Machine Learning-Based Selection of Efficient Parameters for the Evaluation of Seismically Induced Slope Displacements

Farahnaz Soleimani, Ph.D.[1]; Jorge Macedo, Ph.D.[2]; and Chenying Liu[3]

[1]Dept. of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA. Email: soleimani@gatech.edu
[2]Dept. of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA. Email: jorge.macedo@gatech.edu
[3]Dept. of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA. Email: cliu662@gatech.edu

ABSTRACT

Seismically induced slope displacements (D) are often used as a performance index in the seismic design of slope systems. Although previous studies have developed predictive relationships for estimating D, a comprehensive selection of the most efficient parameters to explain D, considering the linearity and nonlinearity in D, has not been thoroughly performed. This study uses modern machine learning-based techniques including Stepwise Selection, LASSO, and Random Forest to identify the influential features in the estimation of D in shallow crustal tectonic settings. The detected influential features are the system's yield coefficient (k_y), initial fundamental period (T_s), earthquake moment magnitude (M_w), peak ground velocity (PGV), and the degraded spectral acceleration at $1.3T_s$ [$S_a(1.3T_s)$]. Moreover, the results indicate that there is no significant gain in accuracy beyond five features. The detected significant parameters provide insight and a basis for developing more efficient prediction models.

INTRODUCTION

Disaster assessment is an essential step to improve the resilience of civil infrastructure and lifeline systems. For example, seismically-induced displacements in slope systems such as landslides can affect lifelines (Olsen et al. 2015). In this context, this study focuses on the selection of efficient parameters for estimating seismically-induced slope displacements (D) as elaborated in detail subsequently. Currently, the procedures that are commonly used in the seismic performance evaluation of slope systems include: (I) Newmark-based slope displacement analyses, (II) pseudo-static slope stability analyses, and (III) advanced numerical procedures.

Procedure II is still used in practice mostly in projects that have low associated risk (e.g., Rennat and Miller 1997; Gerath et al. 2010; Kavazanjian et al. 2011) and based upon regulators' request. In contrast to procedure III, which requires specific geotechnical information and may be computationally intensive, procedure I is still preferred in engineering practice because of its simplicity and yet providing reliable estimates of the seismic performance of slope systems. The procedures in I are based on seismic sliding block displacement analysis that can be performed following three procedures: (1) by either considering rigid-sliding blocks (e.g., Newmark 1965; Watson-Lamprey and Abrahamson 2006; Saygili and Rathje 2008; Du and Wang 2016), or (2) considering a decoupled approximation for the dynamic response of the slope (e.g., Bray and Rathje 1998), or alternatively (3) based on fully coupled stick-slip sliding blocks (e.g., Bray and Travasarou 2007; Macedo et al. 2017; Bray et al. 2018; Bray and Macedo 2019; Macedo et al. 2020).

In order to generate *D* realizations, we used a fully coupled stick-slip sliding block model (Macedo 2017). The *D* realizations can be used to build *D* models or understand the parameters that govern *D*. An essential step in developing a *D* model is the selection of efficient parameters that can explain the trends in *D*. These parameters are often selected among candidates such as the earthquake magnitude (M_w), rupture distance (*CIstD* or *R*), and ground motion intensity measure parameters (*IMs*) such as spectral accelerations (S_a), peak ground acceleration (*PGA*), etc. Previous efforts to develop *D* models have typically used fixed functional forms often including first or second-degree polynomials of *Ln*(*IM*). These studies have selected efficient *IMs* by iteratively varying the *IM* and evaluating the standard deviation of regression models (e.g., Rathje and Saygili 2008). However, there are now well-established machine learning (ML)-based procedures that can identify influential features in a more holistic manner to provide valuable insights to develop ML-based predictive models for *D*. This study focuses on the use of selected ML procedures to better understand the parameters that influence *D*.

GROUND MOTION DATABASE

For this study, we considered the NGA-West2 database (Bozorgnia et al. 2014) for shallow crustal earthquakes. The NGA-West2 database consists of 21,332 three-component ground motion recordings from which we chose 6711 ground motion records (with each record having two horizontal components) with a moment magnitude (M_w) ranging from 5 to 7.9 at *R* less than 200 km. More details on the ground motion records are found in Bozorgnia et al. (2014).

