Cal Poly

Caltech

SCEC
AN NSF+USGS CENTER

UC Irvine

UCLA

UC Santa Barbara

USC

Horizontal-to-Vertical Spectral Ratios from California Sites: Open-Source Database and Data Interpretation to Establish Site Parameters

Pengfei Wang, Paolo Zimmaro*, and Tatiana Gospe
Civil & Environmental Engineering Department
University of California, Los Angeles
(*now at University of Calabria, Italy)

Sean K. Ahdi and Alan K. Yong
U.S. Geological Survey
Golden, CO (SKA) and Pasadena, CA (AKY)

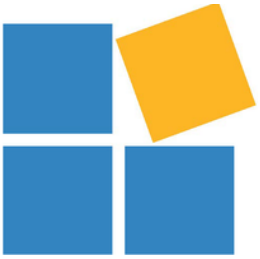
Jonathan P. Stewart (PI)
Civil & Environmental Engineering Department
University of California, Los Angeles

A report on research conducted with support from California Strong Motion Instrumentation Program, California Geological Survey, Agreement 1016-985 and 1018-569.

Report GIRS-2021-06

DOI: 10.34948/N3KW20

University of California, Los Angeles (headquarters)



Natural Hazards Risk & Resiliency Research Center

B. John Garrick Institute for the Risk Sciences

Horizontal-to-Vertical Spectral Ratios from California Sites: Open-Source Database and Data Interpretation to Establish Site Parameters

Pengfei Wang, Paolo Zimmaro*, and Tatiana Gospe

Civil & Environmental Engineering Department
University of California, Los Angeles, California, USA
(* now at University of Calabria, Italy)

Sean K. Ahdi and Alan K. Yong

U.S. Geological Survey, Golden Colorado (SKA) and Pasadena California (AKY), USA

Jonathan P. Stewart (PI)

Civil & Environmental Engineering Department
University of California, Los Angeles, California, USA

A report on research conducted with support from California Strong Motion Instrumentation Program, California Geological Survey, Agreement 1016-985 and 1018-569.

Report GIRS-2021-06
doi:10.34948/N3KW20

Natural Hazards Risk and Resiliency Research Center
B. John Garrick Institute for the Risk Sciences
University of California, Los Angeles (Headquarters)

August 2021

ABSTRACT

Frequency-dependent horizontal-to-vertical spectral ratios (HVSR) of Fourier amplitudes from three-component recordings can provide information on one or more site resonant frequencies and relative levels of amplification at those frequencies. Such information is potentially useful for predicting site amplification but is not present in site databases that have been developed over the last 15–20 years for the Next-Generation Attenuation (NGA) projects, which instead use the time-averaged shear-wave velocity (V_S) in the upper 30 m of the site (V_{S30}) as the primary site parameter and are supplemented with basin depth terms where available. As a consequence, HVSR parameters are also not used in NGA ground motion models.

In order for HVSR-based parameters to be used in future versions of site databases, a publicly accessible repository of this information is needed. We adapt a relational database developed to archive and disseminate V_S data to also include HVSR. The database provides relevant microtremor-based HVSR data (mHVSR) and supporting metadata. We consider the most relevant data to be the frequency-dependent mHVSR, where the horizontal is taken as the median component and also as a function of horizontal azimuth (referred to as *polar plots*). Relevant metadata includes site location information, details about the equipment used to make the measurements, and processing details related to windowing, anti-trigger routines, and filtering. We describe the database schema developed to organize and present this information.

The relational database stores mHVSR data, but not site parameters derived from the data. Site parameters of potential interest for modeling purposes include: (1) a binary variable indicating whether an mHVSR plot contains a peak; (2) one or more peak frequencies; (3) peak amplitudes; and (4) peak widths.

We present procedures for peak identification that we believe to be better suited to California conditions than the SESAME (2004) guidelines that are typically applied in current practice. These procedures are informed by analysts' visual assessments and can largely reproduce peak selections developed by relatively "conservative" or "liberal" analysts (producing relatively few or many sites with peaks, respectively). These procedures use tree regression to identify peak-adjacent plateaus in HVSR, which in turn can be used to identify relative peak amplitudes and peak widths that are considered in the proposed peak identification criteria. The algorithm is coded in R and a Jupyter Notebook and performs the operation of peak identification, and for sites with peaks, peak fitting using a Gaussian function. These routines interact with the database via cloud computing, but are not directly part of the database.

We investigate the consistency of mHVSRs derived from velocity seismometers and accelerometers, which shows a high rate of false negatives (missed peaks) from accelerometers, even when used with 24-bit data recorders. This might be due to the relatively high intrinsic noise and low sensitivity of accelerometers. We compare mHVSRs derived from co-located temporary

instruments (as would be used in a site characterization study) vs. permanent instruments (which could be applied to characterize ground motion stations) to evaluate the effectiveness of the latter, enabling us to query microtremors from permanent stations to boost the HVSR database. We find about 60-80% consistency in this case, with no bias in the peak assessment from one type of noise measurement relative to the others.

We compare mHVSR from velocity seismometers to those from earthquake recordings (eHVSRs). We find microtremors and earthquake recordings are consistent for 60-70% of sites, in the sense that both either do or do not have significant peaks, and when peaks are present, they occur at similar frequencies (i.e., differences between frequencies are < 20%). However, for sites where mHVSR identifies a peak, we find a false-positive rate of about 50%. Approximately one-third of those false-positives can be accounted for by the limited frequency range over which eHVSR results are well constrained (low frequencies are often missed). The false-negative rate from mHVSR is very low. Future adjustments of peak identification criteria (making them more conservative) could be undertaken if it were desirable to bring the false-positive and false-negative rates into closer alignment.

These findings are important to consider when contemplating the development or use of site response models derived from HVSR-based parameters such as site frequency. In engineering applications, these parameters will almost always be derived from mHVSRs. However, for model development, it is tempting to use eHVSRs, because such information is most widely available for ground motion stations. Because mHVSRs and eHVSRs do not always match, it is important to derive models solely from mHVSRs to ensure consistency between parameters used in model development and forward applications. Moreover, because many (approximately 70%) of California sites do not have HVSR peaks, it is important for HVSR model development to consider indices on whether peaks are or are not present. Currently available models do not adequately account for these effects.

ACKNOWLEDGMENTS

Funding for this study was provided by California Strong Motion Instrumentation Program, California Geological Survey, Agreement 1016-985 and 1018-569. We gratefully acknowledge this support. We also thank Antony Martin from GEOVision for sharing his expertise on processing the HVSR data. Site data were provided for addition to the database by the fifth author (U.S. Geological Survey), Koichi Hayashi (Geometrics), David Teague (Engeo), and Tristan Buckreis (UCLA); this data sharing and cooperation are deeply appreciated. We thank Silvia Castellaro, Seth Carpenter, Brady Cox, Hiroshi Kawase, Shinichi Matsushima, Robert Nigbor, Stefano Parolai, Marco Pilz, Lisa Schleicher, Jamison Steidl, Alan Yong, and Zhenming Wang for their input on this research. We thank Grace Parker, Oliver Boyd, Morgan Moscetti, Mike Diggles, and Shane Detweiler for their reviews of manuscripts and abstracts related to this project.

The opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the study sponsors, the B. John Garrick Risk Institute, or the Regents of the University of California. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 RESEARCH MOTIVATION

Seismic site response is influenced by several factors, including: resonance, nonlinearity, amplification due to impedance contrasts, and amplification related to wave propagation in sedimentary basins. Ground-motion models predict site response conditioned on relatively simple site parameters such as the time-averaged shear wave velocity (V_S) to 30 m depth (V_{S30}) and the depth to 1 km/s or 2.5 km/s V_S ($z_{1.0}$ or $z_{2.5}$) (Bozorgnia et al., 2014). These models are referred to as ergodic (Anderson and Brune, 1999) even if the site parameters are measured on site. The underlying models are ergodic because they are derived from large global or regional databases, and as such are not site-specific.

Any particular site would be expected to produce site amplification that departs from the ergodic estimate for a variety of reasons related to location-specific geologic conditions. A site amplification model that accounts for the effects of these features on site amplification is non-ergodic (e.g., Atkinson, 2006; Stewart et al., 2017). One common feature of non-ergodic site response is resonance at one (fundamental site frequency, f_0) or more site frequencies (Di Alessandro et al., 2012; Bonilla et al., 2002; Bonilla et al., 1997), which produce peaks that are smoothed out in ergodic models. While not currently used in NGA models nor in general practice, horizontal-to-vertical Fourier amplitude spectral ratio (HVSR) vs. frequency plots have the potential to add this site-specific attribute to predictions of ergodic site response at low cost, relative to non-ergodic procedures. While V_{S30}-based models provide reasonable, first-order estimates of site response over a wide frequency range (e.g., Borcherdt 1994, Choi and Stewart 2005, Seyhan and Stewart 2014, Parker et al. 2019), attributes of HVSR peaks are expected to be effective at describing site amplification for frequencies proximate to f_0 and to have limited utility elsewhere. Hence, the two parameters serve different purposes and we postulate that they can be most effectively utilized together (Cadet et al., 2012; Ghofrani et al., 2013).

Current HVSR-based site amplification models, whether using HVSR parameters solely (e.g., Zhao and Xu 2013; Hassani and Atkinson 2016), or in combination with V_{S30} (e.g., Cadet et al. 2012; Ghofrani et al. 2013; Kwak et al. 2017; Hassani and Atkinson 2018a, 2018b), are derived using HVSR computed from the same earthquake ground motion data that is being predicted by the model. This model development practice is inconsistent with how the models would be used in forward applications, which will typically be for sites without earthquake recordings.

We posit that for the use of HVSR to gain traction in California practice, several technical issues need to be addressed. Among these are the following:

1. Best practices for collecting and analyzing HVSR data need to be developed and agreed upon by the informed technical community.
2. A database of HVSR data, assembled to the extent possible in a manner consistent with best practices, should be provided and made publicly available.
3. Procedures for identifying when HVSR peaks are present and should be used in model development are needed, as well as procedures for characterizing those peaks (i.e., frequency, amplitude, width).
4. The reliability of HVSR peaks as established from a particular noise-based measurement is needed, under the assumption that the measurement is made by a credible analyst. The issue in this case is the repeatability of HVSR when measured from noise with different equipment or at different times.
5. An understanding of the consistency of HVSR peaks as established from earthquake data and noise is needed. Noise-based measurements will dominate practical forward applications, but they are intended to predict earthquake shaking attributes. As a result, consistency between HVSR from these two data sources is desirable.
6. Development of HVSR-based site amplification models conditioned on interpretations of HVSR data (i.e., identification of peaks, peak parameters), in combination with V_{S30} and perhaps sediment depth, is needed.

The aforementioned models derived from ground motion-based HVSR in effect assume that earthquake- and noise-based HVSR are perfectly consistent (Issue 5) and that noise-based HVSR measurements are fully repeatable (Issue 4). This report presents the results of research on the first five issues described above.

1.2 OBJECTIVES, SCOPE, AND REPORT ORGANIZATION

Our long-term research objective is to complete the sequence of work described in the previous section. The work described in this report was completed under a research agreement with the California Strong Motion Instrumentation Program (CSMIP) and was focused on the development of a database of credible HVSR data (Items 1-2 above) and analysis of the data aimed at providing reliable, repeatable HVSR-based site parameters (Items 3-5). Model development work utilizing this data has not been completed as of this writing, and it is anticipated to be the topic of future work.

The HVSR database is an extension of the V_S profile database (PDB), an early version of which is described by Ahdi et al. (2018). Chapter 2 of this report presents the data sources, signal processing procedures, and analysis procedure used to compute HVSR. In Chapter 2, we also

describe external (to the database) routines that can be used to evaluate the presence of peaks and identify HVSR-related parameters used for site response studies.

Chapter 3 presents analysis, based on data compiled to date, of the 4th and 5th issues above. The aim is to evaluate the consistency of HVSR as derived from different data sources, in particular microtremors from permanent instruments and temporary arrays at the same site, and microtremor vs earthquake signals. Implications of the findings for model development are discussed. Chapter 4 summarizes the scope and principle findings of this study and recommends needed future work to facilitate the utilization of HVSR for ground motion prediction in California.

2 HVSR Database

An HVSR relational database has been created as an additional feature of the shear wave velocity profile database. Users can access and visualize it through the website: https://doi.org/10.21222/C27H0V. Locations having HVSR can be identified by selecting, on the left side of website, the "Spectral Ratio" check box. Users can then navigate the map interface to plot and download HVSR data. In the following, we describe what specific data are saved in the relational database behind the website and how we processed, organized, and maintain the data. We also describe a Jupyter Notebook web application that provides more flexibility and functionality for data analysis.

2.1 INSTRUMENT TYPES

The database is structured to allow entry of HVSR data primarily from three sources of recordings:

- Source 1 (mHVSR-T): Temporarily deployed three-component (3C) seismometers—commonly known in practice as the single-station site characterization method—to specifically record microtremors (e.g., Yong et al., 2013; Molnar et al., 2018).
- Source 2 (mHVSR-P): Permanent or semi-permanently-installed 3C high-gain broadband (BB) instruments, with triaxial orientations denoted as HH* components (* corresponds to E for east, N for north, or Z for vertical) (SEED Manual 2012). The instruments are installed in vault-type housings and operated by regional seismographic networks to monitor local and teleseismic ground motions, which include both earthquake and microtremor ground vibrations. Often these instruments continuously stream data that are captured remotely (Ringler and Bastien, 2020).
- Source 3 (eHVSR): Seismic strong motions (SM) (e.g., Hassani et al. 2019).

In the above, mHVSR indicates microtremor HVSR.

mHVSR-T data (source 1) are typically obtained from 3C seismometers, including BB seismometers and geophones with corner frequencies (f_c) typically ranging from 1 to 5 Hz. Most of the recordings in SM databases are from accelerometers, some of which currently operate with

continuous streaming and others of which are triggered. For modern network-operated earthquake monitoring stations, these SM accelerometers are often co-located with relatively sensitive BB seismometers. For such sites, we collect mHVSR data from continuously streamed BB seismometers (source 2). The mHVSR data included in the database are derived from BB seismometers and geophones with $f_c = 2$ Hz. Microtremor measurements from accelerometers has been collected as part of this study, which are analyzed in Section 3.3.1, but were not found to produce usable mHVSR results. As a result, accelerometer-based mHVSR data are not included in the database.

Figure 2.1 describes the bandwidth and gain for sensors considered in the present research, some of which are commonly encountered at permanent ground motion monitoring stations in California. Different colors and line types correspond to different sensors. The sensors shown in Figure 2.1 are (1) BB seismometers (i.e., extensive flat plateau at the sensor's maximum gain, extending to frequencies of 0.1 Hz or lower) and high gain (plateau level of at least 750 V/[m/s]) as well as (2) geophones with more limited bandwidth (f_c~2 Hz) and moderate gain. The first group of BB seismometers include Strekeisen STS-2, Güralp (CMG-3T, CMG3-ESP, CMG-40T), and Nanometrics Trillium (40, 120, and 240), which are commonly the source of mHVSR data streamed continuously from the permanent network monitoring stations (source 2). Such sensors provide good signal resolution for mHVSR measurements. The second group of sensors are geophones with lower gain, the only example of which in Figure 2.1 is the 2-Hz PS-2B geophone (22 V/[m/s]). Similar geophones have been shown to provide good resolution to frequencies as much as a factor of 10 or more below f_c (Strollo et al. 2008; Chatelain and Guiller 2013; Yong et al., 2013). Sensor type is included as metadata in the relational database.

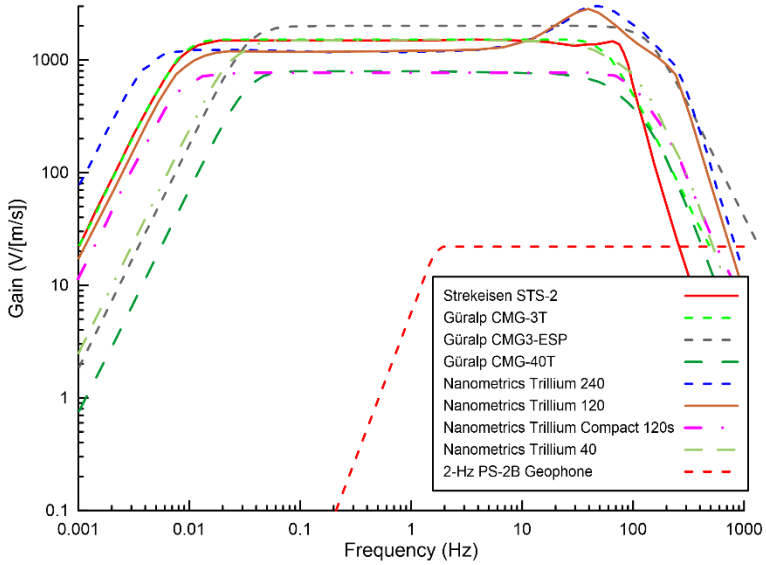


Figure 2.1. Sensor responses for BB seismometers (adapted from IRIS PASSCAL [2021]) and geophones used to obtain microtremor horizontal-to-vertical-spectral-ratios (mHVSR) data for the relational database.

2.2 HVSR SITE INVENTORY

The largest known inventory of mHVSR-T data at strong motion stations is Yong et al. (2013). The study (aka: the American Recovery and Reinvestment Act funded project; hereafter referred to as the ARRA project) presented data from 191 strong-motion stations, the majority of which are located in California (187 stations), with an additional four stations in the central and eastern United States. The ARRA data was provided as time-domain signals, which were processed in the manner described in the next section. Yong et al. (2013) provided 588 mHVSR 3C recording sets for the 191 sites, which is due to multiple measurements at most sites. Another major data source is Geometrics, Inc. (K. Hayashi, personal written communication, 2020) which shared mHVSR data from 638 sites, 281 of which are in California. This data was provided in XML files, which has been extracted and added to the database. Additional mHVSR-T contributions to the database include:

- 33 ground motion accelerograph sites in the Sacramento-San Joaquin Delta region of California (Buckreis et al. 2021).
- 40 ground motion accelerograph sites maintained by the California Strong Motion Instrumentation Program (CSMIP), part of the California Geological Survey (CGS). Reports are from GEOVision (2016; 2018) and Petralogix (2017).
- 24 sites, some of which are ground motion stations and others are of sites of commercial engineering interests, investigated as part of non-ergodic ground motion investigations by ENGEO, Inc. (D. Teague, personal written communication, 2020).

Time series data from the Delta sites was processed as described in Section 2.3. For the CSMIP and ENGEO sites, we obtained mean mHVSR-frequency plots, which were digitized and added to the database.

For mHVSR-P data, we queried three data centers: Incorporated Research Institutions for Seismology (IRIS), Southern California Earthquake Data center (SCEDC), and the Northern California Earthquake Data Center (NCEDC) (IRIS, 2020; SCEDC, 2013; NCEDC, 2014). We sampled continuously streamed data for 402 stations instrumented with high-gain BB seismometers with sampling rates between 80-250 Hz. The time series from these data were processed using procedures in Section 2.3.

Altogether, the database currently contains mHVSR data for 941 California sites, locations of which are shown in Figure 2.2. Many of these sites, including all of the ARRA sites, have mHVSR from both temporary and permanent sources, which causes the number of mHVSR entries (1352) to exceed the number of sites (941). Of the 941 sites with mHVSR, 668 are located in the immediate vicinity of strong motion stations. There are also 363 sites with mHVSR data located outside of California.

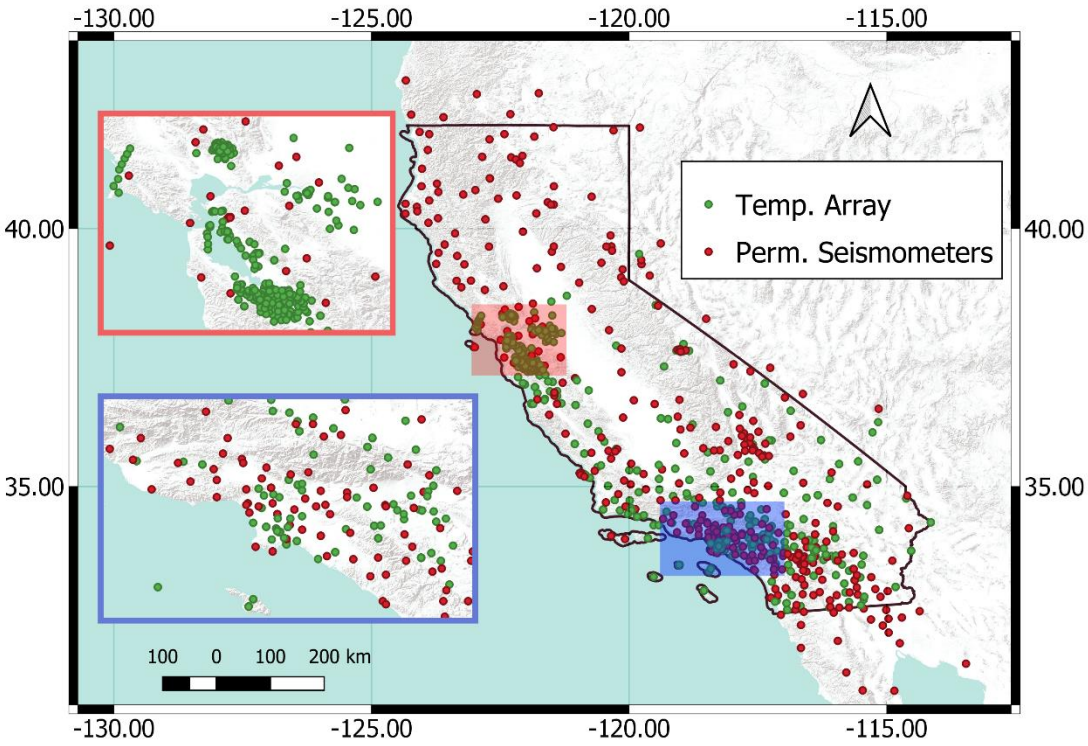


Figure 2.2. Locations of sites in PDB with mHVSR from either temporary deployments (mHVSR-T) or continuously streaming ground motion sensors (mHVSR-P).

Prior to the present effort to assemble an HVSR relational database, a similar database for sites with shear wave velocity (V_S) data had been compiled by Ahdi et al. (2018). That database had around 1,500 V_S profiles in California and 2465 overall. The present work started with the organizational structure (schema) from that database, including the web portal and map interface. Figure 2.3 shows the relative number of V_S profiles and mHVSR sites for California sites in the database as of April 2021. Whereas various techniques have been used to collect profile data since the 1960s, the collection of mHVSR data is much more recent. The abrupt increase in volume of microtremor data is primarily from the present project, mainly using permanently installed BB sensors.

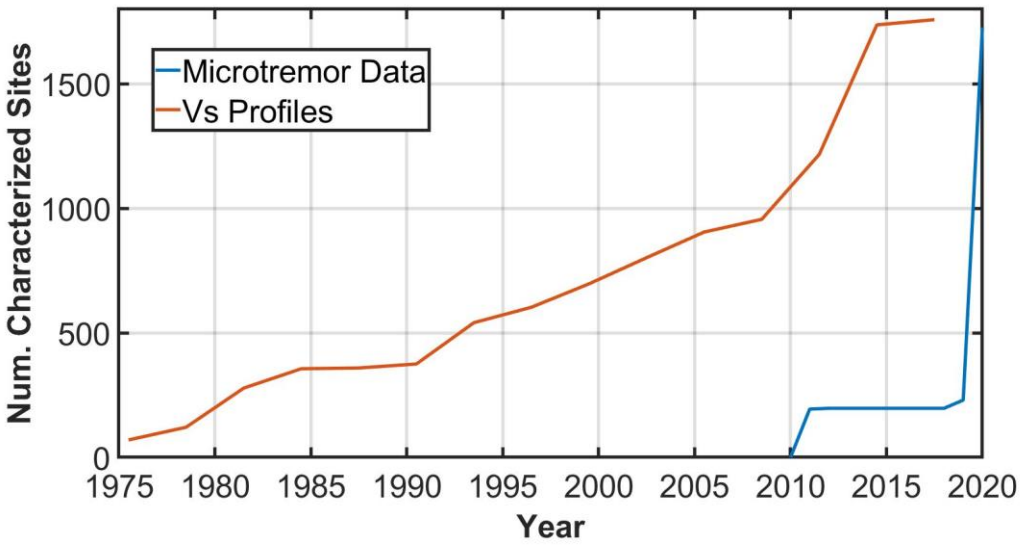


Figure 2.3. Cumulative distribution of V_S profiles and HVSR data in California versus time.

