Machine Learning Approaches for Predicting Ground Motion Directionality Parameters

Victor H. Calderon* and Marcos Burgos†

I. Introduction

The Pacific Earthquake Engineering Research Center (PEER) has promoted the development of models to predict ground motion parameters for seismic design. One of these parameters is the maximum spectral response acceleration in all horizontal orientations ($Sa_{RotD100}$), which represents the observed maximum value of spectral acceleration in any direction at each period. Being the main intensity measure used in conventional structural design to compute design forces, accelerations, and displacements in structural and non-structural elements, according to Stewart et al.[1], by definition, $Sa_{RotD100}$ leads to conservative estimates of demands.

In the last two decades, the Probabilistic Seismic Hazard Assessment (PSHA) methodology has been implemented to characterize site-specific seismic hazards using ground motion models (GMMs). These GMMs predict the mean spectral acceleration spectral response acceleration overall horizontal orientations, denoted as Sa_{RotD50}. Thus, $Sa_{RotD100}$ being a more meaningful parameter for structural design, there have been many efforts to predict its median value that, for a specific period of vibration, is expressed as

$$\mu_{\ln Sa_{RotD100}} = \mu_{\ln(Sa_{RotD100}/Sa_{RotD50})} + \mu_{\ln Sa_{GM}}; \tag{1}$$

and its total variance that can be written as:

$$\sigma^2_{\text{tot},\ln Sa_{RotD100}} = \sigma^2_{\ln Sa_{GM}} \left(\frac{\sigma_{\ln Sa_{RotD100}}}{\sigma_{\ln Sa_{RotD50}}} \right)^2 + \sigma^2_{\ln Sa_{RotD100}/Sa_{RotD50}}, \tag{2}$$

where the median of ratio $\mu_{\ln(Sa_{RotD100}/Sa_{RotD50})}$, the multiplier $\sigma_{\ln Sa_{RotD100}}/\sigma_{lnSa_{RotD50}}$, and the standard deviation of $\ln(Sa_{RotD100}/Sa_{RotD50})$ were previously estimated using regression methods [2]; and $\mu_{\ln Sa_{GM}}$ and $\sigma_{\ln Sa_{GM}}$ were terms that can be estimated using GMMs, related to the *RotD50* orientation.

In the present work, we implement three machine learning models to predict the $Sa_{RotD100}/Sa_{RotD50}$ ratio, and the variance of $\ln(Sa_{RotD100})$, $\ln(Sa_{Rot50})$, and $\ln(Sa_{RotD100}/Sa_{RotD50})$. These models will be trained and tested using observed Sa_{RotD50} and $Sa_{RotD100}$ spectra contained in the PEER's NGA-West 2 database [3].

II. Related Work

Boore et al. [4] proposed $Sa_{RotD100}$ as an orientation-independent seismic intensity measure that links the seismic hazard and the structural response that captures the largest possible spectral amplitude across all directions. Before 2010, the predominantly used intensity measure in GMMs was $Sa_{GMRotDnn}$, the geometric means of spectral accelerations overall rotation angles. Thus, there are various studies have tried to estimate $Sa_{RotD100}$ using $Sa_{GMRotD50}$ [5, 6]. In addition, these studies used subsets of the NGA database, a previous version of the NGA West-2 database, to make their estimations.

In 2014, the PEER incentivized using $Sa_{RotD100}$ as a consistent intensity measure throughout the structural design process. Therefore, Shahi and Baker [2] compute the first estimation of the $Sa_{RotD100}/Sa_{RotD50}$ ratio and its total standard deviation using mixed effects logistic regression and a subset of 3000 seismic records previously selected by Abrahamson et al. to develop their GMM [7]. The results of this study were used in the new GMMs that are included in the 2023 National Seismic Hazard Model [8], which is currently the base for the multi-period response spectra of the ASCE 7-22 [9], that is the design standard for infrastructure across the US.

Even though the NGA West-2 database was released in 2013, it has yet to be formally employed in its entirety to estimate ground motion directionality parameters. The present study estimates some of the aforementioned parameters using ground motions contained in the NGA West-2 database alongside machine learning models, so it can provide an

*Graduate Student, Department of Civil and Environmental Engineering, Stanford University, vcalast@stanford.edu.

†Graduate Student, School of Electrical Engineering and Computer Science, University of Queensland, m.burgossaavedra@student.uq.edu.au

alternative for updating the factors employed by the modern GMMs to estimate $Sa_{RotD100}$.

