17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 27th to October 2nd, 2021

MACHINE LEARNING BASED SEISMIC BEHAVIOR PREDICTION OF RC COLUMNS FROM PAST CYCLIC PUSHOVER TEST DATA

S.P. Rayjada [(1)], J. Ghosh [(2)], M. Raghunandan [(3)]

[(1)] *Research Scholar, Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, India, sprayjada@iitb.ac.in*
[(2)] *Assistant Professor, Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, India, jghosh@iitb.ac.in*
[(3)] *Assistant Professor, Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, India, meerar@iitb.ac.in*

Abstract

To simulate the nonlinear response of reinforced concrete (RC) frame elements under the earthquake shaking, analytical models must capture the full range of structural behavior, from flexural yielding to cyclic stiffness and strength degradation. The lumped plasticity beam-column element model has been widely adopted for modeling such behavior owing to its simplicity and computational efficiency. The numerical modeling parameters for a column are estimated using the trilinear backbone curve that envelops experimental hysteresis results and a deterioration parameter that captures the column's strength and stiffness deterioration over cycles of loading in the experiment. At present, a widely used approach for estimating column modeling parameters utilizes linear regression-based semi-empirical equations developed following the calibration of experimental column test results with varying design and detail. As backbone curve parameters represent complex seismic behavior and depend on many factors, an assumption of static linear underlying relationship in linear regression may not be appropriate. In the last two decades, various machine learning algorithms have been developed to determine such complex relationships and successfully implement various problems. This study utilizes a random forest based approach owing to its capabilities to efficiently handle high-dimensional data, easy parallelization, few tuning parameters, applicability with categorical or continuous variables and less susceptibility of overfitting. In addition to the prediction, the model should be quantified prediction uncertainty to assess deviation of model prediction from the actual value. Typically advanced machine learning approaches only provide pointwise prediction not accompanied by uncertainty estimation. Hence, this study compares various methodology to quantify model uncertainty in random forest prediction. Further, in the linear regression models, uncertainty is described by the constant lognormal standard deviation. The results reveal that as significant heterogeneity prevails in the column properties, the assumption of the constant standard deviation (as typical of linear regression approach) throughout the sample space leads to an inaccurate estimation of uncertainty. While on the other hand, random forest prediction intervals are not constant and vary across the data points based on the data point's proximity to training data. Therefore, uncertainty can be quantified more reliably.

Keywords: *Random Forest; Statistical Learning; Structural Response; Trilinear Backbone Curve; Uncertainty Propagation*

1. Introduction

Several structural component models are developed over the years to idealize the behavior of reinforced concrete (RC) elements under seismic loading. Structural component models can be classified as lumped plasticity models and distributed plasticity models based on how the plasticity is distributed through the member cross sections and along its length. In lumped plasticity models, inelastic deformations are concentrated at the member's ends as nonlinear rotational spring and its complex structural behavior is captured by phenomenological hysteresis rules. In distributed plasticity models, inelastic behavior is either restricted to the specific length of the member in the finite-length hinge approach or it is distributed across the length of the member by fiber section formulation [1]. It is noted that while it seems that a more rigorous distributed plasticity formulation provides the better capability to capture inelastic behavior, they offer model calibration challenges due to a higher number of parameters. Additionally, specific actions such as degradation due to bar bulking can be easily incorporated using simpler phenomenological rules. Further, computational cost and analysis time is significantly higher for advanced analysis such as nonlinear time history analysis. On the other hand, lumped plasticity models are able to capture relevant features faster with a sufficient level of approximation. For example, post-capping negative stiffness related to physical behavior phenomena of concrete crushing, reinforcing bar buckling and fracture, or bond failure can be characterized more effectively by the lumped plasticity model [2]. Hence, the lumped plasticity beam-column element model has been widely adopted to represent such behavior due to its simplicity and computational efficiency.

In lumped plasticity models, it is a critical task to define phenomenological hysteresis rules for the springs provided at the ends of structural elements to accurately characterize the element behavior under the cyclic loading. Several hysteresis rules are developed over the years to simulate the cyclic behavior of RC frames [3–9]. For all hysteresis models, it is necessary to estimate backbone curve parameters and the cyclic deterioration parameters accurately. The backbone curve envelops experimental hysteresis results and deterioration parameters capture the strength and stiffness deterioration of the column over cycles of loading in the experiment. These parameters depend on the material and geometric properties of the member. Often, empirical equations are used for the prediction of the parameters following the calibration of experimental column test results with varying design and details. Such equations establish the relationship between parameters and material or geometric properties [2,10–15]. Panagiotakos and Fardis [16] proposed simplified models for the parameters associated with the yielding and failure of the RC members. Haselton et al. [2] proposed semi-empirical equations for trilinear backbone curve parameters of the ductile RC column mainly failing in flexure. A similar study was extended by Lee and Han [17] for old reinforced concrete columns failing in shear and flexural shear. Many code provisions and guidelines also provide models to estimate model parameters [14,15]. However, these estimates are generalized and often deviate from true value [18].