2.3 PROCESSING PROCEDURES

This section describes the processing procedures used to convert time-domain signals to HVSR. These procedures adopt guidelines from **S**ite **E**ffect**S** assessment using **AM**bient **E**xcitations (SESAME 2004). When time-series data are available (all mHVSR-P data and mHVSR-T data from Yong et al. 2013 and from the Delta region), signal processing for mHVSR was performed using a processing package, *hvsrProc*, written in R (Wang, 2021). That package implements the procedures described here. When time series data was not available, we adopted mHVSR (generally geometric mean) from mHVSR-T data providers. Those analyses were generally performed using the open-source software analysis code Geopsy (Wathelet et al, 2020). Another processing code used in recent applications is *hvsrpy* (Vantassel, 2020), which implements procedures described by Cheng et al. (2020) and Cox et al. (2020). The package operates similarly to Geopsy, but also implements a new automated algorithm for anti-triggering.

2.3.1 Microtremor Measurements

Data processing using the *hvsrProc* script, Geopsy, and *hvsrpy* follow the same general steps, which are outlined in this section. The *hvsrProc* and Geopsy procedures differ from each other only for the steps of (1) combining horizontal components; (2) computing mHVSR amplitudes; and (3) resampling and decimation. The *hvsrpy* procedure differs from the other two mainly in the anti-triggering step.

Number of Windows and Cycles

Microtremor signals used to derive mHVSR are recorded over long durations (typically hours). The mHVSR peak frequency should be greater than 10 divided by the window duration in sec (SESAME, 2004). The total number of significant cycles is defined as $N_{\mathrm{cyc}} = T_{\mathrm{win}} f_0 N_{\mathrm{win}}$, where T_{win} is window length (in sec), f_0 is the frequency (in Hz) of the lowest prominent peak in the HVSR spectrum, and N_{win} is the number of windows used in the mHVSR spectrum computation. It is good practice to have no fewer than 200 cycles in the time series used for Fourier series computations, which effectively sets a minimum signal duration ($T_{\mathrm{sig}} = N_{\mathrm{cyc}}/f_0$). For typical applications in which f_0 is not known at the time the window durations are set, the minimum signal duration can be taken as $T_{\mathrm{sig}} = N_{\mathrm{cyc}}/f_{\mathrm{min}}$, where f_{min} is the minimum frequency where we seek to compute reliable HVSR ordinates. In most cases, f_{min} is taken as 0.1 Hz.

Table 2.1 shows typical values for the above parameters. It is important to note that parameters can be manipulated to ensure that the number of significant cycles is larger than 200. To ensure that low-frequency peaks can be identified, mHVSR-T recordings are usually at least 1 hour in duration (and about 2 hours for the Delta sites) and mHVSR-P data are obtained for at least 2 hours.

Table 2.1. Recommended recording duration, assuming at least N_{cyc} = 200 and N_{min} = 10 [modified from SESAME (2004)].

f_{min} (Hz)	Minimum value for T_{win} (s)	Recommended minimum record duration T_{sig} (s)
0.1	150	3000
0.2	50	1800
0.5	20	1200
1	10	600
2	5	300
5	5	180
10	5	120

Geopsy provides bad sample tolerance and threshold options that help the user optimize the number of windows (Wathelet et al., 2020). These bad sample options allow windows to be selected that do not satisfy the anti-triggering criteria described subsequently. The bad sample tolerance allows the user to define the number of bad samples which can remain in a usable window. Due to the relatively long signal durations used in the present work, we were able to exclude data from questionable time windows, thus obviating the need to apply these options.

Window Overlap; Taper Width and Type of Window

Sometimes the signal duration is not long enough, and the windows may be too short in duration to satisfy suggested window lengths in Table 2.1. To adjust for this, time windows can overlap by a specified percentage (Wathelet et al., 2020). For the present database, overlapping is

not applied as signal durations are long enough to satisfy the suggested durations. We use cosine tapers with a length of 5% of the window length (Chatelain et al., 2008).

Anti-Triggering

"Triggering" refers to a temporary vibration source affecting a signal, which can compromise the accuracy of mHVSR. Ground vibrations from far-field microtremors typically have approximately constant amplitudes in time. In contrast, local vibrations induced by surface sources such as traffic or other anthropogenic sources will produce transient, erratic bursts in the recorded signals. Anti-triggering is used to remove intervals of the signal with potential triggers, with the objective of ensuring approximately constant amplitudes in time.

The presence of potential triggers within a window of the recorded signal is judged based on relative values of the short-term average (STA) and long-term average (LTA) signal amplitudes as well as by visual inspections. The STA and LTA are computed using 5- and 30-sec durations, respectively. The SESAME (2004) guidelines call for the amplitude ratios to be within the range of STA/LTA from 0.1 to 10, although Chatelain et al. (2008) recommend 0.01 to 10.

During signal processing, we look for stationary (i.e., approximately constant amplitude in time) intervals of ambient vibrations. Removing windows with transient signals produces clearer mHVSR peaks and reduces variance. The anti-triggering algorithm is typically applied to both horizontal and vertical components, and if the data from a window is found to be problematic in any component, it is removed for all three components. Additional visual checks are performed in the frequency-domain by examining HVSR for individual windows; windows with amplitudes judged to be obvious outliers (too large or too small) compared to the mean curve are excluded. This additional check can sometimes identify problematic windows that have no obvious deficiencies in the time domain.

While the anti-triggering algorithm can be applied to either the unfiltered or filtered microtremor signals, here we apply it to the raw (pre-filtered) signal (consistent with procedures used in Yong et al., 2013). Within the metadata table we provide the durations used to derive the STAs and LTAs, and the STA/LTA amplitude range.

The *hvsrpy* package uses a different approach to accept or screen out widows to be included in the HVSR analysis (Cox et al., 2020). In this approach, the objective is to minimize the window-to-window dispersion of the identified site frequency f_0. Hence, windows are removed when they increase the variance of f_0. Application of this algorithm may also narrow the shape of the peak itself. This approach does not directly consider "trigger" indicators evident from time- or frequency-domain inspections to identify problematic windows, as applied in other procedures. We have not adopted the algorithm in *hvsrpy*. Our concern is that HVSR dispersion, including near the site frequency, is a natural feature of site signature. The algorithm potentially obscures such features.

Filter

Filtering is applied to reduce low-frequency drift in waveforms. The high-pass corner frequency can be selected manually for each signal in each of the processing codes. The high-pass corner frequency is usually taken around 0.1 Hz, which is near the instrument corner for broadband seismometers and below it for geophones. The corner frequency for each signal is provided as metadata. The upper-bound frequency that can be resolved from the data is the Nyquist frequency and low-pass filtering is not applied.

Theoretically, the application of filtering should not affect mHVSR since the same filter is applied to the horizontal and vertical components. However, some effects of filtering on HVSR are observed, as shown for example in Figure 2.4 for mHVSR at the CE_67910 site below 0.1 Hz. These effects are likely caused by the smoothing operator (next sub-section).

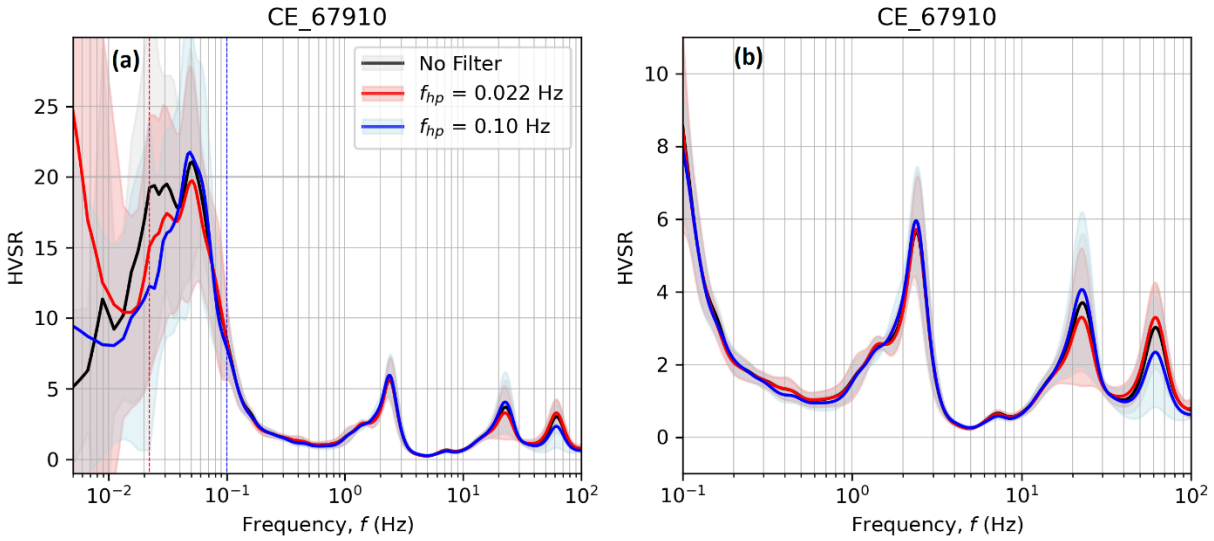


Figure 2.4. Comparison of mHVSR when different high-pass corner frequencies (f_{HP}) are used in record processing. (a) HVSR for wide frequency range (0.0005 to 100 Hz) showing differences below 0.1 Hz; (b) HVSR for more typical frequency range (0.1 to 100 Hz) showing negligible differences.

Smoothing Type and Constant

Spectral smoothing reduces high frequency noise and can facilitate identification of peaks. The Konno and Ohmachi (1998) smoothing filter (KO smoothing), which accounts for variable numbers of points at low frequency, is typically used and is applied to the combined horizontal and vertical components. Geopsy and *hvsrProc* provide different smoothing operators [e.g., KO smoothing, Parzen-window method (Parzen, 1961)] and the smoothing type is a metadata field in the database.

Currently, all HVSRs in the database are processed by KO smoothing, which is applied to the HVSR for each window. The degree of KO smoothing increases as the bandwidth decreases. Chatelain et al. (2008) uses a bandwidth parameter of 40. We typically use a value of 30 and

change this parameter depending on the quality of the data over the range of 20–40. Noisy data requires a lower bandwidth parameter (which produces a greater degree of smoothing).

Horizontal Component Combination Method

Because horizontal ground motions are recorded in two directions, a method to combine these components is required. Our preferred methods of representing horizontal-component motions are (1) median component (RotD50; Boore, 2010), which mirrors applications in ground motion studies and (2) variable-azimuth components. The *hvsrProc* package developed for this research provides these outputs and allows the user to select how many azimuthal HVSRs are calculated (the default is 18, separated by 10 degrees from north to south). Geopsy version 3.4.1 provides the geometric mean and squared average (i.e., square root of the sum of the two squared components).

HVSR Calculation

HVSR is computed as a function of frequency by dividing the smoothed RotD50 (or geometric mean) horizontal-component Fourier amplitudes by the smoothed vertical-component Fourier amplitudes. This computation is performed for each window, the results of which are then averaged across windows to produce a mean curve. This averaging provides smoothing beyond that provided to the individual components (the selection of the KO smoothing parameter accounts for the combined effect). The dispersion of HVSR ordinates across windows can be computed in *hvsrProc* as an arithmetic or natural log standard deviation, whereas Geopsy provides only the natural log standard deviation. Dispersion is represented as arithmetic standard deviation for mHVSR data in the database, based on analyses of data distributions in Section 3.4.

Section 2.4 describes database fields containing HVSR amplitudes for various azimuths (every 10 degrees from true north to south). Subsequently in the report, plots showing the variation of HVSR amplitudes with both frequency and azimuth are referred to as *polar curves*.

Resampling and Decimation

The frequencies (Hz) returned from a Fast Fourier Transform are evenly spaced on a linear scale. In contrast, HVSR is usually plotted on a log scale for frequency (e.g., Figure 2.4). To provide a uniform level of resolution on a log frequency scale, which can be important for the fitting of HVSR peaks, we resample the mean curve using linear interpolation. The impact of this resampling is to increase frequency resolution at low frequencies and decrease resolution at high frequencies (the number of samples does not change).

Decimation can be performed to reduce the size of the arrays stored in the database. For example, a 2-hour duration microtremor survey by a sensor with sampling rate of 100 points per sec, using a window length T_{win} = 150 sec, provides 150 × 100 / 2 = 7500 Fourier Amplitudes for each of 48 windows, plus the mean and mean ± one standard deviation curves, with additional data storage required with azimuthal results. To limit storage requirements, we incorporated a

decimation operator into the *hvsrProc* program, which selects one out of every n points, and investigated the effects of alternate values of n. The value of 10 was found to produce satisfactory results and was used to populate the database for sites where we had time series data.

The resampling and decimation steps can be combined into a single step in which the linear interpolation is applied directly at the desired frequencies (decimated frequency points with evenly spaced on a log scale). This approach is applied in Geopsy (users specify the desired frequency step).

2.3.2 Earthquake Measurements

HVSR can also be measured from earthquake recordings (denoted eHVSR), and the database has been configured to accommodate eHVSR results. The analysis of data for derivation of eHVSR should begin with ground motion processing, such as that used for NGA projects (e.g., Kishida et al., 2020). This should occur outside of the *hvsrProc* package; the time series data can then be imported into the program. Post-processing procedures are then applied to the records within the program in the same manner as for mHVSR, which include the analysis of individual components of the Fourier amplitude spectrum (FAS), the calculation of mean curve and polar curves, and resampling.

The duration of earthquake ground motion recordings is usually about 30 sec to 2 min. Sub-windowing of these signals is impractical, so multiple earthquake events are required to provide multiple "windows" to meet reliability criteria. We recommend a minimum number of earthquake records per site of 10.

As with mHVSR, the *hvsrProc* package can calculate the eHVSR mean curves, polar curves, and their uncertainties. In the case of eHVSR, based on data analyses in Section 3.4, we recommend the use of logarithmic standard deviation terms.

Although the database provides a framework for storing eHVSR data, such data has not yet been uploaded. The collection and dissemination of eHVSR data has not been prioritized because eHVSR is not an independent metric of site condition. The dependence in this case arises from the site parameters being derived from the same data they are intended to predict.

2.4 DATABASE SCHEMA

For the V_S profile database (PDB), a relational database was adopted as the means by which to organize and archive information (Ahdi et al., 2018). This project added mHVSR to the PDB, which required adding some tables to the existing database schema. The database has been developed using the My Structured Query Language (MySQL) relational database management system. Within the natural hazards community, there are many examples of "databases" that consist of non-structured data collections presented in the form of spreadsheets or text files.

Structured relational databases represent a different tool to store data. Relational databases have a hierarchical structure that defines relationships among different tables. Data are stored in tables in a series of fields (or columns). The tables within the database are linked together through primary and foreign keys. Primary keys represent unique identifiers of each entry in a table. Hence, one primary key can only be used once in each table. A foreign key is a field in one table used to identify a record in another table. Foreign keys are used to link different tables to each other. Relational databases were introduced by Codd (1970), and some advantages include avoiding redundancy and null fields, consistency (information is entered only once), and security (if a database crashes, information is saved) (Codd, 1970; Brandenberg et al., 2020).

The tables related to HVSR data in the PDB are listed in Table 2.2. There are two categories: general information and geophysical data. The meaning of the table names in Table 2.2 are described below. Figure 2.5 shows all tables, specific fields, and the primary and foreign keys in each table.

Table 2.2. Different group and table types and the number of fields in the HVSR schema.

Group Type	Table Type	Number of Fields
General	site	18
	citation	5
Geophysical	spectralRatioMeta	28
	hvProcessing	23
	meanCurve	5
	hv_Data	3
	azimuthVariation	3
	polarCurve	5

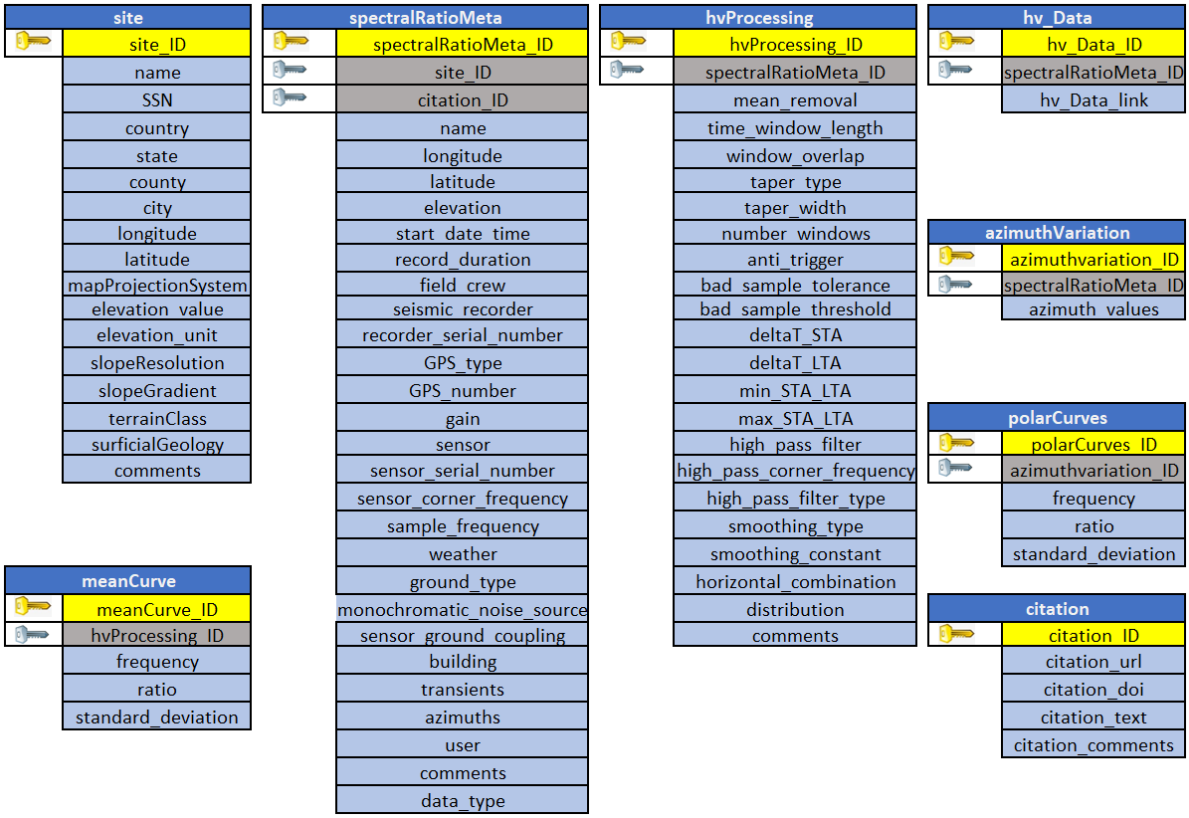


Figure 2.5. Tables, fields, and primary (gold) and foreign (gray) keys in HVSR database schema. Site table is taken from the V_S Profile Database schema (Ahdi et al., 2018).

2.4.1 Site and Measurement Metadata

Site metadata are provided in the site table, which identifies the site by name, provides location data, provides the NGA station sequence number (as used in NGA site tables; Seyhan et al. 2014; Ahdi et al. 2020) for sites with ground motion instruments, and provides various proxies for V_{S30} estimation related to surface geology units (e.g., Wills et al., 2015 for California sites) and terrain classes (Iwahashi and Pike, 2007, which is available globally). The site is identified within the database using the site_ID primary key. This site table is common to HVSR and V_S sites (a site with both types of data would have a single site_ID).

HVSR measurement metadata are provided in the spectralRatioMeta table, which provides information on how HVSR signals were recorded or collected, such as measurement location and time, instrument information, and field notes. Primary key spectralRatioMeta_ID is assigned to a data set and is connected to a site using site_ID as a foreign key. Locations are provided in both the site and spectralRatioMeta tables; the former is for a general location and is used for mapping purposes only, the latter is the specific location of the HVSR measurement. The type of data used to derive HVSR is indicated by the data_type field: 0=microtremors recorded by temporarily deployed single station (mHVSR-T); 1=microtremors recorded by permanently installed BB

instruments (mHVSR-P); and 2=earthquake signals. Some of the columns in the spectralRatioMeta table may be null, depending on data type. For example, for data_type = 2 (eHVSR), the start_date_time will be null as it is not unique. This is because the records used to derive HVSR mean curve and polar curves are collected from multiple different earthquakes without a unique start time. The data set can be referenced by assigning the citation_ID foreign key.

2.4.2 Signal Processing Metadata

The hvProcessing table provides the data processing parameters used to produce HVSR curves. The horizontal_combination field indicates the approach used to combine horizontal components (e.g., "RotD50", "GeoMean", and "SquaredAverage" can be entered as text strings). The smoothing_type field indicates the smoothing operator (e.g., "KonnoOmachi" and "Parzen", again as text strings). The type of HVSR ordinate distribution across windows (e.g., "normal" or "lognormal") is given as text strings in the distribution field, and affects the types of mean and standard deviation used in the meanCurve table. Other parameters are as described in the Section 2.3. The primary key is the hvProcessing_ID and the foreign key is spectralRatioMeta_ID.

2.4.3 HVSR Data Table

The hv_Data table is used to store links to the raw time series data from which HVSR curves are calculated. The links are to servers that are external to the database. When time series data are obtained from public data repositories (e.g., the ARRA data and data downloaded from IRIS or other network servers), the data are cited using a citation_ID and does not appear in hv_Data. When time series data are not permanently available on servers (e.g., from testing performed by an individual investigator or consulting firms), an hv_Data entry is provided to a digital object identifier (DOI) set up for that specific project. For example, the recorded time series data for sites in the Delta region of California (Buckreis et al. 2021) are archived in the DesignSafe cyberinfrastructure (Rathje et al., 2017) with a DOI; the url to this time series data is provided in hv_Data for the applicable sites. The primary key in hv_Data is hv_Data_ID and the foreign key is the spectralRatioMeta_ID.

2.4.4 Mean Curve Table

The meanCurve table provides HVSR ordinates for the horizontal-component combination method selected in the horizontal_combination field in the hvProcessing table (e.g., RotD50, geometric mean, or squared average). The frequency field is for the frequencies where ordinates are provided. The ratio field is for the mean HVSR ordinate, which may be computed as an arithmetic mean or natural log mean of results across windows. The type of mean should be compatible with the standard deviation of ordinates across windows, which is provided in the

standard_deviation field. The distribution type for which the mean and standard deviation are provided is given by the distribution field in the hvProcessing table. The primary key is the meanCurve_ID and the foreign key is hvProcessing_ID.

When a site is selected through the map interface, a window opens that allows HVSR data to be plotted. This tool shows the mean and mean ± one standard deviation ordinates vs frequency using the data in the meanCurve table. Figure 2.6(a) shows an example RotD50 mHVSR for the CI.GR2 site (Griffith Park Observatory) in Los Angeles, California (figure generated using tools described in the next section).

