Engineering Geology 306 (2022) 106764

Contents lists available at ScienceDirect

Engineering Geology

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

Machine learning techniques for estimating seismic site amplification in the Santiago basin, Chile

J.P. Díaz [a,b], E. Sáez [a,b,*], M. Monsalve [b], G. Candia [b,c], F. Aron [a,b], G. González [b,d]

[a] Department of Structural and Geotechnical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
[b] Research Center for Integrated Disaster Risk Management (CIGIDEN), Santiago, Chile
[c] Facultad de Ingeniería, Universidad del Desarrollo, Santiago, Chile
[d] Departamento de Ciencias Geológicas, Universidad Católica del Norte, Antofagasta, Chile

ARTICLE INFO

Keywords:
Seismic site amplification
Machine learning
Seismic hazard

ABSTRACT

Seismic site amplification and seismic hazard maps are crucial inputs for decision making and risk evaluation in places where seismicity imposes a significant risk to human life and infrastructure. In this work, we propose a novel machine learning (ML) based methodology to integrate qualitative and quantitative data to map the degree of seismic amplification in an area of Chile, one of the most seismically active countries on Earth. Our method uses measurements of surface shear wave velocities (V_{s30}) and predominant frequencies (f_0) combined with gravity anomaly maps to update the geographic extension of seismic amplification classes. Additionally, we trained the predictive models to interpolate and extrapolate V_{s30} and f_0 to the unsampled sites. Applying this method to the Santiago basin resulted in (i) a refined seismic amplification map, and (ii) maps of V_{s30} and f_0 estimated with improved accuracy. The best predictions, obtained by ML techniques and validated through cross-validation, are possibly due to the inclusion of spatial covariates for algorithm training, enhancing the ability of the model to capture the spatial correlations of geological, geophysical and geotechnical data. The estimation of predominant frequencies (f_0) is improved considerably by including gravity as a covariant. The accuracy of the f_0 predictions apparently depends more on the choice of covariates than on the algorithm used, while the V_{s30} predictions are more sensitive to the chosen algorithm. These results illustrate the great potential of machine learning predictive algorithms in digital soil mapping, which surpass traditional geostatistical techniques. The major contribution of this work is to introduce a novel methodology, based on artificial intelligence models, to extend local measurements of site-specific dynamic properties. This information can be used to quantitatively estimate seismic hazard over a regional scale.

1. Introduction

Chile is one of the most seismically active countries in the world. The active continental margin where the oceanic (Nazca) plate subducts under the continental (South American) plate, extends between 18° and 47° S. This active margin has generated some of the largest subduction earthquakes on record (e.g. 9.5 Mw Valdivia 1960 earthquake, Cifuentes, 1989). Observational data show that, along the country, the impact of seismic waves increases in the areas closest to the trench and decreases with distance from the seismogenic source (i.e. Leyton et al., 2010).

Estimations of the recurrence of large historical earthquakes (Mw > 8), adjacent to the Santiago Metropolitan Region (SMR), indicate occurrence approximately every 80 years (Fig. 1); the 8.0 Mw Algarrobo earthquake of 1985 being the last major event (Ruiz and Madariaga, 2018). The SMR concentrates 41.6% of the population and 41.5% of the gross domestic product of Chile and so it is crucial to be able to accurately estimate the ground shaking and minimize the damage caused by major seismic events in the region.

The local geotechnical conditions of a site can induce important seismic amplifications (i.e. Aki, 1988), known as site effects. Site effects can be evaluated through the dynamic characterization of the subsoil, measuring dynamic parameters of the site's response to the pass of seismic waves. In this regard, two parameters appear key to achieve a

* Corresponding author at: Departamento de Ingeniería Estructural y Geotécnica, Pontificia Universidad Católica de Chile, Santiago, Chile.
E-mail addresses: jpdiaz7@uc.cl (J.P. Díaz), esaez@ing.puc.cl (E. Sáez), mauricio.monsalve@cigiden.cl (M. Monsalve), faron@ing.puc.cl (F. Aron), ggonzale@ucn.cl (G. González).

https://doi.org/10.1016/j.enggeo.2022.106764
Received 20 December 2021; Received in revised form 7 June 2022; Accepted 13 June 2022
Available online 17 June 2022

correct geotechnical characterization: 1) the shear wave velocity profile, which allows a primary evaluation of the dynamic response of a site (Tokimatsu, 1997), and 2) the predominant frequency of sites, f_0, defined as the frequency associated with the impedance contrast that predominates in the site (Maringue et al., 2021). In Chile, the seismic code classifies a site based on the value of the harmonic average of the propagation velocity of shear waves in the first 30 m of depth, V_{s30}. Other parameters from in situ tests, such as drilling, standard penetration and laboratory tests, are also required for this classification. Geophysical techniques based on surface waves allow the determination of both V_{s30} and f_0 site parameters in a non-invasive, fast and low-cost manner (Becerra et al., 2015). Several studies have shown that surface wave methods are reliable techniques to evaluate site effects (e.g. Oliveira et al., 2020; Pegah and Liu, 2016).

Seismic site amplification and seismic hazard maps are crucial inputs for decision making and risk evaluation in places where seismicity imposes a significant risk to human life and infrastructure. In the case of SMR, most of the available seismic zonation maps do not consider the dynamic characteristics of the subsoil and have been developed based on observations of seismic damage distribution. Only a few studies in Chile have directly incorporated site specific dynamic properties to generate seismic zonation maps (e.g., Leyton et al., 2011), however, the dynamic properties of sites have for long been routinely included in the generation of site class maps in other countries (e.g. Lee and Tsai, 2008).

The ability of some geological and terrain-based proxies to anticipate V_{s30} values and site class has been also studied (e.g. Forte et al., 2017; Forte et al., 2019; Mori et al., 2020; Stewart et al., 2014). Digital soil mapping (DSM) uses statistical models to generate digital representations of the spatial distribution of soil properties using point soil observations and spatially exhaustive environmental covariates (proxies or independent variables) (McBratney et al., 2003; Scull et al., 2003). In recent decades, DSM has proven successful in producing soil property maps, capturing the main patterns of soil spatial variation (e.g. McBratney et al., 2000; Molnar et al., 2020), however its use in seismic geotechnical engineering in Chile is hampered by limited dynamic site characterizations and data availability.

Spatial interpolation of natural variables is important in many scientific fields. In the 1980s, the kriging geostatistical interpolation technique was introduced, gaining popularity as it had the advantage - unlike previous techniques - of considering the spatial correlation of the data and being able to quantify the interpolation error (Matheron, 1963). The statistical approach of data mining has proved useful in providing tools for DSM. This approach identifies patterns in datasets through statistical methods, transforming information into a perceptible structure for further use (Khaledian and Miller, 2020).

Machine learning (ML) has been increasingly used for spatial interpolation in fields such as soil science and geology (e.g. Li and Heap, 2014; Marzan et al., 2021). ML is highly dependent on the relationship between the target variable and its associated covariates and can produce remarkably accurate results if this correlation is strong (Sekulić et al., 2020). A great advantage of ML over traditional techniques is their ability to capture non-linear associations within the data without having to assume explicit functional forms for these relations (Kohestani et al., 2015). Nevertheless, for many practical applications, it is hard to obtain the large data sets required to train these models. To address this problem, some authors have generated a high number of experimental (laboratory) results representing the real problem to train their models (e.g. Huang et al., 2021) or tested specific machine learning algorithms that use fewer samples to generate predictions (e.g. Huang and Zhao, 2018).

There have been few attempts to use these techniques in the area of seismic geotechnical engineering (e.g. Kim et al., 2021; Yaghmaei-Sabegh and Rupakhety, 2020), and none in Chile to spatially predict dynamic site properties (V_{s30} and f_0). Zhao and Wang (2020) did use ML tools to infer the subsurface stratification and characterize soil property profiles. Kohestani et al. (2015) used ML tools to predict liquefaction potential in soils based on cone penetration tests. Thomson et al. (Thompson et al., 2010; Thompson et al., 2014) used variants of kriging to estimate V_{s30} in Kobe and California. Though there have been several attempts to predict V_{s30} using geostatistical methods, until now - at least in Chile - no tested techniques of ML have been usedto spatially predict dynamic site properties (V_{s30} and f_0). We posit that ML techniques can reasonably predict V_{s30} and f_0 values and improve the accuracy of quantitative seismic hazard assessments in the Santiago basin.

This paper aims to improve the quality and accuracy of seismic zonation maps in the Santiago Metropolitan Region (SMR). We present two results: (i) an updated seismic microzoning of the SMR based on recent measurements of dynamic properties of sites through the basin, and (ii) a methodology which integrates the geology, geophysical data and seismic geotechnical engineering concepts, to predict V_{s30} and f_0 accurately using ML. We compared six predictive algorithms to estimate V_{s30} and f_0: simple kriging, linear regression, elastic net, random forests, artificial neural networks and decision trees. The best predictions obtained were used to generate seismic hazard maps in the study area, through a state-of-the-art software that uses ground motion prediction equations (GMPE), seismicity models and seismic scenarios to assess the

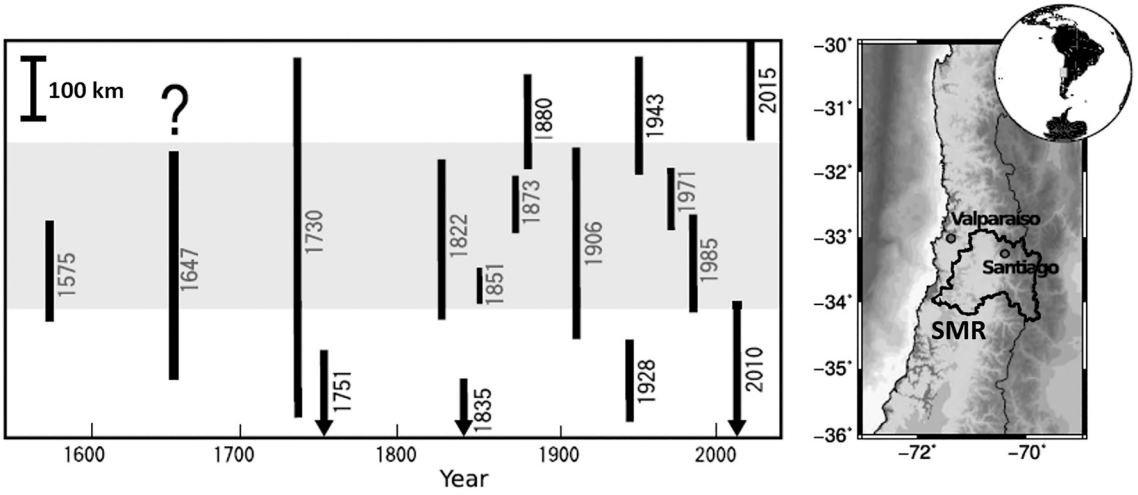

Fig. 1. Historical recurrence of large earthquakes in central Chile. The length of the bars indicates the approximate extent of the rupture that generated each event, while their widths are proportional to the registered magnitudes. The map to the right indicates the location of the Santiago Metropolitan Region (SMR). (Modified from Bravo et al., 2019).

seismic hazard due to both subduction related earthquakes and crustal earthquakes. The major contribution of this work is to introduce a novel estimation method, based on ML, to extend local measurements of a site's dynamic properties in an area of interest. Additionally, the work introduces a data augmentation methodology for enhancing the dataset so that the statistical and machine learning models learn to predict in situations where the dataset is excessive sparse. This information can be used to quantitatively estimate the seismic hazard over a regional scale.