The majority of existing literature provides empirical models or utilizes elementary statistical tools such as the linear regression approach for developing prediction equations. As backbone curve parameters represent complex seismic behavior and depend on many factors, the assumption of underlying static relationship such as linear relationship may not be appropriate. It introduces additional prediction bias and may lead to poor model performance. In recent years, machine learning algorithms have significantly evolved due to their capability to handle complex data and automatically detect the underlying relationship in data. In the field of earthquake engineering, various researchers across the globe have successfully implemented the suitable machine learning techniques for a variety of problems [19–21]. For backbone curve parameter estimation, Luo and Paal [22] utilized a multi-output support vector machine algorithm. The backbone curve considered in the study is bilinear i.e., post-cap softening behavior is not incorporated. Liu and Li [23] proposed Artificial Neural Network model for bilinear and trilinear backbone curves. However, extensive calibration of experimental results is not carried out incorporating cyclic degradation effects.

When applying machine learning algorithms, the primary emphasis is given to achieve better prediction accuracy. However, it is also essential to quantify prediction uncertainty to assess the deviation of model prediction from the true value. Typically advanced machine learning approaches only provide pointwise prediction not accompanied by uncertainty estimation. On the other hand, in approaches proposed

earlier, such as the linear regression approach proposed by Haselton et al. [2], uncertainty is described by the constant lognormal standard deviation. It is noted that the models are fitted on limited experimental column test results and significant heterogeneity prevails in column properties. Thus, the assumption of the constant standard deviation throughout the sample space can lead to an inaccurate estimation of uncertainty. Hence, pointwise estimation of prediction uncertainty is also necessary.

To fill the existing research gap, this study utilizes a random forest regression algorithm to estimate parameters of the hysteresis model proposed by Ibarra et al. [7] based on the material and geometric properties of the member. Random Forest algorithm was first proposed by Breiman [24] and has successfully implemented in various fields due to its capabilities to efficiently handle high-dimensional data, easy parallelization, few tuning parameters, applicability with categorical or continuous variables and less susceptibility of overfitting. At the onset, a description of the backbone curve parameters and energy dispassion parameters for the hysteresis model developed by Ibarra et al. [7] is provided. It is followed by a compilation of column databases and potential predictors. Further, an overview of random forest regression and a discussion of the proposed methodology for hyperparameter optimization, cross-validation, and feature selection are given. Lastly, the description of various approaches to quantify pointwise prediction uncertainty in random forest regression is provided.

2. Backbone Curve and Energy Dissipation Parameters of the Hysteresis Model

As stated earlier, inelastic behavior is captured in the lumped plasticity model by providing nonlinear rotational spring at the member's ends and properties of the rotational springs are defined based on the hysteresis model. This study considers the hysteresis model developed by Ibarra et al. [7]. This model captures flexural response, bond-slip between concrete and reinforcing bars and deterioration behavior of the RC fame member observed under the cyclic loading. This behavior is characterized by a trilinear monotonic backbone curve and energy dissipation parameters that can degrade the backbone curve under cyclic loading. Hence, rotational spring properties depend on the backbone curve and energy dissipation parameters [2,7].

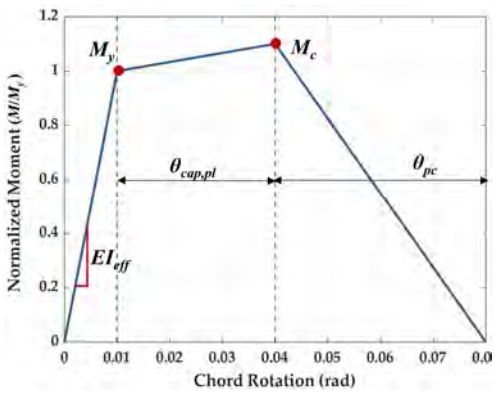


Fig. 1 – Trilinear backbone curve

Trilinear backbone curve can be obtained by estimating various parameters as shown in Fig. 1. Parameters yield strength (M_y) and effective stiffness (EI_{eff}) govern the member behavior up to yielding state. Yield strength (M_y) accounts for behavior longitudinal rebar yielding, concrete cracking (flexure and shear), concrete crushing, while effective stiffness (EI_{eff}) accounts for pre-yield displacements, including concrete cracking and bond-slip. Effective stiffness (EI_{eff}) can be defined in two ways based on deformation levels expected in the analysis - secant stiffness to the member yield strength (EI_y) and the secant stiffness to 40% of the yield strength (EI_{40}) [2]. Effective stiffness values are expressed as a ratio of the gross cross section stiffness (EI_g) i.e. EI_y/EI_g and EI_{40}/EI_g Strain hardening behavior is governed by post-yield stiffness that can be described in terms of a ratio of ultimate to yield strength (M_c/M_y). The capping point is represented by plastic rotation capacity ($\theta_{cap,pl}$). The capping point indicates longitudinal rebar bucking and loss of confinement due to failure of lateral reinforcement. Deterioration in a given cycle of loading is effectively

characterized by the post-peak softening branch of the trilinear monotonic backbone curve Post capping softening behavior is governed by post-capping rotation capacity (θ_{pc}). Deterioration over two subsequent cycles is characterized by peak-oriented hysteresis rules and energy dissipation parameters. Energy dissipation parameters include normalized energy- dissipation capacity (λ) and an exponent term (c) representing the change in the rate of cyclic deterioration. For the columns failing in flexural, the exponent term (c) can assume to be 1.0 [2]. Residual strength is considered to be negligible.