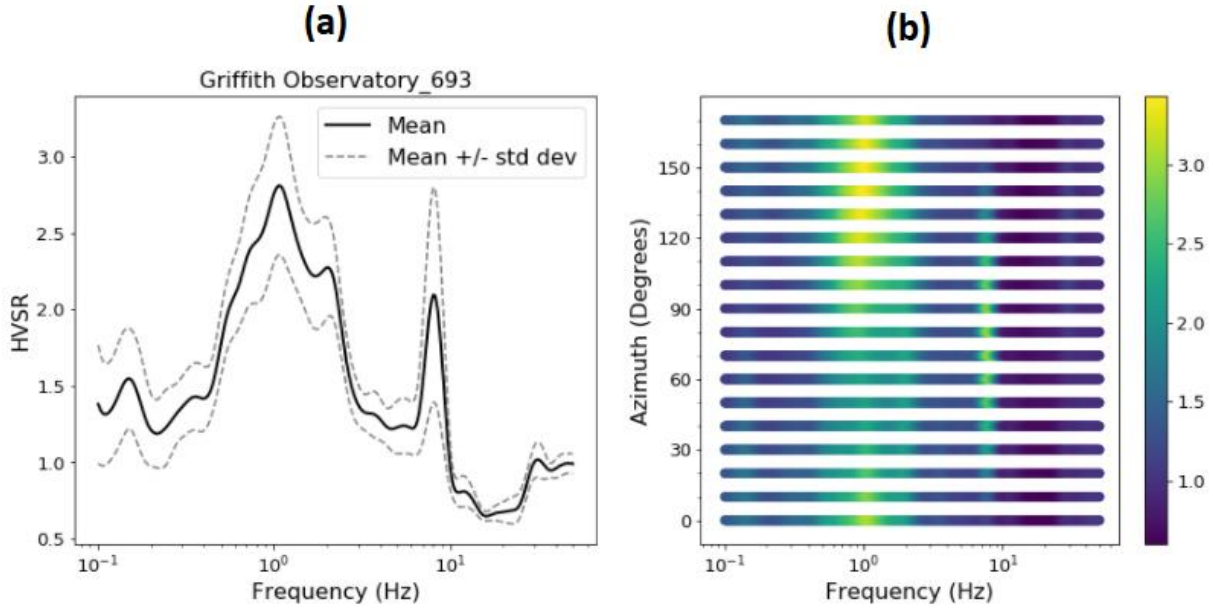


Figure 2.6. HVSR plots for a site near the Griffith Park Observatory in Los Angeles (CI.GR2): (a) RotD50 mean and mean ± one standard deviation HVSR ordinates, and (b) polar plot showing mean of azimuth-dependent mHVSR ordinates.

2.4.5 Azimuth Variation and Polar Curves Tables

The azimuthVariation table includes azimuth values from 0 to 180 degrees in varying increments, typically around 5–10 degrees. The primary key and foreign key are the azimuthVariation_ID and spectralRatioMeta_ID, respectively.

The polarCurves table contains the data (frequency, ratio, standard_deviation) for the azimuthVariation values. The primary key is polarCurve_ID and azimuthVariation_ID is the foreign key. Polar curves are generated by rotating the two horizontal components at selected azimuths. In the database, we typically store HVSR polar curves at 10-degree intervals (i.e. 18 polar curves – 0–180 degrees – for each site). Figure 2.6(b) shows example results of polar curves for the Griffiths Park site.

Polar curves are often used to detect sites where site conditions (possibly topographic effects) may produce amplification effects due to wave-field polarization. For example, site CI.GR2 is located near the crest of a ridge in the Santa Monica Mountains. As shown in Figure 2.7, the ridgeline axis follows an approximate azimuth of 0 deg (north-south), whereas the azimuth of 90 deg is oriented down-slope. The 1 Hz resonance (apparent first mode) clearly visible in Figure 2.6(b) spans across all azimuths but is strongest in the 110-170 deg azimuth range. We have not investigated the origins of this feature, but given its low frequency it is unlikely to reflect local topographical features. On the other hand, the second peak at approximately 7 Hz is centered on a 90-deg azimuth. In this down-slope direction, we expect topographic amplification effects to be strongest (Di Giulio et al., 2009).

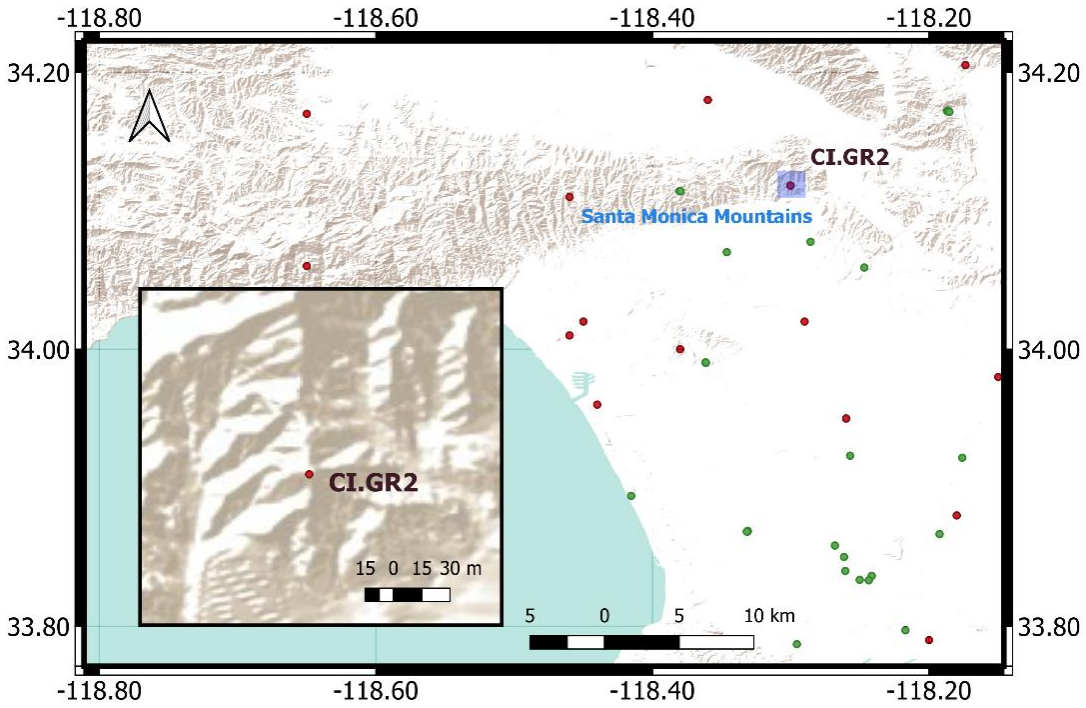


Figure 2.7. Relief map showing location of CI.GR2 station relative to Hollywood Hills ridgeline.

2.4.6 Citation Table

The citation table provides fields to reference URLs, DOIs, text, and comments related to the citation. The primary key is citation_ID. The citation_ID field is used to link metadata to its citation and is a foreign key in the spectralRatioMeta table.

2.5 DATA INTERPRETATION TOOLS

2.5.1 Data Access and Visualization within Database

The website graphical user inference (GUI) for the HVSR database provides mean (arithmetic or natural log) and mean ± one standard deviation HVSR ordinates vs frequency, which can be plotted or downloaded as comma-separated values (.csv) files. Figure 2.8 shows a screen shot of the information available directly from the online HVSR database for the CI.GR2 site. The azimuthal variations of ordinates, while tabulated in the database, are not provided in the downloadable csv file.

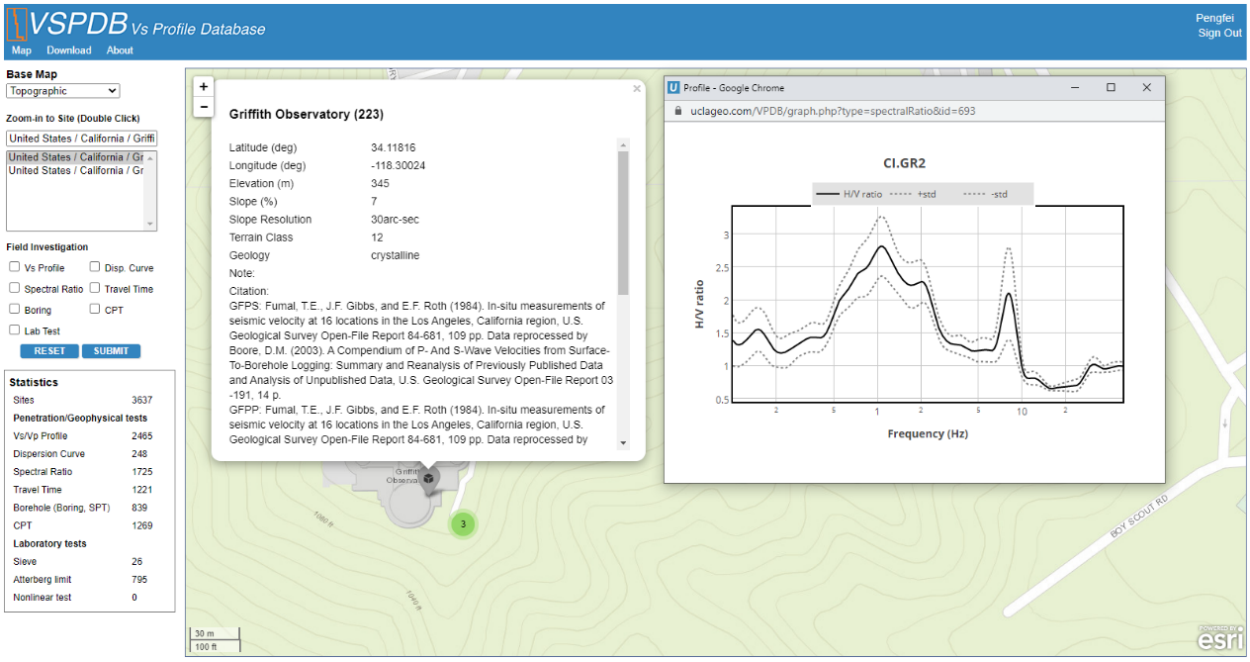


Figure 2.8. Example data windows for site in database with HVSR data.

2.5.2 Data Analysis Using External Tools

To facilitate common applications, the HVSR database is replicated daily to DesignSafe (Rathje et al. 2017) where it can be accessed using Python scripts in Jupyter notebooks. An example Jupyter Notebook tool called *HVSR_Viz_Query* (Wang et al., 2021) demonstrates data interaction, and example output from *HVSR_Viz_Query* for the Griffiths Park site is shown in Figure 2.9. The interpreted parameters include (1) identification of features as peaks; (2) plots of azimuthal variations of HVSR; and (3) for each peak in the median-component HVSR, fitting of a pulse function to evaluate peak frequency, peak amplitude, and width of peak.

Users interact with *HVSR_Viz_Query* via a GUI that is shown in Figure 2.9. The GUI has three segments. The first segment is to allow the user to select the HVSR measurement – the Griffiths Observatory site and Test ID 693 are selected in this example. The second segment is used to set up parameters for plots and outputs, where plots of the mean curve and polar plots are selected. The third segment displays the two requested plots.

In the second segment, the Plot Mean Curve check box is enabled by default, whereas the other three check boxes are disabled by default to speed up querying the data and preparing plots. The figures are dynamically generated from the data based on user requests. Polar curves are provided by checking the Plot Polar Curves check box, as shown in Figure 2.9. To download HVSR data, the Download data button can be clicked and then a link to the data is provided. The second segment also has the following two check boxes:

- Run peak detection algorithm: Used to identify peaks in the HVSR spectral ratios using procedures presented in Section 3.2.
- Manually select frequency range for peak fitting: Used to fit peaks using procedures presented in Section 2.5.3.

At the bottom of the second segment, the text boxes, Minimum frequency (Hz) and Maximum frequency (Hz) define the frequency range used for peak fitting. These text boxes only need to be filled in when the Manually select frequency range for peak fitting check box is selected. The minimum and maximum frequencies control the locations of the flat tails of the peak function (for the example curve in Figure 2.9, these are 0.2 Hz and 3 Hz). Users can either run this notebook as-is if doing so suits their needs, or they can copy this notebook into their own DesignSafe directory where they can implement necessary modifications.

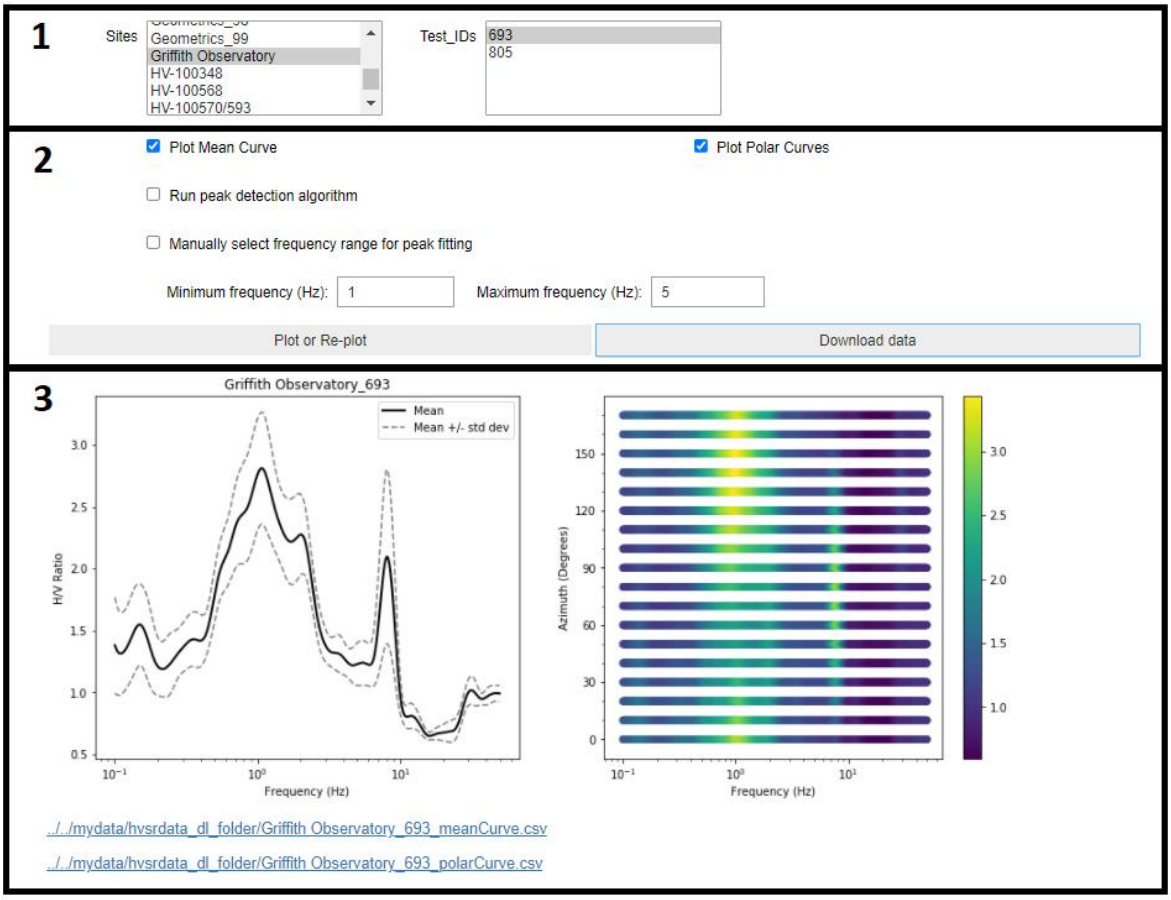


Figure 2.9. Screenshot of Jupyter notebook GUI for data access and analysis.

2.5.3 Peak Identification and Fitting

For mean HVSR plots with a peak, we fit a Gaussian pulse function adapted from Ghofrani and Atkinson (2014) as follows:

$$F_{H/V,i} = c_{0,i} + c_{1,i}\exp\left[-\frac{1}{2}\left(\frac{\ln(f/f_{pi})}{2w_i}\right)^2\right] \quad (2.1)$$

where f_{pi} is the fitted peak frequency, $c_{1,i}$ is the peak amplitude relative to $c_{0,i}$, w_i controls the peak width, $c_{0,i}$ is a frequency-independent constant indicating the amplitude of flat tails, i is the order of peak (only 1 or 2 can be selected), and f is frequency in Hz. The fit is performed using nonlinear regression in Python with the *optimize* function in the *Scipy* package, which minimizes the sum of squared errors. Fitting can also be performed in R using the *TreeReg.R* script, which was developed as part of this project and is included in the DesignSafe application (Wang et al, 2021). Figure 2.10 shows results for the CI.GR2 site, which contains a peak of amplitude 2.8 at frequency 1.05 Hz.

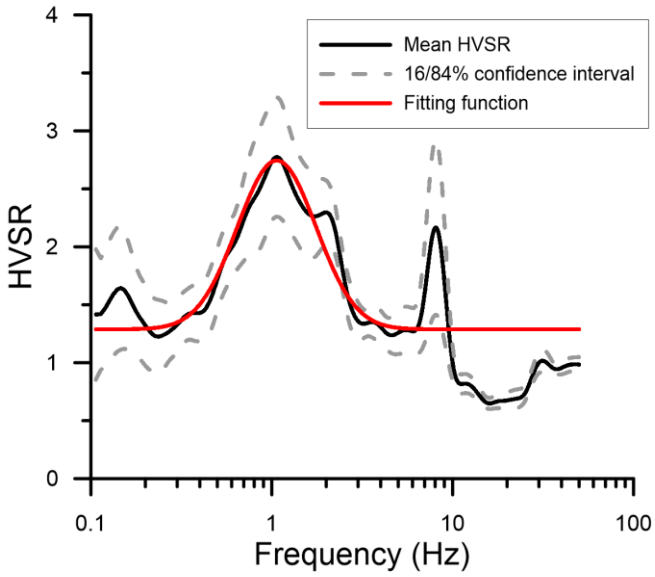


Figure 2.10. RotD50 HVSR for CI.GR2 site with Gaussian fit to the peak using Eq. (2.1).

There are two ways to execute the peak fitting algorithm – via an automated algorithm or via a manual selection process. The automated algorithm is applied by checking Run peak detection algorithm, which identifies whether peaks are present, and if so, fits peaks using the above procedure. Figure 2.11(a) shows an example output in Jupyter for the mHVSR data at the CI.GR2 site. The red line is the Gaussian fit curve and the fitted coefficients are provided. The automated algorithm will be discussed in detail in Section 3.2.

The manual approach is applied by checking Manually select frequency range for peak fitting, which requires the user to input the minimum and maximum frequencies that define the interval in which the peak function is fit to the data. This manual approach may be preferred when users are not satisfied by the outcomes of the automated algorithm. Figure 2.11(b) shows the output of the manual approach for the CI.GR2 mHVSR data with Minimum and Maximum frequencies of 0.2 and 4.0 Hz, respectively. The output coefficients are close to those returned by the automated algorithm in this case.

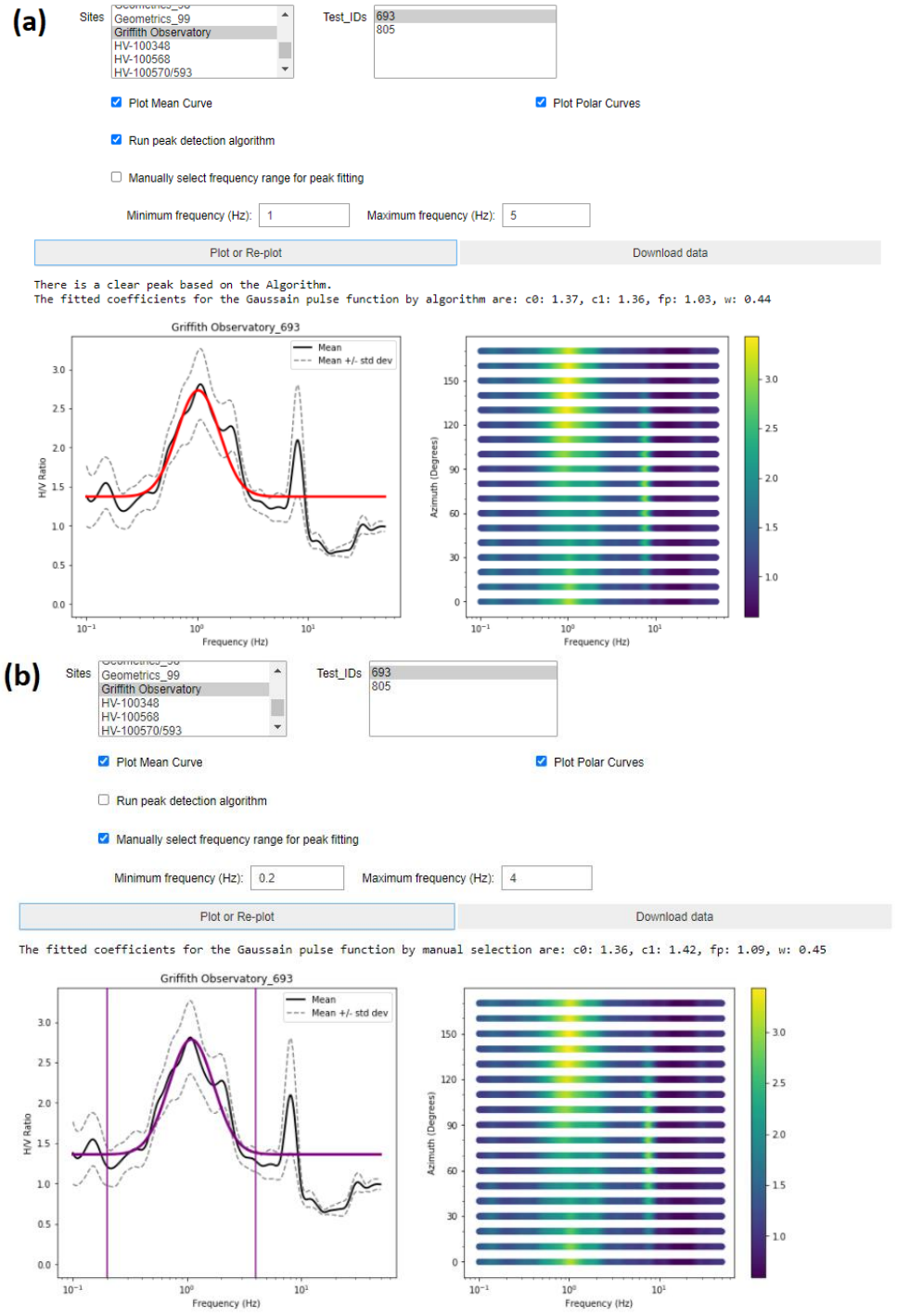


Figure 2.11. Output of Jupyter Notebook for peak detection and fitting for CI.GR2 site. (a) Output when the peak detection algorithm is selected, (b) Output when manual fitting is selected.

3 HVSR Data Interpretation

3.1 INTRODUCTION

Our data interpretation has emphasized two issues. The first is peak identification. We engaged multiple analysts to inspect HVSR spectra from various data sources (mHVSR from permanent and temporary arrays and different instrument types, as well as eHVSR) and identify sites with peaks. Section 3.2 describes the sites and instruments selected for this assessment, variations in outcomes between analysts, and recommended peak selection criteria based on these findings (these criteria modify those of SESAME 2004).

The second issue addressed here is comparisons of HVSR attributes between data sources. Section 3.3 describes the sites selected for these comparisons, the data sources and their availability in the database, and between-method HVSR comparisons of peak occurrence and (for sites with peaks) peak attributes (frequency). Section 3.4 presents the HVSR statistical properties particularly for the amplitude distributions for mHVSR and eHVSR. The chapter is concluded in Section 3.5 with discussion of the significance of the findings.

