16TH EUROPEAN CONFERENCE ON EARTHQUAKE ENGINEERING THESSALONIKI 18 - 21 JUNE 2018

SITE CLASSIFICATION DERIVED FROM SPECTRAL CLUSTERING OF EMPIRICAL SITE AMPLIFICATION FUNCTIONS

Sreeram Reddy KOTHA[1,2], Fabrice COTTON[3,4], Dino BINDI[5]

ABSTRACT

With increasing recorded strong motion data, Ground Motion Prediction Equation (GMPE) developers could quantify empirical site amplification factors ($\delta S2S_s$) as site-specific adjustments, which effectively scale the generic GMPE predictions to precise site-specific predictions. Availing $\delta S2S_s(T)$ vectors ($\Delta S2S_s$) for 588 well-characterized sites recording multiple earthquakes in Japan, we derived a *data-driven* site classification. We first demonstrate that $\delta S2S_s$ is weakly correlated to V_{s30}, confirming V_{s30} to be a poor standalone proxy in capturing linear site-response. As an alternative, we perform *k-mean* spectral clustering of $\Delta S2S_s$ (for T = 0.01s – 2s) to identify 8 site clusters/classes with distinct *mean* site amplification functions ($AF=e^{\Delta S2Ss}$), and a within- cluster site-to-site variability ~50% smaller than the overall dataset variability (ϕ_{S2S}). We conclude that while the *soil-like* clusters can be efficiently distinguished by a combination proxy of V_{s30}-H_{800} ranges, the rock-like clusters can be only distinguished using V_{s10}-H_{800}. Following an evaluation of existing Eurocode 8 scheme, we propose a revised data-driven site classification characterized by bivariate normal distributions of V_{s30}, V_{s10}, H_{800}, and predominant period (T_G) of the site clusters

Keywords: Mixed-effects regression; Ground Motion Prediction Equation; Site classification; Spectral Clustering Analysis; Empirical Site Amplification Functions

1 INTRODUCTION

Current seismic code provisions consider the significant role of local site conditions on earthquake shaking. Their influence is described through appropriate elastic design spectra based on different site categories. The main parameter proposed for soil categorization is the V_{s30}, i.e. the time-averaged shear wave velocity (V_s) in the upper 30 m of the soil profile. This parameter has been introduced by Borcherdt and Glassmoyer (1992) and Borcherdt (1994) as a means to classification of sites for building codes. For example, Eurocode 8 (EC8 Code (2005)) recommends a site classification based on V_{s30}, and two families of spectral shapes depending on the seismic activity level of area (Type I for active areas, and Type II for moderately active areas).

A number of authors (Castellaro et al. (2008), Kokusho and Sato (2008), Lee and Trifunac (2010), Héloïse et al. (2012)) drew attention to the limitations of V_{s30} parameter, which is only a proxy and cannot describe alone the physics of site amplification across a broad period (or frequency) range. Alternative proxies were proposed, coupling information on the shallow impedance and the overall sedimentary thickness. Several recent studies aimed at developing new and more refined site classification

[1]Doctoral candidate, Section 2.6, Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany, sreeram@gfz-potsdam.de

[2]Doctoral candidate, University of Potsdam, Institute of Earth and Environmental Sciences, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany

[3]Head of section, Section 2.6, Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany, fcotton@gfz-potsdam.de

[4]Professor, University of Potsdam, Institute of Earth and Environmental Sciences, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany

[5]Senior researcher, Section 2.6, Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany, bindi@gfz-potsdam.de

schemes taking into account these additional information (e.g., Cadet et al. (2008), Gallipoli and Mucciarelli (2009), Luzi et al. (2011)). For example, Pitilakis et al. (2013) introduced a more refined classification using H_{800} (depth to seismic bedrock with V_s = 800m/s), $V_{s,av}$ (average shear-wave velocity of the soil column) and fundamental period (f_0). In total, Pitilakis et al. (2013) suggested 12 site classes for the two European seismicity classes (Type I and Type II). Defining new classifications schemes is however highly challenging because of a few technical issues:

- Only a minimum *sufficient* number of classes is desirable. The optimal choice of the number of classes is however difficult to define
- Only few studies (e.g., Derras et al. (2016)) tested the relative efficiency of the various site conditions proxies (e.g., H_{800}, f_0, and V_{s30}) to predict soil amplifications
- Site class definitions should avoid unphysical discontinuities in amplification coefficients at the boundaries of adjacent classes. However, such discontinuities are to be expected when using discrete site classes, as opposed to continuous functions of site-response proxies

To resolve some of these issues we explore a new approach to deriving a site classification and site amplification functions. Our aim is to develop a *data-driven* classification scheme with minimal a priori conditions. For this purpose, we adopt the following steps:

1. We take advantage of a high-quality dataset featuring several well-characterized sites recordings multiple earthquakes in a region. In this study, we use the KiK-net dataset built by Dawood et al. (2016), consisting of 1164 shallow crustal events recorded at 644 sites with several geotechnical site parameters available – e.g. V_{s30} and H_{800} values have been directly derived from down-hole measurements of V_S profile
2. The empirical site amplification factors are products of a Ground Motion Prediction Equation (GMPE) mixed-effects analysis. We develop a site-specific GMPE from the selected strong motion dataset as suggested in Kotha et al. (2017a). For the sake of brevity, readers are referred to Kotha et al. (2018) for an extended detail on the dataset, GMPE development and mixed-effects analysis. Here we present only the site classification part of Kotha et al. (2018)
3. The site amplification factors obtained in the above step are subject to spectral clustering analysis to identify sites with similar linear soil response. An optimal number of classes is chosen to minimize the site-to-site variability within each site cluster/class and maximize the dissimilarity of mean amplification functions across clusters
4. In the final step, we check the compatibility of various site-response proxies in explaining the site clusters obtained in the third step. Site-response proxies (H_{800}, V_{s30}, V_{s10}) are not used *a priori* to define the classes, but *posteriori* to characterize the statistical clustering of site-response. We then introduce a revised site classification scheme, mean site amplifications associated with each class, and site-to-site variability of amplification within each site class

2 DATA

In this study, we use the Kiban-Kyoshin network Okada et al. (2004) database compiled by Dawood et al. (2016) for ground motion studies. A step-by-step automated protocol used to systematically process about 157,000 KiK-net strong ground motion recordings obtained between October 1997 and December 2011 is elucidated in Dawood et al. (2016) and related appendices. A *flatfile* with all the metadata and the 5% damped pseudo spectral acceleration (PSA) of the processed records is uploaded to NEEShub (https://nees.org/resources/7849).