GENERATION OF *D* REALIZATIONS

To generate *D* realizations, the ground motions (described in the previous section) are used as inputs in the fully coupled, nonlinear, deformable stick-slip sliding model employed by Bray and Travasarou (2007) and implemented to be used in engineering practice in Candia et al. (2018, 2019). This model is used in the current study with the modification provided by Macedo (2017). In total, around 3 million sliding block analyses are performed. An equivalent-linear viscoelastic modal analysis with strain-dependent material properties is used to capture the dynamic response of the deformable sliding mass.

The assigned yield coefficient (k_y) and initial fundamental period (T_s) values that are used in the model cover a wide range of slope systems. These properties range from 0.01 to 0.8 for k_y and from 0.0 s to 2.0 s for T_s. In addition, for the baseline sliding block, the overburden stress-corrected shear wave velocity (V_{s1}) is set to 270 m/s. For non-zero T_s values, the sliding block height (*H*) ranges from 3 m to 100 m, and V_{s1} varies from 200 m/s to 450 m/s, following the IBC Site Class D or C. Consistent with the definition provided by Bray and Macedo (2019), the *D* values lesser than 0.5 cm are considered as negligible displacements ("zero" *D*) since they imply an acceptable performance for most geotechnical slope systems.

The candidate features that are considered to estimate *D* are listed in Table 1. These candidates cover slope properties, earthquake parameters, and *IMs*. Considering non-zero values, the trends of *D* against selected features are displayed in Figure 1. It is observed that *D* increases with the increase of S_a and M_w, while it decreases with the increase of k_y. Also, *D* is not too sensitive to T_s. However, a thorough investigation is required to indicate the influence of each feature on the slope displacements due to the large variability in the data.

Table 1. List of considered features to estimate *D*

Variable	Feature Description
k_y	System's yield coefficient
T_s	Initial fundamental period of the sliding mass
M_w	Earthquake moment magnitude
V_{s30}	Shear-wave velocity averaged over the 30m depth of soil
$d_{5\text{-}95}$	Average duration (5-95 percentile) of two horizontal components of the ground motion
I_a	Average arias intensity of the two horizontal components of the ground motion
R	Closest distance of the station to the earthquake rupture
$S_a(nT_s)$	Ground motion's spectral acceleration at a degraded period equal to nT_s (n = 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 2.0, 2.5, 3.0)
PGA	Peak ground acceleration
PGV	Peak ground velocity

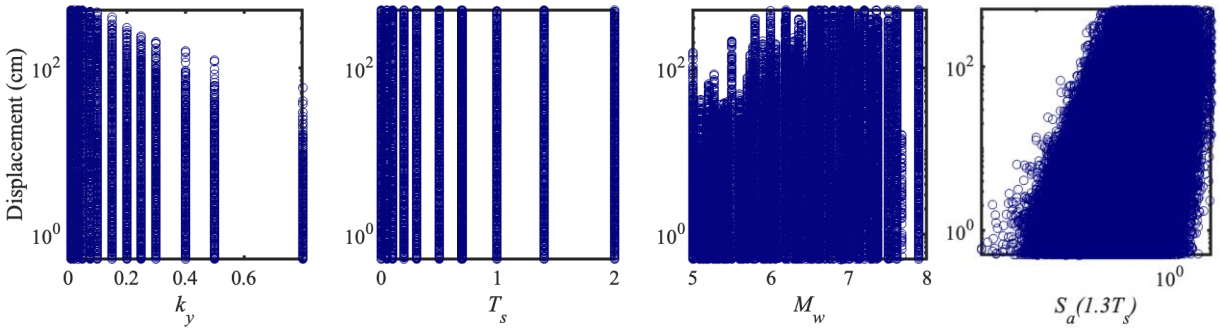

Figure 1. Variation of *D* against a subset of candidate features.

MACHINE-LEARNING-BASED TECHNIQUES

Sparse regression algorithms (Krishnapuram et al. 2005; Cawley et al. 2007; Hastie et al. 2009), that are described in the following, are applied to identify features that contribute the most in the estimation of *D*. Including only the influential features in the predictive models in one hand reduces model complexity, and on the other hand, prevents overfitting that commonly happens when many features are added to the model. Furthermore, Multivariate Regular Regression (MRR) (Berry et al. 1985; Hahs-Vaughn and Lomax 2020) is developed as the base model for the purpose of comparisons. MRR predicts the linear relationship between a response variable and multiple explanatory variables.