2. Methodology

2.1. Santiago basin and available seismic zonation maps

The Santiago basin is located in the center of the SMR (see Fig. 2) and contains an alluvial sedimentary infill which has accumulated between the Main Cordillera and the Coastal Cordillera, reaching maximum depths in the range of 350–500 m (Yáñez et al., 2015). There were two seismic zonation maps developed after the earthquakes of 1985 and 2010 which define seismic classes within the basin representing a seismic microzonation:

1) The 2004 seismic zonation map of the Santiago Metropolitan Region (von Igel et al., 2004) displays the degree of seismic amplification related to the relevant geological classes in the area. The seismic amplification is qualitatively determined based on seismic intensities collected after the March 3rd earthquake of 1985 (Ms 7.8) and available geological information. This map does not incorporate quantitative information related to the dynamic characteristics of sites and was developed predominantly from observations of damage to buildings and infrastructure.
2) The 2011 seismic zoning of the SMR, Chile (Leyton et al., 2011) was a seismic zoning of the Santiago basin carried out based on the surface geology, available measurements of the predominant seismic period (Bonnefoy-Claudet et al., 2009) and the distribution of damage observed following the 2010 Maule earthquake.

2.2. Collected V_{s30} and f_0 measurements

Surface wave geophysical methods were used to measure V_{s30} and f_0 at 312 sites in the Santiago basin (Fig. 2a). In cases in which more than one measurement of V_{s30} and f_0 was available, the uncertainly in the values of these parameters was evaluated to report the level of accuracy of the field measurements.

V_{s30} measurements were obtained through an inversion process of the empirical dispersion curve of each site. These curves were obtained using a method that combines active sources (hammer) with passive sources (ambient noise), with a multichannel analysis approach (Humire et al., 2015). The objective of the surveys was generally to describe the dispersion curve between a wavelength of 10 m and 90 m using a combination of both active and passive methods. The methods used in this study were the f-k (frequency-wave number) method (Kvaerna and Ringdahl, 1986; Lacoss et al., 1969) for active 1D and passive 2D arrays, the SPAC (spatial autocorrelation) method (Aki, 1957) for passive 2D arrays, and the ESPAC (extended spatial autocorrelation) method (Hayashi, 2008) for passive 1D arrays. In sites with more than one estimate of V_{s30}, measurements showed coefficients of variation (CV) between 0% and 35.8%. Furthermore, in 50% of the cases the CV was <2.6%.

To obtain f_0, the horizontal-to-vertical spectral ratio (HVSR) or Nakamura's technique (Nakamura, 1989) was used. This technique estimates the ratio of the Fourier amplitude spectrum between horizontal components and vertical component produced by environmental vibrations. Then, the predominant period ($T_0 = 1/f_0$) is defined by the peak of the HVSR curve (Pastén, 2007), and the amplitude of this peak is defined as A_0. This study used a variation of the Nakamura method (Leyton et al., 2012) which considers fixed windows of 60 s, applying the Stockwell Transform (S-transform) in each of these windows. For f_0, CV between 0% and 46.4% were obtained, and in 50% of the cases this value was lower than 5.6%. For A_0, a CV between 0% and 50.8% were obtained, and in 50% of the cases this value was <12%.

Fig. 2a shows the distribution of the V_{s30} values, while Fig. 3a shows simultaneously the distribution of f_0 and A_0. These sites concentrate near the urban areas of the study area and were selected to improve the definition of transitions between seismic classes. Figs. 2b and 3b show the histograms of the data distribution of V_{s30} and f_0, respectively.

2.3. Degree of seismic amplification index

Chile is currently in the process of improving its seismic classification system for residential buildings. The new classification simultaneously uses the value of V_{s30} and the estimate of the predominant period T_0, which were used as a degree of seismic amplification index in this work.

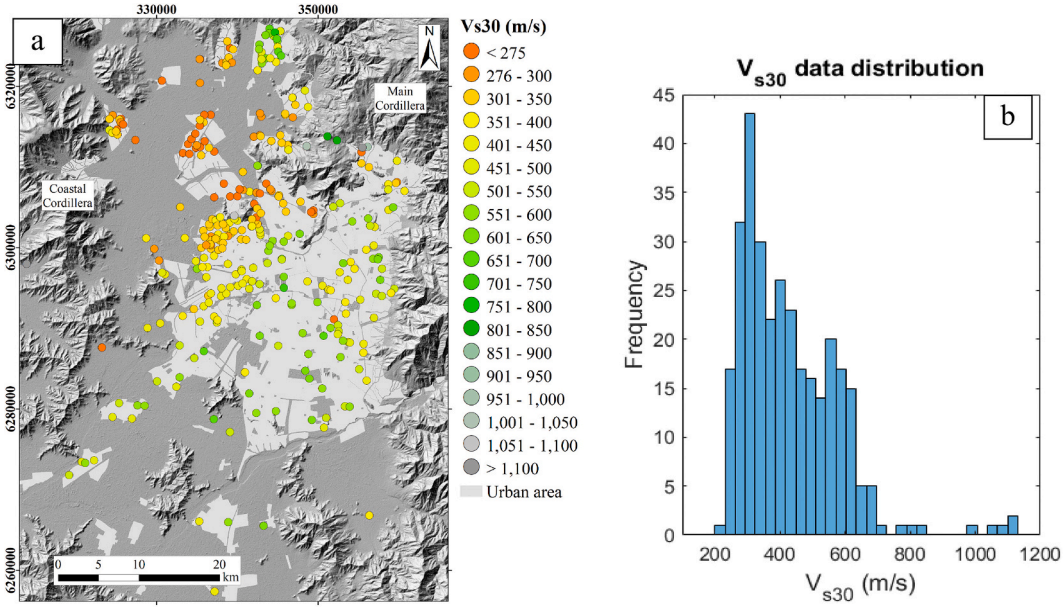

Fig. 2. a) Distribution of V_{s30} in the study area, and b) Distribution histogram of V_{s30}.

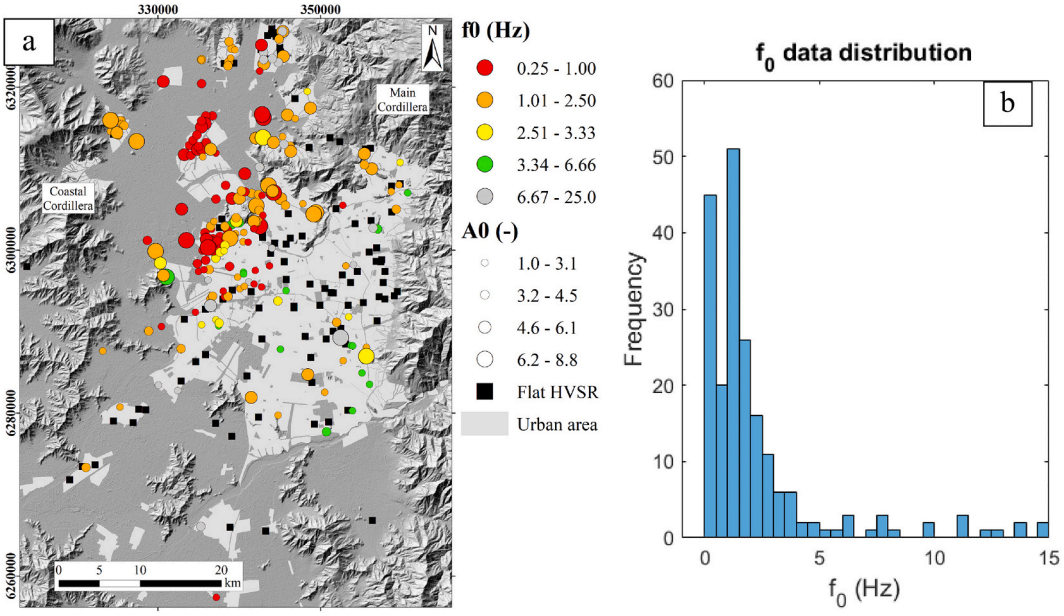

Fig. 3. a) Distribution of f_0 in the study area, and b) Distribution histogram of f_0. Among the 312 sites considered, 101 sites reported flat HVSR curves.

Five indices were defined: A, B, C, D and E, according to the criteria shown in Table 1 (modified from Verdugo et al., 2019).

2.4. Seismic classes update

The current seismic zonation maps were compared against the distribution of the degree of seismic amplification indices from Table 1, to solve the limits of the classes. Five seismic classes were defined in this study, following the degree of seismic amplification indices (A, B, C, D and E), where a seismic class classified as A is the one with the best seismic response (i.e. rock) and a class classified as E is the one with the worst response in terms of seismic amplification expected due to site effects (i.e. very soft and/or deep site).

Additionally, gravity models of the Santiago basin published by Yáñez et al. (2015) were incorporated to fill information gaps in areas where there was insufficient data from V_{s30} and f_0 measurements to update the limits between seismic classes determined by geologic criteria. The direct gravimetric residual is also expected to have a good correlation with f_0 because it provides an idea of depth to a significant change in density or gravimetric contrast (Maringue et al., 2021). Then, classes with a considerable gravimetric anomaly (deep sites), fine granulometries and/or presence of surface volcanic ash were classified as with low seismic response (D or E).

2.5. Prediction of V_{s30} and f_0

This section describes the procedures and considerations used to generate a predictive model of V_{s30} and f_0 in the Santiago basin. First, the database and the covariates used to train the predictive models for each explored algorithm are presented; secondly, the algorithms used are briefly described. Additionally, the methods for validating and evaluating the predictive performance of the models are detailed.