3. Compilation of Column Databases

This study develops separate random forest regression models for backbone curve parameters and energy-dissipation capacity (λ) (called hysteresis model parameters) based on experimental results of RC members under cyclic loading. For that purpose, this study utilizes an experimental column database developed by Berry et al. [25] and calibration results by Haselton et al. [2] for the same database. The calibrated column test set includes specimens failing in flexural (220 tests) failure mode and combined flexure-shear mode (35 tests). Out of 220 column specimens failing in flexural mode, 187 columns are selected based on the judgment after removing repeated data (columns having the same properties).

Based on experimental results, random forest regression models aim to establish the relationship between member properties and hysteresis model parameters. For that purpose, 19 potential predictor variables (features) mentioned in Table 1 encompassing data regarding the material and geometric properties of the member and applied axial loads are selected. An initial screening of features is done based on the correlation coefficient. Five variables P, A_l, P/P_b, s, s/d. with a correlation coefficient higher than 0.75 are removed. So 14 features are utilized for the development of random forest regression models.

Table 1 – Potential predictor variables

Features	Symbol	Mean	Maximum	Minimum
Shear span to effective depth ratio	Ls/H	3.92	7.4	1.5
Axial load ratio	v	0.28	0.80	0.00
Applied axial load (kN)	P	1687	8044	0.00
Ratio of axial load and axial load at the balanced condition	P/P_b	0.93	2.12	0.00
Concrete compressive strength (MPa)	f_c	59.6	118	20.2
Longitudinal reinforcement yield strength (MPa)	f_y	447	587	339
Longitudinal reinforcement ratio	ρ	0.03	0.08	0.01
Longitudinal reinforcement area (mm^2)	A_l	2550	7910	438
Maximum diameter of longitudinal reinforcement (mm)	d_b	17.6	35.8	9.5
Stirrup spacing (mm)	s	75.6	152	25
Rebar buckling coefficient	S_n	9.12	31.1	2.9
Transverse reinforcement area (mm^2)	A_{sh}	212	683	23.2
Transverse reinforcement ratio	ρ_{sh}	0.01	0.03	0.00
Effective transverse reinforcement ratio	ρ_{sheff}	0.09	0.33	0.02
Transverse reinforcement yield stress (MPa)	$f_{y,sh}$	547	1426	253
Indicator variable for possibility of longitudinal rebar slip	a_{sl}	0.73	1.00	0.00
Stirrup spacing to effective depth ratio	s/d	0.27	0.68	0.12
Maximum normalized shear stress (MPa)	V_n	544	2090	68.5
Column cross-sectional area (mm^2)	A	107000	360000	23200

All independent variables are standardized to have a mean of zero and a standard deviation of 1. The lognormal transformation is applied for all target variables i.e., hysteresis model parameters and transformation is suggested by Haselton et al. [2] is applied to the three variables as per equation (1) to ensure compliance with past studies.

$$\begin{aligned} \rho_{sh,eff}' &= \log\left(0.02 + 40\rho_{sh,eff}\right) \\ \rho_{sh}' &= \log\left(0.02 + 40\rho_{sh}\right) \\ v' &= \log\left(0.1 + v\right) \end{aligned} \tag{1}$$

4. Random Forest Regression for Hysteresis Model Parameter Estimation

Random Forest is an ensemble learning algorithm used for classification and regression [24]. Ensemble learning methods are one of the popular applications of machine learning in which models with less predictive capacity (called weak learners) are aggregated to have a better prediction. Decision tree-based ensemble learning methods such as random forest, boosting, bagging, etc. aggregate various trees to produce a more generalized model. A decision tree is a nonparametric regression approach. A decision tree partitions the training data into distinct regions and generates a tree-like graph. The non-partitioned region is called the root node. The divided area is termed as either interior or terminal nodes depending upon further division. A terminal node represents decision/ prediction. Shortcomings of decision trees such as sensitivity to the small change, susceptibility towards overfitting are typically addressed in ensemble learning methods [21,26].

In the random forest, each tree is generated using the subset of data randomly sampled with replacement from the original dataset. To provide additional randomness, a subset of features is used instead of all features. The final result is decided based on the aggregated result of each decision tree. Suppose for the regression problem, the training dataset is represented by $\boldsymbol{T} = \{\boldsymbol{X}, \boldsymbol{y}\}$, where $\mathbf{X} = \{\mathbf{x}_i \in R^m\}_{i=1,2,3,\ldots,n}$ is feature matrix and $\boldsymbol{y} = \{y_i \in R\}_{i=1,2,3,\ldots,n}$ is a continuous dependent variable. m and n represent the number of features and the number of samples respectively. A random forest regression algorithm is presented below:

(a) Firstly, n_{tree} bootstrap sample sets with replacement from the original training dataset $\boldsymbol{T}$ are generated. The percentage of data points included in each bootstrap set is called sample fraction (*SF*). The data points not included in the bootstrap sample are called 'out-of-bag' (OOB) samples and the prediction error of such data points is called 'out-of-bag' (OOB) prediction errors.

(b) For each bootstrap set, regression trees are fitted i.e., n_{tree} trees are created. In this process, instead of utilizing all m features, m_{try} ($m_{try} < m$) features are randomly selected at the splitting node. This exercise is conducted till the stopping criteria - a minimum number of samples in leaf nodes (*nodesize*) is reached.

