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*, *110*, pp. 318—329.

DOI: http://doi.org/10.1016/j.soildyn.2018.01.051

A New Approach to Site Classification: Mixed-effects Ground Motion Prediction Equation with Spectral Clustering of Site Amplification Functions

Kotha, Sreeram Reddy [1, 2], Cotton, Fabrice [1, 2], Bindi, Dino[1]

[1]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany

[2]University of Potsdam, Potsdam, Germany

Corresponding author: sreeram@gfz-potsdam.de

Abstract

With increasing amount of strong motion data, Ground Motion Prediction Equation (GMPE) developers are able to quantify empirical site amplification functions ($\Delta S2S_s$) from GMPE residuals, for use in site-specific Probabilistic Seismic Hazard Assessment. In this study, we first derive a GMPE for 5% damped Pseudo Spectral Acceleration (g) of Active Shallow Crustal earthquakes in Japan with $3.4 \leq M_w \leq 7.3$ and $0 \leq R_{JB} < 600km$. Using k-mean spectral clustering technique, we then classify our estimated $\Delta S2S_s (T = 0.01s - 2s)$ of 588 well-characterized sites, into 8 site clusters with distinct mean site amplification functions, and within-cluster site-to-site variability ~50% smaller than the overall dataset variability (φ_{S2S}). Following an evaluation of existing schemes, we propose a revised data-driven site classification characterized by kernel density 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 take into account 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-based average value of 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 Code (2005) and Rey et al. (2002) recommend 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)) have drawn 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. A number of other proxies (or combinations of proxies) were proposed, coupling information on the shallow impedance and the overall sedimentary thickness. There are several recent studies aimed at developing new and more refined site classification 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. Ideally the site-to-site variability within each site class should be small compared to a less resolved site classification which, to our knowledge, was not quantitatively analyzed. Moreover, enough recorded strong motion data within each class is seldom available to define statistically well-constrained amplifications factors
- 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. There is often little consensus on the way to choose and combine the site proxies
- 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

In order to resolve some of these issues we explore a new approach to derive a new 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 site parameters available – e.g. V_{s30} and H_{800} values have been directly derived from down-hole measurements of V_S profile. Further description of the data set is provided in the section titled Data

2. The empirical site amplification factors are products of a Ground Motion Prediction Equation (GMPE) mixed-effects analysis. Essentially, we develop a site-specific GMPE from the selected strong motion dataset following the steps described in Rodriguez-Marek et al. (2013) and Kotha et al. (2017a). Details on the GMPE development and mixed-effects analysis is provided in section Ground Motion Prediction Equation
3. The site amplification factors obtained in the second step are subject to spectral clustering analysis to identify sites with similar response. An optimal number of classes is chosen to minimize both: the site-to-site variability within each site cluster/class and the similarity of their mean amplification functions. In section Spectral clustering analysis, we provide a description of the technique and its application
4. In the final step, we check the compatibility of various site-response proxies with 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 *a posteriori* to characterize the statistical clustering of site-response. In section Site classification, we introduce the 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 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 our regression:

1) Dawood et al. (2016) remarked that the hypocentral location and M_w obtained from the F-net catalog are more reliable than the values reported in the KiK-net data files. They matched the KiK-net records to F-net earthquakes and classified the match into five categories (A through E) depending on the error margins on location and M_{JMA}. Category A represents the strictest criteria, Category D contains earthquakes that were manually matched, and Category E contains earthquakes for which no match was found. In our study, we choose only the Category A events, which constitute about 89% of the records
2) While most of the GM records in the dataset correspond to subduction earthquakes, we choose only the Active Shallow Crustal (ACRsh) events classified using the Garcia et al. (2012) algorithm. However, to filter out any subduction intra-slab and deep continental events, we chose only the ACRsh events whose F-net reported hypocentral depth is ≤ 35km (as in the H_{ANSR1} criteria of Garcia et al. (2012))
3) Most of the KiK-net sites provide 3-component recordings at both surface and borehole sites. In our study, we use only choose the surface recordings at sites with measured V_{s30} available
4) Each record is associated with a high-pass corner frequency (f_c) which limits the maximum usable period $T_{max} \leq \frac{1}{f_c}$ of the record in a GMPE regression. Since the dataset is compiled from an automatic recording processing procedure described in Dawood et al. (2016), we applied a more conservative limit of $T_{max} = \frac{0.5}{f_c}$. First, we choose only those event and site combinations for which all the 6-component GMs (at surface and borehole) show a Signal-to-Noise ratio (SNR) ≥3 in the bandwidth

f_c – 30Hz. Then, for regression at each spectral period (T) we select only those records whose $T_{max} \geq T$

5) Finally, we choose only the earthquakes with at least three usable records after all the selection criteria above are cleared. In doing so, the number of usable records for the GMPE regression at T = 0.01s falls from 157,000 to 15,896. The number of usable records further decreases to 6462 at T = 2s. The data distribution for GMPE regression at T = 0.01s is shown in Figure 1. 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 $0 \leq R_{JB} \leq 543$km.

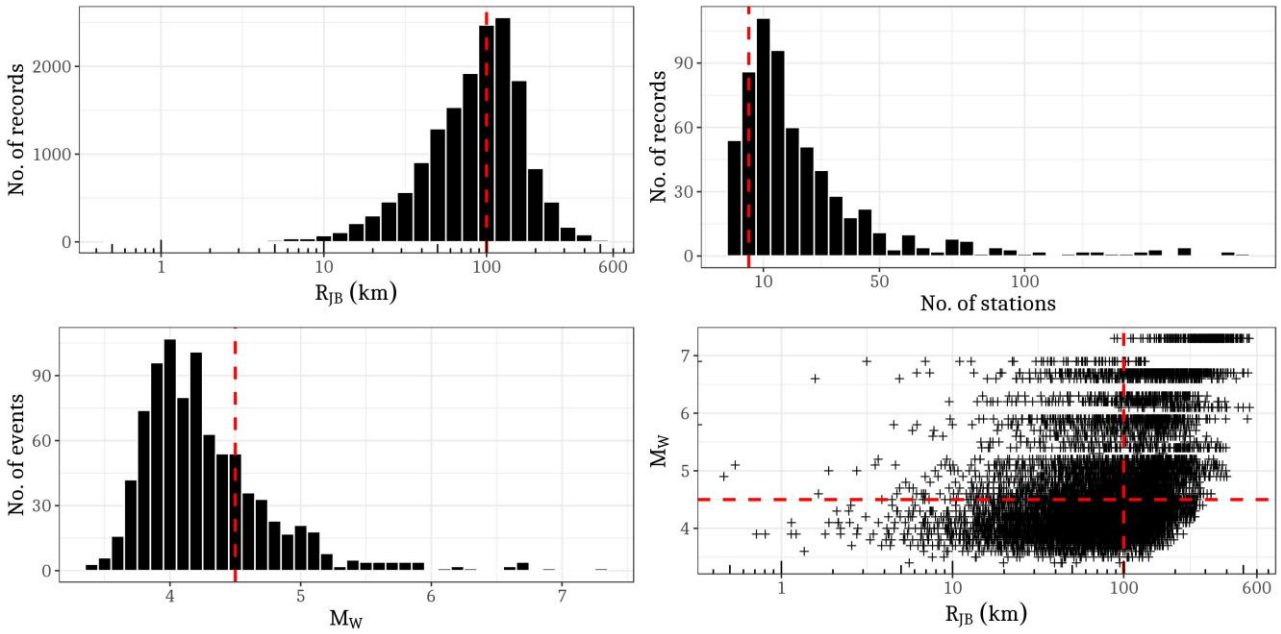

Figure 1: Data distribution following the record selection criteria for GMPE regression at T = 0.01s: (top-left panel) Distance distribution of usable records, (top-right panel) number of records per station, (bottom-left panel) magnitude distribution of usable records, (bottom-right panel) magnitude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

3 Ground Motion Prediction Equation

Using a mixed-effects regression approach (as in Abrahamson and Youngs (1992; Kotha et al. (2016)), we derive a GMPE for the geometric-mean of (5% damped) horizontal Pseudo Spectral Acceleration (PSA) at 33 values of T between 0.01s and 2s.

