Development of data-driven models to predict V_{S30} with mHVSR

Kushal Sharma Wagle

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Civil Engineering

Joseph P. Vantassel, Chair

Adrian Rodriguez-Marek

Russell A. Green

May 6, 2025

Blacksburg, VA

Keywords: mHVSR, V_{S30}, Seismic Site Characterization, Machine Learning

Development of data-driven models to predict V_{S30} with mHVSR

Kushal Sharma Wagle

ABSTRACT

This work investigates the potential of using microtremor horizontal-to-vertical spectral ratio (mHVSR) measurements to predict the time-averaged shear-wave velocity in the upper 30 meters (V_{S30}) using data-driven models. We develop a dataset comprising 536 sites with 2,861 three-component ambient noise recordings from global regions, including New Zealand, Taiwan, Italy, Ecuador, Mexico and the United States. The identically processed three-component ambient noise recordings are used to make mHVSR measurements. To predict V_{S30} from mHVSR, we consider two types of models: low-dimensional models, which use features of the mHVSR curve such as the fundamental site frequency ($f_{0,HVSR}$) and peak amplitude ($A_{0,HVSR}$), and high-dimensional models, which use the entire mHVSR mean curve to predict V_{S30}. In addition, we integrate topographic features from a 1 arc-second digital elevation model (DEM) for both model types using binned elevation as a proxy for geologic composition and relative elevation as a proxy for topography. The low-dimensional models are shown to reasonably predict V_{S30}, coefficient of determination (R^2) up to 0.69 on the testing set, when considering both mHVSR and topographic features. The high-dimensional models are shown to achieve improved accuracy to the low-dimensional models (R^2 of 0.82 on the testing set) when using the mean mHVSR curve regardless of whether the mean mHVSR curve is supplemented with the additional topographic features. These findings demonstrate that while low-dimensional features of the mHVSR curve are informative, leveraging the full shape of the mHVSR curve leads to improved prediction of V_{S30}. Furthermore, the use of the mHVSR mean curve has the additional advantage that it does not

require the extraction of features and can be used at all sites including those with and without resonant peaks. We compare the results with a model developed to predict V_{S30} from remote sensing data and demonstrate the utility of the models by predicting V_{S30} at 1,855 broadband recording stations across North America.

Development of data-driven models to predict V_{S30} with mHVSR

Kushal Sharma Wagle

GENERAL AUDIENCE ABSTRACT

The amount of damage caused by earthquakes is strongly controlled by local ground conditions. Soft ground can dramatically amplify earthquake shaking compared to firm bedrock, making those areas more vulnerable to damage. To understand which areas may be most at risk, engineers use a parameter called V_{S30}, which measures how soft the ground at a site is in the top 30 meters. V_{S30} plays a central role in earthquake hazard maps, building codes, and infrastructure design. However, traditional methods for measuring V_{S30} in the field are expensive, labor-intensive, and spatially limited, leaving large portions of the world unmapped and unprepared.

My research proposes a practical solution: using a method called the Horizontal-to-Vertical Spectral Ratio (HVSR), to estimate V_{S30} in a low-cost, non-invasive way. HVSR involves recording weak, natural ground vibrations, called microtremors, that are present in the ground, even when there is no earthquake. By analyzing the way these vibrations behave, we can learn about the layers of soil and rock beneath the surface without drilling or using heavy equipment.

While HVSR-based models have been developed previously using smaller regional datasets, this thesis develops a global dataset of HVSR and V_{S30} measurements. These data cover diverse geological environments that span multiple continents. Using this rich dataset, I developed machine learning models that combine HVSR with topographic features to produce better

predictions of V_{S30}. The result is a generalized, data-driven set of models capable of predicting V_{S30} from HVSR measurements.

The impact of this work extends beyond earthquake engineering. It supports global efforts in urban planning, infrastructure development, and disaster risk reduction, especially in rapidly growing cities where formal geotechnical studies are lacking. It can enhance the resolution of global seismic hazard maps, inform safer land-use policies, and serve as a foundational layer for climate-resilient infrastructure planning, particularly in areas prone to multiple hazards such as landslides or flooding.

By lowering the barriers to site characterization, this research contributes to a more informed and prepared global community, where science and data can guide decisions that ultimately save lives and reduce economic losses.

DEDICATION

To my parents,

for your unwavering love, endless support, and the countless sacrifices you've made so that I could chase my dreams.

Your belief in me has been my foundation, and your strength has been my guide.

This work is as much yours as it is mine.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors, Dr. Joseph Vantassel and Dr. Adrian Rodriguez-Marek for their unwavering guidance, support, and mentorship throughout the course of this research. Their expertise, patience, and critical insights have been invaluable at every stage of this work. I am truly fortunate to have had the opportunity to learn under their direction.

I'm especially grateful to Dr. Russell Green for his constructive feedback and encouragement that helped shape this thesis. I also extend my thanks to the faculty and staff of the Department of Civil and Environmental Engineering at Virginia Tech for providing an environment that fosters research, learning, and collaboration.

I would like to acknowledge the researchers and institutions worldwide who have made their seismic and geotechnical datasets publicly available or provided to me for this research. This thesis would not have been possible without their openness and contributions to the scientific community.

To my family, your constant love, encouragement, and belief in me have been my greatest source of strength. I am especially grateful to my parents and my sisters for their sacrifices and for always supporting my academic pursuits, even from afar.

Finally, I thank my girlfriend and my friends, both near and far, for grounding me and reminding me to find joy in the process. This journey would not have been the same without you.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This thesis proposes a methodology for estimating the time-averaged shear-wave velocity in the upper 30 meters of the subsurface (V_{S30}) based on the horizontal-to-vertical spectral ratio (HVSR) of ambient seismic noise, also known as microtremors. The V_{S30} parameter is widely recognized as a key proxy for characterizing site effects and is commonly employed in ground motion prediction models (GMMs) for earthquake engineering applications.

Earthquake ground motions refer to the movement of the Earth's surface caused by seismic waves generated during an earthquake. These motions can significantly impact civil infrastructure, causing damage and, in some cases, structural collapse. Ground motions are complex phenomena and are influenced by many factors including earthquake magnitude, fault mechanism, rupture depth, travel path, and local geological conditions. Earthquake engineers and seismologist have sought to parameterize earthquake ground motions using different earthquake characteristics. The objective of parameterization of earthquake ground motions is twofold: a) to facilitate the prediction of future ground motions using earthquake characteristics, and b) to enable the estimation of their effects on the built environment. The parameterization of an earthquake ground motion is generally broken into descriptions of their intensity, duration, and frequency content. Ground motion intensity has been quantified by parameters such as peak ground acceleration (PGA), velocity (PGV), and displacement (PGD) that seek to describe the intensity of the motion. Duration is another critical factor, as longer-duration shaking tends to increase the cumulative energy imparted to a structure, leading to greater damage potential. Measures such as significant duration or bracketed duration are means to quantify the duration of ground motions. An alternative parameter that incorporates both intensity and duration is Arias Intensity. Ground motion frequency content also plays a critical role, as different structures have unique natural

frequencies, also referred to as resonant frequencies, which make them particularly vulnerable to specific frequency ranges of shaking. Ground motion frequency content has been quantified using tools such as pseudo-spectral acceleration (PSA) and Fourier amplitude spectra (FAS).

Ground motions recorded during earthquakes are influenced by three primary factors: the source, the travel path of seismic waves, and local site conditions. Accurate prediction of ground shaking therefore requires understanding and modeling of each of these components. This thesis focuses on the influence of *site effects*—modifications to ground motion due to local geological and geotechnical conditions at a site. Site effects can significantly amplify or attenuate seismic waves, altering their intensity, duration, and frequency content, and in turn contributing to structural damage during earthquakes. While rigorous modeling for site effects requires a detailed characterization of the site's structure to a substantial depth, such a detailed characterization is not practical for most structures and therefore is reserved for only critical structures. As a result, engineers have sought proxies to site structure that can be used to incorporate site effects without requiring a detailed characterization effort.

Historically, the quantification of site effects relied on qualitative or semi-quantitative proxies such as surface geology (e.g., rock, soil), standard penetration test (SPT) blow count (Seed et al., 1976; Borcherdt, 1970), soil thickness, and depth to bedrock (Aki, 1993). However, these proxies were limited by subjectivity, inconsistent definitions, and poor correlation with site effects. Other early approaches included estimating site effects directly from measurements of ambient noise (Nakamura, 1989), which required specialized measurements, lacked standardization, and had limited predictive capacity (SESAME, 2004). Later efforts used detailed characterization of a site's shear wave velocity (V_S) profile, non-linear soil and rock properties from lab measurements, and one-dimensional wave propagation to estimate site effects. While this approach worked well,

the detailed site characterization required was not practical for use at all sites. Therefore, efforts have been made to find proxies to capture features of a site's 1D structure without detailed characterization. The most common proxy currently in use, and the focus of this thesis, is the time averaged shear-wave velocity of a site's top 30 meters (V_{S30}) (Borcherdt 1992, 1994).

V_{S30} has gained widespread acceptance due to its empirical correlation with site amplification and its relative simplicity. It has been incorporated into building codes and standards such as the International Building Code (IBC) and the Minimum Design Loads and Associated Criteria for Buildings and Other Structures (ASCE/SEI 7-22). While V_{S30} represents only the shallow velocity structure (depths < 30 m), it has been shown to serve as a robust parameter for incorporating site effects into ground motion models (GMMs) (Cadet and Duval, 2009; Borcherdt, 2012). It has been extensively used in GMMs, including the Next Generation Attenuation (NGA) models (e.g., Campbell and Bozorgnia, 2014; Abrahamson et al., 2014, Chiou and Youngs, 2014; Boore et al, 2014, Parker et al., 2022). Despite its advantages, directly measuring V_{S30} via borehole logging or geophysical surveys remains costly and labor-intensive. As a result, less than 800 stations have measured V_{S30} values among the total of over 5,500 seismic stations in the combined NGA-West2 and NGA-East networks, indicating almost 85% of the stations do not have measured V_{S30}. Consequently, many efforts have focused on estimating V_{S30} from other readily available data. Early efforts to estimate V_{S30} relied on proxies such as geologic age (Wills and Clahan, 2006), topographic slope (Wald and Allen, 2007), and geologic unit (Parker et al., 2017). While these methods provided practical alternatives, they often failed to account for the frequency-dependent resonance effects of subsurface layers, which are critical for site response analysis. More recently, advanced techniques such as machine learning and remote sensing have been developed to improve V_{S30} prediction. Studies like those by Geyin and Maurer (2023) leverage these modern

approaches to estimate V_{S30} over large regions where direct measurements remain sparse. While an important step forward these methods suffer from large uncertainty and low resolution compared to site-specific measurements. Given the challenges associated with directly measuring V_{S30} at all sites, this work seeks to develop an alternative approach to predict V_{S30} using single-station measurements. In particular, this work sees to answer whether HVSR computed from ambient noise can be used as an alternative for making site-specific estimates of V_{S30}. If successful, V_{S30} predictions could be made using existing broadband recording stations in the United States, many of which are in proximity (< 100 m) of seismic stations. Figure 1 illustrates the spatial distribution of 1,855 broadband stations across North America where such V_{S30} predictions could be made. The spatial coverage of these broadband stations highlights the opportunity to leverage existing ambient noise data at a significant number of seismic stations for HVSR-based V_{S30} estimation.

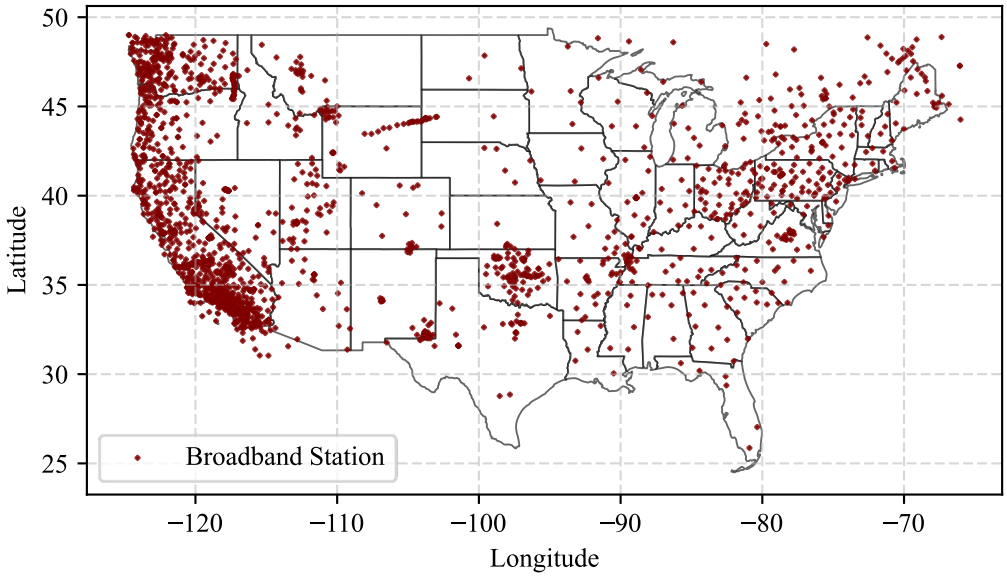

Figure 1: Distribution of broadband station across North America capable of recording continuous microtremor data which could be used to estimate V_{S30}.

The information potential of the relative amplitude of the horizontal and vertical Fourier spectra of ambient noise (also known as microtremors) was initially explored by Nogoshi and Igarashi (1971). Later efforts by Nakamura (1989) compared the ratio of the horizontal to vertical Fourier spectra to empirical estimates of site effects, indicating some level of correlation. These efforts resulted in the development of the microtremor horizontal-to-vertical spectral ratio (mHVSR) as a tool for better understanding site effects. Computing mHVSR requires the measurements of three-component ambient noise to calculate the ratio of the combined horizontal (north-south and east-west) Fourier spectra divided by the vertical spectrum:

$$HVSR = \frac{\sqrt{S_{NS} \cdot S_{EW}}}{S_{vertical}} \quad (1)$$

where S_{NS}, S_{EW}, and $S_{Vertical}$ are the Fourier amplitude spectra of the north-south, east-west, and vertical ground motion components, respectively. In Equation (1) the two horizontal components are combined using the geometric mean (SESAME, 2004) although many alternative approaches have been proposed in the literature. A summary of methods is provided in Vantassel (2025). mHVSR has gained traction for its simplicity and cost-effectiveness. While the full informative potential of HVSR is still debated it is generally agreed that HVSR is an effective means of estimating a site's fundamental frequency (f_0) (SESAME, 2004; Molnar et al., 2022).

Our understanding of the physical principles behind the correlation of HVSR and site response are still evolving. In his paper, Nakamura (1989) proposed that the ratio of the horizontal and vertical components of ambient vibrations could isolate site-specific amplification effects, as the vertical component was assumed to remain relatively unaffected by subsurface layering. However, this assumption was challenged early on by Lermo and Chávez-García (1993), who observed that vertical amplification can occur in deep sedimentary basins such as Mexico City, potentially leading to misinterpretation of HVSR curves. These findings that HVSR may not

directly estimate site effects were later reinforced by the SESAME (Site Effects Assessment Using Ambient Excitations) Project (2001–2004), a European Commission funded initiative, which aimed to standardize HVSR processing and improve its reliability. The SESAME Project findings revealed that in some geological settings, particularly deep sedimentary basins and soft soil profiles, the vertical component can be amplified. This conflicts with Nakamura's original assumption and suggests that the HVSR peak does not always purely represent the fundamental site resonance frequency (Kuo et. al, 2015). Furthermore, the wavefield differences between microtremors, dominated by surface waves and anthropogenic signals, and earthquakes, encompassing a broader spectrum of body and surface waves, can affect the HVSR curve's shape and amplitude (Theodulidis et al., 2006; Pilz et al., 2009). As such, HVSR is now understood to reflect not only SH-wave resonance but also the character of the incoming wavefield and site-specific structural complexities. The SESAME project helped formalize this understanding by issuing critical guidelines for HVSR data acquisition, processing, and interpretation (SESAME, 2004). It also introduced HVSR as a tool for linking empirical observations with theoretical site response and acknowledged HVSR's limitations at hard-rock sites where amplification is minimal (Mucciarelli and Gallipoli, 2001; Bard, 2004).