Forward Stepwise Regression. Stepwise Regression involves an iterative process in which the inclusion or exclusion of the features is tested in an iterative process to satisfy a predefined fitness criterion. In this study, Forward Selection (Miller 2002; Blanchet et al. 2008) is used as one of the main approaches under the category of Stepwise Regression. This selection technique starts with constructing a constant model and sequentially add each variable to the regression model until none of the remaining variables improves the considered criterion.

LASSO . As a modified form of standard linear regression, Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 2011; Soleimani et al. 2017) includes a regularization term in its objective function Equation (1).

$$\min \sum_{i=1}^{q}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \sum_{j=1}^{p} \lambda \left\|\beta_j\right\|_1 \quad (1)$$

In this equation, x represents the explanatory variables and β represents the corresponding coefficients. In addition, q and p represent the total number of observations and the number of considered features, respectively. The variable y expresses the observed response value, and λ is known as the hyper-parameter since its value directly affects the coefficients that are set to zero. In this technique, λ is tuned until an optimal sparsity is obtained in the model. This condition is reached at the minimum residual sum of squares. As expressed in Equation (1), the *L1* regularization (i.e., $\left\|\beta_j\right\|_1$) in LASSO results in a sparse model since some coefficients are set to zero and hence the corresponding features are eliminated from the model.

Random Forest. Decision trees are typically constructed from multiple nodes (representing the cut points) and leaves (decision on the target). Random decision trees are produced by Random Forest (Ho 1998; Liaw and Wiener 2002; Soleimani 2021) to address two common issues with the decision trees: overfitting models and oversized uninterpretable trees. To this end, Random Forest partitions training data into several individual subsets to form multiple random decision trees and takes the average results. This approach leads to reduce the variance. The features located at the topmost nodes are counted as the most significant ones in reducing the impurity of the constructed decision tree. Hence, the tree is pruned below a specific node to identify the most relevant features. In this study, Random Forest with 200 decision trees, is used to estimate features' importance and measure the correlation between the features.

Evaluation Process. In this study, we have used 10-fold cross-validation, which partitions data into similar-sized 10 subsets and assigns an equal number of data points to each subset. In this iterative process, the model is trained on 9 subsets of data and then is validated on the remaining one subset. The evaluation process is performed based on prediction accuracies and the cross-validation errors (CV error) computed according to Equation (2) to Equation (4).

$$\text{Prediction accuracy} = \frac{1}{10}\sum_{j=1}^{10}[1 - \left(\left\|y_{predicted} - y_{observed}\right\|_2 / \|y_{observed}\|_2\right)]_j \quad (2)$$

$$CV\ error = \frac{1}{10}\sum_{j=1}^{10} MSE_j \quad (3)$$

$$MSE = \frac{1}{q}\sum_{i=1}^{q}((y_{observed})_i - (y_{predicted}{}_i))^2 \quad (4)$$

SELECTION OF EFFICIENT FEATURES

The findings of feature selection analysis are summarized in Table 2 in which the selected influential features are marked with "**x**". Figures 2 and 3 elaborate additional key findings in the process of conducting the feature selection techniques. As demonstrated in Figure 2, five is the optimum number of features to predict the value of *D*. Beyond this point, increasing the number of features in the model has a negligible impact on the prediction accuracy. The selected five features by the Forward Selection are k_y, T_s, M_w, *PGV*, and $S_a(1.3T_s)$.

Figure 3 illustrates the path followed by LASSO to satisfy the objective function. This shows how the mean squared error (MSE) is minimized by optimizing the hyperparameter λ in the model. The optimum λ resulting in the minimum MSE is selected for the LASSO model, whereas the plot also shows the lowest MSE value plus one standard error. To minimize the MSE, LASSO typically tends to include as many features in the model as possible and as a result, it generates models with a greater number of features (a total of 18 features in our problem) compared to the other approaches. As highlighted in Table 2, the features selected by Random Forest and Forward Stepwise Selection are both subsets of the 18 features selected by LASSO. Although the number of features identified by LASSO is not ideal for the practical application, we use the results of this approach as a verification for the results obtained by the other techniques.

Furthermore, the level of importance of the considered features is compared in Figure 3 using the Random Forest technique. The importance of each feature is defined as the sum of changes in MSE due to the splits of that feature in the growth of the decision trees. According to this comparison, the k_y has the highest importance, and the next two significant variables are PGV and T_s.

