Data-Driven Prediction of Seismic Drift Response in Buildings

Using Machine Learning and Real Instrumented Building Data

Jason Nketia

CEE, Duke University

Date: December 14, 2025

Abstract

This report investigates the use of machine learning models to predict interstory drift ratios in buildings subjected to earthquake ground motion. Using the PEER NDE1.0 dataset, which contains over 12,500 records of real building response measurements, a structured pipeline was developed involving feature engineering, unit normalization, missing-data handling, and model training using Linear Regression and Random Forest Regression. Because the dataset consists almost entirely of elastic building responses, traditional performance-level classification (IO/LS/CP) is not feasible; instead, continuous drift prediction is pursued. Model performance is evaluated using MAE, RMSE, and R^2, selected for their interpretability and relevance to deformation-based seismic engineering. Results show that PGV, spectral acceleration (SA1), Arias Intensity, and PGA are the most significant predictors of drift. The findings demonstrate the potential and limitations of data-driven approaches for performance-based seismic design.

1 Introduction

Performance-Based Seismic Design (PBSD) relies on accurate prediction of structural response, particularly interstory drift, which is widely recognized as a primary indicator of earthquake-induced damage and deformation demand. Traditional PBSD procedures typically require nonlinear time-history analysis, which is computationally intensive and difficult to scale across large building inventories.

This project investigates whether machine learning (ML) models can estimate drift directly from structural parameters and ground-motion intensity measures. The analysis is based on the NDE1.0 flat-file dataset of real earthquake recordings obtained from instrumented buildings, which includes building metadata, ground-motion measures, spectral parameters, duration metrics, and observed drift responses (1).

Two regression models are examined: a Linear Regression baseline and a Random Forest model. By evaluating prediction accuracy and identifying the most influential structural and seismic predictors, this study assesses the feasibility and limitations of ML-based surrogate models for rapid drift estimation within PBSD workflows.

2 Background

2.1 Seismic Intensity Measures

Earthquake ground motions are commonly described using a set of intensity measures that capture different characteristics of shaking amplitude, frequency content, and duration. Peak Ground Acceleration (PGA), while Peak Ground Velocity (PGV), by contrast, is closely related to displacement response and is often a good predictor of drift.

Peak Ground Displacement (PGD) captures long-period components of ground motion and can affect flexible structures. Energy-based measures such as Arias Intensity, defined as the time integral of squared acceleration (2), quantify the cumulative input energy that a structure experiences, which can affect cyclic deformation demands. Spectral acceleration measures, such as short-period spectral acceleration (SA1), approximate the acceleration response of a single-degree-of-freedom system with a period near the fundamental mode of typical buildings. These spectral values provide a direct link between ground-motion frequency content and modal response.

Duration metrics, including effective duration and the significant duration between 5% and 95% of Arias Intensity (3), describe how long a structure is subjected to strong shaking. Longer durations generally increase cumulative deformation and therefore influence drift demands even when peak intensity measures are moderate.

2.2 Building Dynamic Properties

Structural response to seismic excitation depends strongly on building dynamic properties. The fundamental frequency (F_1) characterizes the primary mode of vibration and is inversely related to stiffness: more flexible structures have lower F_1 values and typically experience larger drift demands (4). The second-mode frequency (F_2) provides additional information about the building's higher-mode behavior, which can contribute to drift in mid-rise structures.

Geometric properties such as building height and the number of stories influence the lateral stiffness distribution and, therefore, the modal periods. Typology (e.g., reinforced concrete, steel, or composite systems) also reflects differences in stiffness, mass, and energy dissipation. Site conditions, commonly represented using the average shear-wave velocity in the top 30 meters (V_{s30}), affect soil amplification: softer soils (lower V_{s30}) can increase shaking intensity at longer periods and thereby elevate drift demands.

2.3 Interstory Drift

Interstory drift is defined as the relative lateral displacement between two consecutive floors divided by the story height. It is a dimensionless measure of deformation and is widely recognized as the primary indicator of damage in buildings during earthquakes. Drift ratios govern nonstructural damage, structural cracking, yielding, and ultimately instability. Within PBSD frameworks, performance levels such as Immediate Occupancy (IO), Life Safety (LS), and Collapse Prevention (CP) are expressed in terms of allowable drift thresholds (5). Accordingly, accurate estimation of drift is essential for assessing structural performance under seismic loading and forms the central objective of this study.

3 Dataset Description

The analysis in this study is based on the NDE1.0 flat-file dataset, which contains 12,503 records of structural response from instrumented buildings subjected to real earthquakes. Each record corresponds to a building–earthquake pair and includes metadata characterizing the building, ground-motion intensity measures, spectral parameters, duration metrics, and response quantities such as drift and peak floor motions. The flat-file used in this project was obtained directly from the publicly accessible repository hosted by ISTerre (1), and its structure and content follow the specifications described in the accompanying dataset publication (6).