In addition to the waveform processing by Dawood et al. (2016), we make a more GMPE specific record selection for the mixed-effects regression (see Kotha et al. (2018) for further details). We choose only the surface recordings at sites with measured V_{s30} available, from Active Shallow Crustal (ACRsh) events with F-net reported hypo-central depth $\leq$ 35km. In doing so, the number of usable records for the GMPE regression at T = 0.01s drops from 157,000 to 15,896. The number of usable records further decreases to 6462 at T = 2s. Figure 1 illustrates the distribution of data used in the GMPE development. In all there are 850 events with $3.4 \leq M_W \leq 7.3$, 641 sites with $106 \leq V_{s30} \leq 2100$m/s, and 15,896 records with $0 \leq R_{JB} < 600$km, which also defines the applicability range of the GMPE.

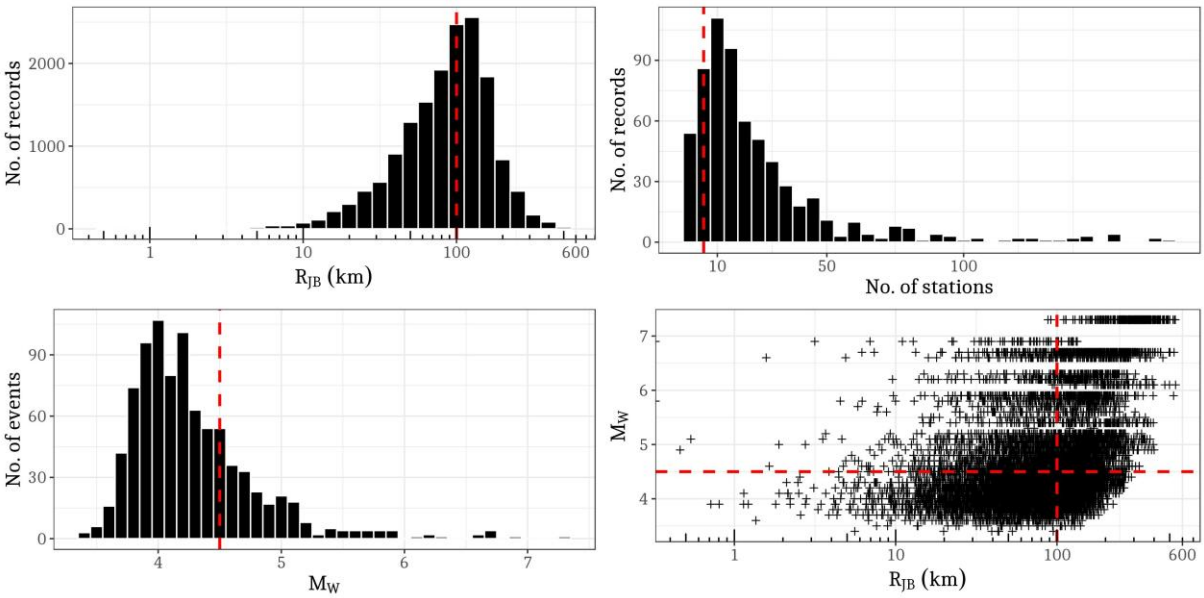

Figure 1: Data distribution following the record selection criteria for GMPE regression at T = 0.01s: (top-left panel) Distance distribution of usable records, (top-right panel) number of records per station, (bottom-left panel) magnitude distribution of usable records, (bottom-right panel) magnitude - distance scatter plot of usable records.

3 GROUND MOTION PREDICTION EQUATION

As in Kotha et al. (2016), we perform mixed-effects GMPE regressions using the LMER algorithm of Bates et al. (2014) implemented in R by Team (2013). Eq. (1) shows the structure of the GMPE derived to predict the geometric-mean of 5% damped horizontal Pseudo Spectral Acceleration (PSA) at 33 values of T between 0.01s and 2s. The parametric functions, $f_R(M_w, R_{JB})$and $f_M(M_w)$, which model the scaling of PSAs with distance (R_{JB}) and magnitude (M_w) at each T, are the GMPE fixed-effects. In the multi-step GMPE regression, we first calibrate the fixed-effects, and then split the residual $\varepsilon = \ln(PSA) - f_R(M_w, R_{JB}) - f_M(M_w)$ into random-effects δB_e and $\delta S2S_s$, and residual $\delta WS_{e,s}$.

$$ln(PSA) = f_R(M_w, R_{JB}) + f_M(M_w) + \delta B_e + \delta S2S_s + \delta WS_{e,s} \quad (1)$$

δB_e is the between-event random-effect quantifying the systematic deviation of observed ground motions associated to an event *e* with respect to the GMPE fixed-effects prediction. $\delta WS_{e,s}$ is the regression residual capturing record-to-record variability, for event *e* and station *s*. $\delta S2S_s$ is the site-specific random-effect for a site *s*, used to scale the GMPE prediction to a site-specific prediction (e.g., Rodriguez-Marek et al. (2013), Kotha et al. (2017a)). The period-dependent random-effects and the residuals follow orthogonal normal distributions as $\delta B_e = \mathcal{N}(0, \tau)$, $\delta S2S_s = \mathcal{N}(0, \phi_{S2S})$ and $\delta WS_{e,s} = \mathcal{N}(0, \phi_0)$, where τ is event-to-event or between-event variability, ϕ_{S2S} captures the site-to-site or between-site variability, and ϕ_0 is the event-and-site corrected or residual aleatory variability. Note that the ϕ_0 in this study is the same as the single-station standard deviation ϕ_{ss} of Rodriguez-Marek et al. (2013). The total aleatory variability of the dataset with respect to a GMPE is $\sigma = \sqrt{\tau^2 + \phi_{S2S}^2 + \phi_0^2}$.

4 EMPIRICAL SITE AMPLIFICATION FUNCTIONS: $\Delta S2S_S$

Figure 2 is the customary residual analysis performed after a GMPE regression to verify if the fixed-effects components capture well the attenuation of PSAs at all magnitudes and distances. In the top panels of Figure 2, we plotted δB_e versus M_w to evaluate the performance of $f_M(M_w)$. We divide the magnitude range M3.4 – M7.3 into 10 magnitude bins, and calculate the mean and 15th-85th percentile error bars on δB_e within each bin. At all periods, the mean δB_e for each bin falls very close to zero, implying no significant trend with M_w and that $f_M(M_w)$ captures the magnitude scaling of PSAs very

well. The bottom panels of Figure 2 show the event-and-site corrected residuals, $\delta WS_{e,s}$ versus the distance metric, R_{JB}. We recall that $f_R(M_w, R_{JB})$ is regressed for data with 0km ≤ R_{JB} < 600km, which is a considerably larger distance range than any GMPE developed for Active Shallow Crustal environments. Such a modeling choice is motivated by our need to minimize the estimation errors of $\Delta S2S_s$, with a large amount of data. Regardless, the $f_R(M_w, R_{JB})$ performs very well at all distances, as indicated by the zero mean $\delta WS_{e,s}$ within each distance bin.

