https://doi.org/10.1007/s12583-021-1560-6

A prediction method of ground motion for regions without available observation data (LGB-FS) and its application to both Yangbi and Maduo earthquakes in 2021

Jin Chen[1, 2], **Hong Tang**[1, 2, *], **Wenkai Chen**[3], **Naisen Yang**[1, 2]

1. *State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing* 100875*, China*
2. *Key Laboratory of Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal University, Beijing* 100875*, China*
3. *Lanzhou Institute of Seismology, China Earthquake Administration, Lanzhou* 730000*, China*

[*] *Corresponding author:* hongtang@bnu.edu.cn

ABSTRACT: Currently available earthquake attenuation equations are locally applicable, and methods based on observation data are not applicable in areas without available observation data. To solve the above problems and further improve the prediction accuracy of ground motion parameters, we present a prediction model referred to as a light gradient boosting machine with feature selection (LGB-FS). It is based on a light gradient boosting machine (LightGBM) constructed using historical strong motion data from the NGA-west2 database and can quickly simulate the distribution of strong motion near the epicenter after an earthquake. Cases study shows that compared with GMPE methods and those based on real-time observation data, the model has a better prediction effect in areas without available observation data and be applied to Yangbi earthquake and Maduo earthquake. The feature importance evaluation based on both information gains and partial dependence plots (PDPs) revealed the complex relationships between multiple factors and ground motion parameters, allowing us to better understand their mechanisms and connections.

KEY WORDS: LightGBM, ground motion parameters, NGA-west2, feature selection, Yangbi, Maduo.

0 INTRODUCTION

Strong ground motion parameters are recognized as the most crucial information in seismic hazard analysis, followed by those of earthquake-resistant structure designs (Jafariavval and Derakhshani, 2020; Sen, 2011). Attenuation relationships are commonly implemented to investigate major ground motion parameters for seismic hazard analysis (Ambraseys and Douglas, 2003). These attenuation relationships show that ground motion parameters are related to certain characteristics, such as magnitude, site, fault, distance from the seismogenic structure, etc. The relationship between ground motion parameters and the abovementioned characteristics is often nonlinear and complicated. There are roughly three types of methods available to try to fit this complex relationship.

The first method is based on traditional ground motion parameter equations (GMPEs). Earlier research to develop GMPEs was reported in the literature of many scholars (Youngs et al., 1997; Youngs et al., 1988; Campbell, 1985; Aptikaev and Kopnichev, 1980). There have also been some GMPEs developed based on the NGA-west2 database in recent years (Abrahamson et al., 2014; Campbell and Bozorgnia, 2014; Chiou and Youngs, 2014; Boore et al., 2013; Idriss, 2013; Abrahamson and Silva, 2008). The second method is based on machine learning. Compared with the traditional GMPE method, the machine learning method can identify implicit relationships between variables more deeply and can better handle nonlinear problems. Therefore, it is a promising method for predicting ground motion parameters. At present, there are related machine learning methods used for ground motion parameter prediction, such as the ANN method (Derras et al., 2014; Derras et al., 2012) and the SVM method (Sonia et al., 2016). Some scholars have combined machine learning methods with other optimization algorithms for model construction, such as the method of combining ANN with the genetic algorithm (Shiuly et al., 2020; Gandomi et al., 2011) or the method of combining ANN with simulated annealing (Alavi and Gandomi, 2011). With the development of artificial intelligence technology, methods based on deep learning have gradually emerged (Derakhshani and Foruzan, 2019). While the deep learning approach is popular, the constructed model is a black-box model, leading to poor result interpretation. The above methods are all based on the strong motion data of historical earthquake cases to construct a model and finally predict the strong motion parameters. The third method estimates ground motion parameters in the area surrounding the epicenter based on real-time observation data from a station after an earthquake occurs. The representative of this method is ShakeMap (Worden et al., 2010). ShakeMap is a system designed for the rapid characterization of the extent and distribution of strong ground shaking following significant earthquakes worldwide.

The main problems of the above methods are as follows. The current earthquake attenuation equations are locally applicable and have poor transferability (Akkar et al., 2014; Kayabali and Beyaz, 2011), and the methods based on observation data are not applicable in areas without stations (Worden et al., 2010). Furthermore, current methods for estimating ground motion parameters are based on the empirical selection of features for model construction, which makes it difficult to adequately fit the relationship between features and parameters (Campbell and Bozorgnia, 2008; Douglas, 2003; Ambraseys et al., 1996). There are many features that affect the ground motion parameters, while there is no research available to specify the importance of features and how different features affect the prediction results.

To solve these problems, a machine learning model is constructed. Due to the utilization of LightGBM and the feature selection method, the final constructed prediction model is referred to as LGB-FS, which has a good transferability and a high prediction accuracy. LightGBM is an improved gradient-enhanced decision tree (GBDT) method (Ke et al., 2017) which has not been used to predict strong motion parameters. We can quickly get the prediction results of the ground motion parameters after earthquake through the model. The model is verified on the actual earthquake cases, and applied to the M_w6.1 earthquake occurred in Yangbi County, Dali Prefecture, Yunnan Province on May 21st (abbreviated as Yangbi earthquake) and the M_w7.3 earthquake occurred in Maduo County, Guoluo Prefecture, Qinghai Province on May 22nd (abbreviated as Maduo earthquake).

1 DATASETS

1.1 Raw data

The data presented in this paper are from the NGA-west2 database. In 2003, the Pacific Earthquake

Engineering Research Center (PEER) began to develop a new ground motion attenuation relationship for describing shallow crustal earthquakes in the western part of the United States, providing a common ground motion record database, namely, the NGA-west database. The NGA-west2 database is an updated version of the previous generation. The "flatfiles" of the NGA-West2 database used in the development of the GMPEs are publicly available on the PEER website at http://peer.berkeley.edu/ngawest2/databases/ (Bozorgnia et al., 2014). The NGA-west2 flatfile is considered to be a good quality data in the field of strong motion observations. The raw data include 599 earthquakes and 21,539 strong motion data. Table 1 shows the available parameters and screened out from the NGA-west2 database.

Table 1. Seismic parameters in NGA-west2 database

Parameters	Unit	Parameters	Unit
Magnitude	M_w	Fault length	km
Strike	degree	Epicenter distance	km
Dip	degree	Hypocenter distance	km
Rake	degree	Joyner-Boore distance	km
Hypocenter latitude	degree	Closest distance	km
Hypocenter longitude	degree	Vs30	m/s
Hypocenter depth	km	Station latitude	degree
Fault width	km	Station longitude	degree

1.2 Data preprocessing and descriptive statistics

The types of earthquake magnitudes used in this study are moment magnitudes. Characteristic information such as magnitude, dip angle, strike, slip angle, location and focal depth should not be null. We cleaned the raw data based on above two criteria. Finally, the cleaned data contain 305 earthquakes and 12,892 strong motion data records, which are all used for model construction and verification. The geographical locations of these earthquakes are shown in Fig. 1. They are mainly distributed along two main seismic belts, i.e., the circum-Pacific seismic belt and the Mediterranean-Himalayan seismic belt. The magnitudes of 305 historical earthquake cases range between M_w 3.2 and M_w 7.9.

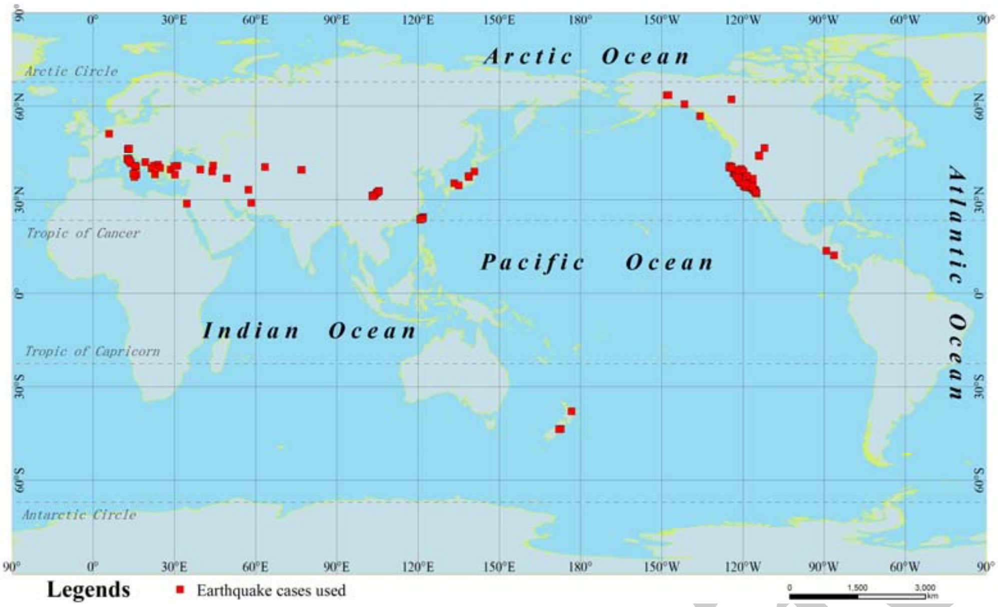

Figure 1. Distribution map of earthquake cases used in model construction

According to previous research, natural logarithm processing of peak ground motion parameters is performed. This needs to be done due to the large scale of variation of the original seismic peak parameters and the huge magnitude difference between their maximum and minimum values. Logarithm operations do not change the nature and correlation of data but rather compress the scale of variables and make the data more stable. We selected 16 characteristics strongly related to the ground motion parameters from the NGA-west2 database, and expressed their mathematical relationship as follows:

$$\begin{Bmatrix} In(PGA) \\ In(PGV) \end{Bmatrix} = \mathrm{f}(Mag, Strike, Dip, Rake, Lat, Lon, Depth, \\ Length, Width, EpiD, HypD, R_{jb}, ClstD, V_{s30}, Lat', Lon'), \quad (1)$$

where *Mag* is the magnitude of the main shock, *Strike* is the strike of the main earthquake fault, *Dip* is the dip angle of the fault, *Rake* is the slip angle of the fault, *Lon* and *Lat* are the longitude and latitude of the epicenter, *Depth* is the focal depth, *Length* and *Width* are the length and width of the main fault plane, respectively, *EpiD* is the epicenter distance, *HypD* is the focal distance, R_{jb} is the Joyner-Boore distance, *ClstD* is the fault distance, V_{s30} is the underground 30 m shear wave velocity, and *Lon'* and *Lat'* are the longitude and latitude of the target point, respectively. The above features can be classified into four major categories, as shown in Tab. 2, which are related to the focal source, fault, site and target points. Furthermore, these features are classified into subcategories and labeled to facilitate subsequent experimental discussion.

