Clustering analysis of probabilistic seismic hazard for the selection of ground motion time histories in vast areas

Published Version:
C. Mascandola, S. Barani, M. Massa, E. Paolucci, D. Albarello (2020). Clustering analysis of probabilistic seismic hazard for the selection of ground motion time histories in vast areas. BULLETIN OF EARTHQUAKE ENGINEERING, 18(7), 2985-3004 [10.1007/s10518-020-00819-x].

Availability:
This version is available at: https://hdl.handle.net/11585/950504 since: 2023-12-12

Published:
DOI: http://doi.org/10.1007/s10518-020-00819-x

Journal : SmallCondensed 10518 | Article No : 819 | Pages : 20 | MS Code : 819 | Dispatch : 12-3-2020

Bulletin of Earthquake Engineering
https://doi.org/10.1007/s10518-020-00819-x

ORIGINAL RESEARCH

Check for updates

Clustering analysis of probabilistic seismic hazard for the selection of ground motion time histories in vast areas

C. Mascandola[1,2] · S. Barani[3] · M. Massa[1] · E. Paolucci[4] · D. Albarello[4]

Received: 10 September 2019 / Accepted: 9 March 2020

Abstract
We present a methodology for the selection of accelerometric time histories as input for dynamic response analyses over vast areas. The method is primarily intended for seismic microzonation studies and regional probabilistic seismic hazard assessments that account for site effects. It is also suitable for structural response analyses if one would like to use a fixed set of ground motion records for analyzing multiple structures with different (or unknown) periods. The proposed procedure takes advantage of unsupervised machine learning techniques to identify zones (i.e., groups of sites) with homogeneous seismic hazard, for which the same set of earthquake recordings can be reasonably used in the numerical simulations. The procedure consists of three steps: (1) data-driven cluster analysis to identify groups of sites with comparable seismic hazard levels for a specified mean return period (MRP); (2) for each zone, definition of a single, reference uniform hazard spectrum (UHS) corresponding to the MRP of interest; (3) selection of a set of accelerometric recordings that are consistent with the magnitude-distance scenarios contributing to the hazard of each zone, and meet the spectrum-compatibility requirement with respect to the reference UHS. An application of the procedure in the Po Plain (Northern Italy) is described in detail.

Keywords Probabilistic seismic hazard · Seismic hazard disaggregation · Seismogram selection · Cluster analysis

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10518-020-00819-x) contains supplementary material, which is available to authorized users.

✉ C. Mascandola
claudia.mascandola@ingv.it

[1] Istituto Nazionale di Geofisica e Vulcanologia - Sezione di Milano, Via Alfonso Corti 12, 20133 Milan, Italy

[2] Dipartimento di Scienze della Terra, Università di Pisa, Via S. Maria 53, 56126 Pisa, Italy

[3] Dipartimento di Scienze della Terra, dell'Ambiente e della Vita (DISTAV), Università degli Studi di Genova, Corso Europa 26, 16132 Genoa, Italy

[4] Dipartimento di Scienze Fisiche, della Terra e dell'Ambiente (DSFTA), Università degli Studi di Siena, Via Laterina 8, 53100 Siena, Italy

Journal : SmallCondensed 10518	Article No : 819	Pages : 20	MS Code : 819	Dispatch : 12-3-2020

1 Introduction

Selection of accelerometric time histories is a key step of many applications in the field of engineering seismology, such as structural response analyses and ground response assessments. Although most of these studies are target-specific (i.e., structure- or site-specific), risk mitigation strategies adopted by local governments often require the evaluation of the dynamic response of multiple targets (e.g., strategic structures for emergency management, critical facilities, sites susceptible to amplification effects, landslides) spread over wide areas. These areas may present significant hazard variability depending on the contributions of both local and distant earthquake sources. Hence, reference probabilistic seismic hazard estimates (e.g., national seismic hazard maps) are at the foundations of the selection of sets of ground motion recordings in many practical applications. Indeed, earthquake recordings are often required to be consistent with the reference hazard at the target and to capture the inherent variability of the expected ground motion.

In the field of engineering seismology, the selection of ground-motion time histories is an important task of seismic microzonation, which aims at identifying and characterizing all potential geohazards within an area, such as ground motion hazard, liquefaction hazard, landslide hazard, and fault displacement hazard (e.g., Sitharam and Anbazhagan 2008; Ansal et al. 2010; SM Working Group 2015). In particular, most dynamic analyses aim at quantifying site amplification. Hence, seismic microzonation provides the basis for refined seismic hazard assessments and risk analyses on a scale that goes beyond that of the single, specific location (Barani et al. 2020). Furthermore, the selection of ground motion records is the basis of structural engineering studies for performance-based design or, more generally, for the vulnerability assessment of structures and infrastructures (e.g., Silva et al. 2019).

The present work aims at fulfilling one of the needs of studies that require the realization of dynamic analyses over wide areas or for a large number of structures with different (or unknown) periods. Specifically, it deals with the selection of sets of accelerometric time histories for extensive dynamic response analyses. To this end, we take advantage of unsupervised clustering algorithms to zone areas into groups of sites characterized by similar seismic hazard, here expressed in terms of uniform hazard spectra (UHSs) corresponding to a specified return period. For each group (i.e., zone), a set of ground motion recordings is then selected. Conceptually, our approach is similar to that of Rota et al. (2012) where the authors propose a zonation of the entire Italian territory based on the similarity of the design acceleration response spectra provided by the Italian building code (Ministero delle Infrastrutture e dei Trasporti 2008). In that study, the zoning was based on a trial-and-error procedure that defines sets of elastic response spectra that simultaneously satisfy specific conditions on four target parameters, three of which concur in the definition of the spectral shape (according to the mathematical formulation given by the Italian norms), and one quantifies the deviation δ of each spectrum from a reference one (Iervolino et al. 2008). Compared to the procedure of Rota et al. (2012), the clustering approach proposed here presents the advantage of removing the subjectivity in the choice of the conditions to be applied on some target parameters. Indeed, unsupervised machine learning algorithms make inferences from datasets using only the input (data) vectors (e.g., uniform hazard spectra for a number of sites). Analyst's expertise only helps to establish the appropriate number of clusters, which should reflect the regional variability of the hazard. The interpretation of the regional seismic hazard, both in terms of hazard maps and in terms of geographical distribution of the magnitude (M) and distance (R) scenarios contributing the

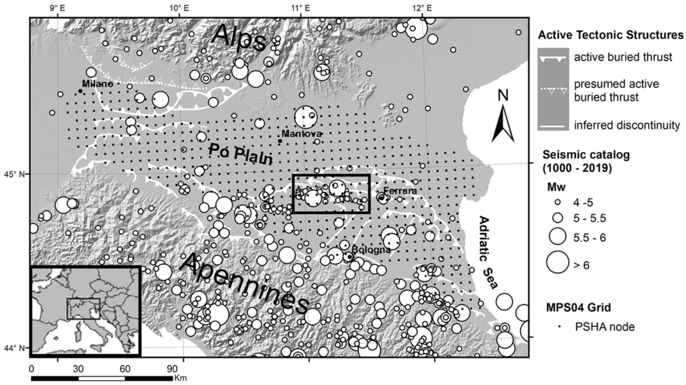

Fig. 1 Distribution of seismicity ($M_w > 4$) in the study area. Historical epicenters are from Rovida et al. (2016), while instrumental seismicity is from Osservatorio Nazionale Terremoti (http://terremoti.ingv.it/). Active tectonic structures are shown in background (Martelli et al. 2017). The black rectangle identifies the area of the 2012 Emilia seismic sequence. The black dots indicate the nodes of the computational grid of the Italian seismic hazard assessment considered in the clustering analysis

most to the hazard, is essential to refine the number of clusters found through the application of specific statistical approaches. In the present study, we examine the reliability of three conventional techniques; namely, the elbow method (e.g., Sugar 1998; Sugar et al. 1999), the average silhouette method (Rousseeuw and Kaufman 1990), and the gap statistic approach (Tibshirani et al. 2001).

The clustering procedure introduced above is presented through an application in the Po Plain in Northern Italy. The study area is shown in Fig. 1, which displays the seismicity distribution ($M_w > 4$) from 1000 to 2019. In this area, severe ground motion amplification effects were observed and described in several scientific studies released following the 2012 Emilia seismic sequence (e.g., Priolo et al. 2012; Massa and Augliera 2013; Luzi et al. 2013; Paolucci et al. 2015; Mascandola et al. 2017; Laurenzano et al. 2017). Since then, local governments have funded extensive seismic microzonation activities, and other significant research efforts have been spent in Italy with the aim of overcoming conventional probabilistic seismic hazard estimates for reference rock conditions (e.g., Barani and Spallarossa 2017; Mascandola et al. 2017, 2019; Barani et al. 2020). In light of these considerations, as well as the regional variability of seismic activity, the Po Plain is a suitable area where the potentiality of unsupervised clustering algorithms can be tested to select input time series for dynamic analyses.