Table 2. The list of identified significant features using Forward selection, LASSO, and Random Forest

All features	Significant features		
	Forward Selection	LASSO	Random Forest
k_y	**x**	**x**	**x**
T_s	**x**	**x**	**x**
d_{5-95}			
I_a			
M_w	**x**	**x**	**x**
R		**x**	
V_{s30}		**x**	
PGA		**x**	
PGV	**x**	**x**	**x**
$S_a(1.5T_s)$		**x**	
$S_a(2.0T_s)$		**x**	
$S_a(2.5T_s)$		**x**	
$S_a(3.0T_s)$		**x**	
$S_a(1.3T_s)$	**x**	**x**	
$S_a(1.4T_s)$		**x**	
$S_a(1.6T_s)$		**x**	
$S_a(1.7T_s)$			
$S_a(1.8T_s)$		**x**	
$S_a(1.0T_s)$		**x**	**x**
$S_a(1.1T_s)$		**x**	**x**
$S_a(1.2T_s)$		**x**	

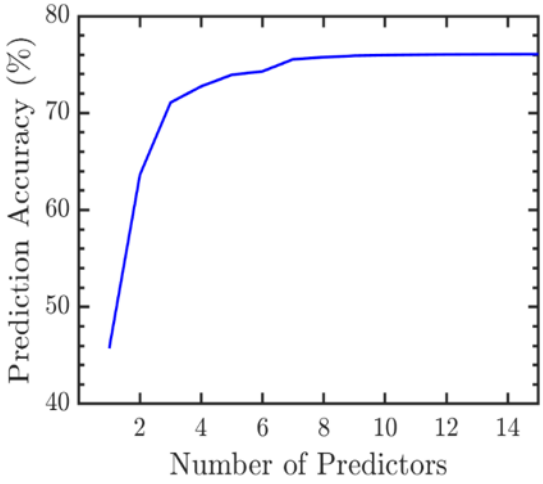

Figure 2. Optimal number of features defined by Forward Selection.

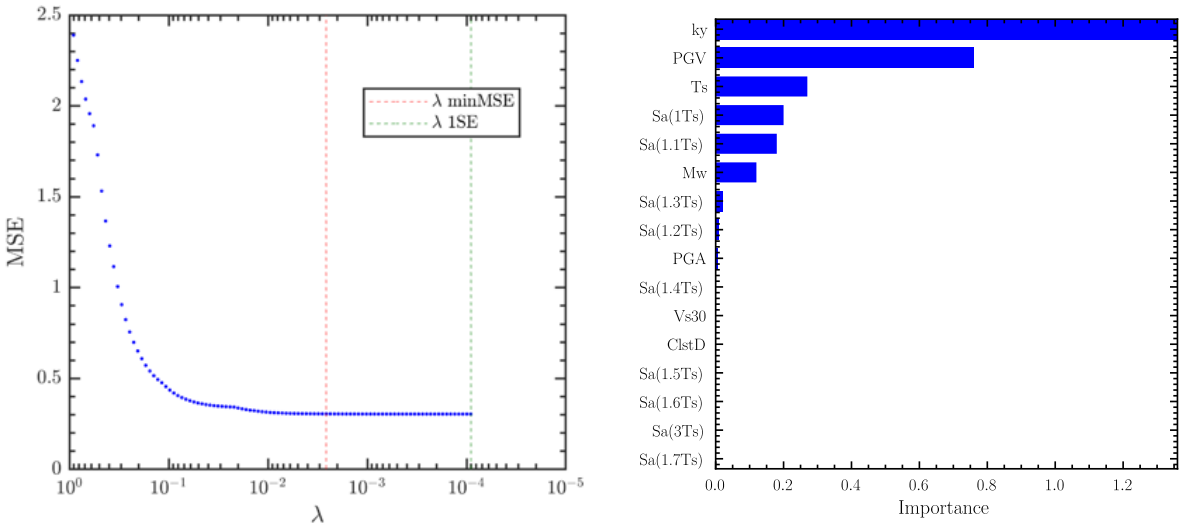

Figure 3. Highlights of the implemented feature selection techniques; (left): tunning the hyperparameter by LASSO; (right): the level of importance of the features found by Random Forest.