(c) Prediction for a given data point is given by averaging predictions from each decision tree.

4.1 Hyperparameter tuning

In the field of machine learning, the model learning process is governed by hyperparameters. Hyperparameters are tuned based on training data to achieve enhanced performance. For Random forest regression, four hyperparameters - n_{tree}, *SF*, m_{try}, *nodesize* can influence the model performance. As stated earlier, n_{tree} is the number of trees in the forest. A sufficiently larger value of n_{tree} is required to stabilize error and to reduce the chances of overfitting. The second hyperparameter, sample fraction (*SF*), represents the percentage of data points included in each bootstrap set. The lower value of *SF* can introduce more bias, while the higher value of *SF* may increase the chance of overfitting. The number of features used during each splitting is m_{try} that determines the diversity among each tree. Lastly, *nodesize* - *a* minimum number of samples contained in terminal nodes controls the tree's depth, i.e., the complexity of the random forest. Hyperparameters are optimized by minimizing OOB error using a grid search algorithm. In the grid search algorithm, the hyperparameter grid is created by varying each hyperparameter in a specific range. Hyperparameters set giving minimum OOB error is selected.[27]

4.2 Feature selection

Out of many features, it is necessary to select a subset of features that contribute significantly. The relative influence of each feature is determined using the impurity-based feature importance metric [28–30]. For each variable used for the node split, the mean squared error (MSE) is calculated before and after the split. Predictive accuracy after the node split is known as 'node impurity' and a feature that reduces this impurity is considered more important than those features that do not. The reduction in MSE for each feature across all the trees is determined and the feature with the most significant accumulated impact is considered the more important. A backward elimination approach is used to select the optimum set of features. In this approach, a series of random forest models are fitted by eliminating the least important variable every time. A feature set with the highest prediction accuracy is selected [29].

4.3 Cross-validation

It is necessary to assess model performance for unseen data i.e. observations on which model is not created. When new data is not available, some portion of existing data is separated, called validation data and the model is trained on for the rest of the data. Model performance can be susceptible to how training and validation subsets are divided. To overcome this issue, cross-validation (CV) procedures are employed. For this study, *k*–times repeated random sampling cross-validation is used. In this approach, data is randomly divided into training and validation sets for *k*-times and model performance is assessed each time [31].

4.4 Model performance indices

To assess the prediction accuracy of the random forest models two performance indices - coefficient of determination (R^2) and root mean squared error (RMSE) are used. Both matrices are defined as,

$$RMSE = \sqrt{\frac{\sum (y-\hat{y})^2}{l}} \tag{2}$$

$$R^2 = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2} \tag{3}$$

where, l represents validation sample size, $\hat{y}$ represents the model prediction, y represents observed response and $\bar{y}$ represents mean of the observed prediction.

4.5 Stepwise model training methodology

A comprehensive methodology is proposed to train a model incorporating hyperparameter optimization, feature selection and cross-validation procedures described in previous subsections. Fig. 2 provides an overview of the procedure and the stepwise description of the methodology is given below.

(a) Firstly, all m (Here m =14 as per section 3) features are selected for each hysteresis model parameter.
(b) 10- times repeated random sampling cross-validation is performed. Firstly, 80% of data is randomly selected as the training set and 20% of data is chosen as the validation set. This exercise is repeated for 10 times that will result in 10 train-validation set pairs called.
(c) For each 10 train-validation set pairs, a random forest model is fitted for the training set by optimizing hyperparameters using a grid search algorithm.
(d) For each of the fitted random forest models, validation set and training set accuracy is estimated and averaged.
(e) Similarly, for each fitted random forest model, an impurity-based feature importance score is estimated and averaged.
(f) A feature with the least average feature importance score is eliminated.
(g) Steps proposed in (b) - (f) is repeated with m-1, m-2,…, 1 features. A feature set with the highest cross-validation accuracy is selected.

(h) For the optimal feature set, we already have cross-validation results from the previous steps based on which cross-validation accuracy is reported. Out of 10 train-validation set pairs, one pair is randomly selected to develop the final model for future predictions.