The procedure consists of three steps. In the first one, a cluster analysis is carried to out in order to divide the study area into group of sites (i.e., clusters or zones) characterized by similar hazard levels (i.e., similar UHSs). In the second step, a target UHS is defined for each cluster. Its shape should represent the general pattern of the UHSs associated with the sites belonging to that cluster, and therefore embeds the contributions to the hazard from the same group of seismogenic sources (or from sources with similar seismic potential). Finally, the target spectra defined at the previous step are used as reference spectral shapes for the selection of groups of accelerometric recordings. In the present study, we show an application considering natural accelerograms. In particular, since different *M–R* scenarios contribute to the hazard in a given zone (i.e., cluster), we will show how to take these contributions into consideration in the selection of sets of natural accelerograms that cover a hazard-consistent range of aleatory variability in magnitude and source-to-site distance, and meet the spectrum-compatibility requirement with respect to the reference UHSs. Depending on the scope of work, the last two steps can be clearly adapted to handle other types of target spectra (e.g., conditional mean spectra instead of uniform hazard spectra), accelerograms (i.e., artificial or synthetic), selection techniques (e.g., attempting to match or not specific record properties, such as magnitude and distance), and spectral matching criteria. Interested readers on these topics can refer to the articles of Bommer and Acevedo (2004), Baker and Cornell (2006), Watson-Lamprey and Abrahamson (2006), Kottke and Rathje (2008), Iervolino et al. (2009), Buratti et al. (2011), Corigliano et al. (2012), Burks et al. (2015), Baker and Lee (2018) and Tsioulou et al. (2019).

2 Methodology

The grouping of sites with similar ground motion hazard is carried out here by using the *k*-means algorithm, which was first proposed by Lloyd (1957). The *k*-means procedure is one of the most commonly used unsupervised machine learning technique for partitioning a given data set into a specified number K of groups of objects that are similar to each other, commonly termed as 'clusters' (e.g., Wagstaff et al. 2001). In simple words, given a set of N objects each having measurements on P 'attributes' (i.e., variables), the *k*-means algorithm assigns observations to a given cluster k ($1 \leq k \leq K$) so as to minimize the total intra-cluster variation (or within-cluster sum of squares) through an iterative relocation scheme. Precisely, if C_k denotes the set of n_k objects in cluster k, the total within-cluster sum of squares is defined as (e.g., Stenley 2006):

$$SSE = \sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} \left(x_{ij} - \bar{x}_j^{(k)} \right)^2 \tag{1}$$

where SSE stands for error sum of squares, x_{ij} indicates a generic observation, and $\bar{x}_j^{(k)}$ is the centroid value for the jth variable in the cluster C_k:

$$\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij} \tag{2}$$

A detailed description of the *k*-means is given by Lloyd (1957, 1982), Forgey (1965), MacQueen (1967), and Hartigan (1975). In this study, we use the algorithm of Hartigan and Wong (1979) implemented in the R software environment (R Core Team 2017), which requires as input only the array of observations and the number K of clusters.

In this study, we partition the nodes (with the associated hazard values) of the computational grid considered in the Italian seismic hazard assessment (MPS Working Group 2004; Stucchi et al. 2011). Specifically, we consider $N = 596$ nodes covering the Po Plain area (Fig. 1). For each node, an object is defined by its relevant UHS for a given mean return period (MRP). Each object is characterized by $P = 11$ attributes, each of which corresponds to a spectral ordinate of the UHS (i.e., 5%-damped spectral acceleration $S_a(T)$ for an oscillator period T in the range 0–2s). Note that if one is interested in a specific spectral range, then the input array should only include data in that specific interval, as the association of a point to a cluster is influenced to some extent by the input data.

A preliminary but fundamental step of the clustering analysis is the choice of the number K of clusters. In the present analysis, this choice is guided by a qualitative interpretation of the regional seismic hazard (both in terms of hazard maps and in terms of magnitude and distance maps obtained from hazard disaggregation) and through the use of statistical techniques aimed at finding the optimal value of K. A guide to the choice of K is presented in the next sub-section with reference to the study area.

2.1 Determination of the number of clusters

As stated above, the number of clusters should reflect the regional variability of the hazard in the study area. Therefore, in order to determine the appropriate value of K for the cluster analysis, a preliminary examination of seismic hazard maps and hazard disaggregation results should be coupled with the application of specific statistical techniques designed for this purpose. Statistical approaches can be used to corroborate the approximate and, to some extent, subjective information about the number of clusters provided by the hazard maps. Conversely, these latter may help to refine the results from statistical techniques. As mentioned in the introduction, three alternative techniques are examined below.

Before discussing the hazard results, it is worth specifying that the hazard maps shown in the following are obtained by interpolation of the hazard values computed within the framework of the Italian seismic hazard assessment (MPS Working Group 2004; Stucchi et al. 2011). Two return periods (i.e., 475 and 2475 years), corresponding to different limit states, are considered in order to examine the sensitivity of clustering to the MRP. As is known, indeed, the contribution from closer, large magnitude scenarios increases with increasing return period (e.g., Iervolino et al. 2011), thus affecting the shape of the UHSs used in the cluster analysis. The disaggregation maps are based on the results of Barani et al. (2009). We consider the maps for PGA (i.e., $T=0$s) and S_a(2s) only, as these spectral ordinates were shown to be well representative of the M–R contributions at short-to-medium and medium-to-long periods, respectively (Barani et al. 2009). Note that disaggregation results are expressed here in terms of mean values of magnitude ($\bar{M}$) and distance ($\bar{R}$). Mean values are preferred to their modal counterparts (M^* and R^*) as, in the case of multi-modal joint M–R distributions (or M and R distributions in the case of 1D disaggregation), the mean implicitly captures the contributions to the hazard from multiple, dominating scenarios of magnitude and distance in a single metric. On the other hand, although the mode has the advantage of representing the event that most likely generates the exceedance of the target ground motion level at the site considered, it evidently loses information relative to possible secondary peaks in the M–R distributions. Hence, maps of modal M and R may be uninformative about multiple contributions to the hazard.

Figure 2 shows the seismic hazard maps and disaggregation maps for the study area. From top to bottom, the panels in the left column show the geographic distribution of the PGA values for an MRP of 475 years, and the corresponding scenarios of $\bar{M}$ and $\bar{R}$ obtained from the 2D disaggregation (i.e., M–R disaggregation) of the mean annual rate of exceedance of the 475-year PGA. The panels in the right column show the same maps but for the 475-year S_a(2s). The PGA hazard map (Fig. 2a) and the corresponding disaggregation map for $\bar{R}$ (Fig. 2e) identify at least three zones, each of which identifies a group of sites apparently characterized by similar hazard. The first area encompasses the central sector of the Po Plain, with PGA values comprised between 0.075 and 0.1 g, and $\bar{R}$ spanning between about 20 and 60 km. Here, $\bar{M}$ is in the range 4.5–5.5 (Fig. 2c). A second area includes the marginal Po Plain sectors near Milan and the Adriatic coast, which present PGA values lower than 0.075 g and $\bar{R}$ values ranging between about 60 and 120 km. According to the geographic distribution of $\bar{R}$, these marginal areas may form a single zone. However, as

Journal : SmallCondensed 10518	Article No : 819	Pages : 20	MS Code : 819	Dispatch : 12-3-2020