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

In eq. (1), the parametric functions $f_R(M_w, R_{JB})$and $f_M(M_w)$ capture the scaling of PSAs with distance and magnitude, respectively, and they are referred to as fixed-effects. 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. $\delta S2S_s$ is the site-specific random-effect for a site *s*, which can be used to scale the GMPE prediction to a site-specific prediction (e.g., Rodriguez-Marek et al. (2013), Kotha et al. (2017a)). $\delta WS_{e,s}$ is the regression residual capturing record-to-record variability (combination of *e* and *s*), and can be investigated for other repeatable effects (e.g., Boore et al. (2014), Kotha et al. (2017b), Baltay et al. (2017)). The period dependent random-effects and the residuals follow

orthogonal normal distributions as $\delta B_e = \mathcal{N}(0,\tau)$, $\delta S2S_s = \mathcal{N}(0,\varphi_{S2S})$ and $\delta WS_{e,s} = \mathcal{N}(0,\varphi_0)$, 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 $\sigma = \sqrt{\tau^2 + \Phi_{S2S}^2 + \Phi_0^2}$.

It is worth noting that eq. (1) does not include any site-response component in its fixed-effects, unlike the usual practice of including a parametric function of V_{s30}. The site-specific random effects $\delta S2S_s$ absorb all the site-specific response, and serve as the empirical site-specific amplification functions in our proposed site classification scheme. Given that the large fraction of data used in constraining $\delta S2S_s$ is from events with $M_w < 5$ and $R_{JB} > 25$km (Figure 1), these empirical amplification functions capture only the *average* linear soil response. Therefore, our site classification is solely based on classification linear soil response. To constrain the non-linear soil response in $\delta S2S_s$ would require more data from larger and closer events.

4 Parametric regression

We develop a GMPE following a multi-step approach where we first calibrate the magnitude-dependent distance scaling function $f_R(M_w, R_{JB})$ and then use distance-corrected observations to calibrate the magnitude scaling function $f_M(M_w)$. At each step, we perform a mixed-effects regression using LMER algorithm of Bates et al. (2014) implemented in R Development Core Team (2010), estimating both the δB_e and $\delta S2S_s$ random-effects. In doing so, we ensure that the regression coefficients are unbiased by a well-recorded events or sites (e.g.Boore et al. (2014), Kotha et al. (2017a)).

4.1 Distance scaling: $f_R(M_w, R_{JB})$

The first step in our multi-step regression procedure is to derive the distance scaling component $f_R(M_w, R_{JB})$ of eq. (1). Figure 2 shows the observed PSAs at T = 0.02s, 0.2s and 2s against the Joyner-Boore distance metric (R_{JB}). The scatter plot is color coded according to magnitude ranges of the observations to identify magnitude dependence of distance scaling.

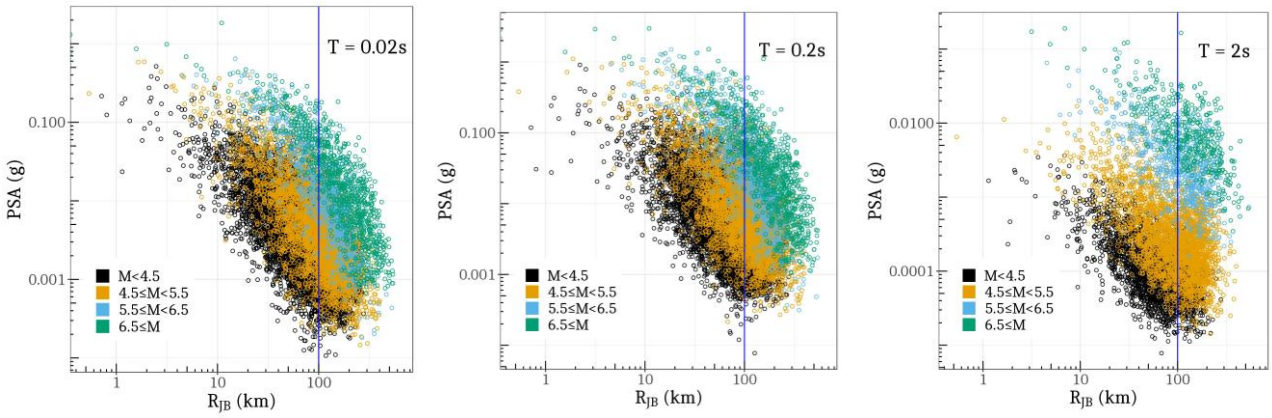

Figure 2: Distance scaling of Geometric Mean of 5% damped horizontal Pseudo Spectral Accelerations at T = 0.02s, 0.2s and 2s.

The distance scaling of PSA shows a magnitude-dependent near-source saturation effect, extending to about 5km for $M_w \leq 4.5$ and up to 20 Km for $M_w \geq 6.5$ events. This effect is generally modeled by

introducing the so-called effective-depth or *h*-parameter in the GMPE distance scaling fixed-effect component. For an equivalent point-source simulation of finite ruptures, we adopted the effective-depth formula (eq. 2) derived by Yenier and Atkinson (2015) from a global dataset of well-recorded earthquakes (including California, Italy, Japan, New Zealand, Taiwan, and Turkey).

$$\ln(h) = 2.303\ (\max[(-0.05 + 0.15M_w), (-1.72 + 0.43M_w)]\ ) \qquad (2)$$

Figure 2 also suggests that short (0.02s) and intermediate (0.2s) period PSAs attenuate more rapidly beyond 100 km. To capture the apparent anelastic attenuation of high frequency PSAs, we introduce a hinge-distance in the definition of $f_R(M_w, R_{JB})$ and model it as in eq. (3). Note that the only magnitude dependence in distance scaling is from h (eq. 2). The coefficients c_1, c_2, and c_3 estimated in this step are held constant for the later steps of GMPE regression.

$$f_R(M_w, R_{JB}) = \begin{cases} c_1 \ln\sqrt{R_{JB}^2 + h^2} & R_{JB} < 100km \\ c_1 \ln\sqrt{100^2 + h^2} + c_2 \ln\left(\frac{R_{JB}}{100}\right) + c_3(R_{JB} - 100) & R_{JB} \geq 100km \end{cases} \qquad (3)$$

4.2 Magnitude scaling: $f_M(M_w)$

The recorded PSAs corrected for $f_R(M_w, R_{JB})$ yield the expected PSA at reference distance R_{ref} = 1km. Per-event averages of distance scaled PSA, hereafter PSA_{Rref}, are the near-source ground motions corrected for attenuation effects. The GMPE magnitude scaling component $f_M(M_w)$ is then derived through a weighted mixed-effects regression of the PSA_{Rref}.