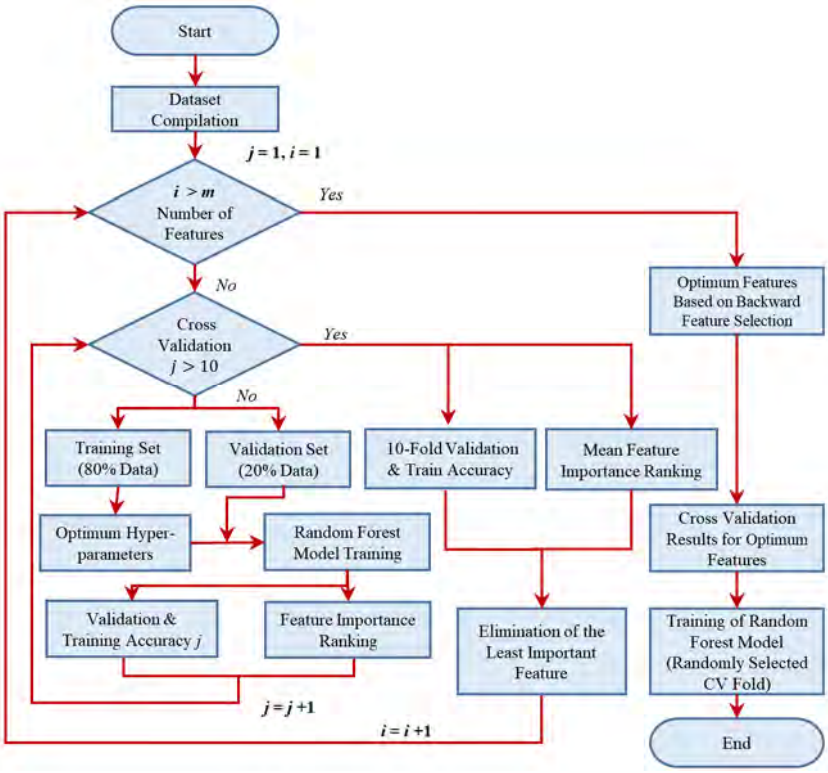


Fig -2 Proposed methodology for random forest regression

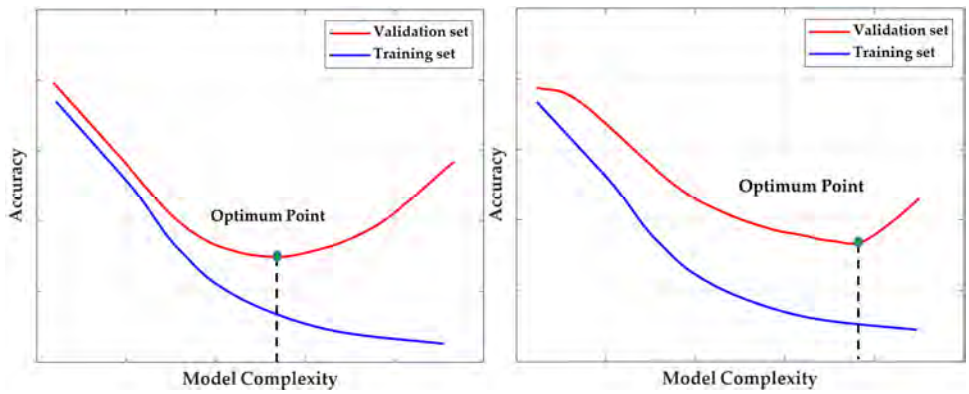
4.6 Model training results

Cross-validation accuracy for each of the hysteresis model parameters is presented in Table 2. As discussed previously, the final model is fitted for the optimum set of the feature for one of the training sets (80% data). The final model's validation accuracy is calculated and presented in Table 2.

Table 2 – Random forest regression model performance

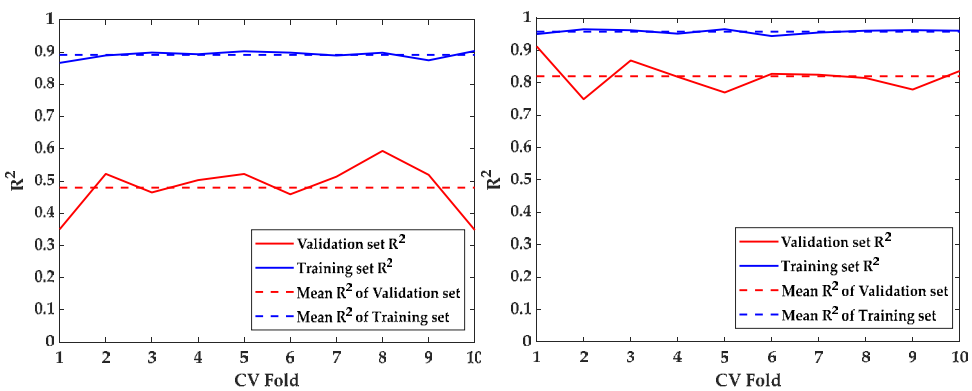
Hysteresis Model Parameter	Cross-Validation Results				Accuracy of the Final Validation Set			
	R^2		RMSE		R^2		RMSE	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation
M_y	0.97	0.89	0.09	0.17	0.99	0.95	0.10	0.15
EI_y/EI_g	0.96	0.83	0.11	0.23	0.97	0.82	0.12	0.23
EI_{40}/EI_g	0.91	0.65	0.15	0.29	0.93	0.72	0.14	0.30
M_c/M_y	0.80	0.27	0.05	0.10	0.89	0.29	0.05	0.11
$\theta_{cap,pl}$	0.88	0.47	0.25	0.53	0.94	0.53	0.23	0.55
θ_{pc}	0.85	0.24	0.40	0.86	0.92	0.42	0.40	0.88
λ	0.69	0.36	0.30	0.55	0.91	0.35	0.24	0.39

