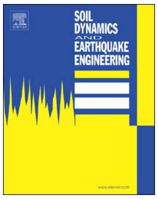
ARTICLE IN PRESS

Soil Dynamics and Earthquake Engineering xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Soil Dynamics and Earthquake Engineering

journal homepage: www.elsevier.com/locate/soildyn



A new approach to site classification: Mixed-effects Ground Motion Prediction Equation with spectral clustering of site amplification functions

Sreeram Reddy Kotha[a,b,*], Fabrice Cotton[a,b], Dino Bindi[a]

[a] Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany
[b] University of Potsdam, Potsdam, Germany

ARTICLE INFO

Keywords:
Mixed-effects regression
Ground Motion Prediction Equation
Site classification
Spectral clustering analysis
Empirical site amplification functions

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 $0 \leq R_{JB} < 600$km. Using k-mean spectral clustering technique, we then classify our estimated $\Delta S2S_s$ $(T = 0.01 - 2\,s)$ 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

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 [1,2] as a means to classification of sites for building codes. For example, Eurocode 8 [3] and [4] 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 [5–8] 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., [9–11]). For example, Ref. [12] introduced a more refined classification using H_{800} (depth to seismic bedrock with $V_s = 800$ m/s), $V_{s,av}$ (average shear-wave velocity of the soil column) and fundamental period (f_0). In total [12] 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., [13]) tested the relative efficiency of the various site-response 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:

---

[*] Corresponding author at: Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Helmholtzstraße 6/7, Building H 6, room 115, 14467 Potsdam, Germany.
E-mail address: sreeram@gfz-potsdam.de (S.R. Kotha).

https://doi.org/10.1016/j.soildyn.2018.01.051
Received 31 July 2017; Received in revised form 16 January 2018; Accepted 28 January 2018

Please cite this article as: Sreeram Reddy, K., Soil Dynamics and Earthquake Engineering (2018), https://doi.org/10.1016/j.soildyn.2018.01.051

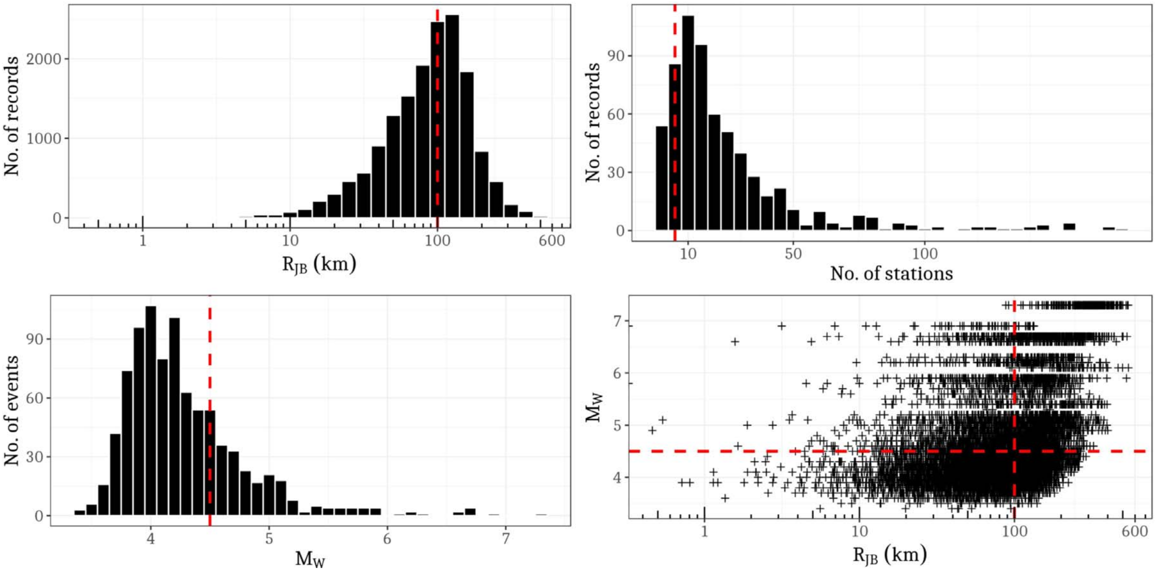


Fig. 1. Data distribution following the record selection criteria for GMPE regression at T = 0.01 s: (top-left panel) Distance distribution of usable records, (top-right panel) number of records per station, (bottom-left panel) magnitude distribution of usable records, (bottom-right panel) magnitude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

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 [14], 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 dataset is provided in the Section 2.
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 [15,16]. Details on the GMPE development and mixed-effects analysis is provided in Section 3.
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 5, 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 6, 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 [17] database compiled by [14] 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 [14] 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 [14], we make a more GMPE specific record selection for our regression:

1) Ref. [14] 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–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 [18] 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 ≤ 35 km (as in the H_{ANSR1} criteria of [18]).
3) Most of the KiK-net sites provide 3-component recordings at both surface and borehole sites. In our study, we use only 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 [14], 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 – 30 Hz. 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 = 0.01 s falls from 157,000 to 15,896. The number of usable records further decreases to 6462 at T = 2 s. The data distribution for GMPE regression at T = 0.01 s is shown in Fig. 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.

3. Ground Motion Prediction Equation

Using a mixed-effects regression approach (as in [19,20]), we derive a GMPE for the geometric-mean of (5% damped) horizontal Pseudo Spectral Acceleration (PSA) at 33 values of T between 0.01 s and 2 s.

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

In Eq. (1), the parametric functions $f_R(M_w, R_{JB})$ and $f_M(M_w)$ capture