The widespread use of HVSR has prompted the development of many open-source software tools to facilitate HVSR processing. One of the most significant advancements was the development of Geopsy (Wathelet et. al., 2020), an open-source software suite launched during the SESAME project and initially released in 2005. Geopsy marked a significant leap in HVSR processing. It streamlined the computation of HVSR from ambient noise and enabled joint analysis of HVSR with other geophysical methods such as multichannel analysis of surface waves (MASW). Building on these foundational efforts, the development of automated HVSR analysis

tools gained momentum. The HVSR Incorporated Research Institutions for Seismology (IRIS) Station Toolbox (Bahavar et al., 2020) was introduced to support station-level HVSR processing. Around the same time, the Python package hvsrpy was developed, enabling automated window rejection, peak identification, and uncertainty quantification for large datasets (Vantassel, 2020; 2025). This automation trend continued with the creation of cloud-based tools such as HVSRweb, which allowed users to upload ambient noise data for real-time HVSR processing (Vantassel et al., 2021). Machine learning frameworks further advanced the field, with studies using deep neural networks (DNNs) to classify HVSR curves and predict Vs profiles, reducing subjectivity in peak selection (Pan et al., 2022). The introduction of the HV Noise and Earthquake Automatic Analysis (HVNEA) software package (Vassallo et al., 2023) further streamlined the automatic computation of HVSR from continuous seismic recordings. Most recently, the AutoHVSR algorithm (Vantassel et al., 2023) exemplifies the integration of machine learning into HVSR analysis, using models trained on diverse datasets to fully automate HVSR processing.

Recent studies suggest that mHVSR curves contain implicit information about subsurface velocity structure, including V_{S30} (Kuo et al., 2015; Molnar et al., 2017). However, the relationship between mHVSR and V_{S30} remains non-unique and regionally variable. We hypothesize that data-driven models, developed using machine learning (ML), may be capable of identifying latent patterns in mHVSR spectra and linking them to V_{S30} values across diverse geological settings. This thesis develops a novel ML framework to predict V_{S30} directly from mHVSR data, bypassing traditional proxies and enhancing the accessibility of site-specific seismic hazard assessments.

The thesis continues by presenting a detailed literature review on previous HVSR-V_{S30} correlations, followed by a journal paper manuscript where new ML models for predicting V_{S30} from mHVSR are presented. The manuscript includes the details of data collection, data

preparation, and ML model development. The ML model's predictions are compared with V_{S30} predictions from existing models using remote sensing data. The manuscript also includes a section where the best ML models are employed to estimate the V_{S30} of the broadband stations with recordings of ambient noise in North America. The manuscript is followed by the engineering applications of this thesis and the conclusion.

Chapter 2: Literature Review

Over the past two decades, researchers have investigated the use of the microtremor and earthquake HVSR (i.e., mHVSR and eHVSR, respectively) to estimate a site's V_{S30}. The investigations have focused primarily on estimating V_{S30} from the site's resonant frequency measured using HVSR. The physical basis behind the development of such models is that softer sites (i.e., sites with lower V_{S30}) should have a lower resonant frequency whereas stiffer sites should have higher resonant frequency. Here, a key distinction is made between two terminologies adopted in the literature for describing a site's resonant frequency: f_{peak} and $f_{0,HVSR}$. f_{peak} refers to the highest amplitude peak observed in the HVSR curve. In contrast, $f_{0,HVSR}$ refers to the fundamental (i.e., the lowest) frequency peak observed in the HVSR curve, regardless of its amplitude. In some cases, f_{peak} and $f_{0,HVSR}$ may be the same (e.g., only one peak is present), however in general they are not the same. Two example mHVSR mean curves where f_{peak} and $f_{0,HVSR}$ are and are not the same are shown in Figure 2. In the following literature review we use these terms consistently to avoid confusing the reader even when these terms may not have been used by the original authors. The remainder of this chapter reviews the major studies that have attempted to correlate HVSR with V_{S30}. Each study is described in terms of its methodology, accuracy levels, and strengths and weaknesses. The chapter closes with a brief overview of how this thesis aims to address the limitations of previous studies.

1.1. Early Foundations and Key Concepts

The pioneering work and foundational concepts for correlating HVSR measurements with V_{S30} emerged from a series of studies that bridged seismic site characterization, ambient noise analysis, and earthquake engineering. Nakamura's (1989) work established mHVSR as a tool for estimating a site's resonant frequencies. While Nakamura's original work focused on estimating site response

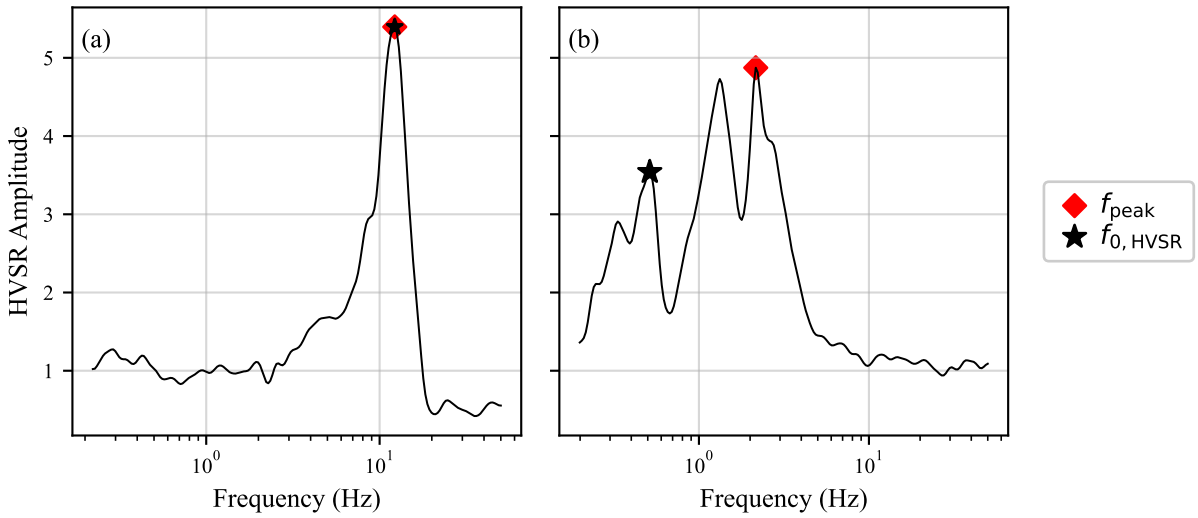

Figure 2: Two example mHVSR mean curves where f_{peak} and $f_{0,HVSR}$ (a) are the same and (b) are different. The data in both panels are from Yong et al. (2013).

directly using mHVSR rather than indirectly using V_{S30}, his efforts laid the groundwork to link HVSR-derived parameters to a site's stiffness profiles. Ibs-von Seht and Wohlenberg (1999) demonstrated that $f_{0,HVSR}$ could be used to estimate sediment thickness through indirect assumptions of a site's stiffness with depth. Their approach while effective at sites with relatively simple geology, assumed idealized one-dimensional layering, which limits its application at sites with complex geology (e.g., lateral heterogeneity or multiple impedance contrasts). This limitation underscored the need for more nuanced interpretations of HVSR curves, especially when deeper velocity contrasts (e.g., bedrock at >30 m) dominate over the shallow structure. For instance, Castellaro and Mulargia (2009) pointed out that the relationship between HVSR peak frequency and V_{S30} can be complex because HVSR can be influenced by deeper impedance contrasts that overshadow the upper 30-meter velocity structure. This highlights one of the primary challenges for predicting V_{S30} from HVSR: while HVSR is sensitive to the stiffness of a site's upper 30 m it can also be influenced by much deeper structure that are not correlated to V_{S30}.

1.2. Microtremor vs. Earthquake HVSR

In the literature HVSR has been computed from both ambient noise (mHVSR) and earthquake recordings (eHVSR). While mHVSR is often favored for its convenience, requiring no specific seismic event, some researchers have argued that eHVSR may be able to better capture a site's seismic behavior (Ghofrani and Atkinson, 2014; Ahn et al., 2021). However, eHVSR is not practical for most forward applications because it depends on the availability of seismic events that may not be available or are rare (e.g., in areas of low to moderate seismicity). While some have argued that eHVSR and mHVSR can be used interchangeable, recent work by Vantassel et al. (2024) indicates that mHVSR and eHVSR mean curves are similar only about 60% of the time. As a result, models developed using eHVSR as input cannot be assumed to operate to the same level of accuracy when supplied with mHVSR measurements. These findings align with earlier studies (e.g., Theodulidis et al., 2006; Pilz et al., 2009) that similarly noted amplitude discrepancies and occasional frequency shifts between mHVSR and eHVSR recordings. These results illustrate that while eHVSR may offer a more direct measure of how a site behaves under strong ground shaking it is not necessarily a practical characterization tool as it requires a permanent instrument and the occurrence of multiple earthquake events. mHVSR, in contrast, can be recorded using short deployments (between 20 minutes and 2 hours) of temporary instruments that can be done on demand.

1.3. Global Research on HVSR– V_{S30} Correlations

This section discusses existing published correlations to predict V_{S30} from HVSR in chronological order. The existing correlations between HVSR and V_{S30} are summarized in Figure 3. As mentioned previously, prior studies have used both f_{peak} and $f_{0,HVSR}$ for developing predictive models and so we list both on the figure's horizontal axis as the one plotted is model dependent.

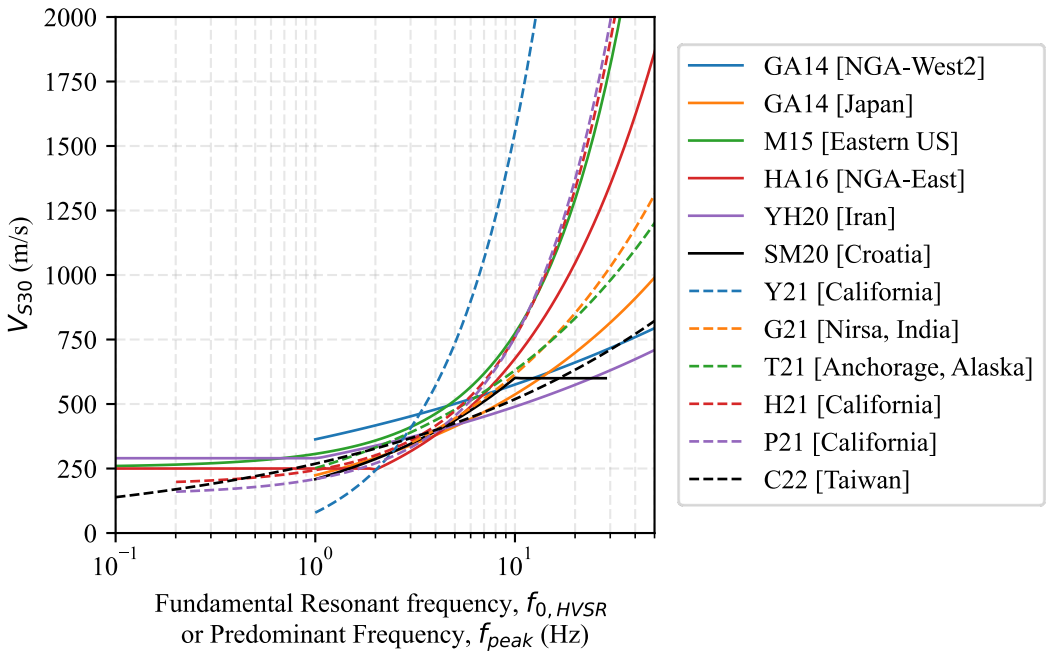

Figure 3: Existing correlations between HVSR resonant frequency using either f_{peak} or $f_{0,HVSR}$ and V_{S30}. The legend abbreviates each reference based on the author(s), the year of publication, and the geographic region for which the model was developed.

Ghofrani and Atkinson (2014) [GA14] conducted a study to assess whether eHVSR could serve as an alternative or supplementary proxy for V_{S30} in site classification and seismic site response analysis. The study utilized three ground-motion databases: NGA-West2, KiK-net, and K-net datasets. These datasets included records from strong-motion stations across diverse geological settings, enabling a broad evaluation of the relationship between HVSR peak parameters and V_{S30}. Notably, rather than use the traditional FAS-based HVSR, they used the ratio of the 5%-damped pseudospectral acceleration (PSA). Statistical regression was employed to correlate the HVSR-derived features of peak frequency (f_{peak}) and peak amplitude (A_{peak}) to V_{S30}. A total of six regression models were developed, three from the Japanese KiK-net and K-net databases using f_{peak}, A_{peak}, and a combination of both features, and three from the NGA-West2 database, following the same approach. The coefficient of determination (R^2) for these models was 0.30, 0.23, and 0.42, respectively, for the Japanese dataset, and 0.10, 0.23, and 0.30 for the NGA-

West2 dataset, indicating limited correlation between HVSR parameters and V_{S30}. Due to dataset limitations, the Ghofrani and Atkinson (2014) models are only applicable for f_{peak} values greater than 1 Hz. The best performing Ghofrani and Atkinson (2014) models from the Japan and NGA-West2 datasets which used both f_{peak} and A_{peak} are shown in Figure 3.

McNamara et al. (2015) [M15] conducted a study to investigate the correlation between eHVSR $f_{0,HVSR}$ and measured V_{S30} values in the Central Virginia Seismic Zone (CVSZ) and surrounding regions in the Eastern United States [EUS]. The primary goal was to assess the utility of HVSR-derived resonance frequencies as a practical proxy for estimating V_{S30}, especially in regions with sparse geophysical measurements. The study used ambient noise recordings to compute HVSR curves and focused on 21 out of 66 sites where clear spectral peaks were observed. This study utilized FAS-based HVSR from ambient noise. A simple linear regression model was developed correlating $f_{0,HVSR}$ to V_{S30}, resulting in correlation coefficient (R) of 0.89 and a standard deviation of residuals of 78.91 m/s, indicating a strong predictive capability within the CVSZ. The authors emphasized that the method is effective only when a distinct HVSR peak is present, limiting its applicability at sites with weak or complex resonance behavior, often the case for bedrock or borehole installations. Although the model demonstrated strong predictive performance, it is developed with a very small dataset and its regional specificity and dependence on observable spectral peaks constrain its generalizability to other geologic settings.

Hassani and Atkinson (2016) [HA16] investigated the applicability of $f_{0,HVSR}$ as a proxy for V_{S30} in Central and Eastern North America (CENA). Their study aimed to develop a new correlation between $f_{0,HVSR}$ and V_{S30} using the NGA-East database, which includes 1,379 seismic stations. However, only 535 stations had recorded at least three events, and among these, only 315 stations had identifiable $f_{0,HVSR}$ values, representing only 23% of the NGA-East dataset. The

ground-motion database consisted of 5,783 three-component records from 225 seismic events, mostly with moment magnitudes between 2.5 and 4.5 and recorded at hypocentral distances up to 800 km. They computed eHVSR from the 5% damped PSA rather than conventional Fourier spectra. The authors developed a predictive equation for V_{S30} using $f_{0,HVSR}$ using a bilinear regression. The proposed model exhibited lower uncertainty (standard deviation of 0.14 log10 units) than other proxy-based approaches such as topographic slope and surface geology, which typically have standard deviation of 0.25log10 units. While work by Hassani and Atkinson (2016) represented another significant contribution to the prediction of V_{S30} from HVSR, it was limited by the small fraction of stations (315) where $f_{0,HVSR}$ could be reliably determined of which only 41 had a direct measurement of V_{S30}.