The results shown in Table 2 can be seen interpreted under the light of model complexity and sample size. Simple models often miss out important relationships between predictor and output. On the other hand, the complex model can fit random noise in the training data rather than intended outputs. Typical learning curves representing the trend of model error with model complexity is shown in Fig. 3 (a). For the random forest model, the complexity depends on the complexity of each tree. Hyperparameter *nodesize* majorly governs the tree complexity. Complexity is higher for a tree having a higher depth or small *nodesize* and vice versa. For example, for the extreme case when *nodesize* is one, terminal nodes include training data. In that case, if hyperparameter *SF* is 0.632, then approximately 63.2 % of trees will predict training data exactly. Hence training error is extremely low. In the opposite case, when *nodesize* is large, the prediction of each tree will be the average of data points present in the terminal node. In that case, training error will be higher. For validation error, there is an optimum point where validation error is minimum. Hyperparameter tuning is necessary to achieve optimum model complexity. As mentioned earlier, this study utilizes the cross-validation approach to tune the hyperparameters. Hence, it can be concluded that validation accuracy achieved is optimum. For this study, the optimum point is located in the region of higher complexity, as shown in Fig. 3 (b), leading to significantly lower training errors. This shows the importance of splitting data into train-validation sets for machine learning algorithms. Often it is observed that the accuracy achieved by training data is reported as model accuracy, particularly for the approaches with low flexibility such as linear regression. But accuracy of unseen data is the correct representation of the model's predictive capability. It is also noted that at the optimum point, the difference between validation and training set error is higher for the parameters related to strain hardening and post-capping regions. Models remember training data well but the prediction is poor for the unseen data that indicates the model's incapability to generalize the unlined complex relationship. This difference can be reduced through the availability of additional data in the future.



(a) (b)

Fig. 3 – (a) Typical learning curves (b) Learning curves for this study



(a) (b)

Fig. 4 – Cross-validation results for (a) $\theta_{cap,pl}$ (b) EI_y/EI_g.

Fig. 4 represents the accuracy of each cross-validation fold results for $\theta_{cap,pl}$, and EI_y/EI_g. A significant difference in the accuracy is observed among the folds. Hence, models are sensitive to the train-validation split which indicates heterogeneity in the data. In other words, the number of samples available in certain parts of sample space are less as compared to the other parts. Uncertainty in those regions is larger as we have less information regarding that region. So pointwise prediction of uncertainty is necessary as described in the subsequent section.

5. Uncertainty in Random Forest Regression

As mentioned earlier, the model prediction uncertainty estimates are not always directly available for the machine learning algorithms. However, they may be obtained by introducing additional techniques into the modeling process. In the last decade, efforts have been made to quantify uncertainty in random forests by constructing confidence intervals and prediction intervals [32–37]. Confidence interval accounts for the uncertainty in the model's conditional mean estimate with a certain level of confidence while prediction interval is associated with the inference of the new observation, not with the model prediction. When the assumption is made that data are independent and identically distributed (i.i.d.) α - percentile confidence or prediction interval is given by

$$I_\alpha = \hat{y} + d_\alpha(\hat{s}) \tag{4}$$

where $\hat{y}$ represents model prediction and $\hat{s}$ represents error margin. It is assumed that errors follow some distribution generally normal or t-distribution denoted by d . d_α represents the α - percentile of assumed distribution. Hence, for a given prediction, the error margin and related d_α are estimated and α - percentile confidence or prediction interval (I_α) is generated. For example, in the case of linear regression with known parameters, regression error is assumed to be normal and regression standard deviation is an estimator of the error margin. Based on which α - percentile prediction interval is created. As regression standard deviation is constant throughout the sample space, the prediction interval is constant.

It is noted that prediction intervals have practical implication than confidence interval because prediction interval is associated with the accuracy of the response variable itself, not just the accuracy of the mean value of the response estimated by the model. Hence, this study utilizes the four prediction interval estimation approaches for the random forest. Details of each approach are presented in subsequent sections.

5.1 Quantile regression forest (QRF)

QRF proposed by Meinshausen [32] provides the conditional distribution of the response for a given feature vector. As mentioned earlier, the prediction of a tree is an average of samples located in terminal nodes. In QRF for a given feature vector, in addition to the average value if the entire distribution of response values at the terminal node is returned. Based on this distribution α - percentile prediction interval is created. Hence, this approach gives response distribution without any assumption of the error distribution.

5.2 Prediction interval based on 'out of bag' prediction error

The study proposed by Zheng et al. [36] aims to determine prediction intervals based on error distribution. As stated in the previous section that each tree in the random forest is constructed from bootstrap samples. Hence, some data points are not included in the given tree. For that tree, this 'out-of-bag' data point is analogous to the validation data point. Approximately $\left(1-n^{-1}\right)^n \approx \mathrm{e}^{-1} \approx 0.368\,(36.8\%)$ of total trees do not contain a particular training data point. For the given data point, a random forest is developed by aggregating the rest of the 63.2 % trees. Error in this prediction can be treated as a proxy of the prediction error. If this exercise is repeated for all training data and the distribution of error is achieved. Consequently, for a given training data set, one common error distribution is achieved and it is assumed that the error for the new prediction will follow the same distribution. Hence prediction interval determined based on error distribution is constant for all predictions. However, the interval may be asymmetric. Instead of determining the constant error distribution throughout the feature space, Lu and Hardin [38] proposed a unified framework that

provides conditional distribution for a given feature vector by weighting the quantiles. Weights are determined based on the training observations' proximity to the test observation.

5.3 An approximate method for the prediction interval

Random forest prediction is given by the average of individual tree predictions. Similarly, the standard deviation for individual tree prediction can be obtained. Coulston et al. [37] proposed an approximate method in which this standard deviation tree predictions can be used as error margin as per equation (4). Further, the bootstrap method is proposed to determine the empirical distribution of d_α. Based on which α-percentile prediction interval is created.

