Title
A Data-driven Building Seismic Response Prediction Framework: from Simulation and Recordings to Statistical Learning

Author
SUN, HAN

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Data-driven Building Seismic Response Prediction Framework: from Simulation and Recordings to Statistical Learning

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Civil Engineering

by

Han Sun

2019

ABSTRACT OF THE DISSERTATION

A Data-driven Building Seismic Response Prediction Framework: from Simulation and Recordings to Statistical Learning

by

Han Sun

Doctor of Philosophy in Civil Engineering

University of California, Los Angeles, 2019

Professor John Wright Wallace, Chair

Structural seismic resilience society has been grown rapidly in the past three decades. Extensive probabilistic techniques have been developed to address uncertainties from ground motions and building systems to reduce structural damage, economic loss and social impact of buildings subjected to seismic hazards where seismic structural responses are essential and often are retrieved through Nonlinear Response History Analysis. This process is largely limited by accuracy of model and computational effort. An alternative data-driven framework is proposed to reconstruct structure responses through machine learning techniques from limited available sources which may potentially benefit for "real-time" interpolating monitoring data to enable rapid damage assessment and reducing computational effort for regional seismic hazard assessment. It also provides statistical insight to understand uncertainties of seismic building responses from both structural and earthquake engineering perspective.

The dissertation of Han Sun is approved.

Henry J. Burton

Ali Mosleh

Jonathan Paul Stewart

John Wright Wallace, Committee Chair

University of California, Los Angeles, 2019

2019

To my parents ...

For their encouragement on science

Table of Contents

List of Figures

List of Tables

Acknowledgement

The research presented in this dissertation is supported by National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation research grant number 1538866. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the authors and do not necessarily reflect the views of the sponsor.

I would like to thank my Ph.D. advisor, Dr. John W. Wallace for his continuous support and encouragement. I would also like to express my sincere gratitude to Dr. Henry V. Burton for his dedicated guidance and mentoring. Besides, I sincerely thank Dr. Jonathan P. Stewart and Dr. Arash A. Amini for their technical advices.

My special thanks to my parents for their unconditional support and sacrifices. I dedicate this dissertation to the memory of my grandmother Longxian Hu, who set a perfect example for me on honesty, bravery and diligence and is always in my heart. At last, I would like to acknowledge my wife, Xin Qian, for her professional suggestions and all the time we share from the very beginning.

VITA

2008-2012 B.Sc. in Civil Engineering, Hong Kong Polyethnic University, Hong Kong

2012-2014 M.Sc. in Civil Engineering, University of Michigan, Ann Arbor, MI

2014-2019 Ph.D. candidate in Civil Engineering, University of California, Los Angeles, CA

2016-2018 M.Sc. in Statistics, University of California, Los Angeles, CA

1. Introduction

1.1 Overview

Earthquakes cause significant structural damage that can result in consequential economic loss due to operational downtime and monetary cost related to inspections and repair. In particular, in seismically active regions such as California, Japan, and New Zealand are facing critical challenges related to lower magnitude, more frequent earthquakes resulting in extensive efforts to complete post-earthquake inspection on large inventories of buildings. In addition, there is an increasing probability of high magnitude earthquake occurrences on active faults (e.g., San Andrews Fault, Puente Hills, etc.) that are likely to create significant challenges for the impacted areas related to post-earthquake inspections, repair and retrofit, and recovery as observed in Los Angeles, Taiwan and Mexico City [1–3].

Besides the time, effort, and expense of inspection, which is typically accomplished visually, building damage identification can also be achieved by Structural Health Monitoring (SHM). In specific, SHM involves remote sensing devices installed in the building to retrieve structural response histories (e.g., accelerations, displacements, strains) subjected to the seismic hazard (ground shaking), to determine Engineering Demand Parameters (EDP) such as peak floor acceleration and peak story drift ratio and use these data to identify and assess the importance of the damage. However, due to cost constraints related to installation and maintenance, SHM systems typically involve a relatively limited number of sensors in a limited number of buildings; to date, the vast majority of installations involve using accelerometers. Therefore, a major challenge associated with SHM systems involves data interpretation in buildings with sensors, as well as the lack of sensors in some buildings. An attractive option to address some of these issues involves reconstructing building seismic responses to gain additional value from SHM

installations, in both instrumented and non-instrumented buildings.

In addition to damage identification needs for seismic building responses subject to earthquake shaking, modern seismic risk assessment frameworks such as Performance-Based Earthquake Engineering (PBEE) developed by the Pacific Earthquake Engineering Research (PEER) Center [4–6] typically require conducting Nonlinear Response History Analysis (NRHA) to acquire EDPs from a suite of ground motions [7,8]. It is time consuming and computationally challenging to conduct such evaluations for complex structural systems, and an even more challenging to scale this effort up to a cluster of buildings or an inventory of buildings for a city or regional risk assessment [9,10]. Alternative approaches currently available within the earthquake engineering field to address these issues are to either estimate EDPs using site Intensity Measures (IM) such as Spectrum Acceleration (SA) instead of running NRHA simulations or to reduce model complexity. The former approach has been used since 1999 [11] as part of HAZUS estimation procedure [12] and applied in estimating earthquake losses [13–15]. The latter approach has been applied for regional seismic risk assessment studies (e.g., [16]).

An alternative approach is presented here to interpolate EDPs for a cluster of buildings at various scales using limited data sets from instrumented buildings. The framework presented aims to gain additional insight from instrumented buildings and also mitigate the computational effort of scenario-based risk assessment. To accomplish these goals, modern machine learning models are applied to predict seismic building responses at specified locations. The framework includes issues related to collecting and filtering data obtained from both simulation results and sensor recordings, scenario design, and model performance evaluation. Depending on the data available, a variety of scenarios are developed.

1.2 Research Significance

The objective of the research reported is to develop a data-driven framework that predicts building seismic responses for a given building located at a given site using available recordings from co-regionally located buildings subjected to the same event, and recordings from actual buildings subjected to prior events. The goals of the research are to enable rapid 1) damage identification for individual buildings that are either instrumented or un-instrumented during a seismic hazard; 2) regional risk estimation for a given seismic event, and 3) risk assessment for a range of domains, including for individual buildings, limited (local) and large (city or regional) clusters of buildings, as well as various building types using models with various complexities. The data presented, mythologies applied, and evaluation approaches considered are used to demonstrate that the machine learning framework proposed provides valuable tools to address SHM and risk assessment challenges facing the earthquake engineering community.

An illustration connecting the framework with a seismic event is provided in Figure 1.1. Earthquakes are caused by fault moving as a consequence of tectonic plates moving. The shaking generates seismic waves and travels through rock layers underneath the ground and soil layers. The shaking wave also travels upwards and is altered by soil layers and reaches the ground surface as is referred to as a ground motion response history, $y_{ground}(t)$, typically measured by accelerometers located on the ground surface. The shaking is further altered by the building system and produces seismic building response histories, $y_{roof}(t)$ as measured at the roof level and $y_{\text{int}ermediate}(t)$ as measured at intermediate levels. A number of buildings are instrumented with accelerometers at ground level and along the building height to record these building responses at seismic active regions provided by organizations such as United States Geological Survey and California Geological Survey. By retrieving these response data and converting them to EDPs,

building physical damage and its consequential monetary loss can be estimated. The core of this dissertation is to construct a framework that simulates the above-mentioned sophisticated scenario through data-driven means.

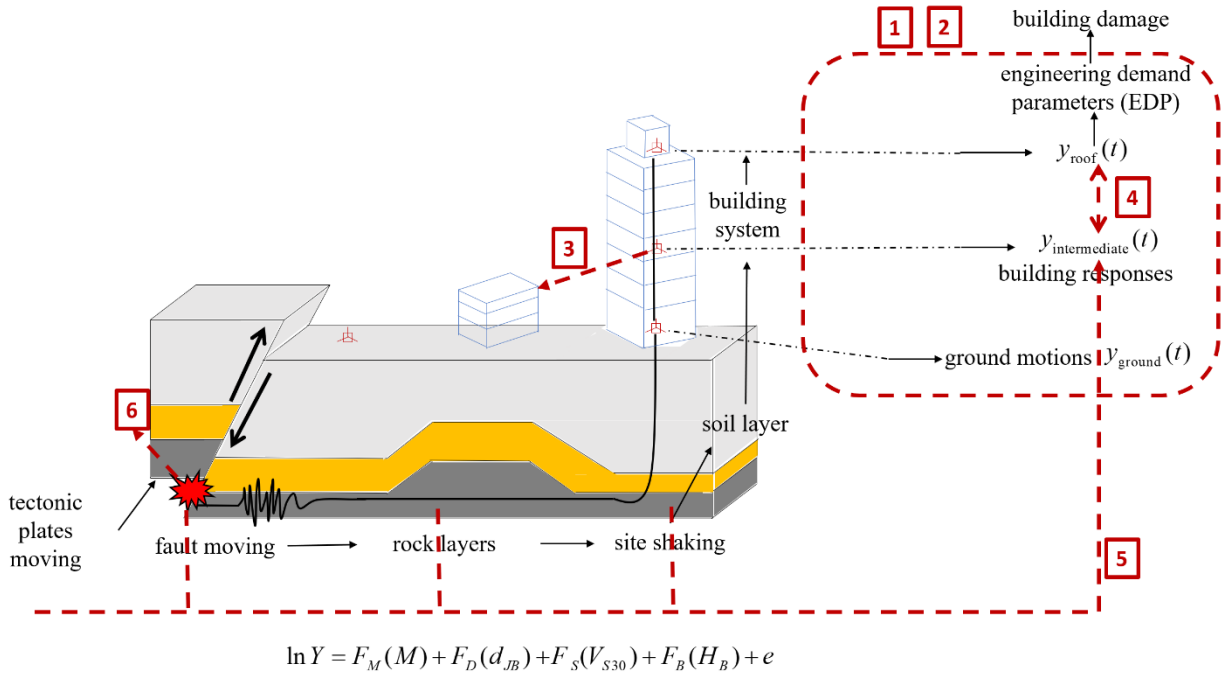


Figure 1.1 Proposed data-driven building seismic response prediction framework

1.3 Outline of the Thesis

Seven chapters are used to present the proposed framework for building seismic response prediction. The red boxes with numbering on Figure 1.1 are covered in the referenced chapter to achieve the goal.

Chapter 1 provides the background and motivation for the research, as well as an outline of the dissertation (red box 1). A review of key literature most closely related to the proposed framework is presented in Chapter 2. Chapter 2 (red box 2) includes details to assess the viability of applying machine learning to solve earthquake engineering challenges by examining the nature of the existing problems, addressing important characteristics of machine learning algorithms, and

demonstrating applications using archived data. Chapter 3 (red box 3) presents an interbuilding interpolation scheme using Kriging to predict peak seismic response demands for a portfolio of reinforced concrete moment frame buildings subjected to a seismic event. Three historical scenario earthquakes (1994 Northridge, 1999 Chi-Chi and 2000 Tottori) are used to evaluate model performance. Chapter 4 (red box 4) expands the interpolation scheme by reconstructing EDP profiles of four representative tall buildings along building heights by utilizing two kernel-based machine learning methods, kernel ridge regression, and kernel support vector regression, respectively. A rigorous model evaluation technique, non-replacement bootstrap, is used to demonstrate the different approaches. Chapter 5 (red box 5) expands the previous two seismic demand reconstruction methodologies from scenario-based (subjected to a single event) to a generalized prediction model that incorporates event characteristics of structural and site dissimilarity with a mixed-effect statistical model. This approach is validated using data collected for California buildings between 1984 and 2018. Chapter 6 (red box 6) includes a discussion of several alternatives for event terms considered in the generalized prediction model including Bayesian and Frequentist approaches. Chapter 7 provides a summary and the primary conclusions.

2. A Critical Examination of the Viability of using ML to Address Structural Engineering

2.1 Introduction

The term Machine Learning (ML) is no foreigner to the structural engineering research community as pioneer applications, such as applying neural network [17,18] and regression analysis [19,20], can be tracked to early 1990s. Since ML methods at that time were often only treated as a method to maps nonlinear patterns, as discussed in [21], ML applications merely serves as an alternative to physics-based models. For instance, Ghaboussi, et. al. [17] developed a neural network mapping for experimental biaxial stress-strain relationship of concrete; however, the physics-based models of concrete were well-developed and it was much easier to interpret the data using physics-based models versus the ML model [17]. As early as 1997, Reich [22] provided a comprehensive review associated with the application of ML to solve civil engineering problems, from data collection, model creation, results evaluation using advanced ML techniques such as cross-validation and bootstrap. Due to computational speed and memory space issues, early ML models were restricted to solving classes of problems with limited data such that the prediction performance were merely comparable with physics-based models. Furthermore, ML models are typically considered to be black-box solutions that often lack the framework to gain physical insight.

Starting from late 2000s, driven by the decade long boost of development of Structural Health Monitoring (SHM) and rapid growth of computational power and data storage, the structural engineering community has gradually engaged in "big data" problems. The challenge to interpret large data sets collected through recordings from field monitoring and to extract information from data retrieved through numerous or complex computer simulations has brought

ML back to the front stage. Recent success stories for ML in many other fields and the growing number of ML application publications in structural engineering that have been initiated in recent years, has drawn more attention and interest to critically examine opportunities to use ML to solve structural engineering problems. Two major items have spurred this development. First, the amount and diversity of data have increased significantly. With the improvement in computational methods and structural modeling tools, many large complex structures can be analyzed by finite element method for multiple hazard levels or types. This significant increase in the problem scale and complexity can be challenging for physics-based models. On the other hand, ML methodologies have vastly improved with the state of art methods such as Convolutional Neural Network (CNN) [23,24], and feature fusion techniques [25–27] and these methods have been applied to solve structural engineering problems. With the increased data size and complexity, data-driven methods may be advantageous relative to traditional physical-based models in terms of capability and efficiency. For instance, the ability to assess structural performance for a regional and city scale models [27,28], as opposed to focusing on individual structural components and systems as has been common until recently, for multi-level hazard assessment for a range of intensities, is a case where ML is very attractive [8].

The modern ML applications in structural engineering field can be categorized into the following four general areas:

1. to predict structural nonlinear responses and damage;

2. to interpret experimental data and formulate empirical relations between structural properties and responses;

3. to retrieve information about structural characteristics through visual media;

4. to recognize patterns from SHM.

The objective of this chapter is to present ML methods and assess their ability to address current strucutral engineering challenges by examining the nature of existing problems, characteristics of ML methods, and current research directions. Specifically, challenges associated with collecting data, preparing sophisticated model simulations, and enhancing computation tools are reviewed. An introduction of ML is presented with brief descriptions of their mathematical formulations and capabilities. Following which, domain specific motivations to use ML methods in the structural engineering field are demonstrated with discussions on their corresponding challenges across the above mentioned four main perspectives, each supplemented with examples from recent research publications. Finally, future directions and opportunities to apply ML methods, including data source, model interpretability, and extrapolating ability are presented.

2.2 Brief Introduction of Machine Learning

2.2.1 General Formulation of ML Setting

Systematic approaches to create data-driven models appropriate to solve engineering problems have already been developed, i.e., surrogate modeling [29]. In the ML context, surrogate models can be treated as a subset of ML models that have a complex objective function to minimize (i.e., non-convex function). Instead of directly solving the function, an alternative mechanism is used to find a solution. For instance, the logit function converts the non-convex 0/1 loss of a classification objective function into a logistic loss such that it can be solved through logistic regression. To avoid confusion, this chapter will focus on the classic formulation of ML using a universal language instead of discussing more broadly defined data-driven methods.

ML problems are typically categorized into two major tasks, supervised and unsupervised learning. The former is formulated using the ground truth, or the labeled data, i.e., actual damage classes of a pool of buildings, while the latter is not. Therefore, their applications are dependent

on the objective. Supervised learning can be further expanded into two sub-categories, (a) regression, where labeled data are continuous variables, and (b) classification, where labeled data are discrete class tags. Therefore, the goal of supervised learning is to generate a function that approximates the labeled data based on input observations. An unsupervised task, on the other hand, is to extract the inherent structure information from the given input observations, i.e., to search for clusters within the data.

For supervised learning, the dataset contains a collection of feature variables, a matrix X with dimension $n \times p$, where n is the total number of observations (data points) and p is the number of features (independent variables); and a labeled response variable y , a vector of size $n \times 1$ representing labels of each observation. Similarly, a dataset for unsupervised learning includes a feature matrix X but not the response variable y. The objective of supervised learning is to solve the generalized optimization problem by minimizing an empirical risk defined in Equation (2-1) [27].

$$\arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i;\theta)) + \lambda\Omega(\theta) \qquad (2\text{-}1)$$

Where y_i is the response variable for observation i . $f(x_i;\theta)$ is the approximated response from the ML model based on the feature x_i and θ represents the set of model parameters. ϕ is a loss measure between the true value y_i and the predicted value $f(x_i;\theta)$. $\lambda\Omega(\theta)$ is a regularization term that adjusts model complexity by restricting the parameter set θ through $\Omega(.)$. The objective is to find the best set of model parameters, $\hat{\theta}$, that minimizes the empirical risk over training data with the regularization penalty considered. Equation (2-1) is a convenient generalized form and can be adopted by many supervised learning methods including ordinary least squares, ridge regression, LASSO regression, logistic regression, and kernel regression. Depending on

which ML method is adopted, the minimization problem can be solved with either a closed form solution, through gradient based methods, or convex approximation.

The objective function for unsupervised learning is shown in Equation (2-2), where α is the set of model parameters that characterizes a learning structure for the given dataset. The loss function ϕ quantifies the costs to assign a data point x_i to a particular cluster. ML methods that can be generalized by Equation (2-2) include K-means and K Nearest Neighbors [30].

$$\arg\min_{\alpha} \frac{1}{n}\sum_{i=1}^{n}\phi(x_i;\alpha) \tag{2-2}$$

In statistical learning theory, the objective function expressed in Equations (2-1) and (2-2), is defined as an empirical risk over the training dataset, donated as $I_n(f)$ for a given model f. The theoretical objective of ML problem is to minimize the expectation in Equation (2-3), which is an integration over entire data space. Limited by the amount of data sampled from the space, the ideal case is often approximated by minimizing the empirical risk of training data instead.

$$\arg\min_{\alpha} I(f) = \int_{X\times Y}\phi(y, f(x))\,p(x,y)dxdy \tag{2-3}$$

2.2.2 Feature Engineering

Prior to training the model, applied ML methods always involve the process of selecting and extracting features which are found to influence model performance, improve training efficiency, and increase flexibility; all are necessary to tackle ML problems. A successful ML application usually deploys a standard algorithm; however, it also typically includes the ability to adjust features to achieve best possible prediction performance.

Feature selection can be categorized into three methods: filter, wrapper, and embedded. The filter method ranks original features according to an importance measure score such as scores from a Chi-square test or correlation coefficients between individual feature and response variable

and selects a subset of original features. The wrapper method involves recursively including or excluding features from current pool and selecting the best performing pool. Embedded methods are those ML algorithms which incorporate automatic feature selection, e.g., LASSO and ridge regression. Both filter and wrapper methods are good at avoiding overfitting issues by reducing model complexity and improving training efficiency by reducing duplicated features. Feature extraction consists of two major tasks that increase the effectiveness of ML models: 1) conducting dimension reduction over data through methods such as Principle Component Analysis (PCA), which performs linear mapping from the original data space to a lower dimensional space such that data variance over each orthogonal component is maximized; 2) transforming data into a higher dimensional space such that the patterns become sparse and separable, e.g., kernel-based ML algorithms [31].

Besides the general feature engineering techniques mentioned in the prior paragraph, specific feature designs have been proven to be very successful for domain-specific problems. For instance, the use of HAAR-like features achieved human-level accuracy with far less computational effort [32] for face recognition; SIFT [33] features are very effective for object detection within images, and the HOG [34] features are particularly good at human detection. However, these domain-specific feature engineering techniques require considerable trial and error testing and are designed to only work for very specific problems and data structures. Neural network approaches and the associated deep learning approaches are extremely popular in that they automate feature engineering to achieve state-of-the-art level performance in many pattern recognition and data mining domains. This approach has emerged due to the increase in computation power in recent years.

2.2.3 Model Training and Performance Metrics

There are many well-established procedures for ML model training that attempt to achieve stable and effective prediction performance for new data (extrapolation) given a training dataset. One common strategy is k-fold cross validation (also discussed in [22]) that randomly splits a dataset into k different subsets and trains the model k times using the k^{th} subset as testing data and the remaining k-1 subsets as training data. Among all k models, the best performing model over the testing dataset is selected. This procedure largely reduces overfitting on the training dataset. Another popular technique to avoid overfitting is Bootstrap, which randomly samples a subset of the data with replacement and trains the model M times. The final model is selected as an average over the predicted results (regression) or based on majority vote (classification) [35] from the M models. Both Bootstrap and k-fold cross validation effectively reduces model variances are unbiased and are the primary training techniques used in developing data-driven models presented in later chapters. These training procedures are evaluated by using various performance metrics for model selection. For example, performance metrics of binary classification models include accuracy, precision and recall [36] whereas performance metrics for multi-class classification models typically include using a confusion matrix and top-k class accuracy as shown in ImageNet [37] literature. Regression models are usually evaluated by Root Mean Squared Error, Median Absolute Error, and Median Absolute Relative Deviation as presented by Burton, Sun et. al. [27,28,38].

2.3 Motivation of Using Machine Learning in Structural Engineering Field

For a long time, the structural engineering community has been developing models for real-world structures by conducting physics-based simulations and constructing prototypes to investigate uncertainties using laboratory tests. The traditional research path involves interpreting

results from the models and/or tests and combining observations (empirical and/or analytical) to develop guidelines, standards, or codes of practice for structural design, construction and retrofit. The approaches involved range from statistical curve fitting to physics-based analytics using computer simulations. Classic examples include the steel moment connection and steel shea wall system development by Qian and Astaneh-Asl [39,40] and a structural wall and coupling beam development by Wallace [41]. This common research path usually includes considerable engineering judgement and commonly requires some level of consensus to become standard practice. Limitations of this approach are apparent, such as the lack sufficient model complexity to accurately represent the dataset which leads to empirical approximations and, sometimes, cumbersome rules and step functions where small changes in design variables lead to undesirable outcomes. In addition, these approaches often include relatively crude factors (e.g., load and capacity-reduction) to account for uncertainties in the confidence intervals based on incomplete information.

Due to the rapid improvement in computational resources (speed, memory, storage, visualization), current mechanics-based analyses, such as finite element analysis using OpenSees [42] and Perform3D [43], are capable of simulating complicated structure systems from relatively simple nonlinear behavior subjected to loadings from multiple hazard types (earthquakes, hurricanes) at various load intensities [44,45]; these analyses tend to generate a lot of structural response simulation data. A study by Burton et. al. [46], where building seismic performance was assessed for a set of archetypical structural models representing existing wood frame buildings in Los Angeles using four different retrofitting schemes with thousands of Nonlinear Response History Analyses (NL-RHA). In addition, evaluation of structural safety has expanded from collapse prevention of single buildings to more diverse and comprehensive studies that involve

assessment of a spectrum of damage states for a range of structure system types, which then enables multi-variable risk assessment studies using tools such as HAZUS [11,12,47]. Guan et. al. [48] performed intensive Incremental Dynamic Analysis (IDA) to assess the seismic performance of an innovative self-centering steel moment frame and to assess economic risk using the FEMA P-58 procedure [49]. In addition to evaluating individual structures, the risk assessment could also be applied across the regional geo-spatial dimension by analyzing structural responses at the portfolio scale level. Regional loss assessment is demonstrated by DeBock et. al. in [14] and Sun et. al. [28], where a set of concrete moment frame buildings are subjected to an actual earthquake scenario at a number of sites distributed over an urban region. It is shown that the current structural research in hazard mitigation is moving towards to more sophisticated hazard scenarios that generate extreme volumes of data, which are not suitable for some of the traditional approaches that are less capable or less efficient at extracting information from the multiple dimensional data space.

With the continuing growth of experimental data accumulated in the structural engineering field and the use of more advanced technologies to capture additional data and data types from experiments (i.e., digital image technique to capture stress distribution on concrete slender wall [50]), traditional data interpolation techniques are insufficient to explore and identify data patterns for a broad range of structural parameters among a large set of experiment events. A recent database developed by Abdullah and Wallace [51] collects detailed, organized, and parameterized information from more than 1000 reinforced concrete wall tests available in the literature, which is the largest of its kind. Traditional statistical tools are not ideal to examine data trends in such a large dataset. On the other hand, modern ML techniques have advantages, such as reduced influence of outliers, expanded dimensional parameter spaces (higher model complexity), and

improved predictions.

Another emerging trend is the development and expansion of SHM over the last several decades to the point where rapid assessment based on measured data can provide valuable decision variables. The concept of SHM is to remotely monitor the status (e.g., level of damage) within a structure using advanced sensing technologies based on data collected during an event. The process involves data cleaning, feature selection, and statistical model development for damage detection, all of which can be achieved using modern ML algorithms. As early as 2006, Farrar and Worden discussed some of the challenges facing by SHM, such as over-prediction of damage, which is an application of statistical pattern recognition. They noted that key challenges include structural damage is often localized whereas monitoring is often accomplished using global measurements, where it is difficult to identify what has influenced the minor changes in the system responses (lack of uniqueness, in space and time), especially given the relatively sparse data measurements that were common at the time. With more computation power and with state-of-the-art ML models, these challenges are being overcome. For example, the use of kernels to extract key features from both space and time domain and develop a discriminant model for damage detection by Santos et. al. [26], incorporation of state-of-the-art deep learning CNN models to make use of image data [52], the application of random forest, an advanced decision tree algorithm, for assessing post-earthquake structural safety of buildings [8].

The National Hazards Engineering Research Infrastructure (NHERI) is a platform that provides a network of research laboratories located at universities around the country that conducts experiments and collects experimental data in various forms related to hazards such as images of observed physical damage (photos, video) and measured data from a wide-range of sensor types from tests conducted in that involve water systems, energy and communication systems,

residential , commercial, and industrial buildings, other infrastructure (dams, bridges, roads, tunnels). This broad range of applications provides a tremendous opportunity for the structural engineering community to search for data-driven approaches to help address some most challenging problems in our field.

The preceding discussion provided a general overview of use of modern ML algorithms, and the opportunities to expand the use of structural response history data obtained from sensors, images to identify structural and non-structural damage, and of ML models. Four critical aspects associated with these challenges are discussed in the following sections.

2.4 Examples of Machine Learning Applications in structural engineering

A board range of relatively recent ML publications in structural engineering field are summarized based on their application into the four major categories: 1) predicting nonlinear structural responses and damage; 2) experimental data interpretation and empirical fitting; 3) information extraction from visual media; and 4) pattern recognition for structural health monitoring.