III. Dataset and Features

The NGA-West2 database compiles seismic records from shallow crustal earthquakes in active tectonic regions, expanding on the earlier NGA-West database by incorporating data from global seismic events. It includes over 21,000 three-component records, covering a moment magnitude range of 3 to 7.9 and rupture distances between 0.05 and 1,533 km. The database provides spectral response data for 111 periods (0.01 to 20 seconds) and includes $Sa_{RotD100}$ and Sa_{RotD50} for most records.

To ensure data reliability, the cleaning process addressed inconsistencies and outliers, focusing on variables such as $Sa_{RotD100}$, Sa_{RotD50}, and the lowest usable frequency. Invalid entries (e.g., NaN values, −999, or zeros) were removed consistently across variables. The dataset was refined to 105 periods up to 10 seconds, emphasizing the engineering range. Outliers, defined as records with $Sa_{RotD100}/Sa_{RotD50}$ ratios outside 1 to $\sqrt{2}$, and records with source-to-site distances (R_{rup}) exceeding 200 km, were excluded to focus on near-field events relevant to engineering.

Frequency constraints were addressed by excluding spectral acceleration values for periods beyond the lowest usable frequency of 0.1 Hz. After this cleaning process, over 17,700 records remained, which were used to compute target variables by calculating the median and variance for each period strip (See Figure 1), resulting in 105 observations per target variable.

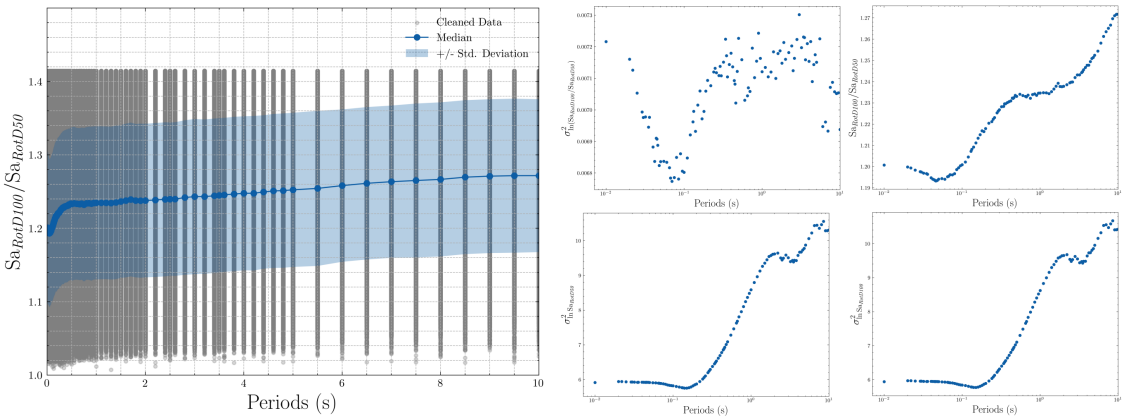

Fig. 1 Distribution of target variable observations across periods.

The findings of Shahi and Baker [10] show that the $Sa_{RotD100}/Sa_{RotD50}$ ratio is nearly independent of parameters such as the shear-wave velocity in the top 30 m (V_{S30}), the closest distance to rupture (R_{rup}), and earthquake magnitude. Although the study notes a slight dependency on R_{rup}, V_{S30} and magnitude exhibit negligible influence on this ratio. In this study, we assume the target variables are independent of V_{S30}, R_{rup}, and magnitude, leaving period as the sole feature variable. However, formal verification of this assumption remains pending.

IV. Methods

A. Objective Function and Performance Metric

Since this is a non-linear regression problem, our main objective is to minimize the Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,, \tag{3}$$

where y is the vector of target values, $\hat{y}$ is the vector of predicted values, and n is the number of examples.

We chose MSE as the loss function because it focuses on minimizing prediction errors and is not heavily influenced by outliers or extreme values. Additionally, since the ratio target variables have a relatively narrow range of possible values, e.g. $Sa_{RotD100}/Sa_{RotD50}$ is between 1 and $\sqrt{2}$, MSE effectively pushes the model to reduce small differences between the target values and the predicted ones.

The performance metric to evaluate our trained models is the coefficient of determination R^2. The evaluation is done using test data, and the value of R^2 ranges between 0 to 1, with 1 indicating a perfect fit.

B. Regression Models

We selected three regression models commonly used in research and practice for their proven ability to handle nonlinearity: Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

Support Vector Regression (SVR), referred to here as **SVM** for generality, modifies Support Vector Machines for regression by balancing prediction accuracy and model simplicity. It introduces a tolerance range to ignore minor errors, focuses on larger deviations, and uses kernel functions like Radial Basis Function (RBF) and Sigmoid to capture complex, nonlinear patterns. SVR also incorporates regularization to reduce overfitting, making it well-suited for small datasets like ours.