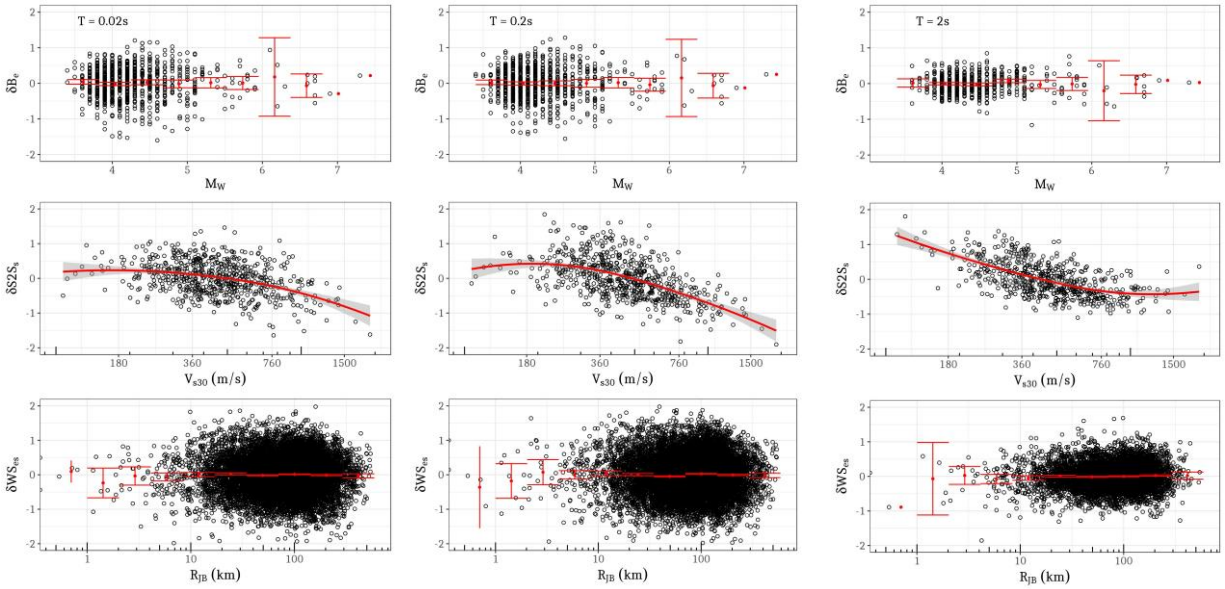

Figure 2: Random-effects and residual plots for GMPE evaluation at T = 0.02s, 0.2s, and 2s. In each panel, δB_e is plotted against M_W, $\delta S2S_s$ against V_{s30}, and $\delta WS_{e,s}$ against R_{JB} to check if random-effects and residuals show a systematic trend with predictor variables

Note that eq. (1) does not include any site-response component in its fixed-effects, unlike the standard practice of including a parametric function of V_{s30} among the fixed-effects of a GMPE (e.g. Bindi et al. (2017)). Instead, the site-specific random-effects $\delta S2S_s$ are set to *absorb* all the site-specific response (as in Kotha et al. (2016)). The middle panels of Figure 2 showing $\delta S2S_s$ (at T = 0.02s, 0.2s, 2s) versus V_{s30} (in log-scale) is the most important plot of this section. Since a site-response component is deliberately left out of the fixed-effects in GMPE, the random-effects $\delta S2S_s$ show a trend with V_{s30} (unlike δB_e and δWS_{es}). However, at T = 0.02s (and T = 0.2s), the gradient of LOESS fit (Local regression of scatterplots by Cleveland (1979)) for $\delta S2S_s$ versus ln(V_{s30}) is close to zero at V_{s30} < 600m/s, implying that high frequency soil response is very weakly correlated to V_{s30} (also in Seyhan and Stewart (2014)). For longer periods (T = 2s), although a steeper gradient at V_{s30} < 200m/s indicates some relevance of V_{s30} as site-response proxy, it appears that low frequency response of stiffer soils cannot be captured with V_{s30} alone. Our observations suggest that V_{s30} may not be an ideal proxy for the strongly period-dependent linear site-response, implying a site classification scheme based solely on V_{s30} needs to be replaced. $\delta S2S_s$ are site-specific period-dependent amplification factors, therefore, vectors of $\delta S2S_s$ for T = 0.01s – 2s resemble empirical site-specific amplification functions ($\Delta S2S_s$). $\Delta S2S_s$ as in this study can be used to adjust the generic GMPE PSA response spectra (e.g., Rodriguez-Marek et al. (2013), Kotha et al. (2017a)) and Conditional Spectra (e.g. Kotha et al. (2017b)), to yield their site-specific counterparts. As a proposed alternative, instead of modelling period-dependent relationship between $\delta S2S_s$(T) and V_{s30}, in this study, we use the vectors $\Delta S2S_s$, capturing the site-response in the entire range T = 0.01s – 2s, to develop a new empirical site-response classification. However, given that the large fraction of data used in constraining $\Delta S2S_s$ is from events with M_w < 5 and R_{JB} > 25km (as seen in Figure 1), these empirical amplification functions capture only the *average* linear soil response at a site. To constrain the non-linear soil response in $\Delta S2S_s$ would require more data from larger and closer events.

5 SITE CLASSIFICATION

With empirical site amplification functions $\Delta S2S_s$ of 641 sites available, the next step is to develop a data-driven site classification scheme to classify sites with similar linear soil response under seismic excitation, i.e., similarity of their $\Delta S2S_s$. The sequence of steps involves: verifying if the data indeed shows a tendency for classification, the degree of classification in terms of an optimal number of classes, the amount of variability that can be explained with classification, and finally, if the statistical classification is physically meaningful. In a preliminary check, a h-index of 0.87 indicated a strong clustering tendency of the $\Delta S2S_s$ dataset, when 0 is for uniform distribution and 1 for highest degree of clustering Hopkins and Skellam (1954). Therefore, we pursue a data-driven site classification based on spectral clustering of $\Delta S2S_s$

5.1 Spectral clustering analysis

For the purpose of identifying clusters of similar amplification functions, we adopt the spectral clustering analysis – a type of unsupervised machine learning aimed at extracting hidden patterns/structures from large amounts of unlabeled multidimensional data. In this case, the $\Delta S2S_s$ vectors are the multidimensional data points, site clusters/classes are the hidden structures, and *mean* amplification functions of site classes are the patterns that characterize these structures. The steps involved in spectral clustering are as follows:

1) **Preparing the data:** $\Delta S2S_s$ vectors of all the sites to be clustered must be of equal length, therefore we only select the 588 sites (of the 641 sites with measured V_{s30}) with $\delta S2S_s$ available at all periods in the range T = 0.01s -2s. $\Delta S2S_s$ vectors of the 588 sites are then normalized with the period-dependent ϕ_{S2S}
2) **Choice of clustering technique:** There are several advanced clustering techniques depending on the amount of supervision (a priori information) that is input and the knowledge that is being queried. We chose a basic partitioning algorithm: the k-means clustering technique by MacQueen (1967). The clustering algorithm (available in the R library as *ClusterR*) by Lampros Mouselimis (2017) is applied on the 588 $\Delta S2S_s$ vectors to partition them into k clusters such that the Total Within Sum of Squares (WSS) of clusters is iteratively minimized
3) **Selection of the number of clusters k:** We use two indices to guide the selection of optimal *k*: Total Within Sum of Squares (WSS) and the Gap statistic that compares the WSS change with that expected under an appropriate null reference distribution of the data (see Tibshirani et al. (2001) for more details on this statistic). After testing different selections for the number of clusters, we found that *k*=8 provides an acceptable WSS reduction without introducing large overlaps among the clusters (Figure 3).