2.4.1 Predicting Structural Nonlinear Responses and Damage

The ability to use ML to predict using either regression or supervised learning provides a data-driven alternative to determine structural responses or classify damage states for structures under various scenarios. Structural demand data usually comes from a large set of NL-RHA, although it is becoming more common that such data might be retrieved from instrumented buildings. Depending on the objective, these NL-RHA were conducted using models ranging from 1) a single type of structure whose structural characteristics may vary; 2) a single or pool of structure models subjected to a defined hazard at different intensity levels; and 3) a cluster of structure models in a regional assessment scenario.

To detect structural damage, Zhang et. al. [8] applied random forest for damage detection for a 4-story reinforced concrete building case study using the response demand output from NL-RHA, whereas Bagriacik et. al. [53] applied logistic regression, boosting regression trees and random forest to identify individual structural pipe damage status. To predict structural seismic responses using regression algorithms for a single structure, example applications include [54], where Support Vector Regression (SVR) was used to predict seismic engineering demand parameters using NL-RHA data from a single degree of freedom system, a four-story building and a bridge pile, whereas Soleimani et. al. [55] applied LASSO regression to determine uncertain parameter significance seismic demand prediction for an irregular bridge. To estimate demand parameters that are critically related to structural damage, Burton et. al. proposed a framework for aftershock collapse vulnerability using mainshock intensity, structural responses, and physical damage indicators [38], whereas Mangalathu, et. al. estimated seismic vulnerability and fragility curves for skewed bridges [56].

For a regional setting, Sun et. al. [28] reconstructed peak structural response engineering demand parameters by utilizing structural and spatial dissimilarities within in a seismic event using kriging models, and further extended the framework using kernel-based ML regression algorithms, kernel ridge regression, and kernel SVR to reconstruct a profile of responses for engineering demand parameters [27]. Burton et. al. [10] presents a conceptual framework for modeling post-earthquake housing recovery with building-level damage limit states which could be scaled up to an event scenario at the region level.

2.4.2 Experimental Data Interpretation and Empirical Fitting

One advantage of adopting modern ML algorithms is that it is possible to extract data patterns from complex structural system behavior and identify structure damage without human

inspection, which is the most common approach used. Ghiasi et. al. [25] conducted an experiment on a four-story prototype structure and mimicked damage scenarios by removing braces at each floor. A SVM classifier with a combined parameter kernel function was applied to predict structural damage scenarios using acceleration sensor response history signals at each floor as input features.

Besides the complex tasks of damage detection mentioned above, which is not viable through traditional empirical data fitting in a laboratory testing, ML also shows superior prediction performance on traditional tasks. For example, Vu and Hoang [57] proposed a kernel-based SVM model to map factors that influenced the punching shear capacity of fiber-reinforced polymer (FRP) from a dataset of laboratory tests performed on FRP-reinforced concrete slabs and demonstrated that the approach outperformed traditional formula-based methods based on RMSE. Naeej et. al. [58] applied a simple ML algorithm, decision tree, to predict lateral confinement coefficients in reinforced concrete columns to show that the approach outperformed traditional empirical formula. These two examples, one that deploys relatively complex SVM algorithm, and one that uses a simple decision tree, both demonstrate that ML algorithms can provide more accurate information than traditional empirical formulas to fit patterns for large datasets.

2.4.3 Information Extraction from Visual Media

Traditionally, structural damage detection at the structural component level through visual media has been accomplished by converting image data into strain distribution data using Digital Image Correlation (DIC), a mechanism based on a technique that captures local deformations using cross-correlation between randomly painted dots on a concrete surface, such as [59], [60] and [61]. The limitation is clear: 1) DIC requires a randomly painted dot field over the surface of interest, which is, for the most part, only practicable in laboratory setup; 2) a consistent length of image

sequence is necessary for calculating cross-correlations between painted dots; and 3) the converted damage indication data, the local strain distributions, only accounts for a subset of the information available from an image.

Vision-based techniques can extract additional information from image data. For example, Yeum and Dyke [62] deployed HOG and HAAR feature-based object detection schemes with a sliding window search to identify cracks near bolts from structure images. The HOG [34] and HAAR feature [32] classifiers are mainly texture feature extractors and are generally not as successful at image feature extraction as the CNN technique, which is extremely efficient in extracting both marco- and micro-scale features from images. It is superior in terms of unrestrained image input, accessibility and fast inference speed compared to traditional image processing techniques. Due to the flexibility of CNN for supervised learning problems, it has been deployed quite extensively to solve vision-based automatic damage detection problems. Cha et. al. [24,63] applied a region-based deep learning scheme to detect different damage types and concrete cracks in images of structures. Kong and Li [64] applied CNN to detect crack opening in steel bridge components under repetitive loadings from video clips. Other applications for damage detection using images include [65–68].

2.4.4 Pattern Recognition for Structural Health Monitoring

Pattern recognition is an essential part of SHM in order to evaluate the tremendous amount data collected from various sources and to monitor structural condition to enable assessment in a timely manner. One major challenge noted in the literature is the ability to identify critical features that are associated with structure damage. Traditional ways include wavelet transformation and Fast Fourier Transform (FFT) which are limited because these methods are highly sensitivity to outliers. Data-driven approaches using ML are found to be more robust in dealing with complex

structure systems and various sources of input signals [69]. According to Worden and Manson [70], there are four levels of application for the data-driven approaches in SHM to: 1) detect if damage exists in the structure; 2) extract the probable location of the damage; 3) estimate the extent of the damage, and 4) infer information regarding the safety of structure. ML classification algorithms are suitable for level 1 e.g., [25,26,71], which adopt supervised ML algorithms for damage detection. While [72] applied an unsupervised, nonlinear principle component learning algorithm analysis together with auto-associative neural network to select features to predict stiffness changes from a synthetic bridge model (level 3). Another example of feature extraction using ML is by Rafiei and Adeli [73] where restricted Boltzmann machine was applied for damage related feature extraction followed with a classification neural network to detect damage in a 38-story reinforced concrete building.

ML algorithms may provide superior performance for these damage-related feature extraction examples compared to traditional physics-based methods given its capability in pattern recognition [73,74].

2.5 Discussion in Applying Machine Learning to Structural Engineering

Previous discussions cover a wide range of ML applications at four major fronts of strucutral engineering community:

1. To predict structural responses and damage for stochastic excitation, such as seismic and wind loading, or stochastic structural characteristics;

2. To interpret experimental data where the test setting and scenarios are complex such that physics-based models either perform poorly or are limited to a particular configuration or problem type (versatility);

3. To serve as feature extraction for structural visual media data such as images and videos;

4. To recognize patterns from SHM data to inform structural system status.

In this section, we intend to critically discuss three domain challenges that we are facd in utilizing ML in structural engineering, data source, model interpretability and extrapolation fidelity.

2.5.1 Data Source

One big contributor of the success of ML is the access to more data. Although the amount of data required to achieve ideal performance for ML models depends on the problem scenario and goal, it is essential to have sufficient data that the sampled group could represent the true distribution. Traditional data from structural engineering is often limited in its quantity and diversity, e.g., NL-RHA structural responses from a particular building model, cyclic behavior of a specific structural component. In these cases, it is not ideal to employ ML to capture the entire physical phenomenon behind finite element simulation or laboratory testing. On the other hand, there are ML applications that utilize more diverse structural response data across spatial domain such as [27], which, however; does not cover a diverse range of structural dissimilarities. The challenge is to collect a subset of data from the true data distribution that is representative in diversity and quantity such that ML algorithms are able to discover the underlying patterns of structural system within the domain objective. The shortage of data can be compensated for by incorporating domain knowledge within the ML model such that additional domain knowledge is introduced to reduce the complexity of the model space and consequently reduce necessary amount of data to fit the ML model. In addition, some statistical procedures may also aid through data augmentation, e.g., Monte Carlo simulation, to generate synthetic data. Future efforts are likely to examine both options, i.e., collecting more diverse real and synthetic data and incorporating domain knowledge for model design.

Another critical concern is data fidelity, which is a common issue for data-driven models.

However, the impact of outlier data might be more critical to certain ML algorithms, suc as logistic regression, due to their low robustness to noise. There are anomaly detection procedures for ML, such as using clustering techniques (DBSCAN [75], K-means [76]) and Z-score [77]; however, it also is necessary to include structural engineering domain-specific outlier detection procedures in ML applications. Ideally, the data-driven anomaly detection should be merged into physics-based outlier detection. In other words, the universal data filtering procedure of ML models should be modified with structural engineering domain knowledge, which has not been demonstrated in most recent research works.

2.5.2 Model Interpretability

One of the most significant challenges associated with a ML model is interpreting the physics meaning of the model parameters. A commonly held view is that a ML model is a black box, i.e., there are no physical bounds that can be derived from a data-driven model. One option to address this issue is to deploy ML evaluation procedures (e.g., k-fold cross validation) over traditional statistical learning models to enable some degree of interpretability [28]. In addition, some recent efforts have demonstrated the potential of introducing domain knowledge into ML algorithms by incorporating a physics-based loss function, e.g., to embed hard conditions with a langrage multiplier into the loss function [78,79]. This approach provides a means to explain some of the ML model by adding a physics-based law into the objective function (Equation (2-1) and (2-2)). In [80], a spectrum of approaches are discussed that leverage the wealth of domain knowledge to improve performance data-driven models by adding domain knowledge, such as including theory guided model design or learning and regularization. However, combining ML and physics-based models remains a challenging problem that will be explored with future research.

2.5.3 Model Extrapolation Ability

The ability to extrapolate the results of a ML model is another challenge and is also referred to as overfitting. Standard ML procedures involve data-based extrapolation by using training/testing split, k-fold cross validation, bagging and bootstrap, as well as other means. In addition, random forest provides a robust method to avoid overfitting given its stochastic procedure in generating trees in both feature and data space. For structural engineering, it is important to identity and apply domain knowledge to help avoid overfitting issues. The combination of a data-driven procedure and domain knowledge, similar to dealing with data sparsity, may provide a powerful combination. Although extrapolation has been extensively studied in the ML field, it could be more critical in applications of the strucutral engineering field given the high sensitivity of some strucutral responses. For example, to simulate a regular 40-story tall building, thousands of parameters are needed for configuration such as strucutral component geometry dimensions, material properties, external loadings, and construction conditions such that complex ML is almost unavoidable. Consequently, complex ML models require large amounts of data, better noise filtering processes, and careful model tuning to reduce the effects of overfitting.

2.6 Summary

In this chapter, application of ML within civil engineering since the early 1990s, when the approach was introduced, are reviewed. ML applications in structural engineering field were initially motivated as an alternative to physics-based modeling to map nonlinear experimental data, later deployed for damage detection as a data-driven pattern recognition tool in SHM, and recently expanded to multiple fronts due to the large volume of data now available in structural engineering problems and the rapid growth of computational power, which consistently affect the viability to apply ML to solve strucutral engineering challenges.

As the strucutral engineering community has considered various approaches to developing ML models, ML models were discussed earlier this chapter using a universal setting related to model formulation, feature engineering, model training, and performance-driven tuning. This brief discussion summarizes current knowledge related to ML and identifies the complimentary steps required to develop a ML model. In addition, the various ML algorithms are grouped under the empirical risk objective from learning theory to regularize them into optimization formulations. This discussion is provided to help establish a consistent scenario setting or utilization of ML.

Subsequently, the motivations behind the interest in applying ML to strucutral engineering research and example applications from the literature were organized into four major categories: 1) strucutral response regression and damage classification; 2) experimental data fitting; 3) information extraction from visual data, and 4) pattern recognition in monitoring data. The literature provides various examples on the application of ML and demonstrates its potential; however, these examples share a common impression that the adoption of ML in structural engineering is still at early stage and the lacks approaches for comprehensive performance evaluation and universal model development procedure. Therefore, with regard to the above mentioned four categories, a detailed discussion is included on potential future directions for ML in three critical areas, namely: data source, model interpretability, and extrapolation ability. Data source is the primary concern when developing and applying data-driven models. The degree of to which the data represent the true distribution is directly related to resulting model capabilities, which can be improved by increasing the diversity and fidelity of the training data. Another critical concern is associated with the ability to interpret of ML models, which is, in-part, a challenge within our community due to lack of knowledge of ongoing developments within the data-driven field. Some recent research has indicated that progress is being made in this area and that ML

offers opportunities to develop models with domain-localization in structural engineering. The last topic is related to the ability to extrapolate the results of a ML model and problems associated with overfitting. Various of ML techniques are proven to be very effective against overfitting; however, it remains as a challenge to adopt these techniques for strucutral engineering problems. It should be noted that these three challenges are not independent. For example, the overfitting issue could be addressed by increasing data diversity and is therefore connected to data source. Making use of domain knowledge could potentially help with the ability to interpret model results and also reduce the amount of data needed.

To summarize, there are a vast number of opportunities within strucutral engineering research to address current challenges with data-driven ML models. However, challenges to develop such models exist such as data source, model interpretability and extrapolation. Creating ML models specific for strucutral engineering may be one future direction.

3. Inter-Building Interpolation of Peak Seismic Response using Spatially Correlated Demand Parameters

3.1 Introduction

During the period following a major earthquake, building owners, users and local and state officials are tasked with deciding whether damaged buildings can be re-occupied and reused. This decision can be informed by different types of post-event condition assessments. Seismic instrumentation and monitoring are one tool that is used to provide rapid damage detection and inform system-level evaluations. Accelerometers are placed in carefully chosen locations in plan and along the height of the building. Translational floor accelerations recorded during earthquakes are then used to compute velocity, displacement and story drift demands [81]. Using these recorded and computed response demands, a rapid diagnosis can be performed using one of several alternative damage measurement techniques. Response measurements from instrumented buildings can also be used to conduct real-time assessment of earthquake-induced damage, economic losses and social impacts [82].

Numerous studies have developed tools and techniques for interpolating structural response demands measured through seismic instrumentation, along the height of a building, i.e. intra-building response interpolation [83,84]. However, no research is available that has explored the possibility of inferring structural demands in non-instrumented buildings using response measurements from instrumented buildings, i.e. inter-building interpolation. With the decreasing costs of remote sensing technologies [85,86], seismic instrumentation of new buildings is becoming increasingly popular [87,88]. The city of Los Angeles, which has the only mandatory seismic instrumentation program in a US city, requires accelerometers be installed at the base, mid-level, and roof to obtain a permit for all buildings over ten stories as well as for buildings over

6 stories with an aggregate floor area exceeding 60,000 ft^2 [89]. There also is a more extensive instrumentation program required for buildings that utilized alternative design approaches. The city also has an extensive program for monitoring the equipment currently installed in approximately 400 buildings. For buildings at where seismic instrumentation is not mandated (e.g. buildings located outside the City of Los Angeles or buildings within Los Angeles not meeting the criteria), some owners choose to voluntarily install sensors [87,90,91]. As such, while still small, the fraction of buildings with instrumentation within large U.S. west coast city-centers such as Los Angeles, San Francisco and San Diego, continues to increase. This research seeks to leverage the expansion of seismic instrumentation by developing a spatial cross-correlation-based model for inter-building interpolation of peak seismic response demands.

The previous studies that are relevant to the work presented in this paper include those focused on characterizing the spatial correlation of ground motion intensity measures (IMs) and engineering demand parameters (EDPs). Spatial correlation of ground motion intensities is needed to predict their joint distribution at the sites of a portfolio of buildings or infrastructure systems. Wang and Takada [92] developed a macro-spatial correlation model of IM residuals using empirical data from historical Japanese and Taiwanese earthquakes. Regression analyses were used to formulate a one-parameter exponential decay function that relates site separation distance to the auto-covariance of peak ground velocity. Goda and Hong [93] used records from California and the Chi-Chi, Taiwan earthquake to study the spatial correlation of peak ground accelerations and pseudo-spectral acceleration responses. Both inter- and intra-event variability were considered and empirical equations were developed to predict correlations considering site separation distance and period of vibration of single-degree-of-freedom systems. In two follow-up studies, Goda and Atkinson [94,95] developed similar models using strong motion data from the K-NET, KiK-net

and SK-net databases.

Baker and Jayaram [96] used geostatistical tools to develop spatial correlation models of spectral acceleration residuals obtained from Next Generation Attenuation ground motion prediction models. The correlation structure of spectral acceleration residuals is represented using semi-variograms, which is a measure of the dissimilarity between two spatially random variables. Predictive equations were developed to compute correlation of spectral acceleration residuals based on site separation distance and the period of interest. This work was extended by Loth and Baker [97] to develop spatial cross-correlation models of spectral acceleration residuals corresponding to different periods.

The research work that is most relevant to the current study was conducted by DeBock et al. [14,98], who investigated the spatial correlation in building seismic response demands. EDP datasets for six reinforced concrete moment frame buildings were generated from Nonlinear Response History Analyses (NRHA) using ground motions from two historical (1994 Northridge and 1999 Chi-Chi Taiwan) and two hypothetical (Puente Hills Fault and ShakeOut rupture) scenario earthquakes. The spatial correlation (same building at different sites) and cross-correlation (different building at different sites) structure of local (component deformations) and global (drift demands and floor accelerations) EDPs were examined. Correlation patterns were related to the characteristics of the building and the scenario event that generated the ground motion data. The relationship between correlation of EDPs and spectral accelerations was also investigated. Based on the findings with regards to this relationship, an approach to account for correlation of structural response in building portfolio loss assessment was presented.

This paper introduces a methodology for using spatial and structural correlation to interpolate seismic response demands across buildings. As shown in Figure 3.1, NRHAs are used

to obtain response demands for a building portfolio subjected ground motions recorded from a historical earthquake. Correlation patterns are established for the recorded peak EDPs based on spatial and structural dissimilarity. Semi-variograms are used to model the spatial and structural similarity dependence of the peak EDP-correlations. The Kriging method is then used to construct the inter-building seismic response interpolation model. Various aspects of the model performance are evaluated by randomly and repeatedly dividing the site locations for a single scenario into training and testing datasets through non-replacement Bootstrap sampling. The training data represents the locations of known response demands and is used to construct the interpolation model, which is then evaluated against the response demands in the testing data. The inter-building interpolation model can be used to rapidly assess scenario earthquake impacts for an urban setting where only a subset of the buildings is instrumented.

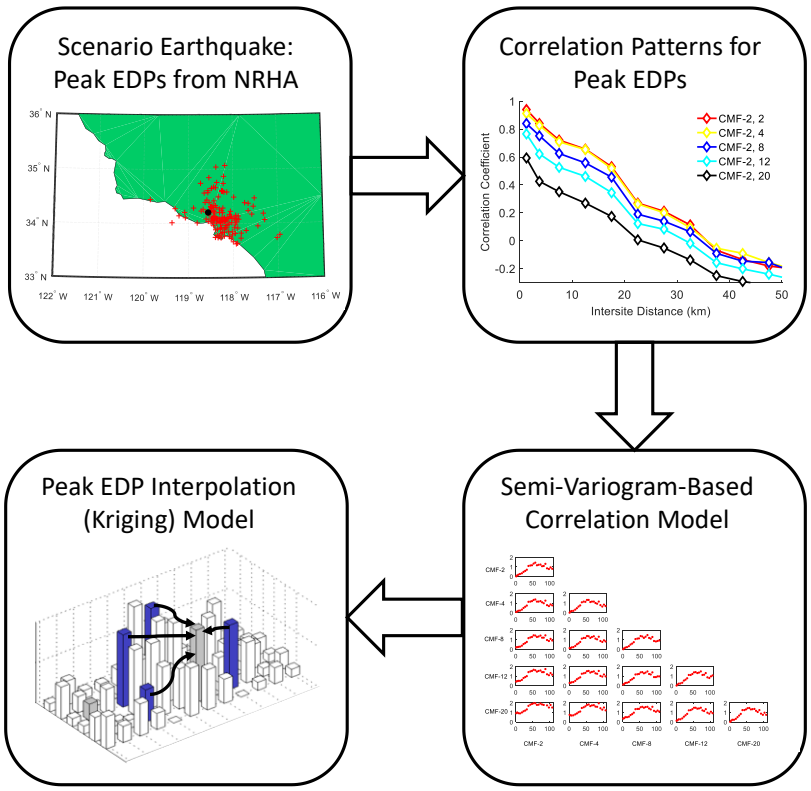


Figure 3.1 Conceptual illustration of methodology used to develop inter-building interpolation model of peak structural responses

3.2 Measuring Structural Response Correlation for A Portfolio of Buildings Subjected A Scenario Earthquake

3.2.1 Scenario Earthquakes and Ground Motions

To demonstrate the framework, scenario based EDPs are generated from NRHA. Three historical events are used as the scenario earthquakes (Table 3.1), with the Mw 6.7 1994 Northridge earthquake used as the primary event to demonstrate the model approach. The Mw 7.6 1999 Chi-Chi and the Mw 6.6 2000 Tottori earthquakes are used to evaluate the effect event characteristics on the model performance. These three events were chosen because of the availability of large numbers of recorded ground motions and their use in previous studies on spatial correlation of ground motion intensity measures and structural response demands [14,93,96–98]. Orthogonal pairs of recorded ground motion histories were obtained from the PEER NGA-West2 database [99] at each site for the three events.

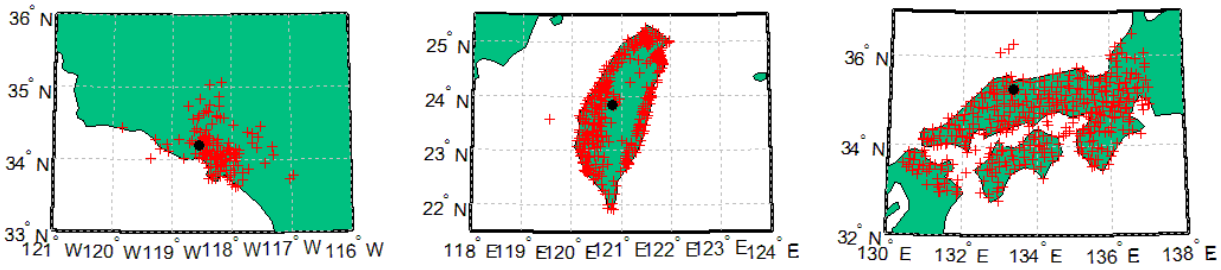
Maps of the region and the locations of the ground motion recording stations (red crosses) and the epicenters (black dot) for the three events are shown in Figure 3.2. It is worth noting that the spatial coverage of the recording stations is different for the three events, which may affect the resolution of each earthquake scenario and model performance. This issue is revisited later in the paper when the spatial correlation patterns and performance of the interpolation models across events are evaluated.

The relevant details of the three scenario earthquakes is summarized in Table 3.1 including the event magnitude, number of ground motion pairs, rupture distance and V_{S30} range. Of the three events, Northridge recorded the highest level of shaking with a maximum geometric mean peak ground acceleration (PGA) of 2.01g compared to 1.15g and 1.13g for the Chi-Chi and Tottori earthquakes, respectively. For all three events, the upper bound of the 90% confidence interval for

each earthquake scenario is less than half of the maximum PGA. This indicates that the PGA levels are highest near the epicenter and PGA decays rapidly in all directions. This is further illustrated in Figure 3.3, which shows exponential functional form trend lines that capture the relationship between the PGA and the rupture distance of the recording stations. The rate of decay in the PGA with rupture distance is highest for the Northridge earthquake. Chi-Chi, which is the event with the largest magnitude, has the slowest rate of decay in PGA as the rupture distance increases. As described in Section 3.2.2, a similar spatial pattern is observed in the building responses. These observations are also relevant to the performance of the Kriging model, which is introduced in Section 3.4.

Table 3.1 Summary of historical earthquakes used to develop interpolation model

Earthquake	Mw	No. of ground motion pairs	V_{S30} range (m/s)	Rupture distance range (km)	Geometric mean PGA range (g)
Northridge	6.7	152	161 to 2016	5.2 to 147.6	0.047 to 2.01
Chi-Chi	7.6	400	124 to 1526	0.3 to 172.2	0.008 to 1.15
Tottori	6.6	414	111 to 2100	1.0 to 333.2	0.001 to 1.13



(a) (b) (c)

Figure 3.2 Locations of ground motion recording stations (red crosses) and epicenter (black dot) for the (a) 1994, Northridge; (b) 1999, Chi-Chi; (3) 2000, Tottori earthquake

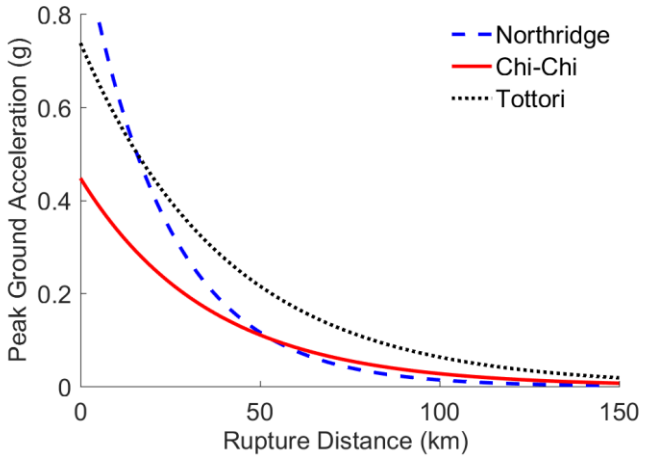


Figure 3.3 Exponential trend line showing the relationship between the geometric mean PGA and rupture distance

3.2.2 Structural Models, Response Simulations and Engineering Demand Parameters

Five modern, code-conforming RC moment-resisting frame buildings are used for this study. The buildings and structural models, which are adopted from those developed by Haselton [100], include 2-, 4-, 8-, 12- and 20-story RC moment frames designed based on the provisions of ASCE 7-05 [101] and ACI 318-02 [102] for a site in Los Angeles. The buildings were chosen to incorporate a broad period-range, such that the effect of structural dissimilarity is considered in the response interpolation model. Two-dimensional numerical models of the buildings developed by Haselton et al. [103] in OpenSees [42] are utilized for the NRHA. Each model consists of three bays of moment-resisting RC frames. The destabilizing effect of loads on the gravity frames are included by using a P-Δ column. The beams and columns are modeled as elastic elements with nonlinear flexural hinges with a trilinear backbone curve and hysteretic rules developed by Ibarra et al. [104]. More details on the design and structural modeling can be found in Haselton et al. [103,105]. Table 3.2 summarizes the building ID, number of stories and first mode period of each structure.