3.1 Feature Definitions

Table 1 summarizes the primary variables used in the drift-prediction models. These features were selected based on their established relevance to seismic deformation demand, including velocity-based measures (PGV), spectral accelerations near the first-mode period (SA1), energy-based measures (Arias Intensity), and duration metrics associated with cumulative damage potential. Building properties such as height and fundamental frequency (F_1) capture structural stiffness and modal characteristics, while the drift ratio serves as the engineering demand parameter of interest.

Table 1: Summary of Key Variables Used in the Drift Prediction Models

Variable	**Definition**	**Units**
Vs30	Average shear-wave velocity (site condition proxy)	m/s
Height	Building height	m
No._of_story	Number of stories	–
Typology	Structural system (encoded as dummy variables)	–
F1	First natural frequency	Hz
F2	Second natural frequency	Hz
PGA	Peak Ground Acceleration	g
PGV	Peak Ground Velocity	m/s
PGD	Peak Ground Displacement	m
AI	Arias Intensity	cm/s
SA1	Spectral Acceleration (short period)	g
SA2	Spectral Acceleration (longer period)	g
Effective	Effective duration of strong motion	s
Significant$_{5-95}$	Significant duration (Arias intensity, 5–95%)	s
Drift	Maximum interstory drift ratio (target variable)	–

3.2 Data Cleaning and Preparation

The raw flat-file contains building metadata, event information, ground-motion measures, spectral ordinates, frequency estimates, and several response quantities. Prior to model development, the dataset was cleaned and reformatted to produce a consistent, analysis-ready feature matrix. The primary steps were as follows:

- **Unit normalization**: Several variables were provided in non-SI units (e.g., cm, cm/s, cm/s^2). These were converted to meters, meters per second, and units of g for consistency and interpretability.

- **Feature selection**: Variables representing *observed* structural response—such as peak top acceleration (PTA), peak top velocity (PTV), and peak top displacement (PTD)—were removed to avoid data leakage, since the goal is to predict drift using only input parameters.

- **Handling missing values**: Features with limited missing data (e.g., V_{s30} or F_1) were imputed using median values, ensuring statistical consistency without altering distributional characteristics.

- **Encoding categorical variables**: Building typology was converted into dummy variables through one-hot encoding.

- **Column restructuring**: The original flat-file stores variable names and units in a multi-row header. These were consolidated into single-level column names using a dictionary mapping derived from the header rows.

The resulting cleaned dataset contains only predictor variables that represent building properties or ground-motion characteristics, along with the drift ratio as the target variable. This ensures that the machine learning models are trained exclusively on physically meaningful inputs suitable for predictive modeling.

4 Methodology

4.1 Model Development

The objective of this study is to develop data-driven surrogate models capable of predicting maximum interstory drift from structural and ground-motion parameters. Because the drift values in the NDE1.0 dataset fall almost exclusively within the Immediate Occupancy range, all classification models (Linear and Random Forest classifiers) were discarded due to the complete absence of meaningful class diversity. Only regression models were retained, as drift remains a continuous variable even within the elastic regime. Consequently, the analysis focuses solely on continuous drift prediction.

Two supervised regression models were implemented:

1. **Linear Regression**: This model establishes a baseline for understanding the approximate linear relationships between drift and the selected predictor variables. In the elastic range, many structural response quantities scale linearly or quasi-linearly with ground-motion intensity and modal properties (4), making Linear Regression a meaningful benchmark for evaluating predictive performance.

2. **Random Forest Regression**: Random Forests are ensemble models that average the predictions of multiple decision trees to reduce variance and capture nonlinear interactions (7). This model is well suited for heterogeneous predictor sets, such as the combined structural and seismic parameters used here. Random Forests can model threshold behaviors, nonlinear dependencies, and variable interactions that Linear Regression cannot represent.

Both models were trained using an 80/20 split of the cleaned dataset. The Random Forest model was trained on unscaled features, while the Linear Regression model used standardized predictors to avoid numerical dominance by variables with large magnitudes.

4.2 Evaluation Metrics

Model performance was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). These metrics were selected to reflect both engineering relevance and statistical robustness.

- **MAE**: Provides the average absolute difference between predicted and observed drift. Because drift ratios in this dataset are small in magnitude and physically interpretable, MAE offers a clear measure of typical prediction error.

- **RMSE**: Penalizes larger errors more strongly than MAE. This is critical for seismic engineering applications, where occasional underestimation of drift—even by small absolute amounts—could be structurally significant. RMSE, therefore, evaluates the model's reliability in higher-demand cases.