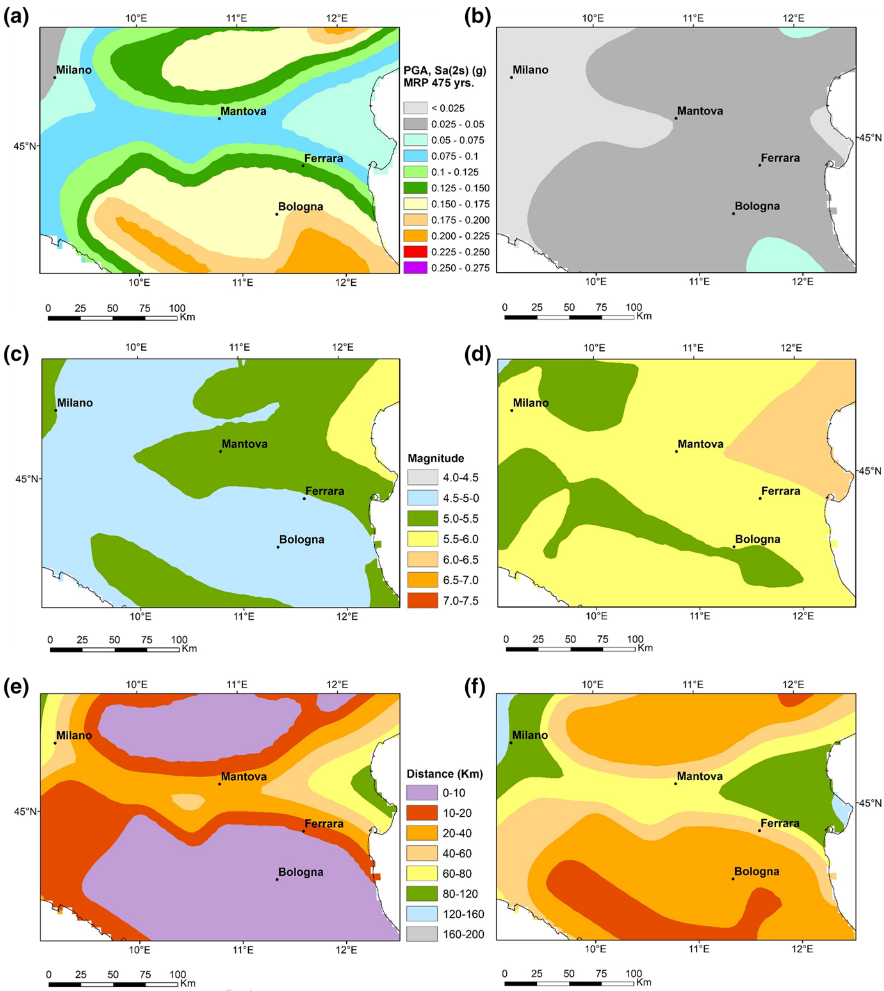

Fig. 2 Left column: PGA hazard map for a mean return period of 475 years (**a**) and corresponding disaggregation maps of mean magnitude (**c**) and distance (**e**). Right column: same as left column but for 2s spectral acceleration, $S_a(2s)$

suggested by the map of $\bar{M}$ in Fig. 2c, they may represent two distinct group of sites, with the Adriatic one (to the east) controlled by stronger, distant events. The third zone includes both the Alpine foothills to the north and the southern Po Plain sector towards the Apennines, and presents PGA values up to about 0.2 g. Here, the PGA hazard is dominated by nearby scenario events ($\bar{R} < 20$ km) with $\bar{M}$ less than 5.5. Possibly, a transition zone that separates the northernmost and southernmost portions of these two sectors from the central Po Plain can be identified. It presents 475-year PGA values comprised between approximately 0.1 and 0.15 g and $\bar{R}$ values that are up to about 20 km.

Similar considerations can be done by analyzing the maps for $S_a(2s)$, particularly the disaggregation map for $\bar{R}$ (Fig. 2f). Again, this latter map suggests setting the number K of clusters to 3 or 4. Hence, changing the spectral period does not seem to affect the choice of the number of clusters to be adopted in the subsequent clustering analysis (at least in the

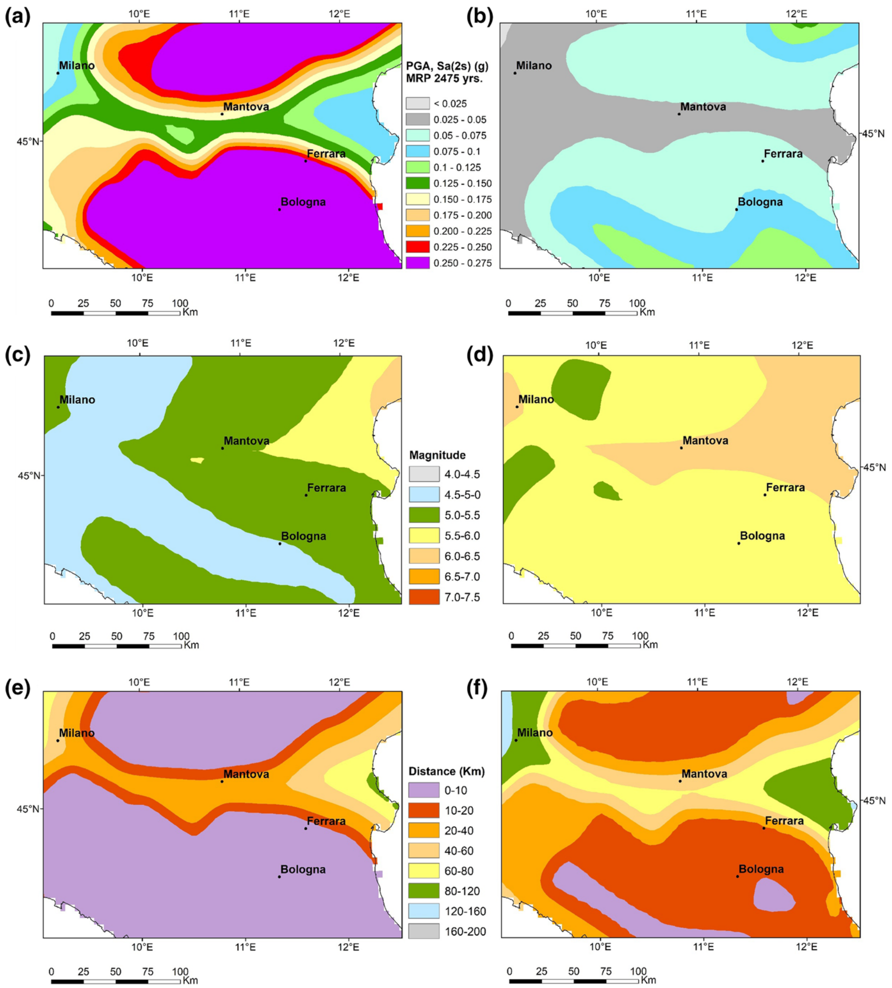

Fig. 3 Left column: PGA hazard map for a mean return period of 2475 years (**a**) and corresponding disaggregation maps of mean magnitude (**c**) and distance (**e**). Right column: same as left column but for 2s spectral acceleration, $S_a(2s)$

area considered). The same holds true for the return period. Indeed, analyzing the maps in Fig. 3, which refer to an MRP of 2475 years, leads to the same conclusions about K as for an MRP of 475 years. Note that this does not guarantee that, for a fixed value of K, clustering analyses carried out on input data corresponding to different MRPs yield the same point-cluster association. However, we will observe in the next section that, at least for the MRPs considered, the cluster composition is little sensitive to changes in the MRP of the UHSs used in input.

In order to strengthen the observations above, statistical techniques can be applied to constrain the value of K. An example is presented in Fig. 4. Specifically, the figure compares the outcomes from three standard techniques: the elbow method (e.g., Sugar 1998; Sugar et al. 1999), the average silhouette method (Rousseeuw and Kaufman 1990), and the gap statistic approach (Tibshirani et al. 2001). Again, we use the functions provided in the

Journal : SmallCondensed 10518 | Article No : 819 | Pages : 20 | MS Code : 819 | Dispatch : 12-3-2020

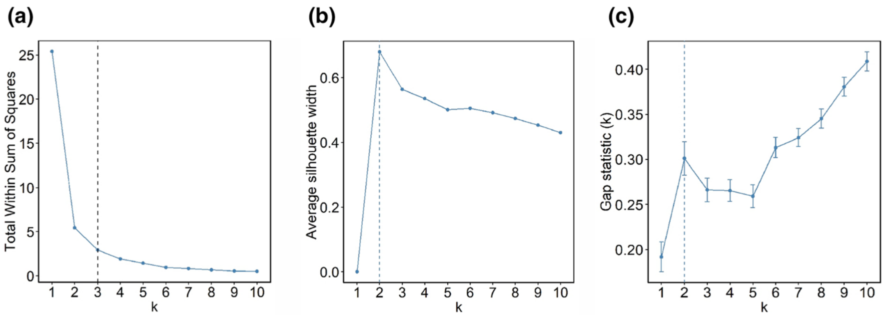

Fig. 4 Comparison of different optimization algorithms for the determination of *K*. **a** Elbow method; **b** average silhouette method; **c** gap statistic method. The vertical dashed line in each diagram indicates the optimal number of clusters

R software environment (R Core Team 2017). The first two methods define the most appropriate value of *K* on the grounds of an optimization criterion, such as the minimization of the within-cluster sum of squares (see Eq. 1) or the maximization of the average silhouette over a range of possible values for *K*. The latter approach consists of comparing the (logarithmic) within-cluster sum of squares for different values of *K* with that expected under an appropriate reference null distribution. Mathematically, this is expressed by Eq. 3 in Tibshirani et al. (2001), which defines the so-called gap function $Gap(k)$. The optimal number of clusters for the given data set is the smallest k such that $Gap(k) \geq Gap(k+1) - s_{k+1}$, where s_k is the error associated with the expected value of the within cluster variation (indicated by the error bars in Fig. 4c). Figure 4 shows that the three methods provide optimal values of *K* equal to 2 or 3. In particular, the elbow method indicates a value in the range 2–4 (identified by the bending of the curve in Fig. 4a), thus corroborating the conclusions drawn from the analysis and interpretation of the hazard and its disaggregation, which suggest values of *K* equal to 3 or 4.