5.4 Prediction interval results

95th percentile prediction intervals for random forest based on the four methods discussed earlier are constructed for the final validation set for $\theta_{cap,pl}$, and EI_y/EI_g and centered values are presented in Fig. 5 In addition, prediction interval based on linear regression-based semi-empirical equations proposed by Haselton et al. [2] is developed. While developing prediction intervals using QRF, quantile values were calculated for each node. Often tree node size is very small which leads to unusual distribution. This drawback is eliminated in 'out-of-bag' error-based prediction intervals. Linear regression-based prediction interval is symmetric and constant while the method proposed by Zheng et al. [36] provides an asymmetric constant prediction interval. As discussed earlier, that data is highly heterogeneous. Providing constant prediction interval will penalize the regions where we are more confident i.e. in the regions where we have training data. It is evident from Fig-5 that by weighing 'out-of-bag' error based on training data as suggested by Lu and Hardin [38], the prediction intervals change significantly as compared to constant prediction intervals. Prediction intervals based on an approximate method by Coulston et al. [37] follow a similar trend with the method proposed by Lu and Hardin [38]. However, prediction intervals are narrower for Coulston et al. [37] approach for these cases.

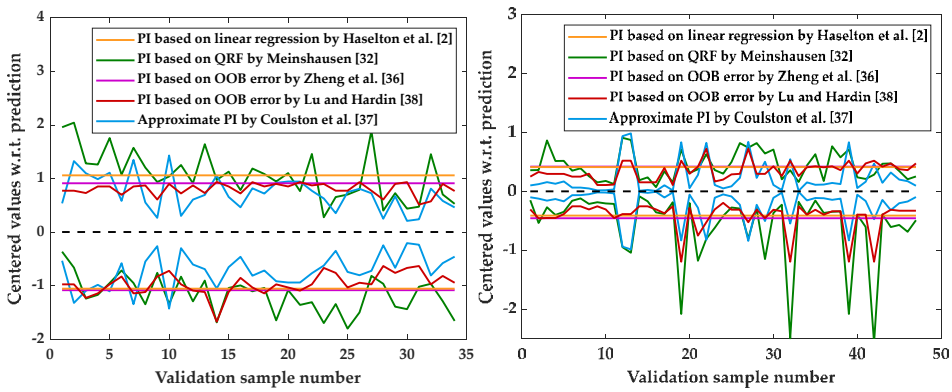


Fig. 5 - Prediction intervals (PI) for (a) $\theta_{cap,pl}$, (b) EI_y/EI_g.

6. Conclusion

This study utilizes a random forest regression approach for hysteresis model parameter estimation. The comprehensive methodology has been proposed to develop random forest-based regression models incorporating hyperparameter optimization, feature selection and cross-validation. The results show the importance of splitting data into training and validation sets which is not adopted for existing linear regression-based approaches. It was observed that models remember training data well but the prediction is poor for the unseen data that indicates the model's incapability to generalize the unlined complex relationship. This difference can be reduced through the availability of additional data in the future.

The majority of literature related to the application of machine learning algorithms focuses on prediction accuracy. However, estimation prediction uncertainty is also an important aspect. This study compiles various approaches for quantifying uncertainty in random forest regression and uncertainty is quantified as prediction interval. Results reveal that the prediction interval can change as much as 50% for certain data points, as compared to the prediction interval based on constant standard deviation. It is also noted that the versatile distribution of the response is obtained for each data point and prediction intervals are constructed based on this distribution. This approach provides flexibility, but on the other hand, it creates an obstacle for easy adoption by researchers and stakeholders. Hence, future study will explore methodologies to provide closed-form prediction uncertainty estimate at each data point. In addition, a framework will be formulated to propagate the pointwise prediction uncertainty in seismic fragility and risk assessment.

7. Acknowledgments

The first author would like to acknowledge the support of the Ministry of Human Resource Development, Government of India through the Prime Minister's Research Fellowship (PMRF). The second author would like to acknowledge the funding from Science and Engineering Research Board (statutory body under the Department of Science and Technology, India) through Grant No. MTR/2019/000287 for this work.

8. References

[1] Deierlein GG, Reinhorn AM, Willford MR (2010): Nonlinear structural analysis for seismic design a guide for practicing engineers, NEHRP seismic design technical brief, NEHRP, New York.

[2] Haselton CB, Liel AB, Taylor-lange SC, Deierlein GG (2016): Calibration of model to simulate response of reinforced concrete beam-columns to collapse, **113** (6), 1141–52.

[3] Elwood J (2004): Modelling failures in existing reinforced concrete columns, *Canadian Journal of Civil Engineering*, **31** (5), 846–59.

[4] Aschheim M, Hernández-Montes E, Vamvatsikos D (2019): *Design of Reinforced Concrete Buildings for Seismic Performance Practical Deterministic and Probabilistic Approaches*, CRC Press.

[5] Takeda T, Sozen M, Nielsen N (1970): Reinforced concrete response to simulated earthquakes, *Journal of the Structural Division*, **96** (12), 2557–73.

[6] Sezen H, Chowdhury T (2009): Hysteretic model for reinforced concrete columns including the effect of shear and axial load failure, *Journal of Structural Engineering*, **135** (2), 139–46.

[7] Ibarra LF, Medina RA, Krawinkler H (2005): Hysteretic models that incorporate strength and stiffness deterioration, *Earthquake Engineering and Structural Dynamics*, **34** (12), 1489–511.