Table 1
Definition of the degree of seismic amplification index.

Index category	First criterion: V_{s30} (m/s)	Second criterion: T_0 (s)
A	≥ 900	< 0.15 or flat HVSR
B	≥ 500	< 0.30 or flat HVSR
C	≥ 350	< 0.40 or flat HVSR
D	≥ 180	< 1.00 or flat HVSR
E	< 180	

2.5.1. Data and choice of covariates

Proper choice of training covariates for ML predictive models is key to obtaining reasonable and accurate estimates in DSM. In this work, we chose the covariates shown in Table 2. Punctual covariates such as terrain slope, topographic elevation, and geological typology are included, based on previous work, that showed an improvement in the

Table 2
Description of the covariates used to train the models and predict the values of V_{s30} and f_0 in Santiago basin.

Covariate	Definition	Unit
Slope	Maximum rate of elevation change between each pixel.	°
Elevation	Elevation above sea level according to the DEM.	M
Seismic class	Seismic class of the Degree of Seismic Amplification Map that contains the evaluation point.	–
Gravity*	Residual Bouguer anomaly measured at site.	mGal
External seismic class	Seismic class of the Degree of Seismic Amplification Map that does not contain the evaluation point but is the closest to it.	–
Edge distance	Inverse of the minimum distance between the evaluation point and the seismic class that contains it.	1 / km
Distance to closest observation**	Inverse of the minimum distance between the evaluation point and the observation closest to it.	1 / km
V_{s30} closest to observation**	Value of V_{s30} in the closest observation to the evaluation point.	m / s
HVSR peak in the closest observation**	Closest observation to the evaluation point has a peak in the HVSR curve (1) or does not present a peak (0).	Binary
Nearest predominant frequency**	Predominant frequency (f_0) measured at the closest observation to the evaluation point.	Hz
HVSR amplitude**	Amplitude of the HVSR curve (A_0) at the closest observation to the evaluation point.	–

* Covariate was used only in the Santiago basin area where the gravimetric study was carried out.

** Covariates were calculated for the 6 closest observations to each evaluation point.

performance of predictive models in predicting V_{s30} (Wills and Clahan, 2006; Wald and Allen, 2007). Slope and elevation were obtained from a digital elevation model (DEM) of 12.5 m resolution available from public satellite data (https://asf.alaska.edu), while the surface geology covariate was obtained directly from the geological maps.

Despite the success of ML predictive models in DSM, most of these approaches do not consider the possible spatial correlation between the observed data and focus mostly on punctual covariates, thus they do not fully exploit the available spatial information. Several recent investigations have shown that the inclusion of spatial covariates (in addition to punctual ones), such as distance and inverse of distance to neighboring observations, considerably improve predictions of ML models in DSM (Beguin et al., 2017; Deng et al., 2020). Therefore, we used a combination of punctual and spatial covariates to train the models, as indicated in Table 2. Fig. 4 describes the spatial covariates chosen in this work. Note that since gravimetry did not cover the entire study area, we worked in two independent areas: a zone with gravimetric information and a separate zone without gravimetric information. The predictive models for both areas differed only in the inclusion of the gravimetric covariate in their training.

Although several measurements were available at some sites, which made it possible to generate uncertainly indicators, this information was available for only 35% of sites. To simplify the training of the models, the pair V_{s30} and f_0 leading to the most conservative classification according to the criteria in Table 1, was selected for training.

2.5.2. Predictive methods

There are numerous methods for predicting soil properties from a sample data set; in this paper, geostatistical and ML predictive methods are compared. The geostatistical method tested was Simple Kriging (SK), while the ML methods were Linear Regression (LR), Elastic-Net (EN), Random Forests (RF), Artificial Neural Networks (ANN) and Decision Trees (DT). Hyperparameters (tuning parameters) were used to setting the algorithms. These variables can be adjusted by trial and error until a minimum amount of error is obtained when the predictions are validated. The way in which the predictive algorithms used in this article are described in Table 3.

To predict V_{s30} and f_0, a total of 47 models were tested. The models and their hyperparameter settings are shown in Table 4.

2.5.3. Predictive performance evaluation

All models were programmed in Python. The first treatment to the original database was the application of the Data Augmentation

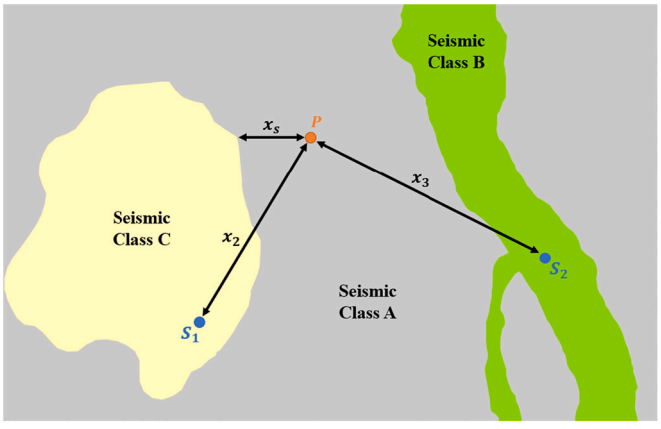

Fig. 4. Example of the spatial covariates associated with a point *P* in the study area. In this case, point *P* is in seismic class A, x_g is the shortest distance from *P* to the boundaries of this seismic class. Seismic class C is the closest class to point *P*. On the other hand, x_2 and x_3 are the two shortest distances to *P* of $S_1$and S_2, where V_{s30} and f_0 are known and are considered spatial covariates associated with the point *P*.

Table 3
Description of predictive algorithms used in this research.

Algorithm	Description
SK	A generalized least-squares regression algorithm that assigns weights for the surrounding measured values to derive a prediction for each location. These weights, in addition to being based on the distance between the measured points and the prediction site, are based on the general spatial arrangement of the measured points (Filho et al., 2017). This method considers that the spatial fluctuation of the mean of the observations is unknown, but constant (Thompson et al., 2010).
LR	Fits a linear model with coefficients to minimize the sum of the squares in the difference between the observed and predicted values by the linear approximation (Hutcheson and Sofroniou, 1999).
EN	Combines two linear models: (i) the Ridge method, that addresses some of the problems of Linear Regression by imposing a penalty on the size of the coefficients; and (ii) the Lasso method, that estimates sparse coefficients. EN learns from its shortcomings to improve the regularization of statistical models and is useful when there are multiple features which are correlated with one another (Friedman et al., 2010).
RF	Randomly selects a group of observations from the larger set to build a decision tree that is associated with this group. The process is repeated to build multiple decision trees based on different observation sets. Typically, two-thirds of the observations are used for algorithm training, and the rest are used to test model error. RF randomly permutes the arrangement of the covariates in the selection of the observation groups, considering all the possibilities of arrangement of covariates. Finally, the predictions are based on the average of the results produced from thousands of decision trees. It is currently the most widely used ML algorithm in DSM, and it often shows excellent potential when it comes to spatial data (Boulesteix et al., 2012; Deng et al., 2020).
ANN	These mimic biological neural networks, building a set of nodes called artificial neurons, forming a network. Through multiple layers of the network, information is transmitted from one neuron to another. The connection between neurons consists of weights that define the network architecture, organize the layers, and adjust the parameters to learn from the data. Training the network consists of comparing the input to the output and calculating a residual, then the algorithm goes back through the layers to fit the equation of the network and recalculate the residual. This process is repeated until a minimum residual is reached. It is a common and longstanding algorithm used in DSM (Behrens et al., 2005~~; Were et al., 2015~~).
DT	These models divide the data space and fit a simple prediction model within each partition. A decision tree is the graphical result of each partition. DT are intended for dependent variables that take continuous or ordered discrete values (Breiman et al., 2017)

Table 4
Tested models and hyperparameter settings.

Algorithm	Number of tested models	Hyperparameter	Settings
SK	1	Power of the inverted distances	1
LR	1	–	–
EN	2	Total penalty value (α)	1, 2
		Penalty ratio (ρ)	0.5
RF	25	Number of trees	20, 40, 50, 80, 100
		Proportion of variables considered (%)	10, 20, 40, 60, 80
ANN	17	Number of neurons per layer	10, 20, 30, 50, 75, 100
		Maximum number of layers	5
		Maximum number of neurons	100
		Trigger function	tanh*
		Training method	lbfgs*
DT	1	–	–

* 'tanh' refers to the hyperbolic tangent function; 'lbfgs' refers to an implementation of the BFGS quasi-Newton method for nonlinear optimization. For more information about the hyperparameters of the ML models used in this paper visit https://scikit-learn.org/stable/.

technique by artificially increasing the initial number of observations, as well as their covariates, to obtain a larger training dataset, while preserving the associations present in the original data (Padarian et al., 2019). Data Augmentation has been shown to reduce variance and overfitting, improve robustness, and mitigate bias of ML models (Roudier et al., 2020; Shorten and Khoshgoftaar, 2019; Zhong et al., 2020). This technique is especially beneficial for ANN models, because they are particularly sensitive to small sample sizes (Khaledian and Miller, 2020). Additionally, this technique is expected to allow training of predictor models under difficult prediction scenarios, such as sites with few or no measurements in their proximity.

After testing with data augmentations of 10, 20, 30, 40 and 50 times the size of the database, the largest increase was applied because marginally better results were consistently obtained for both V_{s30} and f_0. All models were trained with 90% of the augmented database (training sets) by cross-validation and validated with the remaining 10% (testing sets). To ensure that the comparisons between the models were valid, the same set of covariates was kept for the training of all predictive models (see Table 2). For all models, the root mean squared error (RMSE) and the root relative mean squared error (RRMSE) were calculated. Additionally, the predictive models of f_0 permitted an evaluation of the probability of a peak in the HVSR curve. For those obtained probabilities <60%, a flat HVSR curve was assumed. The error rate in the prediction of this probability (ErrRate) was also quantified. These errors were calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i' - y_i)^2} \quad (1)$$

$$RRMSE = \sqrt{\frac{1}{n}\sum \left(\frac{y_i' - y_i}{y_i}\right)^2} \quad (2)$$

$$ErrRate = \frac{\sum |b_i' - b_i|}{n} \quad (3)$$

where $i = 1$. n is the i-th iteration and n is the total number of tests, y_i' is the i-th predicted value and y_i is the i-th observed value. b_i' is a binary value equal to 1 if the i-th prediction has a peak in the HVSR curve and equal to 0 otherwise, while b_i is also a binary value equal to 1 if the i-th observation has a peak in the HVSR curve and equal to 0 otherwise. Fig. 5 shows the main steps in the training of the predictive models of V_{s30} and f_0.