3 Clustering analysis

The results of a clustering analysis are conveniently displayed through the so-called PCA score plots. Specifically, each object is typically depicted as a point in a 2D space where the axes are the first two principal components (hereinafter indicated as Dim1 and Dim2) determined via Principal Component Analysis (PCA) (e.g., Wilks 2011). In brief, the PCA applies an orthogonal linear transformation that converts a set of *P* possibly correlated variables (i.e., the spectral ordinates of the UHSs) into a smaller number of uncorrelated variables, called principal components. The first principal component accounts for the greatest variance in the data, and each succeeding component accounts for the highest variance possible under the constraint that it is orthogonal to the preceding ones. In practice, this allows reducing dimensionality of the original data set to just a few dimensions that allow a simple but effective description of the data.

Figure 5 shows the results of the cluster analysis carried out on the UHSs for an MRP of 475 years assuming $K=3$. The PCA score plot in Fig. 5a clearly indicates that the three clusters are different based on Dim1, which explains 96.7% of the point variation, while Dim2 describes 2.9% only. The UHSs belonging to each cluster are shown

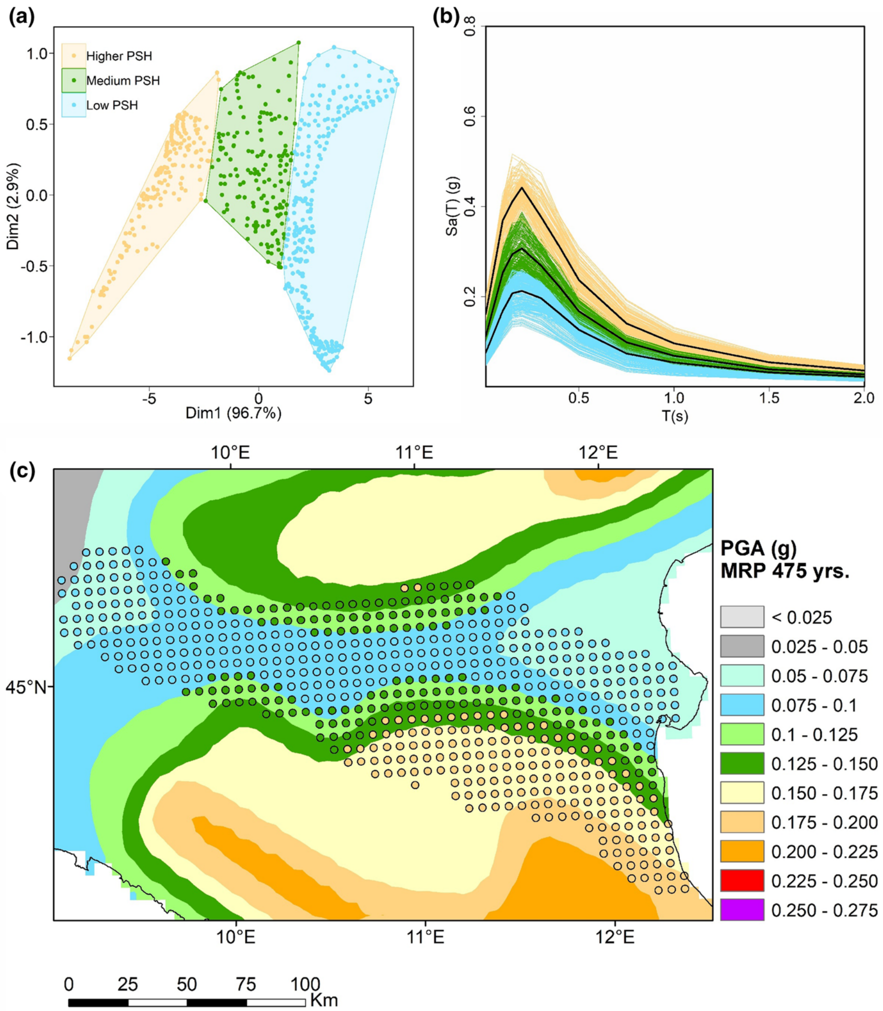

Fig. 5 Results of the clustering analysis for a mean return period of 475 years and $K=3$: **a** PCA score plot. Dim1 and Dim2 indicate the principal component axes; **b** UHS clusters including the centroid spectra (thick black lines); **c** geographical distribution of the UHS clusters (i.e., zones) superimposed on the 475-year return period PGA hazard map

in Fig. 5b (by using the same colors as in Fig. 5a) along with the UHS corresponding to the cluster centroid (in black). Note that, by definition, the latter may not correspond to any of the UHSs in the cluster. Finally, Fig. 5c shows the geographical distribution of the nodes belonging to the three clusters. The map highlights that the zonation deriving from the cluster analysis reflects the observations following the analysis of the hazard results only partially. Whereas the clustering algorithm identifies the narrow transition zone that separates the Alps foothills and the southernmost portion of the plain from the central sector (green dots in Fig. 5c), it fails to distinguish the marginal areas around Milan and near the Adriatic coast (where the hazard is controlled by distant scenarios

Journal : SmallCondensed 10518 | Article No : 819 | Pages : 20 | MS Code : 819 | Dispatch : 12-3-2020

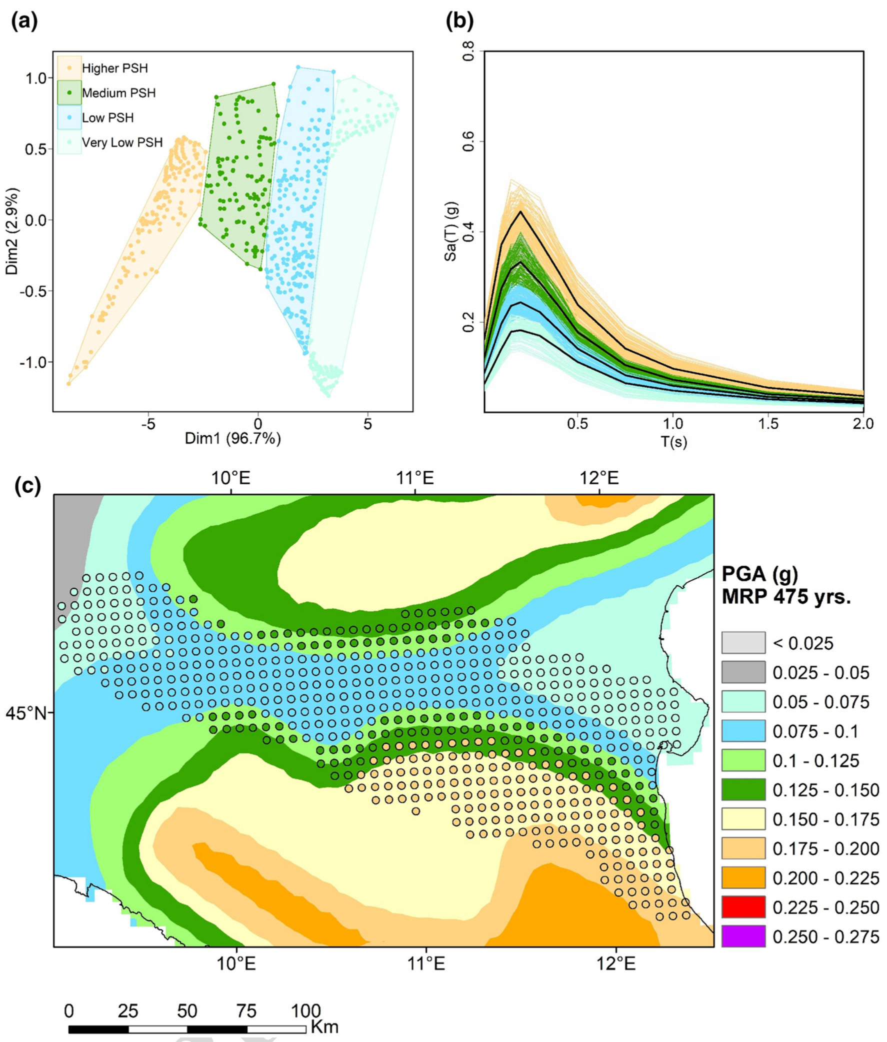

Fig. 6 Results of the clustering analysis for a mean return period of 475 years and $K=4$: **a** PCA score plot. Dim1 and Dim2 indicate the principal component axes; **b** UHS clusters including the centroid spectra (thick black lines); **c** geographical distribution of the UHS clusters (i.e., zones) superimposed on the 475-year return period PGA hazard map

up to 120 km) from the central Po Plain (where the contribution of moderately distant events, from about 20 to 40 km distance, is prevalent).