[8] Clough RW, Johnston SB (1966): Effect of stiffness degradation on earthquake ductility requirements, *Proceedings Second Japan National Conference on Earthquake Engineering*.

[9] Lee CS, Han SW (2021): An accurate numerical model simulating hysteretic behavior of reinforced concrete columns irrespective of types of loading protocols, *International Journal of Concrete Structures and Materials*, **15** (1), 1–16.

[10] Huang H, Burton HV, Sattar S (2020): Development and utilization of a database of infilled frame experiments for numerical modeling, *Journal of Structural Engineering*, **146** (6).

[11] Tariq H, Jampole EA, Bandelt MJ (2021): Development and application of spring hinge models to simulate reinforced ductile concrete structural components under cyclic loading, *Journal of Structural Engineering*, **147** (2).

[12] Dai KY, Yu XH, Lu DG (2020): Phenomenological hysteretic model for corroded RC columns, *Engineering Structures*, Elsevier. **210**.

[13] Li G, Yu L, Dong Z (2019): Simplified collapse analysis model for RC frames with cyclic deterioration behaviors, *Journal of Earthquake Engineering*, 1–27.

[14] FEMA (2009): Quantification of building seismic performance factors, FEMA P-695 Washington, DC, USA.

[15] American Society of Civil Engineers (2017): Seismic evaluation and retrofit of existing buildings, ASCE/SEI, 41-17, USA.

[16] Panagiotakos TB, Fardis MN (2001): Deformations of reinforced concrete members at yielding and ultimate, *ACI Structural Journal*, **98** (2), 135–48.

[17] Lee CS, Han SW (2018): Computationally effective and accurate simulation of cyclic behaviour of old reinforced concrete columns, *Engineering Structures*, **173** 892–907.

[18] Han SW, Lee CS, Zambrana MAP, Lee K. (2019): Calibration factor for asce 41-17 modeling parameters for stocky rectangular RC columns, *Applied Sciences*, **9** (23), 19–22.

[19] Sun H, Burton H V., Huang H (2021): Machine learning applications for building structural design and performance assessment: state-of-the-art review, *Journal of Building Engineering*, **33**.

[20] Xie Y, Ebad Sichani M, Padgett J, DesRoches R (2020): Machine learning applications in earthquake engineering: literature review and case studies. *Proceedings of 17th World Conference on Earthquake Engineering World Conference on Earthquake Engineering*, Sendai, Japan.

[21] Hastie T, Tibshirani R, Friedman J (2009): *The Elements of Statistical Learning Data Mining, Inference, and Prediction,* Springer, 2nd edition.

[22] Luo H, Paal SG (2018): Machine learning–based backbone curve model of reinforced concrete columns subjected to cyclic loading reversals, *Journal of Computing in Civil Engineering*, **32** (5).

[23] Liu Z, Li S (2019): Development of an ann-based lumped plasticity model of RC columns using historical pseudo-static cyclic test data, *Applied Sciences*, **9** (20).

[24] Breiman L (2001): Random forest, *Machine Learning*, **45**.

[25] Berry M, Parrish M, Eberhard M (2004): PEER structural performance database user's manual.

[26] Breiman L, Friedman J, Stone CJ, Olshen RA (1984): *Classification and Regression Trees,* CRC Press.

[27] Probst P, Wright MN, Boulesteix AL (2019): Hyperparameters and tuning strategies for random forest, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **9** (3), 1–15.

[28] Tyralis H, Papacharalampous G, Langousis A (2019): A brief review of random forests for water scientists and practitioners and their recent history in water resources, *Water*, **11** (910).

[29] Xu Z, Lian J, Bin L, Hua K, Xu K, Chan HY (2019): Water price prediction for increasing market efficiency using random forest regression: a case study in the western united states, *Water*, **11** (2), 1–20.

[30] Biau G, Scornet E (2016): A random forest guided tour, *Test*, **25** (2), 197–227.

[31] Kuhn M, Johnson K (2013): *Applied Predictive Modeling with Applications in R*, Springer.

[32] Meinshausen N (2006): Quantile regression forests. *Journal of Machine Learning Research*, **7** (6), 983–99.

[33] Mentch L, Hooker G (2016): Quantifying uncertainty in random forests via confidence intervals and hypothesis tests, *Journal of Machine Learning Research*, **17** (1), 1–41.

[34] Wager S, Hastie T, Efron B (2014): Confidence intervals for random forests: the jackknife and the infinitesimal jackknife, *Journal of Machine Learning Research*, **15** (1), 1625–51.

[35] Sexton J, Laake P (2009): Standard errors for bagged and random forest estimators, *Computational Statistics and Data Analysis*, Elsevier B.V. **53** (3), 801–11.

[36] Zhang H, Zimmerman J, Nettleton D, Nordman DJ (2020): Random forest prediction intervals, *The American Statistician*, **74** (4), 392–406.

[37] Coulston JW, Blinn CE, Thomas VA, Wynne RH (2016): Approximating prediction uncertainty for random forest regression models, *Photogrammetric Engineering and Remote Sensing*, **82** (3), 189–97.

[38] Lu B, Hardin J (2021): A unified framework for random forest prediction error estimation, *Journal of Machine Learning Research*, **22** (8), 1–41.