Table 2. Feature categories related to ground motion parameters

Category		Label	Features
Major	Sub		
Focal source	Basic	*FOB*	*Mag, Depth,*
	Location	*FOL*	*Lat, Lon*
	Distance	*FOD*	*EpiD, HypD*
Fault	Basic	*FAB*	*Strike, Dip, Rake,*
	Scale	*FAS*	*Length, Width*
	Distance	*FAD*	R_{jb}, *ClstD*

Site	/	*S*	*Vs₃₀*
Target	Location	*TL*	*Lat', Lon'*

2 METHODOLOGY

2.1 Evaluation indicators

The four evaluation indicators used in the article include the coefficient of determination (R^2) (Nagelkerke, 1991), mean absolute error (MAE), mean absolute percent error (MAPE) and root mean squared error (RMSE). These indicators are calculated by the following formulas in the testing data sets:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(x_i - y_i)^2}{\sum(x_i - \bar{x})^2}, \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{N}(|x_i - y_i|)}{N}, \quad (3)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\left|\frac{x_i - y_i}{x_i}\right|\right), \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N}\sum(x_i - y_i)^2}, \quad (5)$$

where N, x_i, y_i, and $\bar{x}$ are the total number of data, the observation value, the prediction value and the average value of observation values, respectively. SS_{res} represents the sum of residuals, and SS_{tot} represents the total sum of squares. It is known that the square error represents the dispersion degree of the numerical value, and the larger the value is, the more discrete it is. Our goal is to adopt an indicator that can measure the quality of regression fitting, which is not affected by numerical discreteness. We avoid the affect by "division" (SS_{res}/SS_{tot}).

2.2 LightGBM algorithm

LightGBM is a fast, high-performance gradient boosting framework based on the decision tree algorithm developed by Microsoft Research Asia, which can be used in classification and regression tasks. The algorithm uses a histogram optimization algorithm and a leaf-wise algorithm with depth limitations.

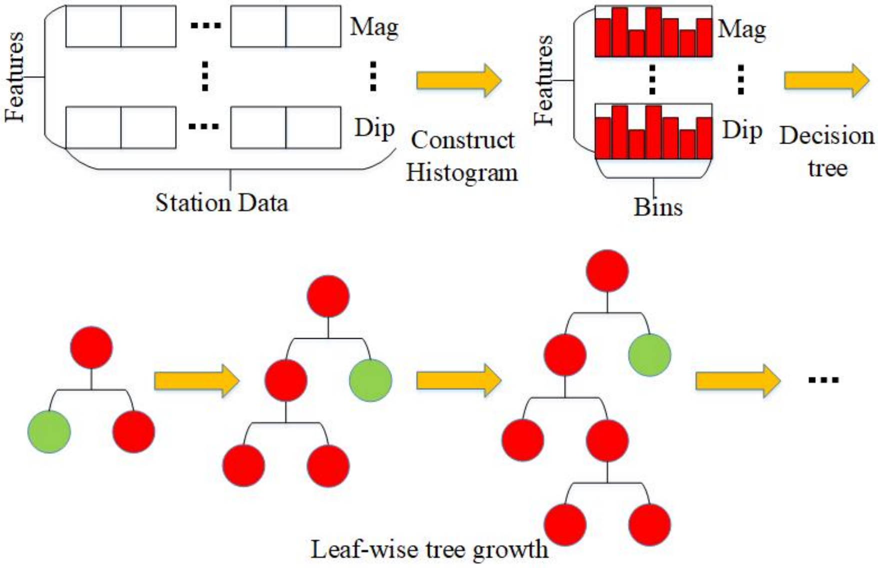

Figure 2. Histogram-based decision tree algorithm. The upper part of the figure is the construction process of each feature histogram, and the lower part indicates the process of leaf-wise tree growth.

Boosting algorithms are based on tree models, such as XGBoost, using a presorting algorithm for feature selection and tree splitting. The presorting algorithm computationally consumes many resources. As shown in Fig. 2, LightGBM uses a histogram algorithm to discretize continuous floating-point features into k discrete values and construct a histogram with a width of k. Then all training data are traversed, and the cumulative statistics of each discrete value in the histogram are counted. It is necessary to traverse the histogram to find the optimal segmentation point based on the discrete values of the histogram when selecting features. The use of histogram algorithm will greatly save time because the number of bins is much less than the number of station data points. The histogram will also have a regularization effect, effectively preventing the model from overfitting and improving the accuracy of the model.

LightGBM uses a leaf-wise growth strategy with depth restrictions instead of the level-wise growth strategy currently used by most GBDT algorithms. Based on the level-wise strategy, the leaves within the same layer can be split at the same time to control the model's complexity and make the model difficult to overfit. Level-wise splitting is an inefficient strategy because it treats the leaves within the same layer indiscriminately, which could introduce many unnecessary calculations. In fact, many leaves do not need to be searched and split because the split gain is low. Leaf-wise is a more efficient strategy. Each split identifies the leaf with the largest split gain from all the current leaves. For each tree node, the information gained after splitting can be expressed by the change in entropy as follows:

$$G(D,A) = Entroy(before) - Entroy(after) = \sum_{i=1}^{N} -p_i log_2 p_i - \sum_{a\in Values(A)} \frac{|D_a|}{D} En(D_a), \qquad (6)$$

where $Entroy(before)$ represents the information entropy of collection D, p_i is the ratio of the number of categories i to the total number of D, N is the number of categories, a is the value of feature A, and D_a is a subset of collection D.

Compared with the level-wise strategy, the leaf-wise strategy can greatly reduce errors and obtain better accuracy when the number of splits is the same. The leaf-wise strategy may lead to a deeper decision tree and thus overfit the model. Therefore, LightGBM adds a maximum depth limit on the basis of the leaf-wise strategy to prevent overfitting while ensuring a high efficiency.

2.3 Geographic feature coding

From the ground motion attenuation equation, the strong motion parameters are closely related to the magnitude factor and the distance factor. Theoretically, the longitude and latitude of the epicenter combined with the longitude and latitude of the target point can reflect the variation in strong motion parameters with distance. In the prediction of ground motion parameters, the longitude and latitude position of the target point has not been used in previous studies. However, the direct use of longitude and latitude as features will cause some geographic problems. Tobler (1970) once proposed the law of spatial correlation, which stated that "all things are related, but near things are more related than distance things"; this is also the first law of geography. From the point of view of geography, if a location is close in space, the corresponding ground motion parameters will be similar. As shown in Fig. 3, after the transformation of geographical features, the geographical feature space shifts from being two-dimensional to three-dimensional. In fact, the points in the yellow circle are geographically closer to the points in the blue circle and farther away from the points in the green circle. If there is no transformation, the point in the yellow circle is closer to the point in the green circle in the feature space, which is inconsistent with reality.

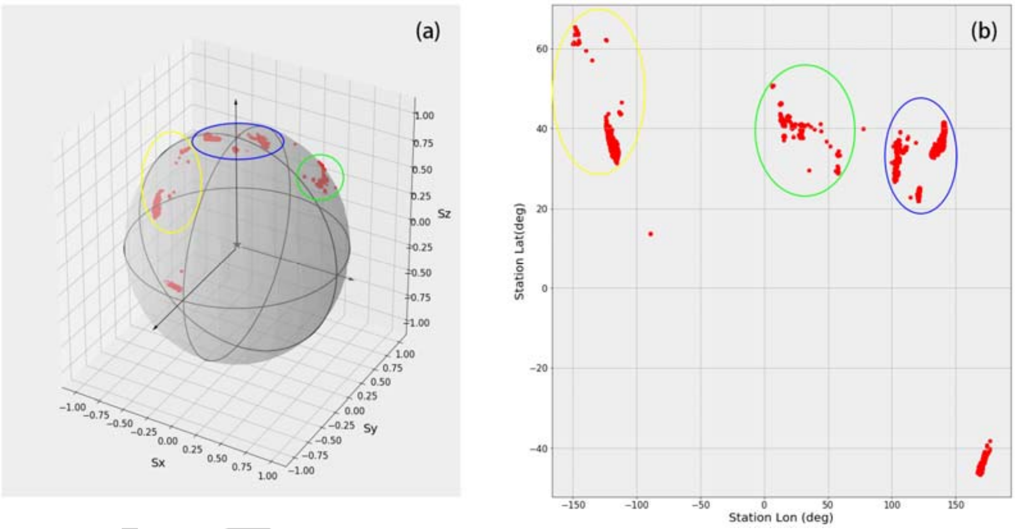

Figure 3. The feature space before and after the transformation of geographical features. Subplot (a) shows the distribution of earthquake cases before the feature transformation in the spherical coordinate system, and subplot (b) shows the distribution of earthquake cases in the plane coordinate system.