RECOMMENDED EFFICIENT PARAMETERS TO ESTIMATE *D*

The performance of the implemented feature selection techniques is compared in Table 3 in terms of prediction accuracy and CV error. It is found that LASSO adopted many features, without a significant improvement in the model performance.

Overall, Random Forest provided the best-performed model among the applied techniques with the highest prediction accuracy of 0.795 and the lowest CV error equal to 0.336. On the other hand, the Forward Stepwise Selection technique with five features performed close to Random Forest in terms of both the prediction accuracy (0.733) and the CV error (0.375). Moreover, Forward Selection exhibits comparable results to those of the full linear model.

In summary, although both Random Forest and Forward Stepwise Selection proved that efficient predictive models with high prediction accuracy and low cross-validation error can be achieved by reducing the number of features in their models, the Forward Stepwise technique provided a lower-dimensional model. Random Forest identified six features as the most influential predictors in estimating *D*, while the Forward Stepwise technique revealed no notable improvement in the prediction power beyond five features, as demonstrated in Figure 2. Moreover, in terms of the features selected by Random Forest, two of them are highly correlated (i.e., $S_a(1.0T_s)$, and $S_a(1.1T_s)$). In addition to the insight gained from the feature selection techniques, engineering judgment is required to avoid overfitting by introducing similar or highly correlated predictors and incorporating predictors with physically reasonable meanings in affecting the slope displacements. Thus, the mentioned observations guide us to propose the selected influential features by the Forward Selection (i.e., k_y, T_s, M_w, *PGV*, and $S_a(1.3T_s)$) as the most efficient features to estimate *D*.

Table 3. Comparison of the feature selection techniques

ML Techniques	No of Features	Prediction Accuracy	Cross-validation Error
Linear Regression Full Model	21	0.756	0.363
Forward Stepwise Selection	5	0.733	0.375
Least Absolute Shrinkage and Selection Operator (LASSO)	18	0.756	0.363
Random Forest	6	0.795	0.336

CONCLUSION

For the seismic performance assessment of slope systems, analytical procedures are typically used to estimate seismically-induced slope displacements (*D*). This study explores the more efficient features in predictive models of *D*, using novel ML-based techniques. In this regard, the NGA-West2 shallow crustal ground motion database was used to generate *D* realizations, which were subsequently used in the feature selection. The candidate features considered in this study include slope parameters (e.g., the yield coefficient), earthquake parameters (e.g., magnitude and distance), intensity ground motion parameters (e.g., spectral accelerations), and site conditions.

Beyond validating our analysis approach, our findings imply that optimal prediction accuracy and cross-validation error can be obtained by using a reduced number of features in *D* models. More particularly, a noticeable improvement in the prediction accuracy was not observed by including more than five features (recommended as the optimal number of features). These features are highly correlated with the estimated *D* values. We found the most efficient features are system's yield coefficient (k_y), its fundamental period (T_s), earthquake moment magnitude (M_w), peak ground velocity (*PGV*), and ground motion spectral acceleration at degraded periods equal to 1.3*Ts*. Therefore, we recommend them to be included in the predictive models for *D*.

REFERENCES

Berry, W. D., Feldman, S., and Stanley Feldman, D. (1985). *Multiple regression in practice (No. 50).* Sage.

Blanchet, F. G., Legendre, P., and Borcard, D. (2008). Forward selection of explanatory variables. *Ecology*, 89(9), 2623-2632.

Bozorgnia, Y., Abrahamson, N. A., Atik, L. A., Ancheta, T. D., Atkinson, G. M., Baker, J. W., Baltay, A., Boore, D. M., Campbell, K. W., Chiou, B. S. J., and Darragh, R. (2014). NGA-West2 research project. *Earthquake Spectra*, 30(3), 973-987.

Bray, J. D., Macedo, J., and Travasarou, T. (2018). Simplified procedure for estimating seismic slope displacements for subduction zone earthquakes. ASCE *Journal of Geotechnical and Geoenvironmental Engineering*; 144(3).

Bray, J. D., and Macedo, J. (2019). Procedure for Estimating Shear-Induced Seismic Slope Displacement for Shallow Crustal Earthquakes. ASCE *Journal of Geotechnical and Geoenvironmental Engineering*; 145(12).

Bray, J. D., and Rathje, E. R. (1998). Earthquake-induced displacements of solid-waste landfills. ASCE *Journal of Geotechnical and Geoenvironmental Engineering*; 124(3):242–53.