Table 3.2 Design information for concrete moment frame buildings

Building ID[a]	Building ID[b]	No. of stories	1st mode period (sec)
2064	CMF-2	2	0.66
1003	CMF-4	4	1.12
1011	CMF-8	8	1.71
1013	CMF-12	12	2.01
1020	CMF-20	20	2.63

[a]From Haselton [100,103,105]
[b]For use in this paper

NRHAs are conducted for the five building cases using the unscaled ground motions from each of the three scenario earthquakes. At any given site, each structural model is analyzed using a pair of ground motions; one for each of the two horizontal orthogonal directions. Since the layout and design of the moment frames are symmetric and identical along the two principal axes of the building, this approach captures the directional effects of the ground motion on the structural response correlations. The EDPs of interest include the peak story drift ratio (PSDR) and floor accelerations (PFA) along the building height, both of which have been shown to be strongly correlated with structural damage and direct economic losses [106]. Figure 3.4 shows exponential trend lines of PSDR (Figure 3.4a) and PFA (Figure 3.4b) versus rupture distance for the CMF-2 building, which provide insight into the spatial patterning the response demands. Overall, PSDR decays at a faster rate with rupture distance when compared to PFA and PGA, with the Northridge event producing the highest rate of decay. PFA decays at a rate that is even slower than PGA with Chi-Chi again having a slower rate of decay than both Northridge and Tottori.

Table 3.3 summarizes EDP statistics for all five buildings subjected to the Northridge ground motions. The mean PFA is approximately 0.30 for all five structures. The mean PSDR

generally decreases with building height, ranging from 1.06% in CMF-2 to 0.37% in CMF-20. The relatively small differences (less than 20%) in the coefficient of variation (CV) across the five structures suggests similar spatial patterning of PSDRs. In contrast, the spatial variation of PFAs is much more significant. The lowest CV, 0.40, is observed in CMF-2 and increases to 0.82 in CMF-20. As described further in Section 3.4, the spatial and structural patterning of peak response demands is one of the factors that affects the overall performance of the interpolation model.

Table 3.3 EDP statistics of all five buildings for original intensity scenario, Northridge earthquake

Building ID	PSDR[a] (%)		PFA[b] (g)	
	Mean	COV[c]	Mean	COV[c]
CMF-2	1.06	1.7	0.32	0.40
CMF-4	0.88	1.6	0.31	0.67
CMF-8	0.68	1.7	0.27	0.70
CMF-12	0.51	1.5	0.29	0.81
CMF-20	0.37	1.4	0.31	0.82

[a] Peak Story Drift Ratio
[b] Peak Floor Acceleration
[c] Coefficient of Variation

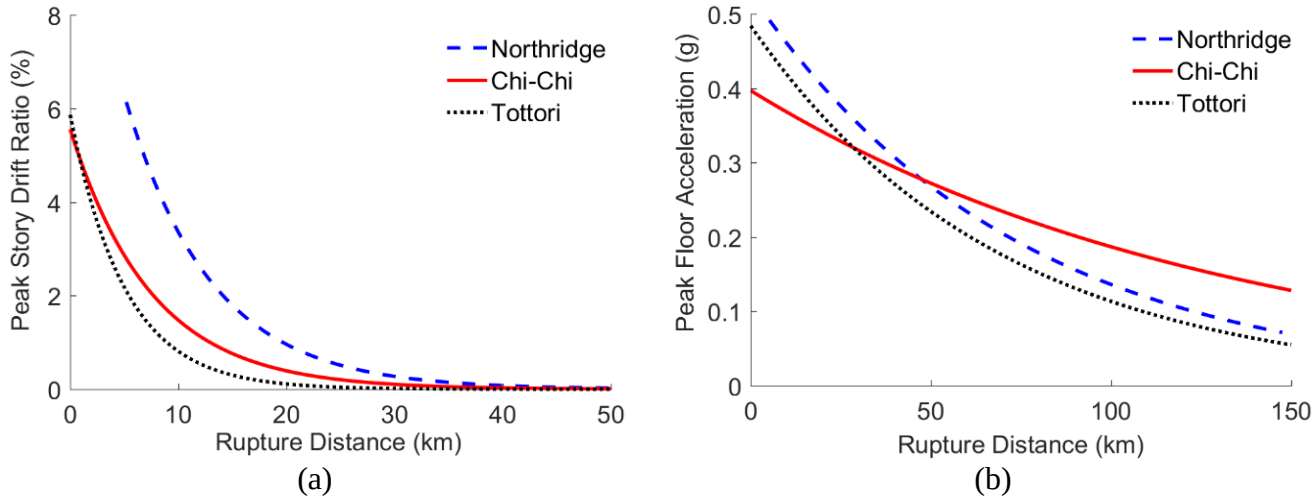


Figure 3.4 Exponential trend lines showing the relationship between the (a) PSDR and (b) PFA for CMF-2

3.3 Spatial Correlation of Peak EDPs from Structural Response Simulations

The proposed interpolation model relies on measurements of the spatial correlation of PSDRs and PFAs obtained from NRHA, which is performed on structures located at different sites, using ground motions from a single earthquake scenario. Also of interest is how ground motion directionality and site and structural dissimilarity affect this correlation. For each earthquake scenario, site pairs and their corresponding inter-site distances are established. The site pairs are then placed in bins corresponding to their inter-site distances. Each bin is defined by an upper, lower and median inter-site distance, the latter of which is taken as the median between the upper and lower bounds. An inter-site distance range (upper minus lower inter-site distance) of 2.5 km is used for first two bins and 5 km is used for the remaining bins to obtain better resolution for closer site pairs. The site pairs within a given bin are used to define two vectors X and Y, which represent natural log value of the maximum EDP recorded at each site. The correlation between maximum EDPs conditioned on the median inter-site distance, *D*, is computed using the following equation [98].

$$\rho_{X,Y}(D) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{3-1}$$

Where μ_X and μ_Y are the mean values and σ_X and σ_Y are the standard deviations of the vectors X and Y respectively. Several types of correlations are examined using the EDP dataset. Self-correlation is based on site pairs that have the same structure and cross-correlation is based on site pairs with different structures [98]. The difference between the self- and cross-correlation is a measure of the effect of structural dissimilarity on the correlation pattern. The effect of ground motion directionality is considered by comparing the correlation pattern of the maximum response demands corresponding to the two orthogonal directions to the pattern obtained by considering each individual direction separately.

The scatter plots in Figure 3.5 show how the median inter-site distance affects the correlation between the natural log of the PSDR for different site pairs. The plots are obtained from the NRHA results for CMF-2 subjected to the ground motions from the Northridge earthquake. Ground motions that cause structural collapse, which is defined by the occurrence of dynamic instability [11] are excluded from the results. Figure 3.5a through 5d correspond to median inter-site distances of 1.25 km, 3.75 km, 7.5 km and 12.5 km respectively. The plots show that the scatter around the line representing perfect correlation increases with the median inter-site distance.

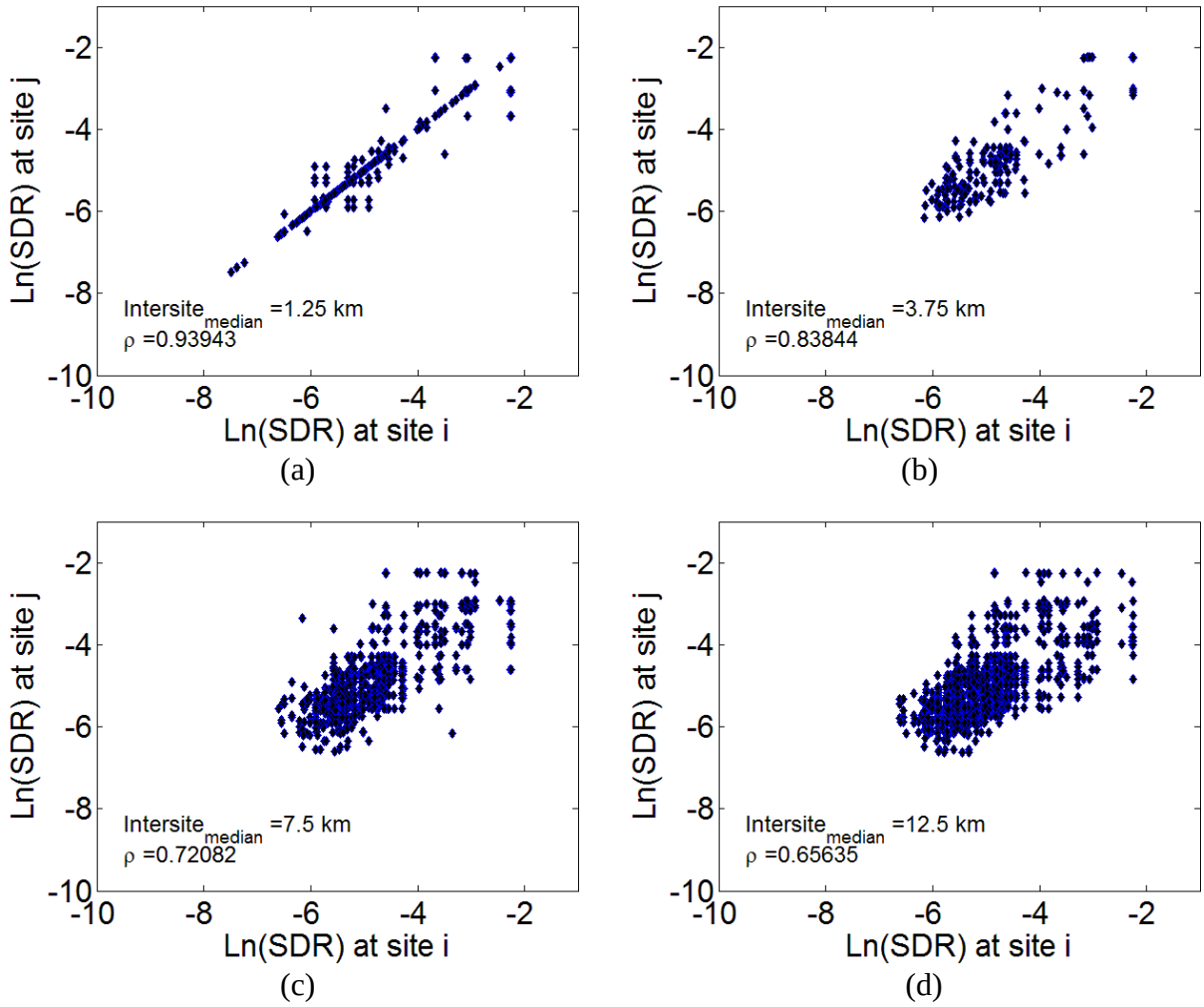


Figure 3.5 Scatter plot showing the natural log of PSDRs for various site pairs with median inter-site distances of (a) 1.25 km, (b) 3.75 km, (c) 7.5 km and (d) 12.5 km for the CMF-2 structure subjected to the Northridge ground motions

A more concise illustration of the relationship between the median inter-site distance and the maximum EDP correlations is presented in Figure 3.6. The effect of inter-site distance on PSDR correlations (ρ_{PSDR}) is shown in Figure 3.6a, which is also corresponds to the CMF-2 structure subjected to the Northridge ground motions. Included are the self-correlation relationships obtained from considering the maximum response for the two orthogonal directions as well as each individual direction. It can be observed that the inter-site distance has a strong

influence on the correlation coefficient, which ranges from almost 1.0 to 0.2 for median inter-site distances of 1.25 km and 22.5 km respectively. Moreover, this influence is unaffected by the directionality of the ground motions. Figure 3.6b shows that $\hat{D}$ also has a stronger effect on the site pair correlation coefficients for PFAs (ρ_{PFA}) compared to that of PSDR. Figure 3.7 shows how the median inter-site distance affects the self-correlation between the natural log of PSDR and PFA considering all possible site pairs. For each of the five buildings, a plot is obtained from the NRHA results after subjecting the structure to ground motions from the Northridge earthquake. Ground motions that cause structural collapse, which is defined by the occurrence of dynamic instability [11] are excluded from the results. The effect of inter-site distance on PSDR and PFA correlations is shown in Figure 3.7a and b, respectively. The correlations are computed by taking the maximum response corresponding to the two orthogonal ground motions for a given site. However, it should be noted that the correlation pattern was very similar when the peak response for each direction was considered separately as observed in Figure 3.6, highlighting the fact that ground motion directionality did not have a significant effect on the correlation patterns.

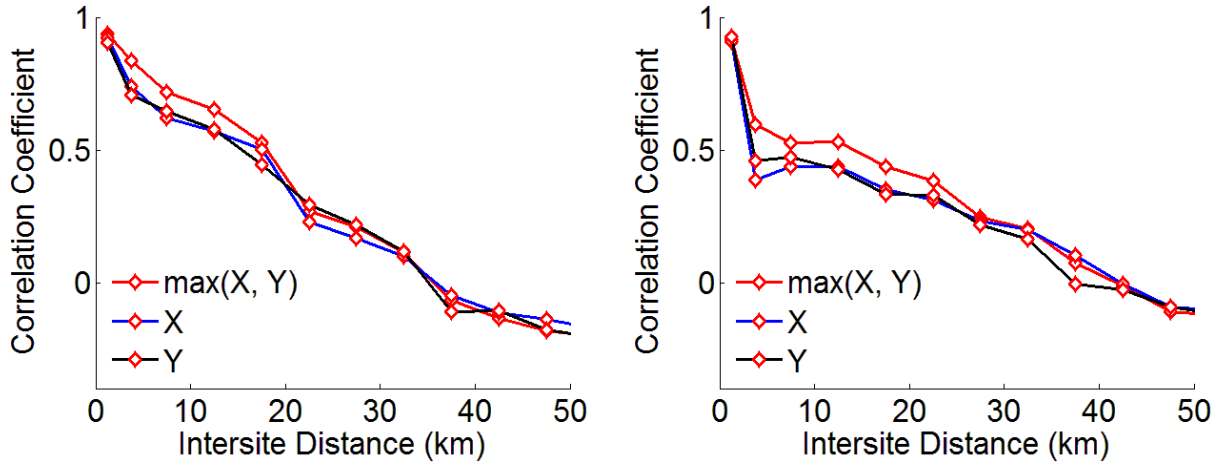


Figure 3.6 Correlation coefficients for maximum (a) PSDRs and (b) PFAs versus the median inter-site distance for the CMF-2 structure subjected to the Northridge ground motions

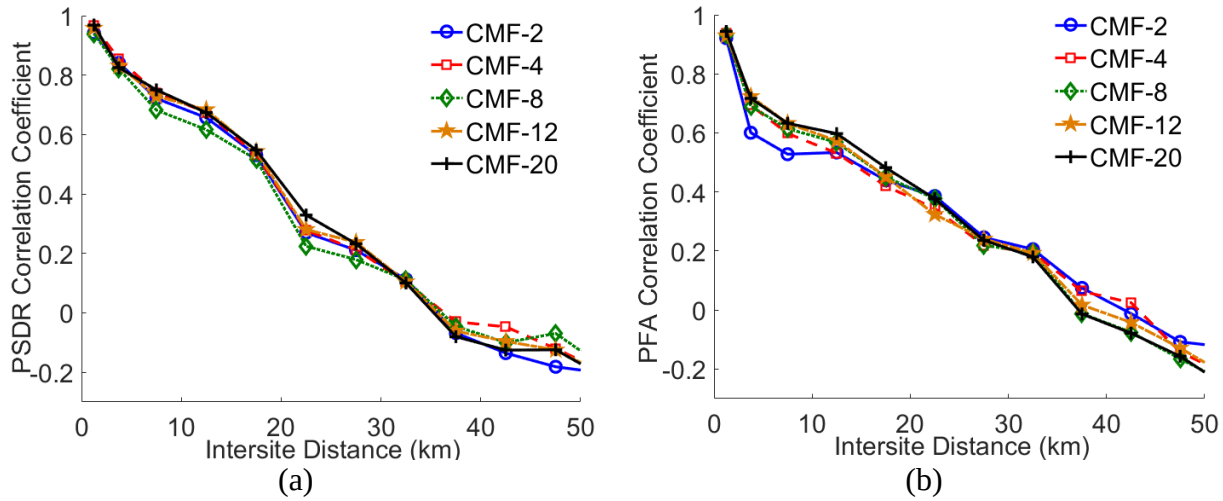


Figure 3.7 Self-correlation coefficients for (a) PSDR and (b) PFA versus the median inter-site distance of all structures subjected to the Northridge ground motions

Figure 3.8a and b compare the PSDR and PFA self-correlation for the CMF-2 and -20 structures respectively. The PSDR self-correlation is generally higher than PFA for lower inter-site distances (0km to 20km); however, the trend is reversed for site-pairs with larger inter-site distances (20km to 50 km). Similar trends were observed for the Chi-Chi and Tottori scenario events with slight differences in the inter-site distance at which the trend reverses. Low inter-site distances are dominated by site pairs near the epicenter, where low frequency ground motions, which have a greater effect of PSDR compared to PFA, are prevalent. This explains the higher level of correlations observed for PSDR at low inter-site distance pairs. Higher inter-site distances generally correspond to site pairs with one location near the epicenter and the other much further away. The lower PSDR correlations can be explained by the faster rate of decay observed in Figure 3.4 compared to PFA.

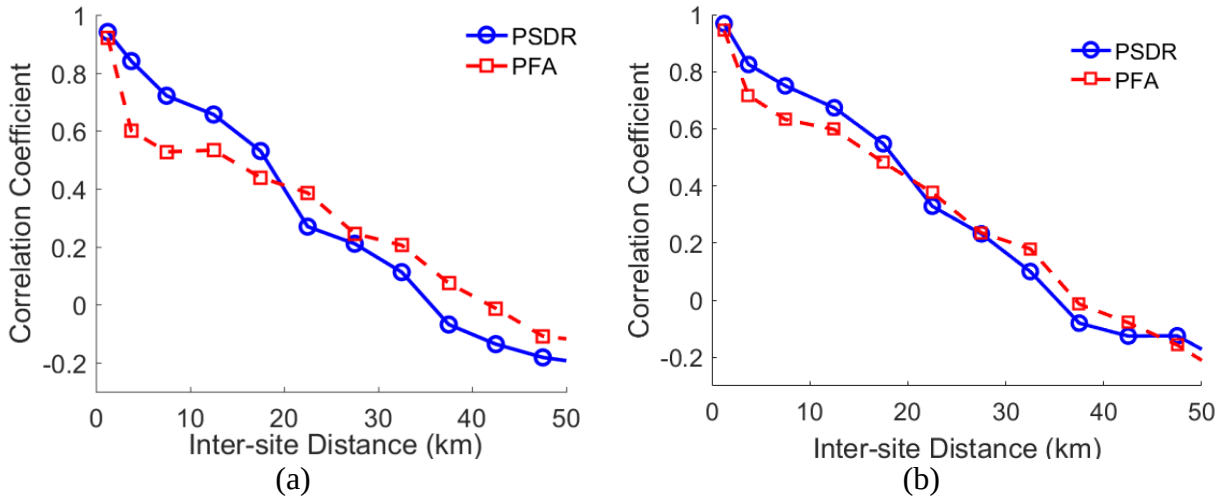


Figure 3.8 PSDR and PFA self-correlation coefficients versus the median inter-site distance for the (a) CMF-2 and (b) CMF-20 structures subjected to the Northridge ground motions

The correlation patterns shown in Figure 3.6, Figure 3.7 and Figure 3.8 were generated using site-pairs with the same structure at the two locations. However, a more realistic representation of the effect of inter-site distance on EDP correlations is obtained by using different structures at each site for a given pair (cross-correlations). Plots of the PSDR correlation coefficients versus the median inter-site distance for site-pairs with the CMF-2 and CMF-20 structures at each location are shown in Figure 3.9. The comparison between the self- and cross-correlations serves as a measure of the effect of structural dissimilarity. The plots in Figure 3.9 are obtained from NRHA in which both structures are subjected to all the Northridge ground motions. For each site pair, two sets of spatial cross-correlations can be computed. For instance, if denotes the peak EDP recorded in structure $B1$ located at site a, Y_{B1} in structure $B1$ located at site b and vice versa, one set of correlation coefficients is obtained from pairing the maximum EDP for X_{B1} and Y_{B2}, $(\rho_{X_{B1},Y_{B2}})$, and another from pairing X_{B2} and Y_{B1}, $(\rho_{X_{B2},Y_{B1}})$. This allows the evaluation of whether switching the location of structures in a site pair affects the spatial cross-correlation. Figure 3.9a and b show that the effect of switching the building locations on the spatial cross-

correlations is negligible for PSDR and slight for PFA. However, Figure 3.9c and d (site-pairs with CMF-2 and CMF-20) show that, as structural dissimilarity increases, the effect of switching on spatial cross-correlations, especially for PSDR, becomes much more significant.

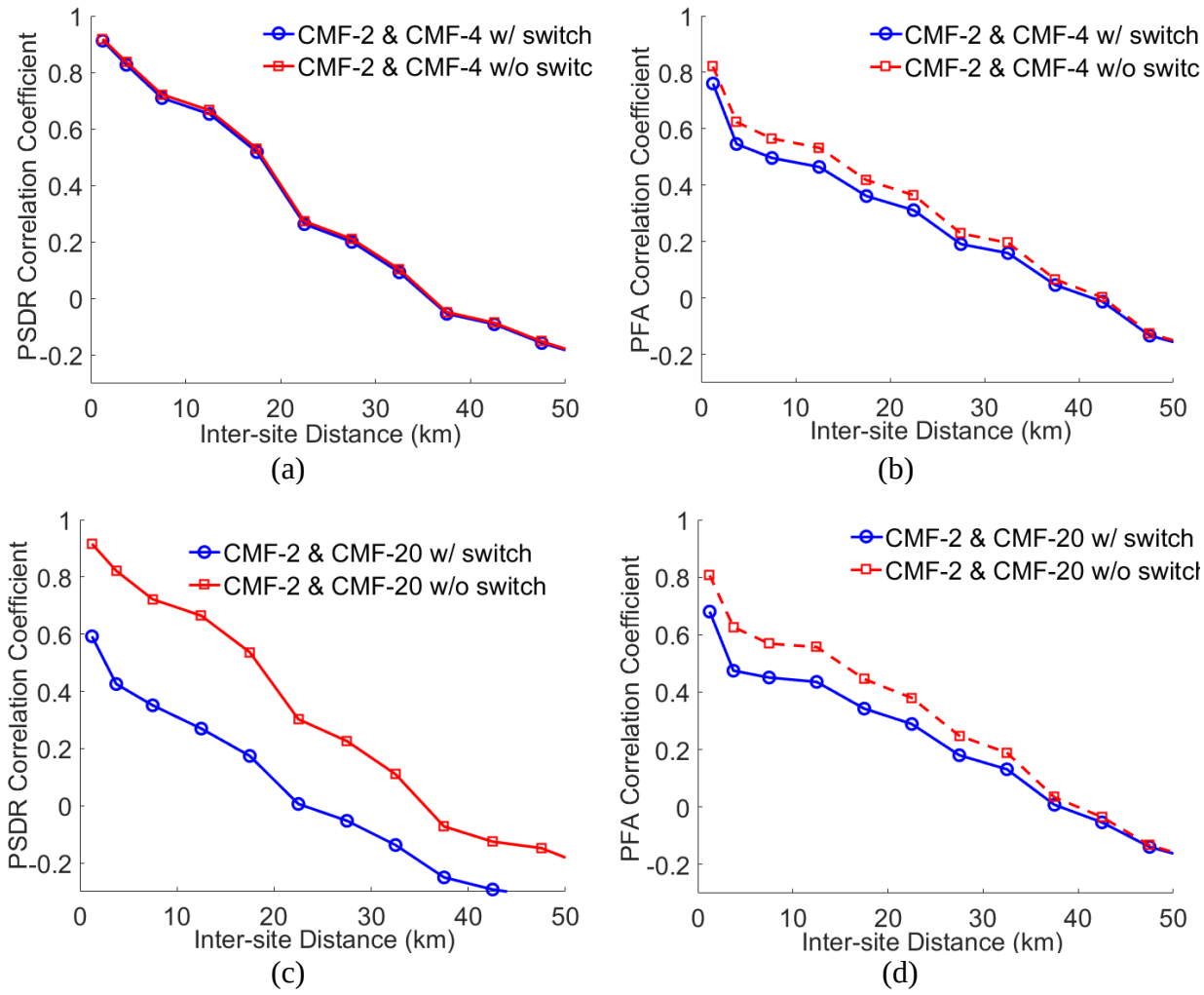


Figure 3.9 (a) PSDR and (b) PFA cross-correlation patterns for site-pairs with the CMF-2 and CMF-4 structures and (c) PSDR (d) PFA for site-pairs with the CMF-2 and CMF-20 structures all subjected to the Northridge ground motions

Figure 3.10 shows plots of ρ_{PSDR} (Figure 3.10a) and ρ_{PFA} (Figure 3.10b) versus $\hat{D}$ between CMF-2 and all other structures for the Northridge earthquake. Here, the difference in height or fundamental period of two structures is seen to have a significant effect on PSDR cross-

correlation trends. Figure 3.10a shows that the CMF-2 self-correlation and the cross correlation between CMF-2 and CMF-4 are almost identical. However, for a median inter-site distance of 1.25km, the CMF-2 self-correlation is about 42% higher than the cross correlation between CMF-2 and CMF-20. Figure 3.10b shows that, again, the effect of structural dissimilarity on PFA cross-correlations is less pronounced than PSDR but still noticeable. The smaller effect of structural dissimilarity on spatial correlations for PFA compared to PSDR is consistent with Table 3.3, which shows that the mean PFA is generally the same across all structures while the mean PSDR varies by as much as a factor of more than two. Figure 3.11 compares the correlation patterns of PSDR (Figure 3.11a) and PFA (Figure 3.11b) across the three seismic events. The figures show that, at inter-site distances on the order of 10 km or less, the correlation patterns for both PSDR and PFA are very similar for the three events. However, as the inter-site distance increases beyond 10 km, the correlation patterns (PSDR and PFA) for Chi-Chi and Tottori bifurcate from that of Northridge, with the latter having a much faster rate of decrease in the correlation level. Again, these observations are consistent with the results shown in Figure 3.4, which shows that the rate of decay in both PSDR and PFA with rupture distance is highest for the Northridge earthquake.

Prior studies have attributed the differences between spatial correlation patterns across different scenario earthquakes to the event magnitudes. For example, DeBock et al. [14] suggested that the higher levels of correlation observed for Chi-Chi when compared to Northridge are due to the larger magnitude. However, the correlation comparison between Chi-Chi (Mw 7.6) and Tottori (M_w 6.6) does not necessarily support this conclusion. An alternative explanation is suggested in this study. The Northridge earthquake ground motion data were obtained from 152 recording stations spread over an area of approximately 18,500 km^2. In contrast, 400 recording stations spread over an area of approximately 36,200 km^2 are included in the Chi-Chi dataset and the

Tottori dataset was obtained from 414 recording stations located over an area of approximately 112,000 km^2. This suggests that the similar EDP correlation patterns for Chi-Chi and Tottori are more likely due to the similarities in the spatial coverage and density of the ground motion recordings rather than the event magnitudes.