In order to separate the eastern and western marginal areas, the cluster analysis is repeated assuming $K=4$. The results are shown in Fig. 6, which consists of the same three panels of Fig. 5. In this case, besides the transition zone, the clustering algorithm allows distinction between the central plain sector and the marginal zones to the east and west. Thus, with $K=4$, four zones are clearly identified:

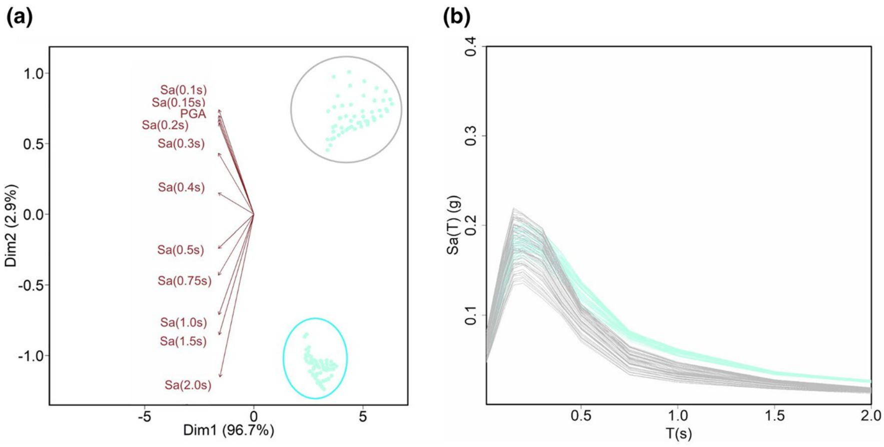

Fig. 7 Clustering of the UHSs for a mean return period of 475 years belonging to the lowest seismic hazard area (i.e., western and eastern marginal Po Plain sectors): **a** biplot showing simultaneously the PCA score plot for $K=2$ and the influence (i.e., loading) of each spectral ordinate to Dim1 and Dim2; **b** UHS clusters

- one, with higher seismic hazard, corresponding to the southern Po Plain sector towards the Apennines foothills;
- one, with moderate hazard, that marks the transition between the central Po Plain and the Alpine foothills to the north, and between the central and the southern plain towards the Apennines;
- one, characterized by low hazard, encompassing the central plain;
- one, with very low hazard, including the marginal plain sectors near Milan and the Adriatic coast.

A closer analysis of Fig. 6a indicates that this latter zone can be further separated into two distinct clusters based on the second principal component (Dim2). A focus on this zone is shown in Fig. 7. Figure 7a presents a biplot (Gabriel 1971) summarizing in a single figure the PCA score plot and the loading plot derived from the PCA. This latter plot displays the $P=11$ attributes of the data matrix as vectors pinned at the origin of the Dim1 and Dim2 axes (i.e., Dim1 = 0 and Dim2 = 0), and shows how strongly each attribute of the UHSs (i.e., spectral ordinate) influences the first and second principal components. The absolute value of the vector projection on each axis shows how much weight each attribute has on that principal component. Thus, it is possible to observe that all attributes have nearly the same influence on Dim1, while short- (PGA to 0.2s) and long- (1 to 2s) period accelerations have the stronger influence on Dim2. The wide angle between these two groups of vectors indicates that the corresponding variables are negatively correlated, and justifies the subdivision of the original cluster into two distinct partitions. Specifically, one cluster consists of the nodes near Milan and its surroundings (light blue points in the grey circle), which present a higher hazard at short periods (i.e., objects with greater positive values of Dim2). The other includes the nodes towards the Adriatic coast (light blue points in the light-blue circle) which, conversely, are characterized by a higher hazard at longer periods (i.e., objects with lower negative values of Dim2) due to the greater contribution from the stronger, distant events associated with the seismic sources in northeastern Italy. Figure 7b shows the UHSs

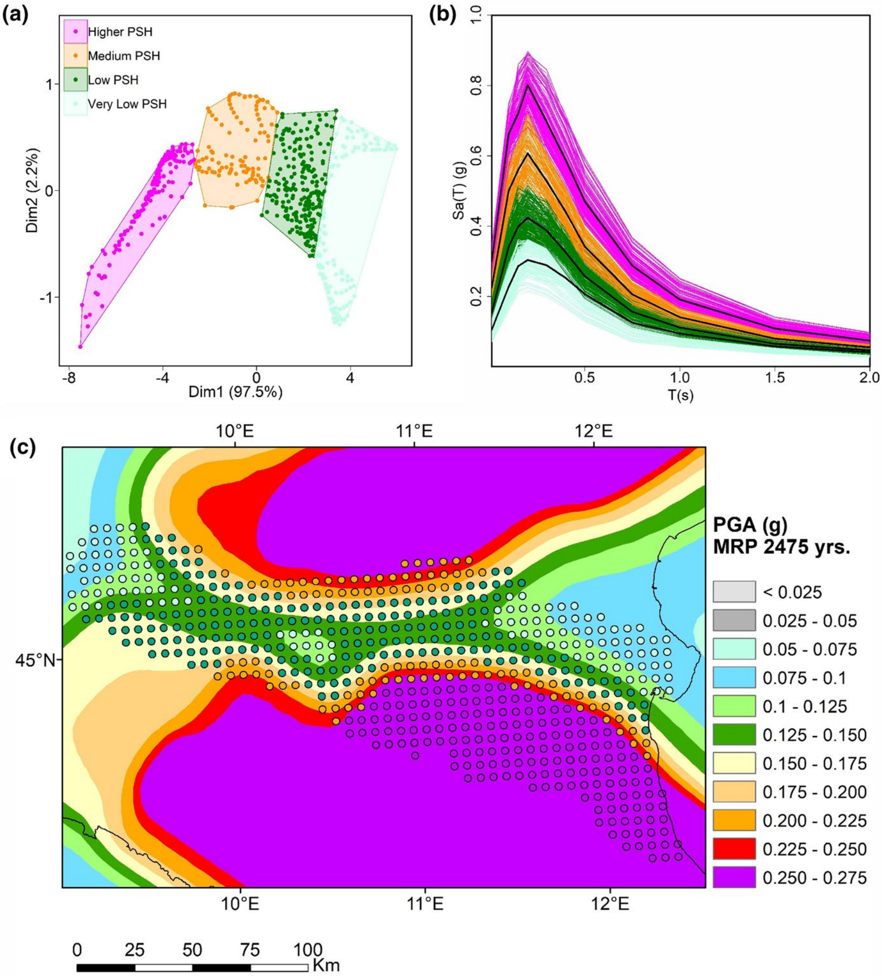

Fig. 8 Results of the clustering analysis for a mean return period of 2475 years and $K = 4$: **a** PCA score plot. Dim1 and Dim2 indicate the principal component axes; **b** UHS clusters including the centroid spectra (thick black lines); **c** geographical distribution of the UHS clusters (i.e., zones) superimposed on the 2475-year return period PGA hazard map

associated with these two clusters. Despite these considerations, for the sake of simplicity, in the subsequent applications we will assume the results of the cluster analysis for $K = 4$. In fact, the extra effort in the ground motion selection stage, which would be implied by this further clustering, appears poorly justified given the very low hazard that characterizes the easternmost and westernmost sectors of the Po Plain.

Finally, in order to show the sensitivity of clustering to the choice of return period, Fig. 8 presents the results of the cluster analysis carried out on the same nodes of Figs. 5 and 6 but for an MRP of 2475. Again, we set $K = 4$. Except for very few points (particularly evident are the light blue nodes in the middle of the Po Plain), the clusters are very similar to those defined for an MRP of 475 years (see Fig. 6). Despite this mild sensitivity of clustering to the choice of return period, slight variations in the composition of clusters may occur. In other words, changing the return period does not insure the complete preservation of the node-cluster association. Therefore, if one is interested in a specific MRP, it is advisable to perform the cluster analysis for that MRP.

4 Selection of accelerometric recordings

Once zones have been defined, the analyst has to face the issue of selecting proper ground motion time histories. Two possible strategies can be adopted in order to account for the different *M–R* scenarios contributing to the hazard of each zone. In both cases, the centroid UHS is assumed as reference. The former, which is based on standard engineering practice, consists of three major steps: (1) selection of the site (or sites) presenting the spectral acceleration hazard closer to the centroid spectrum (in the entire range of periods covered by the UHS or in a specific range); (2) disaggregation of the mean annual rate of exceedance of the spectral acceleration value (or values in the case of multiple response periods) of interest; (3) selection of a group of accelerometric recordings that are consistent with the disaggregation results and some other pre-defined requirements (e.g., prevalent style of faulting in the study region and surroundings, compatibility with the reference spectrum). If the conditional mean spectrum (CMS) is the preferred target in the selection of ground motion records (this is in most structural response analyses), the 2D disaggregation at the second stage will be replaced by the *M–R–*ε disaggregation, where ε indicates the number of standard deviations by which a given value of the logarithmic ground motion, $\log(S_a(T))$, differs from the mean value predicted a ground motion attenuation equation for a given magnitude-distance pair. Then, the $\bar{M}$-$\bar{R}$-$\bar{\varepsilon}$ triplet will be used to compute the CMS. Finally, ground motion records will be selected with spectral shapes that match the conditional mean spectral shape. Interested readers on this topic can refer to Baker and Cornell (2006) and Baker (2011).