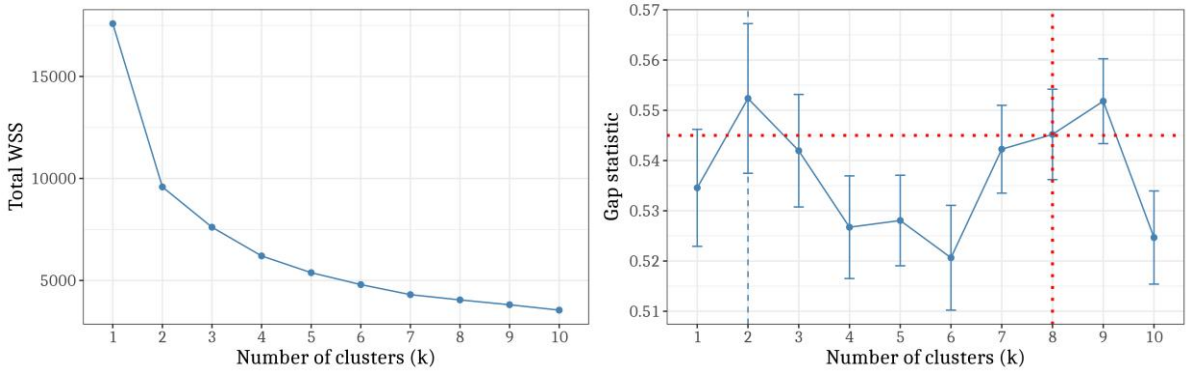

Figure 3: Optimal number of clusters based on Total Within Sum of Squares (WSS, left panel) and Gap statistic (GS, right panel). The WSS metric reduces with increasing number of clusters, but the optimal number of clusters is when Gap statistic is maximized – in which case the WSS is low and the inter-cluster distance is high.

In a way our approach is inverse of the current practice, where the number of site classes (e.g. 5 soil

classes - A, B, C, D, E in EC8, preferred site-response proxy (e.g. V_{s30}), and parametric ranges of selected proxy (e.g. sites with V_{s30} > 800m/s as EC8 Class A) are fixed a priori – and then, the available strong motion data is grouped and processed within each class to derive empirical site amplification functions. In this approach, we first derived the empirical site amplification functions ($\Delta S2S_s$) of the 588 sites, and then classified them into 8 k-means clusters. We now present the site clusters and their mean amplification functions. Later, we investigate and identify site-response proxies that can effectively characterize these eight site classes.

5.2 Site clusters

The eight site clusters partitioning the 588 sites in the dataset are visualized in Figure 4, and the number of sites in each cluster along with within-cluster sum of square (WCSS) are provided in Table 1. In the left panel is the 2D visualization of the k-mean clusters. Regarding the two dimensions, the visualization algorithm performs a principal component analysis (PCA) in which the higher dimensional $\Delta S2S_s$ vectors are reduced to two principal dimensions Kassambara and Mundt (2016). The distance along each dimension can be interpreted as how similar or dissimilar are any two cluster means. For instance, cluster 6 is farthest from cluster 8 along Dim1, and is closest to cluster 7. To interpret this separation, we refer to the more familiar plot in the right panel of Figure 4.

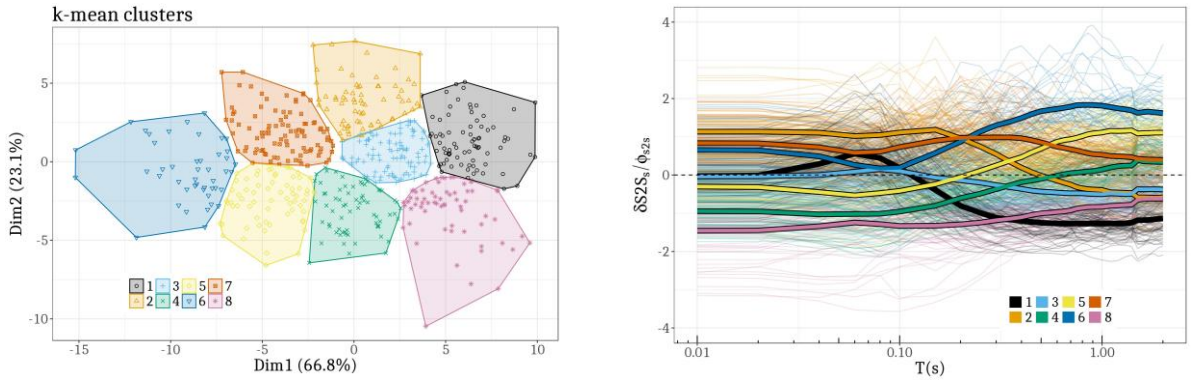

Figure 4: (left panel) Visualization of k-mean clustering, where each polygon is a cluster and each point within is a site ($\Delta S2S_s$). Dim1 and Dim2 are variables derived from a Principal Component Analysis of $\Delta S2S_s$ vectors, which together describe 89.9% of data variability. (right panel) Normalized $\Delta S2S_s$ of 588 sites in thin lines, and cluster-specific normalized mean $\Delta S2S_s$ overlaid as thick lines – color coded according to cluster number

Normalized $\Delta S2S_s$ vectors of the 588 sites in the dataset are plotted in the right panel of Figure 4. Each thin translucent lines corresponds to a single site, while the thick overlaid lines represent the cluster-specific mean normalized $\Delta S2S_s$ vectors, for the period range T = 0.01s – 2s. These are used to develop the empirical site amplification functions associated with the site clusters/classes derived in this study. Observing the two plots in Figure 4: the mean normalized $\Delta S2S_s$ for cluster 8 is well below zero for the entire period range. While cluster 7, which is diagonally the farthest from cluster 8 in the left panel of Figure 4, shows the opposite behavior. The same logic can be applied to cluster 1 and 5, and so on. The following sections presents the practicality of the clusters as site classes.