2.5.4. Probabilistic hazard assessment

A PGA map consistent with a return period of 475 years was developed in the *Seismic-Hazard* software (Candia et al., 2019) which

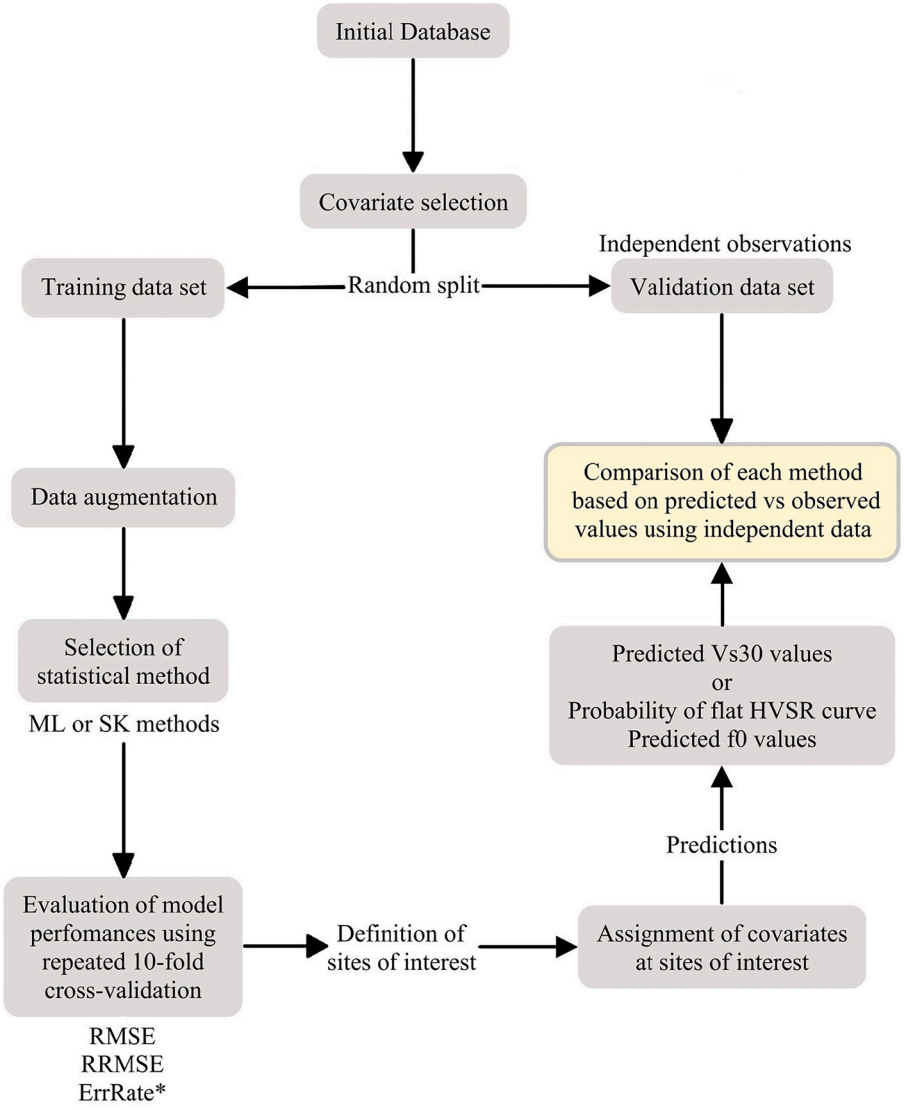

Fig. 5. Flowchart showing the main steps of the modeling process of V_{s30} and f_0 used in this study. *ErrRate is only calculated in the estimation of f_0.

computes hazard-consistent ground motion parameters (e.g., PGA, PSA) at a single site or distributed over a large region, using state-of-art seismicity models and rigorous account of scientific uncertainties. We adopted the Poulos et al. (2019) source model for subduction earthquakes, which uses the Slab 1.0 model (Hayes et al., 2012) to account for the contact surface between the Nazca and South American plates. Additionally, the seismicity model includes four crustal faults reported in the GEM global Active Faults catalog (Styron and Pagani, 2020) located within a 200 km radius of the study area and the Diablo Fault (also known as Baños Morales Fault) located towards the east boundary of the basin. A logic tree of 3 ground motion models was defined, giving greater weight to the Montalva et al. (2017) model, as it collects local knowledge and incorporates large earthquakes (Mw > 8.0) that occurred in Chile in the period 2010–2017, and uses V_{s30} to estimate PGA. The seismicity from crustal sources was modeled with the PCEnga attenuation law (Macedo and Candia, 2020) which considers magnitudes between 4 and 8, closest distance to the rupture plane of <200 km and V_{s30} values between 300 m/s and 1000 m/s. Three PGA maps are presented to illustrate the influence of different V_{s30} realizations in the PGA distribution. Finally, to discuss PGA changes due to uncertainty of the predictive models, a sensitivity analysis is performed on 30 observed and estimated V_{s30} values not used for training.

3. Results

3.1. Geophysical survey

This section presents the results of the dynamic characterization and seismic classification of the sites within the Santiago basin. The combination of the parameters V_{s30} and f_0 in the sampled sites allows us to assign an index of the degree of seismic amplification to each site, as indicated in Table 1. The seismically classified sites are shown in Fig. 6. Sites rated A are generally rock outcrops, with very high V_{s30} (> 900 m/s) and flat HVSR curves (without a clear peak). The A_0 value can be considered as an indicator of the predominant impedance contrast of the site (Pilz et al., 2010; Leyton et al., 2013). It should be noted that there were only a few measurements performed in sites rated A. This is because of the challenges involved in accessing remote areas with flat rock outcrops, required to deploy large arrays of sensors (~100 m long). Sites rated B show high values of V_{s30} (exceeding 500 m/s), where the soils correspond mainly to alluvial fans and fluvial gravels. These sites also have mostly flat HVSR curves. C sites correspond typically to alluvial fans composed by gravels with a higher content of fines and sandy sites. These sites are located mainly nearby the Main Cordillera and the Mapocho river, to the east and southwest of Santiago, respectively. In these sectors, the HVSR curves are also mostly flat, showing that in general there are no predominant frequencies or clearly defined impedance contrasts. The sites D and E, composed by fine-grained, sedimentary deposits, are more prone to seismic amplification where V_{s30} tends to be <350 m/s, f_0 show low values (< 1 Hz), and large values of A_0 are observed.

3.2. Seismic class zoning map

The joint analysis of the site classification with the collected maps of seismic response and gravity model, resulted in an updated seismic microzoning of the Santiago basin (Fig. 6A).

The seismic classes and site amplification indexes show good correlation with specific geologic units present in the area (Fig. 6B). For instance, the class with the best seismic response (A) encompasses the Mesozoic-Cenozoic igneous basement, along the Coastal Cordillera and the high Andes Mountain front, flanking the western and eastern sides of the Santiago basin, respectively. The sites of classes B and C are mostly constrained to the basin infill in the central and southern parts of the map. The sediments filling up the basin are composed by a thick (>500 m) sequence of well compacted alluvial and fluvial strata from mass removal of the two mountain fronts and long-lived deposition of the main rivers traversing the basin (Yáñez et al., 2015). The sites and

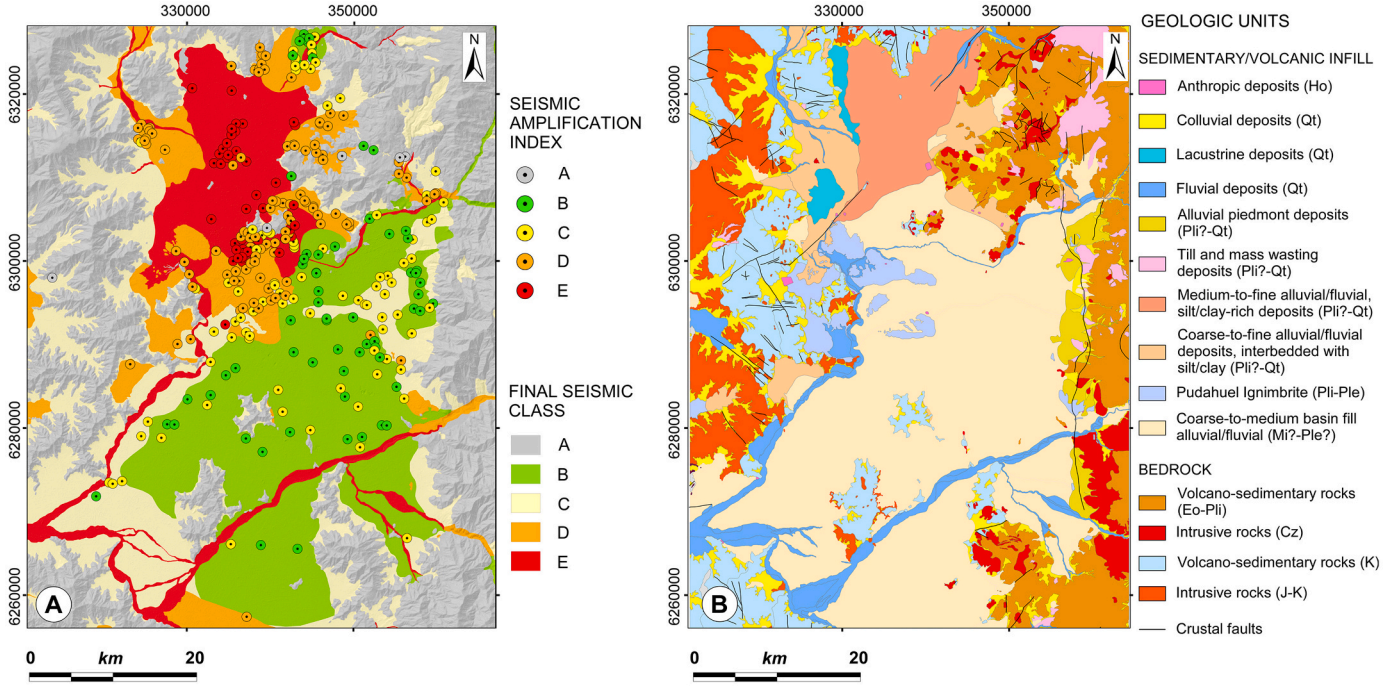

Fig. 6. A. Map of the degree of seismic amplification indexes in measured sites (colored circles) and microzoning of the Santiago de Chile basin, the latter adapted from the maps by Von Igel et al. (2004) and Leyton et al. (2011). B. Geologic map of the study area compiled and reclassified from the works by Wall et al. (1999), Sellés and Gana (2001), and Espinoza et al. (2019). Geologic time abbreviations are: Jurassic (J), Cretaceous (K), Cenozoic (Cz), Eocene (Eo), Miocene (Mi), Pliocene (Pli), Pleistocene (Ple), Quaternary (Qt) and Holocene (Ho). For both maps the coordinates are projected to the UTM zone 19S (cartesian in meters), using the parameters of the WGS84 geoid.

classes with the worst seismic response (D and E) are in general restricted to the northern side of the basin and to the recent fluvial deposits along the path and floodplains of lowland rivers (Fig. 6). The fluvial/alluvial deposits in the northern side, which make up most of class E, are made of poorly consolidated fine-grain silt and clay, interbedded with gravel, sand, and ash (Leyton et al., 2011). This could explain their much higher seismic amplification compared to the rest of the basin infill, characterized by coarser, well compacted gravels. Of special attention is the Pliocene-Pleistocene Pudahuel ignimbrite, a 20 m thick layer of volcanic ash with lithics and pumice categorized as class E, which blankets part of the relief of the mountains and hills flanking the western, and partially the eastern, side of the basin. The poorly consolidated tuff and ash layers of this unit can be also found interbedded with the basin sedimentary infill (Leyton et al., 2011) (Fig. 6B).