Note that Vs30 was evaluated as a potential second measure of site-dissimilarity to be incorporated in the interpolation model. However, Vs30 was found to have a very small correlation with the response demands at different sites even when controlling for the effect of rupture distance. Moreover, no clear trend was found between the response demand correlations, which were used as the basis of the interpolation models, and the differences in Vs30 across site pairs.

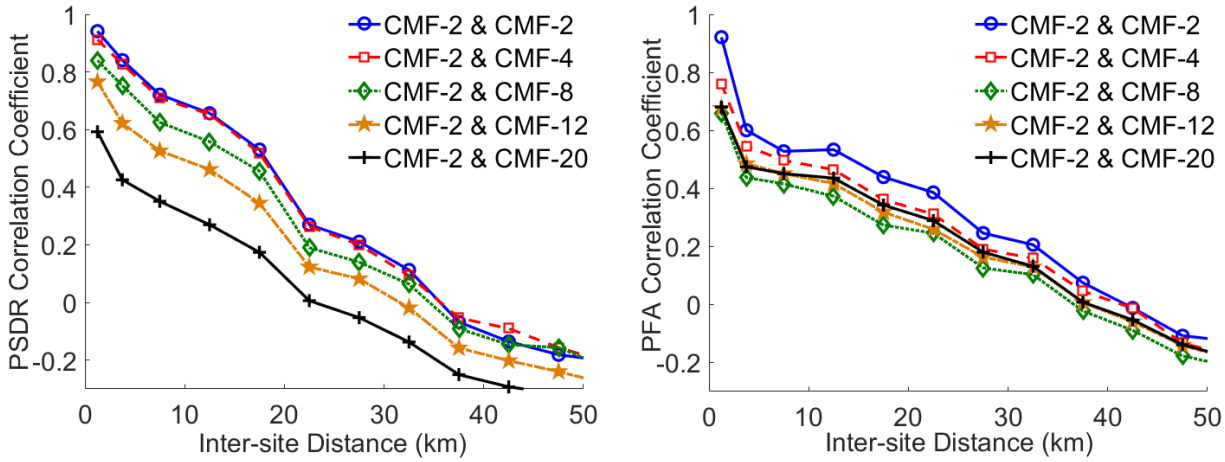


Figure 3.10 Correlation coefficients for (a) PSDRs and (b) PFAs versus the median inter-site distance for CMF-2 paired with all other structures for the Northridge earthquake

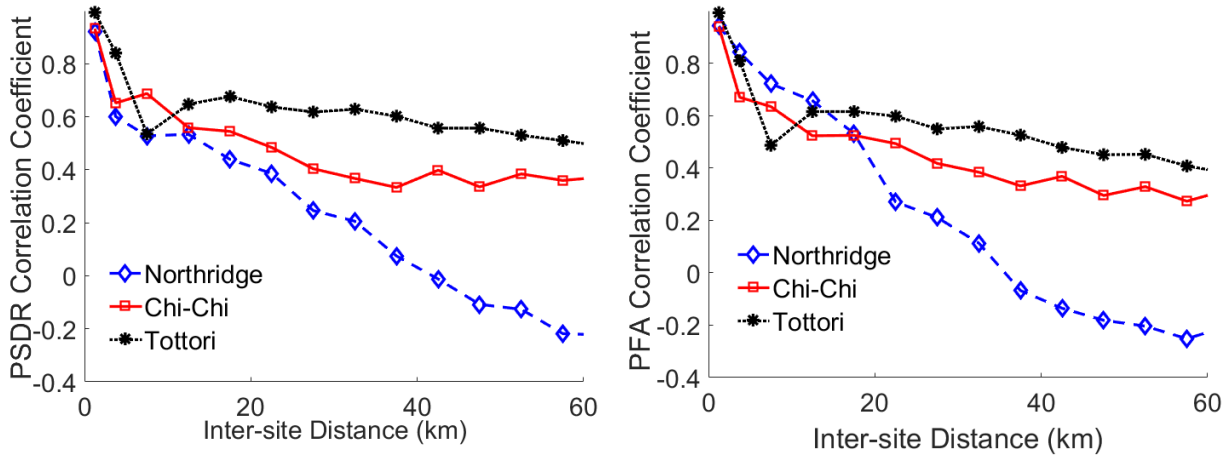


Figure 3.11 Comparing CMF-2 self-correlation pattern for (a) PSDRs and (b) PFAs across the three events

3.4 Modeling Spatial Correlation of EDPs Using Semi-variograms

Semi-variograms are used to visually represent the relationship between the semi-variance of random measurements and the distance between them [107]. Originally developed as a geostatistical tool, semi-variograms have been used to in prior studies to model the correlation between spatially distributed ground motion intensities [96,97]. The current research work extends the application of semi-variograms to model the correlation of peak seismic response demands in spatially distributed buildings. For a random measurement, Z, which is described as a function of its location, s, the semi-variogram for a pair of locations separated by distance h, $\gamma(h)$, is mathematically represented as half the square of its expected value at locations s, (E[Z(s)]), and $s+h$, (E[Z(s+h)]) or the difference between the non-spatial variance for the same pair of locations, $C(0)$, and $C(h)$ respectively [108].

$$\gamma(h) = \frac{1}{2} E[Z(s+h) - Z(s)]^2 = C(0) - C(h) \qquad (3\text{-}2)$$

Note that Equation (3-2) is based on the second-order stationarity assumption, which

implies that E[Z(s)] is constant over the spatial domain. In prior studies, when semi-variograms were used to model the spatial correlation of ground motion intensities, the residual intensity (difference between actual and predicted intensity) represented the random measurement of interest. This residual is assumed to take on a standard normal distribution, which meant that the expected value was always zero and therefore satisfied the second-order stationary assumption. In the current study, the random measurement of interest is the absolute value of the peak structural response demand, whose expected value is location-dependent because both the shaking intensity and structure of interest varies from one location to the next. However, as Goovaerts [108] stated, "stationarity is not a characteristic of the phenomenon under study, but is a decision made by the user, not a hypothesis that can be proven or refuted from data". In other words, the random measurement of interest does not have to follow the second-order stationary assumption. By defining $Z(s) = \mu + \sigma$, μ or $(E[Z(s)])$, can be assumed as a constant over domain s and the spatial variation considered in the residual term σ. This approach is used in the application of Ordinary Kriging in the current study, where the assumption of a constant mean peak response demand is used as a mathematical abstraction for the purpose of developing the Kriging model. Moreover, an alternative interpolation method, Universal Kriging, which does not rely on the second order stationary assumption, is also used. In Universal Kriging, the mean value of the random measurement is computed from ordinary linear or nonlinear regression using the location of each site as the predictors.

The relationship between $\gamma(h)$ and the correlation coefficient between $Z(s)$ and $Z(s+h)$, $\rho(h)$, is described using the following equation [97].

$$\rho(h) = 1 - \frac{\gamma(h)}{C(0)} \tag{3-3}$$

In Equation (3-3), the semi-variogram contains information about the spatial correlation of the random measurement as well as the non-spatial variance, which, in the current study, is used to represent structural dissimilarity. The robust estimator proposed by Cressie and Hawkins [109] and shown in Equation (3-4), is used to construct the semi-variogram in lieu of the classical estimator, to minimize the occurrence of outliers that result from non-normality of the random measurement, which can affect the smoothness of the semi-variogram.

$$\gamma_{robust}(h) = \frac{\left\{ \frac{1}{N(h)} \sum_{N(h)} | X_a - Y_b |^{1/2} \right\}^4}{0.457 + \frac{0.494}{N(h)}} \tag{3-4}$$

Where X_a and Y_b represent residual values at sites a and b respectively with separation distance h. $N(h)$ is the number of pairs within each site-separation bin. By using a fourth order semi-variogram, $\left\{ \frac{1}{N(h)} \sum_{N(h)} | X_a - Y_b |^{1/2} \right\}^4$, the effect of outliers can be reduced. The denominator $0.457 + \frac{0.494}{N(h)}$ is used to ensure that the estimator is unbiased [109].

Figure 3.12 shows semi-variograms obtained from the classical (a) and Cressie and Hawkins estimator (b), using a dataset of PSDRs that was generated from the CMF-2 structure placed at site-pairs with different inter-site distances. Figure 3.12a and b both show that the semi-variance increases with inter-site distance, which is consistent with the inverse of the correlation trend that is typical of semi-variograms. As the inter-site distance increases, the dissimilarity in the EDP measurements, which is reflected in the semi-variogram value, also increases. Comparing Figure 3.12a and b, it can also be observed that the Cressie and Hawkins estimator performs better than classic estimator. For inter-site distances greater than 60km, the former produces a much

smoother (less outliers) semi-variogram. The Cressie and Hawkins estimator is therefore used in the remainder of this study.

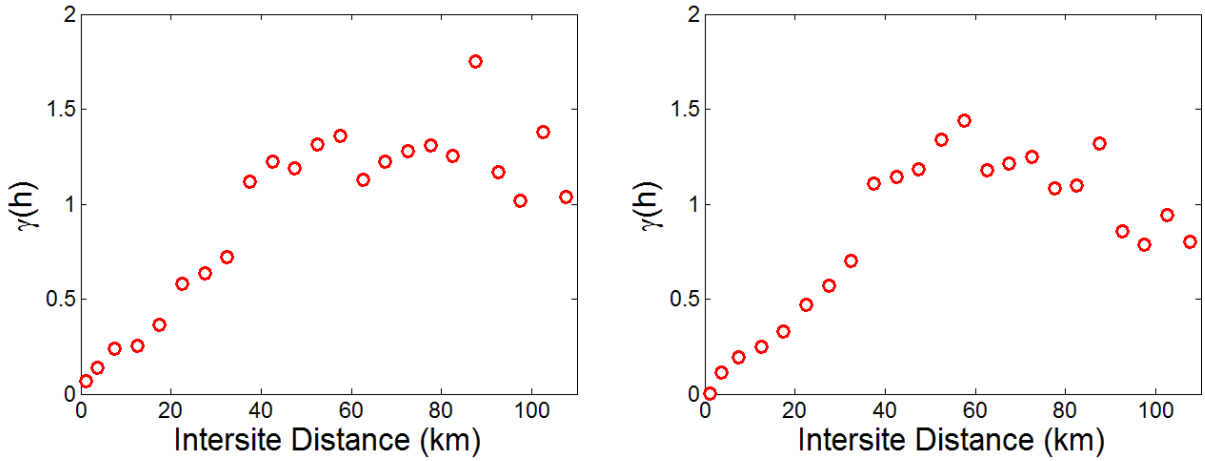


Figure 3.12 Self semi-variogram for CMF-2 constructed using (a) classical and (b) Cressie and Hawkins estimator for the Northridge ground motions

To facilitate the formulation of the Kriging prediction model, the semi-variogram must be described by a continuous positive semi-definite function. The positive semi-definiteness of the covariance function ensures that the variance of any linear combination of the random measurements is non-negative. Admissible models that have been used in previous studies include the Gaussian, exponential, and spherical functions [96,110], which are described in Equations (3-5), (3-6) and (3-7), respectively.

(i) Gaussian variogram model:

$$\hat{\gamma}(h) = (s-n)\left\{1-\exp\left(\frac{h^2}{R^2}\right)\right\}+n \tag{3-5}$$

(ii) Exponential variogram model:

$$\hat{\gamma}(h) = (s-n)\left\{1-\exp\left(-\frac{h}{R}\right)\right\}+n \tag{3-6}$$

(iii) Spherical variogram model:

$$\hat{\gamma}(h) = (s-n)\left\{1.5\left(-\frac{h}{R}\right) - 0.5\left(-\frac{h}{R}\right)^3\right\} + n \qquad (3\text{-}7)$$

Where the nugget, n, is the value of the empirical semi-variogram corresponding to zero inter-site distance, $\gamma(0)$, which is used to represent structural dissimilarity. The sill, s, is the maximum value of the semi-variogram and the range, R, represents the inter-site distance at which the difference between the semi-variogram and the sill becomes negligible. In this study, the sill is obtained from a Gaussian Kernel function [97], which is described in Equation (3-8).

$$K(y_k, \gamma(h_i)) = \sum_{h_i} \exp\left(-\frac{(\gamma(h_i) - y_k)^2}{\sigma}\right) \qquad (3\text{-}8)$$

Where $y_k = 0.01, 0.02, ..., 2.0$, are the sampling points of the Kernel and σ is the Kernel bandwidth. The sill of semi-variogram is determined by maximizing Equation (3-9). As σ is a constant, its value is taken as 0.3 for better visualization of the Kernel function on Figure 3.13. Once an optimal sill is found, a weighted sum of squares loss function is used to find the optimal range, R [110]. The weights are defined by $\frac{1}{h_i}$ at each h_i such that smaller $\gamma(h_i)$ have higher weights. The weighted sum of squares loss objective function is taken as:

$$\min WSS(R_k) = \sum_i \frac{1}{h_i}\left(\hat{\gamma}(h_i) - \gamma_{R_k}(h_i)\right)^2 \qquad (3\text{-}9)$$

Where the optimal range, R, is found by minimizing Equation (3-9) over a range of R_k. An example fitted $\hat{\gamma}(h_i)$ is shown in Figure 3.13. The exponential fitted variogram described by Equation (3-6) is found to be most stable and produces the best Kriging prediction performance. As the inter-site distance increases, the dissimilarity in the EDP measurements, which is reflected in the semi-variogram value, also increases.

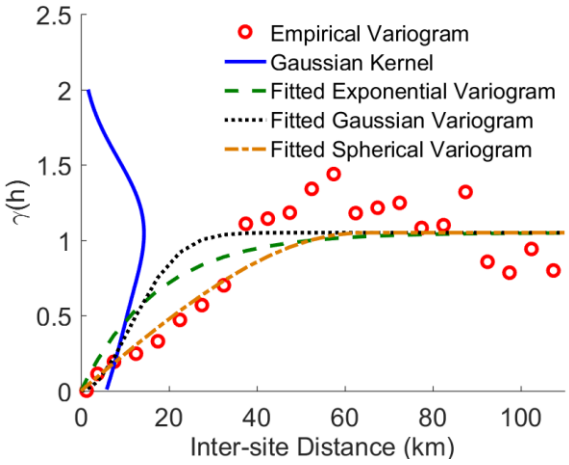


Figure 3.13 Empirical and analytical self-semi-variogram for CMF-2 subjected to the Northridge ground motions

The semi-variograms in the current study are assumed to be geometrically isotropic, which means that the semi-variogram values are direction-independent. Using building CMF-2 as an example, the relationship between the site-pair direction and the semi-variogram value is also shown using a rose plot in Figure 3.15 generated by Figure 3.14 using nine evenly distributed directional semi-variograms. The blue circles are mean values of semi-variograms over each direction and the black crosses are mean plus and minus one standard deviation. Perfect geometric isotropy would result in a circular rose plot. Despite the rose plot shown in Figure 3.14 does not form a perfect circle, it shows that the dataset of self-correlation for building CMF-2 results in a near-perfect circle at the mean level. Similar results were obtained for other cases with inter-site distances less than 50 km in Northridge scenario. This observation is consistent with the fact that, for the Northridge earthquake ground motions the EDP correlations are close to or less than zero at inter-site distances greater than 50 km, which, as a result, is used as the cut-off distance in the correlation model. When interpolating the response demand for a given building, the responses in buildings located at inter-site distances greater than the cutoff distance are not considered.

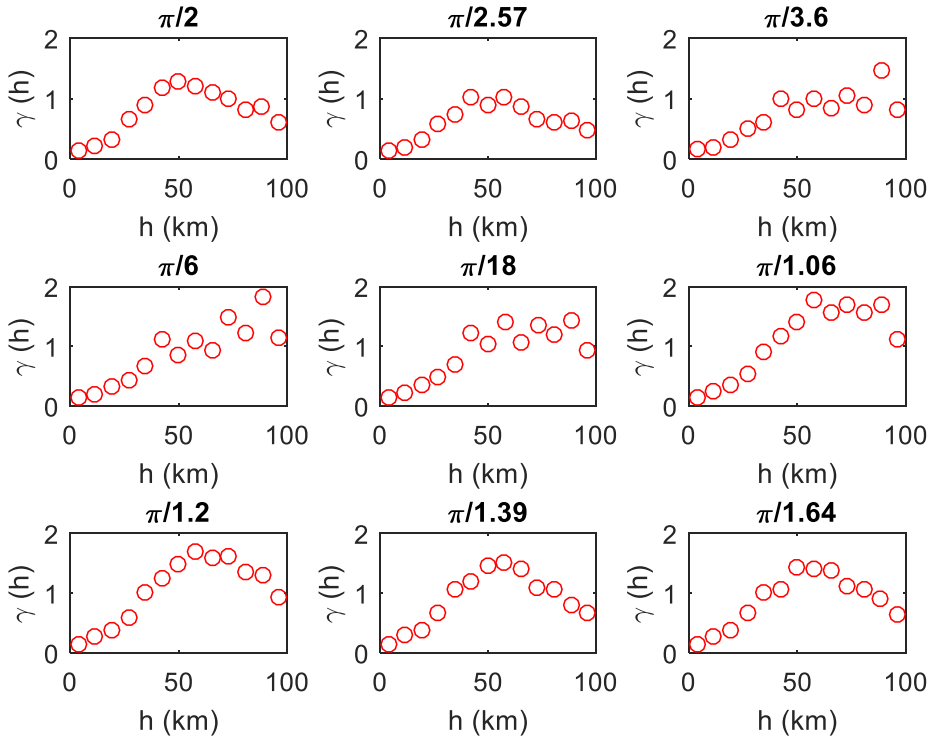


Figure 3.14 Self semi-variogram of PSDR for building CMF-2 with nine evenly distributed directions for the Northridge ground motions

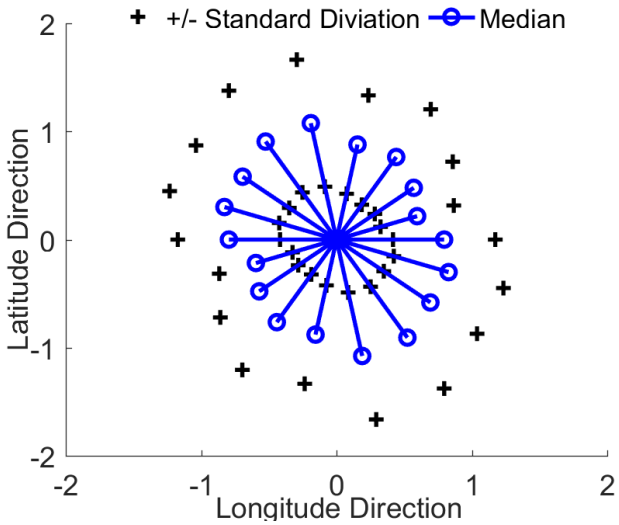


Figure 3.15 Rose plot of self-semi-variogram for building CMF-2 using nine evenly distributed directions for the Northridge ground motions

Cross semi-variogram matrices are developed to incorporate structural dissimilarity by pairing EDPs recorded from different buildings and sites for a single scenario earthquake. Figure 3.16 shows the PSDR cross semi-variogram matrix for the CMF buildings subjected to the Northridge ground motions. The elements of the cross semi-variogram, $\gamma_{B1,B2}(h)$, can be used to compute the EDP correlation for buildings $B1$ and $B2$, which are separated by distance h. The diagonal elements of the cross variogram matrix capture the site-dissimilarity for the same building at multiple locations. Non-diagonal elements capture both site and structural dissimilarity. To quantitatively visualize the effect of structural dissimilarity, the nugget versus the building 1st mode period for each cross-variogram of Figure 3.16 is plotted in Figure 3.17. It can be observed that nuggets between low- and mid-rise buildings (CMF-2 to CMF-12) generally have a linear trend. This suggests that the structural dissimilarity represented by the nugget is linearly related to the first-mode period for low- and mid-rise concrete moment frame buildings. However, the nugget effect for CMF-20 varies nonlinearly with the first-mode period, suggesting that taller buildings have a different pattern of structural dissimilarity. Similar relationships between the 1st mode period of the structure and the optimal sill and range were observed (not shown due to space constraints), which are also needed for the semi-variogram model. While not a direct measure of structural dissimilarity, they represent the effect of structural dissimilarity on spatial correlations. Empirical models can be generated for each of the three semi-variogram fitting parameters (nugget, sill, and range) as a function of a structure's first-mode period. By doing so, the cross semi-variogram matrix can be generated without the EDP dataset, thereby reducing the computational expense of NRHA.

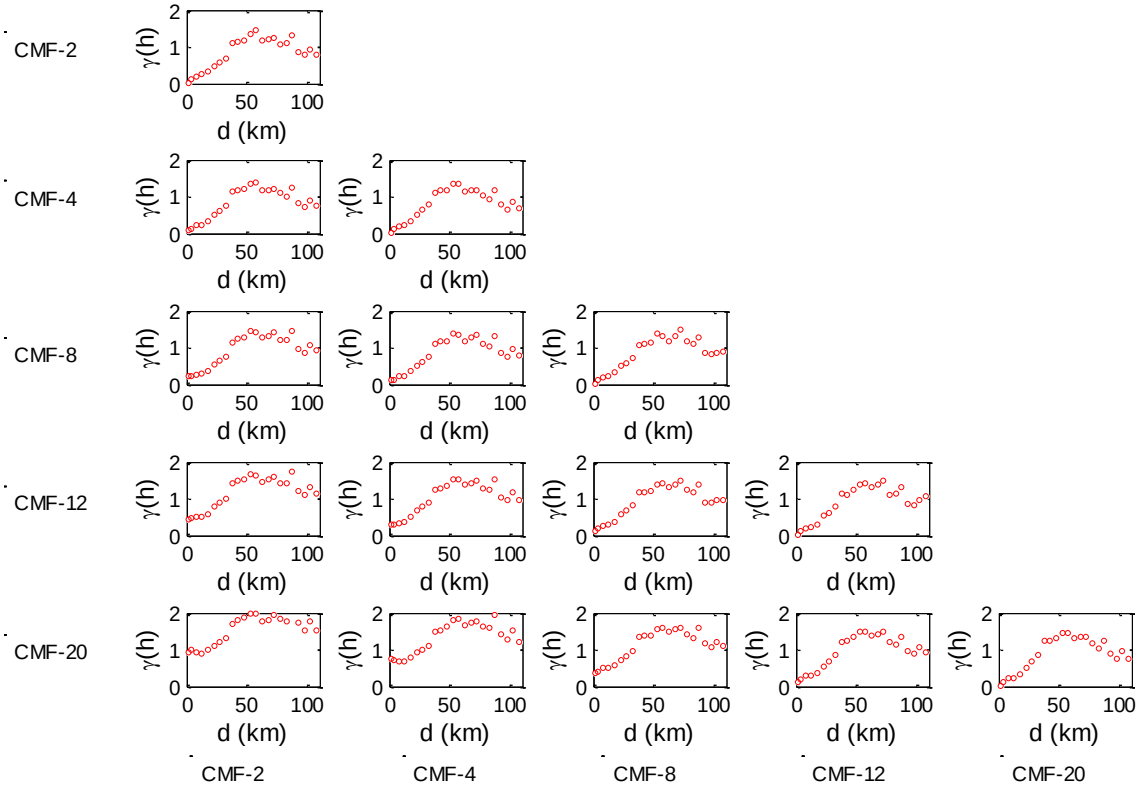


Figure 3.16 Cross semi-variogram obtained from Northridge ground motions

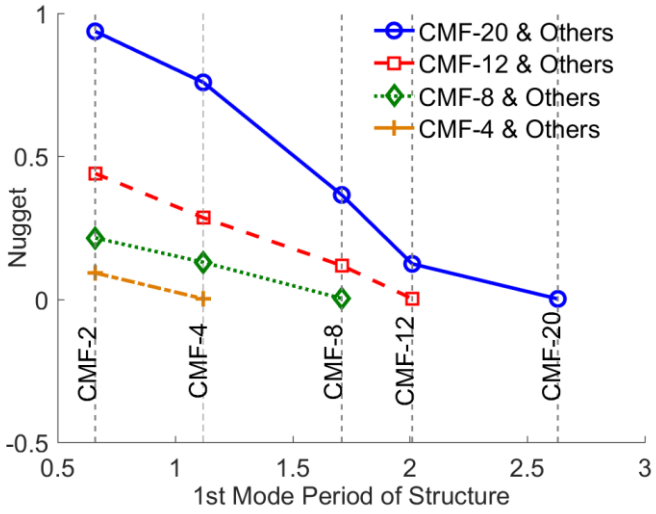


Figure 3.17 Nugget effect of cross semi-variogram obtained from the Northridge ground motions

3.5 Kriging Model for Interpolating Peak Structural Responses

Kriging, which has been widely used in environmental science and geo-statistics, is a spatial interpolation algorithm that performs prediction at locations where a spatial variable is unknown by using a pre-calibrated semi-variogram function. A random variable $Z(s)$, which is defined over a spatial domain $s_1, s_2, \ldots, s_n$, can be described as $Z(s_1), Z(s_2), \ldots, Z(s_n)$. The current study is concerned with formulating a Kriging model to predict peak structural response demands (the random variable of interest) at location s_0, which is denoted by $\hat{Z}(s_0)$, using the weighted sum of the known response demands [111].

$$\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i) \tag{3-10}$$

Equation (3-10) is subjected to the constraint $\sum_{i=1}^{n} w_i = 1$, where w_i is the weight applied to the known response demand at location s_i, $Z(s_i)$. An optimization problem is then defined to

determine the weight at each location by minimizing the mean square error of the prediction obtained by Equation (3-10), σ_e^2.

$$\min \sigma_e^2 = E\left(Z(s_0) - \sum_{i=1}^{n} w_i Z(s_i) \right)^2 \tag{3-11}$$

Equation (3-12) can be obtained by substituting the semi-variogram in Equation (3-2) into the minimization objective function shown in Equation (3-11), the proof of which is described in detail by Cressie [111].

$$\min \sigma_e^2 = 2\sum_{i=1}^{n} w_i \gamma(s_0 - s_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(s_i - s_j) \tag{3-12}$$

To incorporate the constraint $\sum_{i=1}^{n} w_i = 1$, the Lagrange multiplier, λ, is introduced to the minimization objective function.

$$\min \sigma_e^2 = 2\sum_{i=1}^{n} w_i \gamma(s_0 - s_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(s_i - s_j) - \lambda\left(\sum_{i=1}^{n} w_i - 1 \right) \tag{3-13}$$

By taking the derivative of σ_e^2 in Equation (3-13) with respect to λ and each w_i, the optimal weight vector W for this minimization problem can be solved using the Ordinary Kriging system shown in Equation (3-14):

$$\begin{bmatrix} \hat{W} \\ \lambda \end{bmatrix} = \begin{bmatrix} Var_{Z(s_i)} & 1 \\ 1^T & 0 \end{bmatrix}^{-1} \begin{pmatrix} Cov_{Z(s_i),Z(s_0)} \\ 1 \end{pmatrix}$$

$$= \begin{bmatrix} \gamma(s_1 - s_1) & \gamma(s_2 - s_1) & \cdots & \gamma(s_n - s_1) & 1 \\ \gamma(s_1 - s_2) & \gamma(s_2 - s_2) & & \gamma(s_n - s_1) & 1 \\ \vdots & \vdots & \ddots & & \vdots \\ \gamma(s_1 - s_n) & \gamma(s_2 - s_n) & & \gamma(s_n - s_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{pmatrix} \gamma(s_0 - s_1) \\ \gamma(s_0 - s_2) \\ \vdots \\ \gamma(s_0 - s_n) \\ 1 \end{pmatrix} \tag{3-14}$$

Where $Var_{Z(s_i)}$ is an $n \times n$ matrix of the variances between the observed peak response

demands and $Cov_{Z(s_i),Z(s_0)}$ is an $n \times 1$ vector of covariances between observed and unknown peak response demands. $Var_{Z(s_i)}$ and $Cov_{Z(s_i),Z(s_0)}$ can be obtained from the semi-variogram models described in previous section. 1^T is a $1 \times n$ vector of ones, which is associated with the constant mean (second-order stationary) assumption used in the semi-variogram. Estimates of the weighting vector, $\hat{W}$, and the Lagrange multiplier, λ, are obtained from Equation (3-14). Since the location of each of the unknown peak response demands is different at each location, s_0, the weights are different each site. Therefore, the Ordinary Kriging system needs to be solved for each $\hat{Z}(s_0)$ using Equation (10). The prediction variance can also be determined by Equation (3-15).

$$\sigma_e^2 = \sum_{i=1}^{n} w_i \gamma(s_i - s_0) + \lambda \qquad (3\text{-}15)$$

This variance provides a convenient way to evaluate the confidence interval of the prediction model. It should also be noted that the Kriging estimator shown in Equation (3-10) is unbiased. Using these two properties, the predicted peak respond demand can be described by a mean value as well as a range, which is specified by upper and lower bounds.