Yaghmaei-Sabegh and Hassani (2020) [YH20] investigated the correlation between eHVSR and V_{S30} using a dataset from the Bureau of Housing and Research Center (BHRC) of Iran comprising 560 seismic stations. These stations recorded 30 seismic events with moment magnitudes (M_w) ranging between 3.7 and 7.7 and epicentral distances ranging from 1.7 km to 297 km. V_{S30} values were obtained from borehole log reports provided by BHRC. The eHVSR was computed using the geometric mean of 5% PSA. To develop the V_{S30} prediction model, the study employed a bilinear regression model, where V_{S30} was correlated with f_{peak} for frequencies above 1 Hz. For frequencies below 1 Hz, V_{S30} was set to 290 m/s. The R^2 value for the model was not provided, however the standard deviation of the model was reported to be 0.097 log10 units.

Stanko and Markušić (2020) [SM20] developed an empirical relationship between f_{peak}, bedrock depth, and V_{S30} for Croatia. They collected and analyzed over 300 microtremor measurements between 2014 and 2019 from various geological sites across Croatia. For V_{S30}, 62 active geophysical measurements were conducted using the multichannel analysis of surface

waves (MASW) and shear-wave seismic refraction. A trilinear regression model was developed between f_{peak} and V_{S30}. At high frequencies (f_{peak} greater than 30 Hz) the V_{S30} is a constant 600 m/s, at low frequencies (f_{peak} less than 1 Hz) the V_{S30} is a constant 200 m/s, the model is log-linear in between. The V_{S30} for flat HVSR curves (i.e., no clear f_{peak} present) was assumed constant at 800 m/s. The error in V_{S30} estimation was within 10–15%, with larger discrepancies observed at shallow rock sites where it was difficult to constrain the depth and stiffness of bedrock layers.

Although it was not the main focus of their research, Yong et al. (2021) [Y21] investigated the correlation between f_{peak} and V_{S30}. The study utilized data from 81 seismic stations that are part of the California Integrated Seismic Network (CISN). The dataset included between 126 and 458 horizontal component recordings (east–west and north–south) per station. The study does not propose an explicit predictive model linking f_{peak} to V_{S30} but does compare their trends. The study compiled f_{peak} values from mHVSR and eHVSR methods and found a weak correlation between the f_{peak} from both sources and V_{S30}, which suggested that f_{peak} alone may not be sufficient for accurately predicting V_{S30} without additional site parameters.

Hudson et al. (2021) conducted a detailed inter-method and inter-analyst study to evaluate how different mHVSR processing techniques and peak-selection approaches influence the correlation between f_{peak} and measured V_{S30} values at stations across California. Using one hour of ambient microtremor data recorded from 136 BH (high gain broadband) and 156 HH (high gain, high sample rate broadband) sensor sites with known V_{S30} values, the authors generated mHVSRs based on three spectral domains: FAS, Power Spectral Density (PSD), and 5%-damped PSA. Two analysts, Hudson [H21] and Palmer [P21], independently processed the data using different workflows; Hudson applied python library scipy (Virtanen et al., 2020) based peak picking, while Palmer used a MATLAB implementation of the Zhu et al. (2021) method for automated peak

selection. Their separate analyses yielded two distinct f_{peak}–V_{S30} correlations based on the HH data: H21 model tended to identify lower-frequency peaks, while P21 identified higher-frequency peaks more frequently, often resulting in a steeper regression slope. Despite the methodological differences, both models reproduced trends similar to those of M15 and Y20 reinforcing the predictive utility of f_{peak} in estimating V_{S30}. However, the study underscored that the choice of spectral representation (FAS being preferred) and the analyst's subjective or algorithmic selection of peaks substantially affect the robustness and transferability of f_{peak}–V_{S30} models, highlighting the need for standardized protocols in HVSR-based site classification.

Thornley et al. (2021) [T21] conducted a comprehensive study on the applicability of eHVSR and Standard Spectral Ratio (SSR) methods to estimate the V_{S30} in Anchorage, Alaska. The study used earthquake recordings from 35 strong-motion stations across Anchorage. The dataset included 1,727 three-component recordings from 95 earthquakes collected between 2004 and 2019. The earthquakes originated from various sources covering a range of magnitudes and epicentral distances. The V_{S30} estimates were obtained from 16 strong-motion stations, incorporating both surface measurements and one downhole velocity profile. Additional V_{S30} data from 19 borehole measurements and 22 microtremor-based estimates were used to enhance the study's applicability. T21 developed three models as a part of the study: correlating f_{peak} to V_{S30}, A_{peak} to V_{S30} and both parameters to V_{S30}. The reported standard deviations for each of these models are 0.10, 0.16 and 0.09 log10 units respectively. These relationships are more or less similar to the NGA-West2 models developed by Ghofrani and Atkinson (2014), although the f_{peak}-V_{S30} correlation slope is slightly larger for T21. The developed correlations are only applicable above 1 Hz.

The study by Gupta et al. (2021) [G21] focuses on the seismic site characterization and site response of Nirsa, India, using HVSR and Vs profiling techniques to estimate V_{S30}. The study surveyed 78 observation points across a 4 km² area. The HVSR method was employed to determine $f_{0,HVSR}$ and site amplification, while the multichannel simulation with one receiver (MSOR) technique was utilized at 16 locations for Rayleigh wave phase velocity dispersion analysis. The shear wave velocity models were obtained via a guided Monte Carlo inversion method. The study developed two empirical models between $f_{0,HVSR}$ and V_{S30}, as well as A_{peak} and V_{S30}, providing an equation for V_{S30} estimation in similar geological settings. The developed models have reported R^2 values of 0.63 and 0.40, respectively. While the study presents a comprehensive methodology for seismic site characterization, several limitations are apparent. Foremost, the accuracy of V_{S30} estimates is highly dependent on the validity of the inversion process and the assumed model parameters. The MSOR approach, while useful, is known to carry significant uncertainties due to its reliance on a single-receiver setup and assumptions such as a 1D horizontally layered medium and a fixed Poisson's ratio, which can introduce errors in the derived dispersion curves and subsequent velocity models. Also, the study does not provide a detailed validation of its findings against in-situ borehole V_{S30} measurements, which would have strengthened its approach. Furthermore, the model was developed based on the data acquired over a very small area, which might not generalize to other geologic settings.

Chen et al. (2022) [C22] conducted an extensive study in Taiwan to investigate the relationship between mHVSR measurements and V_{S30}. Their research aimed to develop a high-resolution V_{S30} map for Taiwan using microtremor measurements. The dataset consisted of measured and estimated V_{S30} values. To estimate V_{S30}, the study inverted the mHVSR under the Diffuse Field Assumption (DFA) (Sánchez-Sesma et al. 2011; García-Jerez et al., 2013), which

accounts for the contributions of Rayleigh, Love, and body waves in mHVSR (García-Jerez et al., 2016). The researchers performed joint inversion of mHVSR and dispersion curves at selected locations where 136 microtremor arrays had been deployed, to attempt to mitigate the non-uniqueness of HVSR-based inversions. They validated their inversion results using 816 V_{S30} measurements from Taiwan's engineering geologic database (EGDT), obtained from borehole logging and other geophysical surveys. They developed a linear regression model using data from 1,242 selected sites to predict V_{S30} based on the f_{peak}. This model yielded an R^2 value of 0.40, indicating a relatively weak correlation. To enhance their predictive performance, a new parameter, H_R, was introduced as a proxy for V_{S30}. The parameter H_R represents the ratio of the average mHVSR amplitude between high-frequency (HMH) and low-frequency (HML) bands, with a central frequency of 2.0 Hz used as the threshold in this study. Incorporating H_R into the model improved the R^2 value to 0.47. Further refinement was achieved by incorporating the stations elevation into the predictive model using modified elevation (E_m), which is computed as the maximum of the station's elevation and 5 meters. The final model, which correlated V_{S30} with both H_R and E_m, exhibited a significantly improved R^2 value of 0.71, indicating that topographic features play a critical role in V_{S30} estimation. While Chen et al. (2022) effectively reduced the non-uniqueness issue in mHVSR inversion by using microtremor array constraints, potential shortcomings include spatial bias as the 136 microtremor arrays used to constrain mHVSR inversion were performed over regional geologic conditions.

Chapter 3: A Set of Data-Driven Models to Predict V_{S30} from the Horizontal-to-Vertical Spectral Ratio of Microtremors

Abstract

We investigate the potential of using microtremor horizontal-to-vertical spectral ratio (mHVSR) measurements to predict the time-averaged shear-wave velocity in the upper 30 meters (V_{S30}) using data-driven models. We develop a dataset comprising 536 sites with 2,861 three-component ambient noise recordings from global regions, including New Zealand, Taiwan, Italy, Ecuador, Mexico and the United States. We consider two different types of predictive models: low-dimensional models which use features of the mHVSR mean curve such as the fundamental site frequency ($f_{0,HVSR}$) and peak amplitude ($A_{0,HVSR}$), as well as high-dimensional models which use the entire mHVSR mean curve to predict V_{S30}. In addition, we integrate topographic information from a 1 arc-second digital elevation model (DEM) for both model types using binned elevation as a proxy for geologic composition and topographic position index as a proxy for topography features. The low-dimensional models are shown to reasonably approximate V_{S30} when using mHVSR and topographical features (coefficient of determination, R^2, up to 0.69 on the testing set). The high-dimensional models, which use the full mHVSR mean curve, outperform the low-dimensional models (R^2 of 0.82 on the testing set) regardless of whether the mHVSR mean curve is supplemented by the topographical features. These findings demonstrate that while low-dimensional features such as $f_{0,HVSR}$ and A_{peak} are informative, leveraging the full shape of the HVSR curve significantly improves the ability to estimate V_{S30}. We compare the results of our data-driven models with those developed previously using remote sensing from sites in our testing set. Our top performing model substantively outperforms the current state of the art (R^2 of 0.56

versus 0.03). Finally, we apply our models to predict V_{S30} at 1,855 broadband recording stations across North America for use in future ground motion hazard studies.

Introduction

Local site conditions significantly influence seismic ground motions as they travel to the ground surface. These modifications, known as *site effects*, are a crucial piece in seismic hazard assessments because of their influence on the amplitude, duration, and frequency content of surface shaking. One approach to quantify site effects is through site response that numerically propagates earthquake shaking in rock through a model of the site to predict surface shaking. While the methods for performing site response are well understood and widely used, developing the necessary site-specific non-linear properties is not trivial. Instead, most ground motion models quantify site effects through regressions using site-specific proxies. The most widely used proxy for site effects is the time-averaged shear wave velocity of the upper 30 meters (V_{S30}; Borcherdt, 1994; Building Seismic Safety Council, 1997; European Committee for Standardization, 2004). While V_{S30} may not fully capture the complexity of site amplification (Castellaro et al., 2008; Seyhan and Stewart, 2013) it provides a tractable solution for predicting site effects. Furthermore, although V_{S30} is significantly easier to obtain than a site's full non-linear dynamic properties, it still requires specialized measurements. As a result, nearly 85% of the seismic stations in the combined NGA-West2 and NGA-East database lack a directly measured V_{S30} value (Seyhan et al., 2014; Goulet et al., 2021). Consequently, there is a pressing need for cost-effective and reliable techniques to estimate V_{S30} across diverse geological settings for use in improving predictions of seismic hazard. This paper's objective is to develop an approach to estimate V_{S30} from single-station ambient noise measurements. More specifically, this paper presents a set of models to

predict V_{S30} from site-specific microtremor horizontal-to-vertical spectral ratio (mHVSR) measurements.

Building upon prior efforts to relate the horizontal-to-vertical spectral ratio (HVSR) of earthquakes and microtremors to V_{S30}, this paper aims to develop a set of data-driven models using machine learning (ML) that can be used globally to relate mHVSR measurements to site-specific V_{S30}. The paper presents background on different approaches from the literature to estimate V_{S30} when site-specific measurements of shear-wave velocity (Vs) are not available. This is followed by a description of the data used in this study, including how the mHVSR were processed and prepared for model training. Two categories of models to predict V_{S30} from mHVSR are presented: low-dimensional models, which use scalar features of the mHVSR mean curve, and high-dimensional models, which use the full mHVSR mean curve directly. These models are compared with existing proxy-based approaches to determine V_{S30} using sites from the model's testing set where the V_{S30} has been measured independently. The models are then used to estimate V_{S30} for 1,855 broadband recording stations located across North America for use in future ground motion hazard studies.

Background

Several efforts have been made previously to estimate V_{S30} using qualitative or semi-quantitative proxies such as surface geology (e.g., rock vs. soil; Wills et al., 2000), standard penetration test (SPT) blow counts (Seed et al., 1976; Borcherdt, 1970), geologic age (Wills and Clahan, 2006), and topographic slope (Wald and Allen, 2007). While these methods offered practical alternatives, they are limited by weak correlations with V_{S30}. As will be discussed in detail later, recent efforts have been made to correlate estimates of a site's resonant frequency from

non-invasive HVSR measurements with V_{S30} (e.g., Ghofrani and Atkinson, 2014; Hassani and Atkinson, 2016; Stanko and Markušić, 2020).

The HVSR technique discovered by Nogoshi and Igarashi (1971) and later popularized by Nakamura (1989) has been used for understanding a site's seismic behavior. While the full informative potential of HVSR is a matter of debate, it is generally agreed that it is an effective tool for measuring a site's fundamental resonance frequency (f_0). HVSR measurements involve dividing the combined Fourier amplitude spectra of the horizontal ground motion components by the Fourier amplitude of the vertical component. HVSR can be computed from either microtremors (mHVSR) (Nakamura, 1989; SESAME, 2004) or earthquake recordings (eHVSR) (Lermo and Chávez-García, 1993; Field and Jacob, 1995), with mHVSR offering the practical advantage of not requiring earthquakes to occur during data acquisition. Multiple studies have shown that HVSR features hold promise for estimating the near-surface velocity structure, including V_{S30} (e.g., Ghofrani and Atkinson, 2014; Hassani and Atkinson, 2016).

Most existing models to predict V_{S30} from HVSR rely on only the HVSR curve's resonant frequency. Here we clarify that two different measures of a HVSR curve's resonant frequency have been defined, these are: $f_{0,HVSR}$ and f_{peak}. $f_{0,HVSR}$ is the fundamental (i.e., the lowest) resonant frequency as observed in the HVSR curve. In contrast, f_{peak} is the frequency of the peak of the HVSR curve with the highest amplitude. While the two may be similar in some circumstances (e.g., the HVSR curve has only a single peak) they are generally not the same. In the following discussion we use these terms consistent with their definitions provided here, even if the original authors used different terminology. Regardless of how the resonant frequency from HVSR is defined, studies have shown that the relationship between the resonant frequency and the site's

subsurface structure is not unique. For example, deep impedance contrasts can dominate the HVSR peak, masking the influence of the upper 30 meters (Castellaro and Mulargia, 2009; Parolai et al., 2002).

This limitation has led to most existing HVSR–V_{S30} correlations having been developed and validated within limited geological settings and/or over limited frequency ranges (e.g., Ghofrani and Atkinson, 2014; Hassani and Atkinson, 2016; Stanko and Markušić, 2020; Yaghmaei-Sabegh and Hassani, 2020).