3.2 PEAK IDENTIFICATION

HVSR plots can generally be classified as containing no peaks, clear peaks, or ambiguous intermediate cases. A clear peak generally indicates the site has strong impedance contrast(s) near one or more modal frequencies (e.g., Tuan et al., 2011). When there is no peak present in an HVSR, this suggests the site is either underlain with a sediment-filled depth profile that lacks a significant impedance contrast or it is a rock site with nearly depth-invariant near-surface velocities. It is necessary to identify this peak feature in HVSR correctly and explore its consistency among various data sources for the further model development.

3.2.1 Site Selection

We selected sites for the present work with the objective of developing pairs of essentially co-located HVSR from different instrument types or different ground vibration sources. These data

are used to evaluate and develop peak identification methods (this section) and to evaluate differences in HVSR attributes between instruments and/or vibration sources (Section 3.3). Three pairs of HVSR data sources were compared:

(1) mHVSR-P from broadband seismometers versus mHVSR-P from strong motion accelerometers with 24-bit recorders: 122 sites (Section 3.3.1)
(2) mHVSR-P from broadband seismometers versus mHVSR-T from temporary velocity transducers: 98 sites (Section 3.3.2)
(3) mHVSR from permanent or temporary seismometers versus eHVSR from strong earthquake records: 297 sites (Section 3.3.3)

Many sites are shared between these comparison groups. Combining all these sites, there are 345 unique sites as shown on the map in Figure 3.1.

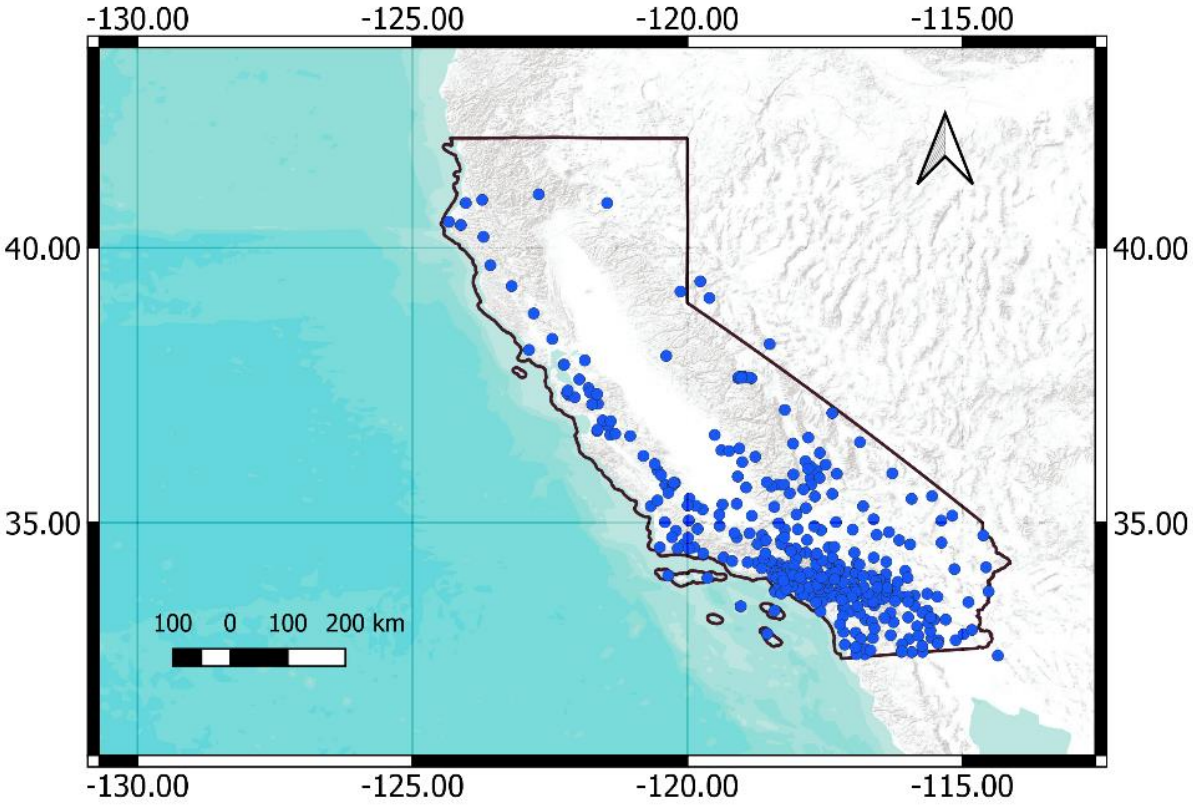


Figure 3.1. Locations of the selected sites for HVSR comparison.

Figure 3.2 shows a histogram of V_{S30} values for the 345 selected sites, which are based on measurements where available and a proxy-based relationship conditioned on surface geology and ground slope otherwise (Wills et al. 2015). The V_{S30} range is 150 m/s to 1200 m/s with a median of 443 m/s. Table 3.1 lists the surface geological units at the sites based on the California statewide geology map of Wills et al. (2015). The sites sample 12 geological units including alluvium of various ages and surface gradients (Qal1, Qal2, Qal3, Qoa, and QT) as well as Tertiary and

crystalline bedrock units (Kjf, sp, Tsh, Tss, and Tv). About 50% of the sites are on rock units and 50% are on sediments.

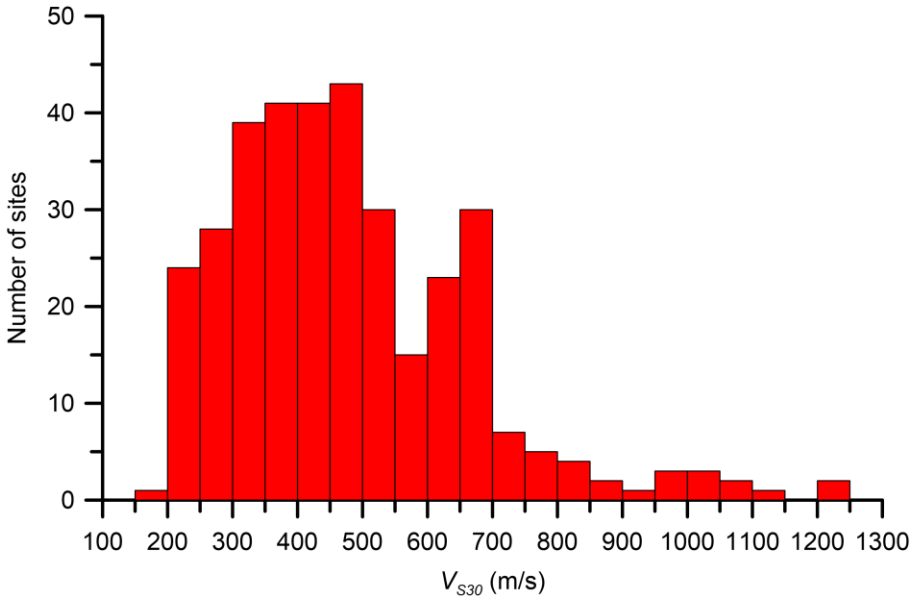


Figure 3.2. Histogram of V_{S30} for the selected sites.

Table 3.1. The statistics of the selected sites aggregated by surface geological units.

Geological unit	Number of sites
Qal1	23
Qal2	36
Qal3	47
Qoa	52
QT	13
crystalline	89
Kjf	14
Kss	4
sp	4
Tsh	15
Tss	23
Tv	21

Note: Four of the selected sites are located outside the geologic map and are not reflected in the table.

3.2.2 Peak Identification from Analysts

The principal use of HVSR for site response applications is to identify site resonances based on peaks in spectral ratio plots (Field and Jacob, 1993, 1995; Theodulidis et al., 1996; Bonilla et al., 1997, 2002; Satoh et al., 2001; Cadet et al., 2012). As such, the process by which an HVSR plot is judged to have or not have a peak is fundamental to its application.

We originally attempted to apply an algorithm to identify peaks from SESAME (2004), details of which are reviewed in Section 3.2.3. This was undertaken using a subset of the sites described in Section 3.2.1, consisting of 140 stations (Wang, 2020). The result was that very few (~20) sites were identified as having peaks. We did not consider this result to be reliable, in part because visual inspection of the data suggested that there were far more sites with peaks.

Since established protocols (SESAME 2004) proved ineffective, our approach was to engage multiple analysts to study the HVSR spectra and develop their own assessments of which sites had peaks or no peaks. Their findings are used to (1) evaluate the principal sources of uncertainty that affect peak identification and (2) provide a set of results against which to train an improved peak identification algorithm (Section 3.2.3).

This section describes the criteria by which the different analysts assessed the presence of peaks, shows examples, and provides summary statistics on peak identification. Results are presented for HVSR using four datasets: mHVSR-T, mHVSR-P from broadband seismometers and accelerometers, and eHVSR. The processing of data from these sources is described in Section 2.3. HVSR results and peak assessment results for each site are provided in Appendixes A-C and summarized in Table S1 in Appendix D. Appendixes A-C are organized for three sets of sites as described in Section 3.2.1.

Analysts' Qualitative Criteria

Four analysts reviewed the plots in Appendixes A-C, which are the first, second, third, and fourth authors of this report. Each were asked to visually assess the presence of peaks. No strict criteria for peak identification were mandated, but analysts tended to consider one or more of four different criteria:

- The amplitude of the peak should be high, either in an absolute sense or relative to amplitudes at neighboring frequencies without peaks.
- Peaks should not be too wide (span too large of a frequency range).
- Uncertainty in the peak amplitude should not be too large.
- The peak should not occur at the low- or high-frequency limits of the spectral ratio plot.

Each of these factors is also reflected in SESAME (2004) guidelines. Table 3.2 summarizes the degree to which each criterion was considered in the assignment of peaks. The entries in Table 3.2. were completed by the analysts themselves (with some editing for consistency).

Table 3.2. Qualitative criteria considered by Analysts 1-4

Criteria	Analyst 1 (PW)	Analyst 2 (JPS)	Analyst 3 (PZ)	Analyst 4 (SKA)
Peak amplitude value	>= ~ 1.5	--	>= ~ 1.5	--
Peak amplitude relative to amplitudes at neighboring frequencies	>= ~ 1.5	>= ~ 2	Note 3	Note 3
Peak width	< one log cycle of freq.	--	--	< one log cycle of freq.
Peak amplitude and/or frequency uncertainty	--	Note 1	--	--
Peak location relative to plot frequency limits	Note 2	Note 2	--	Note 2

-- Considered, but no specific criteria applied

[1] To be considered a peak, there should be a local high in the mean – one standard deviation spectral ratio amplitude at the peak frequency (although specific amplitude criteria are not enforced)

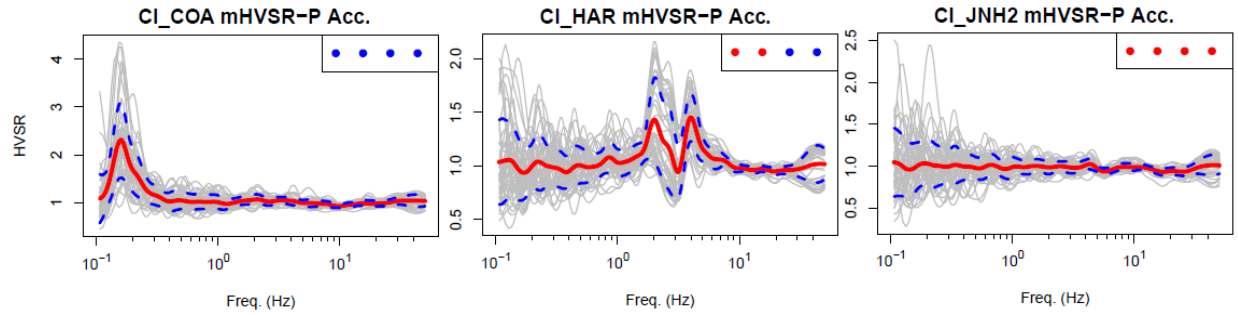
[2] Peaks that occur at or near the frequency limits of the plot are neglected

[3] To be considered a peak, it should stand out on both sides from amplitudes at neighboring frequencies (>= ~ 1.25) and there should be clear ascending and descending branches

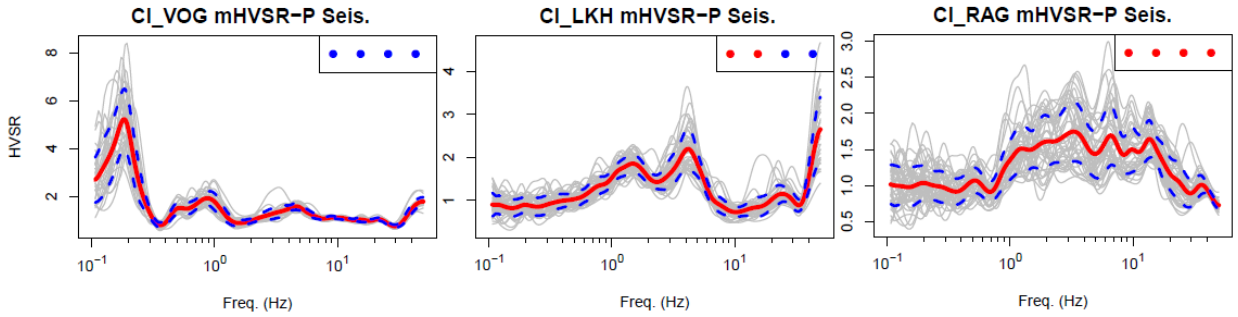
From the table, we see that Analyst 1 and 2 consider similar criteria, but Analyst 1 is somewhat more "liberal" in the assignment of peaks, particularly in regard to relative amplitude thresholds. Accordingly, we would expect Analyst 1 to assign more peaks than Analyst 2. Analysts 3 and 4 apply more liberal criteria, particularly in regard to relative amplitudes (e.g., factor of 1.25 versus 2.0 by Analyst 2). Analysts 3 and 4 tend to identify peaks if there are clear ascending and descending branches on either side. Thus, we expect more peaks to be identified by Analyst 3 and 4 than by Analysts 1 and 2.

Example Results

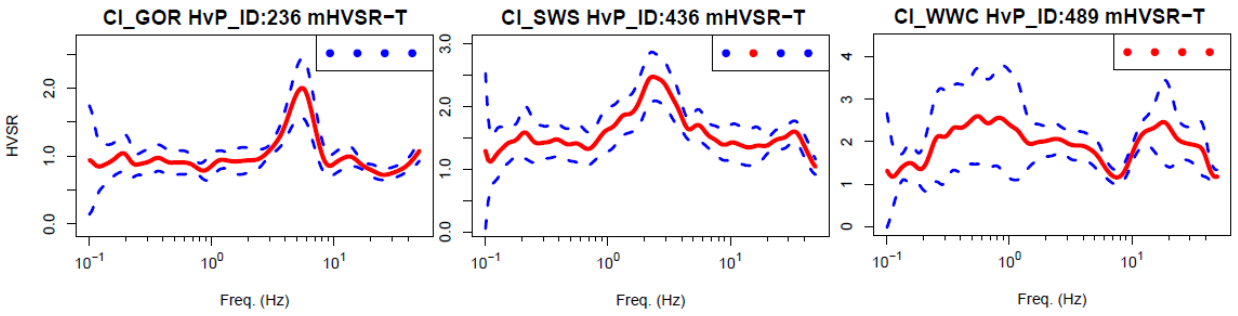
Figure 3.3. presents example results. The figure has four rows and three columns, with each "cell" being a different site. The four rows correspond to different instruments or different vibration sources as follows: (a) mHVSR from permanent strong motion accelerometers (denoted as mHVSR-P Acc.); (b) mHVSR from permanent BB seismometers (denoted as mHVSR-P Seis.); (c) mHVSR from temporary seismometers (denoted as mHVSR-T); and (d) eHVSR from seismometers or accelerometers (minimum of 10 earthquakes) (also denoted as eHVSR). The three columns are differentiated by the peak identification results as follows: (left) all analysts assign a peak; (center) mixed peak assignments; (right) all analysts assign no peak. The assignment or non-assignment of a peak to a given site is marked in the figures by blue dots (indicating a peak was assigned) and red dots (indicating a peak was not assigned). One dot is shown for each of the four analysts, with the exception of eHVSR where only three analysts made peak assignments.



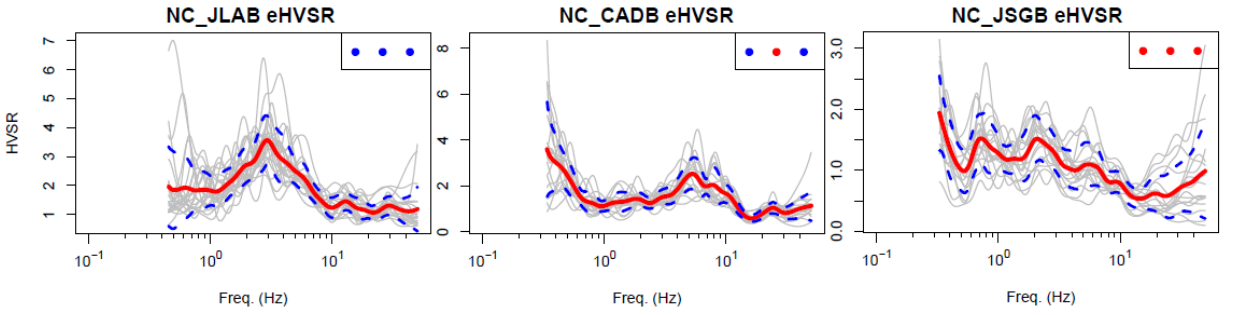
(a) mHVSR-P from accelerometers



(b) mHVSR-P from broadband seismometers



(c) mHVSR-T



(d) eHVSR

Figure 3.3. Example results for HVSR measurements showing clear peaks (left), ambiguous (center), and no peaks (right). The assignment or non-assignment of a peak to a given site is marked in the figures by blue dots (indicating a peak was assigned) and red dots (indicating a peak was not assigned). One dot is shown for each analysts (4 for mHVSR, 3 for eHVSR). The four types of HVSR measurements are (a) mHVSR-P Acc.; (b) mHVSR-P Seis.; (c) mHVSR-T; and (d) eHVSR (at least 10 events/window).

In Figure 3.3, for each cell, a series of distinct HVSR results are shown as gray lines – these correspond to different windows for mHVSR and different earthquakes for eHVSR (details in Section 2.3). In the case of mHVSR-T (Part c), windowed HVSRs are not plotted (the results were pulled from the database, where results for individual windows are not stored). In the case of eHVSR, the frequency range is relatively limited (beginning near 0.4 Hz) because results are not plotted below the median of high-pass corner frequencies from the earthquake ground motion processing. The red lines represent the mean amplitude and blue lines represent the mean ± one standard deviation, which are computed across windows for each frequency. These are computed as arithmetic means and standard deviations based on the data distribution (Section 3.4).

The sites in the left column each have "clear" peaks in which the amplitude is at least 2.0 and higher, by at least a factor of two, than peak-adjacent ordinates at neighboring frequencies (e.g., < 0.2 Hz and > 8 Hz for CI.GOR in Part c of the figure). Sites in the center column all have clearly visible local highs and lows in the HVSR spectrum, but these features are either small in absolute amplitude (Part a, CI.HAR) and/or relative amplitude (Part c, CI.SWS; Part d, NC.CADB). Depending on the amplitude criteria applied by different analysts, peaks may or may not have been assigned; in the cases shown in the center column of Figure 3.3, Analysts 3 and 4 assigned peaks, Analysts 2 did not, and Analyst 1 made mixed assignments. Sites in the right column are essentially flat in some cases (Part a, CI.JNH2) or have local highs and lows similar to the center column. However, those local peaks are either very low in amplitude (Part d, NC.JSGB) or overly wide to be considered as distinct peaks (Part b, CI.RAG; Part c, CI.WWC).

These examples illustrate differences in the assessments made by the four analysts. Modest differences in the qualitative criteria for judging peaks change the outcome of peak assessments for many sites, as well as other downstream outputs. The following sub-section examines systematically the consistency and differences of analyst outcomes across sites.

Inter-Analyst Comparisons

In this sub-section, we synthesize the HVSR peak identification assessments. Pooling the mHVSR results from permanent and temporary velocity seismometers (example results in Parts b and c of Figure 3.3), there are 312 site/data pairs that were assessed by the four analysts. Among this population, 35% have peaks assigned by all analysts, 43% are ambiguous (peaks assigned by some and not by others), and 22% have no peaks assigned by all analysts.

Table 3.3 breaks down inter-analyst consistencies and differences for each analyst pair (note that Analyst 4 did not assess eHVSR peaks). The table has four sub-tables, (a)-(d) corresponding to different instrument types and vibration sources (as in Figure 3.3). Within each sub-table, we compare assessments between analysts by showing proportions of sites in four categories:

(1) Peak – Peak (P-P) sites where both analysts identify clear peaks in the HVSR;

(2) Peak – No Peak (P-N) sites where the first analyst identifies a clear peak but the second does not;
(3) No Peak – Peak (N-P) sites where the second analyst identifies a clear peak but the first does not; and
(4) No Peak – No Peak (N-N) sites where both analysts do not identify clear peaks.

Inter-analyst consistency is represented in Table 3.3 by a large sum of the (P-P) and (N-N) cells. We refer to an analyst as "conservative" when they apply relatively strict peak assignment criteria (Analysts 1-2) and as "liberal" when they are more lenient (Analysts 3-4). In Table 3.3, the first analyst is more conservative than the second when (N-P) proportion is higher than (P-N) proportion, and vice versa.

The results in Table 3.3 show inter-analyst consistency rates that vary from 59% to 92%. Liberal analysts are consistent with each other (83-86%, mean 85%) as are conservative analysts (83-92%, mean 87%). Liberal-conservative agreement rates are somewhat lower (59-92%, mean 76%). These results demonstrate that there is a significant population of sites in California (about 40%) where peak identification is ambiguous, and that for those sites, it is not unlikely that different expert analysts will reach different conclusions.

Given the above, there would be practical value in having algorithms that can identify peaks. Such algorithms offer the potential for HVSR interpretation efficiency (less analyst time) and should be tunable to accommodate relatively liberal or conservative peak selection criteria. This topic is addressed in the next subsection.

Table 3.3. The variability of inter-analysts on peak identification from four data sources, (a) mHVSR-P from broadband seismometer, (b) mHVSR-P from strong motion accelerometer, (c) mHVSR-T, and (d) eHVSR.

(a)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Analyst 1 (PW)	Peak	-		36%	10%	46%	0%	46%	0%
	No Peak			5%	49%	35%	19%	20%	34%
Analyst 2 (JS)	Peak	-		-		40%	0%	40%	0%
	No Peak					41%	19%	25%	35%
Analyst 3 (PZ)	Peak	-		-		-		64%	16%
	No Peak							1%	19%
(b)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Analyst 1 (PW)	Peak	-		10%	3%	13%	0%	12%	1%
	No Peak			5%	82%	25%	62%	11%	76%
Analyst 2 (JS)	Peak	-		-		15%	0%	15%	0%
	No Peak					23%	62%	8%	77%
Analyst 3 (PZ)	Peak	-		-		-		23%	15%
	No Peak							0%	62%
(c)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Analyst 1 (PW)	Peak	-		49%	13%	62%	0%	39%	1%
	No Peak			4%	34%	23%	15%	18%	42%
Analyst 2 (JS)	Peak	-		-		53%	0%	37%	1%
	No Peak					32%	15%	20%	42%
Analyst 3 (PZ)	Peak	-		-		-		56%	13%
	No Peak							1%	30%
(d)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)			
		Peak	No Peak	Peak	No Peak	Peak	No Peak		
Analyst 1 (PW)	Peak	-		29%	8%	37%	0%		
	No Peak			4%	59%	33%	30%		
Analyst 2 (JS)	Peak	-		-		33%	1%		
	No Peak					37%	29%		

Note: Analyst 4's assessment on peak identification for eHVSR is not yet complete so their results are missing in sub-table (d).