3.3. Predictive models comparison and resulting maps

This section shows the prediction performances of the 6 predictive algorithms of V_{s30} and f_0, shown in Tables 5 and 6. These results were obtained from the test sets defined for cross-validation, as explained in Section 2.5.3.

In the prediction of V_{s30}, LR is the best performing algorithm in those sites where the gravimetric covariant is available, with an RMSE of 68.4 m/s and an RRMSE of 17.6%, followed by RF and DT. Similarly, when the gravimetric covariant is not available, the best performing algorithm was LR, with an RMSE of 70.5 m/s and an RRMSE of 17.8%, followed by RF and DT. The spatial distribution of RMSE across the study area is provided as supplementary material.

In the prediction of f_0, RF is the best performing algorithm in those sites where the gravimetric covariant is available, with an RRMSE of 45.6%, an RMSE of 0.13 Hz and an ErrRate of 21.3%, followed by EN and LR (ANN is discarded as having too large an ErrRate). On the other hand, when the gravimetric covariant was unknown, the best performing algorithm was LR, with an RRMSE of 164.8%, an RMSE of 2.43 Hz and an ErrRate of 22.3%, followed by RF and SK. It can be noted that when the gravimetry covariant was available, the estimation error reduced considerably.

Figs. 7 and 8 show the distribution of V_{s30} and f_0 predicted by the best resulting models. In the case of the f_0 prediction, only the area where gravity modeling is available is shown because the error outside of this zone was too high. The distribution of the values of both parameters are consistent with the observations shown in Section 2.2. Once the dynamic characterization of sites for the entire Santiago basin were available, it was possible to proceed with the Seismic Hazard assessment to obtain the PGA map. Fig. 9 shows the estimate of PGA in the study area for the settings described in Section 2.5.4. This map was developed using a uniform 350 m square grid across the Santiago basin.

3.4. PGA sensitivity

To measure the uncertainty associated with the prediction of V_{s30} with the three best predictive algorithms, we randomly chose 30 sites where this parameter was measured (Fig. 10). None of them were used in the training of the models. In general, the predictions of V_{s30} show a good fit for V_{s30} <500 m/s. Above these values, the three models tend to underestimate $V_{s3}0$. This underestimation of Vs30 for stiff sites is reflected in an overestimation of PGA for a design scenario when PGA is calculated with the predictive algorithms. This overestimation grows when the sites become more stiff, reaching PGA values of about 15% higher than those calculated from the measured V_{s30}. The main reason is that the number of sites used to train the model under 500 m/s is much larger (96% of the database) than the data available over this value of V_{s30}.

Table 5
Cross-validation of the best models for each algorithm for the V_{s30} predictions.

	Gravimetric covariant included		Gravimetric covariant not included	
Algorithm	RMSE (m/s)	RRMSE (%)	RMSE (m/s)	RRMSE (%)
SK	185.6	37.5	233.4	41.1
LR	**68.4**	**17.6**	**70.5**	**17.8**
EN	124.6	27.5	141.9	30.0
RF	79.6	19.4	86.3	20.4
ANN	208.7	39.7	234.5	46.5
DT	98.4	22.9	116.7	26.7

Note: The bolds show the best performances obtained among the algorithms.

Table 6
Cross-validation of the best models for each algorithm for the f_0 predictions.

	Gravimetric covariant included			Gravimetric covariant not included		
Algorithm	RMSE (Hz)	RRMSE (%)	ErrRate (%)	RMSE (Hz)	RRMSE (%)	ErrRate (%)
SK	0.155	54.8	28.7	2.917	217.9	29.1
LR	0.132	50.8	**20.8**	**2.432**	**164.8**	22.3
EN	**0.111**	47.2	32.1	2.938	224.1	32.2
RF	0.130	**45.6**	21.3	2.502	183.1	**20.9**
ANN	0.135	48.7	41.3	3.716	214.3	45.5
DT	0.162	58.0	27.0	3.976	423.0	31.3

Note: Figures in bold show the best performances obtained among the tested algorithms.

4. Discussion

4.1. Seismic microzoning

The results shown in Section 3.2 were obtained from an integrated approach that uses geology, geophysics, and earthquake geotechnical engineering information, combining geophysical characterization of sites with seismic response maps. The result was a refined seismic microzoning that considers site effects on the seismic response of the soils in the Santiago basin (Fig. 6).

Some differences are observed in the seismic classes with respect to prior maps, mainly in the central zone of the study area (see Von Igel et al., 2004; Leyton et al., 2011). Nevertheless, despite the seismic class updates, the refined seismic microzoning shows consistency with previous maps. The main explanation for the zoning differences relates to the new combination of qualitative (i.e. geology) with substantial quantitative information obtained using the geophysical techniques (Fig. 5).

4.2. Prediction of V_{s30} and f_0

Regarding the prediction of V_{s30}, for the sites where the gravimetric covariant is available, the three best models had similar performance. The best performance was obtained with LR, followed by RF and DT. Between the first (LR) and third (DT) best model, the difference of RMSE and RRMSE is only about 30 m/s and 5%, respectively. The rest of the models show larger errors. When gravity data were not available, the models exhibited more dispersion. Between the first and third best model, the difference of RMSE and RRMSE is about 46 m/s and 8.9%, respectively. Apparently, the V_{s30} predictions are sensitive to the chosen algorithm, however ML algorithms continue to outperform traditional SK.

These results show that the use of ML algorithms to predict V_{s30} provide reliable approximations with reasonable uncertainty, improving the capabilities of the SK geostatistical algorithm.

Regarding the prediction of f_0, for those sites within the boundaries of the gravity model, the algorithms provide quite stable results. Between the best (EN) and worst (DT) model in terms of RMSE, the difference of RMSE and RRMSE is only about 0.05 Hz and 12%, respectively. Similar to the V_{s30} estimations, ML algorithms outperform

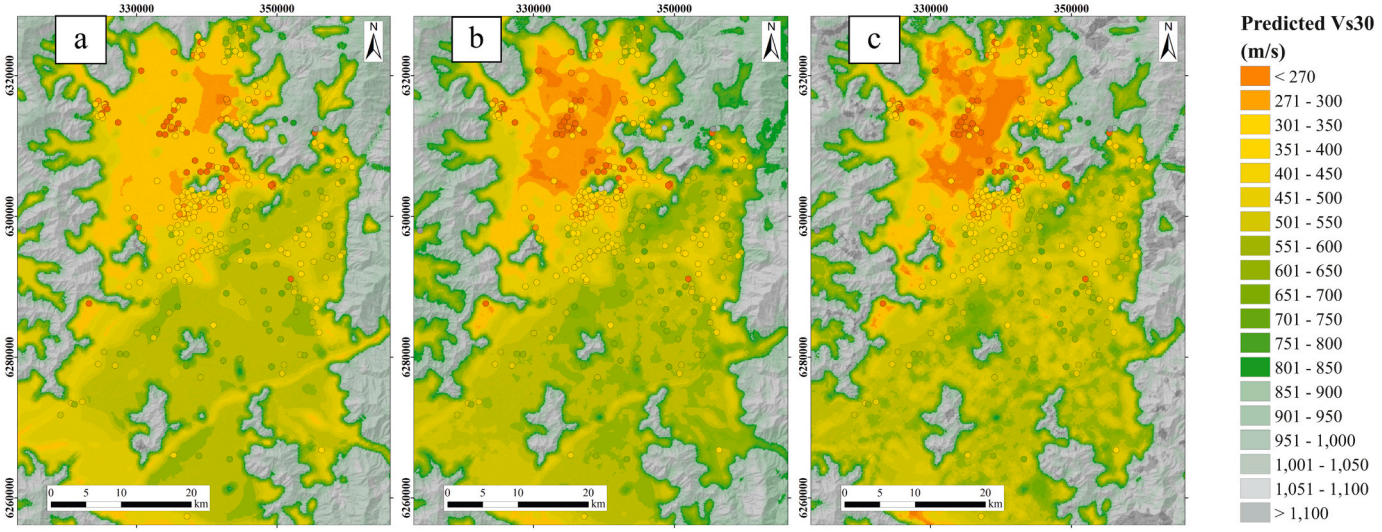

Fig. 7. V_{s30} prediction maps using a) LR model, b) best RF and c) DT model. Circles show the observed V_{s30} in the same color scale.