Figure 4 summarizes existing correlations between HVSR resonant frequency (i.e., $f_{0,HVSR}$ or f_{peak}) and V_{S30} from the literature. The first such model was developed by Ghofrani and Atkinson (2014) using fpeak and the corresponding peak amplitude (A_{peak}) in their regression models developed using PSA-based eHVSR from the NGA-West2 dataset and the Japanese KiK-net and K-net datasets. McNamara et al. (2015) used ambient noise recordings from 21 sites in Central Virginia Seismic Zone (CVSZ) and derived a strong correlation between $f_{0,HVSR}$ and V_{S30} at 21 sites with clear peaks, though limited by regional scope and small sample size. Hassani and Atkinson (2016) utilized $f_{0,HVSR}$ from eHVSR in the NGA-East database to construct a bilinear regression model using data from 315 stations, offering lower uncertainty than geologic or topographic proxies, albeit for a limited number of stations with usable data. Yaghmaei-Sabegh and Hassani (2020) employed eHVSR derived from PSA at 560 stations in Iran, fixing V_{S30} at 290 m/s for f_{peak} below 1 Hz and using bilinear regression otherwise. Stanko and Markušić (2020) analyzed over 300 mHVSR recordings across Croatia and developed a trilinear model correlating f_{peak} to V_{S30}, assigning fixed values 800 m/s for flat, 600 m/s for f_{peak} above 30 Hz and 200 m/s for f_{peak} below 1 Hz. Yong et al. (2020) compiled f_{peak} values from both mHVSR and eHVSR at 81 California Integrated Seismic Network (CISN) stations in California and found weak correlation

with V_{S30}, concluding that f_{peak} alone is insufficient as a predictive parameter. Hudson et al. (2021) analyzed 156 broadband stations in California, producing two separate mHVSR–V_{S30} correlations (denoted as P21 and H21 in Figure 4) using different analysts and spectral representations, demonstrating that f_{peak}–V_{S30} relationships are sensitive to peak-picking methods and domain like FAS, PSA or Power Spectral Density (PSD). Thornley et al. (2021) developed three eHVSR-based models using data from 35 strong-motion stations in Anchorage, Alaska correlating f_{peak}, A_{peak}, and their combination to V_{S30} applicable only for f_{peak} above 1 Hz. Gupta et al. (2021) used HVSR and MSOR inversion at 78 sites in India to relate $f_{0,HVSR}$ and $A_{0,HVSR}$ (amplitude corresponding to $f_{0,HVSR}$) to V_{S30}, though model accuracy was limited by multichannel simulation with one receiver (MSOR) assumptions and small spatial coverage. Finally, Chen et al. (2022) conducted a high-resolution study in Taiwan and developed correlations to predict V_{S30} using data from 1,242 sites. Chen et al. (2022) introduced the ratio of average HVSR amplitudes in high- and low-frequency bands as an additional predictive parameter, which significantly improved predictive power when combined with another parameter related to ground elevation. These strategies reflect a growing recognition that incorporating multiple HVSR features, including the absence of peaks, is necessary to improve the robustness of V_{S30} estimation across diverse geologic settings. Table 1 presents a summary of existing predictive models, the size of the database underlying each study, the parameters used to develop each correlation, and the accuracy of each model in terms of R^2. Taken together, these findings underscore that the prediction of V_{S30} from HVSR may be able to be further improved by increasing the data and the number of features used to develop the predictive model.

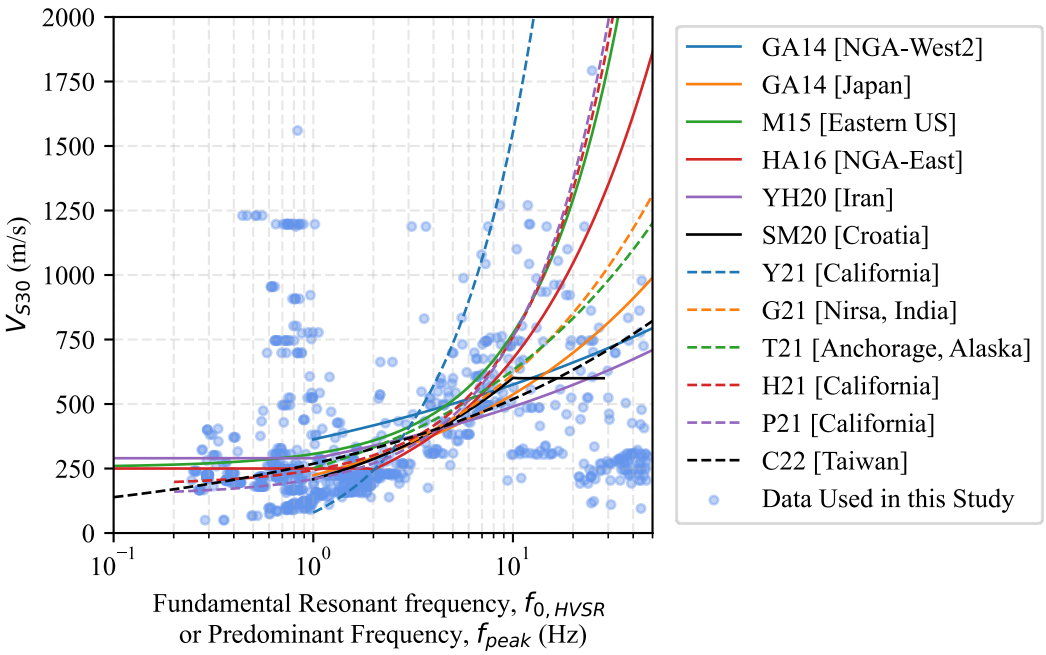

Figure 4:Comparison of existing models for predicting V_{S30} from HVSR resonant frequency ($f_{0,HVSR}$ or f_{peak}) relative to the dataset of 1,148 mHVSR with a clear resonance collected for this study. For each existing model the legend indicates the author(s), the year, and the region for which the model was developed. Additional details for each of the models shown is summarized in Table 1.

Table 1: Summary of existing models for predicting V_{S30} from HVSR. Each model is summarized in terms of study, literature reference, whether the model used eHVSR or mHVSR, the predictive feature(s), dataset size, and model accuracy in terms of R^2.

Study	Reference	eHVSR or mHVSR	Predictive Feature(s)	Dataset Size	R^2
GA14 [Japan] [1]	Ghofrani and Atkinson, 2014	eHVSR	f_{peak}	Not Reported	0.30
GA14 [Japan]	Ghofrani and Atkinson, 2014	eHVSR	A_{peak}	Not Reported	0.23
GA14 [Japan] [1]	Ghofrani and Atkinson, 2014	eHVSR	f_{peak}, A_{peak}	Not Reported	0.42
GA14 [NGA-West2] [1]	Ghofrani and Atkinson, 2014	eHVSR	f_{peak}	4080	0.10
GA14 [NGA-West2]	Ghofrani and Atkinson, 2014	eHVSR	A_{peak}	4080	0.23
GA14 [NGA-West2] [1]	Ghofrani and Atkinson, 2014	eHVSR	f_{peak}, A_{peak}	4080	0.30
M15 [EUS]	McNamara et al., 2015	eHVSR	$f_{0,HVSR}$	21	Not Reported
HA16 [NGA-East]	Hassani and Atkinson, 2016	eHVSR	$f_{0,HVSR}$	41	Not Reported
YH20 [Iran] [1]	Yaghmaei-Sabegh and Hassani, 2020	eHVSR	f_{peak}	560	Not Reported
SM20 [Croatia] [2]	Stanko and Markusic, 2020	mHVSR	f_{peak}	>300	Not Reported
Y21 [CA] [3]	Yong et al., 2020	eHVSR, mHVSR	f_{peak}	458	Not Reported
H21 [CA]	Hudson et. al, 2021	mHVSR	f_{peak}	156	Not Reported
P21 [CA]	Hudson et. al, 2021	mHVSR	f_{peak}	156	Not Reported
G21 [Nirsa, India] [4]	Gupta et al., 2021	mHVSR	$f_{0,HVSR}$	78	0.63
G21 [Nirsa, India] [4]	Gupta et al., 2021	mHVSR	$A_{0,HVSR}$	78	0.40
T21[Anchorage, Alaska] [1],[4]	Thornley et al., 2021	eHVSR	f_{peak}	35	Not Reported
T21[Anchorage, Alaska] [4]	Thornley et al., 2021	eHVSR	A_{peak}	35	Not Reported
T21[Anchorage Alaska] [1],[4]	Thornley et al., 2021	eHVSR	f_{peak}, A_{peak}	35	Not Reported

C22 [Taiwan] [4]	Chen et al., 2022	mHVSR	f_{peak}	1242	0.40
C22 [Taiwan] [4]	Chen et al., 2022	mHVSR	H_R	1242	0.47
C22 [Taiwan] [4]	Chen et al., 2022	mHVSR	H_R, E_m	1242	0.71

[1] Applicable only to $f_{peak} > 1$Hz

[2] Applicable only to $f_{peak} > 1$Hz and < 30 Hz.

[3] Includes data from multiple sources.

[4] Includes sites where V_{S30} values were estimated and not independently measured.

Data

A total of 2,861 high-quality three-component microtremor recordings from 536 distinct sites were used in this study. The microtremor recordings were aggregated from 23 previously published and 15 unpublished studies performed around the world, including Europe (Italy), Asia (Taiwan), Oceania (New Zealand), and North and South America (United States, Mexico, Ecuador). The datasets used in this study in terms of location, number of sites, number of recordings, and the associated reference are summarized in Table 2. The majority (91%) of the microtremor recordings were collected as part of microtremor array measurements (MAM), with the balance (9%) being single station recordings. The recording durations varied significantly across sites, ranging from 20 to 540 minutes. To develop the mHVSR–V_{S30} pairs to train the data-driven models, all microtremor measurements were processed identically and associated with a site-specific measurement of V_{S30}. The details of this process are discussed in the following paragraphs.

All the V_{S30} values used in this study were directly measured and were not inferred from existing correlations. The V_{S30} values in the dataset span from 50 to 2,250 m/s, which includes site classes A to E as defined by the American Society of Civil Engineers/Structural Engineering Institute (ASCE/SEI) 7-22. The distribution of V_{S30} values in terms of their implied site class according to ASCE 7-22 is shown in Figure 5. Most of the data (98%) were acquired using surface wave methods, such as Multichannel Analysis of Surface Waves (MASW) and Microtremor Array Method (MAM), with the balance being obtained through invasive PS suspension logging and downhole testing. Among the total, 79 sites include multiple V_{S30} estimates, these different estimates were either the result of multiple measurement techniques or different data interpretations. The majority of these 79 sites with multiple V_{S30} values are the result of different data interpretation, where different plausible interpretations of the experimental surface wave data

resulted in similar but not identical measures of V_{S30} (e.g., Yust et al., 2018). For these sites we have chosen to average the different V_{S30} values at each site and to neglect uncertainty in V_{S30}. We discuss the result of these modeling decisions following the presentation of the models' performance.

Table 2: Summary of the data used in this study, in terms of its location, number of sites, number of recordings, and reference(s). The number of recordings is equal to or larger than the number of sites because most studies included more than one ambient noise measurement. For some datasets more than one reference is provided.

Location	Number of Sites*	Number of Recordings	References
Italy	92	92	Felicetta et al., 2023
New Zealand (NZ)	111	111	Barker, 1996; Kaiser & Louie, 2006; Wotherspoon et al, 2015; McVerry, 2011; Wotherspoon et al., 2016; Deschenes et al., 2018; Cave 2020; Wotherspoon et al., 2023; Stolte et al, 2023; Hill et al., 2023
Christchurch, NZ	25	422	Teague et al., 2018
Wellington, NZ	23	212	Cox and Vantassel, 2018; Vantassel et al., 2018
Mississippi Embayment, USA	10	252	Wood and Himel, 2019
Mexico City, Mexico	25	156	Wood et al., 2020
Taiwan	45	45	Kuo et al., 2016
USA and Ecuador	32	219	Rahimi and Wood, 2022; Rahimi and Wood, 2023; Wood and Rahimi, 2022; Nikolaou et al., 2017
California, USA	161	493	Yong et al., 2013
Texas, USA	35	859	Yust et al., 2018

*Some sites are replicated across datasets. These duplicate sites have been removed to obtain the 536 distinct sites that were used in this study.

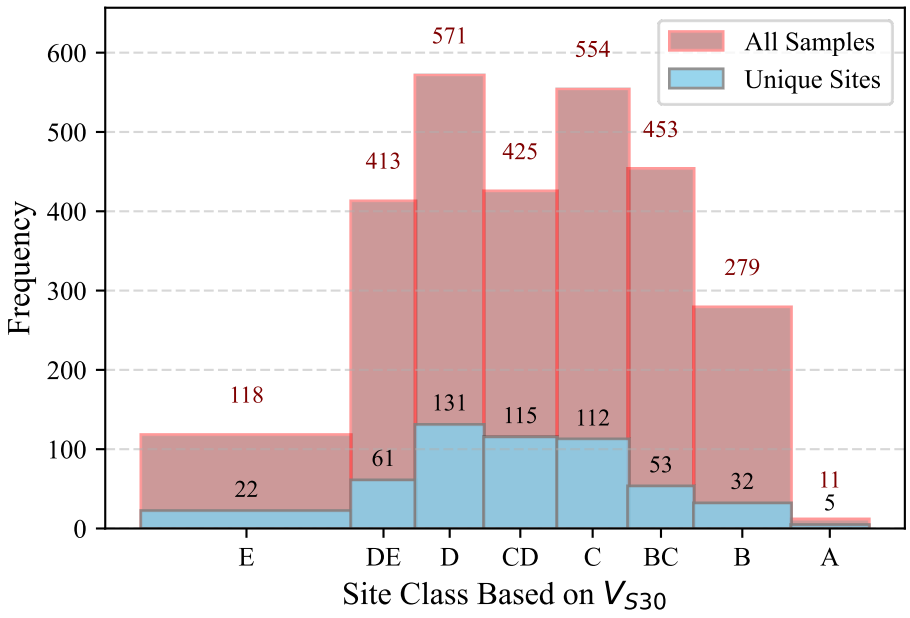

Figure 5: Distribution of data used in this study in terms of the V_{S30} -based Site Class as per ASCE/SEI 7-22, comparing all samples (shaded red) with unique sites (shaded blue) from the resampled mHVSR dataset. The red bars represent all available records, including repeated measurements at the same site, while the blue bars show the number of distinct sites. The x-axis is plotted on a logarithmic scale, and bar widths are defined by the actual V_{S30} range thresholds for each ASCE/SEI site class. Annotated values above each bar indicate the sample count within each class.

All the mHVSR data was processed identically following a semi-automated approach. To begin, each three-component microtremor recording was trimmed to equalize the duration of the three components using the ObsPy (Beyreuther et al., 2010) Python package. These trimmed recordings were then batch processed using the AutoHVSR algorithm (Vantassel et al., 2023). The details of the AutoHVSR algorithm are documented in Vantassel et al. (2023), but for the benefit of the reader we briefly summarize the associated processing steps as follows. The three-component ambient noise recordings were divided into 35 windows of equal length and each window was linearly detrended. A Tukey window (Harris, 1978) with a taper ratio of 20% (i.e., 10% per side) was applied to each window before the time series was transformed to the frequency domain. The horizontal components were combined using the geometric mean (SESAME, 2004). Konno and Ohmachi (1998) spectral smoothing with a bandwidth of 40 was applied at 256

logarithmically spaced frequencies between 0.05 Hz and 50 Hz to the merged horizontal and vertical components. Note that the frequency range was truncated according to the time domain window length such that each window contained at least 15 significant cycles. In cases where horizontal and vertical components had mismatched time steps or sampling rates, frequency-domain resampling was applied to align them before spectral calculations.