Random Forest improves prediction accuracy and minimizes overfitting by averaging outputs from multiple decision trees, each trained on randomly sampled subsets of the data. Features are randomly selected at each split, reducing tree similarity and enhancing generalization. This method effectively captures nonlinear patterns through the hierarchical structure of decision trees.

XGBoost builds an ensemble of decision trees, iteratively refining residuals to improve predictions. It uses a regularized objective function combining MSE and model complexity to prevent overfitting. Key parameters include the learning rate, maximum depth, and subsampling ratios. XGBoost's ability to handle sparse data and utilize parallel processing ensures scalability and efficiency for large datasets.

V. Experiments and Results

The experimentation phase aims to train, evaluate, and select the model for the four ground motion directionality parameters. It consists of three stages. The first stage splits the dataset into two groups: 80% for training and 20% for testing, as illustrated in Table 1. This task uses the library `train_test_split` from Scikit Learn [11].

Dataset	Number of Observations
Total dataset (100%)	105
Training dataset (80%)	84
Testing dataset (20%)	21

Table 1 Dataset split and observation count.

The second stage uses the training dataset to train three model architectures: Random Forest, Support Vector Machine, and XGBoost. The grid search algorithm is used for hyper-parameter tuning by trying every combination from a set of hyper-parameter values. Additionally, since the training dataset is limited to only 84 observations, cross-validation is used to calculate an accurate and efficient negative mean MSE. The model with the hyper-parameter set that shows the maximum mean negative MSE is selected as the best model and represents the model architecture. This task uses the library `cross_validate` from Scikit-Learn. The Random Forest model was evaluated using 1,440 distinct combinations of hyperparameter values, while the Support Vector Machine (SVM) model explored 300 combinations. Similarly, the XGBoost model underwent testing with 432 different hyperparameter configurations. The tested hyperparameter values are shown in Appendix A.

The third stage compares the three machine learning architectures by using the testing dataset. The comparison is evaluated by using R squared and MSE, aiming to select the model architecture with the highest performance. This task uses the libraries `r2_score` and `mean_squared_error` from Scikit Learn.

For each of the four different ground motion parameters for seismic design, the experimentation phase delivers 3 main results: the best set of hyper-parameter values, the comparison plot of the hyper-parameter values combinations and the comparison of the three machine learning architectures.

A. Variance of ln Sa_{RotD50}

The Random Forest model achieves the best results, with performance metrics detailed in Table 2, and the final prediction curve shown in Figure 2.

Model	MSE	R^2 Score
Random Forest	0.002244	0.999328
SVM	0.004926	0.998524
XGBoost	0.007361	0.997794

Table 2 Model performance comparison based on MSE and R^2 score.

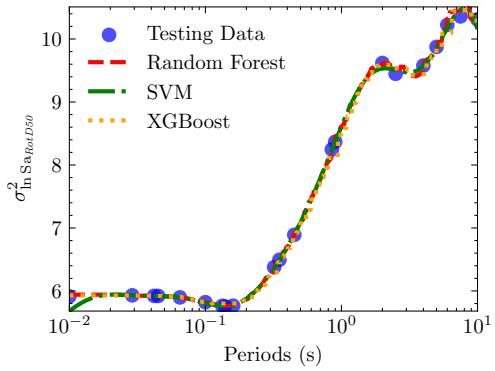

Fig. 2 Predictions and testing data for $\ln(Sa_{RotD50})$.

B. Variance of $\ln Sa_{RotD100}$

The Random Forest model also excels for this parameter, as presented in Table 3, with prediction curves in Figure 3.

Model	MSE	R^2 Score
Random Forest	0.002317	0.999325
SVM	0.007526	0.997808
XGBoost	0.008461	0.997536

Table 3 Model performance comparison based on MSE and R^2 score.

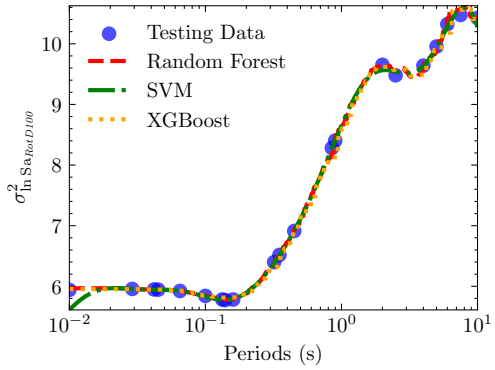

Fig. 3 Predictions and testing data for $\ln(Sa_{RotD100})$.