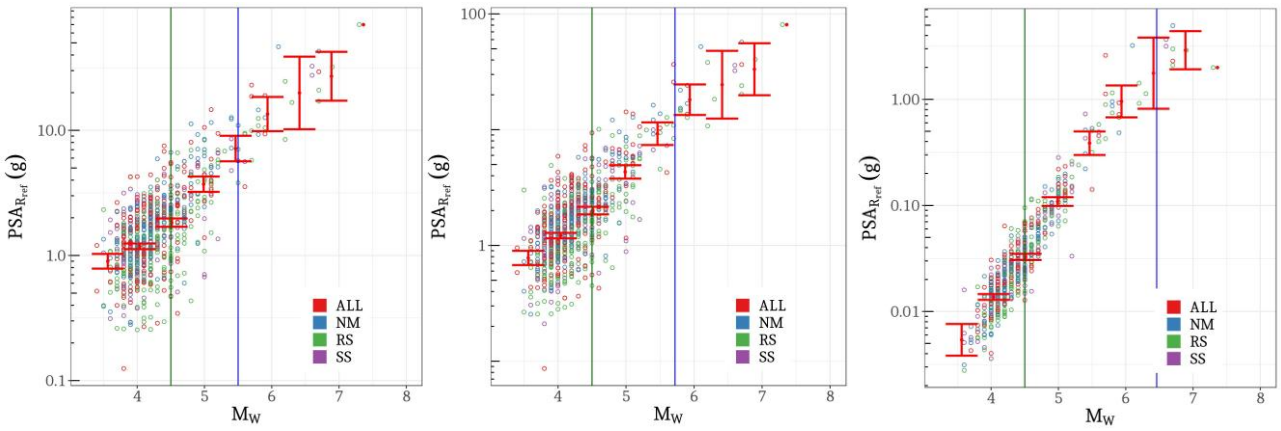

Figure 3: Parametric analysis for magnitude scaling at T = 0.02s, 0.2s, 2s. PSA_{Rref} (g) are the per-event averages of recorded PSAs, corrected for distance scaling (eq. 3). Colors indicate the focal mechanism of events: ALL – unknown mechanism, NM – Normal faulting, RS – Reverse-slip fault, SS – Strike-slip fault.

Figure 3 shows PSA_{Rref} against M_W, color coded according to the focal mechanism of the events. Earlier, Zhao et al. (2016) reported that ACRsh events with Normal (NM) focal mechanism produce higher amplitudes compared to Strike-slip (SS) and Reverse (RS) events. However, our non-parametric analysis showed no clear distinction of observed PSA_{Rref} with rupture focal mechanism. In addition, our parametric mixed-effects regression showed statistically insignificant variability of magnitude scaling with focal mechanism. Therefore, in this GMPE, we chose to not include any focal mechanism term.

Several recent GMPEs developed for applicability over a wide magnitude range (e.g. 3 < M_w < 8), adopted piece-wise linear (or a high degree polynomial) expressions in their magnitude scaling components. For example, Kenneth W. Campbell and Bozorgnia (2014) allowed three period-independent break-points in their magnitude scaling relation at M_w = 4.5, 5.5 and 6.5, Boore et al. (2014) used a single period-dependent magnitude break-point ranging from M_w = 5.5 for T ≤ 0.1s up to M_w = 6.2 for T ≥ 0.4s, while Zhao et al. (2016) used a single period-independent break-point at M_w = 7.1. Piece-wise magnitude scaling expressions allow variability of magnitude scaling gradient in different M_w ranges. Especially for imbalanced datasets with several small events and fewer large events, such artificial break-points in magnitude scaling ensure that PSAs scale differently for small and large magnitude ranges. For instance, in Figure 3, we notice that PSA_{Rref} (T = 0.02s) scales more gradually (less positive gradient) for M_w ≤ 4.5 compared to M_w > 4.5. Similarly, PSA_{Rref} (T = 2s) scale rapidly for M_w < 6.5 compared to M_w ≥ 6.5 range, where the scaling gradient is close to zero.

Based on our non-parametric analysis, we formulated our $f_M(M_w)$ to have two break-points: 1) at reference magnitude M_{ref} = 4.5 to separate the numerous small events (M_w < 4.5) from fewer intermediate – large events (M_w ≥ 4.5), and 2) a period-dependent hinge-magnitude (M_h), to allow over-saturation of PSA_{Rref} for large events. The period-dependence of M_h is inspired by reasoning of Boore et al. (2014), where visual inspection of NGA-West2 data suggested M_h to increase with period from M_h = 5.5 at T ≤ 0.1s to M_h = 6.2 at T ≥ 0.4s. However, unlike Boore et al. (2014), we allowed M_h to monotonically increase beyond T = 0.4s to reach M_h = 6.46 at T = 2s. Stochastic simulations (e.g. Schmedes and Archuleta (2008)) and empirical evidence suggest that ground motions saturate (or even over-saturate) at close distances from large magnitude events. To account such physical effects, we formulated $f_M(M_w)$ for our GMPE as in eq. (4).

$$f_M(M_w) = a + \begin{cases} b_1(M_w - M_{ref}) & M_w < M_{ref} \\ b_2(M_w - M_{ref}) & M_{ref} \le M_w < M_h \\ b_2(M_h - M_{ref}) + b_3(M_w - M_h) & M_h \le M_w \end{cases} \quad (4)$$

4.3 Random-effects: δB_e and $\delta S2S_s$

During the regression of $f_R(M_w, R_{JB})$ and $f_M(M_w)$ we allow the algorithm to estimate conditional values of δB_e and $\delta S2S_s$. These *intermediate* estimates are however not the final random-effects of the GMPE, but only to prevent a well-recorded event or site from biasing the fixed-effects coefficient estimates (as discussed in Kotha et al. (2017a) and Stafford (2014)). The final estimate of random-effects are obtained after correcting the observed PSAs for both magnitude and distance scaling effects (as in Boore et al. (2014)). For a record from e^{th} event at s^{th} site the residual $\varepsilon_{e,s} = ln(PSA_{e,s}) - f_R(M_e, R_{e,s}) - f_M(M_e)$ is split into random-effects δB_e and $\delta S2S_s$, and event-and-site corrected residual $\delta WS_{e,s}$. These random-effects and residuals can be further investigated to evaluate the GMPE performance and/or to identify new predictor variables that can be modelled as fixed-effects.

Figure 4 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. If in case the fixed-effects components are biased by a particularly well-sampled magnitude-distance bin in the dataset, then we should expect to see artifacts in the random-effects and the residuals. In the top panels of Figure 4, we plotted δB_e versus M_W to evaluate our choice 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 4 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 modeling choice is motivated by the need to constrain the site terms with a large amount of data. Nevertheless, our $f_R(M_w, R_{JB})$ performs very well at all distances, as indicated by the zero mean $\delta WS_{e,s}$ within each distance bin.