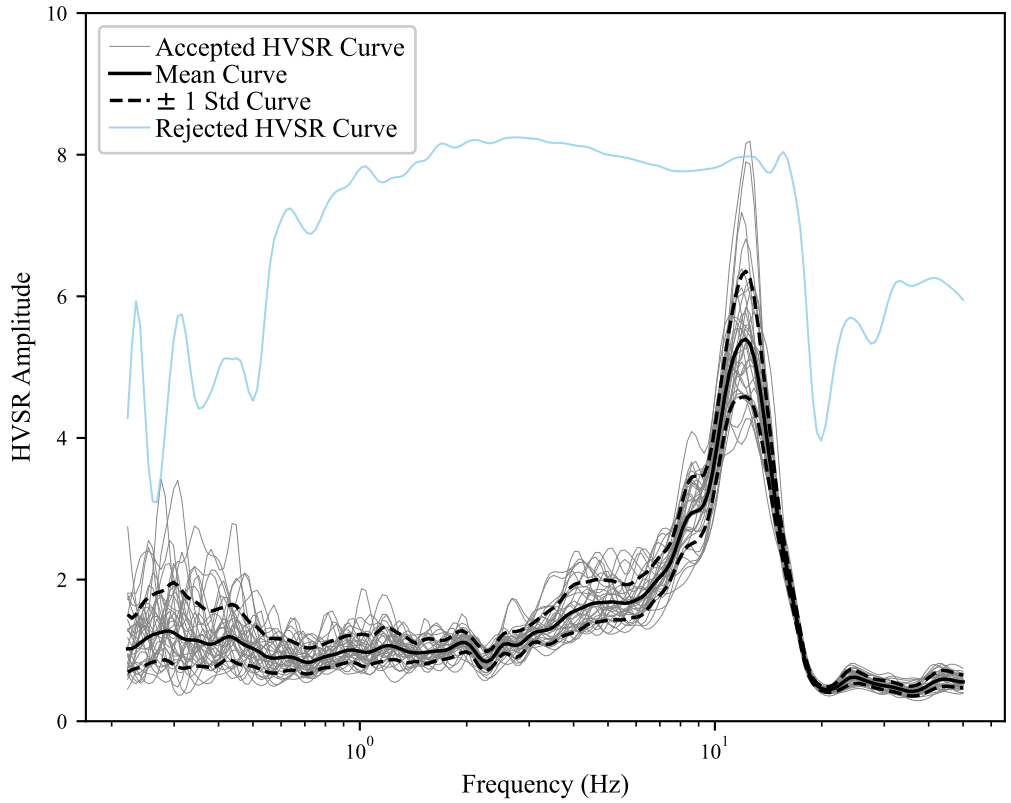

Figure 6: Example microtremor Horizontal-to-Vertical Spectral Ratio (mHVSR) after the application of manual window rejection. The accepted mHVSR from each time window is shown in gray, rejected mHVSR time windows in light color. The lognormal mean curve is shown with a solid black line and the one standard deviation range with dashed black lines.

Following initial automated processing using AutoHVSR, a manual review and window rejection step was implemented to enhance the stability and reliability of the computed mHVSR. Although this step is subjective and analyst dependent, this step is important to ensure the removal of outlier mHVSR that can bias statistical measures on the mHVSR resonant peaks and mean curve. Manual window rejection involved visual inspection and elimination of time windows in

the mHVSR domain that were substantially different from the other time windows recorded at the site. These time windows are likely contaminated by transient noise sources, such as vehicle noise or other human activity. These artifacts can severely distort the HVSR curve and lead to misidentification of the fundamental resonance frequency. An example of the manual window rejection process, highlighting both accepted and rejected windows, is shown in Figure 6.

To supplement the HVSR data, parameters obtained from surface topography were computed for each site. To account for terrain-related variability in subsurface conditions, topographic position index and elevation were computed for each mHVSR test location. The topographic position index (TPI) is defined as the difference between the site's point elevation and the mean elevation within a 1,500-meter radius, capturing the site's relative elevation compared to its surrounding landscape (Rai et al., 2017). Sites with positive TPI values, often associated with ridges or elevated features, are typically underlain by shallow or exposed bedrock, resulting in higher V_{S30} values. Conversely, negative TPI values, found in valleys or depressions, may indicate the presence of thicker alluvial deposits, often corresponding to lower V_{S30}. We also use elevation as an additional proxy. We acknowledge that the correlation between elevation and V_{S30} is less clear. However, as will be shown later, elevation improves the model's predictive ability. The authors hypothesize that this is because elevation serves as a generic proxy for regional geology (e.g., coastal vs. mountain range). TPI and elevation were determined using the 1 Arc-second Global DEM obtained from the USGS EarthExplorer platform (USGS, 2024) using the open-source software QGIS (QGIS.org, 2024). We note that topographic features have been used previously to predict V_{S30}. Wald and Allen (2007) introduced the use of topographic slope as a seismic site condition proxy, which was later validated and refined by Allen and Wald (2009). A modified elevation measure was used by Chen et al. (2022) in a predictive model along with HVSR

features. Geyin and Maurer (2023) combined remote sensing derived terrain attributes, including relative elevation, to predict V_{S30}. These studies support the use of topographic metrics in combination with mHVSR to predict V_{S30}. Our findings (presented later) demonstrate that the use of topographic features improve predictions of V_{S30} when used in combination with mHVSR.

Development of Predictive Models with Machine Learning

This study adopts a machine learning (ML)-based approach to explore the relationships between mHVSR and V_{S30}. Traditional regression models, such as those correlating the $f_{0,HVSR}$ or f_{peak} to V_{S30} (recall Figure 4), are often constrained by assuming a linear or log-linear relationship. In contrast, ML algorithms offer the flexibility to model non-linear relationships and heterogeneous feature types making them well-suited for capturing other patterns in the mHVSR. With the increasing availability of large mHVSR and V_{S30} datasets, ML provides a scalable framework capable of learning relationships that are difficult to formulate analytically. Once trained, these models are computationally efficient, allowing for the rapid prediction of V_{S30} from distributed microtremor measurements. In this study, a suite of predictive models was developed, using supervised learning, ranging from low-dimensional models using a limited number of features extracted from the HVSR mean curve, to high-dimensional models leveraging the entire HVSR mean curve.

Low-Dimensional Models

The low-dimensional models were developed using only those mHVSR with clear resonance peaks. A clear peak was defined by SESAME (2004) as a single dominant peak with an amplitude greater than 2, narrow bandwidth, minimal secondary peaks, and stability across multiple time windows after Konno-Ohmachi smoothing. The data available included 1,148 samples from 281 distinct sites. Each fundamental resonant peak was described in terms of its

frequency ($f_{0,HVSR}$) and amplitude ($A_{0,HVSR}$). The $f_{0,HVSR}$ versus V_{S30} for the data used for developing the low-dimensional models are shown relative to the existing V_{S30} predictive models in Figure 4. In addition, to the resonant frequency features, TPI and elevation were used to support generalization across varying terrain types. While TPI performed well as a feature, early models using station elevation strongly overfit to the training data. To mitigate this overfitting, elevation was binned into 500-meter intervals. As mentioned previously, while elevation can help segment broad geologic regions, they are not expected to reliably correlate with V_{S30}. Through the binning process we transform elevation into a coarse contextual feature, consistent with our expectation that is serves as a coarse proxy for geology. In addition, we explored various other features extracted from the mHVSR curve using feature engineering. However, for the sake of brevity we will only present our final selection, spectral skewness. Spectral skewness computed from the mHVSR curve was used to quantify the asymmetry of the HVSR curve around the fundamental frequency ($f_{0,HVSR}$). This metric reflects how the amplitude spectrum is distributed across frequencies and can be linked to subsurface conditions. Sites with low V_{S30} (softer soils) tend to produce sharper, more symmetric peaks, while stiffer sites with higher V_{S30} often show broader or skewed spectral shapes, resulting from more complex layering or impedance transitions. The features used for the low-dimensional models are summarized in Table 3.

The dataset was split 80 / 20 into training and testing sets, respectively, using stratified sampling based on ASCE/SEI 7-22 site class, ensuring balanced representation of V_{S30} categories across sets. To prevent information leakage, the testing set was not used until the model was finalized.

To improve model stability and meet the assumptions of linear modeling several feature transformations were applied. $A_{0,HVSR}$ was log-transformed to reduce skewness and stabilize

variance, while $f_{0,HVSR}$ underwent a Box-Cox transformation (Box and Cox, 1964), appropriate for strictly positive, non-Gaussian data. The Box-Cox transform ensured that the residuals from linear regression were normally distributed. Features such as spectral skewness and TPI were standardized via z-score normalization to ensure proportional influence among features. The target variable, V_{S30}, was log-transformed. Final predictions were back transformed using the exponential function to recover values in m/s. For the tree-based, feature transformations were not applied as these models are robust to unscaled and non-Gaussian inputs. Nevertheless, the log-transformation of V_{S30} was retained across all model types to ensure uniformity in all model predictions.

Table 3: Features used for low-dimensional modeling in terms of their name, description, and range.

Name	Description	Range
$f_{0,HVSR}$	Fundamental resonant frequency from HVSR	0.25 - 47.56
$A_{0,HVSR}$	Amplitude of the resonant frequency	2.13 - 22.91
Binned Elevation	Elevation obtained from 1 arcsec DEM binned into 500m interval	-71 - 2690*
Topographic Position Index	Relative elevation of the point compared to average elevation of 1500m radius	-165.22 - 153.80
Spectral Skewness	Quantity of asymmetry of HVSR curve around f0,HVSR	-0.46 - 4.68

*Provided range is the actual range of elevation, although it was binned for training.

The simplest low-dimensional model considered was a multivariate linear regression (MLR), trained using five-fold stratified cross-validation based on ASCE/SEI 7-22 site class. The use of stratification for the five-fold cross validation for the grid search helps maintain the proportion of each class across folds, which is essential for consistent model development. Model performance was evaluated using coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), with metrics averaged across all five folds. The MLR model achieved an

R^2 of 0.49, RMSE of 0.42 ln units, and MAE of 0.32 ln units on the testing set indicating moderate explanatory power given the feature set and non-linear nature of the underlying relationships.

To better capture potential non-linear relationships, three tree-based models were implemented: Decision Tree (DT; Quinlan, 1986), Random Forest (RF; Breiman, 2001), and Gradient Boosted Trees (GBT; Friedman, 2001), with the latter trained using the extreme gradient boosting (XGBoost) framework (Chen and Guestrin, 2016) for enhanced speed and regularization. These models were tuned using a grid search over specified hyperparameter ranges with each hyperparameter combination being evaluated using five-fold stratified cross-validation based on ASCE/SEI 7-22 site class. The selected hyperparameters for each model are summarized in Table 4, while

Table 5 presents the comparative performance metrics across all models on the testing set. The tree-based methods are shown to consistently outperform MLR, highlighting the benefits of non-linear modeling techniques when working with HVSR data for V_{S30} prediction.

For mHVSRs lacking clear resonance peaks, a low-dimensional model could not be developed, as topographic features alone do not exhibit a direct physical relationship with V_{S30}. As evidence of this Figure 7, shows the distribution of V_{S30} values from the 1,713 mHVSR measurements without a clear peak. The V_{S30} values shown in Figure 7, range from 50 to 2,250 m/s preventing any simplified approach for reliably inferring V_{S30} from a mHVSR curve using the fact that it does not have a clear peak as the sole predictive feature. For use in practice, the reader is referred to the high-dimensional models developed in this study as an alternative.

Table 4: Considered and selected hyperparameters for the decision tree (DT), random forest (RF) and extreme gradient boosted trees (GBT) algorithms.

Model	Hyperparameter	Hyperparameter Range	Selected Hyperparameter
MLR	L2 regularization	0.0, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000	0

DT	Tree Depth	2,4,6,10	4
	Minimum Samples Required to Split an Internal Node	3,10,30,100	10
	Minimum Samples Required at a Leaf Node	3,10,20,30	20
	Maximum Number of Leaf Nodes	3, 10,20,50,100	20
	Pruning Parameter (Cost-Complexity Alpha)	0.0, 0.01, 0.05, 0.1, 0.3	0.01
RF	Number of Trees	10, 30, 50, 100	50
	Tree Depth	3, 6, 8, 10	6
	Minimum Samples Required to Split an Internal Node	3, 5, 10	5
	Minimum Samples Required at a Leaf Node	3, 5, 10	3
	Pruning Parameter (Cost-Complexity Alpha)	0.01, 0.1, 0.3, 1, 3, 10	3
GBT	Number of Boosting Rounds (Trees)	10, 30, 50, 70, 100, 150, 250	100
	Learning Rate	0.001, 0.003, 0.01, 0.03, 0.1, 0.3	0.03
	Maximum Tree Depth	3, 5, 9, 12	5
	Minimum Sum of Instance Weights in Child Nodes	3, 5, 10	3
	Row Sampling Rate per Tree	0.5, 0.7, 0.9	0.7
	Feature Sampling Rate per Tree	0.5, 0.7, 0.9	0.9
	L1 regularization	0.0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0	10
	L2 regularization	0.1, 0.3, 1.0, 3.0, 5.0, 10	5

Table 5: Performance comparison of the selected model across the unbiased testing set in terms of R^2, RMSE and MAE. The prediction errors are in natural log space.

Model	**R^2 (ln)**	**RMSE (ln m/s)**	**MAE (ln m/s)**
MLR	0.49	0.42	0.32
DT	0.57	0.38	0.25
RF	0.69	0.32	0.21
GBT	0.68	0.32	0.21

Single mode ANN	0.82	0.27	0.18
Dual mode ANN	0.83	0.27	0.17

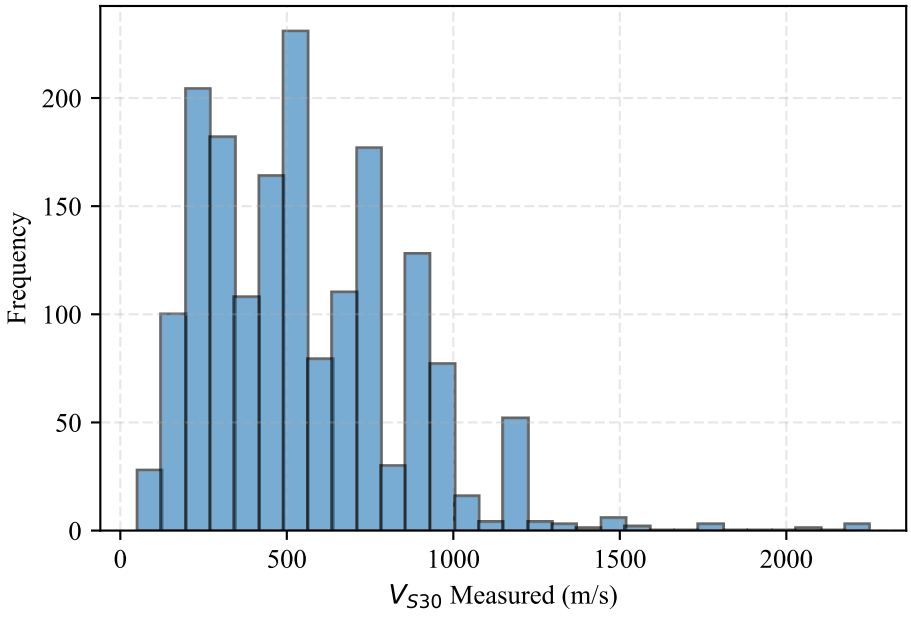

Figure 7: Histogram for the measured V_{S30} of 1713 samples (61% of total samples), which did not have clear resonance peak in the mHVSR mean curve.

High-dimensional Models

The development of the high-dimensional models used a subset of the dataset which consisted of sites with mHVSR measurements in the frequency range of 0.3 to 50 Hz, independent of whether a clear resonance peak was present. The dataset included 2,824 samples collected from 531 unique sites. The lognormal mean mHVSR curve was downsampled to 35 logarithmically spaced frequencies between 0.3 Hz and 50 Hz. Reducing the number of mHVSR input frequencies substantially reduced the size of the artificial neural network (ANN) models (Rosenblatt, 1958) accelerating the model development process while maintaining the general shape of the mHVSR curve. Consistent with what was done for the low-dimensional models the target variable, V_{S30}, was log-transformed. To ensure balance between the training, validation, and testing sets stratification was performed on both V_{S30} site class and the presence/absent of a clear resonant

peak. This dual stratification resulted in a very small number of samples in the Site Class A and B bins, therefore we chose to combine these bins for stratification. We developed two high-dimensional models: one using just the mHVSR mean curve (referred to as the single mode ANN model, hereafter) and the other with both the mHVSR mean curve and the topographic features TPI and binned elevation (referred to as the dual mode ANN model, hereafter).

