9</td>
<td>279.8</td>
<td>173.8</td>
<td>143.8</td>
<td>117.1</td>
<td>135.8</td>
<td>212.8</td>
</tr>
</tbody>
</table>

Fig. 3 - Clusters for the entire sequences of the three dimensions.
Conclusions. Our methodology allows to find groups of curves which are spatially homogeneous, overcoming some limitations arising with more traditional techniques based on cross correlations.

In this robust approach waveforms are clustered after a registration, by mean of which shape and phase variations are accommodated using derivatives and elastic distances.

We adopt clustering techniques based on modified band depth for warped curves, that retain shape and other features of the data, in order to identify similarity of waveforms. We show how these methodologies can be joined together to carry out classification of spatially correlated curves, as the produces clusters are also well identified in the space, preserving their natural structure.

Acknowledgements This work is supported by National grant MIUR, PRIN-2015 program, Prot.20157PRZC4: Complex space-time modeling and functional analysis for probabilistic forecast of seismic events.

References