3.2.3 Algorithmic Peak Identification

In this section, we first describe SESAME (2004) guidelines and a few examples of their application, which demonstrate that the guidelines are overly restrictive for peak identification, even relative to the most conservative criteria in Section 3.2.2. New criteria are suggested that can be implemented using the tools introduced in Section 2.5.

SESAME Guidelines

SESAME guidelines (SESAME, 2004) provide a procedure for the identification of peaks that first considers three criteria that assess the reliability of the HVSR curve and then considers six conditions intended to establish the presence of a clear HVSR peak. The first two criteria for the reliability of HVSR curves constrain the minimum required number of sub-windows and duration; these requirements are accounted for in the query and processing procedures described in Section 2.3. Hence, the additional procedures used to identify peaks are the third reliability criterion and the six conditions, which are listed in Table 3.4.

Table 3.4. Reliability criterion and conditions for peak identification from SESAME (2004)

Parameters	SESAME
Reliability 3: $f_{peak} > 0.5$ Hz, $f \in [0.5 f_{peak}, 2 f_{peak}]$	$\sigma_A(f) < 2$
Reliability 3: $f_{peak} < 0.5$ Hz, $f \in [0.5 f_{peak}, 2 f_{peak}]$	$\sigma_A(f) < 3$
Clear 1: $f \in [0.25 f_{peak}, f_{peak}]$	$A_{H/V}(f) < 0.5 A_{peak}$
Clear 2: $f \in [f_{peak}, 4 f_{peak}]$	$A_{H/V}(f) < 0.5 A_{peak}$
Clear 3:	$A_{peak} >= 2$
Clear 4: peak of SD curve f_{peak} $[A_{H/V}(f) - \sigma_A(f)]$	within $[f_{peak}/1.05, 1.05 f_{peak}]$
Clear 4: peak of SD curve f_{peak} $[A_{H/V}(f) + \sigma_A(f)]$	within $[f_{peak}/1.05, 1.05 f_{peak}]$
Clear 5: $f_{peak} < 0.2$ Hz	$\sigma_f < 0.25 f_{peak}$
Clear 5: $f_{peak} \in [0.2, 0.5)$ Hz	$\sigma_f < 0.2 f_{peak}$
Clear 5: $f_{peak} \in [0.5, 1.0)$ Hz	$\sigma_f < 0.15 f_{peak}$
Clear 5: $f_{peak} \in [1.0, 2.0)$ Hz	$\sigma_f < 0.1 f_{peak}$
Clear 5: $f_{peak} > 2.0$ Hz	$\sigma_f < 0.05 f_{peak}$
Clear 6: $f_{peak} < 0.2$ Hz	$\sigma_A(f_{peak}) < 3$
Clear 6: $f_{peak} \in [0.2, 0.5)$ Hz	$\sigma_A(f_{peak}) < 2.5$
Clear 6: $f_{peak} \in [0.5, 1.0)$ Hz	$\sigma_A(f_{peak}) < 2$
Clear 6: $f_{peak} \in [1.0, 2.0)$ Hz	$\sigma_A(f_{peak}) < 1.78$
Clear 6: $f_{peak} > 2.0$ Hz	$\sigma_A(f_{peak}) < 1.58$

In Table 3.4, f_{peak} is the peak frequency of interest (there could be multiple f_{peak} values in a single HVSR plot); f is the independent frequency; $A_{H/V}(f)$ is the amplitude of the HVSR mean at frequency f; A_{peak} is the amplitude at f_{peak}; $\sigma_A(f)$ is the standard deviation of $A_{H/V}(f)$ at f; $\sigma_A(f_{peak})$ is the standard deviation of $A_{H/V}(f)$ at f_{peak}; and σ_f is the standard deviation of f_{peak}. In Table 3.4, the rows labelled Reliability 3, Clear 5, and Clear 6 are f_{peak}-dependent. The greater is f_{peak}, the more stringent are the standards for establishing a peak as reliable and clear.

Figure 3.4 shows two example mHVSR plots for which all four analysts assigned peaks: for site CI.BBR, the peak was assigned at 0.7-0.8 Hz; for site CI.BOR, the peak was assigned at 12-15 Hz. However, for both of these sites, multiple conditions specified by SESAME are not satisfied, and as a result these sites would not be assigned as having peaks. Table 3.5 presents our assessments of reliability criterion 3 and the six conditions for the two sites – the table provides the calculated statistics and evaluates those statistics relative to SESAME guidelines. Both sites fail two conditions. Wang (2020) previously documented similar issues for 28 out of 53 clear peak sites considered in his research (the total number of sites in that study with mHVSR was 140).

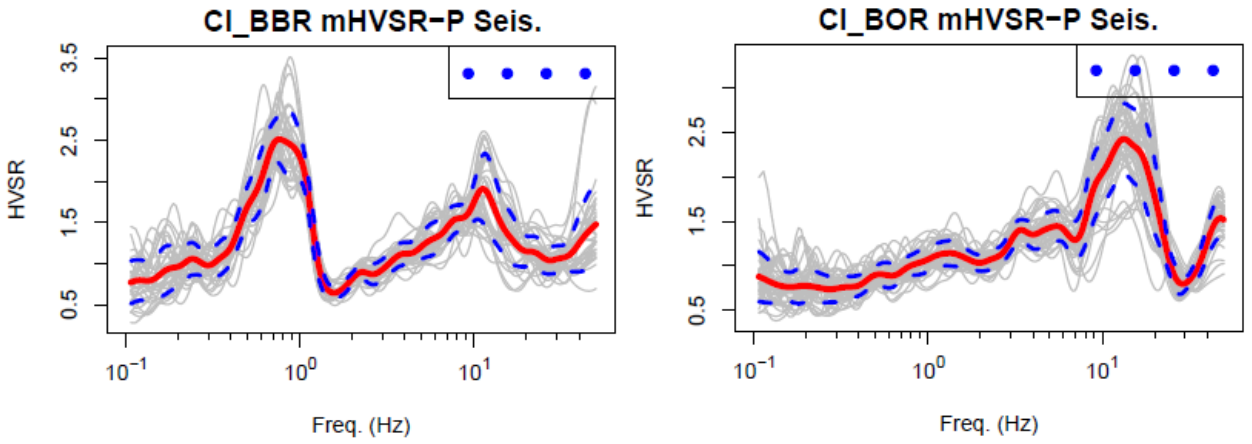


Figure 3.4. Example sites that are identified with peak presence by analysts but not by SESAME guidelines

For CI.BBR, the conditions of Clear 4 and Clear 5 are not satisfied. This is because the HVSR shows larger uncertainty than SESAME's requirements. For CI.BOR, the conditions of Clear 1 and Clear 5 are not satisfied. Clear 1 fails because the peak amplitude is not significantly (factor of 2) larger than the ordinates at neighboring frequencies between 0.25 f_{peak} and f_{peak}.

In the next two subsections, we describe alternative procedures for peak identification. The first maintains the SESAME framework and modifies thresholds for some condition, whereas the second is formulated differently to better capture the analyst criteria (Table 3.2).

Table 3.5. Peak identification reliability criteria and conditions from SESAME (2004)

Site	Parameters	Calculated Statistics	Decision
CI_BBR	f_{peak}	0.75 Hz	
	Reliability 3: $\sigma_A(f)$	0.38	< 2, pass
	Clear 3: A_{peak}	2.52	> 2, pass
	Clear 1: min($A_{H/V}(f)$)	0.97	0.97 < 0.5 A_{peak} =1.26, pass
	Clear 2: min($A_{H/V}(f)$)	0.65	0.65 < 0.5 A_{peak} =1.26, pass
	Clear 4-1: peak frequency of mean - SD curve, f_{peak} [$A_{H/V}(f)$ - $\sigma_A(f)$]	0.73	0.73 ∈ [f_{peak} /1.05 = 0.71, f_{peak}*1.05 = 0.79], pass
	Clear 4-2: peak frequency of mean + SD curve f_{peak} [$A_{H/V}(f)$ + $\sigma_A(f)$]	0.82	0.82 ∉ [f_{peak} /1.05 = 0.71, f_{peak}*1.05 = 0.79], ***fail***
	Clear 5: σ_f	1.34	1.34 ≮ 0.15* f_{peak}=0.11 Hz, ***fail***
	Clear 6: $\sigma_A(f_{peak})$	0.29	0.29 < 2, pass
CI_BOR	f_{peak}	13.29	
	Reliability 3: $\sigma_A(f)$	0.46	< 2, pass
	Clear 3: A_{peak}	2.44	> 2, pass
	Clear 1: min($A_{H/V}(f)$)	1.29	1.29 ≮ 0.5 A_{peak} =1.22, ***fail***
	Clear 2: min($A_{H/V}(f)$)	0.79	0.79 < 0.5 A_{peak} =1.22, pass
	Clear 4-1: peak frequency of mean - SD curve, f_{peak} [$A_{H/V}(f)$ - $\sigma_A(f)$]	13.33	13.33 ∈ [f_{peak} /1.05 = 12.66, f_{peak}*1.05 = 13.95], pass
	Clear 4-2: peak frequency of mean + SD curve f_{peak} [$A_{H/V}(f)$ + $\sigma_A(f)$]	13.24	13.24 ∈ [f_{peak} /1.05 = 12.66, f_{peak}*1.05 = 13.95], pass
	Clear 5: σ_f	1	1 ≮ 0.05* f_{peak}=0.7 Hz, ***fail***
	Clear 6: $\sigma_A(f_{peak})$	0.39	0.39 < 1.58, pass

Adjusted SESAME Guidelines

In a similar manner to what is shown in Table 3.5, Wang (2020) studied 28 sites that have peaks based on visual inspection but are not identified as having peaks using SESAME criteria. The principal SESAME condition that screens such sites from clear peak designations is Criterion 5 (denoted "Clear 5" in Tables 3.5-3.6). Accordingly, one way to improve peak identification criteria is to remove Criterion 5, as shown in Table 3.6. Some modest adjustments to Criteria 1-4, in each case relaxing the criteria, are also suggested in Table 3.6. Application of the adjusted SESAME conditions to the two example sites in Figure 3.4 identifies clear peaks for both. Moreover, their application for the 140 sites considered by Wang (2020), 53 of which have clear peaks based on visual inspection (from Analysts 1-2), identifies 41 (was 20 by original SESAME) sites as having clear peaks and identifies 1 (was 1 by original SESAME) sites with peaks when visual inspection suggests no peaks (false positives).

Table 3.6 Adjusted peak identification conditions (modifications are relative to SESAME, 2004)

Parameters	Current SESAME	Adjusted SESAME
Reliability 3: $f_{peak} > 0.5$ Hz, $f \in [0.5 f_{peak}, 2 f_{peak}]$	$\sigma_A(f) < 2$	$\sigma_A(f) < 2$
Reliability 3: $f_{peak} < 0.5$ Hz, $f \in [0.5 f_{peak}, 2 f_{peak}]$	$\sigma_A(f) < 3$	$\sigma_A(f) < 3$
Clear 1: $f \in [0.25 f_{peak}, f_{peak}]$	$A_{H/V}(f) < 0.5 A_{peak}$	$A_{H/V}(f) < 0.6 A_{peak}$
Clear 2: $f \in [f_{peak}, 4 f_{peak}]$	$A_{H/V}(f) < 0.5 A_{peak}$	$A_{H/V}(f) < 0.6 A_{peak}$
Clear 3:	$A_{peak} >= 2$	$A_{peak} >= 1.6$
Clear 4: peak of SD curve f_{peak} $[A_{H/V}(f) - \sigma_A(f)]$	within $[f_{peak}/1.05, 1.05 f_{peak}]$	within $[f_{peak}/1.15, 1.15 f_{peak}]$
Clear 4: peak of SD curve f_{peak} $[A_{H/V}(f) + \sigma_A(f)]$	within $[f_{peak}/1.05, 1.05 f_{peak}]$	within $[f_{peak}/1.12, 1.12 f_{peak}]$
Clear 5: $f_{peak} < 0.2$ Hz	$\sigma_f < 0.25 f_{peak}$	-
Clear 5: $f_{peak} \in [0.2, 0.5)$ Hz	$\sigma_f < 0.2 f_{peak}$	-
Clear 5: $f_{peak} \in [0.5, 1.0)$ Hz	$\sigma_f < 0.15 f_{peak}$	-
Clear 5: $f_{peak} \in [1.0, 2.0)$ Hz	$\sigma_f < 0.1 f_{peak}$	-
Clear 5: $f_{peak} > 2.0$ Hz	$\sigma_f < 0.05 f_{peak}$	-
Clear 6: $f_{peak} < 0.2$ Hz	$\sigma_A(f_{peak}) < 3$	$\sigma_A(f_{peak}) < 3$
Clear 6: $f_{peak} \in [0.2, 0.5)$ Hz	$\sigma_A(f_{peak}) < 2.5$	$\sigma_A(f_{peak}) < 2.5$
Clear 6: $f_{peak} \in [0.5, 1.0)$ Hz	$\sigma_A(f_{peak}) < 2$	$\sigma_A(f_{peak}) < 2$
Clear 6: $f_{peak} \in [1.0, 2.0)$ Hz	$\sigma_A(f_{peak}) < 1.78$	$\sigma_A(f_{peak}) < 1.78$
Clear 6: $f_{peak} > 2.0$ Hz	$\sigma_A(f_{peak}) < 1.58$	$\sigma_A(f_{peak}) < 1.58$

While these adjustments to SESAME conditions are promising, shortcomings remain, as follows:

1. The original and adjusted conditions are challenging to automate, because the analyst typically needs to visually identify the most reliable f_{peak}. The reason for this manual identification is because the most appropriate f_{peak} may not always correspond with the highest point on the HVSR.
2. SESAME conditions 1 and 2 (labelled in Tables 3.5-3.6 as "Clear 1" and "Clear 2") define relative ordinate amplitudes (i.e., amplitudes relative to the amplitude at f_{peak}, which is A_{peak}) as the average amplitude from f_{peak} to frequencies higher or lower than f_{peak}. As a result, ordinates on the peak itself are used in the normalization. We prefer the use of portions of the HVSR spectrum that are "away" from the peak, either plateaus or low points before the start of adjacent peaks.
3. A range of frequencies is required for Gaussian peak fitting, using procedures given in Section 2.5.2. The SESAME guidelines (original and adjusted) do not provide such a

range. While a range can certainly be selected manually in the SESAME framework, manual selection is incompatible with an automated framework.

These shortcomings motivated us to propose new peak identification criteria amenable to algorithmic implementation, as described next.

Proposed Peak Detection Procedure

We propose here a peak detection procedure that is inspired by certain aspects of the SESAME guidelines, but not bound by the historical framework, especially with regards to conditions in that framework that have been found to be relatively ineffective when applied to the California HVSR data. Our aim is a procedure that can be implemented in an automated manner (i.e., a coded algorithm), although human inspection of results is still recommended.

The procedure was developed to capture the features that were considered by the four analysts and summarized in Table 3.2. One of the most critical features in Table 3.2 is whether a peak amplitude is sufficiently high. For this purpose, we prefer the relative amplitude criteria (second row in Table 3.2), but also consider absolute amplitude criteria (first row in Table 3.2).

The principal challenge in implementing the relative amplitude criteria is defining the amplitudes of those portions of the HVSR spectrum adjacent to, but off of, the peak in question. Such features may be plateaus or local low points before the start of another peak. To define those amplitudes, we implement a regression tree (Breiman et al. 1984), which is a predictive modelling approach in machine learning. Figure 3.5 illustrates the regression tree approach as applied to HVSR. The spectrum is approximated by a series of non-overlapping horizontal lines (i.e., a step function). As shown in Figure 3.5, the step length is influenced by the *cp* parameter (i.e., complexity parameter); larger values of *cp* effectively increase the penalty in tree regression for complex models (i.e., models with many steps). Accordingly, large *cp* values (e.g., 0.1 in left frame) produce relatively wide steps and a crude fit, whereas smaller *cp* produces narrower steps and a tighter fit. If *cp* is too large, the fit is relatively poor, whereas if *cp* is too small, the step functions capture too many small peaks, which is not amenable to defining a stable peak-adjacent plateau amplitude.

Selection of the preferred value of *cp* is subjective. While in principle it could be established using an optimization algorithm, it is unclear how to formulate an objective function that accounts for both the accuracy and reliability of the fit. Therefore, we use trial and error with visual inspection, and on this basis find that *cp* = 0.005 (central frame in Figure 3.5) strikes a reasonable balance between accuracy and reliability in the R-script implementation (i.e., with *TreeReg.R*). The tree regression routines in R and Python are not identical, and slightly smaller *cp* parameters are required in Python – in that case *cp* = 0.002-0.003 produces similar results to *cp* = 0.005 in R. The *cp* parameter may take on different values if the regression tree is run on a linear frequency scale; we have used it with a natural log frequency scale.

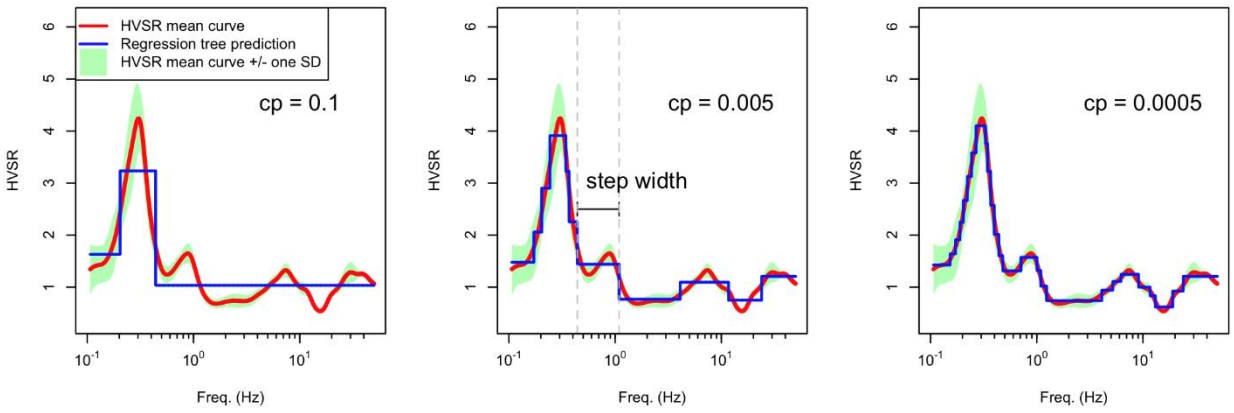


Figure 3.5. Schematic illustration of regression tree fitting of HVSR data with different values of the complexity parameter, *cp*. For this figure, tree regression was performed with TreeReg.R.

Critical step attributes are the step_width (log difference of maximum and minimum frequency in the step) and step amplitude. Step_width is illustrated in Figure 3.5. Because our primary interest in the application of tree regression is defining peak-adjacent steps, it is necessary to screen out very short step widths as might be found on the two sides of the peak (e.g., two such steps occur on the left side of the peak in the middle frame of Figure 3.5). Examples of peak-adjacent steps are those with $f < 0.2$ Hz and $f \approx 0.42$-1.0 Hz in the middle frame of Figure 3.5. To facilitate step screening, we define step_jump as a minimum width that is required for use in defining a peak-adjacent amplitude. A step is screened if step_width < step_jump.

Figure 3.6 shows a flow chart that implements a step screening process, ultimately producing a left_ratio (left peak-adjacent step amplitude ÷ peak amplitude) and right_ratio (right peak-adjacent step amplitude ÷ peak amplitude). The procedure in the flow chart follow the steps below.

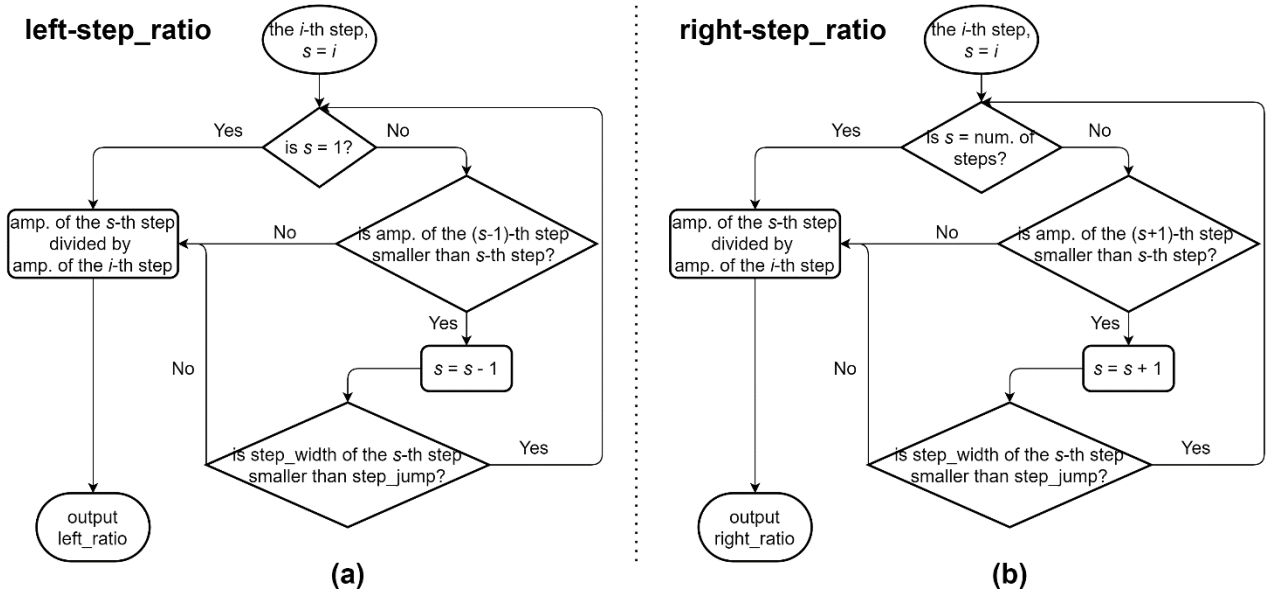


Figure 3.6. Flowcharts illustrating step screening process that leads to calculation of left_ratio and right_ratio. In the flowchart, i is the step index and s is a secondary index used to check the amplitudes of steps immediately to the left or right of i.

Task 1 – Compute Left-Adjacent Step Ratios (R_ℓ): This operation proceeds as follows:

- Let i denote the step.
- Starting at $i = 1$, $R_{\ell,1}$ is the step amplitude divided by itself, hence it is 1.0.
- For $i = 2$ and subsequent steps up to the number of steps (N_{step}), $R_{\ell,i} = 1.0$ is assigned if the amplitude of the preceding step $(i - 1)$ is larger than that at step i. Otherwise, find the first descending step to the left of step i for which step_width > step_jump. The index of this step is denoted i_ℓ.
- The ratio of the amplitude of step i_ℓ to the amplitude of step i is $R_{\ell,i}$.

Task 2 – Compute Right-Adjacent Step Ratios (R_r): This operation proceeds as follows;

- The procedure is similar to Task 1.
- In this case for the last step ($i = N_{step}$) $R_r = 1.0$.
- For $i < N_{step}$, $R_{r,i} = 1.0$ is assigned if the amplitude of the next step $(i + 1)$ is larger than that at step i. Otherwise, find the first descending step to the right of step i for which step_width > step_jump. The index of this step is denoted i_r.
- The ratio of the amplitude of step i_r to the amplitude of step i is $R_{r,i}$.

Task 3 – Identify potential peak steps: Step i is a peak if its amplitude is larger than those of steps $i - 1$ and $i + 1$. This is equivalent to requiring $R_{\ell,i} < 1$ and $R_{r,i} < 1$.