To ensure stable and effective training, all mHVSR features (i.e., amplitudes) were log-transformed. The single mode ANN model was implemented using a sequential architecture that began with an input layer with 35 neurons, one per mHVSR frequency. Each hidden layer consisted of a dense layer with He normal initialization (He et al., 2015) and L2 regularization (L2 = 0.001) to reduce overfitting (Ng, 2004). Batch normalization (Ioffe and Szegedy, 2015), applied before the activation function, was used to stabilize and accelerate training. The ReLU (rectified linear unit) activation function (Nair and Hinton, 2010) was used after each batch-normalized layer to introduce non-linearity. To encourage generalization, dropout (Srivastava et al., 2014) was applied after each activation function, randomly deactivating a given percentage of neurons during training. The output layer consisted of a single neuron with a linear activation function for predicting log-transformed V_{S30}. The model was trained using the Adam optimizer (Kingma and Ba, 2014), which combines adaptive learning rates with momentum for greater efficiency. Training used a batch size of 32 for up to 200 epochs, with early stopping (patience = 15) and learning rate reduction (factor = 0.5) triggered by validation loss, with a lower learning rate bound of 1×10^{-6}. A five-fold cross-validation strategy was used to evaluate model generalization. Evaluation metrics of R^2, RMSE, and MAE were averaged across folds. A hyperparameter tuning grid was used to select the optimal architecture as summarized in Table 6.

Table 6: Architecture selection and hyperparameters tuning with grid search for the high-dimensional models.

Model	Hyperparameter	Hyperparameter Range	Selected Hyperparameter
Single mode ANN	Architecture	[64, 128, 64, 32, 1], [128, 256, 128, 64, 1], [32, 64, 32, 1]	[128, 256, 128, 64, 1]
	Activation function	ReLU, LeakyReLU	ReLU
	Dropout	0.2, 0.3, 0.4	0.3
	Learning Rate	0.001, 0.005, 0.01	0.005
	Epochs	150, 200	200
	Batch size	16, 32	32
Dual mode ANN	HVSR architecture	[64, 128, 64], [128, 256, 128], [16, 32, 16]	[128, 256, 128]
	Topographic data architecture	[128, 1], [128, 64, 1]	[128, 1]
	Learning Rate	0.001, 0.005, 0.01	0.001
	Dropout	0.2, 0.3, 0.4	0.4
	Epochs	150, 200	200
	Batch size	16, 32, 64	64

The dual mode ANN was developed to further explore the distinct contributions of spectral and topographical features. This model consisted of two branches: one processing the 35 resampled HVSR amplitudes, and the other handling topographic index and elevation bin encodings. mHVSR features were log transformed as before while the topographic features i.e. TPI and encoded elevation bins were z-score normalized .Each branch followed a similar design of dense layers with He initialization and L2 regularization, batch normalization, and ReLU activation, with dropout applied in the HVSR branch. The two streams were then concatenated and passed through a shared dense layer before outputting the log-transformed V_{S30} via a linear activation function. This architecture enabled the model to jointly learn from the shape of the mHVSR mean curve and the contextual topographic information. Training procedures, optimizer settings, and callbacks for the multi-input model were consistent with those used in the single mode ANN model. The architecture along with the hyperparameter tuning performed with the grid search is summarized in Table 6. As with other models, testing set metrics are included in the performance comparison summary (

Table 5).

Results and Discussion

The performance of the six models developed for predicting V_{S30} from mHVSR is presented in Figure 8 for the testing set. The high-dimensional models performed significantly better than the low-dimensional models. The dual mode ANN model (R^2 = 0.83) performed essentially the same as the single mode ANN model (R^2 = 0.82) on the testing set. Of the low-

dimensional models Random Forest ($R^2 = 0.69$) and Gradient Boosted Trees ($R^2 = 0.68$) performed nearly identically followed by Decision Tree ($R^2 = 0.57$) and lastly Multivariate Linear Regression ($R^2 = 0.49$). These results highlight the strength of neural networks in capturing complex nonlinear relationships between mHVSR curve and V_{S30}. The minimal performance gain of the dual mode ANN model suggests that topographic features of TPI and binned elevation, intended to represent terrain effects (e.g., ridge, valley, plain) and general geologic setting, are likely already encoded in the shape and amplitude of the mHVSR curve.

Although the single mode ANN and dual mode ANN models demonstrated strong performance across the full V_{S30} range, caution is advised when interpreting very high V_{S30} predictions. For example, refer to the dual mode ANN model prediction that exceeds 2,000 m/s in Figure 8. The dataset used for training contained fewer than 11 samples with a V_{S30} above 1,500 m/s (Figure 5), with the maximum V_{S30} across all measurements being 2,250 m/s. As a result, the model's exposure to high-velocity conditions, typically associated with hard, shallow bedrock is limited, resulting in potential for overconfident extrapolation by the models in this range. To avoid misinterpretation, we recommend caution when model predictions exceed 1,500 m/s, consistent with the majority of training data. Predictions beyond this threshold should be treated as highly uncertain, and users are encouraged to carefully consider the validity of these values based on local geology, remote-sensing based predictions, and if necessary, supplement with direct geophysical measurements. This is particularly important in engineering design or seismic hazard analysis, where overestimation of site stiffness could result in unconservative decisions.

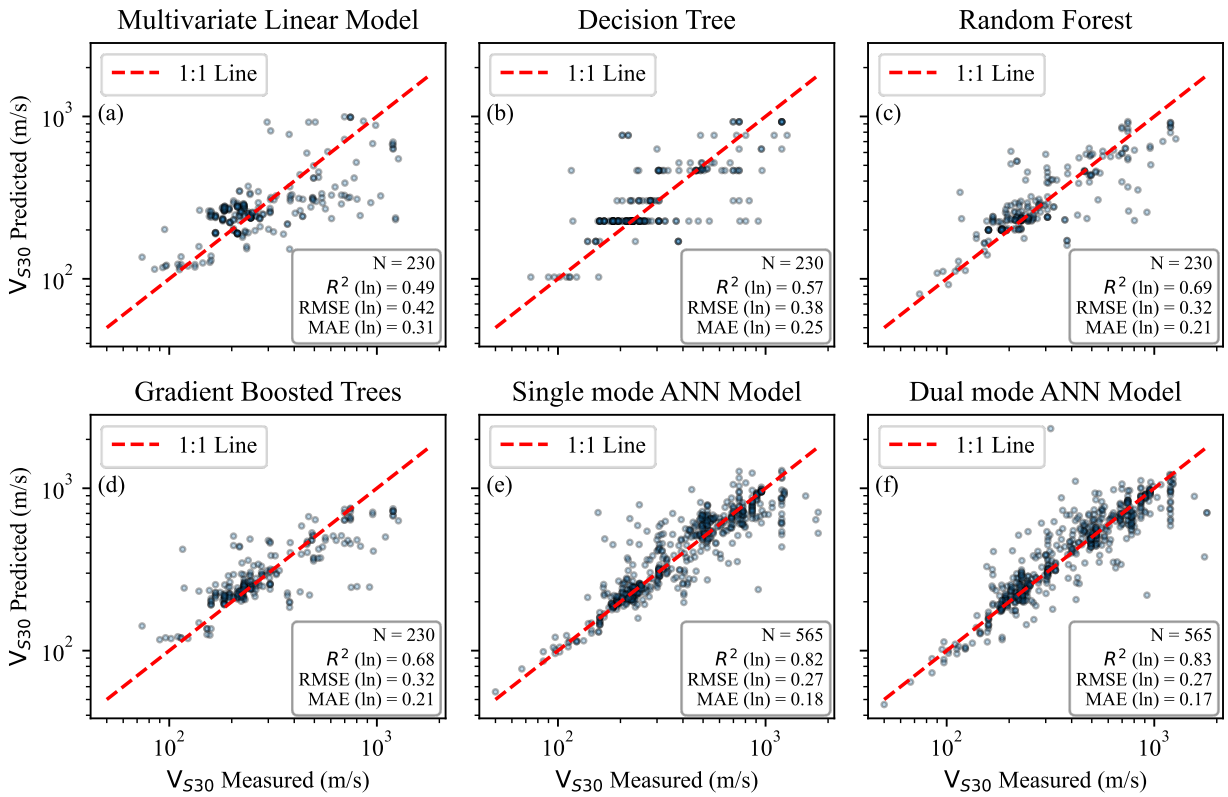

Figure 8: Comparison of V_{S30} prediction performance for the six models on the testing dataset. The models include (a) Multivariate Linear Regression, (b) Decision Tree, (c) Random Forest, (d) Gradient Boosted Trees, (e) Single mode Artificial Neural Network (ANN), and (e) Dual mode ANN. Each panel shows the predicted V_{S30} versus the measured V_{S30} on a log-log scale. The red dashed line indicating the 1:1 reference. Each panel include the number of tested samples (N) and the models R^2 , RMSE and MAE values (in natural log space) as a quantitative measure of predictive accuracy.

To compare our results against the state of the art for V_{S30} prediction, we compare our single mode ANN model's prediction of V_{S30} for the testing set with those from the remote-sensing model developed by Geyin and Maurer (2023). We compare the models at 356 stations from the testing dataset, note that these 356 are a subset of the 565 stations in our testing data as they were located in the contiguous United States (i.e., where the Geyin and Maurer (2023) has been developed). As shown in Figure 9, the single mode ANN model retained reasonable accuracy with an $R^2 = 0.56$ on this subset of the testing set. While the accuracy is on this subset is, somewhat reduced from that of the full testing set it is still considered acceptable. In contrast, the Geyin and Maurer (2023) model yielded a strikingly low $R^2 = 0.03$. This underperformance is visually

corroborated by the lack of clustering around the 1:1 line and the wide scatter across the V_{S30} range. The density of points in Figure 9 shows that many predictions are systematically biased from the true V_{S30} values, particularly for sites with V_{S30} > 600 m/s. This observation is further supported by the residual (Figure 10), which reveals a wide and irregular distribution of errors for the Geyin and Maurer (2023) model. Unlike the single mode ANN model, whose residuals are tightly clustered around zero, the Geyin and Maurer (2023) model shows substantial spread and multiple peaks, indicating inconsistent predictions and frequent deviations from actual V_{S30} values. One potential explanation of the underperformance of the Geyin and Maurer (2023) model is that it was trained on geospatial predictors derived from coarse-resolution remote sensing data (~90–1,000 m), such as slope, terrain ruggedness, and generalized geology. While these features correlate with V_{S30} at large scales, they may fail to capture local subsurface variability, particularly in geologically complex regions. Additionally, their model training data was biased toward flat, soft-soil sites, with limited representation of steep slopes or hard-rock conditions. This reduces its ability to generalize beyond the dominant terrain types in the training set. In contrast, our mHVSR-based model leverages site-specific measurements This highlights a key limitation of relying solely on surface indicators, such as slope, landform, or roughness, when estimating V_{S30}, particularly in areas where local subsurface conditions vary rapidly, or soil-to-bedrock transitions are shallow and abrupt. As Geyin and Maurer (2023) note, while such geospatial predictors are useful at large scales, they lack mechanistic links to V_{S30} and may miss fine-scale heterogeneity, especially in data-sparse or topographically complex regions.

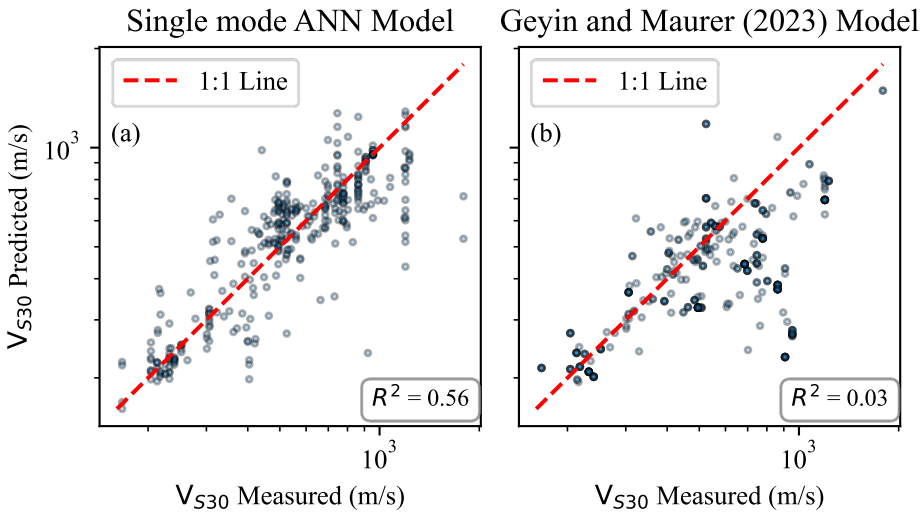

Figure 9: Comparison of V_{S30} prediction performance for (a) the mHVSR-based single mode ANN model and (b) the remote-sensing-based model developed by Geyin and Maurer (2023), using a subset of 356 stations within the contiguous United States. The dashed red line denotes the 1:1 line representing perfect prediction.

While the mHVSR-based models show promising accuracy and outperform remote-sensing-based approaches in predictive power, one limitation of this study is the inability to quantify prediction uncertainty. We could not rigorously quantify the prediction uncertainty in our models due to limitations in the availability and reliability of V_{S30} measurement error across the dataset. Out of 2,861 total recordings, only 979 (about one-third) had standard error values derived from multiple independent V_{S30} estimates. The majority (58%) of these estimates often came from alternative interpretations of the same experimental data and the balance (42%) from different measurement techniques. While these 979 samples allowed for direct computation the standard error in V_{S30}, the remaining two-thirds of the data required assuming a fixed uncertainty value. While estimating some reasonable upper bound for Vs30 standard error was possible (we estimated it at approximately 0.14 ln units), when we compared it to the model errors (approximately 0.39 ln units), it contributed a very small portion to the overall model's prediction uncertainty when assuming the measurement and model errors are uncorrelated (approximately 0.4 ln units).

Furthermore, we acknowledge that even our estimates of measurement error, could be uncertain, as these do not reflect the true uncertainty in V_{S30} estimates that may result from factors like equipment, processing workflows, or local heterogeneity. Consequently, as the measurement errors are only partially representative of the actual data quality and the measurement error we do have contribute a relatively small amount to the overall model prediction uncertainty, we concluded that we cannot confidently compute the total uncertainty in predictions with statistical rigor.

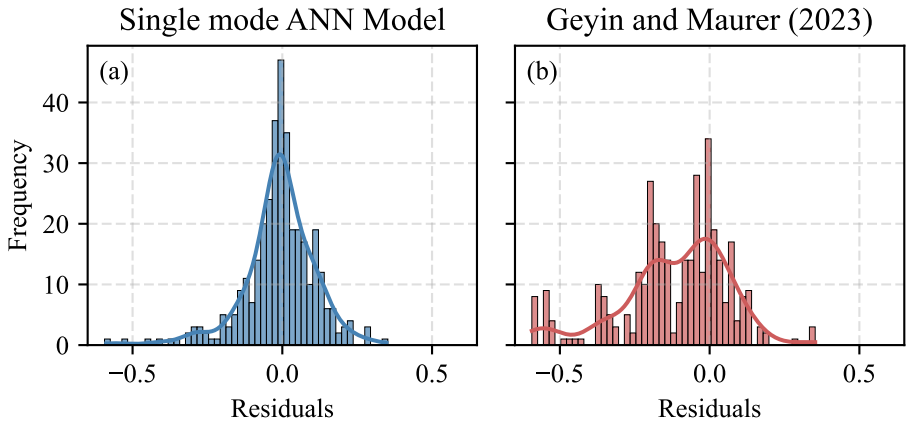

Figure 10: Histogram with Kernel Density Estimate (KDE) of residuals in log10 space for (a) the single mode ANN model and (b) the Geyin and Maurer (2023) model, computed as log10(Predicted) - log10(Measured).