A modified Kriging algorithm, Universal Kriging, is also applied, which estimates the mean value of the random measurement at a particular location, using a regression trend with respect to the coordinates over the spatial domain. The second-order stationary assumption is therefore waived by computing the mean peak response demands at each site from Ordinary Least Squares (OLS) regression using the site coordinates as predictors using Equation (3-16):

$$Z(s_i) = \mu + \beta_1 s_i^{long} + \beta_2 s_i^{lat} + \beta_3 s_i^{long} s_i^{lat} + \beta_4 (s_i^{long})^2 + \beta_5 (s_i^{lat})^2 + \sigma \qquad (3\text{-}16)$$

Universal Kriging can be interpreted as a two-step process. In Equation (3-16), the $\beta' s$ are coefficients obtained from OLS (step 1). Both linear and 2nd order terms of the coordinates (s_i^{long},

longitude; s_i^{lat}, latitude) are incorporated into the regression equation. The remaining terms after regression are μ and σ, which are determined using Ordinary Kriging (step 2). By combining the two steps, regression and Ordinary Kriging, the Universal Kriging system is obtained as shown in Equation (3-17).

$$\begin{bmatrix} \hat{W} \\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma(s_1-s_1) & \gamma(s_2-s_1) & \cdots & \gamma(s_n-s_1) & s_1^{long} & s_1^{lat} & s_1^{long}s_1^{lat} & (s_1^{long})^2 & (s_1^{lat})^2 & 1 \\ \gamma(s_1-s_2) & \gamma(s_2-s_2) & & \gamma(s_n-s_2) & s_2^{long} & s_2^{lat} & s_2^{long}s_2^{lat} & (s_2^{long})^2 & (s_2^{lat})^2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(s_1-s_n) & \gamma(s_2-s_n) & \cdots & \gamma(s_n-s_n) & s_n^{long} & s_n^{lat} & s_n^{long}s_n^{lat} & (s_n^{long})^2 & (s_n^{lat})^2 & 1 \\ s_1^{long} & s_2^{long} & \cdots & s_n^{long} & 0 & 0 & 0 & 0 & 0 & 0 \\ s_1^{lat} & s_2^{lat} & \cdots & s_n^{lat} & 0 & 0 & 0 & 0 & 0 & 0 \\ s_1^{long}s_1^{lat} & s_2^{long}s_2^{lat} & \cdots & s_n^{long}s_n^{lat} & 0 & 0 & 0 & 0 & 0 & 0 \\ (s_1^{long})^2 & (s_2^{long})^2 & \cdots & (s_n^{long})^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ (s_1^{lat})^2 & (s_2^{lat})^2 & \cdots & (s_n^{lat})^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^{-1} \begin{pmatrix} \gamma(s_0-s_1) \\ \gamma(s_0-s_2) \\ \vdots \\ \gamma(s_0-s_n) \\ s_0^{long} \\ s_0^{lat} \\ s_0^{long}s_0^{lat} \\ (s_0^{long})^2 \\ (s_0^{lat})^2 \\ 1 \end{pmatrix}$$

(3-17)

The performance of the Ordinary and Universal Kriging models are evaluated and compared in the next section.

Model Performance Evaluation

In this section, the overall performance of the peak structural response interpolation model is evaluated. The effect of key assumptions and model characteristics (e.g., level of dissimilarity of buildings, size of training versus test dataset, building distribution, earthquake scenario, and response parameter) on the model performance is also assessed. For a given scenario earthquake, the model performance is evaluated using non-replacement bootstrap, which randomly separates the entire dataset into a training and a testing subset. The model is generated using the training dataset and applied to predict the peak structural responses at the locations of the testing subset. The error measure between true and predicted response demand is described by the median

absolute relative deviation (MARD) shown in Equation (3-18).

$$MARD = Median\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right|\right) \tag{3-18}$$

Where y_i is the nature log of the true response demand, $\hat{y}_i$ is the natural log of the predicted response demand, $i = 1,2,\cdots,n_t$ and n_t is number of testing data points [38]. By computing the MARD for multiple cases of randomly generated testing and training data subsets, a probabilistic distribution of the prediction error is obtained.

The performance of the self-interpolation model serves as a measure of the ability of both Kriging models to interpolate peak structural response demands across similar buildings located at different sites. In other words, the same structure type is used at each testing and training site. An illustration of the self-interpolation results obtained by applying Ordinary Kriging to peak story drift demands for the CMF-2 structure is shown in Figure 3.18. PSDRs and PFAs from the Northridge scenario are used to construct a cross semi-variogram matrix and the diagonal terms are used for the self-interpolation. The training and testing data subsets are comprised and 80% and 20% of the full dataset, respectively.

Figure 3.18 shows a scatter plot, which is used to compare the observed and predicted PSDRs, with the points that lie on the diagonal line representing a perfect prediction. The correlation coefficient between the true and predicted peak PSDRs is computed to be $\rho = 0.89$, which suggests that the interpolation model performs well for the CMF-2 self-interpolation case. The size of each point indicates the mean prediction distance, which is the average distance between the location of an unknown response demand and the locations of known response demands used to generate the prediction. The correlation coefficient between the prediction accuracy and mean prediction distance in most cases is between 0.1 to 0.3, suggesting that the

mean prediction distance does not have a strong effect on model performance.

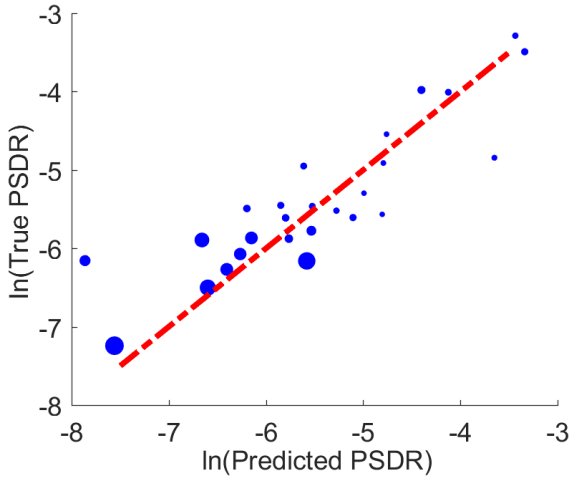
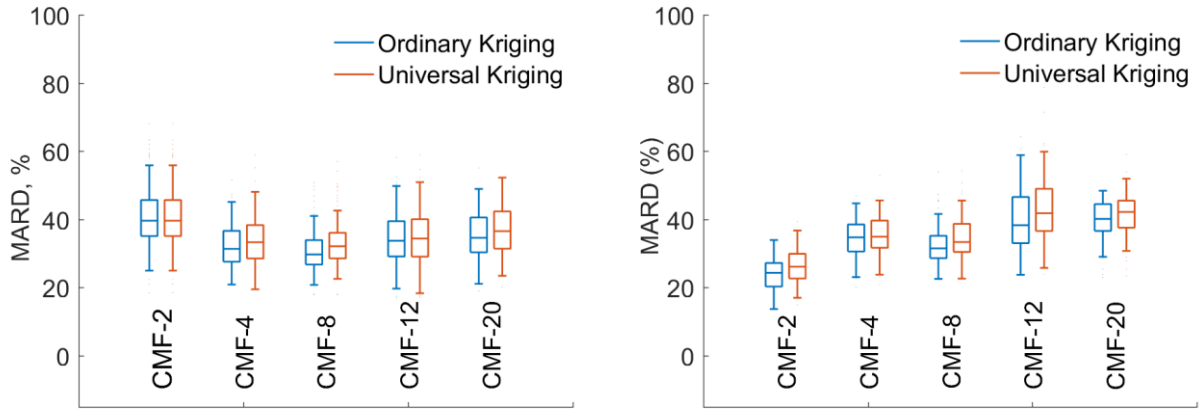


Figure 3.18 PSDR self-prediction result for the CMF-2 structure subjected to the Northridge ground motions

As noted earlier, a more robust way to evaluate model performance is through non-replacement Bootstrap, which repeats the prediction process using randomly generated training and testing data subsets to produce a probabilistically distributed prediction error. Figure 3.19 shows the PSDR (Figure 3.19a) and PFA (Figure 3.19b) self-interpolation results for all five structures obtained from both Ordinary Kriging and Universal Kriging. The performance assessment results are presented in terms of box plots, which show the distribution of MARD values obtained from the Bootstrap procedure using the 80%-20% split for the training and testing data. The upper and lower bound of each box represent 25% and 75% percentile of MARD, respectively, for 100 repeated Bootstrap procedures. The center line is the median MARD value. The individual points are used to represent outliers within the set of MARD values and upper and lower lines represent the error range (excluding outliers). The error distribution from Bootstrap is particularly useful for evaluating the model performance as the median MARD is a measure of accuracy and the range of the box shows stability of the model. Figure 3.19a shows that, for PSDR,

there are only minor differences between the accuracy and stability of the self-prediction model across different structures. However, while the median MARD for PSDR and PFA are not very different, the model performance for the latter appears to worsen as the structure height increases. The accuracy and stability of the Universal Kriging model is on par with that of Ordinary Kriging for the PSDR, and slightly less accurate for the PFA prediction. Subsequent results are presented for Ordinary Kriging only.



(a) (b)

Figure 3.19 Bootstrap evaluation of (a) PSDR and (b) PFA self-prediction model for all structures

To evaluate the ability of the Kriging models to interpolate peak response demands between structurally dissimilar buildings (cross-interpolation), the location and number of each building type must be pre-determined. The accuracy and stability of the cross-interpolation models are evaluated using a randomly generated building location and distribution (therefore equal percentage of each building type in full dataset). The off-diagonal terms of the cross semi-variograms (Figure 3.16) are used to interpolate peak response demands in structurally dissimilar buildings. Figure 3.20 summarizes the Northridge scenario PSDR cross-interpolation performance. The cross-interpolation performance between the CMF-20 and the other four structures is shown

in Figure 3.20a. It shows that both the accuracy and stability of the cross-interpolation decreases with the difference in building period. Figure 3.20b shows the median MARD values for the cross-interpolation between all five structures. For cross-interpolation from the CMF-20 to the CMF-2 and -4 structures (and vice versa), the median MARD value is approximately twice that of the cross-interpolations among the other structures.

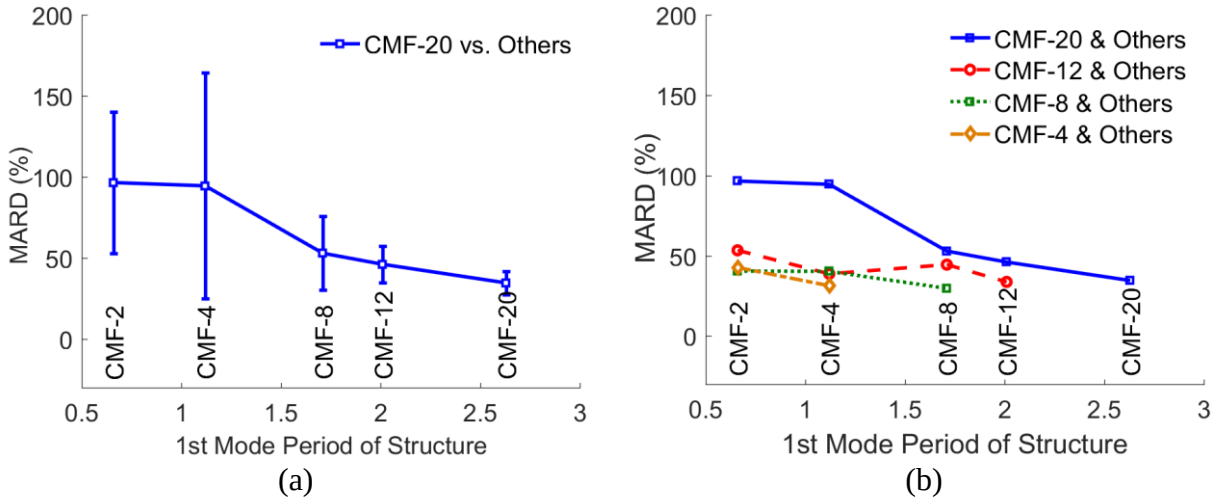


Figure 3.20 PSDR cross-prediction model performance (a) between CMF-20 and all other structures and (b) for all structure combinations (only median MARD)

The previously described model performance evaluation cases are based on 80% training data and 20% testing data. However, for practical applications of the proposed interpolation model, the number of known building responses is constrained by the fraction of instrumented buildings within a portfolio and availability of sensors within each building. As such, it is necessary to understand the effect of the training-testing data split on the model performance. Figure 3.21 shows the effect of the percentage of the responses used as training data (analogous to the fraction of instrumented buildings in a portfolio) on the median MARD value for self- (CMF-2) and cross-interpolation (CMF-12 and CMF-20 structures). For the self-interpolation, the percentage of training data does not appear to have a significant impact on the model performance. However, for

the cross-interpolation cases, the median MARD decreases as the percentage of training data increases. The increase in the median MARD value at 80% training data is mainly due to overfitting. This result suggests that the need for an adequate proportion of instrumented buildings is more important for predictions across structurally dissimilar buildings.

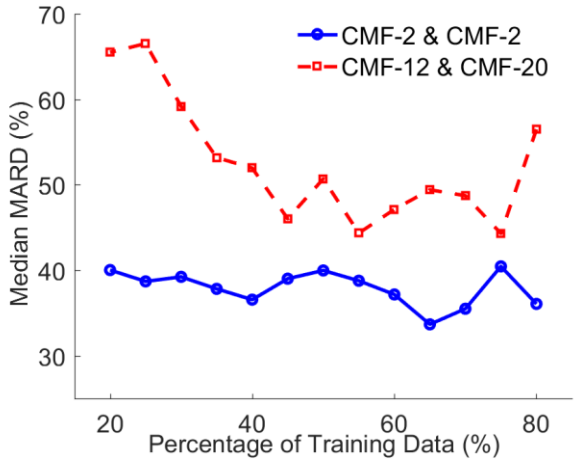


Figure 3.21 Median MARD versus training dataset size

Previous studies suggest that the spatial correlation of the spectral acceleration corresponding to a building's first mode period (SA_{T1}) is a good predictor of the correlation in structural response [17, 19]. Moreover, SA_{T1} is generally more likely to be retrievable after an earthquake (compared to measure structural responses) since it is generated from ground motion recordings which are generally more ubiquitous. These two pieces of information lead to the question of whether the peak response interpolation models can be constructed using semi-variograms of SA_{T1} instead of measured responses. To investigate this question, a cross semi-variogram of SA_{T1} was constructed and used to formulate the PSDR interpolation model. The performance of the IM-based (solid line) and SDR-based (dash line) semi-variogram models is compared in Figure 3.22, which shows the median MARD values for 100 Bootstrap interpolations versus the structure's first-mode period. For the self-correlation cases, the performance of the PSDR- and IM-based semi-variogram interpolation models is almost identical. However, more

often than not, the PSDR-based model has a lower median MARD than the IM-based model.

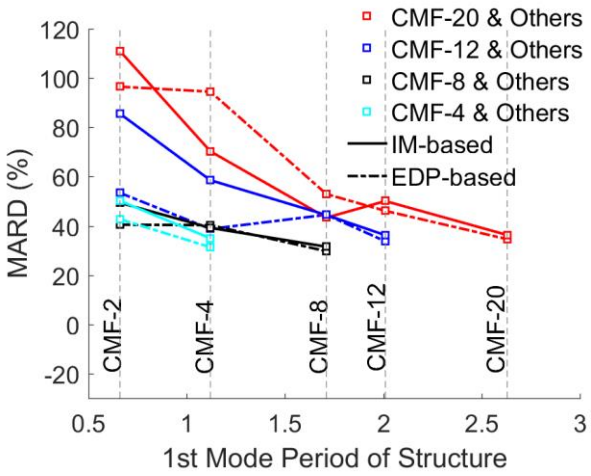


Figure 3.22 Comparing performance of cross-interpolation model between all structures for IM- and PSDR-based semi-variograms

Figure 3.23 evaluates the effect of the response demand level on interpolation model performance, while comparing the overall performance across the three events. The response demand level is increased by uniformly scaling the ground motions for each event by factors of 0.5, 1.5, 2, 2.5 and 3. The PSDR and PFA self-interpolation performance for CMF-2 is compared in Figure 3.23a and b respectively. Both figures clearly show that the median peak EDPs generated by the Tottori event are much lower than those obtained from Northridge and Chi-Chi. However, it should be noted that the demands at sites near the epicenter are similar for all three events. The much lower median response demands produced by the Tottori event is a result of the much larger distribution of the geo-location of the sites where ground motions were recorded (as discussed in 3.2.1). Recall that the interpolation model relies of EDP correlations that decay with inter-site distance. The much larger spatial coverage for the Tottori sites compared to Northridge and Chi-Chi means that most of the site pairs have greater inter-site distances and low or inadmissible (due to cutoff distance) PSDR and PFA correlations, which adversely affects the model performance. The fact that the model constructed using Chi-Chi data generally outperforms the one developed

using Northridge data is consistent with the relative correlation trends shown in Figure 3.11.

Figure 3.23a also shows that, initially, an increase in the PSDR level increases the model accuracy (as evidenced by the lower median MARD). However, for all three events, the performance of the PSDR and PFA models peak at a particular EDP level, beyond which the performance worsens (as evidenced by an increase in the median MARD).

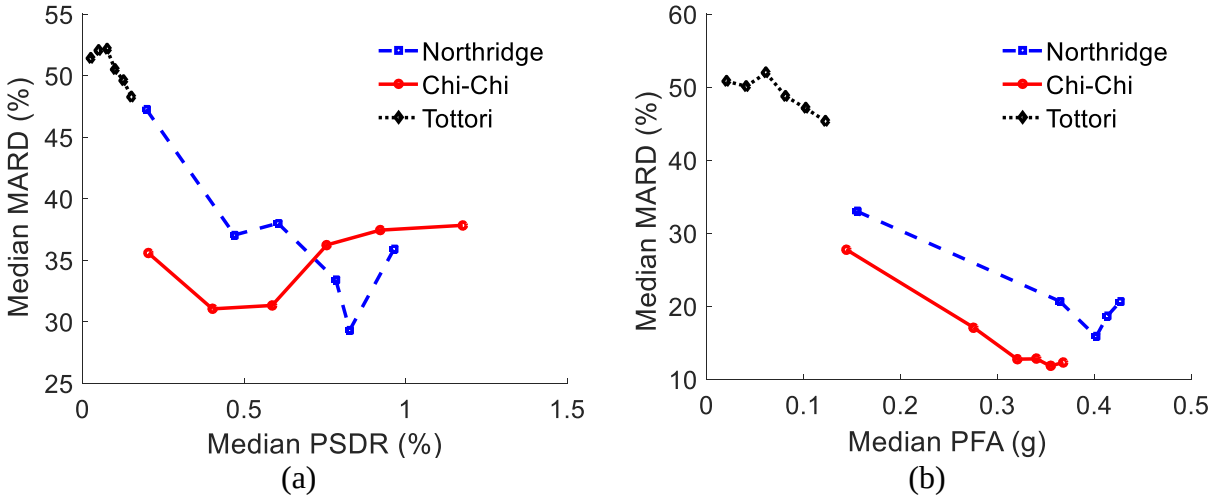


Figure 3.23 Effect of response demand level on model performance for(a) PSDR and (b) PFA self-prediction for CMF-2

Recall that the geographical areas covered by the Northridge, Chi-Chi and Tottori ground motion dataset are 18,500 km^2, 36,200 km^2 and 112000 km^2, respectively. All of the model performance evaluation results presented up to this point are based on using the full dataset (response demands for all sites) for all three events. To investigate how the spatial distribution of the building locations relative to epicenter affects the model prediction accuracy, the building pools were grouped based on whether they fall within concentric circles of varying radii, R. For the self-interpolation case corresponding to CMF-2 buildings, one hundred Bootstrap procedures with an 80-20 training-testing data split were conducted and the median MARD for each case is reported as shown in Figure 3.24. For the Northridge event, whose baseline dataset has the smallest

spatial coverage relative to the other events, there is very little variation in the self-interpolation accuracy as the range of buildings being included (the value of R) increases. However, an observable trend is noted for Chi-Chi and Tottori with the latter having the strongest trend. For both these events, the model performance initially increases to a peak accuracy level after which it worsens. At R = 80 km, the Tottori model performs on par with that of Northridge and Chi-Chi. However, there is a sharp increase in the median MARD of the Tottori model beyond R = 80 km, which is consistent with the much worse performance (compared to Northridge and Chi-Chi) observed when the full dataset (roughly R = 190 km) is used (Figure 3.23).

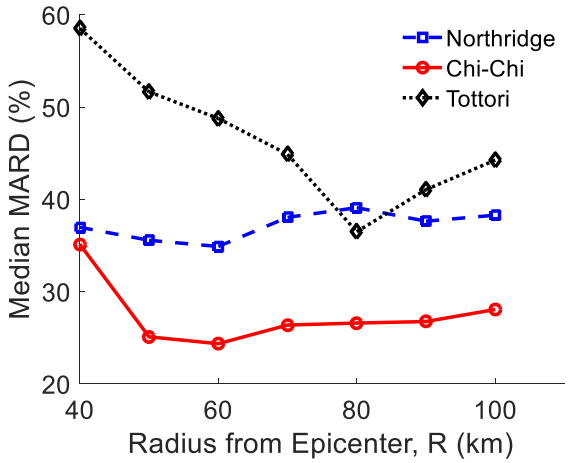


Figure 3.24 Effect of number and spatial clustering of buildings on the performance of the interpolation model for the Northridge earthquake scenario

3.6 Summary

This study extends the application of geostatistical tools to develop models to interpolate peak seismic response demands across structurally similar and dissimilar buildings. Site-specific ground motion recordings were obtained for the Northridge, Tottori and Chi-Chi earthquake scenarios and nonlinear response history analyses (NRHA) were conducted to obtain peak story drift ratios (PSDRs) and peak floor accelerations (PFAs) for five modern code-conforming reinforced concrete moment frame structures ranging in height from 2- to 20 stories. The NRHA

results were used to characterize the spatial self- (same building at different sites) and cross-correlation (different building at different sites) of response demands for the three earthquake scenarios. Spatial self- and cross-correlation of response demands were modeled using semi-variograms, which provide a measure of spatial and structural dissimilarity. The semi-variograms were used to develop Kriging models to construct inter-building interpolation models of peak response demands.

Self- and cross-interpolation models were developed using both Ordinary and Universal Kriging. The former relies on an assumption of 2nd order stationarity of spatially distributed response demands. This assumption is adopted even though both PSDRs and PFAs are expected to vary with the level of ground shaking. We argue that the 2nd order stationary assumption is used as a mathematical abstraction and the spatial variation of response demands was considered in the residual term. Alternatively, Universal Kriging, which was also applied in this study, does not rely on the 2nd order stationarity assumption. Instead, the mean value of the random measurement is obtained from regression using the site coordinates as predictors.

Various aspects of the interpolation model performance were investigated including (a) the effect of structural dissimilarity, (b) the proportion of known response demands used to train the model (relative to the entire building stock), (c) the use of intensity-measure based semi-variograms to develop the Kriging model, (d) the effect of the geographic area covered by the building stock, and (e) the magnitude of the response demands. The performance of the self-interpolation model was generally consistent across the five structure types. For the cross-interpolation models, the performance worsened as the difference in the fundamental period of the structure increased. For the specific structure types considered in this study, the cross-interpolation models for the 2-, 4-, 8- and 12-story structures were generally on par with the performance of the

self-interpolation models. The same can be said for 8-, 12- and 20-story structures. The interpolation between the 2- and 20-story and 4- and 20-story structures was much worse than the self-interpolation models. The proportion of known building responses (relative to the total building stock) did not have a significant effect on the accuracy of the self-interpolation model. However, the performance of the cross-interpolation model deteriorated as the fraction of known response demands decreased. Self-interpolation models constructed using IM-based semi-variograms were shown to perform on-par with semi-variograms constructed using correlated response demands. However, for cross-interpolations, the EDP-based semi-variogram model generally performs better.