- R^2: Measures the proportion of variance in drift explained by the model. Due to the limited spread of drift values in the elastic dataset, moderate R^2 values (approximately 0.5–0.6) still indicate meaningful predictive ability. R^2 complements the error metrics by assessing how well the model captures the underlying physical relationships.

Together, these three metrics provide a balanced assessment of accuracy, robustness, and explanatory power for data-driven drift prediction, enabling a fair comparison between the linear and nonlinear regression models.

5 Results

5.1 Model Performance

Table 2 summarizes the models' performance on the test set. Although both models show small absolute errors, reflecting the low magnitude of drift values, the Random Forest model achieves modest improvements in MAE, RMSE, and R^2.

Table 2: Regression Performance for Drift Prediction

Model	**MAE**	**RMSE**	R^2
Linear Regression	5.48e–05	1.98e–04	0.495
Random Forest Regression	3.76e–05	1.84e–04	0.564

The Linear Regression model explains approximately 49.5% of the variance in drift, which is reasonable given the narrow distribution of drift ratios. The Random Forest model increases the explained variance to 56.4%, indicating that nonlinear relationships and interactions among predictors contribute modestly to drift behavior, even within the elastic regime.

Because predictor variables were standardized, the Linear Regression coefficients reflect the relative strength of each feature's linear association with drift.

5.2 Feature Importance from Random Forest Regression

Figure 1 shows the feature importance ranking from the Random Forest model. Unlike Linear Regression, which identifies only global linear trends, Random Forest regression captures threshold effects, nonlinear dependencies, and interaction terms between seismic and structural variables.

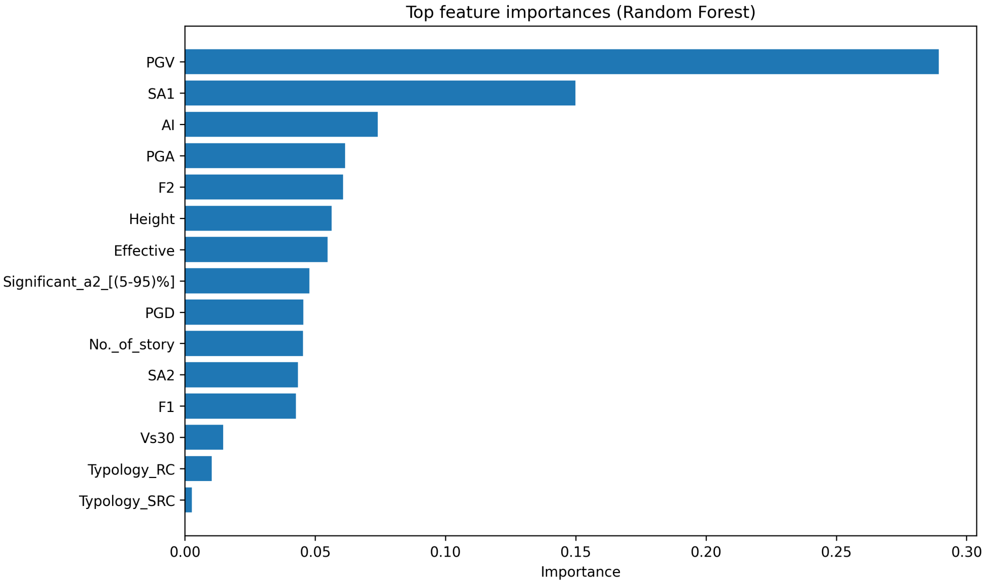

Figure 1: Feature importance ranking from the Random Forest Regression model.

The dominant predictors closely align with established physical intuition:

- **PGV** is the strongest predictor of drift, reflecting the tight relationship between velocity demand and displacement response.
- **SA1** captures first-mode spectral acceleration, which is directly tied to modal displacements in elastic systems.
- **Arias Intensity (AI)** measures the cumulative energy input, influencing cyclic deformation even when peak measures are moderate.
- **PGA** contributes through high-frequency acceleration pulses.
- **Modal frequency** (F_2) indicate building flexibility: lower frequencies correspond to more deformable structures and higher drift.
- **Height** and **number of stories** also appear prominently, consistent with taller or more flexible structures experiencing greater drift.

The agreement between model-derived importance and structural dynamics theory lends credibility to the regression results and demonstrates that even simple data-driven models can meaningfully identify the primary drivers of elastic drift.