Although the site-to-site variability of the *M–R* contributions to the hazard is generally small within the same cluster (indeed, the UHS similarity directly comes from the similarity of the *M–R* scenarios contributing to the hazard at the multiple nodes within a zone), the previous approach does not guarantee for a complete and effective representation of all scenarios contributing to the hazard at the zone scale. To have a composite picture of such contributions in a zone, the analyst can lump (by stacking and normalizing) the magnitude-distance contributions obtained from the disaggregation of the seismic hazard at all computation nodes belonging to that zone, and select the time histories accordingly. A price is clearly paid in terms of computational time. For a specific attribute (i.e., spectral acceleration for a given oscillator period), the stacked contribution U_k for a particular *M–R* scenario (such that $m_1 < M < m_2$ and $r_1 < R < r_2$) relative to the *k*th cluster is given by:

$$U_k\left(m_1 < M < m_2, r_1 < R < r_2\right) = \frac{\sum_{i \in C_k} U_i\left(m_1 < M < m_2, r_1 < R < r_2\right)}{\sum_{h=1}^{n_M} \sum_{j=1}^{n_R} \sum_{i \in C_k} U_i\left(m_1 < M < m_2, r_1 < R < r_2\right)} \tag{3}$$

where C_k indicates again the set of n_k objects belonging to the *k*th cluster, n_M and n_R are the numbers of magnitude and distance bins considered in the hazard disaggregation, and $U_i\left(m_1 < M < m_2, r_1 < R < r_2\right)$ indicates the *M–R* contribution associated with the *i*th object in the cluster, namely (e.g., Barani et al. 2009):

Journal : SmallCondensed 10518	Article No : 819	Pages : 20	MS Code : 819	Dispatch : 12-3-2020

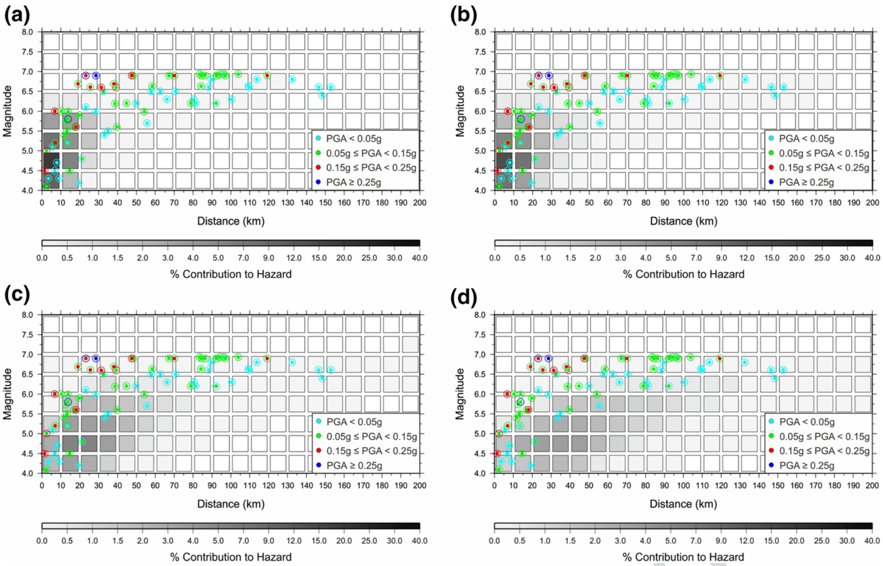

Fig. 9 Stacked *M*–*R* contributions to the 475-year PGA hazard for the four zones identified by the clustering analysis presented in Fig. 6: **a** moderate-to-high seismic hazard zone; **b** low-to-moderate hazard zone; **c** low hazard zone; **d** very low hazard zone. Contributions are normalized so that they sum up to one. Distributions of the accelerometric records listed in Table E1 of the electronic supplement are superimposed: NS and EW seismogram components are indicated by dots and (empty) circles, respectively

$$U_i\left(m_1 < M < m_2, r_1 < R < r_2\right) = \frac{\sum_{N_S} \int_{m_1}^{m_2} \int_{r_1}^{r_2} P\left[Y > y^* | m, r\right] f_{M,R}(m,r) dmdr}{y^*} \tag{4}$$

where υ is the mean annual rate of earthquake occurrence above a minimum threshold magnitude for each one of the N_S seismic sources considered in the hazard assessment, $f_{M,R}(m,r)$ is the joint probability density function of magnitude and distance, $P\left[Y > y^* | m, r\right]$ is the conditional probability of exceeding a particular value y^* of a ground motion parameter Y for a given magnitude m and distance r, and y^* is the mean annual rate of exceeding y^*.

Figure 9 shows the distribution of the stacked *M*–*R* contributions to the 475-year PGA hazard for the four zones identified by the clustering analysis presented in Fig. 6. The same is shown in Fig. 10 for the 475-year $S_a(2\text{s})$ hazard. In a similar way, stacked M–R–ε distributions can be determined if one would like to adopt the CMS as target instead of the centroid spectrum or, more generally, if one is interested in selecting records attempting to match also the ε values representative of the zone hazard.

In order to define a set of 10 accelerograms for each zone, a preselected dataset of 150 natural time histories, corresponding to 75 earthquakes recorded at accelerometric sites characterized by a shear wave velocity, V_S, greater than or equal to 750 m/s,[1] is

[1] We assume a negative tolerance of 50 m/s with respect to the standard definition (i.e., $V_S = 800$ m/s) given in the European and Italian norms (Comitè Europèen de Normalisation 2004; Ministero delle Infrastrutture e dei Trasporti 2018).

Journal : SmallCondensed 10518	Article No : 819	Pages : 20	MS Code : 819	Dispatch : 12-3-2020

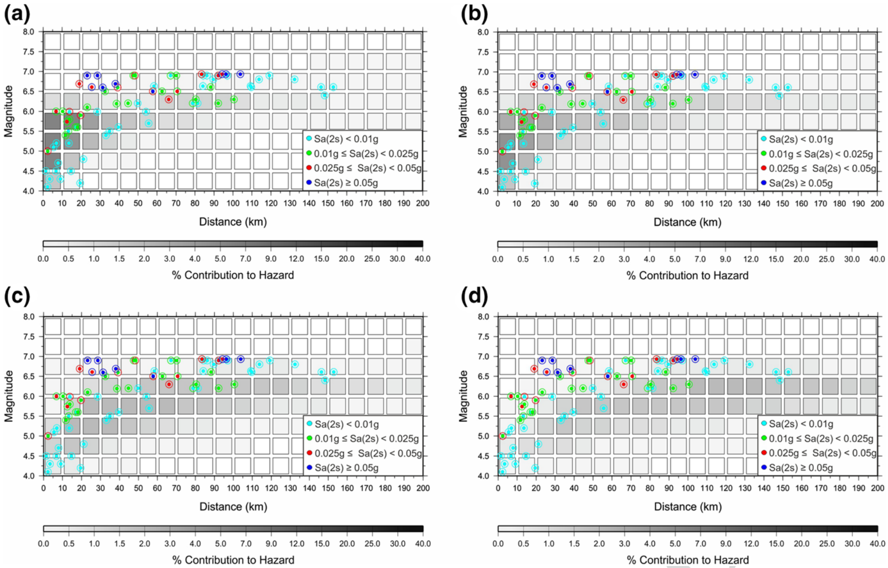

Fig. 10 Stacked *M–R* contributions to the 475-year $S_a(2s)$ hazard for the four zones identified by the clustering analysis presented in Fig. 6: **a** moderate-to-high seismic hazard zone; **b** low-to-moderate hazard zone; **c** low hazard zone; **d** very low hazard zone. Contributions are normalized so that they sum up to one. Distributions of the accelerometric records listed in Table E1 of the electronic supplement are superimposed: NS and EW seismogram components are indicated by dots and (empty) circles, respectively

considered according to such distributions. The records were selected from the European Strong Motion (ESM) database (Luzi et al. 2016) and the NGA-West2 database (Ancheta et al. 2014) taking care of discarding pulse-like, saturated, and jagged seismograms. Selection was constrained in the PGA range 0.015–0.5 g. The dataset is superimposed to each panel in Figs. 9 and 10 making distinction between the NS and EW ground motions. The corresponding earthquake features (e.g., magnitude, source-to-site distance) are listed in Table E1 of the electronic supplement. The relative 5%-damped spectra are shown in Fig. 11.