C. Median $Sa_{RotD100}/Sa_{RotD50}$

The Random Forest model remains the top performer, with results summarized in Table 4 and the prediction curve depicted in Figure 4.

Model	MSE	R^2 Score
Random Forest	4.38e-07	0.999244
XGBoost	1.42e-06	0.997545
SVM	4.78e-06	0.991759

Table 4 Model performance comparison based on MSE and R^2 score.

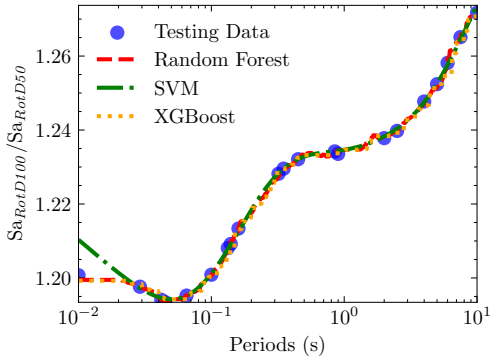

Fig. 4 Predictions and testing data for $Sa_{RotD100}/Sa_{RotD50}$.

D. Variance of ln ($Sa_{RotD100}/Sa_{RotD50}$)

The Random Forest architecture emerged as the best model for this ground motion parameter, as shown in Table 5, outperforming SVM and XGBoost, which failed to capture the ratio variance patterns despite low MSE values. A second approach involving input scaling before training was also unsuccessful. Given the low variance of $\ln(Sa_{RotD100}/Sa_{RotD50})$, its contribution to the total variance of the GMM is negligible (see Equation (2)). Thus, we recommend using a constant value for this parameter across all periods or omitting its contribution. Figure 5 illustrates the final prediction curves of each machine learning model compared to the testing data.

Model	MSE	R^2 Score
Random Forest	2.73e-09	0.844236
SVM	1.76e-08	-0.000388
XGBoost	1.80e-08	-0.027814

Table 5 Model performance comparison based on MSE and R^2 score.

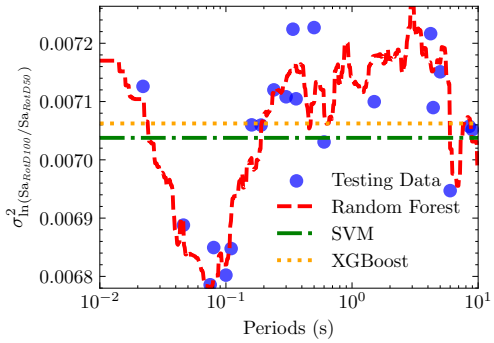

Fig. 5 Predictions and testing data for ln ($Sa_{RotD100}/Sa_{RotD50}$).

VI. Conclusion and Future Work

In conclusion, the Random Forest architecture demonstrated the best performance among the tested hyperparameters for each of the four ground motion directionality parameters. Specifically, it explains 99.93% of the variability in the variance of ln Sa_{RotD50}, 99.93% in the variance of ln $Sa_{RotD100}$, 99.92% in the median $Sa_{RotD100}/Sa_{RotD50}$, and 84.42% in the variance of $\ln(Sa_{RotD100}/Sa_{RotD50})$. Overall, except for the latest one, all machine learning models achieved strong results, with low mean squared error (MSE) and high R^2 scores.

The findings also indicate a high correlation between the period of vibration and the target variables. However, the low variance and tightly clustered values of the variance of $\ln(Sa_{RotD100}/Sa_{RotD50})$ led to near straight-line predictions from SVM and XGBoost. Several methods were explored to address this issue, but none proved effective, prompting the suggestion that a constant value for this parameter may be suitable for all periods.

The predicted directionality parameters can be integrated into modern ground motion models (GMMs) to improve estimates of $Sa_{RotD100}$ and its variance. For a consistent comparison with prior studies, future work would benefit from using established regression approaches (such as mixed-effects regression) in conjunction with the NGA-West2 database seismic records.

Readers interested in utilizing the final model developed in this project are encouraged to contact the authors directly to request access.

Group Members Contribution

The work and responsibilities were distributed equally throughout the entire project. Both members worked on the development and depuration of the code and, in the same way, on the result interpretation process.

References

[1] Stewart, J. P., Abrahamson, N. A., Atkinson, G. M., Baker, J. W., Boore, D. M., Bozorgnia, Y., Campbell, K. W., et al., "Representation of Bidirectional Ground Motions for Design Spectra in Building Codes," *Earthquake Spectra*, Vol. 27, No. 3, 2011, pp. 927–937. https://doi.org/10.1193/1.3608001.