Therefore, we need to convert latitude and longitude into Cartesian coordinates. Suppose that the Cartesian coordinates take the center of the earth as the origin, the direction from the origin to the North Pole is the positive direction of the z-axis, and the direction from the origin to the point with both a longitude and latitude of 0 is the positive direction of the x-axis. The process of geographic feature coding (abbreviated as geocoding) can be expressed as follows:

$$\begin{cases} x = Rcos(Lat)\, cos(Lon) \\ y = Rcos(Lat)\, sin(Lon), \\ \quad z = Rsin(Lat) \end{cases} \tag{7}$$

where Lat is the latitude, Lon is the longitude, and R is the radius of the Earth, which is assumed to be 1. After geographic feature coding, the location of the epicenter is represented by the spatial coordinates of (H_x, H_y, H_z) instead of its longitude and latitude, and the location of the target point is represented by the spatial coordinates of (S_x, S_y, S_z). Therefore, the features affecting the ground motion parameters can be expressed as follows:

$$\begin{Bmatrix} In(PGA) \\ In(PGV) \end{Bmatrix} = \mathrm{f}(Mag, Strike, Dip, Rake, H_x, H_y, H_z, Depth,$$

$$Length, Width, EpiD, HypD, R_{jb}, ClstD, V_{s30}, S_x, S_y, S_z). \tag{8}$$

There are 18 features related to ground motion parameters after geocoding.

Table 3. Performance of the predictive model for PGA and PGV before and after geographic feature coding

ID	Features	PGA				PGV			
		R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
1	*FOB, FOL(lat,lon), FOD, FAB, FAS, FAD, S, TL(lat', lon')*	0.965	0.358	0.119	0.479	0.973	0.337	0.995	0.453
2	*FOB, FOL(H_x, H_y, H_z), FOD, FAB, FAS, FAD, S, TL(S_x, S_y, S_z)*	**0.966**	**0.353**	**0.118**	**0.473**	**0.973**	**0.333**	**0.956**	**0.449**

We used a set of control experiments to demonstrate that the transformation of geographical features is effective for improving the model's predictive ability. The different features used in experiment 1 and experiment 2 only include the locations of the epicenter and target points. Table 3 shows that each indicator in experiment 2 is better than that in experiment 1.

2.4 Feature selection

In this paper, we need to further select features before building LGB-FS. The following principles need to be clarified in the process of feature selection. (1) Features with a relatively great importance should be selected. (2) Features with a high correlation are selectively deleted. (3) Features that can be obtained quickly after the earthquake should be selected. All optimized feature selection is performed on the dataset after geographic feature coding. Based on these rules, the 18 features obtained by preliminary screening were optimized.

2.4.1 Feature importance and correlation

The importance of each feature was calculated. The algorithm used in the article is LightGBM, which is an improved gradient boosting machine. The importance of a feature is calculated based on the total information gain generated when it is used as a feature of tree splitting. For a single decision tree, the feature importance can be defined as follows:

$$SFI(X_i, T) = \sum_{t \in T} \Delta I(X_i, t), \tag{9}$$

where $\Delta I(X_i, t)$ represents a reduction in impurity due to the split on feature X_i at node t of tree T (Breiman et al., 1984). The node impurity $I(t)$ for the regression can be defined as follows:

$$I(t) = \sum_{i \in t} \frac{(y_i - \bar{y})^2}{N(t)}, \tag{10}$$

where y_i represents observation i in node *t*, $\bar{y}$ is the mean of all observations in node *t*, and *N(t)* is the number of observations in node *t*. For the global importance of a feature, each tree needs to be considered, so the average of all trees is taken as the global importance of the feature, which can be defined as follows:

$$FI(X_i) = \frac{1}{M}\sum_{m=1}^{M} SFI(X_i, T_m), \tag{11}$$

where M is the number of trees and T_m is the m^{th} tree (Friedman, 2001; Tuv et al., 2009). After normalization of these gains, the order of feature importance is obtained as shown in Fig. 4(a). The most important features are the magnitude, fault width, Joyner-Boore distance, closest distance to the fault, epicenter distance, etc.

While building the model, we considered possible strong correlation between the features; otherwise, it may make the accurate estimation of the model's weight parameters difficult. If there is a strong linear relationship between the features, their effects on the dependent variable will be indistinguishable thus the results cannot be clearly explained, which reduces the interpretability of the black-box model. We used the Pearson correlation coefficient as the correlation coefficient between features, which can be defined as follows:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}, \tag{12}$$

where X and Y represent two features, σ_X and σ_Y represent the standard deviation of two features, $\bar{X}$ and $\bar{Y}$ are the mean values of two features, respectively, and i represents the i^{th} data point in the dataset. The value range of the correlation coefficient is [-1, 1]. According to Fig. 4(b), some features have high correlation, especially several distance parameters (*EpiD, HypD, R_{jb}, ClstD*). These distance parameters are related to either the focal source or fault. The difference between *EpiD* and R_{jb} in the near field is relatively large, and the difference decreases with increasing distance, depending on the scale of the fault (*Length* and *Width*) and the relative position of the epicenter on the fault. The abovementioned difference will also affect the estimation of ground motion parameters (Wen et al., 2017; Yenier et al., 2008). To eliminate this effect as much as possible, we consider the features not only related to the fault but also related to the focal source when selecting features. According to the feature importance map and feature correlation analysis, we can roughly understand those features that have a greater impact on the ground motion and the correlation between the features. However, how each feature independently affects the ground motion parameters needs to be further explored.

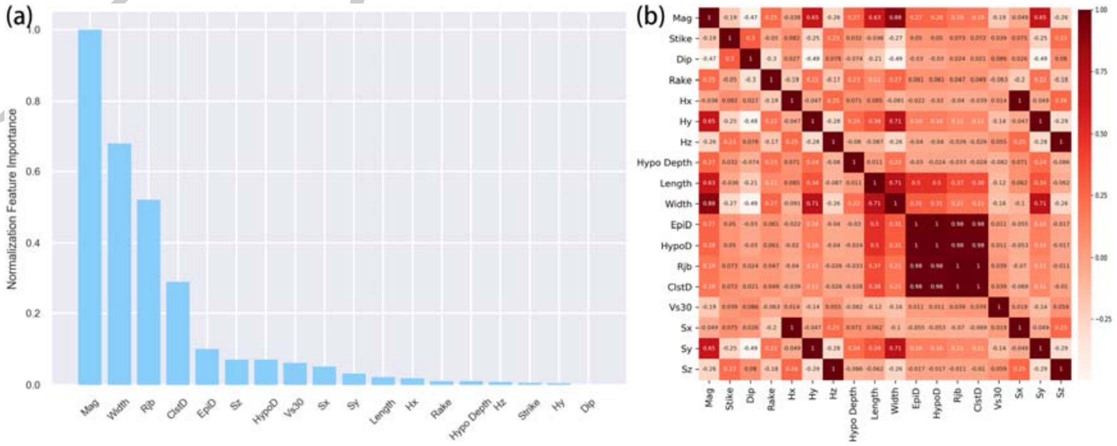

Figure 4. Normalized feature importance and feature correlation graph. The greater the correlation between two features is, the darker the color of the corresponding square. The correlation coefficients between features are also marked.

2.4.2 Nonlinear dependence between features and predictions

The partial dependency plot (PDP) shows the marginal effect of a feature on the predicted output of a previously fitted model (Friedman, 2001). A partial dependency plot can indicate whether the relationship between targets and features is linear, monotonic, or more complex (Molnar, 2019). The partial dependency function is defined as follows:

$$\widehat{f_{x_S}}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)\, dP(x_C), \tag{13}$$

where x_S is the feature selected for the partial dependency analysis, x_C are the other features, and $\hat{f}$ is the fitted model. For a well-fitted LightGBM model, if x_S is the magnitude, x_C are the other 17 features. According to Fig.4, magnitude is the feature that has the most significant impact on the prediction results of all features. As the magnitude increases, the predicted ground motion parameters increase monotonically, and the trend is approximately linear. Several other features that contribute significantly to the ground motion parameters are R_{jb}, *ClstD*, and the site effect parameter Vs_{30}. These features are negatively correlated with the ground motion parameters. The influence of the fault scale parameters on ground motion depends mainly on the fault width. The fault width is mainly distributed in areas with small values. When the fault width increases to a certain limit, the ground motion parameters will increase considerably. In addition, the location of the target points in the ground motion field also affects the ground motion parameters within in a certain value range.

Based on the analysis above, it can be concluded that the characteristics of magnitude, distance, site effect, and fault scale play the most important role in the prediction of ground motion parameters. It is also showed that several distance parameters have high correlation coefficients with each other. Correspondingly, Fig. 5 shows that the change trends of these parameters are generally consistent and negatively correlated with the ground motion parameters. Therefore, these two phenomena can also be mutually explained. Through the analysis of the PDP, we can disassemble the black box model and simplify the relatively complex nonlinear relationship, which can better reveal the influence of each variable on the dependent variable and accordingly improve the interpretability of the model.