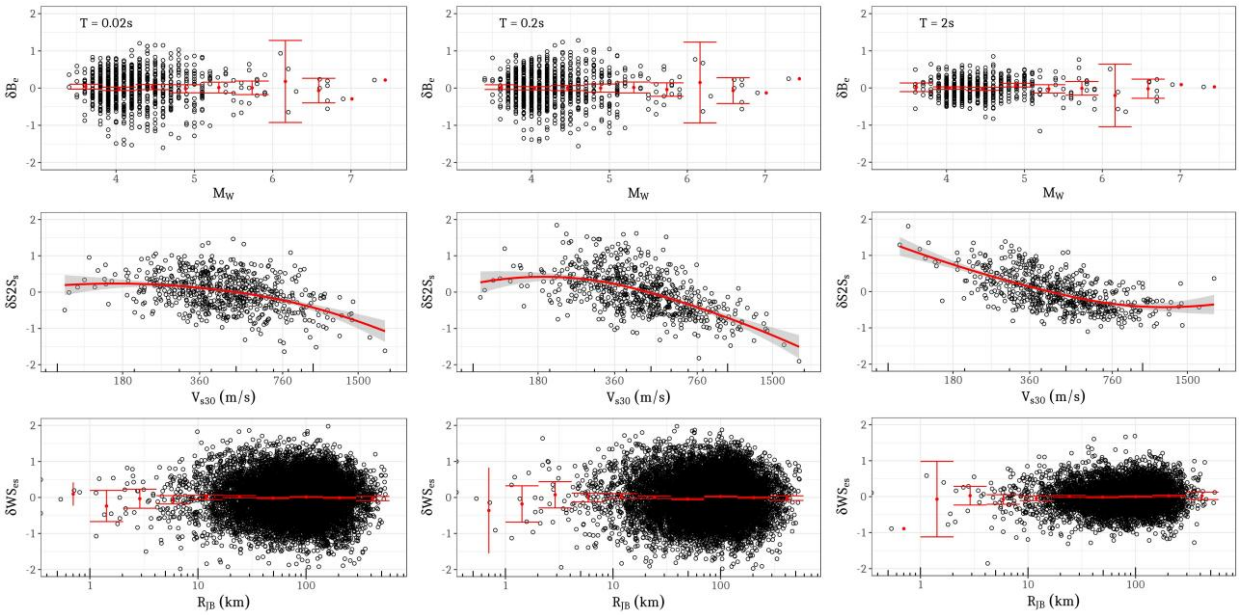

Figure 4: Random-effects and residual plots for GMPE evaluation at T = 0.02s, 0.2s, and 2s. In each panel, δB_e is plotted against M_W, $\delta S2S_s$ against V_{s30}, and $\delta WS_{e,s}$ against R_{JB} to check if random-effects and residuals show a systematic trend with predictor variables

The middle panels of Figure 4 showing $\delta S2S_s$ 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}. However, at T = 0.02s (and T = 0.2s) gradient of the LOESS fit (Local regression of scatterplots by Cleveland (1979)) of $\delta S2S_s$ versus V_{s30} is close to zero for V_{s30} < 600m/s , implying that high frequency soil response is 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 better relevance of V_{s30}, it appears that low frequency response of stiffer soils may not be captured with V_{s30} alone. Our observations suggest that V_{s30} may not be an ideal proxy of linear site-response. In the later sections, we use these inferences in developing alternative approaches to empirical site-response modelling.

4.4 Results

In Figure 5 we plot the distance scaling fixed-effects component of the GMPE against the observed PSAs corrected for the between-event and between-site random-effects, i.e. $\ln(PSA_{\delta B_e + \delta S2S_s}) = lnPSA_{e,s} - \delta B_e - \delta S2S_s$, where e and s are the indices of event and site respectively. The scatter of data around the GMPE median is the record-to-record variability $\delta WS_{e,s}$.

At all periods, the magnitude-dependent *h*-parameter appears to capture well the near-source saturation of PSAs. Kenneth W Campbell (1981) observed that the near-source attenuation of high frequency PSAs (e.g. PGA) is weakly dependent on magnitude and distance, which is evident from the minor differences in our GMPE median predictions at R_{JB} < 10km for M6.5 and M7.5 events in Figure 5. Secondly, the fixed-effects coefficient c_3 forces an exponential decay of GMPE PSA predictions to mimic the anelastic attenuation of PSAs at far-source distances, R_{JB} > 100km. In the panel for T = 2s, the *bump* in the predicted PSAs at 100km, which is in agreement with the observations as well, indicates a possibility of post-critical reflections at crustal discontinuities and/or transition from body waves at near-source distances to surface waves at far-source distances. This phenomenon is also in agreement with the *kink* at about 90 km observed by Oth et al. (2011) in the attenuation of Fourier amplitude spectra at low frequencies in Japan.

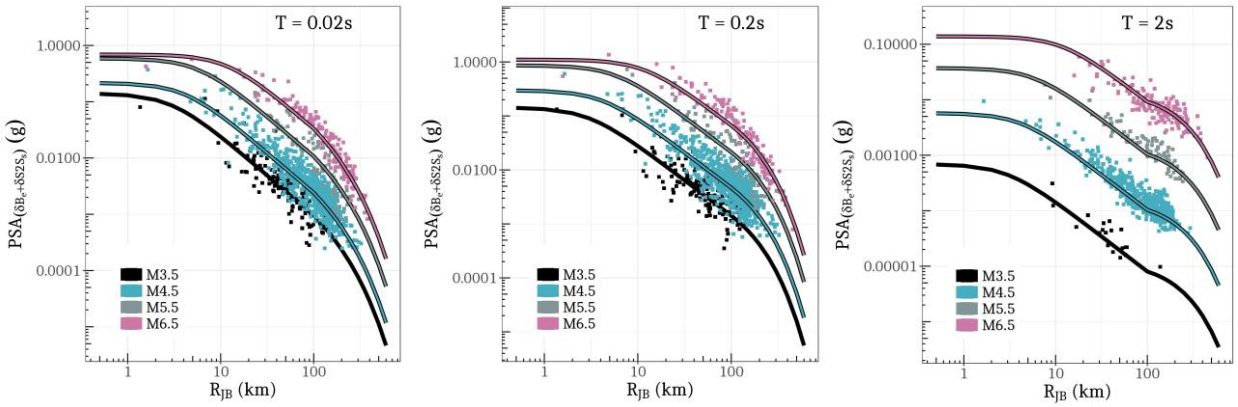

Figure 5: Distance scaling of the GMPE fixed-effects at T= 0.02s, 0.2s, and 2s. Note that the vertical axis shows observed PSAs minus the between-event and between-site random-effects.

Figure 6 shows the magnitude scaling of GMPE fixed-effects versus the observed PSAs corrected for random-effects δB_e and $\delta S2S_s$. Despite two break-points, one at M_w = 4.5 and the other at M_h = 5.5 – 6.5 depending on the period, magnitude scaling for T = 2s appears to be constant for all events $M_w \leq 6.5$, and an over-saturation for M_w > 6.5. Although the over-saturation appears rather strong, the panels corresponding to T = 2s, in our non-parametric analysis of Figure 3 do exhibit a decreasing trend of $PSA_{Rref,}$ while that in residual analysis of Figure 4 shows no bias for M_w > 6.5. The strong over-saturation at short distances is a combined effect of the h-parameter (eq. 2) and the period-dependent M_h (eq. 4).

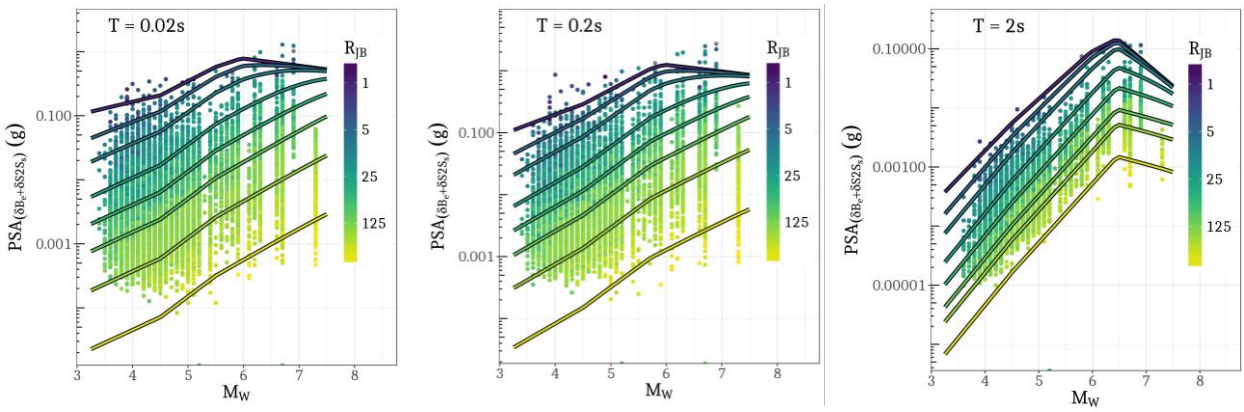

Figure 6: Magnitude scaling of the GMPE fixed-effects at T = 0.02s, 0.2s, and 2s. Note that the vertical axis shows observed PSAs minus the between-event and between-site random-effects.