Applications of the Predictive Model

To illustrate the practical application of the developed V_{S30} prediction models, we extended their use to a large-scale, real-world scenario. Specifically, we applied our models to estimate V_{S30} at 1,855 broadband seismic stations located across North America, with the majority situated within the contiguous United States. These stations are part of various regional and national networks and are equipped to record ambient microtremors, making them ideal candidates for HVSR-based V_{S30} estimation. The data were accessed through the IRIS-DMC and processed using the same process as described above in the Data section.

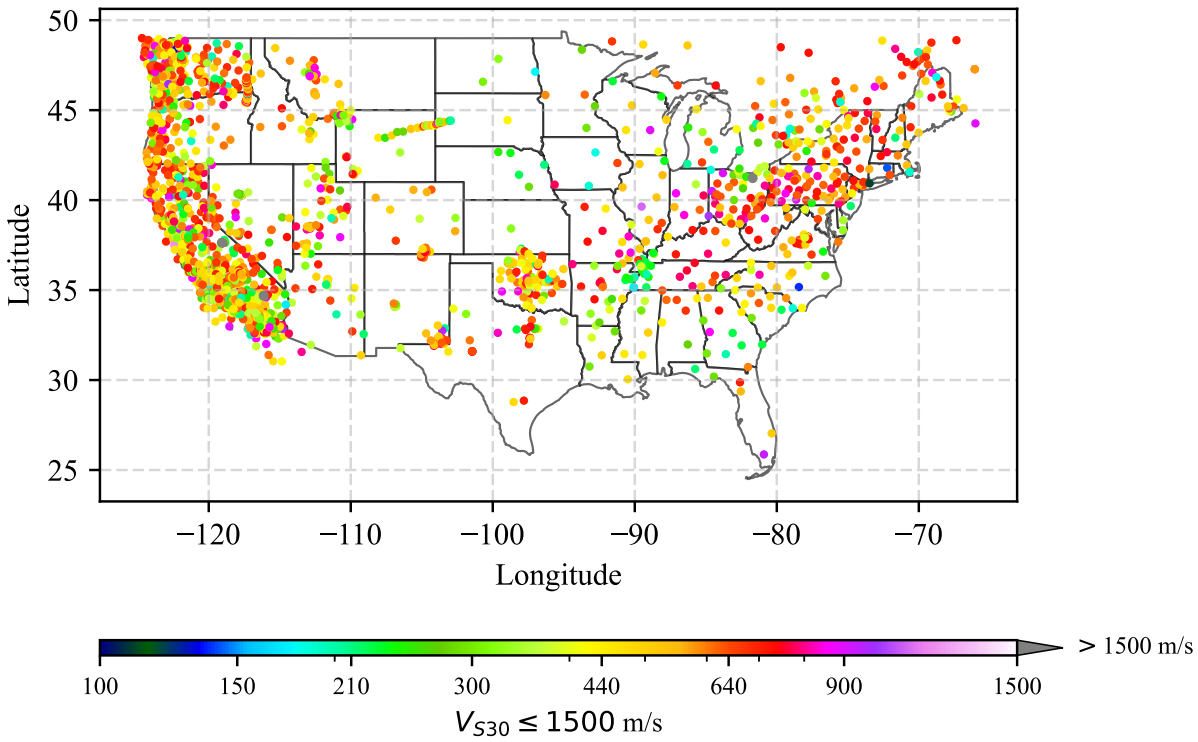

Figure 11: Spatial distribution of 1,855 broadband seismic stations across North America (primarily the contiguous United States) where V_{S30} was predicted using the proposed HVSR-based model. These stations are capable of recording ambient microtremors, enabling application of the method to expand seismic site characterization across regions with sparse direct measurements. Stations with predicted $V_{S30} \leq 1500$ m/s are shown using a continuous colormap as in the color bar below the map, while stations exceeding 1500 m/s are in gray, as pointed by the gray triangle at the right end of color bar.

Figure 11 displays the spatial distribution of these broadband stations, highlighting their coverage across diverse geological and topographic regions. Each point represents a station where V_{S30} was predicted using our models, which combines use of peak resonance features to full spectrum mHVSR features and site-specific metadata. These predictions enable enhanced seismic site characterization across regions where direct geophysical measurements are sparse or unavailable. The V_{S30} predicted with different models is provided in the supplemental material (Table S1). As mentioned in the discussion section, a level of discretion is advised to the users regarding some very high predicted V_{S30} values. This demonstration underscores the scalability of

the proposed framework, which is not only capable of predicting V_{S30} with high accuracy but is also readily transferable to dense networks of seismic stations globally.

Conclusion

This study presents a machine learning-based framework for predicting V_{S30} from mHVSR and topographic features. A comprehensive global dataset was compiled from diverse geological settings. This dataset was used to develop six machine learning models at different levels of complexity including four low-dimensional and two high-dimensional models. The low-dimensional models including multivariate linear regression, decision trees, random forest, and gradient boosted trees. These models resulted in predictions with R2 values of 0.49, 0.57, 0.69 and 0.68 on the testing dataset respectively. The high-dimensional models included a single mode artificial neural network (ANN) that took the mHVSR mean curve as input as well as a dual mode ANN that took the mHVSR mean curve as well as the topographic features of TPI and binned elevation as input. These high-dimensional models achieved R^2 values of 0.82 and 0.83, respectively in ln-transformed space. In short, the random forest and gradient boosted tree models performed nearly identically for the low-dimensional models, and both are recommended for use. The single mode ANN and dual mode ANN also performed nearly identically with the single mode ANN being preferred because of its simple structure. Overall, the single mode ANN is the recommended model for use by the earthquake hazard community, nonetheless, we have made all of the models we developed publicly available.

We compare the single mode ANN model's predictions against the current state-of-the-art remote-sensing-based model developed by Geyin and Maurer (2023). We compare the models at 356 stations in the United States from the single mode ANNs testing set. The single mode ANN model maintained reasonable accuracy ($R^2 = 0.56$), however the Geyin and Maurer (2023) model

performed poorly ($R^2 = 0.03$). This likely due to the Geyin and Maurer (2023) model's reliance on surface proxies and a limited ability to capture dense spatial variability using relatively course remote sensing data. Residual analyses revealed that while the single mode ANN model predictions were unbiased, the Geyin and Maurer model exhibited large biases particularly at sites with high values of V_{S30} (i.e., above 600 m/s).

Despite these promising results, we acknowledge two important limitations. First, the low-dimensional models developed in this study are only applicable to sites with at least one clear resonance peak. For sites lacking such peaks (e.g., those with flat or irregular mHVSR curves), we recommend the high-dimensional models, specifically the single mode ANN, which utilizes the full mHVSR curve between 0.3–50 Hz. However, to use high-dimensional models, user must ensure that their ambient noise recording is at least 30 minutes. Second, we were unable to rigorously quantify prediction uncertainty due to the limited availability of rigorous V_{S30} measurement error quantification. Only one-third of the datasets used in this study included standard error estimates derived from multiple V_{S30} measurements, with the majority of those standard errors being very small (less than 0.1 ln units) and substantially less than the model error. While we acknowledge that the standard error between V_{S30} measurement methods does not rigorously account for V_{S30} uncertainty, we do not have more rigorous measures available. In the future, it may be possible to obtain those measures using an uncertainty-consistent procedure, such as that proposed by Vantassel and Cox (2025) for surface wave measurements. Nonetheless, should V_{S30} be better quantified (i.e., improvement made to the quantification of measurement error) and more accurate V_{S30} predictions from mHVSR become possible (i.e., reduction in model error) in the future, these uncertainty considerations would be beneficial to the earthquake hazard prediction community.

Chapter 4: Conclusion

This thesis presents a comprehensive, data-driven approach for predicting the time-averaged shear wave velocity in the top 30 meters (V_{S30}) using microtremor-based horizontal-to-vertical spectral ratio (mHVSR) measurements. By leveraging a globally compiled dataset and integrating machine learning techniques, including both low- and high-dimensional models, this study advances current capabilities for seismic site characterization. This research demonstrates the utility of ambient noise recordings and publicly available topographic data in predicting V_{S30}. The developed models offer scalable alternatives that can be applied in both data-rich and data-poor regions, helping bridge gaps in seismic hazard assessment and site classification across the globe.

A key outcome of this work is the development of multiple predictive models tailored for different data conditions. The low-dimensional models rely solely on scalar features like the predominant frequency and amplitude of the mHVSR mean curve and are suitable for sites with well-defined resonance. However, these models cannot be used for sites lacking clear mHVSR peaks. To address this, high-dimensional models were developed using the full mHVSR mean curve and auxiliary features such as binned elevation and topographic index. These models show strong generalization capabilities across diverse geological conditions, including those with and without distinct resonance, making them applicable to a wider range of site conditions.

The engineering significance of this research lies in its direct applicability to seismic design, urban planning, and infrastructure resilience. V_{S30} is a key proxy for site conditions and is used in nearly all existing ground motion prediction equations (GMPEs), soil amplification studies, and seismic microzonation. In regions without direct V_{S30} measurements, especially in developing countries or remote areas, this framework allows rapid, non-invasive, and cost-effective estimation of site conditions. It enhances the capacity to conduct regional seismic hazard assessments, which

are crucial for updating building codes and improving earthquake resilience of lifelines and infrastructure.

Moreover, this research contributes to global consistency in V_{S30} estimation. While many prior studies focused on regional models fine-tuned to local geological settings, this work uniquely consolidates data across continents, offering a generalized model capable of robust cross-regional application. This is particularly valuable for large-scale hazard models like those used by the USGS or GEM (Global Earthquake Model), which require consistent and reproducible input parameters across different regions.

In addition, the integration of machine learning into geophysical workflows introduces opportunities for automation, enabling real-time or near-real-time V_{S30} mapping as new seismic data becomes available. This can support emergency response efforts following earthquakes by rapidly updating ground motion maps, guiding rescue operations, and informing post-event engineering assessments. The use of mHVSR, which can be collected with simple three-component sensors and does not require active sources, further democratizes access to subsurface site condition assessment.

In summary, this thesis not only develops scientifically robust models for V_{S30} prediction but also sets the foundation for its practical implementation in engineering, hazard mitigation, and policymaking. It demonstrates how modern data-driven techniques can complement traditional methods, making seismic site classification more accessible, scalable, and globally consistent.

Bibliography

1. Abrahamson, N. A., Silva, W. J., and Kamai, R. (2014). Summary of the ASK14 ground motion relation for active crustal regions. *Earthquake Spectra*, *30*(3), 1025-1055.
2. Ahn, J. K., Kwak, D. Y., & Kim, H. S. (2021). Estimating V_{S30} at Korean Peninsular seismic observatory stations using HVSR of event records. *Soil Dynamics and Earthquake Engineering*, *146*, 106650.
3. Aki, K. (1993). Local site effects on weak and strong ground motion. *Tectonophysics*, *218*(1-3), 93-111.
4. Allen, T. I., and Wald, D. J. (2009). On the use of high-resolution topographic data as a proxy for seismic site conditions (V_{S30}). *Bulletin of the Seismological Society of America*, *99*(2A), 935-943.
5. Bahavar, M., Spica, Z. J., Sánchez-Sesma, F. J., Trabant, C., Zandieh, A., and Toro, G. (2020). Horizontal-to-vertical spectral ratio (HVSR) IRIS station toolbox. *Seismological Research Letters*, *91*(6), 3539-3549.
6. Bard, P. Y., and Participants, S. E. S. A. M. E. (2004). The SESAME project: an overview and main results. In *Proc. of 13th World Conf. on Earthquake Engineering, Vancouver, BC, Canada, August* (pp. 1-6).
7. Barker, P.R. (1996) A Report on Cone Penetrometer and Seismic Cone Penetrometer Testing at Parkway - Wainuiomata. Barker Consulting, P.O. Box 27-106, Wellington.
8. Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (2010). ObsPy: A Python toolbox for seismology. *Seismological Research Letters*, *81*(3), 530-533.

9. Boore, D. M., Stewart, J. P., Seyhan, E., and Atkinson, G. M. (2014). NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. *Earthquake Spectra*, *30*(3), 1057-1085.
10. Borcherdt, R. D. (1970). Effects of local geology on ground motion near San Francisco Bay. *Bulletin of the Seismological Society of America*, *60*(1), 29-61.
11. Borcherdt, R. D. (1992, November). Simplified site classes and empirical amplification factors for site-dependent code provisions. In *Proc. NCEER, SEAOC, BSSC Workshop on Site Response during Earthquakes and Seismic Code Provisions* (pp. 18-20). University of Southern California, Los Angeles, California.
12. Borcherdt, R. D. (1994). Estimates of site-dependent response spectra for design (methodology and justification). *Earthquake spectra*, *10*(4), 617-653.
13. Borcherdt, R. D. (2012). VS30–A site-characterization parameter for use in building Codes, simplified earthquake resistant design, GMPEs, and ShakeMaps. In *The 15th world conference on earthquake engineering*.
14. Box, G. E., and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211-243.
15. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
16. Cadet, H., and Duval, A. M. (2009). A shear wave velocity study based on the KiK-net borehole data: A short note. *Seismological Research Letters*, *80*(3), 440-445.
17. Campbell, K. W., and Bozorgnia, Y. (2014). NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra. *Earthquake Spectra*, *30*(3), 1087-1115.

18. Castellaro, S., and Mulargia, F. (2009). VS 30 estimates using constrained H/V measurements. *Bulletin of the Seismological Society of America*, *99*(2A), 761-773.

19. Castellaro, S., Mulargia, F., and Rossi, P. L. (2008). VS30: Proxy for seismic amplification? *Seismological Research Letters*, 79(4), 540–543. https://doi.org/10.1785/gssrl.79.4.540

20. Cave, A. (2020). *Site characterisation of the Hamilton basin using surface wave methods* (Doctoral dissertation, The University of Waikato).

21. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785-794).

22. Chen, C. T., Kuo, C. H., Lin, C. M., Huang, J. Y., and Wen, K. L. (2022). Investigation of shallow S-wave velocity structure and site response parameters in Taiwan by using high-density microtremor measurements. *Engineering Geology*, *297*, 106498.

23. Chiou, B. S. J., and Youngs, R. R. (2014). Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra. *Earthquake Spectra*, *30*(3), 1117-1153.

24. Cox, B., J. Vantassel, B. Cox (2018). *Dynamic Characterization of Wellington, New Zealand*. DesignSafe-CI. https://doi.org/10.17603/DS24M6J

25. Deschenes, M. R., Wood, C. M., Wotherspoon, L. M., Bradley, B. A., & Thomson, E. (2018). Development of deep shear wave velocity profiles in the Canterbury Plains, New Zealand. *Earthquake Spectra*, *34*(3), 1065-1089.

26. European Committee for Standardization (CEN). (2004) Eurocode 8: *Design of structures for earthquake resistance – Part 1: General rules, seismic actions and rules for buildings* (EN 1998-1:2004). Brussels: European Committee for Standardization.
27. Felicetta C., Russo E., D'Amico M., Sgobba S., Lanzano G., Mascandola C., Pacor F., Luzi L. (2023) Italian Accelerometric Archive v4.0 - Istituto Nazionale di Geofisica e Vulcanologia, Dipartimento della Protezione Civile Nazionale. doi: 10.13127/itaca.4.0
28. Field, E. H., & Jacob, K. H. (1995). A comparison and test of various site-response estimation techniques, including three that are not reference-site dependent. *Bulletin of the seismological society of America*, *85*(4), 1127-1143.
29. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
30. García-Jerez, A., Luzón, F., Sánchez-Sesma, F. J., Lunedei, E., Albarello, D., Santoyo, M. A., and Almendros, J. (2013). Diffuse elastic wavefield within a simple crustal model. Some consequences for low and high frequencies. *Journal of Geophysical Research: Solid Earth*, *118*(10), 5577-5595.
31. García-Jerez, A., Piña-Flores, J., Sánchez-Sesma, F. J., Luzón, F., and Perton, M. (2016). A computer code for forward calculation and inversion of the H/V spectral ratio under the diffuse field assumption. *Computers and geosciences*, *97*, 67-78.
32. Geyin, M., and Maurer, B. W. (2023). US National VS 30 Models and Maps Informed by Remote Sensing and Machine Learning. *Seismological Society of America*, *94*(3), 1467-1477.