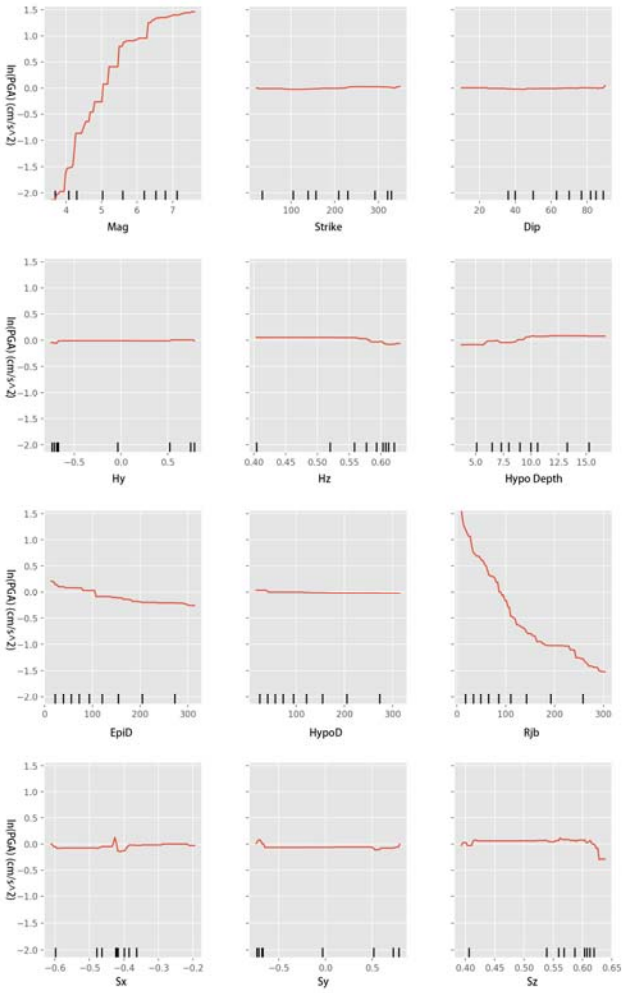

Figure 5. Partial dependence plot for the constructed model. Black bars on the x-axis show the data distribution.

2.4.3 Control experiments

In sections 2.4.1 and 2.4.2, the influence of feature importance, feature relationships and feature dependence on the prediction results were explored. Three sets of control experiments are designed, as shown in Tab. 4 - Tab. 6, to further confirm the features that can improve the performance of the models. These control experiments were carried out on the NGA-west2 dataset, and 1/4 of the data were randomly selected as the test data.

Table 4. Performance of the predictive models with different combinations of features (control variables are *FOD*, *FAD* and *FAS*)

ID	Features	PGA				PGV			
		R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
1	*FOB(Mag),S(Vs_{30}),* *FAB(Rake) FOD(EpiD)*	0.941	0.477	0.153	0.622	0.955	0.447	1.208	0.581
2	*FOB(Mag), S(Vs_{30}),* *FAB(Rake), FOD(HypoD)*	0.940	0.482	0.157	0.629	0.955	0.443	1.159	0.579
3	*FOB(Mag), S(Vs_{30}),*	0.943	0.469	0.146	0.614	0.957	0.434	1.139	0.566

	FAB(Rake),FAD(ClstD)								
4	*FOB(Mag),S(Vs_{30}), FAB(Rake), FAD(R_{jb})*	0.943	0.468	0.144	0.614	0.957	0.435	**1.084**	0.565
5	*FOB(Mag), S(Vs_{30}), FAB(Rake), FAD(R_{jb}),FOD(EpiD)*	0.944	0.466	0.142	0.609	0.958	0.432	1.179	0.562
6	*FOB(Mag), S(Vs_{30}), FAB(Rake), FAD(R_{jb}),FOD(EpiD), FAS(Width, Length)*	**0.945**	**0.463**	**0.141**	**0.605**	**0.958**	**0.428**	1.138	**0.558**

As shown in Tab. 4, the control variables in this set of experiments are *FOD*, *FAD* and *FAS*. The feature combinations in experiments 3 and 4 have been adopted by many scholars in previous studies (Ambraseys et al., 1996; Douglas, 2003; Campbell and Bozorgnia, 2008). Experiments 1-4 reflect the influence of several distance parameters on the results. Among them, the distance parameters related to the fault are better than the distance parameters related to the focal source. Experiments 1, 4, and 5 reflect that increasing the distance parameters will also improve the performance of the models.

Table 5. Performance of the predictive models with different combinations of features (control variables are *FOL*, *TL*, and *FAB*)

ID	Features	PGA				PGV			
		R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
1	*FOB(Mag), S(Vs_{30}), FAD(R_{jb}), FOD(EpiD), FAS(Width, Length)*	0.945	0.463	0.141	0.605	0.958	0.432	1.128	0.561
2	*FOB(Mag), S(Vs_{30}), FAD(R_{jb}), FOD(EpiD), FAS(Width, Length), FOL(lat, lon), TL(lat', lon'), FAB(Rake)*	**0.965**	**0.357**	**0.121**	0.479	0.973	0.336	1.015	0.453
3	*FOB(Mag), S(Vs_{30}), FAD(R_{jb}), FOD(EpiD), FAS(Width, Length), FOL(lat, lon), TL(lat', lon')*	**0.965**	0.358	**0.121**	**0.478**	**0.974**	**0.328**	**0.978**	**0.444**

The control variables of the experiments shown in Tab. 5 are *FOL*, *TL* and *FAB*. The result of experiment 3 is significantly better than that of experiment 1, indicating that the addition of location information will greatly improve the performance of the models. Compared with experiment 3, experiment 2 uses *Rake*, but its performance is worse.

Table 6. Performance of the predictive models with different combinations of features (after geocoding, the control variables are *FAD* and *FAS*)

ID	Features	PGA				PGV			
		R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
1	*FOB, FOL, FOD, FAB, FAS, FAD, S, TL*	**0.966**	0.353	0.118	0.473	0.973	0.333	0.956	0.449
2	*FOB(Mag), FOD(EpiD), FOL(H_x, H_y, H_z), S(Vs_{30}), TL(S_x, S_y, S_z)*	**0.966**	**0.349**	**0.117**	**0.472**	0.973	0.330	0.960	0.448
3	*FOB(Mag), FOL(H_x, H_y, H_z), FAD(R_{jb}), FOD(EpiD), S(Vs_{30}), TL(S_x, S_y, S_z)*	**0.966**	0.350	0.119	**0.472**	**0.974**	**0.326**	**0.869**	**0.441**
4	*FOB(Mag), FOL(H_x, H_y, H_z), FOD(EpiD), FAS(Width, Length), FAD(R_{jb}), S(Vs_{30}), TL(S_x, S_y, S_z)*	**0.966**	0.350	0.119	**0.472**	**0.974**	0.327	0.901	0.442

The features of the experiments in Tab. 6 have been geocoded, and the control variables are FAD and FAS. The overall performance is better than the results in Tabs. 4 and 5. All features were used in experiment 1, but the results were not as good as the results of experiments 2-4. It is not the case that the inclusion of more features yielded better results. The use of too many features may negatively affect the results. Compared with experiment 3, experiment 4 used the *FAS* features, while experiment 2 did not use R_{jb}. The results of experiments 2 to 4 are similar, and the results of experiment 3 are the best.

When the number of features reaches 10, the performance of the model is best. The features used are *FOB(Mag)*, *FOL(H_x, H_y, H_z)*, *FOD(EpiD)*, *FAD(R_{jb})*, *S(Vs_{30})*, and *TL(S_x, S_y, S_z)*, as well as key information related to focal source, fault, site and target point characteristics.

In addition, the location of the epicenter, magnitude and target point can be obtained quickly after an earthquake. The size of the fault can be estimated according to previous research experience. However, it is difficult to guarantee the accuracy of the fault-scale features obtained in this way. Previous studies have shown that the shear wave velocity 30 m underground is often used to represent the site effect. Based on the terrain slope, Heath et al. (2020) launched the global Vs_{30} product with a spatial resolution of 1 km, which can be used as an input of LGB-FS in this study and provides the possibility to calculate ground motion parameters globally. Therefore, in the construction of the model, we finally selected the following 9 factors: *FOB(Mag)*, *FOL(H_x, H_y, H_z)*, *FOD(EpiD)*, *S(Vs_{30})*, and *TL(S_x, S_y, S_z)*. We also constructed two models based on the feature combination of experiment 2 and experiment 4 as comparisons. If accurate fault parameters and R_{jb} are obtained from the corresponding inversion after the earthquake, the model can be improved accordingly.

2.5 Model construction

The prediction model of strong motion parameters is based on the LightGBM algorithm. We randomly divided the original data into 11 equal parts, of which one was used as test data. Another 10 pieces of data were used to construct 10 submodels. Each submodel was constructed with 9 pieces of data as training data and validation data, and the other acted as the test data. The submodels obtained are independent of each other, and their accuracy are inconsistent, which means that there must be differences between them. It is difficult to determine which submodel should be selected as the final model. Therefore, we synthesize all the submodels and average the predictive ability of each model to obtain a comprehensive model. The idea of building the model utilizes the bagging idea of ensemble learning.

3 RESULTS AND DISCUSSION

3.1 Model fitting and testing from the perspective of feature optimization

Because the training of each submodel is independent, the comprehensive model deals with the prediction results of submodels, and it has no independent training process. Therefore, the training curves of each submodel are shown in Fig. 6. When the number of training iterations reaches 2000, the model almost reaches convergence.

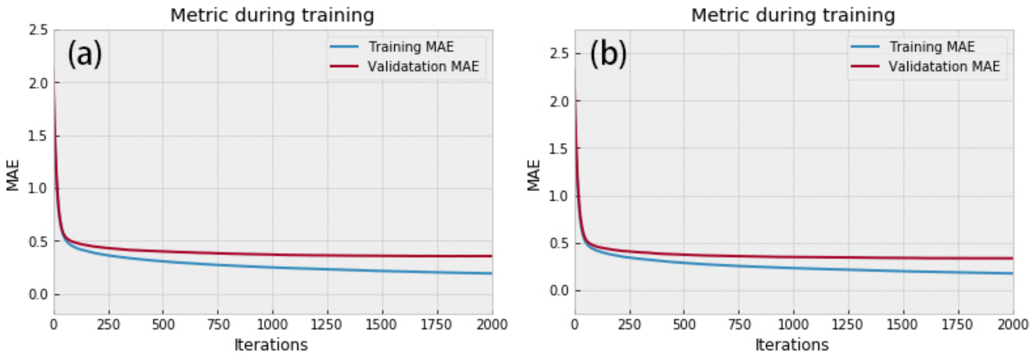

Figure 6. Training curves of the submodels. Subplot (a) is the training curve of the PGA submodel, and subplot (b) is the training curve of the PGV submodel.