Figure 7 shows the GMPE fixed-effects response spectra in the left panel, and random-effects standard deviations in the right panel. The GMPE standard deviations in the left panel are in natural-log scale. φ_{S2S} is the largest among the three components, indicating a large site-to-site variability in the dataset, and also because we chose not to include any site-response parameters (e.g. V_{s30}) in the GMPE fixed-effects component. In the following sections we introduce the clustering approach to site-response modelling, which is expected to reduce the φ_{S2S}, and the total aleatory variability σ to a more reasonable value.

In our mixed-effects GMPE, the magnitude and distance scaling are captured by the fixed-effects $f_M(M_w)$ and $f_R(M_w, R_{JB})$, event-specific adjustments by random-effects δB_e, and record-to-record variability by $\delta WS_{e,s}$. Therefore, each site-specific random-effect $\delta S2S_s$ essentially captures the empirical *mean* site-response of a site in the GMPE regression. For every period at which the GMPE is regressed (T = 0.01s – 2s), a well-recorded site in the dataset has an associated period-dependent $\delta S2S_s$. The scalar $\delta S2S_s$ at a given period can be used to adjust the generic GMPE PSA(T) predictions to yield site-specific $PSA_{ss}(T)$ (e.g. Rodriguez-Marek et al. (2013), Kotha et al. (2017a)). A vector of site-specific $\delta S2S_s$ for a range of T, notated as $\Delta S2S_s$ from hereon, resembles an empirical site amplification function (AF). $\Delta S2S_s$ can be used to adjust the GMPE PSA response spectra, Conditional Spectra (Baker (2010) and Kotha et al. (2017b)), or even the Uniform Hazard Spectra. In this study, the $\Delta S2S_s$ vector for well-recorded sites serve as the empirical site amplification functions.

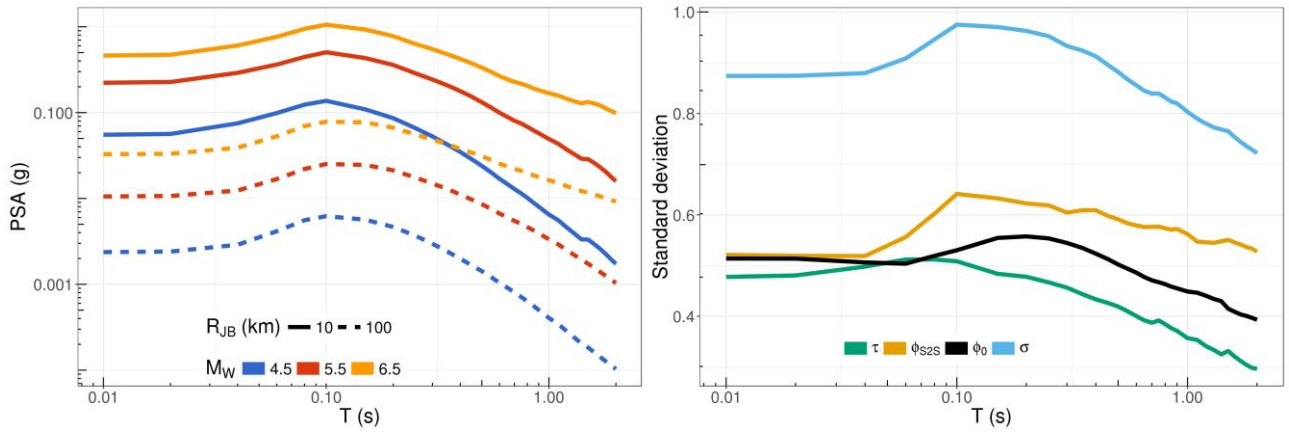

Figure 7: (Left panel) Median response spectra. (Right panel) between-event (τ), between-site (φ_{S2S}), event-and-site corrected standard deviations (φ_0) and the total aleatory variability (σ) of the GMPE in natural-log scale

5 Spectral clustering analysis

With increasing amount of strong motion data, ground motion modelers are adapting advanced statistical tools to analyze large amounts of ground motion data to identify repeatable effects, evaluate parametrization, and quantify better the uncertainties in prediction (e.g. Derras et al. (2012), Ullah et al. (2013), Luzi et al. (2011)). Among the various tools, spectral clustering analysis, a type of unsupervised machine learning, refer to a variety of statistical techniques aimed at extracting hidden patterns/structures from large amounts of unlabeled multidimensional data. Constituted by *n* scalar $\delta S2S_s$ value for *n* spectral periods (T = 0.01s – 2s), the $\Delta S2S_s$ vectors of length *n* are our multidimensional data points. The steps involved in spectral clustering are the following:

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}.
   It is possible to extend the period range of $\Delta S2S_s$ vectors, but the number of sites with a reliable estimate of $\delta S2S_s$ falls rapidly beyond T > 2s, which in turn is controlled by the maximum usable period of T_{max} of a record (in our record selection step). In fact, including long period $\delta S2S_s$ (up to T = 5s) in our analysis resulted in better separation of clusters, but with very few sites in each cluster.
2) **Choice of clustering technique:** There are several advanced machine learning techniques depending on the amount of supervision (a priori information) that is input and the knowledge that is being queried. For our purpose, we chose a basic partitioning algorithm: the k-means clustering technique MacQueen (1967). k-means technique splits the $\Delta S2S_s$ vectors into *k* groups (clusters), where *k* must be specified in advance. The clustering algorithm (available in the R library *ClusterR* by Lampros Mouselimis (2017) and *factoextra* by Kassambara and Mundt (2016)) iteratively partitions the data until the Total Within Sum of Squares (WSS) is 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 8).

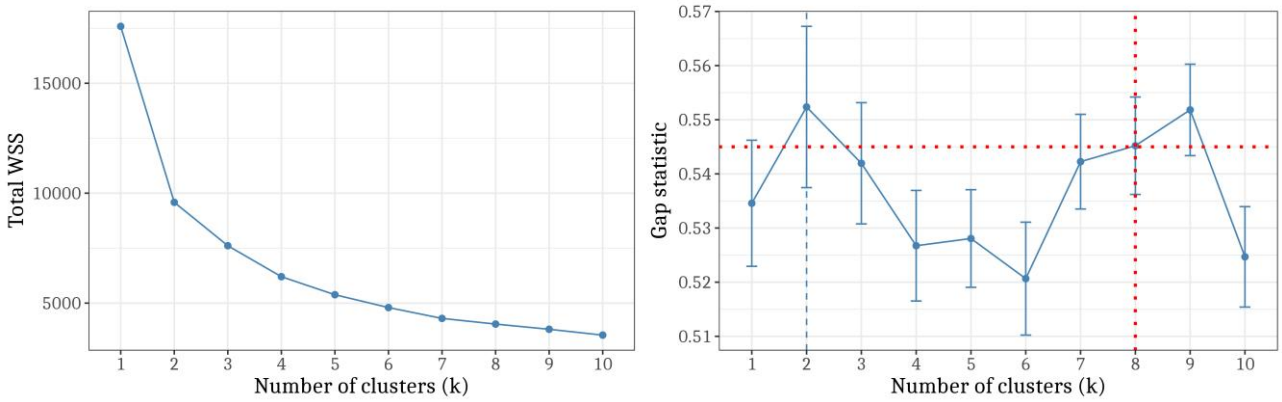

Figure 8: Optimal number of clusters based on Total Within Sum of Squares (WSS, left panel) and Gap statistic (GS, right panel). The WSS metric reduces with increasing number of clusters, but the optimal number of clusters is when Gap statistic is maximized – in which case the WSS is low and the inter-cluster distance is high.

6 Site classification from cluster analysis

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 Eurocode8 Code (2005)), preferred site-response proxy (e.g. V_{s30}), and parametric ranges of selected proxy (e.g. sites with V_{s30} > 800m/s as EC8-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 our 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.