Table 1: Number of stations within each cluster and within-cluster sum of squares (WCSS)

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**
No. of Sites	78	68	101	69	66	45	95	66
WCSS (%)	12	12	13	10	11	12	19	12

5.3 Site amplification functions: Mean and variability

It is customary to present site amplification functions for different site classes with respect to the reference site conditions. In EC8, the reference site conditions are characterized as outcropping rock sites with V_{s30} = 800m/s. In this study, we select the reference site cluster as the one with a relatively *low and*

flat mean $\Delta S2S_s$ vector. In the right panel of Figure 4, cluster 8 shows $\Delta S2S_s$ ideal to be qualified as reference site conditions, since it shows no selective amplification of any period ranges with respect to other sites in the dataset. Note that, until this point we set no a priori criterion on reference site geotechnical conditions (in terms of V_{s30} or other parameters). In the left panel of Figure 5, we show the empirical site amplification functions of the other seven non-reference site conditions with respect to cluster 8. The amplification functions in this plot are estimated from following steps:

1) The normalized $\Delta S2S_s$ of Figure 4 are scaled back to their original random-effect estimates by multiplying them with period-dependent between-site standard deviations ϕ_{S2S}
2) The de-normalized $\Delta S2S_s$ vectors are scaled with respect to the reference cluster 8, so that the reference cluster 8 $\Delta S2S_s$ vector is now a null vector
3) Since the $\Delta S2S_s$ are additive random-effects of a mixed-effects GMPE in natural-log scale, the multiplicative amplification functions would be $AF=e^{\Delta S2S_s}$. In this step, the amplification function of reference cluster 8 becomes a unit vector. For example, if the GMPE predicted PSA(1s) for the reference cluster 8 is 0.1g, and the (multiplicative) amplification factors for cluster 1 through 7 are [0.75, 1.25, 1.25, 1.75, 3.00, 4.50, 2.00], the scaled ground motions would be [0.08g, 0.13g, 0.13g, 0.18g, 0.3g, 0.45g, 0.2g] respectively

The right panel of Figure 5 compares the within-cluster site-to-site variability ($\phi_{S2S,c}$) against the pre-clustered between-site (ϕ_{S2S}) variability of the dataset (for this GMPE). The average reduction in site-to-site variability is approximately 50% with respect to the dataset value, while the reduction for a few clusters in at longer period ranges is larger (up to 70%). Such reduction in variability has a dramatic effect on total standard deviation (σ) and thereby the PSHA based hazard estimates.

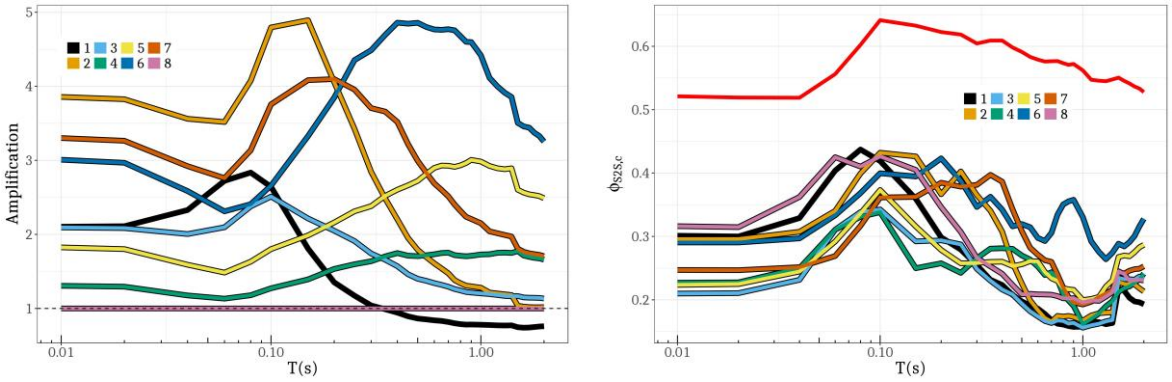

Figure 5: (left panel) Site amplification functions of each cluster scaled with respect to the reference site conditions: cluster 8. (right panel) The within-cluster site-to-site variability $\phi_{S2S,c}$ of the 8 clusters compared to the overall GMPE ϕ_{S2S} prior to clustering (red curve).

Looking at the amplification functions in Figure 5, it is rather clear that the spectral clustering technique distinguishes sites based on their peak amplification period (T_G, analogue to H/V spectral ratio based predominant period of Zhao et al. (2006)) and amplification level at T_G. However, the within-cluster between-site variability $\phi_{S2S,c}$, also reaches its maximum value around its T_G, indicating a large variability in its amplification. Such correspondence between peak amplification and peak variability through T_G was reported in Zhao et al. (2006) for K-Net sites, and for EC8 classification in Cauzzi and Faccioli (2017). The cause for such a parallel is the primary limitation of discrete site classification, where sites with similar T_G but very different AF at T_G are grouped together, resulting in a large site-to-site variability of amplification. In addition, a generic high variability is observed at T = 0.1s, which can be partially attributed to the non-linear transformation from Fourier spectra (frequency domain) to response spectra via convolution with a single-degree-freedom oscillator transfer function (discussed in Stafford et al. (2017) and Kotha et al. (2017b)). Decreasing ϕ_{S2S} trend is observed on either sides of T = 0.1s, but this can be resolved only in Fourier domain, and not with the response spectra used in this study.

5.4 Site-response proxies

Using the empirical site amplification functions and cluster-specific $\phi_{S2S,c}$, the GMPE can be adjusted to predict site class dependent ground motions in hazard assessment; the missing link is *sufficient and*

*efficien*t site response proxies to classify new sites. Dawood et al. (2016) provided the time-averaged shear wave velocity at depth z (m), $V_{s,z}$ for z = 0,5,10,20,30,50,100, borehole depth, and H_{800} –depth to seismic bedrock with V_s = 800m/s. In this study, we attempt characterizing the cluster amplification functions of Figure 5 using only the geotechnical site-response parameters available in the dataset. In process of developing a new site classification scheme, we first evaluate the eight site clusters against the site classification schemes defined in the EC8.

Figure 6 categorizes the clusters based on their distribution of H_{800}, V_{s10} and V_{s30}. For visual guidance, both panels provide guiding lines at H_{800} = 5, 10, 20, 30, and 100m. In the left panel, the x-axis marks the bounding V_{s30} values of EC8 site classification at 180, 360 and 800m/s, along with the revised V_{s30} ranges prescribed in this study. In the right panel, the x-axis is divided at V_{s10} = 200, 300, 400 and 600m/s. We used Figure 6, to evaluate the physical meaning of the clusters, and also to define new site classes in Table 2.