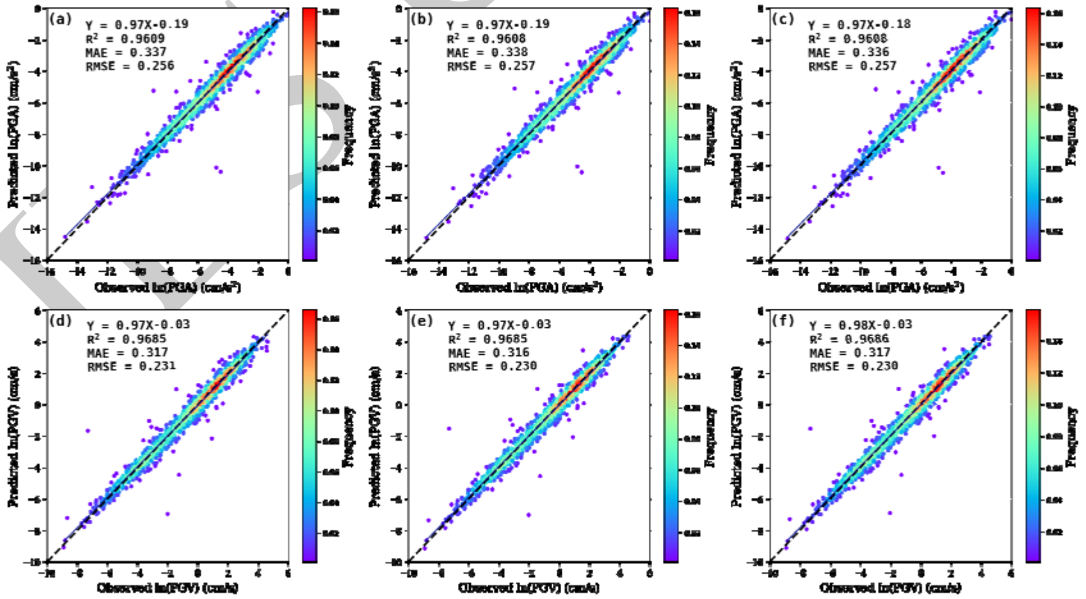

Figure 7. Prediction results of the model on the testing data. Subplots a, b, and c correspond to the PGA prediction results of the three kinds of models, respectively. Subplots d, e, and f correspond to the PGV prediction results of the three kinds of models, respectively. The three columns subplots from left to right

represent the results of the three feature combinations shown as experiments 2-4 in Tab. 6.

According to the exploration of feature selection in section 2.4, we have established three prediction models with different feature combinations, and their feature settings correspond to previous control experiments 2-4 in Tab. 6. The feature settings of the three types models are (1) *FOB(Mag), S(Vs_{30}), FOD(EpiD), FOL(H_x, H_y, H_z),* and *TL(S_x, S_y, S_z)*; (2) *FOB(Mag), S(Vs_{30}), FAD(R_{jb}), FOD(EpiD), FOL(H_x, H_y, H_z),* and *TL(S_x, S_y, S_z)*; and (3) *FOB(Mag), S(Vs_{30}), FAD(R_{jb}), FOD(EpiD), FAS(Width, Length), FOL(H_x, H_y, H_z),* and *TL(S_x, S_y, S_z)*, respectively.

The predicted values of ln(PGA) and ln(PGV) were obtained by estimating the testing data with the constructed prediction model. Regression analysis of the predicted value and the original value shows that the regression curves are nearly straight lines. The R^2 values of the (a)-(c) and (d)-(f) subplots are approximately 0.96 and 0.97, respectively. As shown in Fig. 7, a strong correlation is available between the estimated and actual values, which further manifests the validity of the proposed models. Using the predicted values of the ground motion parameters, their true values can be well represented and explained in the testing set. Most of the testing data were distributed along regression lines; however, individual data deviated considerably, which did not lead to a large impact on the overall predictive ability of the model. The reason for these predicted abnormal values is due to the abnormal data feature values, especially the important magnitude or distance in the feature. However, these outliers are correct and real, rather than errors caused by data entry or processing. The colors in Fig. 7 reflect the frequency distribution of the data. The slopes of the regression lines are all less than 1. Combining the regression lines with the line y=x, LGB-FS favors the underestimation of ground motion parameters.

3.2 LGB-FS versus other existing methods

There are many other methods to predict or estimate the ground motion parameters. Three main types of methods are considered for comparison with LGB-FS. They include regression learning methods, GMPEs and a revised interpolation scheme adopted by ShakeMap.

3.2.1 Regression learning methods

Several models have been developed for the estimation of principal ground motion parameters based on the NGA database that was released in 2003 referred to as NGA-west1. As mentioned before, the extension of this database, known as NGA-west2, is used for modeling in the current study. The new model will be compared with other models constructed based on the NGA-west1 database or NGA-west2 database by different methods, including artificial neural networking combined with simulated annealing (ANN/SA) (Alavi and Gandomi, 2011), genetic programming coupled with orthogonal least squares (GP/OLS) (Gandomi et al., 2011), multiexpression programming (MEP) (Alavi et al., 2011), genetic programming coupled with simulated annealing (GP/SA) (Mohammadnejad et al., 2012) and deep neural networks (DNNs) (Derakhshani and Foruzan, 2019).

Table 7. Performance comparison of the predictive models for PGA and PGV with other methods based on NGA-west1

Related study	Methods	PGA				PGV			
		R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE

Derakhshani and Foruzan (2019)	DNN	0.814	0.395	0.115	0.504	0.808	0.397	0.737	0.503
Alavi and Gandomi (2011)	ANN/SA	0.731	0.460	0.130	/	0.764	0.450	2.170	/
Gandomi et al. (2011)	GP/OLS	0.593	0.488	/	0.637	0.661	0.506	/	0.637
Alavi et al. (2011)	MEP	0.696	0.697	/	0.624	0.686	0.726	/	0.671
Mohammadnejad et al. (2012)	GP/SA	0.704	/	0.144	0.617	0.701	/	2.350	0.648
Our study	**LGB-FS**	**0.882**	**0.284**	**0.039**	**0.374**	**0.889**	**0.287**	**0.606**	**0.389**

Table 7 lists four evaluation indicators of different models based on NGA-west1. To compare with other methods, the feature combination we used includes *Mag, Vs_{30}, Rake, and ClstD*, similar to previous research. It can be seen that the LGB-FS is optimal on all evaluation indicators. PGA and PGV are predicted on the testing dataset by LGB-FS. The R^2 values of the predicted results are above 0.88, the MAE values are less than 0.3, and the RMSE values are less than 0.4. Table 8 lists four evaluation indicators of different models based on NGA-west2. LGB-FS still outperforms the DNN method on this data set. In this set of control experiments, the feature combination of LGB-FS is consistent with the DNN method, and a random 1/4 of the data set is used as the test data.

Table 8. Performance comparison of the predictive models for PGA and PGV with other methods based on NGA-west2

Related study	Methods	PGA				PGV			
		R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
Derakhshani and Foruzan (2019)	DNN	0.910	0.597	1.488	0.789	0.935	0.532	**0.807**	0.708
Our study	**LGB-FS**	**0.943**	**0.469**	**0.146**	**0.614**	**0.957**	**0.434**	1.139	**0.566**

3.2.2 GMPEs and ShakeMap

ShakeMap is a system that can quickly describe the range and distribution of strong ground motion after major global earthquakes. It is used by the USGS and NEIC to develop important global seismograms. The revised interpolation scheme is a weighted-average approach used to incorporate various types of data into the ShakeMap ground motion and intensity mapping framework (Worden et al., 2010). This approach represents a fundamental revision of the existing ShakeMap methodology. The approach allows the combination of direct observations (ground-motion measurements or reported intensities), the conversion of intensity observations to ground motion, and the estimation of ground motions and intensities from prediction equations or numerical models. ShakeMap ground-motion and intensity estimates are an uncertainty-weighted combination of these various data and estimates.

ShakeMap uses GMPEs and monitoring data when estimating ground motion after an earthquake. GMPEs are used to estimate ground motion in some areas where there are no stations or where it is not possible to quickly obtain station data after an earthquake. We selected GMPEs, which were also developed based on the NGA-west2 database, to estimate ground motion. These GMPEs were constructed by different teams, including ASK14 (Abrahamson et al., 2014), BSSA14 (Boore et al., 2013), CB14 (Campbell and Bozorgnia, 2014), CY14 (Chious and Youngs, 2014) and I14 (Idriss. 2013). The scope of application of these GMPEs is different. When calculating ground motion, it is necessary to select the corresponding equations according to the characteristics of earthquake magnitude, source distance, etc.