6.1 Site clusters

The eight site clusters partitioning the 588 sites in our dataset are visualized in Figure 9, 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 9.

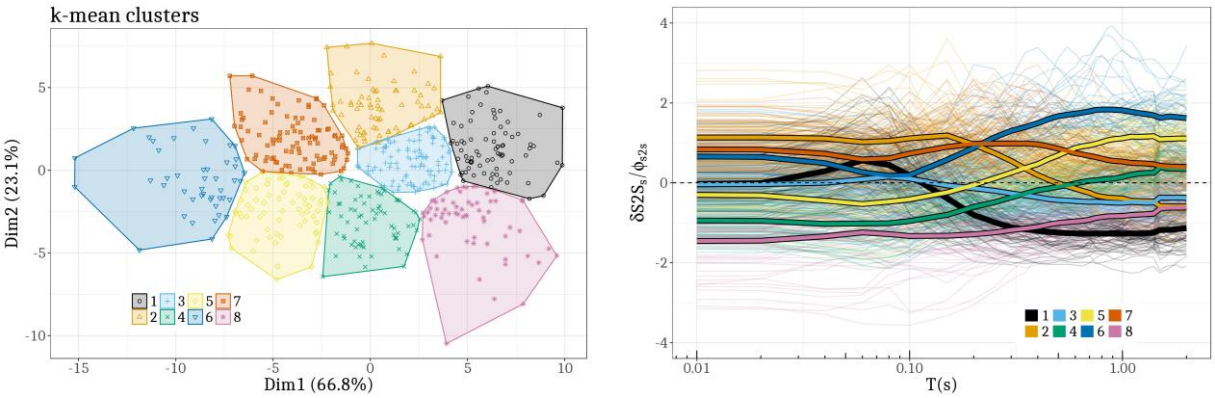

Figure 9: (left panel) Visualization of k-mean clustering, where each polygon is a cluster and each point within is a site ($\Delta S2S_s$). Dim1 and Dim2 are variables derived from a Principal Component Analysis of $\Delta S2S_s$ vectors, which together describe 89.9% of data variability. (right panel) 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 our dataset are plotted in the right panel of Figure 9. 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 to our empirical site amplification functions associated with the site clusters/classes derived in this study. Observing the two plots in Figure 9: 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 9, shows the opposite behavior. The same logic can be applied to cluster 1 and 5, and so on. Since the $\delta S2S_s$ distributions at each period are normally distributed with zero mean, we expect to see such symmetric classification from our procedure. Eventually, these clusters will be validated with geotechnical site-response parameters, based on which we assert that the spectral clustering procedure yields physically meaningful site classification. The following sections presents the practicality of our 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

6.2 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. For example, in EC8 the reference site conditions are characterized as outcropping rock sites with V_{s30} = 800m/s. Probabilistic seismic hazard assessment (PSHA) in a region and associated hazard estimates such as hazard curves and hazard maps are first produced for such reference site conditions, and then adjusted to site-specific estimates using the amplification functions. Our first point of interest is then to identify reference site conditions derived from the site clusters.

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 9, 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). Our selection of reference site cluster is solely based on its relatively flat empirical mean amplification function. In the left panel of Figure 10, 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 9 are scaled back to their original random-effect estimates by multiplying them with period-dependent between-site standard deviations φ_{S2S} of Figure 7
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 our $\Delta S2S_s$ are additive random-effects of a mixed-effects GMPE in natural-log scale, the multiplicative amplification functions would be $AF=e^{\Delta S2Ss}$. 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 10 compares the within-cluster site-to-site variability ($\varphi_{S2S,c}$) against the pre-clustered between-site (φ_{S2S}) variability of the dataset (for our 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 in right panel of Figure 7).

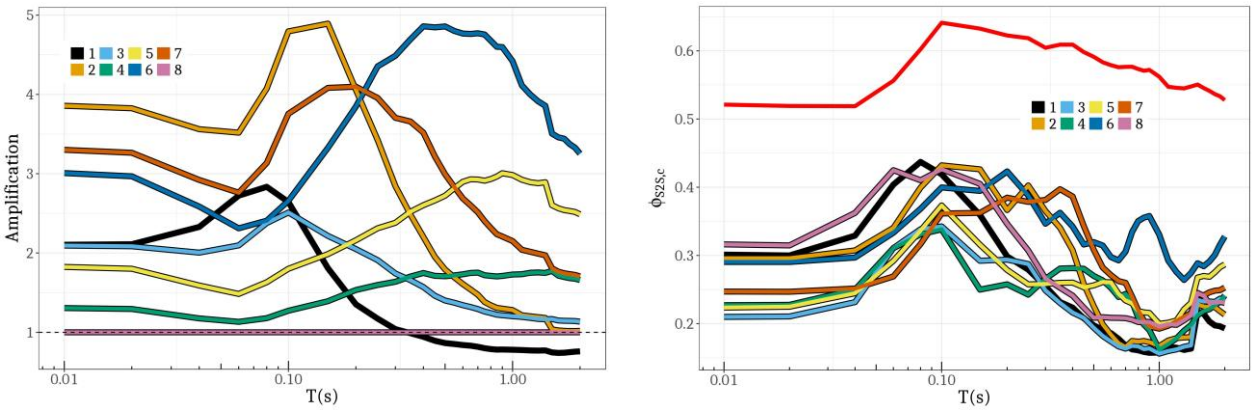

Figure 10: (left panel) Site amplification functions of each cluster scaled with respect to the reference site conditions: cluster 8. (right panel) The within-cluster site-to-site variability φ_{S2S} of the 8 clusters compared to

the overall GMPE φ_{S2S} prior to clustering (red curve). Note that the φ_{S2S} (red curve) is the same as the φ_{S2S} (yellow curve) in Figure 7

Looking at the amplification functions in Figure 10, 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 $\varphi_{S2S,c}$, also reaches its maximum value around its T_G, indicating a large variability in its amplification. For example, cluster 1 shows a peak amplification at its T_G = 0.08s (AF = 2.8 in left panel of Figure 10), which also coincides with the period where its $\varphi_{S2S,1}$ is highest (0.45 in right panel of Figure 10). 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 highly 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.

6.3 Site-response proxies

Using the empirical site amplification functions and cluster-specific φ_{S2S}, the GMPE can be adjusted to predict site class dependent ground motions in hazard assessment; the missing link is sufficient and efficient site response proxies to classify new sites in a PSHA. 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 attempted characterizing the cluster amplification functions of Figure 10 using only the geotechnical site-response parameters available in the dataset. In process of developing a new site classification scheme, we first evaluated our eight site clusters against the site classification scheme defined in Zhao et al. (2006), Association (1980) and Association (1990). For a similar evaluation against the Eurocode 8 site classification, we refer the readers to Kotha et al. (2018).

6.3.1 Predominant period of the soil column: T_G

Current EC8 categorizes five site classes using only V_{s30} ranges, while Zhao et al. (2006) classified the K-Net stations into four sites classes (SC1 through SC4) based on their H/V spectral ratios. Each of these 4 classes is attributed a V_{s30} range, a characteristic range of predominant period T_G, and corresponding NEHRP class Council (2000). For reference, we provide the site classification scheme of Zhao et al. (2006) in Table 2. In this study, we assume that the period at which the cluster amplification functions attain their peak values in Figure 10, are analogues to H/V spectral ratio based T_G of Zhao et al. (2006). Under this assumption, we categorize our eight site clusters based on their inferred T_G (Figure 10) in the right most column of Table 2.