Task 4 – Identify clear peak steps from among potential peaks: The criteria considered by analysts and identified in Table 3.2 are checked to identify clear peaks from among the steps with peaks from Task 3. These checks include:

a) The amplitude of Step i ($A_{step,i}$) should exceed an amplitude threshold, amp_thres (applies the criterion in first row of Table 3.2);

b) The max of $R_{\ell,i}$ and $R_{r,i}$ should be less than the ratio threshold, ratio_thres (applies the criterion in second row of Table 3.2). Smaller values of ratio_thres render this check more conservative;

c) The peak should not be too wide (criterion in third row of Table 3.2), which is checked by taking the natural log difference of the frequency at the left end of step i_r and at the right end of step i_ℓ, which should be smaller than $ln10 \approx 2.3$;

d) The uncertainty of HVSR amplitudes within Step i should not be too large (4th row of Table 3.2). We compute the mean and standard deviation of amplitudes within the step, $\bar{A}_i$ and $\sigma_{A,i}$. We then perform two checks:

$$\bar{A}_i - k \times \sigma_{A,i} \geq A_{i_\ell} \tag{3.1}$$

$$\bar{A}_i - k \times \sigma_{A,i} \geq A_{i_r} \tag{3.2}$$

where k is the number of standard deviations below the mean that is considered for checking whether the peak in Step i is larger than peak-adjacent step amplitudes (A_{i_ℓ} and A_{i_r}). Larger values of k render this check more conservative.

e) An approximate estimate of the peak frequency should be within useful frequency range. For the purpose of this check, this estimate (f_{peak}) is approximated as the geometric mean of the lowest and highest frequency within Step i. We define the minimum frequency, min_freq and maximum frequency, max_freq, as the limits in the HVSR plot. We then require: $f_{peak} > 1.2 \times$ (min_freq) and $f_{peak} <$ (max_freq) $\div$ 1.2.

Task 5 – Select among multiple clear peaks: If more than one clear peak is identified in Task 4, the one with the lowest value of f_{peak} is returned.

Task 6 – Fitting of clear peak and width check: If a clear peak has been identified, Eq. (2.1) is fit to the data as described in Section 2.5.2. The range of frequencies used in the fitting extend from Step $i_{\ell,}$ to Step $i_{r,}$ (not the entire HVSR range).

Tasks 4b and 4d can be applied with different levels of conservatism, depending on selected values of the ratio_thres and k. Table 3.7 provides values of these and other parameters that provide peak identification results broadly similar to those from liberal and conservative analysts for mHVSR (Section 3.2.2). For eHVSR, amplitude uncertainty near the peak ($\sigma_{A,i}$) tends to be larger than for mHVSR. To accommodate this, the uncertainty threshold may be relaxed, as shown in Table 3.7.

Table 3.7. Recommended thresholds for liberal and conservative implementations of the automated algorithm

Cons.	cp	step_jump (natural log)	amp_thres	ratio_thres	k	min_freq (Hz)	max_freq (Hz)
mHVSR	0.005	0.35	1.5	0.7	1	0.1	15
eHVSR	0.005	0.35	1.5	0.7	0.5	0.1	15
Lib.	cp	step_jump (natural log)	amp_thres	ratio_thres	k	min_freq (Hz)	max_freq (Hz)
mHVSR	0.005	0.45	1.15	0.95	0.8	0.1	15
eHVSR	0.005	0.45	1.15	0.95	0.5	0.1	15

In Table 3.8 we apply the conservative thresholds from Table 3.7 and compare the results to those from analysts for the same four instrument types and vibration sources considered in Section 3.2.2. Table 3.9 repeats this exercise for the liberal thresholds.

Table 3.8. Comparison between peak identification results from proposed automated procedure implemented with conservative thresholds and analysts' findings on peak identification for four data sources, (a) mHVSR-P from accelerometer, (b) mHVSR-P from broadband seismometer, (c) mHVSR-T, and (d) eHVSR.

(a)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Algorithm	Peak	11%	4%	11%	4%	15%	0%	14%	1%
	No Peak	2%	83%	4%	81%	23%	62%	9%	76%
(b)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Algorithm	Peak	29%	4%	27%	6%	33%	0%	32%	1%
	No Peak	17%	50%	14%	53%	48%	19%	34%	33%
(c)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Algorithm	Peak	33%	5%	31%	7%	37%	1%	35%	3%
	No Peak	23%	39%	14%	48%	42%	20%	22%	40%
(d)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)			
		Peak	No Peak	Peak	No Peak	Peak	No Peak		
Algorithm	Peak	24%	6%	19%	11%	29%	1%		
	No Peak	13%	57%	14%	56%	41%	29%		

Table 3.9. Comparison between peak identification results from proposed automated procedure implemented with liberal threshold and analysts' findings on peak identification for four data sources, (a) mHVSR-P from accelerometer, (b) mHVSR-P from broadband seismometer, (c) mHVSR-T, and (d) eHVSR.

(a)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Algorithm	Peak	12%	19%	12%	19%	27%	4%	19%	12%
	No Peak	1%	68%	2%	67%	11%	58%	4%	65%
(b)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Algorithm	Peak	42%	32%	36%	38%	65%	9%	57%	20%
	No Peak	4%	22%	5%	21%	16%	10%	9%	14%
(c)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
Algorithm	Peak	57%	17%	48%	26%	69%	5%	54%	16%
	No Peak	4%	22%	5%	21%	16%	10%	3%	27%
(d)		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)			
		Peak	No Peak	Peak	No Peak	Peak	No Peak		
Algorithm	Peak	34%	40%	30%	45%	60%	15%		
	No Peak	3%	23%	4%	21%	11%	14%		

The results in Table 3.8 show that the conservative thresholds implemented in the automated procedure produce consistency rates (i.e., sum of the (P-P) and (N-N) percentages) with Analysts 1 and 2 that range from 72% to 94% (mean of 82%), whereas for the more liberal analysts the range is 52% to 90% (mean 68%). The results in Table 3.9 show that the liberal thresholds produce consistency rates with Analysts 3 and 4 of 71% to 85% (mean 78%), whereas for more conservative analysts the range is 51% to 80% (mean 67%). This shows that the automated procedure can effectively produce analyst results for either threshold type.

The algorithm described here for peak identification and fitting has been implemented in R (*TreeReg.R*) and in a Jupyter Notebook, as described in Section 2.5.2. Recommended *cp* values are 0.005 and 0.002, respectively. These implementations use the conservative thresholds.

Application to Full California Data Set

An advantage of the algorithm is that it can be efficiently applied to all sites in California with mHVSR data, as contained in the database described in Chapter 2. When the procedure is applied to all sites, we find that 27% have peaks. Among sites with surface geology (from Wills et al. 2015) consisting of Mesozoic bedrock, Tertiary bedrock, and Quaternary sediments, the rates with which peaks are encountered are 24%, 27%, and 28%, respectively. As a result, it appears that surface geology is not significantly predictive of peak occurrence. Table S2 in Appendix E indicates for each site whether a peak was identified, and for those sites with peaks, the fitting parameters derived using Eq. (2.1).

3.3 COMPARISONS OF HVSR ATTRIBUTES BETWEEN DATA SOURCES

In the HVSR database, we have processed and stored mHVSR data from temporary arrays (mHVSR-T) and queried data from permanently installed instruments with continuous data streams (mHVSR-P Seis.). In both cases, the data are taken from seismometers. In this section, these mHVSR data are compared with each other, and are also compared to mHVSR from accelerometers (mHVSR-P Acc.) and eHVSR. Such comparisons are critical for establishing the reliability and consistency of HVSR, which is of obvious importance as it is considered for eventual applications in practice. We have investigated these questions using the dataset described in Section 3.2, which consists of 122 sites for the comparison between mHVSR-P Acc. And Seis., 98 sites for the comparison between mHVSR-T and mHVSR-P Seis., and 297 sites for the comparison of mHVSR (combination of mHVSR-T and mHVSR-P Seis.) and eHVSR. The comparisons are made in terms of the presence of peaks and fitted peak frequencies and amplitudes.

3.3.1 Comparison of mHVSR from Strong Motion Accelerometers and Broadband Seismometers

In order for HVSR-based parameters to become widely adopted for use in ground motion prediction, it will be necessary for HVSR data to be available for a large majority of ground motion stations. As explained in Section 1.1, it is desirable for such measurements to be based on data independent of seismic signals (hence mHVSR is preferred).

Many ground motion recording sites (about 550 in CA) have co-located BB seismometers and accelerometers, which is convenient because the seismometers typically have sufficient sensitivity to reliably record ambient vibrations. Signals from these sites provide much of the mHVSR content currently in the database (Chapter 2). However, many other recording sites (about 1100) have only accelerometers, and such sites have produced many of the medium- to large-magnitude recordings in the California portion of the NGA-West2 database (Ancheta et al. 2014). Accordingly, if microtremor data from accelerometers can reliably estimate mHVSR under certain circumstances (e.g., sensors with 24-bit data acquisition units), it would be very useful for the broader effort.

Accordingly, we investigate here the consistency of mHVSR from strong motion accelerometers and BB seismometers. This is done for 122 sites with co-located sensors in which the data acquisition is 24-bit (lower resolution recorders were not considered). Robert Nigbor (personal written communication, 2020) helped us identify the sites with 24-bit recorders for accelerometers. Visual peak assessments were provided by four analysts working independently, and by the algorithm described in Section 3.2.3.

Table 3.10 shows the consistency of peak identification results between instrument types. The format of the table matches that in Section 3.2.2 (Table 3.3), but instead of assessing inter-

analyst consistency, we now show between-instrument consistency. Each 2×2 "block" of results (with cells for P-P, P-N, N-P, and N-N) is for a given analyst, as identified in the column headers. Within this framework, consistency can be judged by the sum of the P-P and N-N cells, which range from 53-60% for analysts and is 65% for the algorithm. More importantly, by comparing P-N and N-P results, we see that the percentage of sites with peaks is much higher for mHVSR-P Seis. than for mHVSR-P Acc. (i.e., P-N percentages are much higher than N-P). This can be visualized through inspection of example results for two sites as shown in Figure 3.7, where peaks are evident in the mHVSR-P Seis. but not in the mHVSR-P Acc. This occurs because the accelerometers, even with 24-bit recorders, are unable to reliably measure ground vibrations from ambient noise.

Table 3.10. Comparison between peak identification results from mHVSR-P Seis. and mHVSR-P Acc. (122 sites)

		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)		Algorithm	
		mHVSR-P Acc.		mHVSR-P Acc.		mHVSR-P Acc.		mHVSR-P Acc.		mHVSR-P Acc.	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
mHVSR-P Seis.	Peak	10%	41%	9%	34%	35%	45%	19%	41%	9%	30%
	No Peak	3%	46%	6%	51%	2%	18%	3%	37%	6%	56%

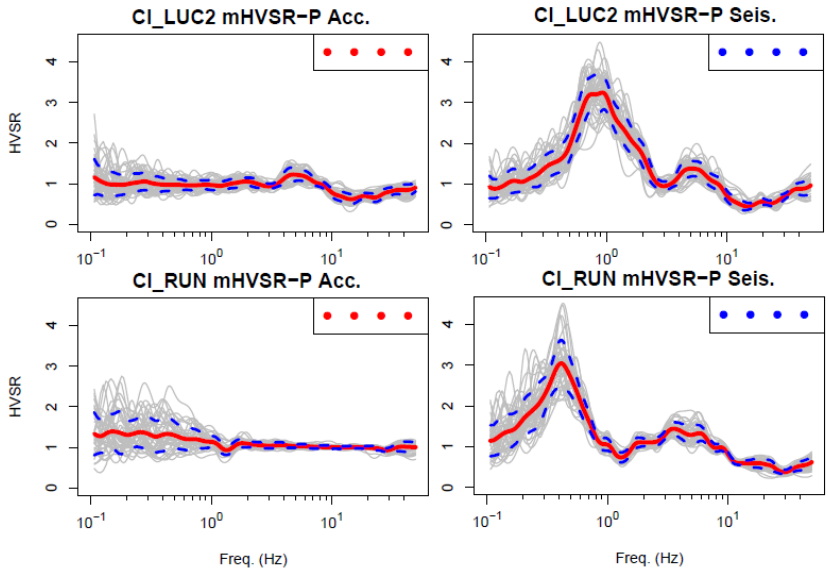


Figure 3.7. Example mHVSR results from accelerometers (mHVSR-P Acc.) and seismometers (mHVSR-P Seis.) for two sites. Peaks evident from the mHVSR-P Seis. are missing from the mHVSR-P Acc.

For the sites identified as having peaks from both instrument types, Figure 3.8 compares the peak frequencies obtained from the fitting process. The peak frequencies generally align for Analysts 1-2 and the algorithm (using the parameters tuned to match outcomes from the conservative group), but there are a significant number of outliers for peaks identified by Analysts 3-4. There is no observable bias in peak frequencies derived from accelerometers vs seismometers.

As a result of the common problem that mHVSR-P Acc. miss peaks from mHVSR-P Seis., we recommend against deriving mHVSR from accelerometers. Unfortunately, this means that mHVSR for such sites will need to be obtained from site visits in which temporary arrays are deployed (mHVSR-T), which is costly.

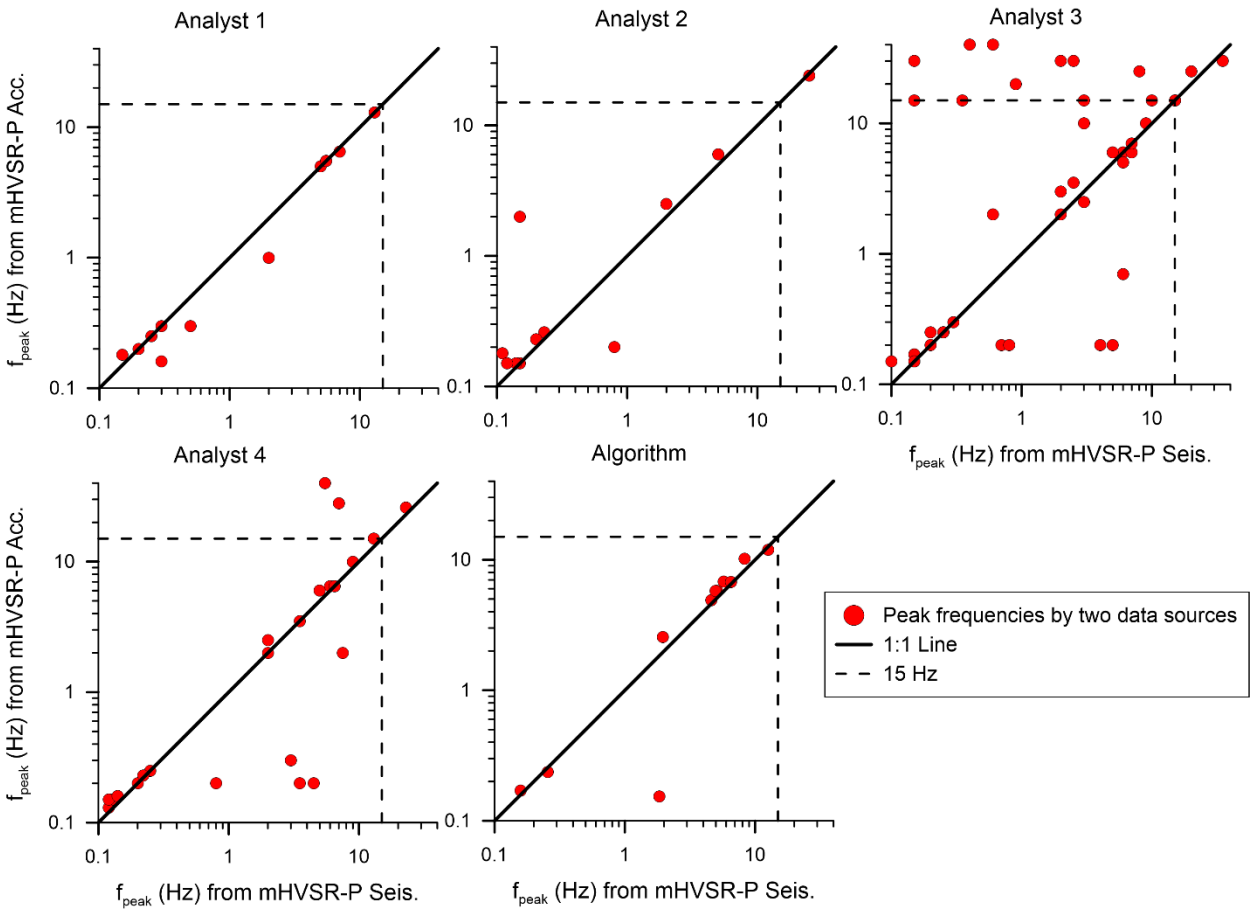


Figure 3.8. Scatter plot of peak frequencies between mHVSR-P Seis. and mHVSR-P Acc.

3.3.2 Comparison of mHVSR from Permanent and Temporary Instruments

There are two main sources of seismometer-based mHVSR in California: data from temporary deployments to measure noise (mHVSR-T) and data from continuously streamed

permanent seismometers (mHVSR-P Seis.). Given that both measure ambient noise and the instrument types are similar, we expect the results to be similar. In this section that hypothesis is tested.

Table 3.11 shows the consistency of peak identification results between mHVSR-T and mHVSR-P Seis. The format of the table matches Table 3.10, in that each 2×2 block corresponds to a particular analyst (or the algorithm). Consistency ranges from 60-82% for analysts (highest for the most conservative analyst) and is 74% for the algorithm. Moreover, percentages in the N-P and P-N cells are generally comparable, meaning that different measurements are not producing systematically higher or lower rates of peak identification. Because the peak occurrence rates from both sources are similar, we have no reason to prefer one mHVSR source over the other, and both are recommended for application.

While the consistent rates in Table 3.11 are relatively high in relation to other comparisons (in Sections 3.3.1 and 3.3.3), even higher rates might be expected given the similarity of the instruments – the principal difference between mHVSR-T and mHVSR-P Seis. is the timing of the measurements and their precise locations at the site. Instrument locations may not exactly coincide because the temporary instruments for mHVSR-T could not always be positioned directly adjacent to the strong motion station (Yong et al, 2013).

To investigate variations of mHVSR with time at a given site, we select three example sites: CI.NSS2 (a site with a clear peak from mHVSR-P Seis. but no peak from mHVSR-T), CI.JNH2 (a site with a clear peak from mHVSR-T but no peak from mHVSR-P Seis.), and CI.OLI (a site with no peak from mHVSR-T and mHVSR-P Seis.). Figure 3.9 presents mHVSR-P Seis. in the first row and mHVSR-T in the second row. The mHVSR-P Seis. are plotted as mean results for 10 separate 2-hour time intervals. Gray shading indicating mHVSR dispersion is plotted for all 10 results, which largely overlap, such that the shading that appears indicates the maximum and minimum limits of dispersion.

The time variations of mHVSR-P Seis. are minimal for site CI.NSS2 and there are no appreciable differences in the peak. The time-variations are relatively significant for sites CI.JNH2 and CI.OLI at frequencies above 1 Hz. Both of these sites have modest-amplitude ordinates in mHVSR-P Seis. (neither are classified as having peaks), such that the time variations of ordinates are relatively visible in the plots. While anecdotal, these results cause us to suggest the following hypotheses regarding time variations: (1) when site have clear peaks, the main features of those peaks may not vary appreciably with time; (2) when sites lack clear peaks, time variations appear appreciable, but generally do not change the overall assessment of a site as having or not having a peak.

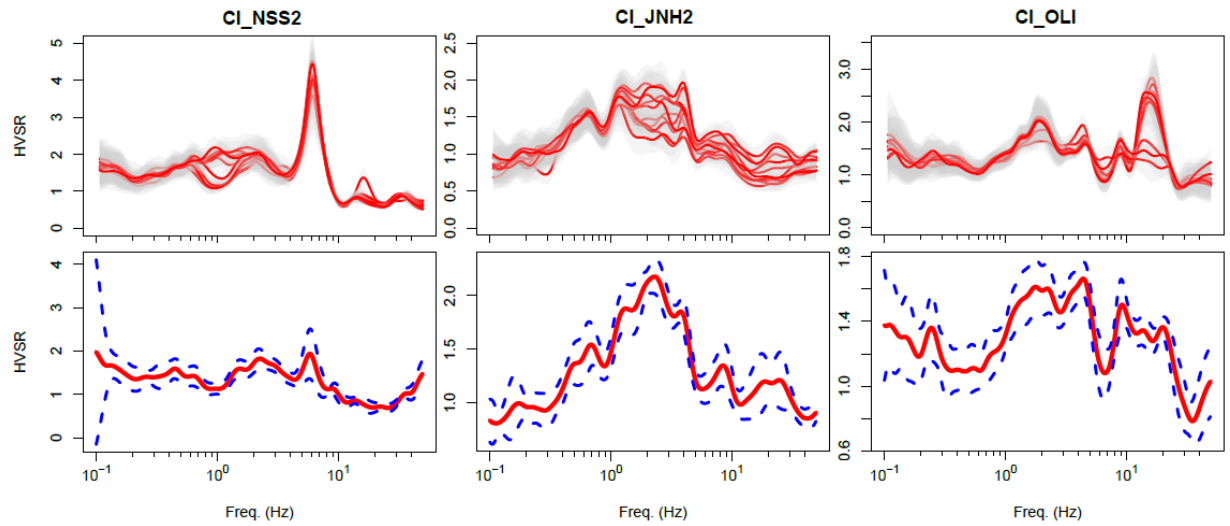


Figure 3.9. mHVSR for three examples sites showing time-variations of mHVSR-P Seis. (top row) and mHVSR-T (bottom row). The time-variable results are shown as 10 means for different dates and the envelop of error bounds.

Table 3.11. Comparison between peak identification results from mHVSR-T and mHVSR-P Seis. (98 sites)

		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Analyst 4 (SKA)		Algorithm	
		mHVSR-T		mHVSR-T		mHVSR-T		mHVSR-T		mHVSR-T	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
mHVSR-P Seis.	Peak	21%	17%	28%	8%	57%	20%	44%	27%	18%	7%
	No Peak	18%	44%	10%	54%	12%	11%	13%	16%	19%	56%

For the sites identified as having peaks from both measurements, Figure 3.10 compares the peak frequencies obtained from the fitting process. The peak frequencies generally align for Analysts 1-2 and the algorithm, but there are a significant number of outliers for Analysts 3-4. There is no observable bias in peak frequencies derived from mHVSR-T vs. mHVSR-P Seis. This general consistency of peak frequencies also confirms the similarity of the two measurements.