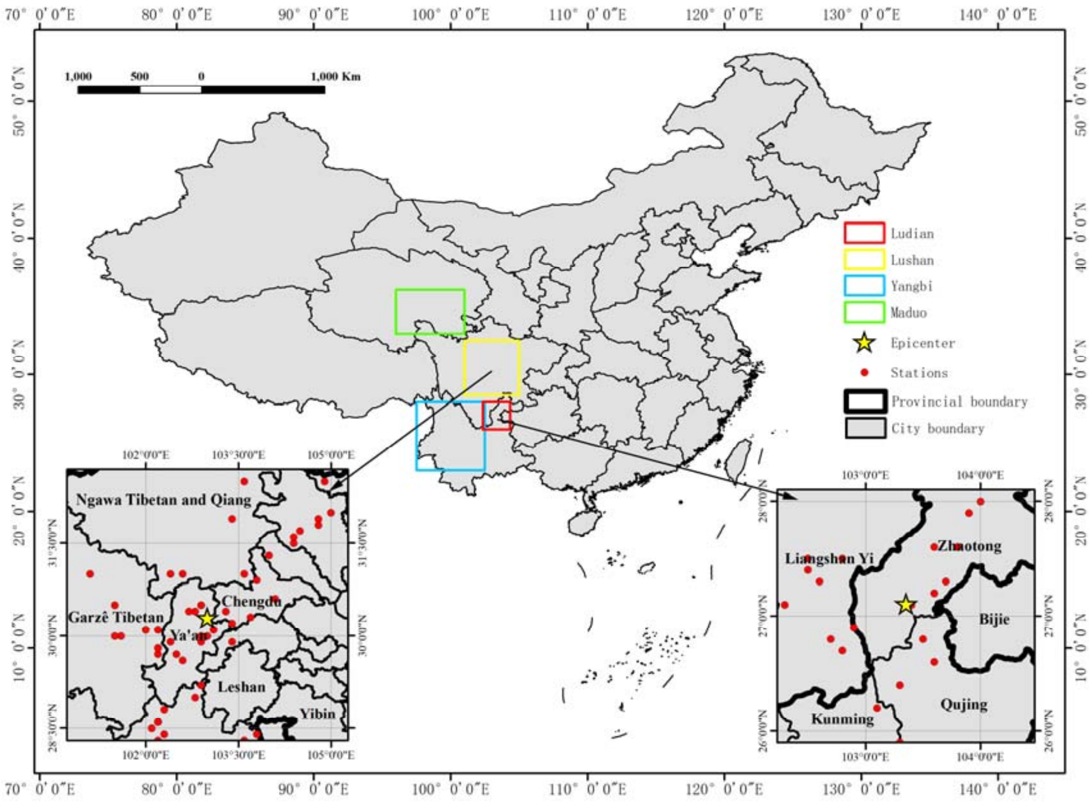

Figure 8. Two verified cases (Lushan earthquake and Ludian earthquake) and two cases to be analyzed (Yangbi earthquake and Maduo earthquake). The red dots represent the locations of monitoring stations around the epicenters, and their observational data are used as verification data for the model.

To compare and analyze LGB-FS and ShakeMap, two historical earthquake cases are selected for experiment. The two cases are the April 20th, 2013 M_w 6.6 earthquake in Lushan, China (abbreviated as Lushan earthquake) and the August 3rd, 2014 M_w 6.2 earthquake in Ludian, China (abbreviated as Ludian earthquake). The ShakeMap estimation results of the two earthquake cases were released by the USGS immediately after the earthquakes and were not revised by the observational data of the monitoring stations. Therefore, these two earthquake cases can be regarded as cases without observation data available. The two earthquake cases shown in Fig. 8 both occurred in China. The figure shows the locations of monitoring stations around the epicenters. The monitoring data of these stations are used as the verification data for all models. Although ShakeMap will revise the initial results based on the observation data collected later, a time interval is required between them. The time interval is often relatively large, which affects emergency rescue efforts after such earthquakes. When using LGB-FS to estimate the ground motion of the two earthquake cases, we used the first feature combination in section 3.1, including Mag, Vs_{30}, $EpiD$, H_x, H_y, H_z, S_x, S_y, and S_z. After an earthquake, accurate fault parameters are often not quickly obtained, which will inevitably affect the calculation of the corresponding fault distance. Therefore, the fault-scale parameters and fault distance are not considered in the model for comparison with ShakeMap. If there is technical support available to obtain accurate fault parameters immediately after an earthquake, we can further modify the model to improve its predictive performance.

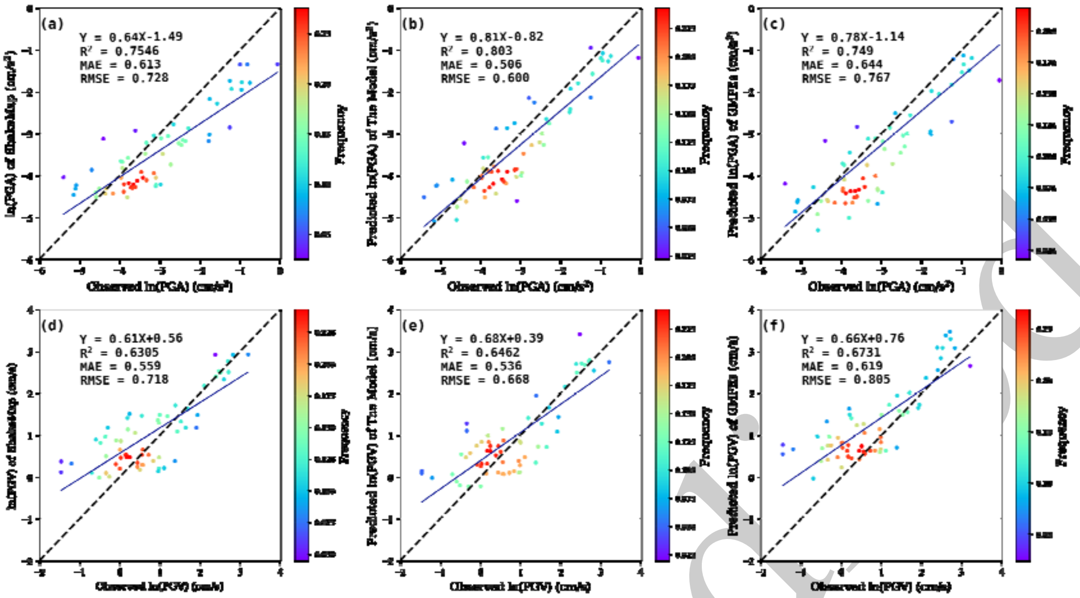

Figure 9. Model comparison of Lushan earthquake on observation stations. Subplots (a)-(c) are the PGA evaluation of ShakeMap, LightGBM and GMPEs, subplots (d)-(f) are the PGV evaluation of ShakeMap, LightGBM and GMPEs respectively. Different colors represent the frequency of test data points distributed around the corresponding observation or predicted value.

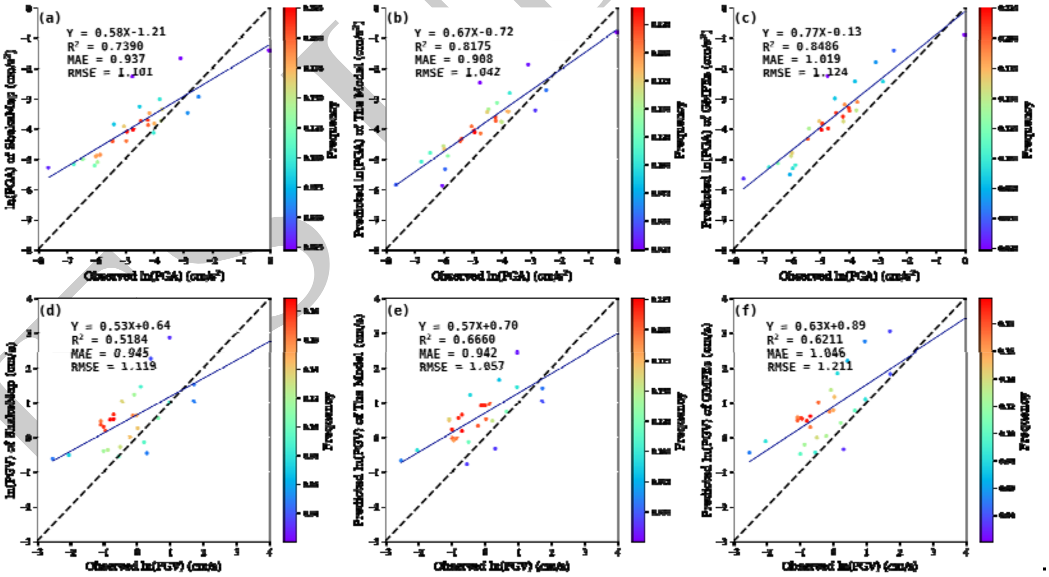

Figure 10. Model comparison of Ludian earthquake on observation stations. Subplots (a)-(c) are the PGA evaluation of ShakeMap, LightGBM and GMPEs, subplots (d)-(f) are the PGV evaluation of ShakeMap, LightGBM and GMPEs respectively. Different colors represent the frequency of test data points distributed around the corresponding observation or predicted value.

In the case of Lushan earthquake, a total of 62 sample points were selected to verify the model estimation results. The observation data of these sample points are from China Earthquake Administration. The ground motion distribution results of ShakeMap, LightGBM and GMPEs are calculated respectively, and the values of the sample points are extracted to calculate evaluation indicators further, which are used for prediction evaluation and eliminating the effect from the discrete verification points. As shown in Fig. 9, the MAE and RMSE of the LightGBM are smaller than those of ShakeMap and GMPEs, which indicates that the model in this paper is more applicable than ShakeMap in the area without observation data available.

Similarly, in the case of Ludian earthquake, a total of 29 sample points were selected to verify the model estimation results. MAE and RMSE were used to evaluate the results in Ludian earthquake. As shown in Fig. 10, for the ground motion parameters, the MAE and RMSE of the LightGBM are smaller than those of ShakeMap and GMPEs.