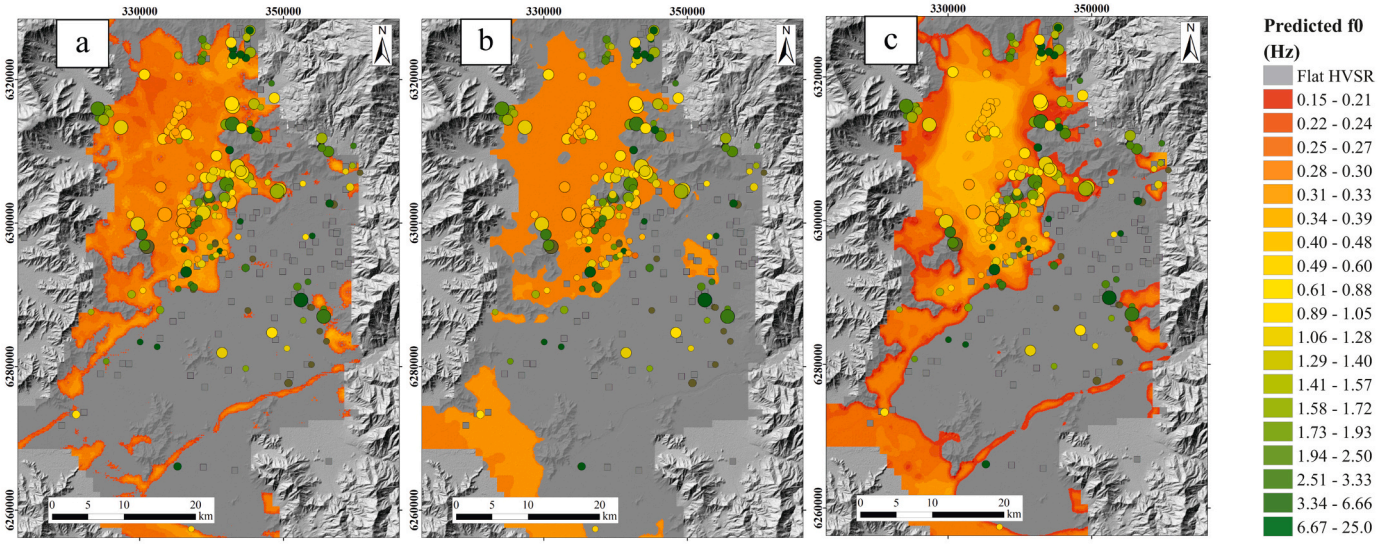

Fig. 8. f_0 prediction maps using a) best RF model, b) best EN model and c) LR model. Circles show the observed f_0 in the same color scale and their sizes are proportional to A_0.

SK. The best performance was obtained with RF, followed by EN and LR. Nevertheless, the RRMSE values are high (about 46% in the best case). This can be explained by the previous step of prediction of the peak of the HVSR curve. This previous step inherently increases the errors in the final prediction of f_0, since there are sites where the real HVSR curve is flat, but the algorithm is not able to identify this situation and, erroneously, provides a numeric estimate of f_0. The predictions of f_0 were highly influenced by this initial step since it was made using regression algorithms and not classification algorithms, which discriminate between finite categories or classes. The use of classification algorithms is outside the scope of this article, but it could be a good opportunity to improve predictions in future work. Unlike the case of V_{s30}, for those sites where gravity information is not available, the predictions of f_0 fail, displaying RRMSE values higher than 100% in all models. Thus, these results are not reliable and are considered unacceptable.

The good performance of the ML algorithm can be explained by the following reasons: 1) a good density of samples in the study area, which allowed an accurate characterization of most of the types of sites; 2) a correct choice of the covariates used for training, because original covariates were developed focused on including information on the spatial distribution of the data, capturing correlations between geological, geophysical and geotechnical information. The use of the Data Augmentation technique allowed original database to be expanded, avoiding overfitting the models and training them to achieve reasonable predictions in complex scenarios. Another possible reason is that some ML models (e.g. RF and DT) are not limited to using only linear combinations of the observations, and can model the nonlinearity between the target variable and the covariates (Appelhans et al., 2015), i.e. inverse distances used in training probably play a non-linear role. Unexpectedly, it was observed that the LR model performed slightly better than the rest, probably because a large number of covariates were used for the training (44 and 45), which would facilitate the prediction as a linear combination of the covariates.

It is also interesting to note that the improvement in predictions, when including the gravimetric covariant, is substantially greater when predicting f_0 than when predicting V_{s30}, even though the predictions with and without gravity were not made in the same sites. In the best models for predicting V_{s30}, RMSE and RRMSE decreased from 70.5 m/s

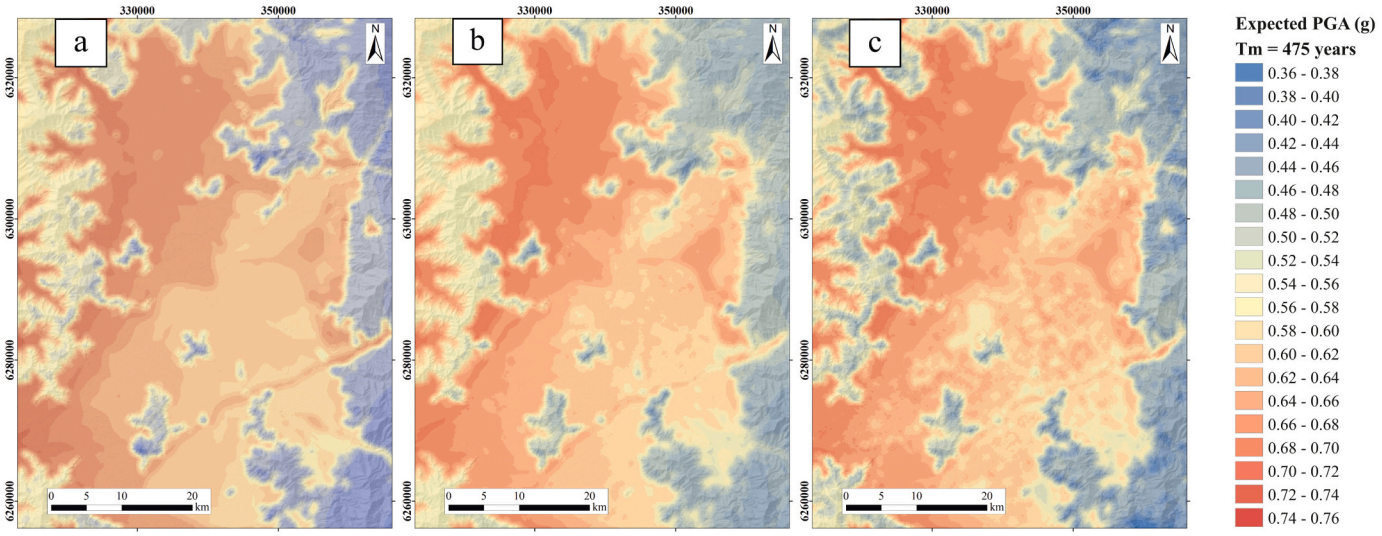

Fig. 9. Expected PGA map for events with T_m = 475 years, based on the estimates of V_{s30} resulting from the a) LR model, b) best RF model, and c) DT model.

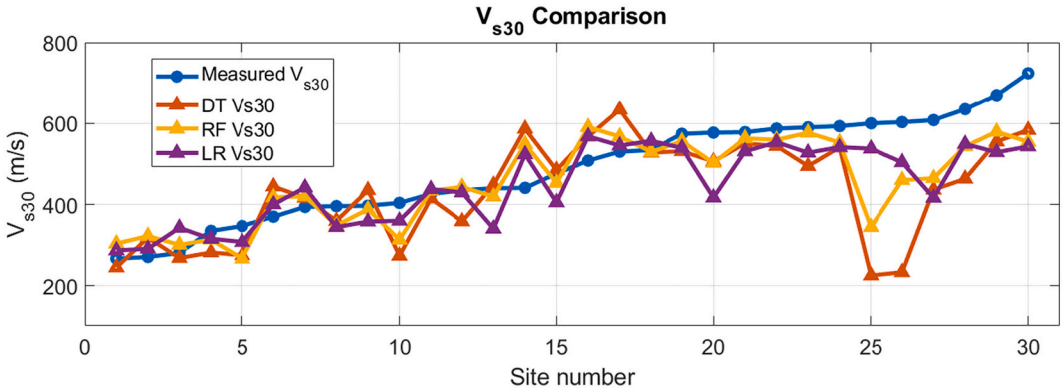

Fig. 10. Uncertainty associated with the predictive models of V_{s30}. Results based on measured V_{s30} are shown in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to 68.4 m/s and from 17.8% to 17.6%, respectively. While, when predicting f_0, RMSE and RRMSE decreased from 2.5 Hz to 0.13 Hz and from 183.1% to 45.6%, respectively. This would be explained because gravimetric residual has a much closer correlation with f_0 than with V_{s30} since it provides an indicator of sediment thickness which often coincides with the depth at which the predominant impedance contrast is located (Maringue et al., 2021). These results suggest that for the same area of interest, including a gravimetric covariate considerably improves the predictions of f_0, and that the predictive capacity of f_0 depends more on the considered covariates than on the algorithms used.

Fig. 7 shows the fit of the predictions to the V_{s30} observations. Quite similar predictions are observed among the best three predictive models. All three models were able to correctly distinguish the rock classes and predict V_{s30} within the basin. In the northwest area, LR is the least conservative predictor, followed by RF, while DT exhibits the lowest values of V_{s30}. In the central and southern sector of the basin, LR is also less conservative than RF and DT. These differences are reflected in the PGA maps (Fig. 9), where the highest expected PGA values are located within the Santiago sedimentary basin and are associated with the most conservative prediction of V_{s30} values (DT model), and the lowest expected PGA values are associated with the model that predicted the highest V_{s30} values (LR model). The rock classes present the lowest expected PGA values, which is consistent with the V_{s30} values observed in these classes. In general, the rock classes to the east of the study area show higher expected PGA values than those located to the east, because they are closer to the main seismogenic source (subduction zone).

The predicted f_0 maps are quite similar. They differ mainly in their ability to predict where the HVSR curve is flat and in the values of f_0 when the HVSR curve has a clear peak. RF is the most accurate, since it identifies the rock classes and the most rigid soils in the study area reasonably well, assigning a flat HVSR curve. It is also the one that best fits the f_0 values observed in rigid and soft soils. EN adequately identifies rock classes but was only able to predict a narrow range of f_0, resulting in an almost bimodal map. LR also identifies rock classes correctly and fits the observations well but is more conservative in the southwestern zone of the study area, delivering low predominant frequencies (deep sites) where soils are known to be rigid. Among the models, RF is the one that best fits the sites with non-flat HVSR curves, correctly identifying the sites classified as D and E, delivering the lowest observed values of f_0 and therefore more prone to seismic amplifications. In general, all models were only able to predict a narrow range of f_0 (0.2 to 0.40 Hz). This is likely due to the previous step that defines the shape of the HVSR curve and to the observed range of f_0 values, which is mostly at frequencies <2 Hz (see Fig. 3b).