Bray, J. D., and Travasarou, T. (2007). Simplified procedure for estimating earthquake-induced deviatoric slope displacements. ASCE *Journal of Geotechnical and Geoenvironmental Engineering*; 133(4):381–392.

Candia, G., Macedo, J., and Magna-Verdugo, C. (2018, June). An integrated platform for seismic hazard evaluation. In *11th US National Conference on Earthquake Engineering*. Los Angeles, USA.

Candia, G., Macedo, J., Jaimes, M. A., and Magna-Verdugo, C. (2019). A New State-of-the-Art Platform for Probabilistic and Deterministic Seismic Hazard Assessment. *Seismological Research Letters*, 90(6), 2262-2275.

Cawley, G. C., Talbot, N. L., and Girolami, M. (2007). Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems*, 19, 209.

Du, W., and Wang, G. (2016). A one-step Newmark displacement model for probabilistic seismic slope displacement hazard analysis. *Engineering Geology*, 205, 12-23.

Gerath, R., Jakob, M., Mitchell, P., and Van Dine, D. (2010). *Guidelines for legislated landslide assessment for proposed residential developments in BC*. Association of Professional Engineers and Geoscientists of British Columbia (APEGBC). British Columbia.

Hahs-Vaughn, D. L., and Lomax, R. G. (2020). *An introduction to statistical concepts*. Routledge.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science and Business Media.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), pp.832-844.

ICC (International Code Council). (2015). *International Building Code*, Washington, DC. International Code Council.

Kavazanjian, E., Wang, J. N., Martin, G., Shamsabadi, A., Lam, I., Dickenson, S. E., Hung, C. J., and Brinckerhoff, P. (2011). LRFD Seismic Analysis and Design of Transportation Geotechnical Features and Structural Foundations-NHI Course No. 130094 Reference Manual Geotechnical Engineering Circular No. 3 (No. FHWA-NHI-11-032). United States. Federal Highway Administration.

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6), pp.957-968.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

Macedo, J. L. (2017). *Simplified procedures for estimating earthquake-induced displacements*. Ph.D. thesis, Dept. of Civil and Environmental Engineering, Univ. of California, Berkeley.

Macedo, J., Bray, J., and Travasarou, T. (2017). Simplified procedure for estimating seismic slope displacements in subduction zones. In *Proc., 16th World Conf. on Earthquake Engineering*.

Macedo, J., Candia, G., Lacour, M., and Liu, C. (2020). New developments for the performance-based assessment of seismically-induced slope displacements. *Engineering Geology*, 277, 105786.

Miller, A. (2002). *Subset selection in regression*. CRC Press.

Newmark, N. M. (1965). Effects of earthquakes on dams and embankments. *Geotechnique*; 15(2):139–60.

Olsen, M. J., Ashford, S. A., Mahlingam, R., Sharifi-Mood, M., and Gillins, D. T. (2015). Impacts of potential seismic landslides on lifeline corridors (No. FHWA-OR-RD-15-06). Oregon. Dept. of Transportation. Research Section.

Rathje, E. M., and Saygili, G. (2008). Probabilistic Seismic Hazard Analysis for the Sliding Displacement of Slopes: Scalar and Vector Approaches. ASCE *Journal of Geotechnical and Geoenvironmental Engineering*; 134(6), 804-814.

Rennat, E., and Miller, S. (1997). Guía ambiental para la estabilidad de taludes de depósitos de desechos sólidos de mina. *Lima: Ministerio de Energía y Minas. Consulta*, 10(08), 2019.

Saygili, G., and Rathje, E. M. (2008). Empirical predictive models for earthquake-induced sliding displacements of slopes. ASCE *Journal of Geotechnical and Geoenvironmental Engineering*; 134(6):790–803.

Soleimani, F., Vidakovic, B., DesRoches, R., and Padgett, J. (2017). Identification of the significant uncertain parameters in the seismic response of irregular bridges. *Engineering Structures*, 141, 356-372.

Soleimani, F. (2021, August). Analytical seismic performance and sensitivity evaluation of bridges based on random decision forest framework. In *Structures* (Vol. 32, pp. 329-341). Elsevier.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), pp.273-282.

Watson-Lamprey, J., and Abrahamson, N. (2006). Selection of ground motion time series and limits on scaling. *Soil Dynamics and Earthquake Engineering*, 26(5), 477-482.