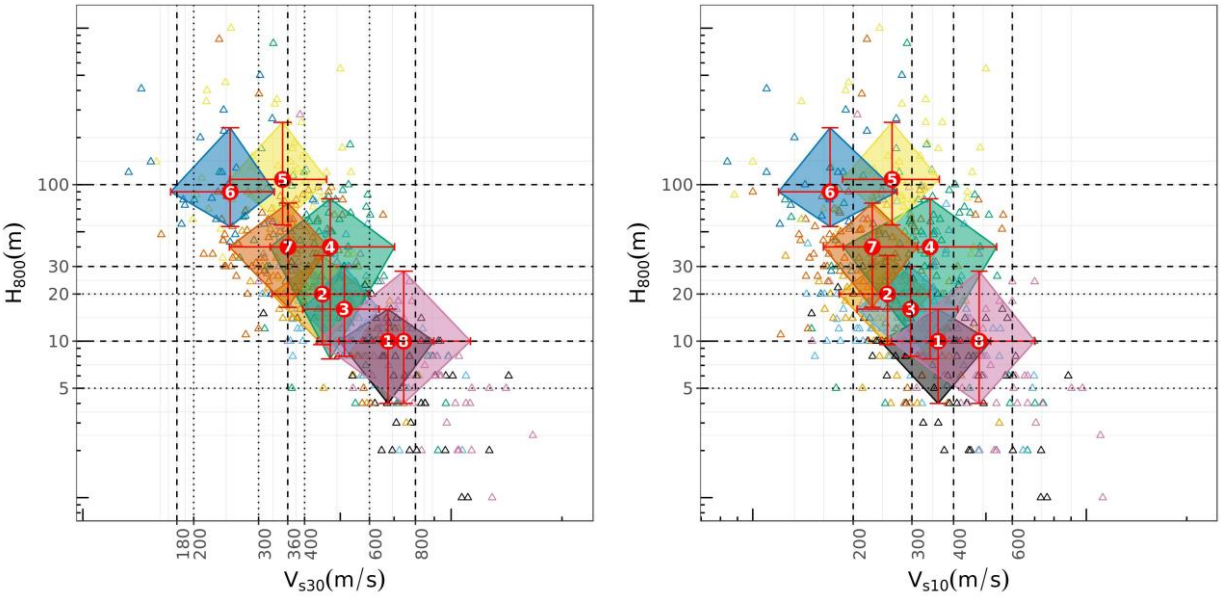

Figure 6: Evaluation of site response proxies in characterizing site clusters: In the left panel is the combination of H_{800} and V_{s30}, and in the right panel, H_{800} and V_{s10}. The polygons bound the 15th-85th percentile ranges of geotechnical parameters of sites within each cluster, while the center (red) marks the 50th percentile value

The colored polygons of Figure 6 represent the 15th-85th percentile ranges of V_{s30}, V_{s10} and H_{800} of each cluster, with the central red markers at 50th percentile values. These plots are only meant to visualize the spread of geotechnical parametric ranges for each cluster, and to compare them against ranges prescribed in EC8. However, the ranges in Table 2 are determined from high-density contours of bivariate normal distribution of the parameters (2D kernel density distribution) used to identify the representative ranges of V_{s30}, V_{s10}, and H_{800} in Kotha et al. (2018) site classification. In addition, we provided a predominant period (T_G) ranges inferred from the peaks in amplification functions of Figure 5.

Table 2: Site cluster characterization based on V_{s10} - V_{s30} - H_{800} ranges

Site Cluster	T_G (s)	V_{s30} (m/s)	V_{s10} (m/s)	H_{800} (m)	EC8
C5	> 1s	300 – 450m/s	200 – 300m/s	> 50m	C + B
C4		400 – 600m/s	300 – 400m/s	30 – 100m	B
C6	0.4 – 1s	200 – 300m/s	< 200m/s	> 50m	C
C7	0.2 – 0.4s	200 – 450m/s	200 – 400/s	30 – 100m	C + B
C3	0.1 – 0.2s	450 – 600m/s	200 – 400m/s	10 – 30m	B
C2		300 – 600m/s	150 – 350m/s		C + B

C1	< 0.1s	450 – 600 m/s	200 – 600m/s	5 – 20m	E
C8	-	> 600m/s	> 600m/s	< 5m	A

The purpose of this exercise was to identify a combination of geotechnical parameters to classify the sites which were otherwise grouped into the same site classes of EC8. From Figure 6, based on geotechnical parametric ranges, we notice that majority of the clusters can be classified into either EC8 class B or class C. More detailed inferences follow:

- Cluster 4 and 5 are constituted of KiK-net sites with 300m/s < V_{s30} < 600m/s and 150m/s < V_{s10} < 400m/s, showing a broad amplification *plateau*, increasing towards longer periods. These two clusters can be set apart from other clusters based on their large T_G. Between cluster 4 and 5, the distinction can be made based on their V_{s10}, V_{s30} and H_{800}. Cluster 4 is nested within class B of EC8, while cluster 5 is on the border between EC8 class B and C. Both these clusters can be distinguished from other similarly placed clusters (according to V_{s30}) based on their stiffer V_{s10} and deeper H_{800} ranges
- Cluster 6 and 7 can be distinguished from other clusters based on their well-defined T_G ranges. Cluster 6 shows a much higher amplification with respect to cluster 5, despite similar H_{800}, due to its systematically lower V_{s30} and V_{s10} ranges. Similarly, cluster 7 despite having the same H_{800} range as cluster 4, shows much higher amplification at a lower T_G due to its softer soil profile – characterized by lower V_{s30} and V_{s10} ranges. Interestingly, cluster 6 and 7 are hard to be distinguished based on their V_{s30} ranges alone. In which case, T_G, V_{s10} and H_{800} as site-response proxies perform significantly better, proving a case against V_{s30} as a standalone proxy. Cluster 6 is nested within class C of EC8, while cluster 7 is on the border between EC8 class B and C. Both these clusters can be distinguished from other similarly placed clusters (according to V_{s30}) based on their softer V_{s10} and deeper H_{800} ranges
- Cluster 8 serves as the reference site cluster, with the highest values of V_{s30} and V_{s10}. Sites in this cluster showed no clear peak amplification (left panel of Figure 5). These sites resemble best the class A of EC8 with V_{s30} > 800m/s
- Cluster 1 with T_G < 0.1s (left panel of Figure 5), is the only one showing a strong amplification at high frequency and de-amplification towards longer periods, with respect to reference site cluster 8. This cluster resembles the class E of EC8 with soft sediment layer of thickness 5 – 20m, overlaying a stiffer soil. Interestingly, Cluster 1 and 8 can be only distinguished based on their H_{800} and V_{s10} ranges, but not V_{s30} – suggesting V_{s10} as a better proxy of high frequency site-response
- Cluster 2 and 3 are separated from the closest resembling cluster 1, based on their longer T_G and larger H_{800} values. However, these two clusters do not appear to differ in their $V_{s10} – V_{s30} – H_{800}$ ranges, as much as they do from other clusters. Given their identical T_G ranges, but radically different amplification levels, we suspect these clusters to differ in the velocity contrast of their soil profiles. A higher impedance contrast results in significantly higher amplification at T_G (see Figure 5), which appears to be the case considering the relatively lower V_{s30} and V_{s10} ranges of cluster 2 against cluster 3. Shear-wave velocity profiles, and additional geotechnical parameters might help in better characterizing the differences among cluster 2 and 3, and cluster 1 as well.