The sensitivity analysis performed in Section 3.4 shows that the estimates of V_{s30} are better for values of V_{s30} <500 m/s, and they get worse when this limit is exceeded. This is because the database contains less information on stiffer sites compared to softer sites, making it difficult for algorithms to predict the V_{s30} value of the most rigid sites. Despite this, the error in the PGA calculation associated with the estimation

error of V_{s30} is small for rigid sites, reaching maximum values of only 15%.

Finally, the results of our work show that: (i) it is possible to generate a refined seismic microzoning in the Santiago basin incorporating quantitative and qualitative information that allows evaluating site effects on soils, and (ii) it is possible to obtain a reasonably good prediction of the dynamic properties of the sites of the Santiago basin using ML predictive algorithms, surpassing the capabilities of traditional geostatistical predictive models.

4.3. Extensions and improvements

The main limitation in the generation of the refined seismic microzoning is related to a very heterogeneous distribution of data from geophysical techniques. This limitation occurred because there were large areas in which only the geological information was known, losing the opportunity to combine qualitative with quantitative information. This situation was mainly observed away from urban areas. Therefore, for future stages of this research, performing dynamic characterizations in the poorly sampled areas, where the uncertainty of dynamic site properties is greater, is recommended,.

The satisfactory results obtained suggest that this methodology could be replicated in other regions of Chile or the world, combining the dynamic properties of sites with information from surface geology, other geophysical techniques and digital elevation models to improve the accuracy of qualitative seismic response maps.

Regarding the predictions of V_{s30} and f_0, the main limitations were: (i) the few measurements available in rock, which made it difficult to train the models in this type of sites, (ii) the limited spatial extend of gravimetric data, and (iii) the difficulty of predicting the existence or not of a peak in the HVSR curve, increasing the errors in the prediction of f_0. The best estimate of f_0 is available as a tool to classify the unsampled sites of the study area, and the methodology to obtain these estimates remains available to be used in future GMPE that include f_0 in their seismic hazard estimates.

Therefore, for future research it is recommended to have more measurements in rock, at least in an amount comparable to observations in other types of soil. It will also be very useful to design a gravimetric experiment that covers the entire area of interest, since it has been shown that it correlates very well with f_0. It would also be useful to study the performance of ML classification algorithms to decide whether the HVSR curves are flat or not, since better results will probably be obtained than using only regression algorithms.

It is important to mention that there should be other combinations of covariates that further improve the estimates of V_{s30} or f_0. This implies the possibility of removing or adding new covariates to the training database of the predictive models. Exploring new combinations of covariates may be necessary for geological contexts, other than the sedimentary basins of Santiago, for example, to generate a predictive model on a much larger spatial scale.

Finally, the distribution of f_0, presented in this article, was not directly used to assess the seismic hazard in the study area, because there is not a sufficiently validated GMPE applicable to Chile that includes this parameter in its calculations. However, GMPE that include f_0 have already been developed (e.g. Kwak and Seyhan, 2020), using a two-stage nonlinear site amplification model derived empirically from records of strong earthquakes in Japan. Those models show that the residual, associated with GMPE that only include V_{s30} as a site parameter, decreases considerably when including the observed values of f_0 at the sites, mainly for spectral periods >0.1 s. In this way, the use of f_0 could strengthen seismic hazard estimates.

5. Conclusions

This work presents two results: (i) a refined seismic microzoning that provides a qualitative estimate of the seismic hazard in the Santiago basin, and (ii) a methodology that uses ML computational tools to estimate dynamic soil properties in areas of the Santiago basin that were not sampled. These results permit the assessment of site effects and the quantitative estimation of local seismic hazard in terms of PGA. A rationale was presented to generate these local seismic hazard estimates for both V_{s30} and the predominant period f_0.

The integration of qualitative information with quantitative data based on geophysical exploration has made it possible to update existing seismic microzoning maps for the Santiago basin and generate more complete predictive models of site specific dynamic properties.

Regarding the predictive algorithms of V_{s30} and f_0, the following can be concluded:

- Five ML algorithms (LR, EN, RF and DT) were compared with a traditional geostatistical algorithm (SK). For predicting V_{s30}, the most robust algorithm was LR, followed by RF and DT. For predicting f_0, the best algorithm was RF, followed by EN and LR.
- The results of all models were verified by cross-validation, obtaining a RMSE in the best prediction of V_{s30} and f_0 of 68.4. m/s and 0.13 Hz, respectively, and a RRMSE in the best prediction of V_{s30} and f_0 of 17.6% and 45.6%, respectively. The spatial distribution of estimated V_{s30} and f_0 is consistent with the available observations.
- The improvement in the estimates of V_{s30} and f_0 by ML algorithms are explained by the inclusion of spatial covariates for algorithm training, helping the techniques capture the spatial correlations of geological, geophysical and geotechnical data. Similar results are well documented in related literature.
- By including the gravimetric residual covariate in the training of the predictive models, a significant improvement was observed in the prediction of f_0, which suggests that both parameters have a strong correlation in sedimentary contexts.
- The predictive capacity of f_0 apparently depends more on the choice of covariates than on the algorithm used, while the V_{s30} predictions are more sensitive to the chosen algorithm.

Machine learning algorithms have shown to be promising tools in the prediction of site-specific dynamic properties. Future work should focus on increasing the database, exploring which combination of covariates gives better predictions in more general geological contexts, testing ML classification tools to reduce uncertainty when estimating f_0, and including f_0 in the estimation of the seismic hazard by utilizing GMPE that include this parameter.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was possible thanks to the ANID/FONDEF/ID19 | 10021 project “Proposal for a guide, mapping and multi-hazard platform for critical decision-making and adaptation to climate change in metropolitan regions and large conurbations of Chile” and the ANID/FONDAP/ 1511017 “Research Center for Integrated Disaster Risk Management - CIGIDEN”. We thank Alex Zúñiga and Francisca Roldán of CIGIDEN and Alejandro Alfaro (SERNAGEOMIN) for the compilation of geological data and Erik Jensen for fruitful discussion about the geology and data treatment. We would also like to thank Prof. John Browning for his assistance in enhancing readability of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enggeo.2022.106764.

References

Aki, K., 1957. Space and time spectra of stationary stochastic waves, with special reference to microtremors. Bull. Earthq. Res. Inst. 35, 415–456.

Aki, K., 1988. Local site effect on ground motion. In: Ke Engineering and Soil Dynamics II-Recent Advances in Ground Motion Evaluation. Geotechnical Special Publication, p. 20.

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Stat. 14 https://doi.org/10.1016/j.spasta.2015.05.008.

Becerra, A., Podestá, L., Monetta, R., Sáez, E., Leyton, F., Yañez, G., 2015. Seismic microzoning of Arica and Iquique, Chile. Nat. Hazards 79. https://doi.org/10.1007/s11069-015-1863-y.

Beguin, J., Fuglstad, G.A., Mansuy, N., Paré, D., 2017. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306. https://doi.org/10.1016/j.geoderma.2017.06.016.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168 https://doi.org/10.1002/jpln.200421414.

Bonnefoy-Claudet, S., Baize, S., Bonilla, L.F., Berge-Thierry, C., Pasten, C., Campos, J., Volant, P., Verdugo, R., 2009. Site effect evaluation in the basin of Santiago de Chile using ambient noise measurements. Geophys. J. Int. 176 https://doi.org/10.1111/j.1365-246X.2008.04020.x.

Boulesteix, A.L., Janitza, S., Kruppa, J., König, I.R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip. Rev. Data Min. Knowl. Disc. 2 https://doi.org/10.1002/widm.1072.

Bravo, F., Koch, P., Riquelme, S., Fuentes, M., Campos, J., 2019. Slip distribution of the 1985 Valparaiso earthquake constrained with seismic and deformation data. Seismol. Res. Lett. 90 https://doi.org/10.1785/0220180396.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification and regression trees. Classif. Regression Trees. https://doi.org/10.1201/9781315139470.

Candia, G., Macedo, J., Jaimes, M.A., Magna-Verdugo, C., 2019. A new state-of-the-art platform for probabilistic and deterministic seismic hazard assessment. Seismol. Res. Lett. 90 https://doi.org/10.1785/0220190025.

Cifuentes, I.L., 1989. The 1960 Chilean earthquakes. J. Geophys. Res. 94 https://doi.org/10.1029/JB094iB01p00665.

Deng, L., Adjouadi, M., Rishe, N., 2020. Inverse distance weighted random forests: Modeling unevenly distributed non-stationary geographic data. In: 2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020. https://doi.org/10.1109/ICACSIS51025.2020.9263208.

Espinoza, L., Marín, M., Pantoja, G., 2019. Peligros de Remociones en Masa tipo Flujo de la Cuenca Alta del Río Maipo, Región Metropolitana de Santiago. Escala 1:100.000. Servicio Nacional (de Geología y Minería).

Forte, G., Fabbrocino, S., Fabbrocino, G., Lanzano, G., Santucci de Magistris, F., Silvestri, F., 2017. A geolithological approach to seismic site classification: an application to the Molise Region (Italy). Bull. Earthq. Eng. 15 https://doi.org/10.1007/s10518-016-9960-1.

Filho, O.A., Soares, W., Irigaray, C., 2017. Mapping of compactness by depth in a quaternary geological formation using deterministic and geostatistical interpolation models. Environ. Earth Sci. 76 https://doi.org/10.1007/s12665-017-6939-4.

Forte, G., Chioccarelli, E., de Falco, M., Cito, P., Santo, A., Iervolino, I., 2019. Seismic soil classification of Italy based on surface geology and shear-wave velocity measurements. Soil Dyn. Earthq. Eng. 122 https://doi.org/10.1016/j.soildyn.2019.04.002.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 https://doi.org/10.18637/jss.v033.i01.

Hayashi, K., 2008. Development of Surface-wave Methods and its Application to Site Investigations. Dk.

Hayes, G.P., Wald, D.J., Johnson, R.L., 2012. Slab1.0: a three-dimensional model of global subduction zone geometries. Journal of Geophysical Research: Solid. Earth 117. https://doi.org/10.1029/2011JB008524.

Huang, Y., Zhao, L., 2018. Review on landslide susceptibility mapping using support vector machines. Catena (Amst). https://doi.org/10.1016/j.catena.2018.03.003.

Huang, Y., Han, X., Zhao, L., 2021. Recurrent neural networks for complicated seismic dynamic response prediction of a slope system. Eng. Geol. 289 https://doi.org/10.1016/j.enggeo.2021.106198.

Humire, F., Sáez, E., Leyton, F., Yañez, G., 2015. Combining active and passive multi-channel analysis of surface waves to improve reliability of V(S,30) estimation using standard equipment. Bull. Earthq. Eng. 13 https://doi.org/10.1007/s10518-014-9662-5.