[2] Shahi, S. K., and Baker, J. W., "NGA-West2 Models for Ground Motion Directionality," *Earthquake Spectra*, Vol. 30, No. 3, 2014, pp. 1285–1300. https://doi.org/10.1193/040913EQS097M.

[3] Ancheta, T. D., Darragh, R. B., Stewart, J. P., Seyhan, E., Silva, W. J., et al., "PEER NGA-West2 Database," Tech. rep., Pacific Earthquake Engineering Research Center, May 2013.

[4] Boore, D. M., "Orientation-Independent, Nongeometric-Mean Measures of Seismic Intensity from Two Horizontal Components of Motion," *Bulletin of the Seismological Society of America*, Vol. 100, No. 4, 2010, pp. 1830–1835. https://doi.org/10.1785/0120090400.

[5] Beyer, K., and Bommer, J. J., "Relationships between Median Values and Aleatory Variabilities for Different Definitions of the Horizontal Component of Motion," *Bulletin of the Seismological Society of America*, Vol. 96, No. 4, 2006, pp. 1512–1522. https://doi.org/10.1785/0120050210.

[6] Watson-Lamprey, J. A., and Boore, D. M., "Beyond SaGMRotI: Conversion to SaArb, SaSN, and SaMaxRot," *Bulletin of the Seismological Society of America*, Vol. 97, No. 5, 2007, pp. 1511–1524. https://doi.org/10.1785/0120070007.

[7] Abrahamson, N. A., Silva, W. J., and Kamai, R., "Summary of the ASK14 Ground Motion Relation for Active Crustal Regions," *Earthquake Spectra*, Vol. 30, No. 3, 2014, pp. 1025–1055. https://doi.org/10.1193/070913EQS198M.

[8] Moschetti, M. P., Aagaard, B. T., Ahdi, S. K., Altekruse, J., Boyd, O. S., Frankel, A. D., Herrick, J., Petersen, M. D., Powers, P. M., Rezaeian, S., et al., "The 2023 US National Seismic Hazard Model: Ground-motion characterization for the conterminous United States," *Earthquake Spectra*, Vol. 40, No. 2, 2024, pp. 1158–1190. https://doi.org/10.1177/87552930231223995.

[9] American Society of Civil Engineers, *Minimum Design Loads and Associated Criteria for Buildings and Other Structures*, asce/sei 7-22 ed., American Society of Civil Engineers, Reston, VA, 2021. URL https://ascelibrary.org/doi/book/10.1061/9780784415788.

[10] Shahi, S. K., and Baker, J. W., "NGA-West2 Models for Ground-Motion Directionality," Tech. rep., Pacific Earthquake Engineering Research Center, University of California, Berkeley, May 2013.

[11] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G., "API design for machine learning software: experiences from the scikit-learn project," *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

A. Hyperparameter Grids

Parameter	Description	Values
n_estimators	Number of trees in the forest	{50, 100, 200, 500}
criterion	Function to measure the quality of a split	{"squared_error", "absolute_error", "friedman_mse", "poisson"}
max_depth	Maximum depth of the tree	{None, 10, 20, 30, 50}
min_samples_split	Minimum number of samples required to split an internal node	{2, 5, 10}
min_samples_leaf	Minimum number of samples required to be at a leaf node	{1, 2, 4}
max_features	Number of features to consider for the best split	{"sqrt", "log2"}

Table A.1 Hyperparameter grid for Random Forest Regressor.

Parameter	Description	Values
kernel	Specifies the kernel type to be used in the algorithm	{"rbf", "sigmoid"}
C	Regularization parameter. Controls trade-off between achieving a low error on the training data and minimizing model complexity	{0.1, 1, 10, 100, 1000}
gamma	Kernel coefficient for "rbf", "poly", and "sigmoid" kernels	{"scale", "auto", 0.001, 0.01, 0.1, 1}
epsilon	Specifies the margin of tolerance in the epsilon-insensitive loss function	{0.1, 0.2, 0.5, 0.01, 0.001}

Table A.2 Hyperparameter grid for Support Vector Machine.

Parameter	Description	Values
n_estimators	Number of boosting rounds (trees)	{100, 200, 500, 1000}
learning_rate	Step size shrinkage to prevent overfitting	{0.01, 0.1, 0.3}
max_depth	Maximum depth of a tree	{3, 5, 10}
subsample	Fraction of samples used per tree	{0.8}
gamma	Minimum loss reduction required to split a node	{0, 1, 5}
reg_alpha	L1 regularization term on weights	{0, 10}
reg_lambda	L2 regularization term on weights	{1, 10}

Table A.3 Hyperparameter grid for XGBoost.