33. Ghofrani, H., and 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 Dynamics and Earthquake Engineering*, *67*, 30-43.

34. Goulet, C.A., Kishida, T., Ancheta, T.D., Cramer, C.H., Darragh, R.B., Silva, W.J., Hashash, Y.M., Harmon, J., Parker, G.A., Stewart, J.P. and Youngs, R.R. (2021). PEER NGA-east database. *Earthquake Spectra*, *37*(1_suppl), 1331-1353.

35. Gupta, R. K., Agrawal, M., Pal, S. K., and Das, M. K. (2021). Seismic site characterization and site response study of Nirsa (India). *Natural Hazards*, *108*(2), 2033-2057.

36. Harris, F.J. (Jan 1978). On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform. Proceedings of the IEEE, 66 (1), 51-83. DOI:10.1109/PROC.1978.10837

37. Hassani, B., and Atkinson, G. M. (2016). Applicability of the site fundamental frequency as a VS 30 proxy for central and eastern North America. *Bulletin of the Seismological Society of America*, *106*(2), 653-664.

38. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

39. Hill, M., Kaiser, A., Wotherspoon, L., & Manea, E. (2023). Using 3D geological models to create maps of estimated Vs30 and site period. New Zealand Society for Earthquake Engineering 2023 Annual Conference.

40. Hudson, K.S., Palmer S.M., Ahdi, S.K., Hassani, B., Toro, G., Yong, A. (2021). Inter-method HVSR bias and the resultant V_{S30}-f_d relationship for measured-V_{S30} stations in the

western United States [Oral presentation]. 2021 Annual Meeting, Seismological Society of America, April 19-23, 2021.

41. Ibs-von Seht, M., and Wohlenberg, J. (1999). Microtremor measurements used to map thickness of soft sediments. *Bulletin of the Seismological Society of America*, *89*(1), 250-259.
42. Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456).
43. Kaiser, A. E., and Louie, J. N. (2006). Shear-wave Velocities in Parkway Basin, Wainuiomata, from Refraction Microtremor Surface Wave Dispersion. New Zealand: GNS Science.
44. Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
45. Konno, K., and Ohmachi, T. (1998). Ground-motion characteristics estimated from spectral ratio between horizontal and vertical components of microtremor. *Bulletin of the Seismological Society of America*, *88*(1), 228-241.
46. Kuo, C. H., Chen, C. T., Lin, C. M., Wen, K. L., Huang, J. Y., and Chang, S. C. (2016). S-wave velocity structure and site effect parameters derived from microtremor arrays in the Western Plain of Taiwan. *Journal of Asian Earth Sciences*, *128*, 27-41.
47. Kuo, C. H., Wen, K. L., Lin, C. M., Wen, S., and Huang, J. Y. (2015). Investigating Near Surface S-Wave Velocity Properties Using Ambient Noise in Southwestern Taiwan. *Terrestrial, Atmospheric and Oceanic Sciences*, *26*(2).

48. Lermo, J., and Chávez-García, F. J. (1993). Site effect evaluation using spectral ratios with only one station. *Bulletin of the seismological society of America*, *83*(5), 1574-1594.

49. McNamara, D. E., Stephenson, W. J., Odum, J. K., Williams, R. A., and Gee, L. (2015). Site response in the eastern United States: A comparison of Vs30 measurements with estimates from horizontal-to-vertical spectral ratios. In J. W. Horton Jr., M. C. Chapman, and R. A. Green (Eds.), *The 2011 Mineral, Virginia, earthquake, and its significance for seismic hazards in eastern North America* (GSA Special Paper 509, pp. 67–79). Geological Society of America.

50. Molnar, S., Cassidy, J.F., Castellaro, S., Cornou, C., Crow, H., Hunter, J.A., Matsushima, S., Sánchez-Sesma, F.J. and Yong, A. (2017). Application of MHVSR for site characterization: State-of-the-art. *Surveys in Geophysics*.

51. Molnar, S., Sirohey, A., Assaf, J., Bard, P.-Y., Castellaro, S., Cornou, C., Cox, B., Guillier, B., Hassani, B., Kawase, H., Matsushima, S., Sánchez-Sesma, F.J., Yong, A. (2022). A review of the microtremor horizontal-to-vertical spectral ratio (MHVSR) method. *Journal of Seismology*, *26*(4), 653-685.

52. Mucciarelli, M., and Gallipoli, M. R. (2001). A critical review of 10 years of microtremor HVSR technique. *Boll. Geof. Teor. Appl*, *42*(3-4), 255-266.

53. Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

54. Nakamura, Y. (1989). A method for dynamic characteristics estimation of subsurface using microtremor on the ground surface. *Railway Technical Research Institute, Quarterly Reports*, *30*(1).

55. Ng, A. Y. (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (p. 78).

56. Nikolaou, S., Vera-Grunauer, X., Gilsanz, R., Luque, R., Kishida, T., Diaz-Fanas, G., Antonaki, N., Toulkeridis, T. Miranda, E., Diaz, V., Alzamora, D., Athanasopoulos-Zekkos, A., Lyvers, G., Morales, E., Lopez, P., Rollins, K., Wood, C., O'Rourke, T., Lopez, S. (2017). GEER-ATC Mw7.8 Ecuador 4/16/16 Earthquake Reconnaissance Part I : Seismological and Ground Motion Aspects, 16th world Conference on Earthquake Engineering, Jan 9-13, Santiago, Chile.

57. Nogoshi, M. and Igarashi, T. (1971). On the amplitude characteristics of microtremor, Part II. *Journal of the seismological society of Japan*, *24*, 26-40.

58. Pan, D., Miura, H., Kanno, T., Shigefuji, M., and Abiru, T. (2022). Deep-Neural-Network-Based Estimation of Site Amplification Factor from Microtremor H/V Spectral Ratio. *Bulletin of the Seismological Society of America*, *112*(3), 1630-1646.

59. Parker, G. A., Harmon, J. A., Stewart, J. P., Hashash, Y. M., Kottke, A. R., Rathje, E. M., Silva, W.J. and Campbell, K. W. (2017). Proxy-based VS 30 estimation in central and eastern North America. *Bulletin of the Seismological Society of America*, *107*(1), 117-131.

60. Parker, G. A., Stewart, J. P., Boore, D. M., Atkinson, G. M., and Hassani, B. (2022). NGA-subduction global ground motion models with regional adjustment factors. *Earthquake Spectra*, *38*(1), 456-493.

61. Parolai, S., Bormann, P., and Milkereit, C. (2002). New relationships between Vs, thickness of sediments, and resonance frequency calculated by the H/V ratio of seismic

noise for the Cologne area (Germany). *Bulletin of the seismological society of America*, *92*(6), 2521-2527.

62. QGIS.org. (2024). *QGIS Geographic Information System*. QGIS Association. http://www.qgis.org
63. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106.
64. Rahimi, S., and Wood, C. M. (2022). Comparison of Wavefield Transformation Techniques for MASW Data Processing. In *Geo-Congress 2022* (pp. 82-91).
65. Rahimi, S., and Wood, C. M. (2023). Reducing Mode Assignment Errors in Surface Wave Inversion for Sites with a Very Shallow Impedance Contrast Using Love Type Surface Waves. In *Geo-Congress 2023* (pp. 173-182).
66. Rai, M., Rodriguez-Marek, A., and Chiou, B. S. (2017). Empirical terrain-based topographic modification factors for use in ground motion prediction. *Earthquake Spectra*, *33*(1), 157-177.
67. Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.
68. Sánchez-Sesma, F.J., Rodríguez, M., Iturrarán-Viveros, U., Luzón, F., Campillo, M., Margerin, L., García-Jerez, A., Suarez, M., Santoyo, M.A. and Rodriguez-Castellanos, A., (2011). A theory for microtremor H/V spectral ratio: application for a layered medium. *Geophysical Journal International*, *186*(1), 221-225.
69. Seed, H. B., Ugas, C., and Lysmer, J. (1976). Site-dependent spectra for earthquake-resistant design. *Bulletin of the Seismological society of America*, *66*(1), 221-243.

70. SESAME (2004). Guidelines for the Implementation of the H/V Spectral Ratio Technique on Ambient Vibrations: Measurements, Processing and Interpretation (pp. 1-62). SESAME European Research Project WP12.

71. Seyhan, E., Stewart, J. P., Ancheta, T. D., Darragh, R. B., & Graves, R. W. (2014). NGA-West2 site database. *Earthquake spectra*, *30*(3), 1007-1024.

72. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

73. Stanko, D., and Markušić, S. (2020). An empirical relationship between resonance frequency, bedrock depth and VS 30 for Croatia based on HVSR forward modelling. *Natural Hazards*, *103*(3), 3715-3743.

74. Stewart, J. P., and Seyhan, E. (2013). *Semi-empirical nonlinear site amplification and its application in NEHRP site factors*. Pacific Earthquake Engineering Research Center.

75. Stolte, A., Wotherspoon, L., Cox, B., Wood, C., Jeong, S., & Munro, J. (2023). The influence of multiple impedance contrasts on mHVSR site period estimates in the Canterbury Plains of New Zealand and implications for site classification. *Earthquake Spectra*, *39*(1), 288-309.

76. Teague, D., Cox, B., Bradley, B., and Wotherspoon, L. (2018). Development of deep shear wave velocity profiles with estimates of uncertainty in the complex interbedded geology of Christchurch, New Zealand. *Earthquake Spectra*, *34*(2), 639-672.

77. Thornley, J. D., Dutta, U., Douglas, J., and Yang, Z. J. (2021). Evaluation of horizontal to vertical spectral ratio and standard spectral ratio methods for mapping shear wave

velocity across Anchorage, Alaska. *Soil Dynamics and Earthquake Engineering*, *150*, 106918.

78. Vantassel, J.P. (2020). jpvantassel/hvsrpy: latest (Concept). Zenodo. http://doi.org/10.5281/zenodo.3666956

79. Vantassel, J.P. (2025). *hvsrpy*: An Open-Source Python Package for Microtremor and Earthquake Horizontal-to-Vertical Spectral Ratio Processing. *Seismological Research Letters* 2025. https://doi.org/10.1785/0220240395.

80. Vantassel, J. P., and Cox, B. R. (2025). An extension to the procedure for developing uncertainty-consistent shear wave velocity profiles from inversion of experimental surface wave dispersion data. *Soil Dynamics and Earthquake Engineering*, *193*, 109329.

81. Vantassel, J., Cox, B., Wotherspoon, L., and Stolte, A. (2018). Mapping depth to bedrock, shear stiffness, and fundamental site period at CentrePort, Wellington, using surface-wave methods: Implications for local seismic site amplification. *Bulletin of the Seismological Society of America*, *108*(3B), 1709-1721.

82. Vantassel, J.P., Cox, B.R., and Brannon, D.M. (2021). HVSRweb: An Open-Source, Web-Based Application for Horizontal-to-Vertical Spectral Ratio Processing. IFCEE 2021. https://doi.org/10.1061/9780784483428.005.

83. Vantassel, J. P., Ilgac, M., Athanasopoulos Zekkos, A., Yong, A., Hassani, B., and Martin, A. J. (2024). Are the Horizontal-to-Vertical Spectral Ratios of Earthquakes and Microtremors the Same?. *Bulletin of the Seismological Society of America*, *114*(6), 3078-3092.

84. Vantassel, J. P., Stolte, A. C., Wotherspoon, L. M., and Cox, B. R. (2023). AutoHVSR: A machine-learning-supported algorithm for the fully-automated processing of horizontal-

to-vertical spectral ratio measurements. *Soil Dynamics and Earthquake Engineering*, *173*, 108153.

85. Vassallo, M., Riccio, G., Mercuri, A., Cultrera, G., and Di Giulio, G. (2023). HV noise and earthquake automatic analysis (HVNEA). *Seismological Society of America*, *94*(1), 350-368.
86. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. and Van Der Walt, S.J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.
87. Wald, D. J., and Allen, T. I. (2007). Topographic slope as a proxy for seismic site conditions and amplification. *Bulletin of the Seismological Society of America*, *97*(5), 1379-1395.
88. Wathelet, M., Chatelain, J. L., Cornou, C., Giulio, G. D., Guillier, B., Ohrnberger, M., and Savvaidis, A. (2020). Geopsy: A user-friendly open-source tool set for ambient vibration processing. *Seismological Research Letters*, *91*(3), 1878-1889.
89. Wills, C. J., & Clahan, K. B. (2006). Developing a map of geologically defined site-condition categories for California. *Bulletin of the Seismological Society of America*, *96*(4A), 1483-1501.
90. Wills, C.J., Petersen, M., Bryant, W.A., Reichle, M., Saucedo, G.J., Tan, S., Taylor, G. and Treiman, J. (2000). A site-conditions map for California based on geology and shear-wave velocity. *Bulletin of the Seismological Society of America*, *90*(6B), S187-S208.

91. Wood, C., A. Himel (2019). *Development of Deep Shear Wave Velocity Profiles at Seismic Stations in the Mississippi Embayment*. DesignSafe-CI. https://doi.org/10.17603/ds2-be10-q668

92. Wood, C., L. Woodfield, R. Rieth (2020). *Dynamic Site Characterization in Mexico City Following the 2017 Mw 7.1 Puebla-Mexico City Earthquake*. DesignSafe-CI. https://doi.org/10.17603/ds2-4kc2-zr63

93. Wood, C., Rahimi, S. (2022). Mapping Subsurface Conditions for Transportation Applications. Arkansas Department of Transportation (ARDOT), Final Project Report.

94. Wotherspoon, L. M., Bradley, B., Hills, A., Thomson, E. M., Jeong, S., Wood, C. M., & Cox, B. R. (2015). Development of deep VS profiles and site periods for the Canterbury region.

95. Wotherspoon, L., Bradley, B., Thomson, E., Cox, B. R., Wood, C. M., & Deschenes, M. (2016). Dynamic site characterisation of Canterbury strong motion stations using active and passive surface wave testing. *EQC Report*, *14*(663), 30.

96. Wotherspoon, L. M., Kaiser, A. E., Stolte, A. C., & Manea, E. F. (2024). Development of the site characterization database for the 2022 New Zealand National Seismic Hazard Model. *Seismological Research Letters*, *95*(1), 214-225.

97. Yaghmaei-Sabegh, S., and Hassani, B. (2020). Investigation of the relation between Vs30 and site characteristics of Iran based on horizontal-to-vertical spectral ratios. *Soil Dynamics and Earthquake Engineering*, *128*, 105899.

98. Yong, A., A. Martin, K. Stokoe, and J. Diehl (2013). ARRA-funded VS30 measurements using multi-technique approach at strong- motion stations in California and Central-

Eastern United States, U.S. Geol. Surv. Open-File Rept. 2013-1102, 1–59, doi: 10.3133/ofr20131102.

99. Yong, A., Cochran, E., Andrews, J., Hudson, K., Martin, A., Yu, E., Herrick, J. and Dozal, J. (2021). VS 30 and dominant site frequency (fd) as provisional station ML corrections (d ML) in California. *Bulletin of the Seismological Society of America*, *111*(1), 61-76.

100. Yust, M. B. S. (2018). *Dynamic site characterization of TexNet ground motion stations* (Doctoral dissertation), University of Texas at Austin.

101. Zhu, C., Weatherill, G., Cotton, F., Pilz, M., Kwak, D. Y., and Kawase, H. (2021). An open-source site database of strong-motion stations in Japan: K-NET and KiK-net (v1. 0.0). *Earthquake Spectra*, *37*(3), 2126-2149.