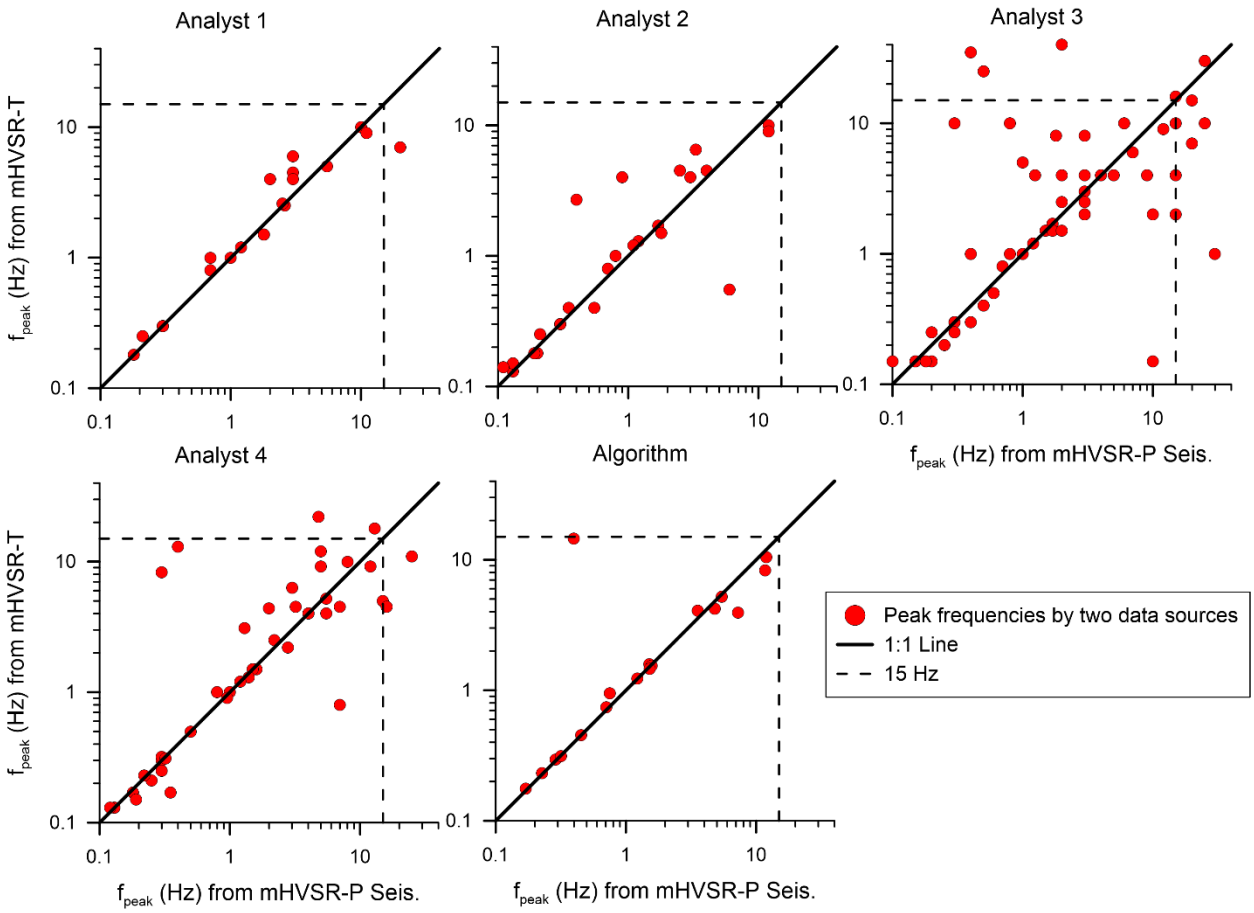


Figure 3.10. Scatter plot of peak frequencies between mHVSR-T and mHVSR-P Seis.

3.3.3 Comparison of eHVSR and mHVSR

Ultimately, the value of HVSR as "site data" is the ability of HVSR-based parameters (like f_{peak}) to predict site response attributes of earthquake ground motions. As mentioned in Chapter 1, while several prior studies have made this connection using eHVSR (Cadet et al. 2012; Ghofrani et al. 2013; Zhao and Xu 2013; Hassani and Atkinson 2016, 2018a, 2018b; Kwak et al. 2017), few have considered mHVSR, which is generally the only information that would be available in forward analyses. Those studies have generally found eHVSR to correlate well with site response, which is not surprising because of circularity in the prediction-validation process. Here we investigate relationships between mHVSR and eHVSR, with specific foci on the consistency with which peaks are identified, and for sites where peaks are found, the characteristics of those peaks. To the extent the eHVSR correlates with actual site response, the relationships explored here would be expected to provide insight into the effectiveness of mHVSR as a means by which to predict earthquake site response.

Table 3.12 shows the consistency of peak identification results between mHVSR and eHVSR for 297 sites. The format of the table matches Tables 3.10-3.11, in that each 2×2 block corresponds to a particular analyst (or the algorithm). Consistency ranges from 60-70% for analysts (highest for the most liberal analyst) and is 69% for the algorithm. The P-N percentages are consistent higher than N-P, which indicates that mHVSR more frequently produces peaks than eHVSR. Interestingly, among the alternate mHVSR peak assessments, the automated algorithm minimizes this false positive rate, although this may be related to its relatively low positive rate generally.

Based on the data in Table 3.12, with the exception of Analyst 3, no-peak sites from mHVSR comprise about 35-63% of the population. Among that population, about 17-19% have eHVSR peaks, which are false negatives with the remainder being true negatives. Hence, the rate of false positives significantly exceeds that for false negatives.

Table 3.12. Comparison between peak identification results from mHVSR and eHVSR (297 sites)

		Analyst 1 (PW)		Analyst 2 (JPS)		Analyst 3 (PZ)		Algorithm	
		eHVSR		eHVSR		eHVSR		eHVSR	
		Peak	No Peak	Peak	No Peak	Peak	No Peak	Peak	No Peak
mHVSR	Peak	31%	34%	25%	27%	66%	25%	18%	19%
	No Peak	6%	29%	9%	39%	4%	5%	12%	51%

Figure 3.11a shows HVSR for two sites with clear peaks from both mHVSR and eHVSR. Figure 3.11b shows HVSR for two sites with clear peaks from mHVSR but no peaks from eHVSR. From these four examples, we observe that when peaks from mHVSR are located away from the low frequency limits of the spectra, they are often also captured by eHVSR, whereas when they occur at low frequencies, they are not seen in eHVSR.

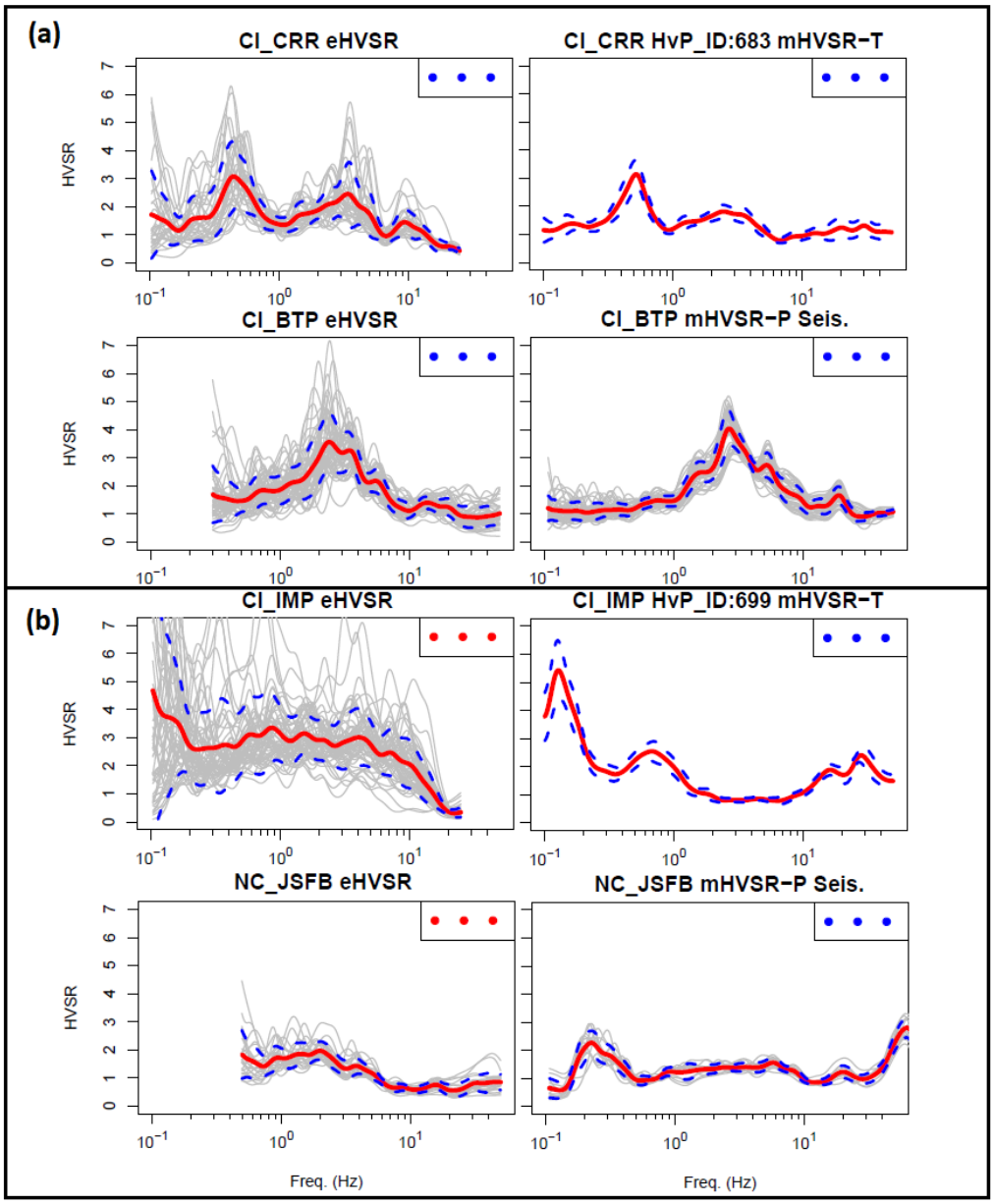


Figure 3.11. Examples of co-located mHVSR and eHVSR for (a) two sites where both shows peaks (P-P) and (b) two sites where mHVSR shows peaks but eHVSR does not (P-N).

With the exception of Analyst 3, the rate of false positives (as in Figure 3.11b) is comparable to the rate of confirmed positives. This implies that a site with an mHVSR peak, as derived using procedures in Section 3.2, has about a 50% probability of not having an eHVSR peak. This is concerning, so we investigate the attributes of peaks for sites with mHVSR peaks. In Figure 3.12, we compare peak frequencies and relative amplitudes (c_1) for sites with mHVSR peaks from the algorithm for P-P sites (peaks are present from both mHVSR and eHVSR) and P-N sites (peaks are present from mHVSR but not eHVSR). The two data sets initially appear to

largely overlap. However, at low frequencies, P-N sites predominate, and overall 31% of the P-N sites have $f_p < 0.5$ Hz. With more earthquake data, many of these sites may be found to have peaks, hence would not be true false positives. Many of the remaining P-N sites (19%) have modest amplitudes ($c_1 \sim 1$), and additional tuning of the peak detection algorithm could relabel them as non-peak sites. However, such adjustments would also likely increase the rate of false negatives. We leave this potential additional tuning of the peak detection algorithm to future work. About 50% of P-N sites have neither low frequencies nor low amplitudes. The causes of these false negatives are not currently understood.

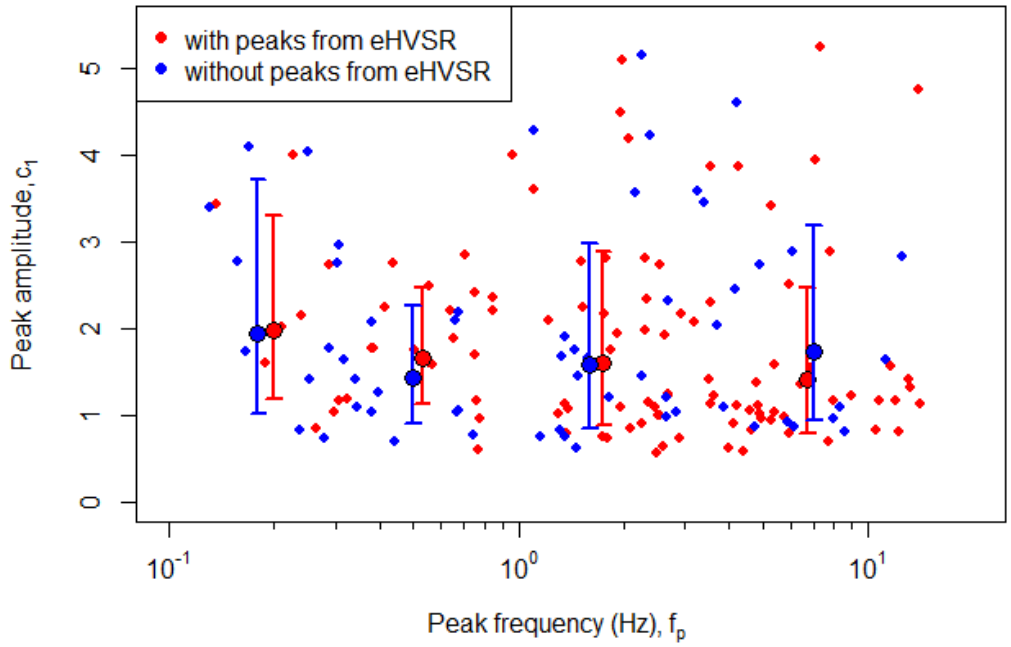


Figure 3.12. Attributes of peaks from mHVSR (peak frequency, f_p; peak amplitude, c_1) segregated among sites with and without peaks from eHVSR

For the sites identified as having peaks from both mHVSR and eHVSR (P-P sites), Figure 3.13 compares the peak frequencies obtained from the fitting process. As in similar plots in prior sections (Figures 3.8 and 3.10), the peak frequencies generally align for Analysts 1-2 and the algorithm, but there are a significant number of outliers for Analyst 3. However, a difference from prior results is that if outliers are excluded, f_{peak} from mHVSR exceeds that from eHVSR by about 3%, suggesting small bias. Such frequency reductions during earthquakes would be expected if the site response was even slightly nonlinear.

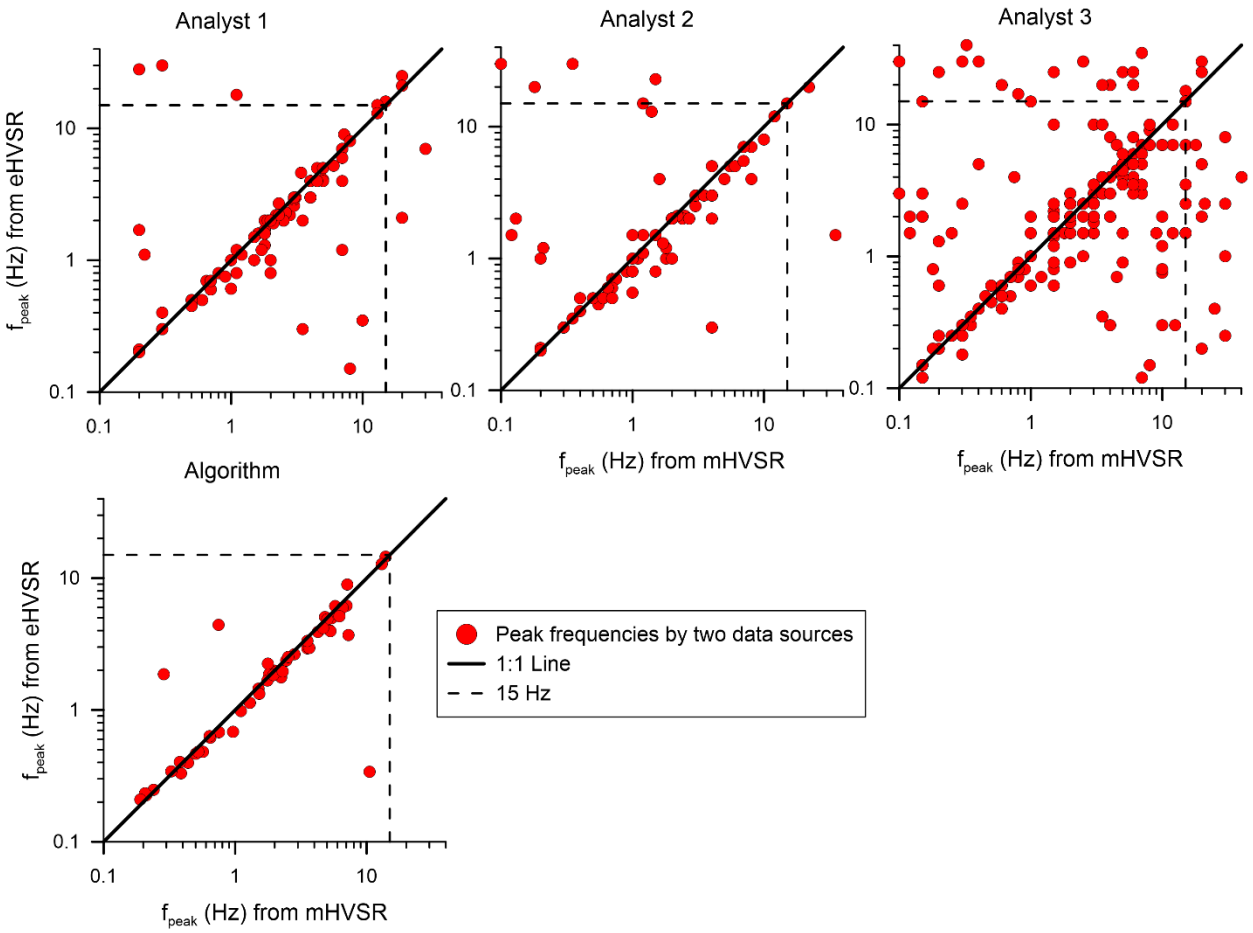


Figure 3.13. Scatter plot of peak frequencies between mHVSR and eHVSR from three analysts and the developed algorithm

Figure 3.14 shows a scatter plot for fitted peak amplitudes (c_1). There are more points above the 45-degree line, which indicates that peak amplitudes from eHVSR are generally slightly larger than those from mHVSR. This finding is consistent with strong motion versus noise comparisons found in soft sites in Mexico (Lermo and Chávez-García, 1994), sites in Iceland (Field et al., 1995), Greece (Atakan et al., 1997), the Garner Valley array in California (Lachet et al., 1996), southern Italy (Theodulidis et al., 1996), and various sites across Europe (Mucciarelli et al., 2003), the Caribbean, and Tehran (Haghshenas et al., 2008).

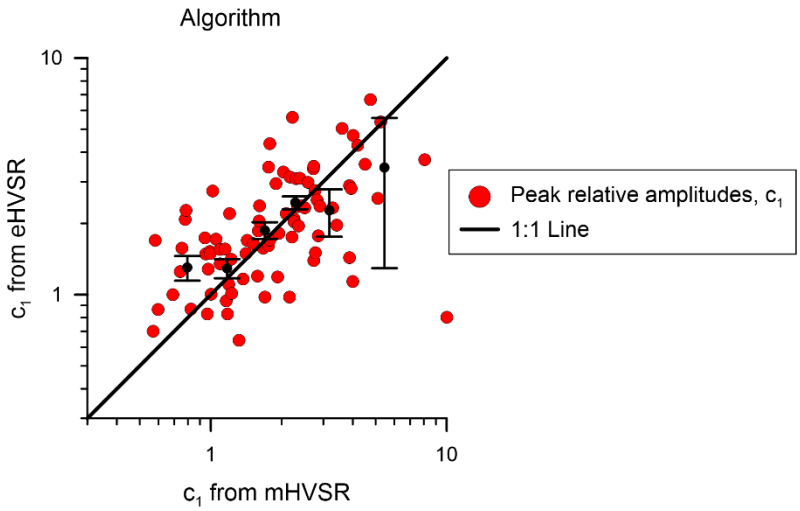


Figure 3.14. Scatter plot of peak relative amplitudes between mHVSR and eHVSR (from algorithm)

Taken together, the mHVSR-eHVSR comparisons suggest there are correlations between them, but also challenges in applying mHVSR for the derivation of site parameters. There are non-negligible numbers of false positives and false negatives. Ultimately, comparisons of mHVSR to actual site response estimates (derived using non-ergodic methods) are needed to judge whether adjustments to the peak selection criteria are needed to bring these into closer alignments. However, despite these challenges, current evidence suggests about a 60-70% success rate for mHVSR identifying site response attributes observed in eHVSR. This leads to two practical conclusions: (1) existing models that are conditioned on eHVSR are likely overly optimistic in terms of the uncertainty reduction that can be gained through the use of HVSR parameters (at least for applications in California) and (2) it is not appropriate to assume equivalence of eHVSR and mHVSR site parameters, and site amplification models intended for practical application should be conditioned on parameters derived from mHVSR.

3.4 HVSR Statistical Properties

We investigate here the statistical properties of HVSR data. In particular, we investigate the shape of the amplitude distributions for mHVSR and eHVSR for different frequencies. For this analysis we use 137 sites with both mHVSR and eHVSR data from continuously streamed seismometers (hence, the data are from the same instruments).

For each site, we compute the difference (i.e., residual) between the HVSR amplitude for a given windowed segment of the data and the mean for that frequency (Section 2.3 describes the data processing procedures, including windowing). After grouping the data for all sites together, residuals were plotted directly and after normalizing using the standard deviations for each site. The latter (normalized residuals) show clearer trends, due to significant differences in standard deviations between sites. Figure 3.15 shows the resulting distributions for mHVSR data at frequencies of 0.3, 2, and 10 Hz. The data are symmetric, indicating an approximately normal distribution, for 2 and 10 Hz. At 0.3 Hz, the data modestly asymmetric, with a heavy positive tail. Figure 3.16 shows the resulting distributions for eHVSR data at frequencies of 0.8, 3, and 10 Hz. The data are asymmetric with heavy positive tails, suggesting an approximately log-normal distribution.

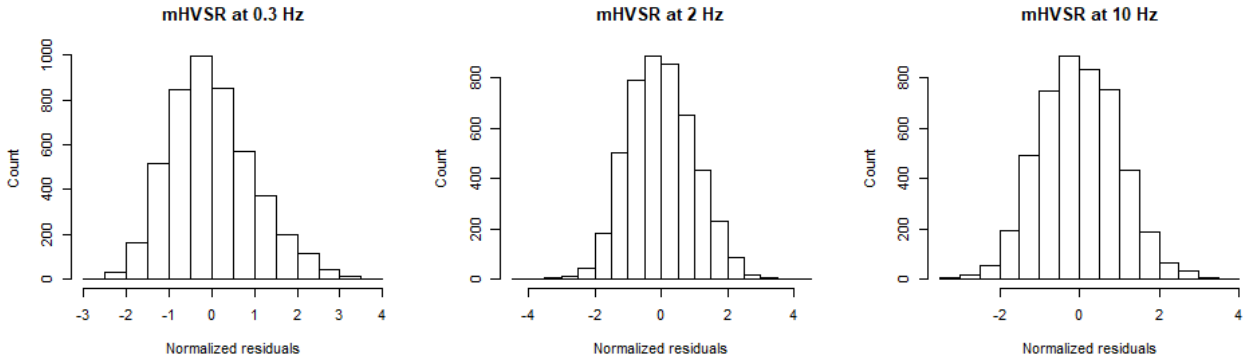


Figure 3.15. Histograms of mHVSR normalized amplitudes for three frequencies.

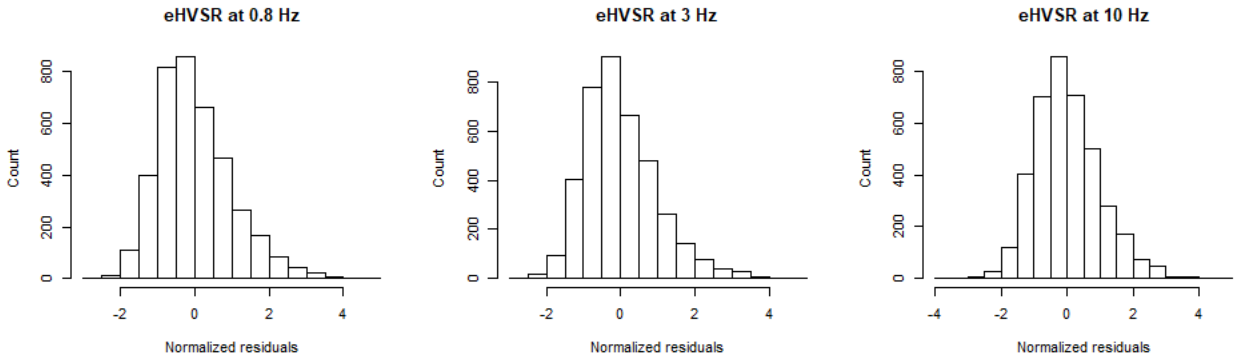


Figure 3.16. Histograms of eHVSR normalized amplitudes for three frequencies.