Table 2: Site classification based on Zhao et al. (2006)

Zhao et al. 2016	T_G (s)	V_{s30} (m/s)	NEHRP	Clusters (this study)
SC1	$T_G < 0.2s$	$V_{s30} > 600m/s$	A + B	C1 + C2 + C3
SC2	$0.2s \leq T_G < 0.4s$	$300m/s \leq V_{s30} < 600m/s$	C	C7
SC3	$0.4s \leq T_G < 0.6s$	$200m/s \leq V_{s30} < 300m/s$	D	C6
SC4	$0.6s \leq T_G$	$V_{s30} < 200m/s$	E + F	C4 + C5

From this comparison of site classification criteria:

1. Cluster 8 does not fit in any of the Zhao et al. (2006) site classes because it showed no clear T_G, which is also the reason it was chosen as a reference site cluster
2. Cluster 7 and 6 show clear peak amplification, and are in accordance with Zhao et al. (2006) proposed classes SC2 and SC3 respectively
3. Cluster 4 and 5 show a broad amplification plateau beyond T > 0.6s, hence grouped into SC4. However, the amplification levels for these two clusters are significantly different, where also the within-cluster variability $\varphi_{S2S,c}$ is relatively very small (~0.2 for T > 0.6s in Figure 10). Based on their low within-cluster variabilities, we chose not merge cluster 4 and 5 into a single cluster
4. Clusters 1, 2 and 3, all show a clearly defined peak amplification at $T_G < 0.2s$. Figure 10 suggests the mean amplification levels for these clusters to be very different in the vicinity of T_G. Reducing to sub-optimal number of clusters (in k-mean clustering) does not merge these clusters, suggesting the need to divide SC1 into three sub-class: C1, C2, and C3 – against merging them into SC1

6.3.2 Time averaged shear-wave velocity of soil column: V_{s10}, V_{s30} and H_{800}

Table 2 shows the compatibility of our clusters with Zhao et al. (2006) classification based on their T_G, while Figure 11 categorizes them further based on their distribution of H_{800}, V_{s10} and V_{s30}. Note that the clusters in Figure 11 are rearranged based on their median H_{800} (depth to engineering bedrock with V_s = 800m/s). For instance, clusters 4, 5, 6 and 7 in the top row have the deepest soil column with median H_{800} > 50m. Clusters 1, 2, 3 and 8 on the other hand are characterized by shallow soil profiles with median H_{800} around 10m. For visual clarity, both panels provide guiding lines at H_{800} = 10, 30, and 100m. In the left panel, the x-axis marks the bounding V_{s30} values of Zhao et al. (2006) site classification at 200, 300 and 600m/s, while in the right panel, the x-axis is divided at V_{s10} = 200, 300 and 400m/s. We used Table 2 and Figure 11, to evaluate the physical meaning of our clusters, and also to define new site classes in Table 3.

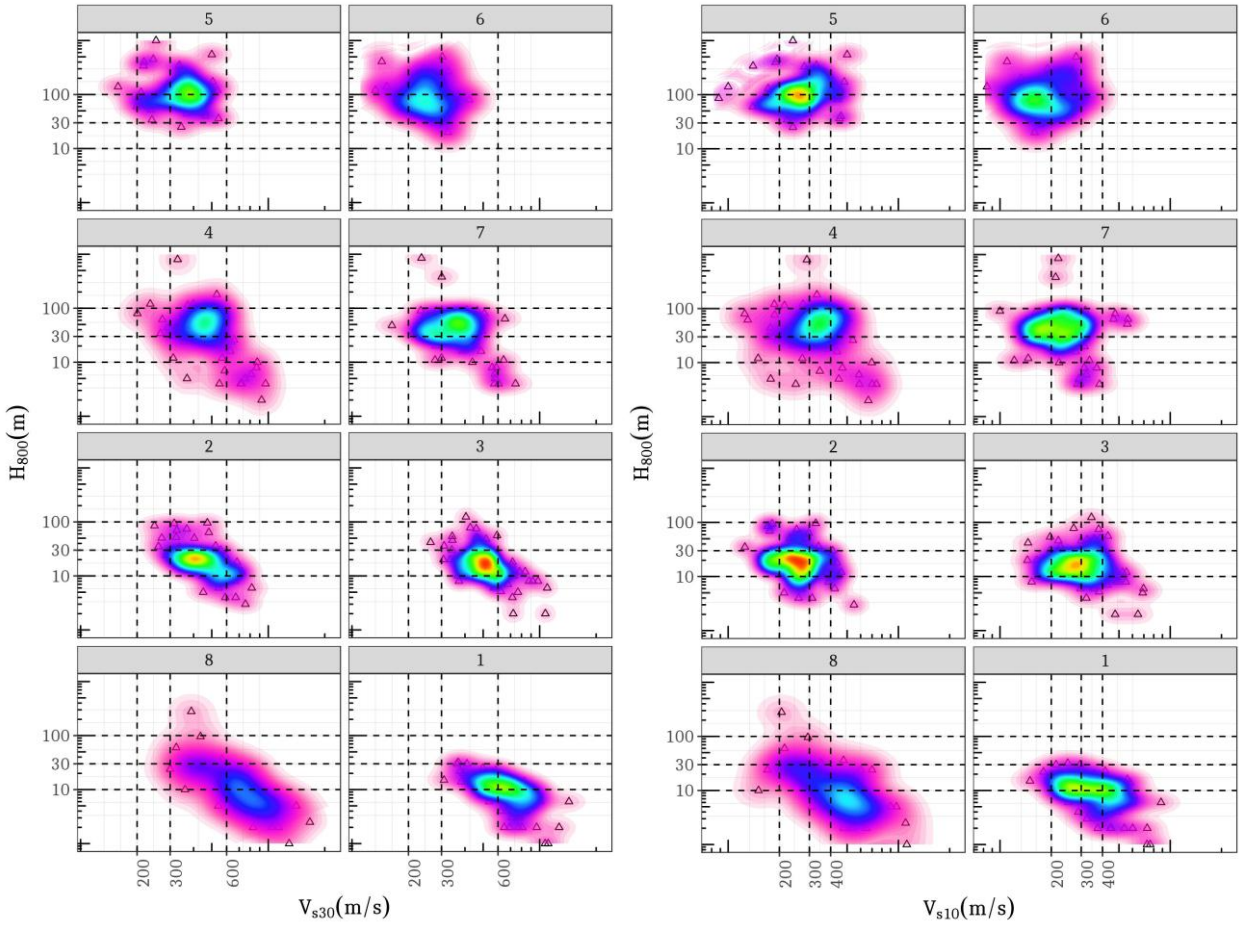

Figure 11: Evaluation of site response proxies in characterizing site clusters: In the left panel is the combination of H800 and Vs30, and in the right panel, H800 and Vs10. The contour colors represent the 2D Kernel (bivariate normal) distribution of geotechnical parameters of sites within each cluster. Warmer colors (Red - Green) imply high density, colder colors (Blue – Purple) mark low density boundaries

The colored contours of Figure 11 represent the 2D kernel density (bivariate normal distribution) of V_{s30} –H_{800} and V_{s10} – H_{800} of each cluster. The brightly colored regions bound the high density of points (sites), which diffuses into outer counters covering low density regions. We use the high density contours (green patches in Figure 11) of the kernel density plot to identify representative ranges of V_{s30}, V_{s10}, and H_{800} combination that characterize our new site classification scheme in Table 3. In addition, we provided a revision of T_G ranges inferred from the peaks in amplification functions of Figure 10.