For each zone, a group of accelerograms is randomly selected so that the average spectrum (i.e., average over the 10 records) differs the least from the reference one (records associated with *M–R* bins with a null contribution to both the PGA and $S_a(2s)$ hazard were omitted from the selection). This allows to meet the spectrum-compatibility requirement indicated by the Italian seismic norms (Ministero delle Infrastrutture e dei Trasporti 2018), which allow positive and negative tolerances (with respect to the reference spectrum) of up to 30% and 10%, respectively. As our major interest is in providing groups of accelerograms that allow covering a wide range of ground motion aleatory variability, which is often recommended in ground response analyses for microzonation purposes and PSHAs at soil sites, we do not scale the time histories to specific acceleration values (e.g., local PGA corresponding to a given MRP) in the selection process. The selected time histories are marked in Table E1 in the electronic supplement by a zone identifier, and the corresponding waveforms are provided (in units of g) in ASCII format. The relative 5%-damped acceleration response spectra are shown in Fig. 11 along with the average spectrum (dashed line) and the reference one (solid black line).

Journal : SmallCondensed 10518	Article No : 819	Pages : 20	MS Code : 819	Dispatch : 12-3-2020

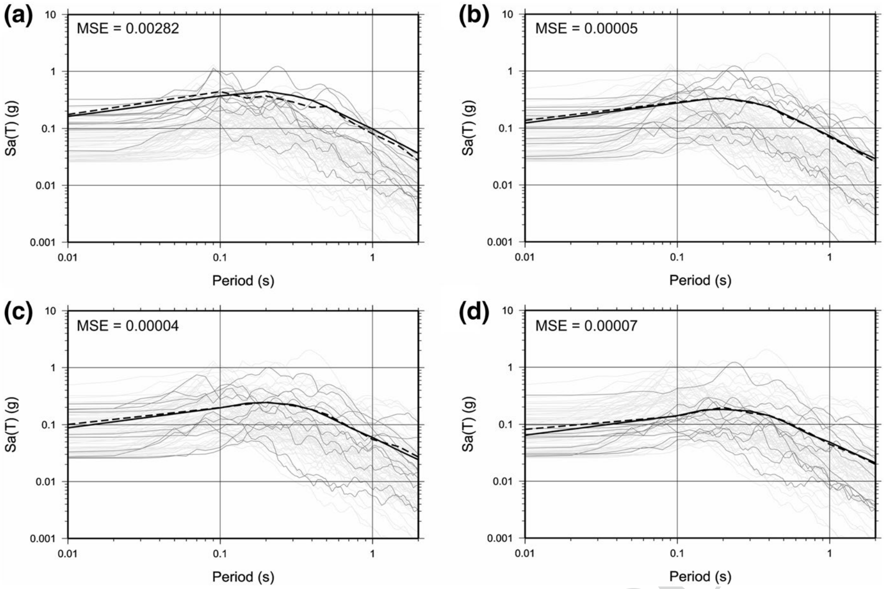

Fig. 11 Sets of acceleration response spectra (5%-damping) selected for the four zones identified by the clustering analysis: **a** moderate-to-high seismic hazard zone; **b** low-to-moderate hazard zone; **c** low hazard zone; **d** very low hazard zone. Spectral accelerations are here displayed as the larger horizontal component of the ground motion (i.e., at each period the larger spectral ordinate of the NS and EW components is chosen) according to the standards adopted by the Italian seismic hazard maps. The 150 spectra collected in the data set used in the random selection process are shown by light gray curves. The selected spectra are displayed in darker gray. The dashed and solid black lines are the average (i.e., average over the 10 selected spectra) and reference spectra, respectively. MSE indicates the mean squared error of the average spectrum (with respect to the reference one)

5 Discussion and conclusions

The paper has presented a methodology for the selection of accelerometric time histories for dynamic response analyses at multiple sites spread over wide areas. Hence, the method is primarily intended for seismic microzonation studies and seismic hazard mapping that accounts for site effects. The method is also suitable for structural response analyses if one would like to use a fixed set of ground motion records for analyzing multiple structures with different (or unknown) periods.

The zoning procedure proposed in the present work relies on a clustering analysis, which was carried out on a set of uniform hazard spectra with the aim of grouping sites with similar seismic hazard and defining, for each group, a target spectrum to be used as reference in the subsequent time history selection. To this end, an unsupervised clustering algorithm was applied, presenting the clear advantage of requiring only the number K of clusters as input. This number can be determined via statistical techniques and should reflect the spatial variability of the hazard in the study region. Hence, a rational analysis of the regional hazard may serve as a guide in the setting of K. We found that the maps guiding the choice of K are those for $\bar{R}$ along with the hazard maps for the ground motion parameter of interest. Maps for $\bar{M}$ appear less informative, although they may be helpful to refine the number of zones (e.g., with reference to our case, to separate the eastern and

Journal : SmallCondensed 10518 | Article No : 819 | Pages : 20 | MS Code : 819 | Dispatch : 12-3-2020

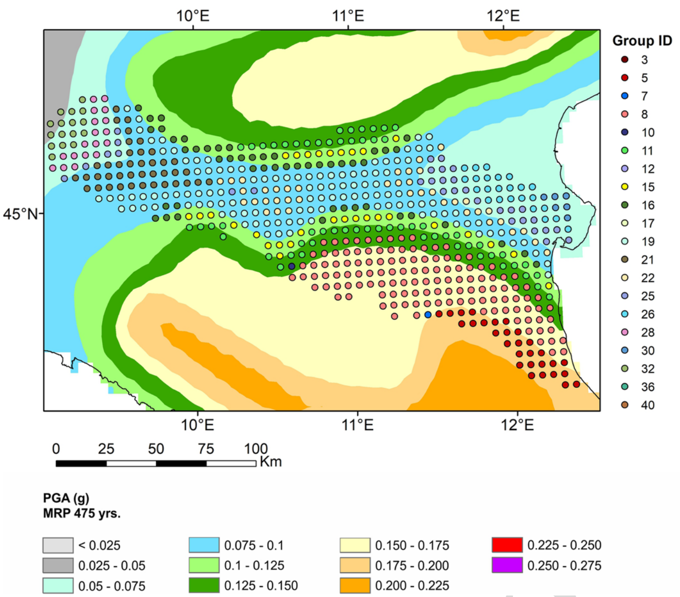

Fig. 12 Mesozonation of Rota et al. (2012). Group ID identifies the groups of ground motion records

western "wings" of the central Po Plain sector). Among the statistical techniques used to determine the value of *K*, the elbow method has provided the outcomes showing the best agreement with the value of *K* suggested by the analysis of the regional hazard.

Once the zones have been identified, accelerometric recordings can be selected according to the magnitude-distance scenarios contributing to the hazard in each zone. To this end, lumping (i.e., stacking and normalizing) the contributions associated with common *M*–*R* classes at all computation nodes belonging to a given zone may be helpful to get a composite picture of the *M*–*R* contributions in that zone. If needed, moreover, earthquake recordings can be selected in order to match some reference spectrum. In the present study, the UHS corresponding to the centroid of each cluster has been assumed as reference, but alternatives can be adopted for the same purpose (e.g., spectrum enveloping all UHSs in a cluster, conditional mean spectrum).

The application of the procedure to the Po Plain area led to the definition of four homogenous zones corresponding to separate areas with homogeneous seismic hazard (see Fig. 6). Compared to the mesozonation of Rota et al. (2012) (Fig. 12), which refer to the same return period considered here (i.e., 475 years), one may notice substantial differences both in the number of clusters (i.e., groups of accelerograms) and in cluster distribution. In particular, comparing Fig. 12 with Fig. 6 immediately reveals that Rota et al. (2012) identify a larger number of clusters, which in some sectors of the Po Plain does not reflect the distribution of the hazard and hazard disaggregation. This is particularly evident in the central plain sector, where the procedure used by Rota et al. (2012) appears to have no clustering power, leading to many sparse nodes (or groups of nodes). This effect can not be attributed to the objects used in input (we recall that Rota et al. (2012) partitioned the design spectra provided by the Italian building code instead of the related UHSs), but to the rationale behind the procedure of Rota et al. (2012), which does not account for the information about the number of clusters provided by seismic hazard maps and hazard disaggregation. If this information is ignored, the number of clusters may be extremely high, without explicit justification in the hazard distribution.

Acknowledgements We are grateful to two anonymous reviewers for their valuable comments and suggestions that have improved the manuscript. Moreover, we are thankful to E. Zuccolo and M. Rota for providing us with the grid displayed in Fig. 12.