We perform a Kolmogorov-Smirnov test of distribution (Marsaglia et al. 2003). This test provides the KS statistic, with lower values indicating the distributions is closer to normal and higher values indicating increasing skewness. Figure 3.17 plots the KS statistic for mHVSR data as a function of frequency using data in arithmetic form (to test for normal distribution) and natural log (to test for log normal distribution). Figure 3.18 repeats this for eHVSR data. The mHVSR data could be represented with either normal or log-normal distributions, with low values of the KS statistic either way. The normal distribution is somewhat preferred for frequencies from about 2 to 15 Hz based on visual inspection of histograms, lower values of the KS statistic, and similar results obtained for individual sites. The selection of the normal distribution informed the use of the arithmetic mean in Section 2.3.2. Figure 3.18 indicates much lower KS statistics for lognormal than normal for the eHVSR data. As a result, for eHVSR, a log-normal distribution is clearly preferred.

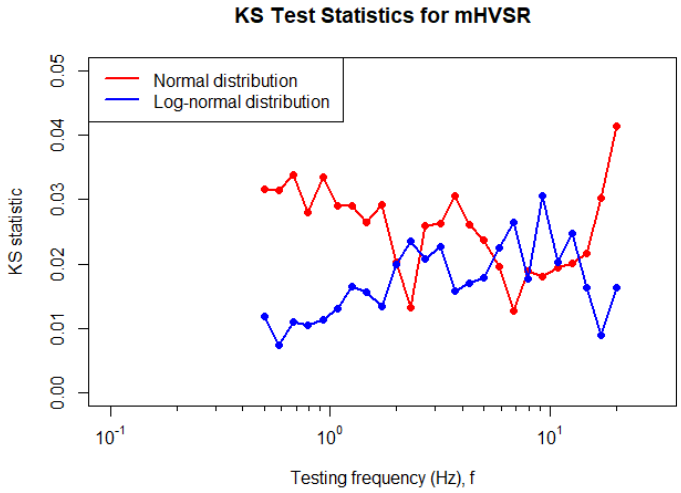


Figure 3.17. KS statistic as function of frequency for mHVSR normalized amplitudes

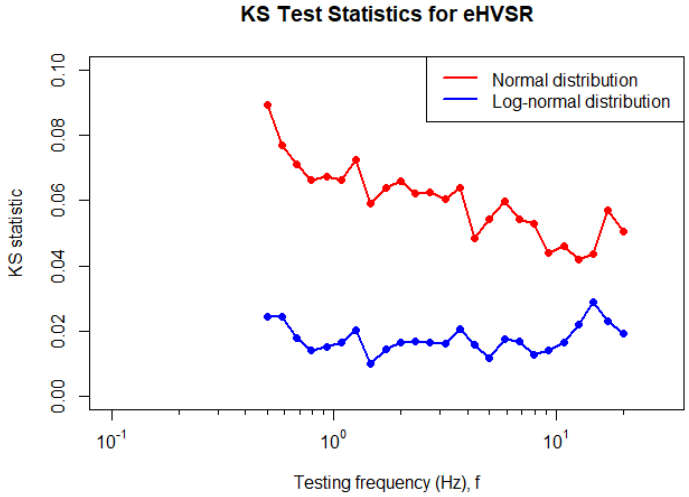


Figure 3.18. KS statistic as function of frequency for eHVSR normalized amplitudes

3.5 Discussion

As explained in Chapter 1, the purpose of measuring and compiling HVSR data is for the derivation of site parameters that can be used in future ground motion models, most likely as an augment to V_{S30}. In a typical forward application (i.e., use of a model to predict ground motions that have not yet occurred), an engineer will measure mHVSR at the site of interest, decide if a peak is present, and if so, identify peak parameters.

In Section 3.2, we have presented procedures for peak identification that we believe to be better suited to California conditions than the SESAME (2004) guidelines that are typically applied in current practice. These procedures are informed by analysts' visual assessments and can largely reproduce peak selections developed by relatively "conservative" or "liberal" analysts (producing relatively few or many sites with peaks, respectively). These procedures use tree regression to identify peak-adjacent plateaus in HVSR, which in turn can be used to identify relative peak amplitudes and peak widths that are considered in the proposed peak identification criteria. The algorithm is coded in R and a Jupyter Notebook and performs the operation of peak identification, and for sites with peaks, peak fitting using a Gaussian function.

In Section 3.3, we describe the consistency of HVSR from different sensors and different vibration sources. These results show that accelerometers, even when used with relatively robust 24-bit data recorders, do not provide a usable basis for mHVSR measurement. The comparisons in Section 3.3.2 show that if mHVSR measurements are made at a slightly different location, and perhaps at a different time, the likelihood of obtaining a significantly different outcome is small but not negligible. Studies of this type, comparing results from multiple noise-based measurements, are relatively rare in the literature, so we are unable to compare to previous findings.

The results in Section 3.3.3 show that if an mHVSR peak is identified, there is only about a 50% chance that a peak will also be present in eHVSR data. This high rate of false positives is influenced by the limited frequency bandwidth of currently available eHVSR data, which does not allow for identification of low frequency peaks that are evident in mHVSR data. This bandwidth issue accounts for about 30% of the false positives, or 15% of the overall population of mHVSR sites with peaks. On the other hand, if no-peak is identified from mHVSR, there is a strong likelihood that the eHVSR also lacks peaks. The consistency of seismic and noise-based HVSR peaks has been studied previously, with most investigators finding consistent results (Lermo and Chávez-García, 1994; Field et al., 1995; Atakan et al., 1997; Lachet et al., 1996; Theodulidis et al., 1996; Mucciarelli et al., 2003; Haghshenas et al., 2008; and Hassani et al. 2019) and a few finding some inconsistent results (Satoh et al. 2001). Our findings are generally more aligned with those of Satoh et al. (2001).

4 Summary and Recommendations

Because HVSR-based parameters are not used currently in ground motion prediction applications, a number of steps are required to build confidence in the efficacy of such parameters, make the required data available to researchers and end users, and produce technically defensible models that can be used in combination with GMMs. This study represents a step in that direction.

We have created an open-source relational database of noise (microtremor)-based HVSR (mHVSR) and associated processing parameters and incorporated this information into an existing community V_S Profile Database (PDB) in the United States (https://doi.org/10.21222/C27H0V, Kwak et al. 2021). The database currently contains mHVSR data from 1303 sites worldwide, 941 of which are in California, and can be scaled up over time to accommodate more data.

Two typical applications of mHVSR data are to identify if a site has peaks, and for sites where peaks are present, establish attributes of the peaks such as peak frequency and amplitude using Eq. (2.1). We develop algorithms for both of these steps that are available in an R script and a Python-based Jupyter Notebook, both of which have been made publicly accessible.

By comparing HVSR data measured using different instruments and from different ground vibration sources, we have shown that mHVSR should not be measured using accelerometers and that different mHVSR measurements from broadband seismometers are relatively consistent with each other in terms of whether peaks are identified and the attributes of those peaks. On the other hand, differences are more frequently encountered between mHVSR and HVSR from earthquake signals (eHVSR).

Among the 941 California sites in the database, 27% have peaks based on the algorithm developed in this study. As such, both site conditions without HVSR peaks are commonly encountered. Accordingly, it will be important for eventual mHVSR-based models to be able to make productive use of not just the attributes of peaks, but also the knowledge of lack of peaks, for site response prediction. We recommend that future research develop such models with suitable conditioning on mHVSR data.

Data and Resources

The HVSR database described in this paper is publicly accessible at the following web portal: https://doi.org/10.21222/C27H0V (Kwak et al. 2021). The data files contained within the portal are mirrored daily to DesignSafe-CI. Data can be plotted directly from the map tools at that site or from a Jupyter Notebook (https://doi.org/10.17603/ds2-nn2e-wm79) on DesignSafe-CI. The Jupyter Notebook also provides functions for peak detection and fitting and generating plots. The *hvsrProc* R package for HVSR processing is available on the Github (https://doi.org/10.5281/zenodo.4724141).

References

Ahdi S.K., Ilhan O., Sadiq S., Bozorgnia Y., Hashash Y.M.A., Park D., Yong A., Stewart J.P. (2018). Development of a United States Community Shear Wave Velocity Profile Database, *5th Conference on Geotechnical Earthquake Engineering and Soil Dynamics (GEESD-V)*, June 10–13, 2018, Austin, Texas.

Ahdi S.K., Ancheta T.D., Contreras V., Kishida T., Kwak D.Y., Kwok A.O.L., Parker G.A., Ruz F., Stewart J.P. (2020). Chapter 5: Site Condition Parameters, in *Data Resources for NGA-Subduction Project, PEER Report No. 2020/02*, Pacific Earthquake Engineering Research Center, UC Berkeley (headquarters).

Ancheta T.D., Darragh R.B., Stewart J.P., Seyhan E., Silva W.J., Chiou B.S.-J., Wooddell K.E., Graves R.W., Kottke A.R., Boore D.M., Kishida T., Donahue J.L. (2014). NGA-West2 database, *Earthq. Spectra.*, **30**, 989-1005.

Anderson J. G., Brune J.N. (1999). Probabilistic seismic hazard assessment without the ergodic assumption, *Seism. Res. Let.*, **70**, 19-28.

Atakan K., Brandsdottir B., Halldorsson P., Fridleifsson G.O. (1997). Site response as a function of near-surface geology in the South Iceland seismic zone. *Nat. Hazards*, **15**(2-3), 139-164.

Atkinson G.M. (2006). Single-station sigma, *Bull. Seismol. Soc. Am.*, **96**, 446–455.

Bonilla L. F., Steidl J.H., Gariel J.-C., Archuleta R.J. (2002). Borehole response studies at the Garner Valley downhole array, Southern California, *Bull. Seismol. Soc. Am.*, **92**, 3165–3179.

Bonilla L.F., Steidl J.H., Lindley G.T., Tumarkin A.G., Archuleta R.J. (1997). Site amplification in the San Fernando Valley, California: Variability of site-effect estimation using the S-wave, coda, and H/V methods, *Bull. Seismol. Soc. Am.*, **87**, 710–730.

Boore D.M. (2010). Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of motion. *Bull. Seismol. Soc. Am.*, **100,** 1830-1835.

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

Bozorgnia Y., Abrahamson N.A., Al Atik L., Ancheta T.D., Atkinson G.M., Baker J.W., Baltay A., Boore D.M., Campbell K.W., Chiou B.S.-J., et al. (2014). NGA-West2 research project, *Earthq. Spectra*, **30**, 973–987.

Brandenberg S.J., Zimmaro P., Stewart J.P., Kwak D.Y., Franke K.W., Moss R.E.S., Cetin K.O., Can G., Ilgac M., Stamatakos J., Weaver T., Kramer S.L. (2020). Next-Generation Liquefaction database, *Earthq. Spectra*, **36**, 939-959.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984). *Classification and Regression Trees*. Wadsworth.

Buckreis T., Winders A., Wang P., Brandenberg S.J., Stewart J.P. (2021). Microtremor Data Collected in Sacramento-San Joaquin Delta Region of California. DesignSafe-CI. doi:10.17603/ds2-dk6t-8610.

Cadet H., Bard P.-Y., Duval A.-M., Bertrand E. (2012). Site effect assessment using KiK-net data: Part 2—Site amplification prediction equation based on f0 and Vsz, *Bull. Earthq. Eng.*, **10**, 451–489. E.

Chatelain J.-L., Guillier B., Cara F., Duval A.-M., Atakan K., Bard P.-Y., and The WP02 SESAME team. (2008). Evaluation of the influence of experimental conditions on H/V results from ambient noise recordings, *Bull. Earthq. Eng.,* **6**, 33–74.

Chatelain J.-L., Guillier B. (2013). Reliable fundamental frequencies of soils and buildings down to 0.1 Hz obtained from ambient vibration recordings with a 4.5-Hz sensor, *Seism. Res. Let.*, **84**, 199-209.

Cheng T., Cox B., Vantassel J.P., Manuel L. (2020). A statistical approach to account for azimuthal variability in single-station HVSR measurements. *Geophys. J. Int.*, **223**, 1040-1053.

Choi Y., Stewart J.P. (2005). Nonlinear site amplification as function of 30 m shear wave velocity, *Earthq. Spectra*, 21 (1), 1-30.

Codd E.F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, **13**(6), 377-387.

Cox B. R., Cheng T., Vantassel J.P., Manuel L. (2020). A statistical representation and frequency-domain window-rejection algorithm for single-station HVSR measurements. *Geophys. J. Int.*, **221**, 2170-2183.

Di Alessandro C., Bonilla L.F., Boore D.M., Rovelli A., Scotti O. (2012). Predominant-period site classification for response spectra prediction equations in Italy, *Bull. Seismol. Soc. Am.*, **102**, 680–695.

Di Giulio G., Cara F., Rovelli A., Lombardo G., Rigano R. (2009). Evidences for strong directional resonances in intensely deformed zones of the Pernicana fault, Mount Etna, Italy. *J. Geophy. Res.*, **114**, B10308.

Field E.H., Clement A.C., Jacob K.H., Aharonian V., Hough S.E., Friberg P.A., Abramian H.A., (1995). Earthquake site-response study in Giumri (formerly Leninakan), Armenia, using ambient noise observations. *Bull. Seismol. Soc. Am.*, **85**(1), 349-353.

GEOVision (2016). Surface Wave Measurements Report, Riverside County, California. Report 16192-01 Rev 2. Prepared for State of California Department of Conservation, California Geological Survey, Strong Motion Instrumentation Program.

GEOVision (2018). Surface Wave Measurements, Santa Clara, Santa Cruz, San Benito, and Monterey Counties. Report 18045-01. Prepared for State of California Department of Conservation, California Geological Survey, Strong Motion Instrumentation Program.

Ghofrani H., Atkinson G.M., Goda K. (2013). Implications of the 2011 M 9.0 Tohoku Japan earthquake for the treatment of site effects in large earthquakes, *Bull. Earthq. Eng.*, **11,** 171-203.

Ghofrani H., Atkinson G.M. (2014). Site condition evaluation using horizontal-to-vertical response spectral ratios of earthquakes in the NGA-West 2 and Japanese databases, *Soil Dyn. Earthq. Eng.*, **67**, 30-43.

Gospe T., Zimmaro P., Wang P., Buckreis T., Ahdi S.K., Yong A., Brandenberg S.J., Stewart J.P. (2020). Supplementing shear wave velocity profile database with microtremor-based H/V spectral ratios, *17th World Conference on Earthquake Engineering (WCEE)*, Paper No. C003010.

Haghshenas E., Bard P.Y., Theodulidis N., Sesame WP04 Team (2008). Empirical evaluation of microtremor H/V spectral ratio. *Bull. Earthq. Eng.*, **6**(1), 75-108.

Hassani B., Atkinson G.M. (2016). Site-effects model for central and eastern North America based on peak frequency, *Bull. Seismol. Soc. Am.*, **106**, 653–664.

Hassani B., Atkinson G.M. (2018a). Site-effects model for central and eastern North America based on peak frequency and average shear-wave velocity, *Bull. Seismol. Soc. Am.*, **107**, 338-350.

Hassani B., Atkinson G.M. (2018b). Application of a site-effects model based on peak frequency and average shear-wave velocity to California, *Bull. Seismol. Soc. Am.*, **108**, 351-357.

Hassani B., Yong A., Atkinson G.M., Feng T., Meng L. (2019). Comparison of site dominant frequency from earthquake and microseismic data in California. *Bull. Seismol. Soc. Am.*, **109**(3), 1034-1040.

Incorporated Research Institutions for Seismology (IRIS), IRIS PASSCAL, https://www.passcal.nmt.edu/content/instrumentation/sensors/sensor-comparison-chart, last accessed <04/04/2021>

IRIS Data Service. https://service.iris.edu/, last accessed <08/01/2021>

Iwahashi J., Pike R.J. (2007). Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*. **86**, 409-440.

Kishida T., Darragh R.B., Chiou B.S.J., Bozorgnia Y., Mazzoni S., Contreras V., Boroschek R., Rojas F., Stewart J.P. (2020). Chapter 3: Ground Motions and Intensity Measures, in *Data Resources for NGA-Subduction Project*, PEER Report 2020/02, J.P. Stewart (editor), Pacific Earthquake Engineering Research Center, UC Berkeley (headquarters).

Konno K., Ohmachi T. (1998). Ground-motion characteristics estimated from spectral ratio between horizontal and vertical components of microtremor. *Bull. Seismol. Soc. Am.*, **88(1),** 228-241.

Kwak D.Y., Stewart J.P., Mandokhai S.-J., Park D. (2017). Supplementing V_{S30} with H/V spectral ratios for predicting site effects, *Bull. Seismol. Soc. Am.*, **107**, 2028-2042.

Kwak D.Y., Ahdi S.K., Wang P., Zimmaro P., Brandenberg S.J., Stewart J.P. (2021). Web portal for shear wave velocity and HVSR databases in support of site response research and applications. UCLA Geotechnical Engineering Group. doi:10.21222/C27H0V.

Molnar S., Cassidy J.F., Castellaro S., Cornou C., Crow H., Hunter J.A., Matsushima S., Sánchez-Sesma F.J., Yong A. (2018). Application of microtremor horizontal-to-vertical spectral ratio (MHVSR) analysis for site characterization: State of the Art. *Surv. Geophys.*, **39**(4), 613–631.

Molnar S., Sirohey A., Bard P-Y., Castellaro S., Cornou C., Cox B., Guillier B., Hassani B., Kawase H., Matsushima S., Sánchez-Sesma F.J., Yong A. (202x). A review of the microtremor horizontal-to-vertical spectral ratio (MHVSR) method, *J. Seism.*, (submitted).

Lachet C., Hatzfeld D., Bard P.Y., Theodulidis N., Papaioannou C., Savvaidis A. (1996). Site effects and microzonation in the city of Thessaloniki (Greece) comparison of different approaches. *Bull. Seismol. Soc. Am.*, **86**, 1692-1703.

Lermo J., Chávez-García F.J. (1994). Are microtremors useful in site response evaluation?. *Bull. Seismol. Soc. Am.*, **84**, 1350-1364.

Marsaglia G., Tsang W.W., Wang J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8/18. doi:10.18637/jss.v008.i18.

Mucciarelli M., Gallipoli M.R., Arcieri M. (2003). The stability of the horizontal-to-vertical spectral ratio of triggered noise and earthquake recordings. *Bull. Seismol. Soc. Am.*, **93**(3), 1407-1412.

NCEDC (2014), Northern California Earthquake Data Center. UC Berkeley Seismological Laboratory. Dataset. doi:10.7932/NCEDC.

Parker G.A., Stewart J.P., Hashash Y.M.A., Rathje E.M., Campbell K.W., Silva W.J. (2019). Empirical linear seismic site amplification in Central and Eastern North America, *Earthq. Spectra*, **35**, 849-881.

Parzen E. (1961). Mathematical considerations in estimation of spectra, *Technometrics*, **3**(2), 167–190.

Petralogix (2017). V_{S30} Site Characterization Report, Los Angeles, Orange, Ventura, San Bernardino, and Riverside Counties. Report 2017-00006.

Rathje E.M., Dawson C., Padgett J.E., Pinelli J.-P., Stanzione D., Adair A., Arduino P., Brandenberg S.J., Cockerill T., Esteva M., et al. (2017). DesignSafe: A new cyberinfrastructure for natural hazards engineering, *Nat. Hazards Rev.*, **18**, 06017001.

Ringler A.T., Bastien P. (2020). A brief introduction to seismic instrumentation: Where does my data come from? *Seism. Res. Let*., **91**, 1074–1083, doi:10.1785/0220190214.

Satoh T., Kawase H., Matsushima S. (2001). Differences between site characteristics obtained from microtremors, S-waves, P-waves, and codas, *Bull. Seismol. Soc. Am.*, **91**, 313–334.

SCEDC (2013). Southern California Earthquake Data Center. Caltech. Dataset. doi:10.7909/C3WD3xH1.

SEED Manual (2012). *Standard for the Exchange of Earthquake Data Reference Manual* (Version 2.4). http://www.fdsn.org/pdf/SEEDManual_V2.4.pdf

SESAME (2004). Guidelines for the Implementation of the H/V spectral ratio technique on ambient vibrations—Measurements, processing and interpretation: European Commission, Project No. EVG1-CT-2000-00026, accessed September 2012, at http://sesame-fp5.obs.ujf-grenoble.fr/

Seyhan E., Stewart J.P. (2014). Semi-empirical nonlinear site amplification from NGA-West 2 data and simulations, *Earthq. Spectra*, **30**, 1241-1256.

Seyhan E., Stewart J.P., Ancheta T.D., Darragh R.B., Graves R.W. (2014). NGA-West2 site database, *Earthq, Spectra*, **30**, 1007-1024.

Stewart J.P., Afshari K., Goulet C.A. (2017). Non-ergodic site response in seismic hazard analysis, *Earthq. Spectra*, **33**, 1385-1414.

Strollo A., Parolai S., Jäckel K.-H., Marzorati S., Bindi D. (2008). Suitability of short-period sensors for retrieving reliable H/V peaks for frequencies less than 1 Hz, *Bull. Seismol. Soc. Am.*, **98**, 671–681.

Theodulidis N., Bard P.Y., Archuleta R., Bouchon M. (1996). Horizontal-to-vertical spectral ratio and geological conditions: The case of Garner Valley downhole array in southern California. *Bull. Seismol. Soc. Am*., **86**(2), 306-319.

Tuan T.T., Scherbaum F., Malischewsky P.G. (2011). On the relationship of peaks and troughs of the ellipticity (H/V) of Rayleigh waves and the transmission response of single layer over half-space models, *Geophys. J. Int*. **184**, 793–800.

Vantassel J. (2020). jpvantassel/hvsrpy: latest (Concept). Zenodo. doi:10.5281/zenodo.3666956.

Wang P. (2020). Predictability and Repeatability of Non-Ergodic Site Response for Diverse Geological Conditions. Ph.D. thesis. UC Los Angeles.

Wang P. (2021). wltcwpf/hvsrProc: First release (Version v1.0.0). doi:10.5281/zenodo.4724141.

Wang P., Zimmaro P., Stewart J.P. (2021). Horizontal-to-Vertical Spectral Ratio Database Access and Analysis. DesignSafe-CI. PRJ-3085. doi:10.17603/ds2-nn2e-wm79.

Wathelet M., Chatelain J.-L., Cornou C., Di Giulio G., Guillier B., Ohrnberger M., Savvaidis A. (2020). Geopsy: A user-friendly open-source tool set for ambient vibration processing, *Seism. Res. Let.*, **91**, 1878-1889.

Wills C.J., Gutierrez C.I., Perez F.G., Branum D.M. (2015). A next generation V_{S30} map for California based on geology and topography, *Bull. Seismol. Soc. Am.,* **105**(6), 3083–3091.

Yong A., Martin A., Stokoe K.H., Diehl J. (2013). ARRA-funded VS30 measurements using multi-technique approach at strong-motion stations in California and central-eastern United States: U.S. Geological Survey Open-File Report 2013–1102, 59 p. and data files, doi:10.3133/ofr20131102.

Zhao J.X., Xu H. (2013). A comparison of V_{S30} and site period as site-effect parameters in response spectral ground-motion prediction equations, *Bull. Seismol. Soc. Am.*, **103**, 1–18.