Several challenges are likely to arise in the real-world implementation of inter-building structural response interpolation models. Ideally, the prediction models should be developed using structural responses recorded from instrumented buildings with different types of structural systems (e.g. shear walls, steel moment frames and braced frames). Given the general lack of availability of recorded earthquake building response, the demands used in this study were generated using NRHA. Moreover, the models presented in this work were developed using a single scenario and used to interpolate responses for that same scenario. This type of model can only be applied in cases where, right after the occurrence of a particular event, an interpolation model is trained using the available response data from instrumented buildings and used to infer responses in un-instrumented buildings, all within the same event. In other words, the models developed in this study can only be applied to the same event that generated the training data. Three different scenario earthquakes were used to evaluate the effect of event characteristics on model performance. To develop a generalized prediction model, training data will need to be obtained from multiple earthquakes to enable direct prediction without using data from a given

event to train for that event. Moreover, to remove the effect of event specific characteristics, the interpolation will need to be performed on EDP residuals (as opposed to absolute value of the EDP). For the dataset used in this study, V_{s30} was not found to have a strong correlation with the response demands. In cases where V_{s30} has a measurable impact on response demands, a second site dissimilarity measure could be incorporated in addition to inter-site distance. In this case, three-dimensional semi-variograms and fitted surface functions with inter-site distance and V_{s30} on the horizontal axes would be developed. The other parts of the proposed methodology would remain the same. Alternatively, V_{s30} can be incorporated by creating a pseudo site dissimilarity measure that is a function of multiple site parameters (inter-site distance, V_{s30}, rupture distance etc.).

4. Reconstructing Seismic Response Demands across Multiple Tall Buildings using Kernel-based Machine Learning Methods

4.1 Introduction

Buildings located in regions with high seismic hazard are often instrumented with sensors (e.g., accelerometers) to enable rapid assessment of damage, safety, and functionality following an earthquake. However, for reasons such as cost and inaccessibility, these sensors are usually placed in a limited number of locations within the structure such that, when earthquake shaking occurs, response demands are measured at isolated locations. For example, the City of Los Angeles requires seismic instrumentation of new buildings with more than ten stories or more than six stories and an aggregate floor area greater than $5574 m^2$. A minimum of three accelerometers must be installed at the base, mid-level and roof. With this type of sensor arrangement, acceleration histories can be recorded at three floor levels. Velocity and displacement histories at the same floor levels can also be computed using the recorded accelerations. To facilitate damage detection, it is desirable to reconstruct (or interpolate) the seismic demands in the unmeasured locations using the measured responses. The accuracy of these reconstructed responses is determined by the accuracy, number, and location of the sensors used to generate the measured responses.

Several approaches to structural response reconstruction can be found in the research literature. Naeim et. al. [82] incorporated cubic spline interpolation in the development of an automated building seismic response analysis and visualization system. Spline interpolation has also been used as the basis for determining the optimal placement of sensors within a building (e.g., [112]). Several structural response reconstruction methods incorporate the concept of transmissibility, which, for a single-degree-of-freedom system, is the ratio between the modulus of the response and imposed motion amplitudes. For a multi-degree-of-freedom system, the

transmissibility matrix can be used to form a relationship between measured and unknown responses [113]. Empirical mode decomposition (EMD) coupled with finite element modeling, also has been used to reconstruct structural dynamic responses in unmeasured locations [114]. The EMD method decomposes measured responses into mode shapes and functions, which are then used to compute the modal responses at the unmeasured locations. Wan et al. [115] extended the EMD method to consider structures with closely spaced modes.

The above-mentioned structural response reconstruction methods enable interpolation of seismic demands measured in specific locations of a structure to other locations within the same structure. However, to support rapid post-earthquake evaluation of building portfolios, seismic response reconstruction both within and across multiple buildings is desirable. To this end, Sun et al. [28] developed a methodology for interpolating peak seismic response demands across buildings. The "interbuilding" interpolation model utilized spatial and structural correlation of response demands in co-regionally located buildings subjected to the same seismic event. Semi-variograms were used to model the correlations and the Kriging algorithm was used to interpolate response demands across multiple buildings. A key limitation of this methodology is that the interpolation is limited to a single location (peak demand) within each building.

In this paper, Kernel based machine learning methods are used to reconstruct response demands across multiple tall buildings (20 to 42 stories) (Figure 4.1). In contrast to the earlier approach by Sun et al. in [28], the current method can reconstruct full-profile response demands using measurements at limited locations in a subset of buildings by employing Kernel Ridge and Support Vector Regression along with the Ordinary Least Squares method. The latter method is used as a point of comparison to assess the accuracy of the two kernel-based methods. Kernel regression algorithms have been proven to be effective in recognizing highly nonlinear data

patterns [31]. The current study deals with structural responses for a cluster of tall buildings where nonlinear spatial patterning is expected both within and across multiple buildings. It is noted that previous studies have adopted the use of machine learning methods for predicting (not reconstructing) structural responses. Examples include using Lasso regression for predicting seismic bridge responses [56], Bayesian updating of deterioration parameters across bridges after spatial interpolation of limited instrumentation data [116] and applying artificial neural networks for estimating seismic vulnerability of skewed bridges [117]. The prediction models presented here are constructed and evaluated using the results of nonlinear response history analyses (NRHAs), which are performed using topographically explicit ground motions from the Northridge earthquake. The proposed multi-building seismic response reconstruction methodology has the potential to enhance the usefulness of remote seismic sensing by enabling rapid assessment of portfolio-scale damage to inform emergency response and recovery-related decision-making.

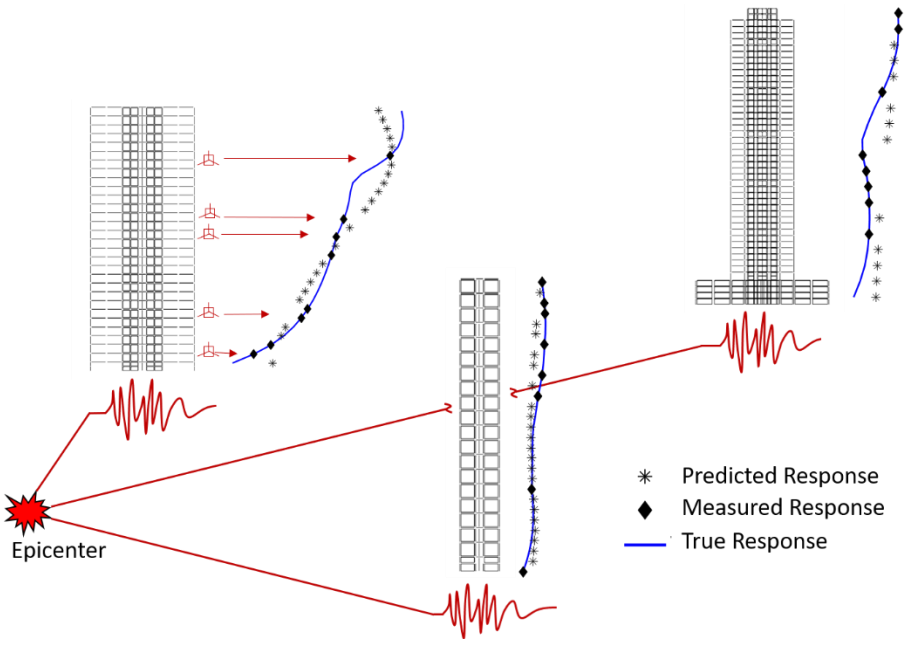
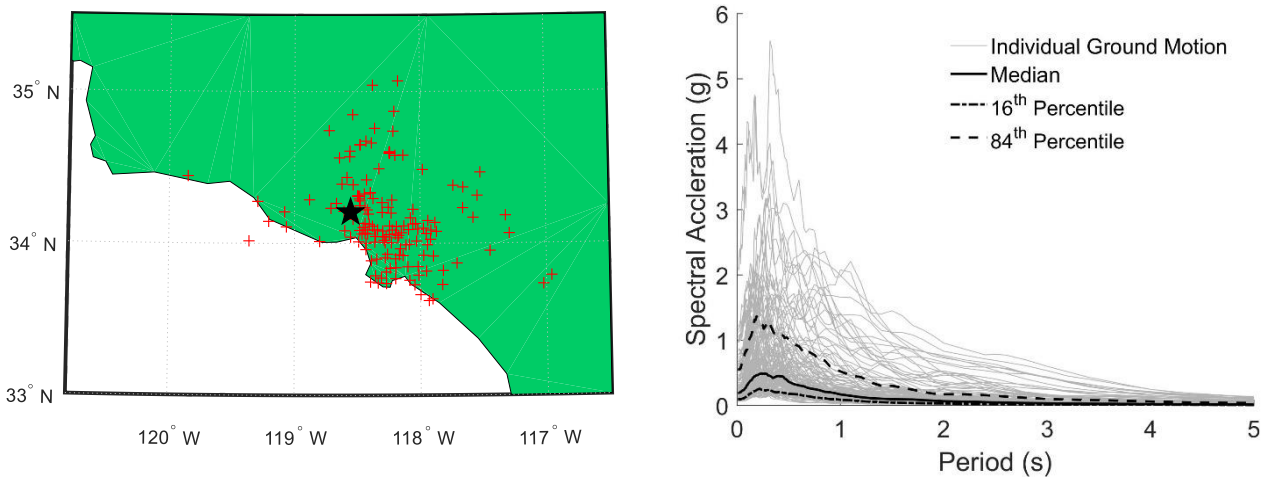


Figure 4.1 Multibuilding seismic response reconstruction framework

4.2 Scenario Earthquake and Ground Motions

The proposed framework is demonstrated using ground motions from the Mw 6.7 1994 Northridge earthquake. Figure 4.2a shows a map of the considered region, the 152 ground motion recording stations (red crosses) and the location of the epicenter (34.2057 N, 118.5539 W). Orthogonal pairs of ground motions corresponding to each recording station were obtained from the PEER NGA-West2 database [99]. Key ground motion properties are summarized in Table 4.1. There is significant variation in the soil properties in the region, as V_{S30} (shear-wave velocity averaged over the top 30m of soil) ranges from 161 m/s to 2016 m/s with a median and coefficient of variation of 347 m/s and 0.56, respectively. The distance from the recording stations to the epicenter ranged from 5.2 km to 147.6 km. More than 90% of the geometric mean peak ground accelerations (PGAs) are less than 0.56g; however, PGA values as high as 2g were recorded at locations near the epicenter.

Response spectra at 5% damping for the 152 pairs of ground motions are shown in Figure 4.2b, including the median, 16^{th} and 84^{th} percentiles. The maximum 84^{th} percentile spectral value is 1.36g, which is observed at 0.19s and drops off to 0.52g at 1s and 0.17g at 2s. The first mode period of the tall buildings considered in this study ranges from approximately 2.5s to 5s, which corresponds to 84^{th} percentile spectral values of 0.20g and 0.03g, respectively. However, due to higher mode effects, the building responses are expected to also be affected by periods as low as 1.37s (3^{rd} mode period of one of the building cases). Periods higher than 5s are also expected to influence the inelastic response.



(a) (b)

Figure 4.2 (a) Station map (red dots are stations; black star is epicenter) and (b) Response spectrum acceleration at 5% damping of all records for the Northridge earthquake

Table 4.1 Summary of 1994 Northridge earthquake

Statistics	V_{S30} (m/s)	Rupture Distance (km)	Geometric Mean PGA (g)
Upper Bound	2016	147.6	2.005
Lower Bound	161	5.2	0.047
Median	347	37.5	0.158

4.3 Description of Buildings and Structural Modeling

4.3.1 Building Descriptions

Four modern code conforming buildings ranging from 20 to 42 stories are used for the current study (summarized in Table 4.2). For three of the buildings, reinforced concrete (RC) shear walls are used for the lateral force resisting system (LFRS) and the fourth building has a dual system with RC shear walls and special moment frames. A key feature of the proposed model is the ability to reconstruct full profile seismic response demands across co-regionally-located buildings with varying degrees of structural dissimilarity. All four buildings used in the current

study have RC shear walls, which is the most common LFRS used in tall buildings constructed since roughly 2000 located in high seismic regions of the United States. However, structural dissimilarities are incorporated through variations in building height and other LRFS elements, such as outriggers and special moment frames. For example, TB-1 (20-story) and TB-4 (42-story) both have RC core walls as their LFRSs but differ in height while TB-2 (30-story) is the only building with an outrigger system. TB-3 and TB-4 have the same height and plan layout; however, in addition to RC core walls, the LFRS of the latter building includes RC special moment frames. Additionally, two different design procedures were used for TB-3 and TB-4. The former is code-based [118] and latter was developed using performance-based design procedures [119]. There are three basement levels in TB-3 and TB-4 but none in TB-1 and TB-2. The LFRS layout in TB-2 is symmetric in plan resulting in the same fundamental 1st and 2nd mode period. The other three buildings have asymmetric LFRS plan configurations. These variations in design procedures, type and configuration of LFRS elements and building layout are explicitly incorporated in the structural models.

Table 4.2 Summary of building characteristics

Building ID	No. of Stories (Basement Levels)	Primary LFRS	Secondary LFRS
TB-1	20 (0)	RC Shear Wall	-
TB-2	30 (0)		Outrigger
TB-3	42 (3)		-
TB-4	42 (3)		RC Moment Frame

TB-1 is designed using seismicity parameters that are based on a hypothetical location in West Los Angeles with $S_S = 1.24g$ (5 percent damped spectral response acceleration parameter at short periods) and $S_1 = 0.48g$ (5 percent damped spectral response acceleration parameter at a

period of 1s [101]), assuming site class D. Four different reinforcement ratios are used for the vertical steel reinforcement, varying from 3% in the confined concrete zone of stories one through five to 0.6% in stories 15 through 20. The length of the special boundary element zone is also varied along the height of the building while the thickness of the core wall is 61.0 cm throughout.

TB-2 is a 30-story coupled RC core wall building with an inelastic slab outrigger system. It was designed by Kim [120] in accordance with the 2012 Intentional Building Code [121], 2013 California Building Code [122] and the Los Angeles Tall Buildings, Structural Design Council (LATBSDC) guidelines [119]. The seismicity is based on a location in downtown Los Angeles with site class C, $S_S = 2.15g$ and $S_1 = 0.72g$. The typical plan configuration of the floor plate can be found in [120]. The core wall is 86.4 cm thick in the first fifteen stories and 61.0 cm thick in the stories above. The vertical reinforcement ratio is 2% within the special boundary element, which varies in length from 76.2 cm to 198.1 cm. The major difference between the structural system of TB-1 and TB-2 is the inclusion of an inelastic slab outrigger system in the latter. The effective slab stiffness is determined using the procedure developed by Kang and Wallace [123]. Other relevant information can be found in Section 4.5 of [120].

The 42-story dual system (TB-4) and coupled core wall only buildings (TB-3) are from the Tall Building Initiative Project [124] and are designed for a site located in the Los Angeles area. The basement levels and a 3-story podium of both buildings are included in the structural model. The thickness of the TB-3 core wall is 24 inches (61.0 cm) in stories 1 through 25 and 53.3 cm in the stories above. For TB-4, the core wall thickness is 61.0 cm in stories 1 through 20 and 45.7 cm in the stories above. Further details about the buildings can be found in [124] and [125] where TB-3 and TB-4 are identified as Building 1A and Building 2B, respectively.

4.3.2 Structural Modeling

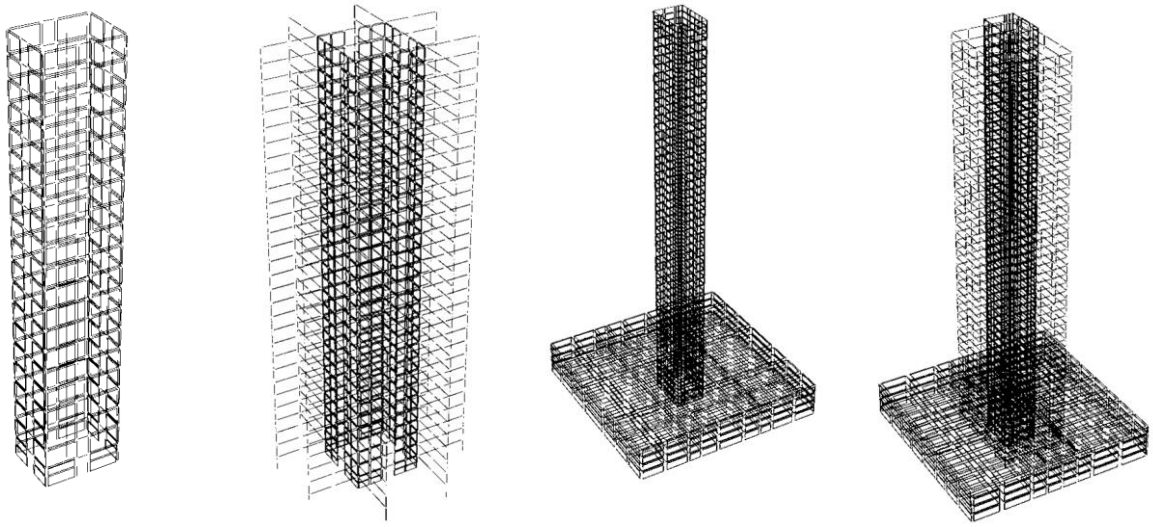
Nonlinear structural models are developed for all four buildings using Perform3D [43]. Core walls, coupling beams, moment frames and outriggers (where they occur) are included in the structural models (Figure 4.3). With the exception of the slab outrigger, the gravity framing is excluded from all four models. A leaning column is placed at the geometric center of the floor diaphragm to account for P-delta effects. Expected material strengths, which are used in all models, are obtained by amplifying the nominal strengths in accordance with Table 4.2 of the LATBSDC guidelines [119]. The elastic modulus is taken to be $E_c = 40{,}000\sqrt{f_c'} + 10^6\ psi$ for high strength concrete and the effective shear modulus is assumed to be $0.2E_c$. Steel and concrete fiber elements are used to model axial-flexural behavior of concrete core walls. The coupling beams are modeled using a shear hinge element calibrated according to the method proposed by [126,127]. A rigid diaphragm is incorporated in all buildings. Monotonic and cyclic degradation parameters of outrigger beams are calibrated based on the results of experiments by [128]. The TB-1, TB-3 and TB-4 models have 2.5% Rayleigh damping corresponding to 20% and 100% of the 1st mode period. For the TB-2 model, 3% Rayleigh damping at 33% and 150% of the first mode period is used. The period of the first three modes of each building are summarized in Table 4.3. TB-1 and TB-2 have similar 1st mode periods at around 2.5s. The 1st and 2nd mode period of TB-2 are the same reflecting the symmetric wall layout. Despite the similarity in the height and floor layout of TB-3 and TB-4, the latter has a moment frame system in addition to the shear wall, which results in a higher initial stiffness and a smaller 1st mode period (4.28s) compared to TB-3 (5.12s). However, the 2nd and 3rd periods of TB-3 and TB-4 are comparable. For all 4 building models, the 1st and 2nd modes are translational and the 3rd mode is torsional.

The abovementioned modeling assumptions are commonly adopted in nonlinear structural

response simulation of tall buildings. It is also worth noting that, while the modeling approach would affect the simulated responses, the proposed data-driven method is model-agnostic.

Table 4.3 Summary of building periods

Building ID	Mode 1 Period (s)	Mode 2 Period (s)	Mode 3 Period (s)
TB-1	2.54	1.76	1.37
TB-2	2.48	2.48	2.18
TB-3	5.12	3.97	2.40
TB-4	4.28	3.89	2.27



(a) (b) (c) (d)

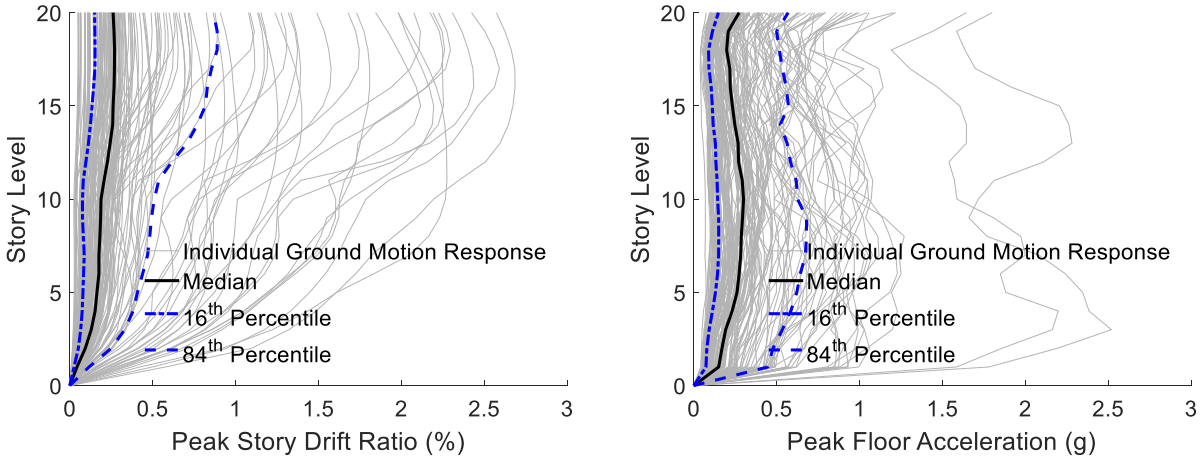
Figure 4.3 Overview of Perform3D model (a) TB-1 (b) TB-2 (c) TB-3 (d) TB-4

4.4 Nonlinear Strucutral Response Simulation

All four structures are analyzed using 152 pairs of unscaled Northridge earthquake ground motions. The peak story drift ratio (PSDR) and peak floor acceleration (PFA) profiles in each direction are the response parameters of interest because they are known to be correlated with structural and non-structural damage [119]. By considering the maximum value of peak response demands across the two orthogonal directions, the effect of asymmetry in the LFRS is considered.

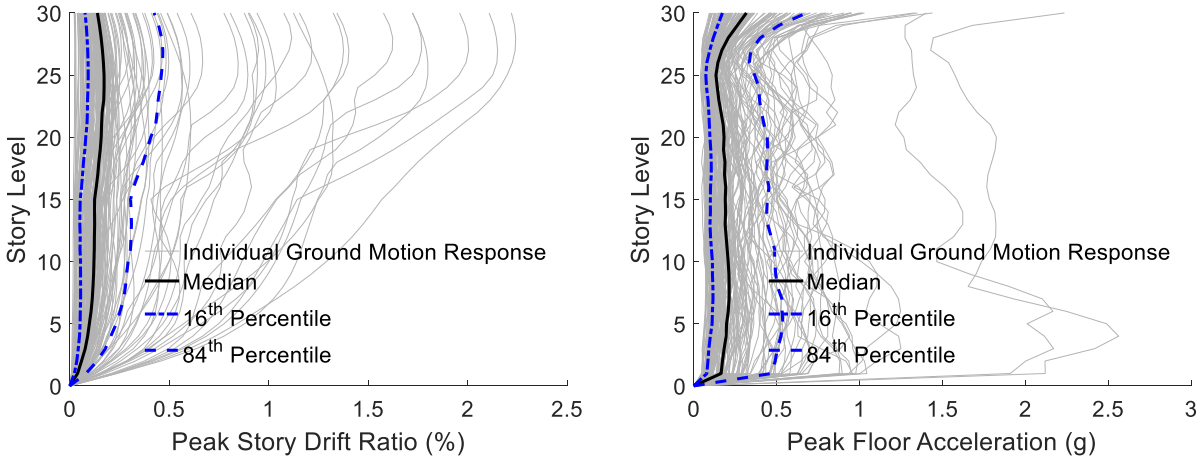
The full profile of PSDR and PFA for the four buildings are shown in Figure 4.4 through Figure 4.7 and coefficients of variation (COV) of the response demands are provided in Table 4.4. For buildings TB-1 (Figure 4.4) and TB-2 (Figure 4.5), the median PSDR ranges from 0.2% to 0.3%, while individual PSDR profiles exceed 2.5% for extreme cases. Figure 4.4 shows that there is a variation in the response profile (along the height) with the demand level. For example, the PSDR profile is relatively uniform at demand levels near the median. Whereas, much higher drift demands are observed in the upper stories at the 84th percentile demand levels. At these higher demand levels, yielding in the 1st story core wall steel reinforcement occurs. This leads to a cantilevered deformed shape with PSDRs increasing almost linearly along the building height and drops off slightly at the uppermost story because of smaller rotations in the core walls.

As described earlier, the core wall thickness in buildings TB-2 and TB-3 changes at stories 16 and 26, respectively, which is reflected in their PSDR profile (Figure 4.5a and Figure 4.6a), especially at higher demand levels. However, as observed in Figure 4.5b and Figure 4.6b, the PFA profile is not influenced by changes in the wall thickness. Figure 4.4a through Figure 4.7a show that the inter-site dispersion of PSDR is generally higher in the upper stories. For example, the average PSDR COV for the four buildings (Table 4.4) is 1.14, 1.24 and 1.29 in the quarter-height, mid-height and the uppermost story, respectively. The reverse trend is observed for PFA, where the average COV for the four buildings ranges from 0.85 at quarter-height to 0.68 at the uppermost story. Comparing Figure 4.6a and Figure 4.7a, the highest PSDR demand between the 15th and 35th stories in TB-3 is about 25% to 35% higher than that of TB-4. This is not surprising given the stronger and stiffer dual LFRS in TB-4 relative to the core-wall only LFRS in TB-3.