3.3 Model application in Yangbi and Maduo earthquakes

Through the verification of the model in Lushan and Ludian earthquake cases, it can be found that the machine learning model constructed is better than GMPEs and ShakeMap.

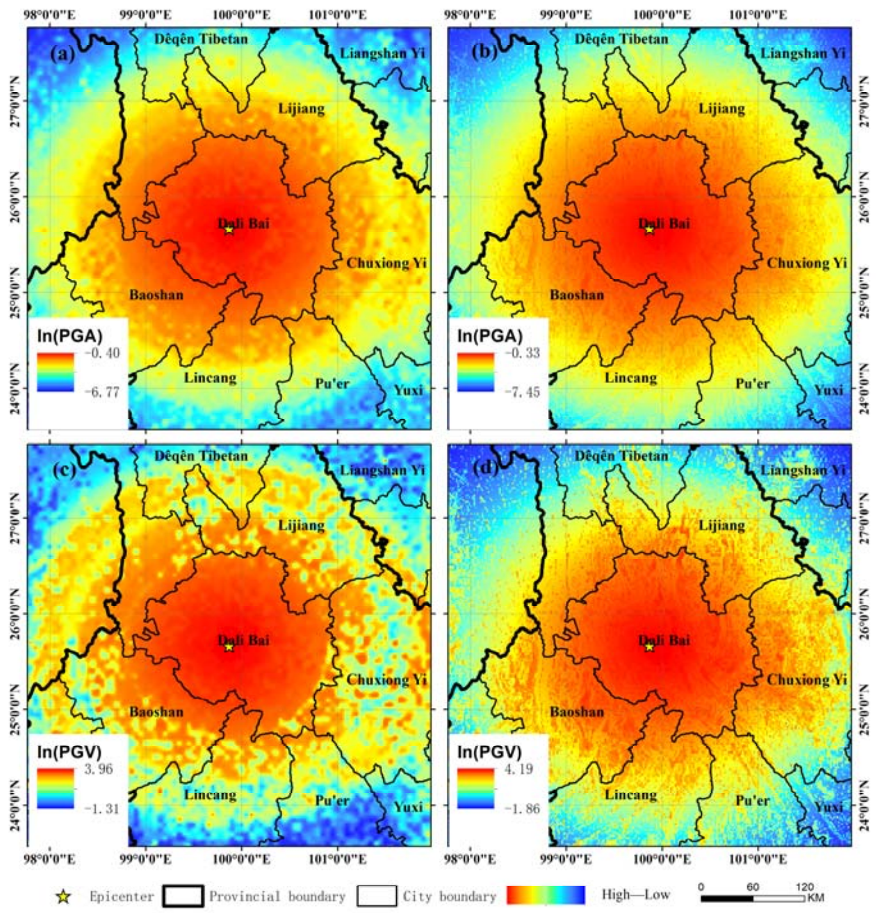

Figure 11. Ground motion parameters distribution of Yangbi earthquake. Subplots (a) and (b) represent the PGA distribution of LGB-FS and ShakeMap. Subplots (c) and (d) represent the PGV distribution of LGB-FS and ShakeMap.

The ground motion distribution of Yangbi earthquake and Maduo earthquake are predicted by using the established models. Since it is difficult to obtain fault related parameters quickly after the earthquake, the feature combination is still the same as before, including the following nine features: *FOB(Mag)*, *FOL(H_x, H_y, H_z)*, *FOD(EpiD)*, *S(Vs_{30})*, *TL(S_x, S_y, S_z)*. Since the data of the stations around the two cases were not obtained in time after the earthquake, we compared the prediction results with those of ShakeMap. As shown in Fig. 8, the four earthquake cases are all located in western China and are very close in geographical location. The model has achieved good results in verifying the earthquake cases, so we think that the model is also applicable to the Yangbi earthquake and the Maduo earthquake.

The distribution of PGA and PGV predicted by the LightGBM and ShakeMap are shown in Figs. 11 and 12. As shown in Fig. 11, we use the LightGBM to predict the ground motion near the epicenter of Yangbi earthquake. Then the prediction results of PGA and PGV models can be obtained and the corresponding ShakeMap results are shown respectively. Several ground motion attenuation equations were used in ShakeMap to form PGA and PGV distribution maps of Yangbi earthquake. The distance factor and site effect factor were considered in these attenuation equations. Therefore, it can be seen that the distribution maps formed do not only change with distance, but also present some irregular changes in some areas.

In the case of Maduo earthquake, ShakeMap used ground motion attenuation equations to calculate the ground motion parameters, and each equation was given weight. Distance and site effect factors are also considered in these attenuation equations. In the Yangbi earthquake, the distribution of ground motion mainly reflected its correlation with distance, and the site effect did not have a significant impact on the overall shape of its distribution. This can also roughly reflect that the underlying surface of the area only changes in individual locations rather than in a wide range.

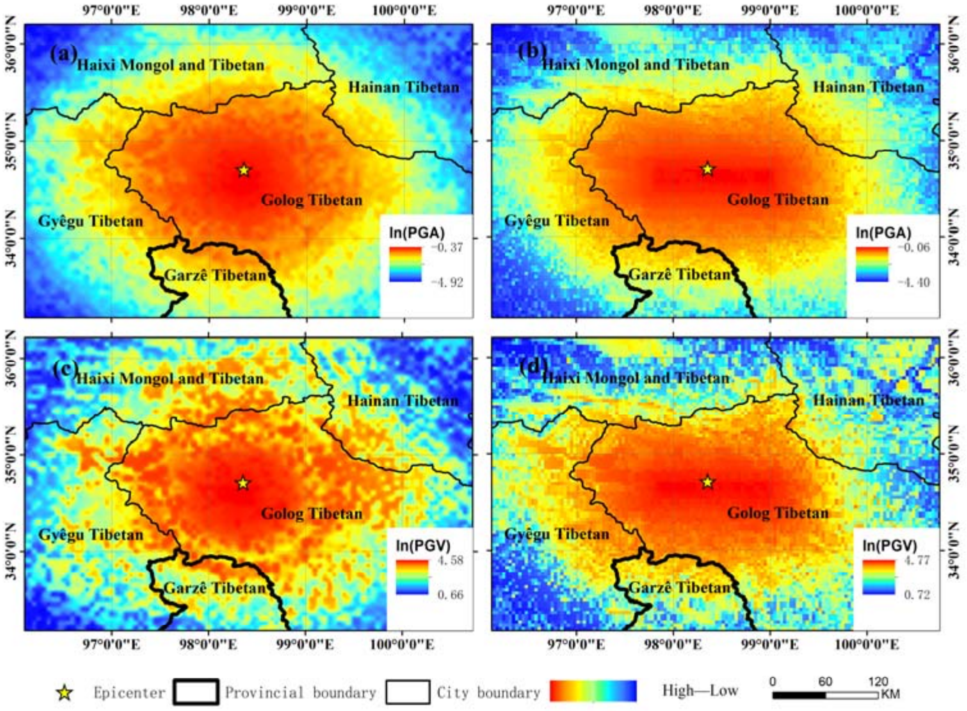

Figure 12. Ground motion parameters distribution of Maduo earthquake. Subplots (a) and (b) represent the PGA distribution of LGB-FS and ShakeMap. Subplots (c) and (d) represent the PGV distribution of LGB-FS and ShakeMap.

Table 9. Correlation analysis between model prediction results and ShakeMap results

	PGA		PGV	
	R^2	σ of residuals	R^2	σ of residuals
Yangbi	0.939	0.264	0.710	0.404
Maduo	0.871	0.300	0.586	0.447

It can be seen from Tab. 9 that the R-square between prediction results and ShakeMap results is larger in PGA distribution than in PGV distribution. The σ of residuals between prediction results and ShakeMap results is smaller in PGA distribution, which shows that the performance of PGA prediction model is more similar to that of ShakeMap. Similarly, the prediction effect of the model is more similar to that of ShakeMap in Yangbi earthquake.

The prediction model is a data-driven model and the distribution range of the data has a great influence on model transferability. Ground motion parameters are mainly functions of magnitude and distance. The magnitude of the training data is distributed between $M_w3.2$ to $M_w7.9$, and the R_{jb} ranges from 0 to 1500km, most of which are within 400km. Regardless of some abnormal values, most of the faults are less than 8.2km in width, and less than 10km in length. In summary, if the features of data are distributed within the above range, the transferability of the model should be better.

4 CONCLUSION

Aimed at resolving the shortcomings of the current methods of earthquake prediction and estimation, we propose a ground motion prediction model based on historical ground motion data and the LightGBM algorithm. Before the construction of the model, the features of the model are optimized and transformed, and then the model is verified by actual earthquake cases and be applied to Yangbi and Maduo earthquake. The following conclusions can be drawn:

(1) The magnitude, fault scale, site effect and distance parameters are the most important factors affecting ground motion. The transformation of the geographical features of the target points can further improve the prediction effect of the model. In addition, the use of a partial dependence plot (PDP) revealed the complex relationships between the various influences and ground shaking parameters for the first time, enabling us to understand the mechanisms and connections between them.

(2) Compared with other machine learning methods and GMPEs, the accuracy of the proposed LGB-FS model is best.

(3) Compared with the method based on real-time observation data (ShakeMap), the LGB-FS model has a better applicability in the region without observation data available. The model in this paper can be used as a supplement to other traditional models in some areas with fewer stations.

Due to the relative uncertainty of the fault parameters, the case validation in this paper does not use the fault-scale parameters, such as Joyner-Boore distance with the best performance in distance parameters. In future research or practical applications, if it is possible to obtain accurate fault parameters, the model can be revised and improved to achieve better results.