6 DISCUSSION AND CONCLUSIONS

In this study we introduce an approach to site classification derived from cluster analysis of empirical site amplification functions. The resulting site classification is aimed to be simple, robust, and data-driven with minimal a priori constrains in terms of relevant site-response parameters. The fundamental requirement for such classification was to derive statistically well-constrained empirical site adjustment functions ($\Delta S2S_s$ vectors). As a first step, we selected a rich dataset featuring several well-characterized sites recording many earthquakes in a region; the KiK-net dataset by Dawood et al. (2016). The next step was to fit a mixed-effects GMPE, whose site-specific random-effects ($\delta S2S_s$) for periods T = 0.01s – 2s constitute the $\Delta S2S_s$ vectors. Given the critical importance of GMPE in our approach, it was necessary that its magnitude and distance scaling fixed-effects components are calibrated very well for a wide magnitude range $3.4 \leq M_W \leq 7.3$, and large distance range $0 \leq R_{JB} < 600$km, so that the estimated

$\Delta S2S_s$ vectors capture predominantly linear soil response and have a low estimation errors. Our primary inference is that $\delta S2S_s$ show a weak correlation relation with V_{s30}. For this dataset, the site-response is not efficiently captured by V_{s30} particularly at short – moderate periods, and is highly variable even for sites with identical V_{s30}. We therefore attempted an alternative approach to classifying sites.

The $\Delta S2S_s$ vectors are site-specific terms filtered for event and path effects with a robust GMPE median, and hence serve as empirical site amplification functions in this study. We chose the k-mean clustering technique to reduce the higher dimensional $\Delta S2S_s$ vectors of 588 sites into 8 clusters (of sites) with similar linear soil response under seismic action. These 8 clusters serve as the site classes in our new classification scheme.

Site amplification functions are usually presented as scaling functions with respect to the reference site conditions. Traditionally, outcropping rock sites with $V_{s30} > 800$m/s are considered as a reference sites. The $\Delta S2S_s$ vectors do not presume any reference site conditions, but instead are additive random-effects (scalar adjustments) to the GMPE fixed-effects median. Technically, the $\Delta S2S_s$ vectors are site-specific deviations from a *hypothetical* reference site, whose response is an average of all sites in the dataset, i.e. a site with null $\Delta S2S_s$ vector. However, for engineering purposes, it is necessary to characterize *real* reference site conditions. We therefore select the cluster of sites whose mean $\Delta S2S_s$ vector is *low and flat*. Meaning, sites in this cluster show a systematic de-amplification over the entire period range, with respect to other sites in the dataset. This unique cluster (cluster 8) represents the reference site conditions in our approach. Essentially, we identified a reference site cluster with no amplification, and seven other clusters with unique non-zero site amplification functions. Additional benefit of clustering technique is the significantly smaller within-cluster site-to-site variability, which is on an average ~50% smaller than the pre-clustered, overall site-to-site response variability of the dataset. This in our opinion is a significant improvement in the context of seismic hazard assessment.

For site amplification functions to be applicable at new sites, we need to develop site-response proxies based on which the new sites can be classified. From this point of view, we are limited by the available geotechnical information at the sites. Among the most prevalently used site parameters in the site-response characterization are the predominant period (T_G), time-averaged shear-wave velocity up to 10m (V_{s10}) and 30m (V_{s30}), and the depth to engineering bedrock with shear-wave velocity V_s = 800m/s (H_{800}). The inferences from this part of the study are enumerated below:

1) Multiple clusters show significantly different site amplifications but similar V_{s30} ranges, suggesting that V_{s30} is insufficient as a standalone proxy in site-response classification
2) Classification based on T_G works well in classifying sites at first order. However, it is insufficient in distinguishing sites with identical ranges of T_G , but different amplification levels at T_G
3) For site clusters with $H_{800} > 30$m and $T_G > 0.2$s, both V_{s30} and V_{s10} perform well in distinguishing the four clusters with moderate – long period amplifications. Clusters (5 and 6) with deepest soil profiles ($H_{800} > 50$m) can be distinguished with their T_G, V_{s10} and V_{s30}, where lower soil stiffness (of cluster 5) translated into lower T_G and a much higher amplification. Similar is the case for clusters (4 and 7) with shallower soil profiles ($30\text{m} < H_{800} < 100\text{m}$)
4) For sites with $10\text{m} < H_{800} < 30\text{m}$ and $0.1\text{s} < T_G < 0.2\text{s}$, we identified two clusters with very similar V_{s30} and V_{s10} distribution, but significantly different amplification levels. These clusters cannot be distinguished with the available geotechnical information. A detailed investigation of their shear-wave velocity profiles may help better distinction of these clusters
5) We identified two clusters with $V_{s30} > 600$m/s that can be separated based on their T_G, V_{s10} and H_{800}. Cluster (1) with lower V_{s10} and a higher H_{800} shows a strong amplification at its $T_G < 0.1$s, while the one with higher V_{s10} and lower H_{800} shows a flat response (cluster 8). Evidently, V_{s30} based classification groups these two very different site types into a unique site class. In our opinion, such misclassification leads to a significant bias and a large variability in response of the so-called reference site class (e.g. $V_{s30} > 800$m/s in EC8 classification). We suggest using at least the V_{s10}, or even better - V_s profiles, to characterize reference site conditions

Our approach is beneficial in identifying hidden site classes, resolving site-to-site variability, and developing efficient site classes from a rich dataset. We note that site types, sparsely represented or not represented at all in the dataset, cannot be identified with data-driven techniques presented here. Despite, we present a scope to revise the existing site classification of EC8. Additional sophistication can be introduced in the clustering analysis, to the point where clusters can be hierarchically divided or merged depending on the available site parametrization in a region. Our next step is to implement this method

on a pan-European dataset provided by Lanzano et al. (2017). The number of clusters, the mean and variability of empirical site amplification functions, and even the relevant site-response proxies may depend on the spatial coverage of regional datasets.