Hutcheson, G., Sofroniou, N., 1999. The Multivariate Social Scientist. In: The multivariate social scientist: Introductory statistics using generalized linear models.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81 https://doi.org/10.1016/j.apm.2019.12.016.

Kim, H.S., Sun, C.G., Lee, M.G., Cho, H.I., 2021. Multivariate geotechnical zonation of seismic site effects with clustering-blended model for a city area, South Korea. Eng. Geol. 294 https://doi.org/10.1016/j.enggeo.2021.106365.

Kohestani, V.R., Hassanlourad, M., Ardakani, A., 2015. Evaluation of liquefaction potential based on CPT data using random forest. Nat. Hazards 79. https://doi.org/10.1007/s11069-015-1893-5.

Kvaerna, T., Ringdahl, F., 1986. Stability of various fk estimation techniques. In: NORSAR Semiannual technical summary, p. 1.

Kwak, D.Y., Seyhan, E., 2020. Two-stage nonlinear site amplification modeling for Japan with VS30 and fundamental frequency dependency. Earthquake Spectra 36. https://doi.org/10.1177/8755293020907920.

Lacoss, R.T., Kelly, E.J., Toksöz, M.N., 1969. Estimation of seismic noise structure using arrays. Geophysics 34. https://doi.org/10.1190/1.1439995.

Lee, C.T., Tsai, B.R., 2008. Mapping Vs30 in Taiwan. Terr. Atmos. Ocean. Sci. 19 https://doi.org/10.3319/TAO.2008.19.6.671(PT).

Leyton, F., Ruiz, S., Sepúlveda, S.A., 2010. Reevaluación del peligro sísmico probabilístico en Chile central. Andean Geol. https://doi.org/10.5027/andgeov37n2-a11.

Leyton, F., Sepulveda, S.A., Astroza, M., Acevedo, P., Ruiz, S., Gonzalez, L., Foncea, C., 2011. Seismic Zonation of the Santiago Basin, Chile. In: 5th International Conference on Earthquake Geotechnical Engineering.

Leyton, F., Ramírez, S., Vásquez, A., 2012. Uso y limitaciones de la técnica de microvibraciones (RHV) en la clasificación sísmica de suelos. In: VII Congreso Chileno de Geotecnia.

Leyton, F., Ruiz, S., Sepúlveda, S.A., Contreras, J.P., Rebolledo, S., Astroza, M., 2013. Microtremors' HVSR and its correlation with surface geology and damage observed after the 2010 Maule earthquake (Mw 8.8) at Talca and Curicó, Central Chile. Eng. Geol. 161 https://doi.org/10.1016/j.enggeo.2013.04.009.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2013.12.008.

Macedo, J., Candia, G., 2020. Performance-based assessment of the seismic pseudo-static coefficient used in slope stability analysis. Soil Dyn. Earthq. Eng. 133 https://doi.org/10.1016/j.soildyn.2020.106109.

Maringue, J., Sáez, E., Yañez, G., 2021. An empirical correlation between the residual gravity anomaly and the h/v predominant period in urban areas and its dependence on geology in andean forearc basins. Appl. Sci. (Switzerland) 11. https://doi.org/10.3390/app11209462.

Marzan, I., Martí, D., Lobo, A., Alcalde, J., Ruiz, M., Alvarez-Marron, J., Carbonell, R., 2021. Joint interpretation of geophysical data: applying machine learning to the modeling of an evaporitic sequence in Villar de Cañas (Spain). Eng. Geol. 288 https://doi.org/10.1016/j.enggeo.2021.106126.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58 https://doi.org/10.2113/gsecongeo.58.8.1246.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97. https://doi.org/10.1016/S0016-7061(00)00043-4.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Molnar, Sheri, Assaf, Jamal, Sirohey, Aamna, Raj Adhikari, Sujan, 2020. Overview of local site effects and seismic microzonation mapping in Metropolitan Vancouver, British Columbia, Canada. Eng. Geol. 270, 105568. https://doi.org/10.1016/j.enggeo.2020.105568. ISSN 0013-7952.

Montalva, G.A., Bastías, N., Rodriguez-Marek, A., 2017. Ground-motion prediction equation for the Chilean subduction zone. Bull. Seismol. Soc. Am. 107 https://doi.org/10.1785/0120160221.

Mori, F., Mendicelli, A., Moscatelli, M., Romagnoli, G., Peronace, E., Naso, G., 2020. A new Vs30 map for Italy based on the seismic microzonation dataset. Eng. Geol. 275 https://doi.org/10.1016/j.enggeo.2020.105745.

Nakamura, Y., 1989. Method for dynamic characteristics estimation of subsurface using microtremor on the ground surface. In: Quarterly Report of RTRI (Railway Technical Research Institute) (Japan), p. 30.

Oliveira, L., Teves-Costa, P., Pinto, C., Gomes, R.C., Almeida, I.M., Ferreira, C., Pereira, T., Sotto-Mayor, M., 2020. Seismic microzonation based on large geotechnical database: Application to Lisbon. Eng. Geol. 265 https://doi.org/10.1016/j.enggeo.2019.105417.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digital soil mapping. SOIL 5. https://doi.org/10.5194/soil-5-79-2019.

Pastén, C., 2007. Respuesta sísmica de la cuenca de Santiago. Tesis para optar al grado de Magíster en Ciencias de la Ingeniería, mención Ingeniería Geotécnica. In: Memoria para optar al título de Ingeniero Civil. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.

Pegah, E., Liu, H., 2016. Application of near-surface seismic refraction tomography and multichannel analysis of surface waves for geotechnical site characterizations: a case study. Eng. Geol. 208 https://doi.org/10.1016/j.enggeo.2016.04.021.

Pilz, M., Parolai, S., Picozzi, M., Wang, R., Leyton, F., Campos, J., Zschau, J., 2010. Shear wave velocity model of the Santiago de Chile basin derived from ambient noise measurements: a comparison of proxies for seismic site conditions and amplification. Geophys. J. Int. 182 https://doi.org/10.1111/j.1365-246X.2010.04613.x.

Poulos, A., Monsalve, M., Zamora, N., de la Llera, J.C., 2019. An updated recurrence model for chilean subduction seismicity and statistical validation of its poisson nature. Bull. Seismol. Soc. Am. 109 https://doi.org/10.1785/0120170160.

Roudier, P., Burge, O.R., Richardson, S.J., McCarthy, J.K., Grealish, G.J., Ausseil, A.G., 2020. National scale 3d mapping of soil ph using a data augmentation approach. Remote Sens. 12 https://doi.org/10.3390/rs12182872.

Ruiz, S., Madariaga, R., 2018. Historical and recent large megathrust earthquakes in Chile. Tectonophysics. https://doi.org/10.1016/j.tecto.2018.01.015.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27 https://doi.org/10.1191/0309133303pp366ra.

Sekulić, A., Kilibarda, M., Heuvelink, G.B.M., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens. 12 https://doi.org/10.3390/rs12101687.

Sellés, D., Gana, P., 2001. Geología del Área Talagante-San Francisco de Mostazal, Regiones Metropolitana de Santiago y del Libertador General Bernardo O'Higgins. In: Chile. Escala 1:100.000. Servicio Nacional de Geología y Minería (Issue N° 74).

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image Data Augmentation for Deep learning. J. Big Data 6. https://doi.org/10.1186/s40537-019-0197-0.

Stewart, J.P., Klimis, N., Savvaidis, A., Theodoulidis, N., Zargli, E., Athanasopoulos, G., Pelekis, P., Mylonakis, G., Margaris, B., 2014. Compilation of a local Vs profile database and its application for inference of Vs30 from geologic- and terrain-based proxies. Bull. Seismol. Soc. Am. 104 https://doi.org/10.1785/0120130331.

Styron, R., Pagani, M., 2020. The GEM Global active Faults Database. Earthquake Spectra 36. https://doi.org/10.1177/8755293020944182.

Thompson, E.M., Baise, L.G., Kayen, R.E., Tanaka, Y., Tanaka, H., 2010. A geostatistical approach to mapping site response spectral amplifications. Eng. Geol. 114 https://doi.org/10.1016/j.enggeo.2010.05.010.

Thompson, E.M., Wald, D.J., Worden, C.B., 2014. A VS30 Map for California with geologic and topographic constraints. Bull. Seismol. Soc. Am. 104 https://doi.org/10.1785/0120130312.

Tokimatsu, K., 1997. Geotechnical site characterization using surface waves. In: Ishihara (Ed.), Proc. 1st Intl. Conf. Earthquake Geotechnical Engineering, p. 3.

Verdugo, R., Ochoa-Cornejo, F., Gonzalez, J., Valladares, G., 2019. Site effect and site classification in areas with large earthquakes. Soil Dyn. Earthq. Eng. 126 https://doi.org/10.1016/j.soildyn.2018.02.002.

Von Igel, B., Naranjo, J., Wall, R., 2004. Respuesta Sísmica de la Región Metropolitana de Santiago. In: (Report IR-04-25). Servicio Nacional de Geología y Minería.

Wald, D.J., Allen, T.I., 2007. Topographic slope as a proxy for seismic site conditions and amplification. Bull. Seismol. Soc. Am. https://doi.org/10.1785/0120060267.

Wall, R., Sellés, D., Gana, P., 1999. Área Til Til-Santiago. Región Metropolitana Escala 1:100.000. In: Servicio Nacional de Geología y Minería.

Wills, C.J., Clahan, K.B., 2006. Developing a map of geologically defined site-condition categories for California. Bull. Seismol. Soc. Am. 96 https://doi.org/10.1785/0120050179.

Yaghmaei-Sabegh, S., Rupakhety, R., 2020. A new method of seismic site classification using HVSR curves: a case study of the 12 November 2017 Mw 7.3 Ezgeleh earthquake in Iran. Eng. Geol. 270. https://doi.org/10.1016/j.enggeo.2020.105574.

Yáñez, G., Muñoz, M., Flores-Aqueveque, V., Bosch, A., 2015. Gravity derived depth to basement in santiago basin, Chile: Implications for its geological evolution, hydrogeology, low enthalpy geothermal, soil characterization and geo-hazards. Andean Geol. 42 https://doi.org/10.5027/andgeoV42n2-a01.

Zhao, T., Wang, Y., 2020. Interpolation and stratification of multilayer soil property profile from sparse measurements using machine learning methods. Eng. Geol. 265 https://doi.org/10.1016/j.enggeo.2019.105430.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v34i07.7000.