ACKNOWLEDGEMENTS

In this research, the results are obtained by programming in Python. The figures are performed using the ArcGIS software and Python. This research has been supported by the National Key R&D Program of China (grant no. 2017YFB0504104) and the National Natural Science Foundation of China (grant no. 41971280). The NGA-west2 data used in the article can be obtained from the following link as http://peer.berkeley.edu/ngawest2/databases/ (last access: 12 July 2021). The NGA-west2 ground motion prediction equations can be obtained from https://peer.berkeley.edu/research/data-sciences/databases (last access: 12 July 2021). The ShakeMap ground motion distribution data in the two earthquake cases are all from the earthquake module of USGS website at https://earthquake.usgs.gov/earthquakes/search/ (last access: 12 July 2021).

REFERENCES CITED

Abrahamson, N. A., Silva, W. J., 2008. Summary of the Abrahamson & Silva NGA ground motion relations. *Earthquake Spectra*, 24(1): 67-97. https://doi.org/10.1193/1.2924360

Abrahamson, N. A., Silva, W. J., Kamai, R., 2014. Summary of the ASK14 ground-motion relation for active crustal regions. *Earthquake Spectra*, 30(3): 1025-1055. https://doi.org/10.1193/070913EQS198M

Akkar, S., Sandikkaya, M. A., Bommer, J., 2014. Empirical ground-motion models for point and extended-source crustal earthquake scenarios in Europe and the Middle East. *Bulletin of Earthquake Engineering*, 12: 389-390. https://doi.org/10.1007/s10518-013-9461-4

Alavi, A. H., Gandomi, A. H., 2011. Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing. *Computer & Structures*, 89(23-24): 2176-2194. https://doi.org/10.1016/j.compstruc.2011.08.019

Alavi, A. H., Gandomi, A. H., Modaresnezhad, M., et al., 2011. New ground-motion prediction equations using multi expression programing. *Journal of Earthquake Engineering*, 15(4): 511-536. https://doi.org/10.1080/13632469.2010.526752

Ambraseys, N. N., Douglas, J., 2003. Near-field horizontal and vertical earthquake ground motions. *Soil Dynamics and Earthquake Engineering*, 23: 1–18. https://doi.org/10.1016/S0267-7261(02)00153-7

Ambraseys, N. N., Simpson, K. A., Bommer, J., 1996. Prediction of horizontal response spectra in Europe. *Earthquake Engineering & Structural Dynamics*, 25(4): 371–400. https://doi.org/10.1002/(SICI)1096-9845(199604)25:4<371::AID-EQE550>3.0.CO;2-A

Aptikaev, F., Kopnichev, J., 1980. Correlation between seismic vibration parameters and type of faulting. In: Proceedings of Seventh World Conference on Earthquake Engineering. 8-13 September, 1980. Istanbul.

Boore, D. M., Stewart, J. P., Seyhan, E., et al., 2013. NGA-West2 equations for predicting response spectral accelerations for shallow crustal earthquakes. In: PEER Report No. 2013. Pacific Earthquake Engineering Research Center, University of California, Berkeley, California.

Bozorgnia, Y., Abrahamson, N. A., Atik, L. A., et al., 2014. Nga-west2 research project. *Earthquake Spectra*, 131(3): 409-444.

Breiman, L., Friedman, J., Olshen, R., et al., 1984. Classification and Regression Trees (CART). *Biometrics*, 40(3): 358. https://doi.org/10.2307/2530946

Campbell, K. W., 1985. Strong motion attenuation relations: a ten-year perspective. *Earthquake Spectra*, 1(4): 759–804. https://doi.org/10.1193/1.1585292

Campbell, K.W., Bozorgnia, Y., 2008. NGA ground motion model for the geometric mean horizontal component of PGA, PGV, PGD and 5% damped linear elastic response spectra for periods ranging from 0.01 to 10 s. *Earthquake Spectra*, 24(1): 139–171. https://doi.org/10.1193/1.2857546

Campbell, K. W., Bozorgnia, Y., 2014. NGA-West2 ground motion model for the average Horizontal components of PGA, PGV, and 5%-damped linear Response Spectra. *Earthquake Spectra*, 30(3): 1087-1115. https://doi.org/10.1193/062913EQS175M

Chiou, B. S. J., Youngs, R. R., 2014. Update of the Chiou and Youngs NGA ground motion model for average horizontal component of peak ground motion and response spectra. *Earthquake Spectra*, 30(3): 1117-1153. https://doi.org/10.1193/072813EQS219M

Derakhshani, A., Foruzan, A. H., 2019. Predicting the principal strong ground motion parameters: a deep learning approach. *Applied Soft Computing*, 80: 192-201. https://doi.org/10.1016/j.asoc.2019.03.029

Derras, B., Bard, P. Y., Cotton, F., et al., 2012. Adapting the neural network approach to PGA prediction: an example based on the KIK-NET data. *Bulletin of the Seismological Society of America*, 102(4): 1446-1461. https://doi.org/10.1785/0120110088

Derras, B., Bard, P. Y., Cotton, F., 2014. Towards fully data driven ground-motion prediction models for Europe. *Bulletin of Earthquake Engineering*, 12(1): 495-516. https://doi.org/10.1007/s10518-013-9481-0

Douglas, J., 2003. Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth Science Reviews*, 61(1-2): 43–104. https://doi.org/10.1016/S0012-8252(02)00112-5

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5): 1189–1232. https://doi.org/10.1214/aos/1013203451

Gandomi, A. H., Alavi, A. H., Mousavi, M., et al., 2011. A hybrid computational approach to derive new ground-motion prediction equations. *Engineering Application of Artificial Intelligence*, 24(4): 717-732. https://doi.org/10.1016/j.engappai.2011.01.005

Heath, D., Wald, D. J., Worden, C. B., et al., 2020. A Global Hybrid Vs$_{30}$ Map with a Topographic-Slope-Based Default and Regional Map Insets. *Earthquake Spectra*, 36(3): 1570-1584. https://doi.org/10.1177/8755293020911137

Idriss, I. M., 2013. NGA-West2 model for estimating average horizontal values of pseudo-absolute spectral accelerations generated by crustal earthquakes. In: PEER Report No. 2013. Pacific Earthquake Engineering Research Center, University of California, Berkeley, California.

Jafariavval, Y., Derakhshani, A., 2020. New formulae for capacity energy-based assessment of liquefaction triggering. *Marine Georesources & Geotechnology*, 38(2): 214-222. https://doi.org/10.1080/1064119X.2019.1566297

Kayabali, K., Beyaz, T., 2011. Strong motion attenuation relationship for Turkey-a different perspective. *Bulletin of Engineering Geology & the Environment*, 70: 467-481. https://doi.org/10.1007/s10064-010-0335-6

Ke, G.L., Meng, Q., Finley, T., et al., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: NIPS'17: Proceeding of the 31st International Conference on Neural Information Processing Systems. December 2017. New York.

Mohammadnejad, A. K., Mousavi, S. M., Torabi, M., et al., 2012. Robust attenuation relations for peak time-domain parameters of strong ground motions. *Environmental Earth Science*, 67: 53–70. https://doi.org/10.1007/s12665-011-1479-9

Molnar, C., 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Leanpub.

Nagelkerke, N. J. D., 1991. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3). https://doi.org/10.1093/biomet/78.3.691

Sen, Z., 2011. Supervised fuzzy logic modeling for building earthquake hazard assessment. *Expert Systems with Application*, 38(12): 14564-14573. https://doi.org/10.1016/j.eswa.2011.05.026

Shiuly, A., Roy, N., Sahu, R. B., 2020. Prediction of peak ground acceleration for himalayan region using artificial neural network and genetic algorithm. *Arabian Journal of Geoscience*, 13(5): 1-10. https://doi.org/10.1007/s12517-020-5211-5

Sonia, T., Pillai, G. N., Pal, K., 2016. Prediction of peak ground acceleration using ϵ-svr, ν-svr and ls-svr algorithm. *Geomatics Natural Hazards & Risk*, 8(2): 1-17. https://doi.org/10.1080/19475705.2016.1176604

Tobler, W. R., 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46: 234–240.

Tuv, E., Borisov, A., Runger, G., et al., 2009. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10(3): 1341-1366. https://doi.org/10.1145/1577069.1755828

Wen, R., Xu, P., Y. Ren, P., et al., 2017. Development of the strong-motin Flatfile. *Earthquake Engineering and Engineering Dynamics*, 37(3): 38-47. (in Chinese with English abstract)

Worden, C. B., Wald, D. J., Allen, T. I., et al., 2010. A revised ground-motion and intensity interpolation scheme for shakemap. *Bulletin of the Seismological Society of America*, 100(6): 3083-3096. https://doi.org/10.1785/0120100101

Yenier, E., Erdoğan, Ö., Akkar, S., 2008. Empirical relationships for magnitude and source-to-site distance conversions using recently compiled Turkish strong-ground motion database. In: The 14th World Conference on Earthquake Engineering. October 12-17, 2008. Beijing.

Youngs, R. R., Day, S. M., Stevens, J. L., 1988. Near field ground motions on rock for large subduction earthquakes. In: Thun, J. L. V., ed., Earthquake Engineering and Soil Dynamics II: Recent Advances in Ground-Motion Evaluation. American Society of Civil Engineers, Reston. 445–462.

Youngs, R. R., Chiou, B., Silva, W. J., et al., 1997. Strong ground motion attenuation relationships for subduction zone earthquakes. *Seismological Research Letters*, 68(1): 58–73. https://doi.org/10.1785/gssrl.68.1.58