7 ACKNOWLEDGEMENTS

We appreciate the efforts of the anonymous reviewer for their review. We would like to thank Prof. John Anderson for his valuable insights in interpreting the results. This research is funded by the SIGMA2 project (EDF, CEA, PG&E, SwissNuclear, Areva, CEZ, CRIEPI) – 2017 - 2021 (http://www.sigma-2.net/)

8 REFERENCES

Bates D, Mächler M, Bolker B, Walker S (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bindi D, Cotton F, Kotha S R, Bosse C, Stromeyer D, Grünthal G (2017). Application-driven ground motion prediction equation for seismic hazard assessments in non-cratonic moderate-seismicity areas. *Journal of Seismology, 21*(5), 1201-1218. doi:10.1007/s10950-017-9661-5

Borcherdt R D (1994). Estimates of site-dependent response spectra for design (methodology and justification). *Earthquake spectra, 10*(4), 617-653.

Borcherdt R D, Glassmoyer G (1992). On the characteristics of local geology and their influence on ground motions generated by the Loma Prieta earthquake in the San Francisco Bay region, California. *Bulletin of the seismological society of America, 82*(2), 603-641.

Cadet H, Bard P Y, Duval A M (2008). *A new proposal for site classification based on ambient vibration measurements and the Kiknet strong motion data set.* Paper presented at the Proceedings of the 14th World Conference on Earthquake Engineering.

Castellaro S, Mulargia F, Rossi P L (2008). VS30: Proxy for seismic amplification? *Seismological Research Letters, 79*(4), 540-543.

Cauzzi C, Faccioli E (2017). Anatomy of sigma of a global predictive model for ground motions and response spectra. *Bulletin of Earthquake Engineering*, 1-19.

Cleveland W S (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association, 74*(368), 829-836.

Code P (2005). *Eurocode 8: Design of structures for earthquake resistance-part 1: general rules, seismic actions and rules for buildings*.

Dawood H M, Rodriguez-Marek A, Bayless J, Goulet C, Thompson E (2016). A Flatfile for the KiK-net Database Processed Using an Automated Protocol. *Earthquake spectra, 32*(2), 1281-1302.

Derras B, Bard P-Y, Cotton F (2016). Site-Condition Proxies, Ground Motion Variability, and Data-Driven GMPEs: Insights from the NGA-West2 and RESORCE Data Sets. *Earthquake spectra, 32*(4), 2027-2056.

Gallipoli M R, Mucciarelli M (2009). Comparison of site classification from VS30, VS10, and HVSR in Italy. *Bulletin of the seismological society of America, 99*(1), 340-351.

Héloïse C, Bard P-Y, Duval A-M, Bertrand E (2012). Site effect assessment using KiK-net data: part 2—site amplification prediction equation based on f 0 and Vsz. *Bulletin of Earthquake Engineering, 10*(2), 451-489.

Hopkins B, Skellam J G (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany, 18*(2), 213-227.

Kassambara A, Mundt F (2016). Package 'factoextra': Extract and Visualize the Results of Multivariate Data Analyses. In.

Kokusho T, Sato K (2008). Surface-to-base amplification evaluated from KiK-net vertical array strong motion records. *Soil Dynamics and Earthquake Engineering, 28*(9), 707-716.

Kotha S R, Bindi D, Cotton F (2016). Partially non-ergodic region specific GMPE for Europe and Middle-East. *Bulletin of Earthquake Engineering, 14*(4), 1245-1263.

Kotha S R, Bindi D, Cotton F (2017a). From Ergodic to Region- and Site-Specific Probabilistic Seismic Hazard Assessment: Method Development and Application at European and Middle Eastern Sites. *Earthquake Spectra, 33*(4), 1433-1453. doi:10.1193/081016eqs130m

Kotha S R, Bindi D, Cotton F (2017b). Site-Corrected Magnitude- and Region-Dependent Correlations of Horizontal Peak Spectral Amplitudes. *Earthquake Spectra, 33*(4), 1415-1432. doi:10.1193/091416eqs150m

Kotha S R, Cotton F, Bindi D (2018). A new approach to site classification: Mixed-effects Ground Motion Prediction Equation with spectral clustering of site amplification functions. *Soil Dynamics and Earthquake Engineering*. doi:10.1016/j.soildyn.2018.01.051

Lampros Mouselimis (2017). ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans and K-Medoids Clustering (Version R package version 1.0.6). Retrieved from https://CRAN.R-project.org/package=ClusterR

Lanzano G, Puglia R, Russo E, Luzi L, Bindi D, Cotton F, D'Amico M C, Felicetta C, Pacor F, WG5 O. (2017). *ESM strong-motion flat-file 2017*. Retrieved from: esm.mi.ingv.it/flatfile-2017/

Lee V W, Trifunac M D (2010). Should average shear-wave velocity in the top 30m of soil be used to describe seismic amplification? *Soil Dynamics and Earthquake Engineering, 30*(11), 1250-1258.

Luzi L, Puglia R, Pacor F, Gallipoli M, Bindi D, Mucciarelli M (2011). Proposal for a soil classification based on parameters alternative or complementary to Vs, 30. *Bulletin of Earthquake Engineering, 9*(6), 1877-1898.

MacQueen J (1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.

Okada Y, Kasahara K, Hori S, Obara K, Sekiguchi S, Fujiwara H, Yamamoto A (2004). Recent progress of seismic observation networks in Japan—Hi-net, F-net, K-NET and KiK-net—. *Earth, Planets and Space, 56*(8), xv-xxviii.

Pitilakis K, Riga E, Anastasiadis A (2013). New code site classification, amplification factors and normalized response spectra based on a worldwide ground-motion database. *Bulletin of Earthquake Engineering, 11*(4), 925-966.

Rodriguez-Marek A, Cotton F, Abrahamson N A, Akkar S, Al Atik L, Edwards B, Montalva G A, Dawood H M (2013). A model for single-station standard deviation using data from various tectonic regions. *Bulletin of the seismological society of America, 103*(6), 3149-3163.

Seyhan E, Stewart J P (2014). Semi-empirical nonlinear site amplification from NGA-West2 data and simulations. *Earthquake spectra, 30*(3), 1241-1256.

Stafford P J, Rodriguez-Marek A, Edwards B, Kruiver P P, Bommer J J (2017). Scenario Dependence of Linear Site-Effect Factors for Short-Period Response Spectral OrdinatesScenario Dependence of Linear Site-Effect Factors for Short-Period Response Spectral Ordinates. *Bulletin of the seismological society of America, 107*(6), 2859-2872.

Team R C (2013). R foundation for statistical computing. *Vienna, Austria, 3*(0).

Tibshirani R, Walther G, Hastie T (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.

Zhao J X, Irikura K, Zhang J, Fukushima Y, Somerville P G, Asano A, Ohno Y, Oouchi T, Takahashi T, Ogawa H (2006). An empirical site-classification method for strong-motion stations in Japan using H/V response spectral ratio. *Bulletin of the seismological society of America, 96*(3), 914-925.