References

Ancheta TD, Darragh RB, Stewart JP, Seyhan E, Silva WJ, Chiou BSJ, Wooddell KE, Graves RW, Kottke AR, Boore DM, Kishida T, Donahue JL (2014) NGA-West2 database. Earthq Spectra 30(3):989–1005

Ansal A, Kurtulus A, Tonuk G (2010) Seismic microzonation and earthquake damage scenarios for urban areas. Soil Dyn Earthq Eng 30:1319–1328

Baker JW (2011) The conditional mean spectrum: a tool for ground motion selection. J Struct Eng 137:322–331

Baker JW, Cornell CA (2006) Spectral shape, epsilon and record selection. Earthq Eng Struct Dyn 35(9):1077–1095

Baker JW, Lee C (2018) An improved algorithm for selecting ground motions to match a conditional spectrum. J Earthq Eng 22(4):708–723

Barani S, Spallarossa D (2017) Soil amplification in probabilistic ground motion hazard analysis. Bull Earthq Eng 15(6):2525–2545

Barani S, Spallarossa D, Bazzurro P (2009) Disaggregation of probabilistic ground-motion hazard in Italy. Bull Seismol Soc Am 99(5):2638–2661

Barani S, Ferretti G, De Ferrari R (2020) Incorporating results from microzonation into probabilistic seismic hazard analysis: an example in western Liguria (Italy). Eng Geol. https://doi.org/10.1016/j.enggeo.2020.105479

Bommer JJ, Acevedo AB (2004) The use of real earthquake accelerograms as input to dynamic analysis. J Earthq Eng 8(SI1):43–91

Buratti N, Stafford PJ, Bommer JJ (2011) Earthquake accelerogram selection and scaling procedures for estimating the distribution of drift response. J Struct Eng 137:345–357

Burks LS, Zimmerman RB, Baker JW (2015) Evaluation of hybrid broadband ground motion simulations for response history analysis and design. Earthq Spectra 31(3):1691–1710

Corigliano M, Lai CG, Rota M, Strobbia CL (2012) ASCONA: automated selection of compatible natural accelerograms. Earthq Spectra 28(3):965–987

Forgey E (1965) Cluster analysis of multivariate data: efficiency vs. interpretability of classification. Biometrics 21:768

Gabriel RK (1971) The biplot—graphic display of matrices with application to principal component analysis. Biometrika 58:453–467

Hartigan JA (1975) Clustering algorithms. Wiley, New York

Hartigan JA, Wong MA (1979) Algorithm AS136: a K-means clustering algorithm. Appl Stat 28:100–108

Iervolino I, Maddaloni G, Cosenza E (2008) Eurocode 8 compliant real record sets for seismic analysis of structures. J Earthq Eng 12(1):54–90

Iervolino I, Galasso C, Cosenza E (2009) REXEL: computer aided record selection for code-based seismic structural analysis. Bull Earthq Eng 8:339–362

Iervolino I, Chioccarelli E, Convertito V (2011) Engineering design earthquakes from multimodal hazard disaggregation. Soil Dyn Earthq Eng 31(9):1212–1231

Kottke A, Rathje EM (2008) A semi-automated procedure for selecting and scaling recorded earthquake motions for dynamic analysis. Earthq Spectra 24(4):911–932

Laurenzano G, Priolo E, Mucciarelli M, Martelli L, Romanelli M (2017) Site response estimation at Mirandola by virtual reference station. Bull Earthq Eng 15(6):2393–2409

Lloyd SP (1957) Least squares quantization in PCM, unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meet., Atlantic City, NJ. Also, IEEE Trans. Inform. Theory (Special Issue on Quantization), vol IT-28, pp 129–137 (1982)

Luzi L, Pacor F, Ameri G, Puglia R, Burrato P, Massa M, Augliera P, Franceschina G, Lovati S, Castro R (2013) Overview on the strong motion data recorded during the May–June 2012 Emilia seismic sequence. Seismol Res Lett 84(4):629–644

Luzi L, Puglia R, Russo E, D’Amico M, Felicetta C, Pacor F et al (2016) The engineering strong-motion database: a platform to access pan-European accelerometric data. Seismol Res Lett 87(4):987–997

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1(14), pp 281–297

Martelli L (coord.), Bonini M, Calabrese L, Corti G, Ercolessi G, Molinari FC, Piccardi L, Pondrelli S, Sani F, Severi P (2017) Carta sismotettonica della Regione Emilia-Romagna e aree limitrofe, scala 1:250.000 (edizione 2016), Con note illustrative, Regione Emilia-Romagna, SGSS; CNR, IGG sez. FI; Università degli Studi di Firenze, DST; INGV sez. BO. D.R.E.AM., Italy (in Italian)

Mascandola C, Massa M, Barani S, Lovati S, Santulin M (2017) Long-period amplification in deep alluvial basins and consequences for site-specific probabilistic seismic-hazard analysis: an example from the Po Plain (Northern Italy). Bull Seismol Soc Am 107(2):770–786

Mascandola C, Massa M, Barani S, Albarello D, Lovati S, Martelli L, Poggi V (2019) Mapping the seismic bedrock of the Po Plain (Italy) through ambient-vibration monitoring. Bull Seismol Soc Am 109(1):164–177

Massa M, Augliera P (2013) Teleseisms as estimator of experimental long period site amplifications: example in the Po Plain (Italy) from the 2011 Mw 9.0 Tohoku-Oki (Japan) earthquake. Bull Seismol Soc Am. 103(5):2541–2556

Ministero delle Infrastrutture e dei Trasporti (2008) Norme tecniche per le costruzioni, D.M. 14 Gennaio 2008, Supplemento ordinario alla Gazzetta Ufficiale No. 29, 4 Febbraio 2008 (in Italian)

Ministero delle Infrastrutture e dei Trasporti (2018) Aggiornamento delle Norme Tecniche per le Costruzioni. Supplemento ordinario alla Gazzetta Ufficiale No. 42 del 20 Febbraio 2018 (in Italian)

MPS Working Group (2004) Redazione della mappa di pericolosità sismica prevista dall’Ordinanza PCM 3274 del 20 marzo 2003. Rapporto conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma

Paolucci E, Albarello D, D’Amico S, Lunedei E, Martelli L, Mucciarelli M, Pileggi D (2015) A large scale ambient vibration survey in the area damaged by May–June 2012 seismic sequence in Emilia Romagna, Italy. Bull Earthq Eng 13(11):3187–3206

Priolo E, Romanelli M, Barnaba C, Mucciarelli M, Laurenzano G, Dall’Olio L, Zeid NA, Caputo R, Santarato G, Vignola L, Lizza C, Di Bartolomeo P (2012) The Ferrara thrust earthquakes of May–June 2012: preliminary site response analysis at the sites of the OGS temporary network. Ann Geophys 55(4):591–597. https://doi.org/10.4401/ag-6172

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rota M, Zuccolo E, Taverna L, Corigliano M, Lai CG, Penna A (2012) Mesozonation of the Italian territory for the definition of real spectrum-compatible accelerograms. Bull Earthq Eng 10(5):1357–1375

Rousseeuw PJ, Kaufman L (1990) Finding groups in data. Wiley Online Library, Hoboken

Rovida A, Locati M, Camassi R, Lolli B, Gasperini P (eds) 2016. Catalogo Parametrico dei Terremoti Italiani (CPTI15). Istituto Nazionale di Geofisica e Vulcanologia (INGV). https://doi.org/10.6092/INGV.IT-CPTI15

Silva V, Akkar S, Baker J, Bazzurro P, Castro JM, Crowley H, Dolsek M, Galasso C, Logomarsino S, Monteiro R, Perrone D, Pitilakis K, Vamvatsikos D (2019) Current challenges and future trends in analytical fragility and vulnerability modeling. Earthq Spectra 35(4):1927–1952

Sitharam T, Anbazhagan P (2008) Seismic microzonation: principles, practices and experiments. Electron J Geotech Eng. Special volume, Bouquet 08

SM Working Group (2015) Guidelines for seismic microzonation. conference of regions and autonomous Provinces of Italy, Civil Protection Department, Rome, (Original Italian Edition: Gruppo di lavoro MS, Indirizzi e criteri per la microzonazione sismica, Conferenza delle Regioni e delle Province autonome - Dipartimento della protezione civile, Roma, 2008, 3 vol. e Dvd). http://www.protezionecivile.gov.it/httpdocs/cms/attach_extra/GuidelinesForSeismicMicrozonation.pdf. Accessed 10 June 2019

Stenley D (2006) K-means clustering: a half-century synthesis. Br J Math Stat Psychol 59:1–34

Stucchi M, Meletti C, Montaldo V, Crowley H, Calvi GM, Boschi E (2011) Seismic hazard assessment (2003–2009) for the Italian building code. Bull Seismol Soc Am 101(4):1885–1911

Sugar CA (1998) Techniques for clustering and classification with applications to medical problems. PhD Dissertation, Stanford University, Stanford

Sugar CA, Lenert LA, Olshen RA (1999) An application of cluster analysis to health services research: empirically defined health states for depression from the sf-12. Technical Report, Stanford University, Stanford

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc: Ser B (Stat Methodol) 63(2):411–423

Tsioulou A, Taflanidis AA, Galasso Carmine (2019) Validation of stochastic ground motion model modification by comparison to seismic demand of recorded ground motions. Bull Earthq Eng 17(6):2871–2898

Wagstaff K, Cardie C, Rogers S, Schrödl S (2001) Constrained k-means clustering with background knowledge. ICML 1:577–584

Watson-Lamprey J, Abrahamson N (2006) Selection of ground motion time series and limits on scaling. Soil Dyn Earthq Eng 26(5):477–482

Wilks DS (2011) Statistical methods in the atmospheric sciences. Academic Press, London