Table 3: 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)**
C5	> 1s	300 – 450m/s	200 – 300m/s	> 50m
C4		400 – 600m/s	300 – 400m/s	30 – 100m
C6	0.4 – 1s	200 – 300m/s	< 200m/s	> 50m
C7	0.2 – 0.4s	200 – 450m/s	200 – 400/s	30 – 100m
C3	0.1 – 0.2s	450 – 600m/s	200 – 400m/s	10 – 30m
C2		300 – 600m/s	150 – 350m/s	
C1	< 0.1s	450 – 600 m/s	200 – 600m/s	5 – 20m
C8	-	> 600m/s	> 600m/s	< 5m

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 Zhao et al. (2006). From Figure 11 and Table 3, we make the following observations:

- 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}. Note that these clusters resembled SC4 of Zhao et al. (2006) based on their T_G (> 0.6s), and were expected to also show a V_{s30} < 200m/s. However, our data consists of very few KiK-net sites with V_{s30} < 200m/s
- 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 8 serves as our 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 10), hence could not be compared with Zhao et al. (2006) classes
- Cluster 1 from 2 and 3, which were nested in SC1 of Zhao et al. (2006), can be resolved based on their T_G values and H_{800} ranges. Cluster 1 with T_G < 0.1s (left panel of Figure 10), is the only one showing a strong amplification at high frequency and de-amplification towards longer periods, with respect to reference site cluster 8. Cluster 1 and 8 can be 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 10), 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

7 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. It is necessary for a variety of reasons:

1) We need as many records as possible for each site, so that the estimated $\Delta S2S_s$ vectors have a low estimation errors
2) It is important to include a large number of small events and long distance recordings, as opposed to the more important (in a hazard perspective) large magnitudes and short distances, to ensure that the $\Delta S2S_s$ vectors capture predominantly linear site response – and not biased by nonlinear response on very soft soils
3) Given that large magnitude events at near source distances may trigger nonlinear site response, and that these scenarios are also subject to other phenomena such as: saturation of ground motions beyond M > 6.5 and at near-source distances R_{JB} < 20km, we were required to model carefully the near-source magnitude and distance scaling. In doing so, we used a period-dependent effective-depth, or the *h*-parameter, proposed by Yenier and Atkinson (2015). For far-source distances, we could capture the apparent anelastic attenuation phenomenon and, what appears to be a combination of post-critical reflections at crustal discontinuities (e.g. Moho reflections) and transition from body waves at near-source to surface waves at far-source distances from the event. We observed that our distance scaling works reasonably well for the large distance range we chose
4) Finally, we observed that the $\delta S2S_s$ show a weak correlation relation with V_{s30}. For this dataset, the site-response: 1) is not efficiently captured by V_{s30} particularly at short – moderate periods, and 2) highly variable even for sites with identical V_{s30}. We therefore attempted our alternative approach to classifying sites based on V_{s30}

Our need to derive a GMPE is to demonstrate that the magnitude, and distance scaling of ground motions (here PSA) are strongly period dependent. Following our exercise, we suggest caution against deriving empirical site amplification functions from response spectra normalized by Peak Ground Acceleration (PGA). The $\Delta S2S_s$ vectors on the other hand, are site-specific terms filtered for event and path effects with a robust GMPE median. In this sense, we consider $\Delta S2S_s$ vectors as more suitable in developing site classification schemes and amplification functions. 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 (in terms of their $\Delta S2S_s$ vectors). 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} > 800m/s are considered as a reference sites. Hazard estimates are made on reference site conditions and then scaled using the soil-site amplification functions. 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 seen as 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} > 30m and T_G > 0.2s, 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} > 50m) 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 (30m < H_{800} < 100m)
4) For sites with 10m < H_{800} < 30m and 0.1s < T_G < 0.2s, 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} > 600m/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 < 0.1s, 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} > 800m/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. It can be extended to the point where the clusters can be hierarchically divided or merged depending on the available site parametrization in a region. However, we note that site types, sparsely represented or not represented at all in the dataset, may not be identified with data-driven techniques as ours. A more flexible, predictive method is in development for application to new pan-European dataset 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.

Acknowledgements

We appreciate the efforts of Dr. Julian Bommer, and the anonymous reviewer, for their impressively detailed 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/)

References

Abrahamson N, Youngs R (1992). A stable algorithm for regression analyses using the random effects model. *Bulletin of the seismological society of America, 82*(1), 505-510.

Association J R (1980). *Specifications for road bridges*. Tokyo, Japan.

Association J R (1990). *Specification for highway bridges*. Tokyo, Japan.

Baker J W (2010). Conditional mean spectrum: Tool for ground-motion selection. *Journal of Structural Engineering, 137*(3), 322-331.

Baltay A, Hanks T, Abrahamson N (2017). Uncertainty, Variability, and Earthquake Physics in Ground-Motion Prediction Equations. *Bulletin of the seismological society of America*.

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

Boore D M, Stewart J P, Seyhan E, Atkinson G M (2014). NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. *Earthquake spectra, 30*(3), 1057-1085.

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.

Campbell K W (1981). Near-source attenuation of peak horizontal acceleration. *Bulletin of the seismological society of America, 71*(6), 2039-2070.

Campbell K W, Bozorgnia Y (2014). NGA-West2 Ground Motion Model for the Average Horizontal Components of PGA, PGV, and 5% Damped Linear Acceleration Response Spectra. *Earthquake spectra, 30*(3), 1087-1115. doi:10.1193/062913eqs175m

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*.

Council B S S (2000). The 2000 NEHRP Recommended Provisions for New Buildings and Other Structures, Part I (Provisions) and Part II (Commentary). *FEMA, 368*, 369.

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.

Derras B, Bard P Y, Cotton F, Bekkouche A (2012). Adapting the neural network approach to PGA prediction: an example based on the KiK-net data. *Bulletin of the seismological society of America, 102*(4), 1446-1461.

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.

Garcia D, Wald D J, Hearne M (2012). A global earthquake discrimination scheme to optimize ground-motion prediction equation selection. *Bulletin of the seismological society of America, 102*(1), 185-203.

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.

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: https://doi.org/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: https://doi.org/10.1193/091416EQS150M

Kotha S R, Cotton F, Bindi D (2018). *Site Classification Derived From Spectral Clustering of Empirical Site Amplification Functions.* Paper presented at the 16th European Conference on Earthquake Engineering, Thessaloniki, Greece.

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.

Oth A, Bindi D, Parolai S, Di Giacomo D (2011). Spectral analysis of K-NET and KiK-net data in Japan, Part II: On attenuation characteristics, source spectra, and site response of borehole and surface stations. *Bulletin of the seismological society of America, 101*(2), 667-687.

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.

R Development Core Team (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rey J, Faccioli E, Bommer J J (2002). Derivation of design soil coefficients (S) and response spectral shapes for Eurocode 8 using the European Strong-Motion Database. *Journal of Seismology, 6*(4), 547-555.

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.

Schmedes J, Archuleta R J (2008). Near-source ground motion along strike-slip faults: Insights into magnitude saturation of PGV and PGA. *Bulletin of the seismological society of America, 98*(5), 2278-2290.

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 (2014). Crossed and nested mixed-effects approaches for enhanced model development and removal of the ergodic assumption in empirical ground-motion models. *Bulletin of the seismological society of America, 104*(2), 702-719.

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.

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.

Ullah S, Bindi D, Pittore M, Pilz M, Orunbaev S, Moldobekov B, Parolai S (2013). Improving the spatial resolution of ground motion variability using earthquake and seismic noise data: The example of Bishkek (Kyrgyzstan). *Bulletin of Earthquake Engineering, 11*(2), 385-399.

Yenier E, Atkinson G M (2015). An equivalent point-source model for stochastic simulation of earthquake ground motions in California. *Bulletin of the seismological society of America, 105*(3), 1435-1455.

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.

Zhao J X, Zhou S, Zhou J, Zhao C, Zhang H, Zhang Y, Gao P, Lan X, Rhoades D, Fukushima Y (2016). Ground-motion prediction equations for shallow crustal and upper-mantle earthquakes in Japan using site class and simple geometric attenuation functions. *Bulletin of the seismological society of America*.