(a) (b)

Figure 4.4 (a) PSDR and (b) PFA distribution for TB-1 subjected to 152 ground motions from the Northridge earthquake



(a) (b)

Figure 4.5 (a) PSDR and (b) PFA distribution for TB-2 subjected to 152 ground motions from the Northridge earthquake

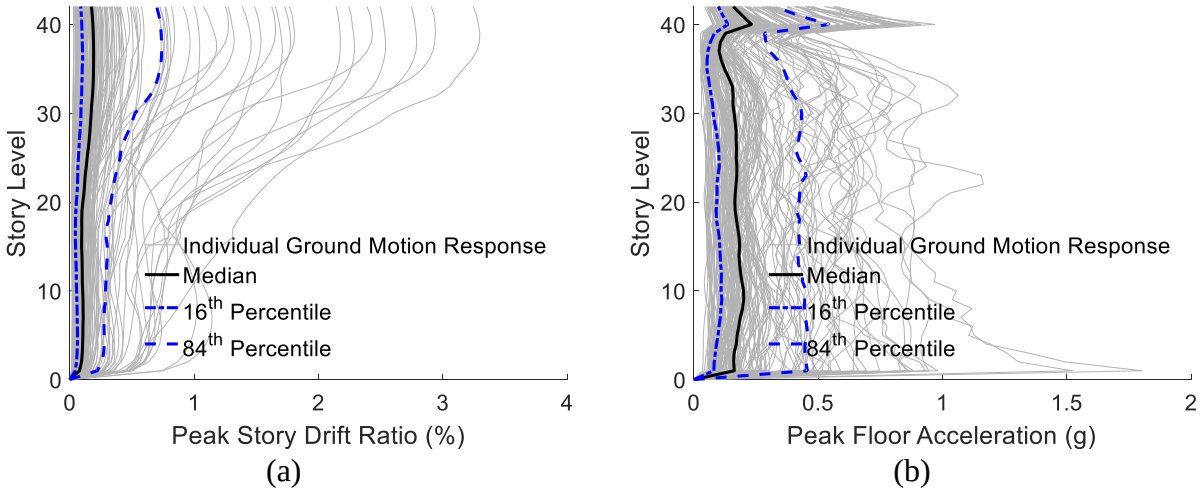


Figure 4.6 (a) PSDR and (b) PFA distribution for TB-3 subjected to 152 ground motions from the Northridge earthquake

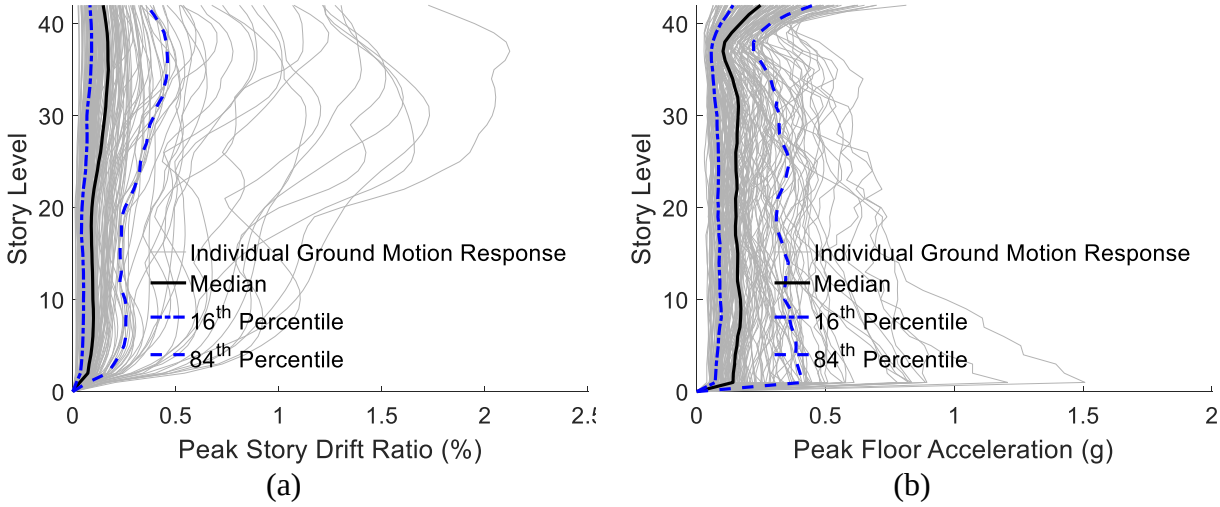


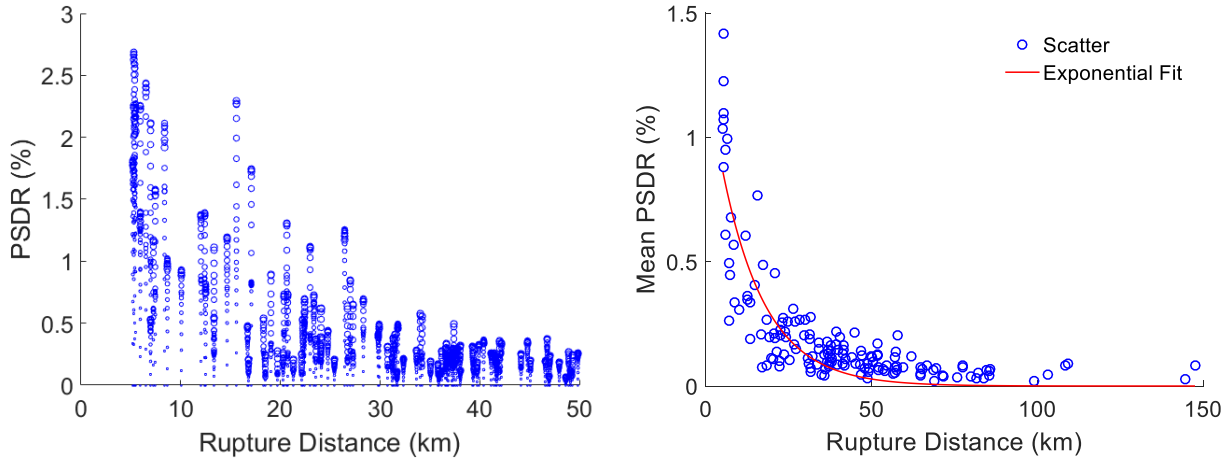
Figure 4.7 (a) PSDR and (b) PFA distribution for TB-4 subjected to 152 ground motions of Northridge earthquake

Table 4.4 Summary of coefficients of variation (COV) along the building height

Building ID	Quarter-height		Mid-height		Roof	
	PFA	PSDR	PFA	PSDR	PFA	PSDR
TB-1	0.98	1.12	0.80	1.23	0.76	1.12

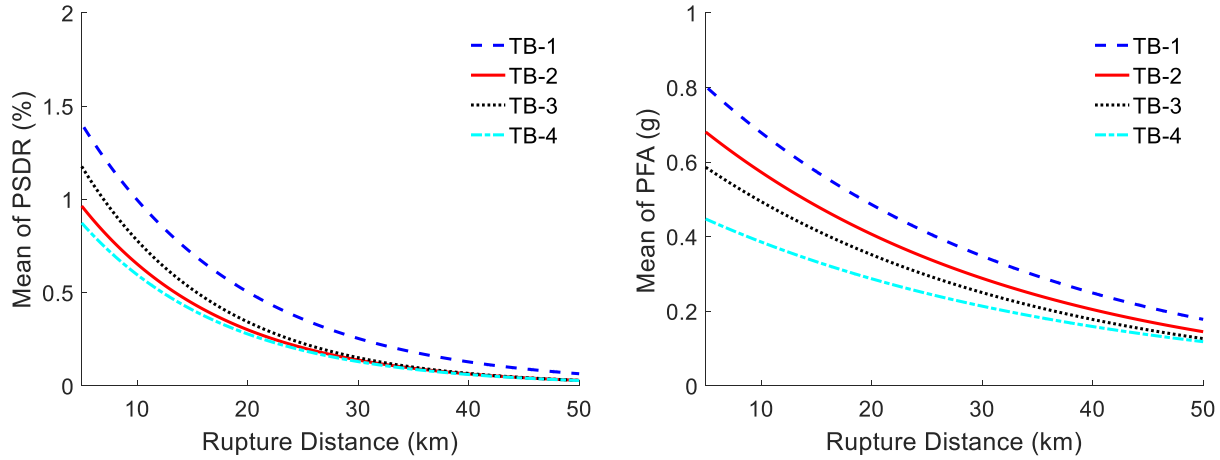
TB-2	0.95	1.10	0.88	1.25	0.77	1.42
TB-3	0.74	1.14	0.78	1.27	0.64	1.47
TB-4	0.74	1.21	0.70	1.35	0.54	1.13

Figure 4.8a shows the PSDR distribution along the height of TB-1 versus the rupture distance (height of floors is indicated by size of scatter). As expected, the demands are generally higher at lower rupture distances. In fact, Figure 4.8b shows that the relationship between the within-building mean PSDR and rupture distance is well represented by an exponential trend line. It can be observed in Figure 4.9a that the mean PSDRs are comparable at higher rupture distances. However, at sites nearer the epicenter, the demands in TB-1 and TB-3 are generally higher, suggesting that the effect of structural dissimilarity varies over the spatial domain. The PFA trendline shown in Figure 4.9b suggests that TB-4 has the lowest PFA demands over the spatial domain (the same is true for PSDR). In summary, the within- and across-building spatial variation of seismic demands are highly nonlinear due to the combined effect of the building configuration and structural properties and the ground motion characteristics. Therefore, nonlinear statistical methods are needed to understand the spatial and structural patterns in the response dataset and formulate the multi-building response reconstruction models.



(a) (b)

Figure 4.8 Plot showing (a) PSDR distribution (along height) and (b) mean PSDR (along height) versus the rupture distance for TB-1



(a) (b)

Figure 4.9 Trendlines for (a) mean PSDR and (b) mean PFA over height with respect to rupture distance

4.5 Description of Machine Learning Models

4.5.1 Statistical Model Design

The primary objective of this study is to develop a machine learning methodology for performing scenario-based reconstruction of full profile seismic response demands for a cluster of buildings (a group of co-regionally located buildings affected by a single seismic event), a subset of which are instrumented in limited locations. The complete (training and testing) dataset of dependent variables includes the peak response demand measurements (PSDR or PFA) in the structures under consideration located at the sites of the recorded ground motions. The number of data points corresponding to a single building/site is equal to (for two horizontal translational directions) the number of stories (for PSDR) or floor levels (for PFA).

The predictors or features used to construct the learning models are carefully selected to represent the factors that are known to affect the spatial distribution of response demands within

and across buildings. The seismic response demand within a building largely depends on the structural characteristics and location of the demand measurements within that building, as well as the properties of the ground motion at the site of interest. There are many structural and ground motion characteristics that can potentially be used as features for the machine learning model. For the current study, the features are chosen primarily because they are readily available and to avoid high levels of correlation among those that are selected. The effectiveness of the selected features is demonstrated later in this paper through a rigorous evaluation of the model performance. The relative importance of the various features is also assessed as part of the overall model appraisal.

The building height (h_{building}), periods of the first three modes (T_1, T_2, T_3) and the wall thickness (t_{wall}) at the location of the response demand are the features used to account for structural dissimilarity. The modal periods are used to capture variations in the dynamic characteristics of the building and the core wall thickness is directly related to the stiffness and response profile of the structure. As observed in the previous section, the response demands within a single building can vary widely along the height, particularly at higher demand levels. Therefore, the location of the response measurement of interest within a building is captured by using the ratio between the floor height of that measurement and the building height (r_{floor}) as a feature. t_{wall}, h_{building} and r_{floor} are also correlated with the actual and measured response profile (along the height of the building). The ground motion and site characteristics (i.e., soil conditions) are represented by the location of the site of interest or the latitude (θ_{latitude}) and longitude ($\theta_{\text{longitude}}$), respectively. Note that the relative spatial location of the buildings as well as the asymmetric properties of ground motions are implicitly considered in the geographical-location-based features. Other site-specific characteristics such as V_{S30}, are known to be highly correlated with geographical

coordinates and therefore not included as features. To summarize, a complete observation of data consists of eight predictors, $X = \left\{\theta_{\text{latitude}}, \theta_{\text{longitude}}, r_{\text{floor}}, h_{\text{building}}, T_1, T_2, T_3, t_{\text{wall}}\right\}$ and the response variable of interest, $Y = \delta_{\text{EDP}}$. A typical statistical regression model can be formulated as a minimization problem that is based on the empirical loss function shown in Equation (4-1).

$$\arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \phi(y_i,\, f(x_i;\theta)) + \lambda\Omega(\theta) \tag{4-1}$$

Where $\phi(.,.)$ is a loss function or error measure between the true response demand y_i and the predicted response demand $f(x_i)$ for observation i. f is a general function term denoting the statistical model formulated by each training observation x_i (the predictor vector of observation i with dimension $1\times p$) which is defined by its model parameters θ. p is the total number of feature variables in each observation ($p = 8$). n is the total number of observations in the training data. A penalty, $\Omega(\theta)$, is added to the empirical loss function to avoid overfitting. The regularization parameter, λ, is used to control the size of that penalty. The goal of model training is to yield a minimum cost defined by Equation (4-1) over f (different statistical models) and its corresponding parameters θ. Details of each applied model are described in the following section.

4.5.2 Ordinary Least Squares

Ordinary Least Squares (OLS) is commonly used in engineering and the physical and social sciences. It is a linear regression method that predicts dependent variables using a series of features. It is included in the current study to provide a point of comparison for the performance of the proposed kernel-based machine learning methods. The model is expressed in matrix form as shown in Equation (4-2) with the predictor X being a matrix of dimension $n\times p$. Y is an $n\times 1$ vector of dependent variables (structural response demands) and ε is an $n\times 1$ vector of residuals, which

represent the difference between the observed and predicted values of Y.

$$Y = \beta X + \varepsilon \tag{4-2}$$

β is a $p \times 1$ vector of the predictor coefficients that is computed using the closed form solution in Equation (4-3) (OLS estimator), which is derived by minimizing the residual sum of squares (RSS): $\min RSS = (Y - X\beta)^T (Y - X\beta)$, which is a matrix form of the objective function described in Equation (4-1). There is no penalty term in the OLS model.

$$\hat{\beta}_{\text{OLS}} = \left(X^T X\right)^{-1} X^T Y \tag{4-3}$$

Once the OLS estimator, $\hat{\beta}_{\text{OLS}}$, is obtained using training data, it can be used to predict building responses at testing data locations by Equation (4-4).

$$f(x_i) = \hat{\beta}_{\text{OLS}}{}^T x_i \tag{4-4}$$

4.5.3 Kernel Ridge Regression

Kernel Ridge Regression (KRR) is an extension of the ridge regression method which incorporates regularization in the OLS model [129,130]. KRR further extends ridge regression by incorporating a nonlinear dissimilarity measure between the observed data and is therefore more suitable for predicting inelastic building response demands [38]. The formulation of KRR in accordance with Equation (4-1) can be expressed in Equation (4-5).

$$\min \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \lambda |f|^2 \tag{4-5}$$

In Equation (4-5), a least sum square is used as the loss function. A regularization constant, λ, is applied to control the model complexity and avoid overfitting. At this point, the concept of a Kernel needs to be introduced to define the prediction model f in Equation (4-6):

$$f(x_i) = \sum_{i=1}^{n} \alpha_i K(x, x_i) \tag{4-6}$$

where α_i is a weight parameter associated with observation i that is to be determined. $K(x, x_i)$ is defined as a Kernel between the observations x and x_i from the training and testing data, respectively, which can also be interpreted as a similarity function between x and x_i. In other words, the Kernel function measures the difference between the two input observations depending on the definition of dissimilarity. Numerous alternative Kernel functions have been developed and used in prior studies and the prediction model performance is typically used to select an appropriate function. Equation (4-7) is a Gaussian Kernel in which σ is a model parameter referred to as the bandwidth. x_i and x_j are the predictor vectors corresponding to the two observations i and j.

$$k(x_i, x_j) = e^{-\left|\frac{x_i - x_j}{\sigma}\right|^2} \tag{4-7}$$

A higher value of σ results in almost linear behavior and vice versa. Also, a Gaussian Kernel generated by lower σ is more sensitive to outliers and boundary data. As shown in Figure 4.10, the Kernel function attempts to capture the similarity between two observations ($x^* = 1$ and x) based on the difference between them. In the KRR model, the dissimilarity between different observations is captured by the chosen Kernel and model parameter being tuned for the best prediction performance. In this paper, Gaussian Kernel (Figure 4.10) is selected because of its positive definite characteristics and better predictive performance.

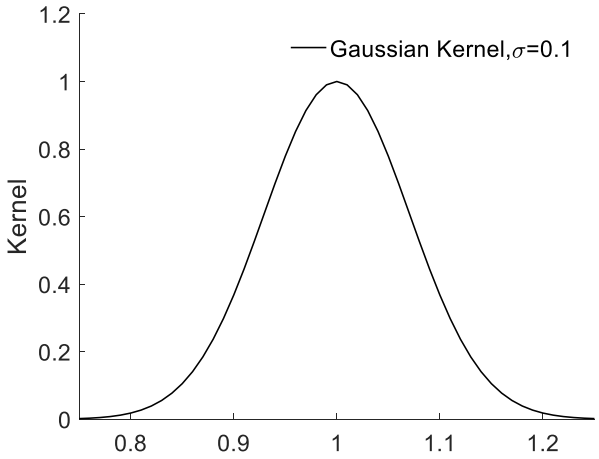


Figure 4.10 Visualization of Gaussian kernel

By implementing Equation (4-6) into Equation (4-5), the latter becomes Equation (4-8), which can be converted to the matrix form in Equation (4-9).

$$\min \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n} \alpha_i k(x_i, x_j) \right)^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \tag{4-8}$$

$$\min \frac{1}{2} \left| Y - K\alpha \right|^2 + \lambda \alpha^T K \alpha \tag{4-9}$$

Y is a vector of all response variables (same as in OLS). K is a matrix of the Kernels between each possible training data pair defined in Equation (4-10). α is an $n \times 1$ weighting vector and the $\frac{1}{2}$ is applied for mathematical convenience.

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix} \tag{4-10}$$

Equation (4-9) can be solved analytically by taking the first derivative with respect to α, which yields Equation (4-11).

$$\alpha = (K + \lambda I_n)^{-1} Y \tag{4-11}$$

Where I_n is an identity matrix of dimension $n \times n$. For a given test dataset of predictors, x_{test}, the dependent variables (structural response demands) can be predicted using Equation (4-12).

$$\hat{f}(x_{test}) = \sum_{i=1}^{n} \alpha_i K(x_i, x_{test}) \tag{4-12}$$

Another more intuitive interpretation of KRR is that the original data space x is transferred into a higher-dimension feature space by $x \to \phi(x)$ where the Kernel is an inner product of the feature representation of x ($K(x,x) = \phi(x)^T \phi(x)$). This nonlinear transformation produces a higher or infinite dimension feature space ($dim(x) \ll dim(\phi(x))$) which increases the sparsity and complexity of the data structure and therefore increases the resolution of the prediction model (better accuracy). Often, the transfer process $x \to \phi(x)$ is very difficult or even impossible. However, KRR does not explicitly require calculation of $\phi(x)$. Instead, it incorporates the Kernel, or the inner product of the feature space, which is usually much easier to obtain. Thus, KRR is computationally effective in transferring data into nonlinear feature space and has a generally better prediction accuracy for nonlinear data containing complex high dimensional patterns.

4.5.4 Kernel Support Vector Regression

Another supervised machine learning method that utilizes the Kernel concept is Kernel Support Vector Regression (KSVR). It is a modified version of the regular Support Vector Regression (SVR) algorithm, which was first developed by Vapnik V. [131] and is widely used in the field of Artificial Intelligence. The derivation of SVR starts with a linear formulation $f(x) = \beta X + b$ where β and b are both $p \times 1$ model parameters. Unlike OLS, which treats b as error or residual term and formulates the cost function to minimize $b^T b$ (RSS), SVR minimizes

the norm value $\frac{1}{2}\beta^T\beta$ with a constraint that the absolute prediction residual of each data point in the training set is within the bound of ε as shown in Equation (4-13) to ensure the maximum distance between decision boundaries. The fundamental difference between KRR and SVR is that the former minimizes prediction error and regularizes model complexity by adding a penalty term to the empirical loss function, while, for the latter, model complexity is regularized and error-minimization is achieved by placing constraints on the loss function.

$$\min \frac{1}{2}\beta^T\beta \tag{4-13}$$

subject to:

$$\forall i : \left| y_i - (x_i'\beta - b) \right| \le \varepsilon$$

The above minimization is a convex optimization problem which may not yield a solution that satisfies the constraints. Thus, two slack variables, ξ_n and ξ_n^*, are introduced to soften the constraints. Equation (4-13) becomes Equation (4-14), which is also called the primal form of SVR [131].

$$\min \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n}(\xi_n + \xi_n^*) \tag{4-14}$$

subject to:

$$\forall i : y_i - (x_i'\beta - b) \le \varepsilon + \xi_n$$

$$\forall i : (x_i'\beta - b) - y_i \le \varepsilon + \xi_n^*$$

$$\forall i : \xi_n \ge 0$$

$$\forall i : \xi_n^* \ge 0$$

The formulation of Equation (4-1) is expressed in Equation (4-14) with additional

constraints where C is a capacity constant, which is the parameter that gives flexibility to model fitting. In other words, it adjusts how well the model fits the training data. Unlike KRR, a closed form solution for the primal form of SVR shown in Equation (4-14) is not possible. However, a computationally tractable formulation, or dual form of SVR, exists and is used to solve the problem. The dual form with kernel implementation can be obtained by applying a Lagrange multiplier to each of the constraints and replacing the inner product of data by kernels for a given training dataset in Equation (4-14), the detail of which can be found in [31].

4.5.5 Model Parameter Tuning

The two nonlinear learning models (KRR and KSVR) require a mathematical search algorithm to find an optimal model parameter set θ (e.g., λ used to regularize model complexity and parameter σ used to adjust the Gaussian Kernel in KRR). As noted earlier, $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_{\text{KRR}}$ can be determined explicitly through a closed-form solution. However, other parameters such as the regularization constant λ (KRR) and β (KSVR) are determined by applying numerical optimizers, including ordinary and stochastic gradient decent and gradient boosting techniques [132,133]. Depending on the size of the training data (n and p), different optimizers have their advantages and disadvantages. A synthesis of the alternative optimization algorithms including their advantages and disadvantages is provided in [134]. To avoid inconsistency in comparing models, a grid search method combined with K-fold cross validation is applied in the training process to obtain the optimal θ.

A predetermined vector of possible values is used to generate an array of parameters, with each element representing a unique set of θ. For example, 5 possible values for λ and 5 possible values for σ would give an array of size 5×5. The entire array is then searched to find the optimal parameter set, which produces the least training error in a K-fold cross validation, which involves

randomly separating the training dataset into k even subsets. Statistical models are trained using subsets 1 through all except the k^{th} subset, which is used for testing. Model parameters are selected from the case producing the least mean squared errors on the k^{th} subset.

4.6 Application and Performance Evaluation of Machine Learning Models

4.6.1 Evaluating Relative Importance of Features (Predictors)

The F-test is applied to determine relative significance of the eight selected features, using the entire dataset of response demands from all four buildings. Two main steps are included in the F-test. First, the correlation between each feature j and the response variable, $\rho_j = \frac{\left(X_j - \bar{X}_j\right)\left(Y - \bar{Y}\right)}{\sqrt{\mathrm{var}\left(X_j\right)\mathrm{var}\left(Y\right)}}$, is calculated, which is then converted to an F-score using Equation (4-15).

$$F - score_j = \frac{2\rho_j}{1-2\rho_j}\left(N-1\right) \tag{4-15}$$

Where N is the total number of observations. Table 4.5 summarizes the F-scores derived from the correlation between each feature and the response variable (PFA and PSDR). For both PSDR and PFA, latitude has the highest F-test score and is therefore the most significant feature. Longitude has the 3rd highest F-score for PSDR and is among the lowest F-scores for PFA. The difference in the significance of the latitude and longitude feature is likely a result of the chosen scenario since, as shown in Figure 4.11, most of the recorded stations do not vary significantly along the longitude direction. As such, the overall conclusion is that the rupture distance is a primary consideration when reconstructing seismic responses across multiple buildings. Normalized floor height has much higher significance for PSDR compared to PFA. This is consistent with the more distinct response profile for the former compared to the latter, which was observed in Figure 4.4 through Figure 4.7. The number of stories is a stronger feature for PFA. These two findings are consistent with the previous observation that PFA has less within-building

variation but varies significantly across different sites. Generally, the modal periods are stronger features for PFA since the initial forces are more sensitive to the elastic stiffness. In contrast, the modal periods have less influence on PSDR, which is more related to nonlinear structural deformations. It should be noted that the F-score is a linear measure between the features and response variable and it may not reveal inherent nonlinear patterns within the data.

Table 4.5 Summary of F-scores

F-score	Longitude	Latitude	Normalized Floor Height[1]	No. of Stories	1st Mode Period	2nd Mode Period	3rd Mode Period	Wall Thickness
PSDR	387	1726	1080	204	34	157	266	178
PFA	88	2073	61	809	446	763	599	327

[1]Normalized by height of building

4.6.2 Model Evaluation

Three different machine learning methods are applied to the dataset. Their performances are comparatively assessed for each scenario dataset using non-replacement Bootstrap (different from the one used for parameter tuning), which provides a statistical distribution of the prediction error as the measure of model accuracy and stability. For each scenario, a designated or randomly generated building type (from the four prototype buildings) is assigned to each site. All of the available seismic response data at each story/floor and at all 152 sites of the recorded Northridge earthquake ground motions, are used to formulate a data pool. Non-replacement Bootstrap is performed by randomly separating the complete dataset into N training and testing subsets. For each randomized partition, the training data subset is used to construct the corresponding statistical model and predict the seismic response demands at the testing dataset locations. An error measure, the median absolute relative deviation (MARD), is calculated to represent the model accuracy and

stability in Equation (4-16).

$$MARD = Median\left(\left|\frac{f(x_i) - y_i}{y_i}\right|\right) \tag{4-16}$$

Where y_i is true response demand, $f(x_i)$ is the predicted response demand, $i = 1,2,\cdots,n_t$ and n_t is number of data points in the testing subset. A probabilistic distribution of the prediction error is obtained by computing the MARD for multiple cases of randomly generated testing and training data subsets [28,135].

4.6.3 Baseline Scenario Dataset

A baseline scenario dataset is assembled using building responses at all 152 sites with the training dataset randomly generated from the stories/floors within the building located at each site. This dataset is generated based on the assumption that (a) all buildings within a target cluster are instrumented and (b) the location of the sensors within each building is randomized. In other words, the training data for this baseline scenario is randomly selected from all stories and buildings at the 152 locations. Other scenario datasets (e.g., training data is at 1st floor, mid-height and roof of each building at each site, a fraction of buildings are instrumented) are evaluated in Section 4.6.4 and 4.6.5.

Two variations of the baseline scenario are considered, which differ based on the building distribution across the considered sites. The self-prediction case assumes the same building is located at each site. This case is unrealistic in its representation of real building clusters. However, it provides a basis for disaggregating the effect of spatial and structural dissimilarity on model performance. The second case, cross-prediction, randomly assigns one of the four tall buildings to each site (shown in Figure 4.11). The training and testing data subsets are taken to be 30% and 70% of the full dataset, respectively. Later in this section, the effect of the size of the training data subset

(relative to the full dataset) on model performance is evaluated. The 30% data subset is used to train each statistical model with a 4-fold cross validation. The low ratio between training and testing data is used to approximate real world scenarios where only a small portion of tall buildings are instrumented at limited locations.