6 Discussion and Conclusion

Discussion

The results of this study demonstrate that several structural and ground-motion parameters exert clear and physically meaningful influences on elastic drift response. The trends captured by the regression models are consistent with classical modal dynamics, where deformation demands scale with both the intensity and the frequency content of ground motion (4). The dominance of PGV and SA1 in both the linear coefficients and the Random Forest importance ranking reflects their direct connection to first-mode displacement response. PGV corresponds closely to velocity-driven deformation, while SA1 approximates spectral acceleration near the fundamental period, reinforcing the importance of modal amplification in elastic drift behavior.

Site conditions also played a measurable role. Although Vs30 exhibited a smaller relative influence compared to intensity measures, its contribution agrees with the concept of site amplification, where softer soils amplify longer-period components of shaking and thereby modify drift demand. Duration metrics, including Effective Duration and Significant Duration (5–95%), were found to be meaningful predictors as well. These variables capture the cumulative effects of repeated inelastic cycles in more nonlinear contexts, but even in the elastic regime, they help explain small increases in drift associated with extended strong-motion shaking.

Building characteristics such as height, number of stories, and modal frequencies also impacted drift, consistent with the height–period relationship. Taller, more flexible structures tend to exhibit lower natural frequencies and larger modal displacements, a trend reflected in both the negative coefficient on F_1 and the elevated importance of height-related variables. These observations reinforce that the data-driven model is consistent with fundamental structural dynamics principles.

Despite these insights, several limitations constrain the generality of the findings. Foremost, the dataset contains almost exclusively elastic drift values, preventing any meaningful classification into PBSD performance levels such as Life Safety or Collapse Prevention. This data imbalance ruled out supervised classification entirely and limited the scope of the predictive models to elastic drift estimation. In addition, several key structural properties—such as stiffness distribution, plan dimensions, and torsional characteristics—are absent from the dataset, restricting the model's ability to account for building-specific response mechanisms. The available ground-motion metadata also lacks full response spectra or directionality measures, limiting the ability to model higher-mode or directional effects.

Future extensions of this work should incorporate nonlinear structural simulations to generate synthetic LS and CP drift values, enabling classifier development and broadening PBSD relevance. Physics-based surrogate models such as mimoSHORSA (8) could also be explored to complement or benchmark the ML models. Further improvements may come from employing more advanced intensity measures, especially duration-modulated spectral parameters or spectral shape indicators, which may better capture the interaction between ground-motion features and structural flexibility.

Conclusion

This study developed a data-driven pipeline for predicting interstory drift in buildings subjected to earthquake ground motions using the NDE1.0 dataset. After eliminating classification models due to the absence of non-elastic drift values, two regression models were evaluated. Linear Regression provided a transparent baseline, while Random Forest Regression captured modest nonlinear patterns, achieving improved predictive accuracy. The analysis identified PGV, SA1, and Arias Intensity as dominant predictors of drift, with duration metrics and building geometry also contributing meaningfully.

Overall, the results show that machine learning can effectively predict elastic drift using readily available structural and seismic parameters, supporting preliminary PBSD assessments. However, the limited variability of drift in the dataset constrains the models' applicability beyond the elastic range. Incorporating nonlinear structural simulations and richer ground-motion metadata would be the next essential steps toward extending data-driven approaches to damage-state prediction and full PBSD workflows.

References

[1] ISTerre – Institut des Sciences de la Terre, "Access to the nde1.0 flatfile dataset." `https://www.isterre.fr/annuaire/pages-web-du-personnel/philippe-gueguen/new-earthquake-data-recorded-in-buildings-nde1-0/article/acces-to-the-flatfile.html`, 2025. Accessed: 2025-12-12.

[2] A. Arias, "A measure of earthquake intensity," in *Seismic Design for Nuclear Power Plants* (R. J. Hansen, ed.), pp. 438–483, MIT Press, 1970.

[3] M. D. Trifunac and A. G. Brady, "A criterion for strong motion duration," *Bulletin of the Seismological Society of America*, vol. 65, no. 3, pp. 581–626, 1975.

[4] H. P. Gavin, "Structural dynamics: Course notes and numerical resources." `https:`

//people.duke.edu/~hpgavin/StructuralDynamics/, 2001–2020. Accessed: 2025-12-12.

[5] FEMA, "Seismic performance assessment of buildings," Tech. Rep. FEMA P-58, Federal Emergency Management Agency, 2018.

[6] A. L. Astorga, P. Guéguen, S. Ghimire, and T. Kashima, "Nde1.0: A new database of earthquake data recordings from buildings for engineering applications," *Bulletin of Earthquake Engineering*, vol. 18, pp. 1321–1344, 2020.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] H. P. Gavin and S.-S. Yau, "Higher-order response surface methods for structural reliability analysis," *Structural Safety*, vol. 30, no. 3, pp. 173–183, 2008.