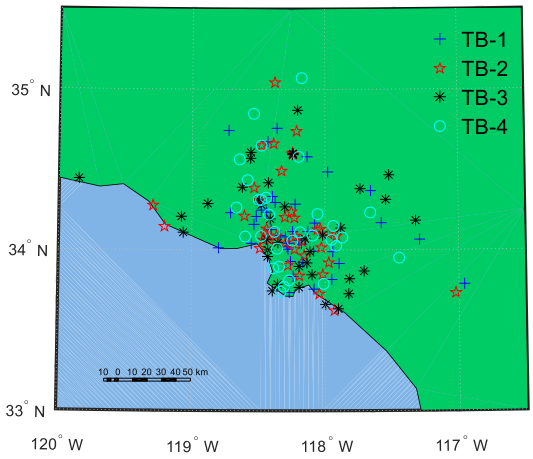


Figure 4.11 Example of randomly distributed building locations

Figure 4.12 shows the full profile PSDR prediction results for individual buildings, TB-1 located at 34.281 N, 118.478 W (Figure 4.12a) and TB-4 (Figure 4.12b) located at 34.296 N, 118.375 W, which correspond to the baseline scenario, cross-prediction case. Shown in each figure are the location of the measured (known) responses, which are included in the training dataset and the true and predicted full profile responses corresponding to the three algorithms. It can be observed that the Kernel based models (KRR and KSVR) are much better at capturing the PSDR profile (within-building pattern and demand levels) compared to OLS. The nonlinear response pattern between the 5th and upper-most stories of TB-1 is accurately captured with the correct curvature direction and demand level. Figure 4.12 shows that the true PSDR profile of the 42-story building has a triple curvature pattern with changes at the 2nd, 12th and 35th story, highlighting the more complex behavior in the taller building. Although the demand levels are not as accurate as

in the TB-1 building, the KRR and KSVR models are able to capture the overall profile. In both buildings, the OLS model does not perform well as it is not able to capture the nonlinear response profiles in tall buildings.

Figure 4.13a and b show PFA prediction results for buildings TB-1 located at 33.790 N, 118.012 W and TB-2 located at 34.070 N, 118.150 W, respectively, from the baseline scenario, cross-prediction case. For TB-1 (Figure 4.13a), the measured responses are dispersed throughout the height of the building and KRR and KSVR accurately capture the highly nonlinear PFA profile. In contrast, for TB-2 (Figure 4.13b), there are no response measurements below the 15th story so the overall performance of the kernel-based models is not as accurate as for TB-1. A similar situation is observed in Figure 4.12b where, above the 32nd story, the PSDR prediction error exceeds 60% because there are no response measurements in this region. This observation highlights the importance of having adequately dispersed sensors in the instrumented building. Recall that for the baseline scenario, the known responses are randomly located along the height of each building. Later in this section, alternative scenarios are explored, where the locations of response measurements are predetermined to improve the overall model performance.

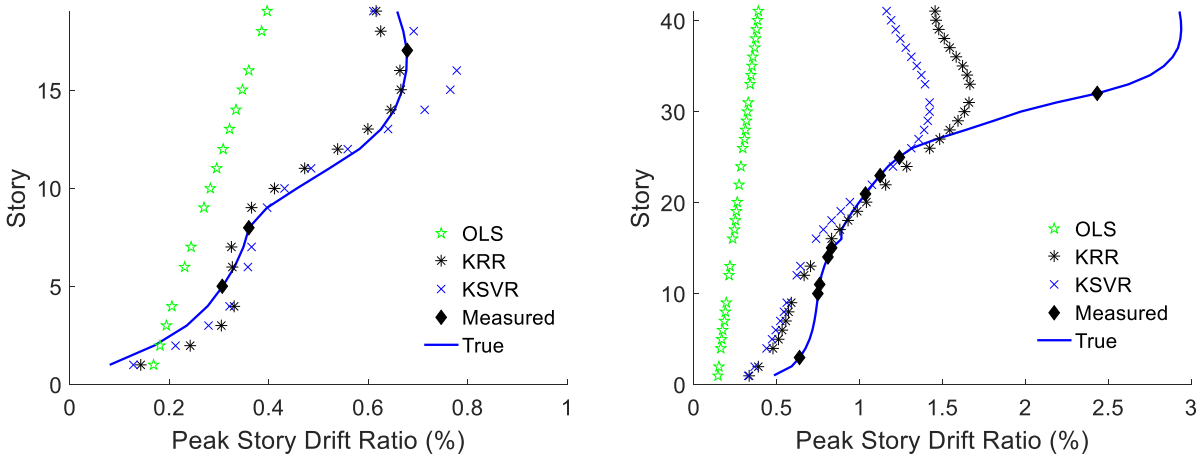


Figure 4.12 PSDR prediction results for (a) TB-1 located at 33.790 N, 118.012 W and (b) TB-4 located at 34.070 N, 118.150 W for the baseline scenario, cross-prediction case

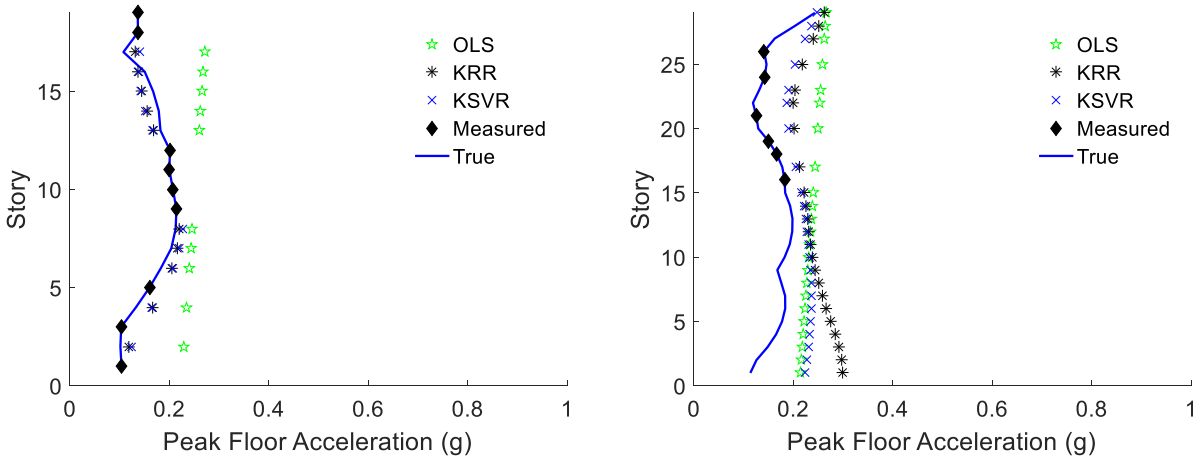
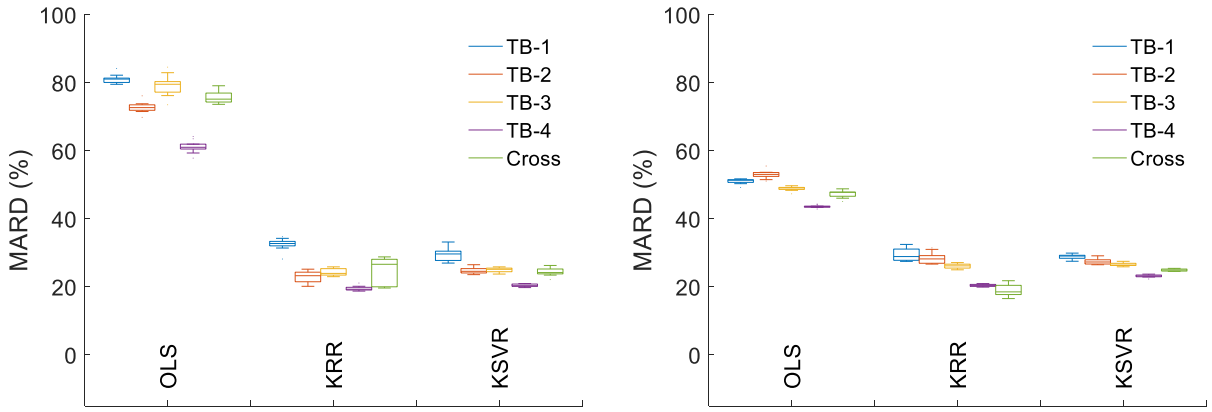


Figure 4.13 PFA prediction results for (a) TB-1 located at 33.790 N, 118.012 W and (b) TB-4 located at 34.070 N, 118.150 W for the baseline scenario, cross-prediction case

To better understand the overall prediction performance of each model, the non-replacement Bootstrap procedure is repeated 10 times for the baseline scenario with uniformly distributed and randomly generated buildings at each location (demonstrated in Figure 4.11). 4-fold cross validation is applied to generate a model for each method using the training data. The trained model is then used to predict the two engineering demand parameters of interest [6] in the testing dataset, and the MARD is computed. The self- and cross-prediction results are summarized in Figure 4.14, showing that the KSVR and KRR methods have much better overall prediction accuracy compared to OLS. In Figure 4.14a, the self-prediction MARD for TB-4 is the lowest among the four buildings, 21% using KRR and KSVR. The self-prediction MARD for TB-1 is 33% for KRR, which is the highest of the four buildings. The self-prediction for TB-2 is the most unstable case of the four buildings as it has the highest difference between the minimum (21%) and maximum (24%) MARD for KRR. Overall, KSVR has the best performance for the PSDR

self-prediction. The cross-prediction model outperforms the self-prediction model in some cases, indicating that the training data for the former model contains a more diverse set of responses.

The PSDR cross-prediction result shown in Figure 4.14a is generated by randomly assigning buildings at each site and generating training data at random locations. Again, KRR and KSVR rank the highest in terms of accuracy at 26.7% and 25.3% MARD, respectively. However, KRR has a much higher MARD dispersion, which ranges from 19.7% at the 25th percentile to 28.9% at the 75th percentile. It can therefore be inferred that KSVR is the best model for both self- and cross prediction. Similar results are observed for the PFA prediction, which are shown in Figure 4.14b, where the overall MARD is similar to the PSDR model. However, it should be noted that MARD is an average measure of model performance. For interpolation over the height of an individual building, the model performance might vary depending on the ground motion intensity and building type.

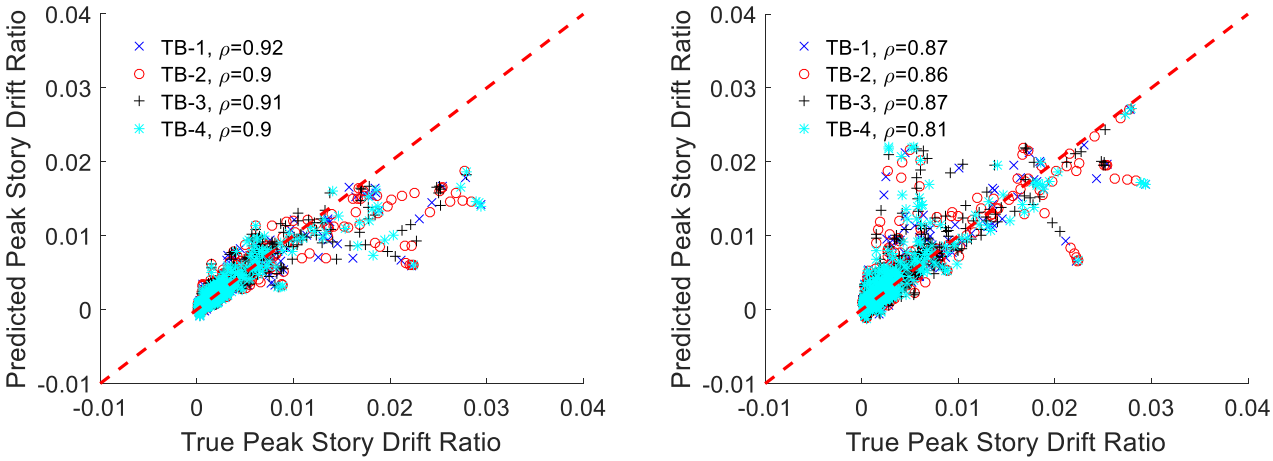


(a) (b)

Figure 4.14 MARD distribution for (a) PSDR and (b) PFA prediction for the baseline scenario

Figure 4.15a shows a plot of the KSVR-predicted versus true PSDR for the baseline scenario, cross-prediction case, corresponding to a single non-replacement bootstrap procedure.

Correlation coefficients between the true and predicted PSDR are approximately 0.90 for all four building types, which indicates good overall performance. The similarity in the correlation across building types indicates that the model performance is not biased based on building type. However, it is observed that the model does not perform as well at PSDRs greater than 1%. For example, where the data points are separated into two bins, PSDRs <= 1% and PSDRs > 1%, the correlation coefficients are 0.90 and 0.41, respectively. This variation in model performance at different demand levels is due to the smaller number of training data points at higher demand levels, which corresponds to buildings closer to the epicenter. An alternative sampling technique, weighted KSVR (KSVRW), is applied to address this issue by weighting each training response data in proportion to its demand level. Figure 4.15b shows that, although there is an improvement in the prediction accuracy for PSDRs greater than 1%, the overall prediction performance drops as indicated by the 6% reduction in the correlation coefficient for each building. Figure 4.16 shows the same plot for PFA, where KSVRW significantly improves the prediction accuracy for data at higher demands (greater than 1g). However, unlike PSDR, the overall prediction performance of each building increases with the correlation coefficient increasing from approximately 0.8 to 0.9 for all buildings.



(a) (b)

Figure 4.15 PSDR cross prediction result for baseline scenario using (a) KSVR and (b) KSVRW

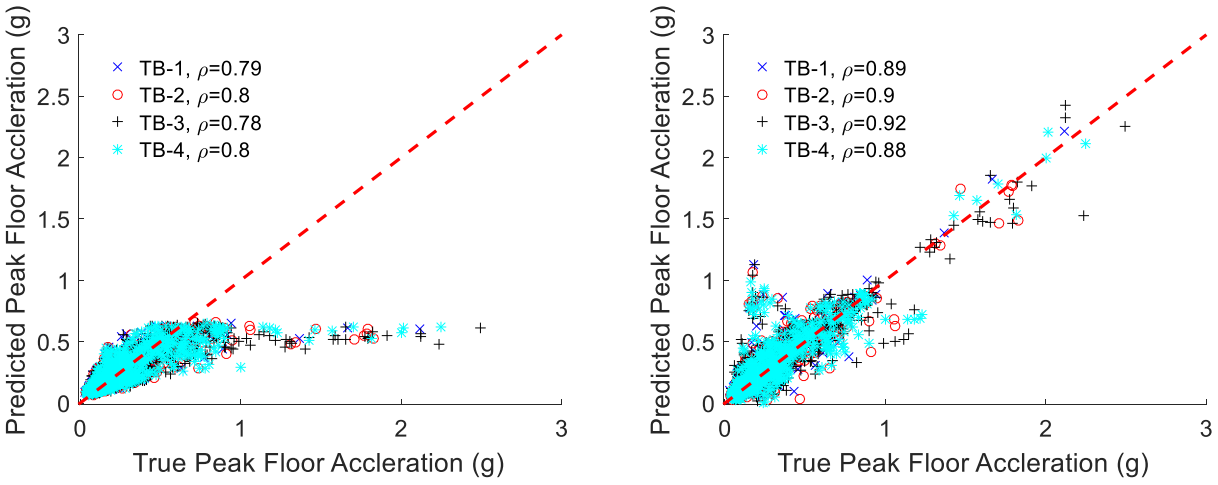
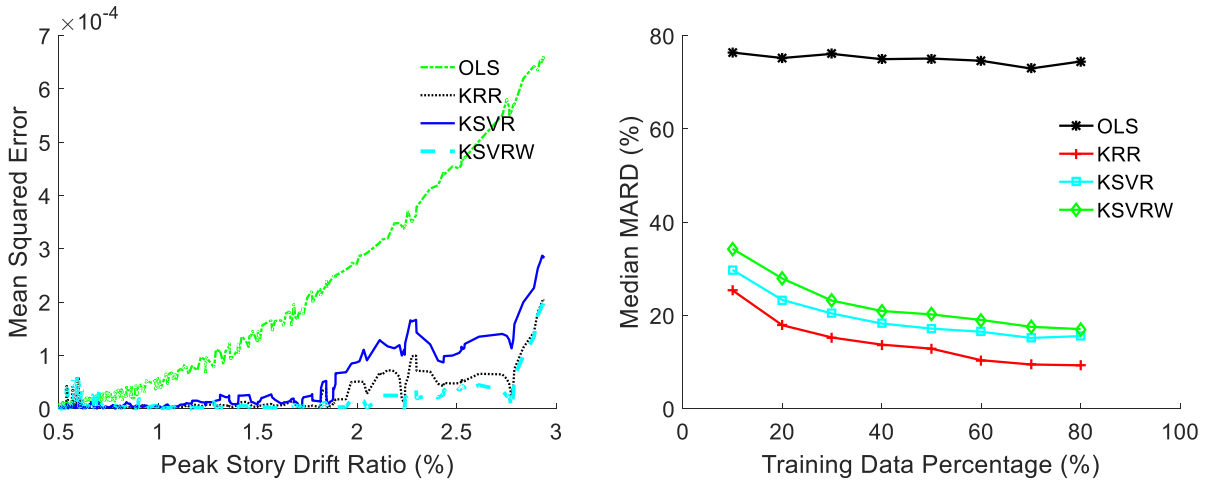


Figure 4.16 PFA cross prediction result for baseline scenario using (a) KSVR and (b) KSVRW

Figure 4.17a shows a plot of the true PSDR versus the mean squared prediction error for each model, which clearly shows that KSVRW has less error than the original KSVR and KRR models for PSDRs over 1.5%. Comparing KSVRW and KSVR, sample weighting is beneficial to reducing prediction error at higher EDP demands. However, as noted earlier, this also increases the overall prediction error. Ultimately, sample weighting seeks to shift the distribution of training data such that it is a better representation of the true data distribution. For example, the increase in the prediction error across all methods at 3% PSDR is due to the smaller amount of training data used in the model at that level. This observation highlights the need to establish an approximate distribution of training data (locations of instruments within and across buildings). In contrast to the Kernel-Based models, OLS shows an exponential increase in the mean squared error with PSDR, which supports the earlier argument that linear models are not suitable for reconstructing seismic response demands.



(a) (b)

Figure 4.17 (a) Ranked PSDR prediction result for the baseline scenario and (b) effect of the training data percentage on model performance of the baseline scenario

Figure 4.17b shows the effect of the size of the training dataset (relative to the total) on the overall model performance. The results shown in this plot are generated by conducting non-replacement Bootstrap with the training data subset ranging from 10% to 80% of the full dataset, using the same randomly generated building distribution. It shows that the overall prediction accuracy generally improves as the size of the training data subset increases. However, for the two Kernel based methods (KSVR and KRR), the model performance stabilizes when the training data subset is approximately 30% of the full dataset. For example, the MARD corresponding to 80% training data (17% MARD) is only 15% lower than the case where the training data subset is 30% (20% MARD) of the full dataset for the KSVR case.

4.6.4 Effect of Within-Building Sensor Location on Model Performance

As noted earlier, the effect of measurement locations within an instrumented building on the prediction model performance is particularly relevant to real world applications. To provide insight into this issue, the non-replacement Bootstrap procedure is applied and used to predict PSDRs for randomized building locations and two instrumentation cases: 1) three EDP

measurements randomly placed along the building height and 2) three located at the first floor, mid-height and roof. 4-fold cross validation is again used to determine model coefficients and the performance of each model is assessed. Plots of the median MARD versus the percentage of instrumented buildings are shown in Figure 4.18a (case 1) and b (case 2). It is interesting that, for the randomly distributed sensor case, KRR produces a very high median MARD (comparable to OLS), ranging from 70% to 80%, which decreases only slightly as the percentage of instrumented building increases. For the same model, a much higher prediction accuracy (35% and 30% median MARD at 40% and 65% of buildings being instrumented, respectively) is obtained for the designated sensor location (Figure 4.18b). For the randomly located sensor case, the median MARD for the KSVR method is about 30% to 50% lower than that of KRR. Like KRR, the KSVR prediction error decreases only slightly as the percentage of instrumented buildings increases. On the other hand, increasing the number of instrumented buildings largely reduces the prediction error for the designated sensor location case. These observations highlight the importance of sensor placement in multi-building seismic response demand reconstruction. However, it is important to note that the observed superior performance of the designated sensor location case is based on a summary statistic (MARD), i.e., it measures the overall performance considering all buildings. There may be stories within individual buildings where the randomized sensor location case performs better.

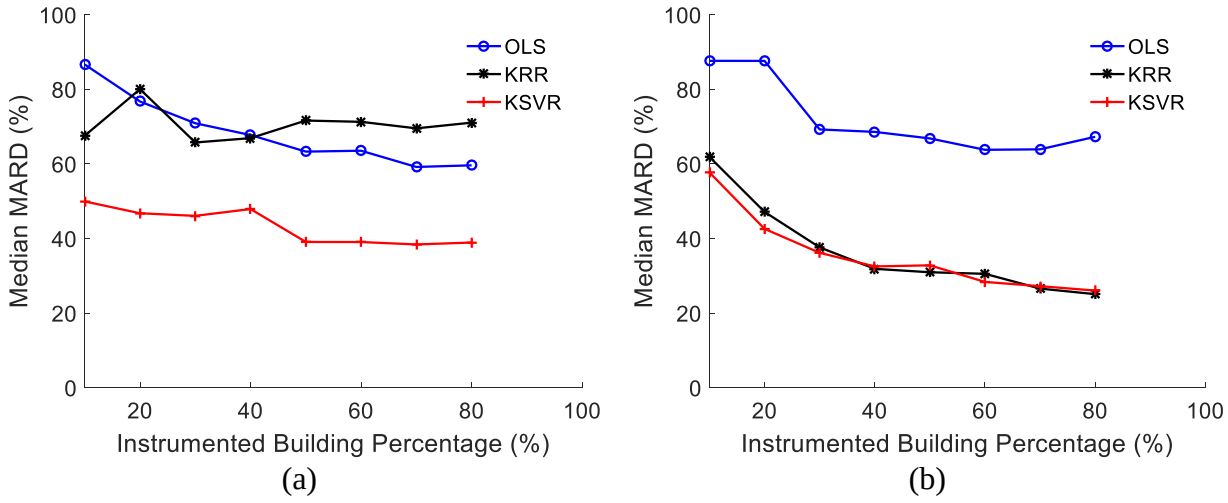
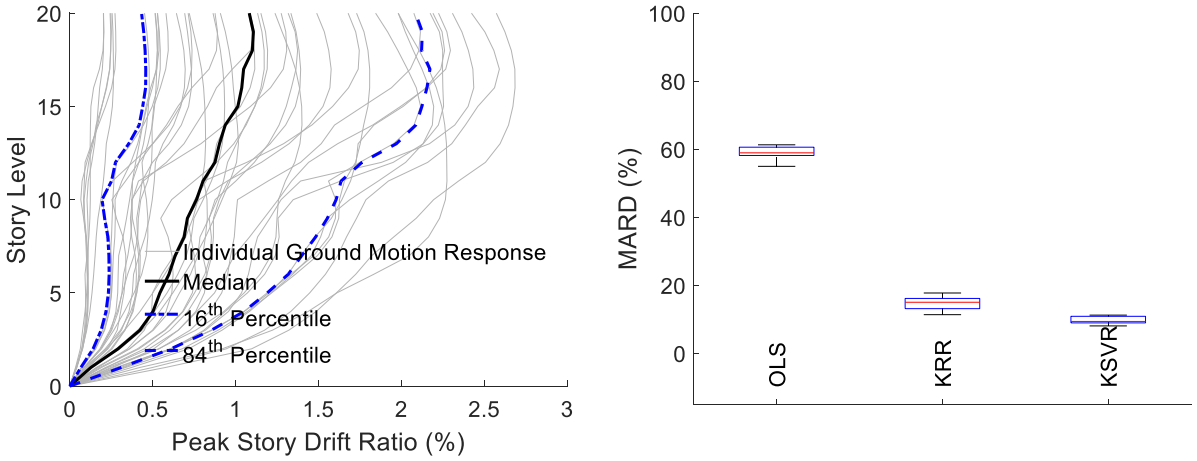


Figure 4.18 PSDR prediction result for (a) randomized and (b) designated sensor locations

4.6.5 Effect of Response Demand Level of Model Performance

As observed in Figure 4.4 through Figure 4.7, the median PSDR for the baseline dataset is relatively low (less than 0.5%), which indicates that most buildings of the 152 respond within the elastic range. To demonstrate their prediction capability for inelastic demands, the kernel-based models are applied to a subset of the baseline dataset that includes 40 buildings located within 23km from the epicenter. The PSDR profile along the height of the TB-1 building is shown in Figure 4.19a. Compared to Figure 4.4, the median PSDR increases to approximately 1% in the upper stories. Prior studies have shown that nonlinear coupling beam rotation and shear wall deformation start to occur at approximately 0.7% PSDR [41,127,127]. This, coupled with the demand levels observed in Figure 4.19a, indicate that most of the buildings in the near-epicenter data subset are subjected to inelastic response demands. Based on the median MARD values from the cross-prediction, the performance of the kernel-based models for the "near-epicenter" data subset is found to be on par with or better than the complete dataset. For example, the median MARD for PSDR is found to be 10% (Figure 4.20b) when the KSVR model is used to perform cross-prediction on the near-epicenter dataset, which is 60% lower compared to the baseline

dataset. Similar improvements are observed for PFA cross-predictions. Overall, the higher demands in the near-epicenter dataset results in lower variance in the within- and across-building spatial pattern and reduced uncertainty in the prediction model. It is important to note the distinction between the variance in the demand level (which is higher for the near-epicenter dataset) and the spatial pattern, where the latter refers to variations in the response profile (e.g. along the height of the building). Figure 4.20 shows the effect of the size of the training dataset on the performance of the kernel-based models, when applied to the near-epicenter dataset. By comparing Figure 4.20 and Figure 4.17b, it is observed that the performance is on par with predictions for the baseline dataset when the fraction of instrumented floors is less than 30% and better than the baseline dataset predictions when the fraction of instrumented floors exceeds 30% of the full dataset.



(a) (b)

Figure 4.19 (a) PSDR distribution for TB-1 subjected to 40 near epicenter ground motions from the Northridge earthquake and (b) MARD distribution for PSDR cross-prediction based on high demand data subset

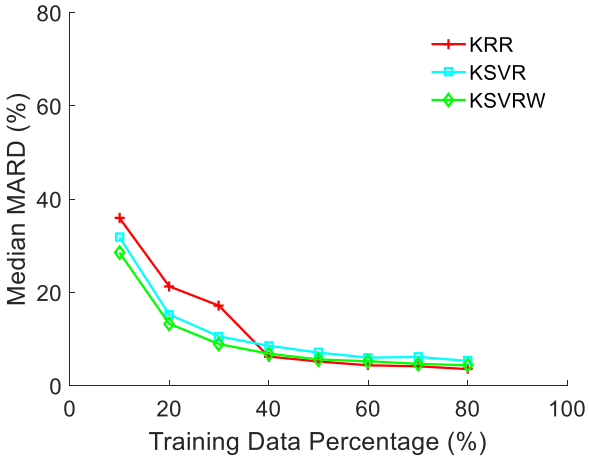


Figure 4.20 Effect of training data percentage on model performance for the high demand data subset scenario

4.7 Summary

A building portfolio seismic response reconstruction methodology is introduced, which uses available demands measured at limited locations within a subset of buildings. The proposed framework is demonstrated using peak floor acceleration (PFA) and peak story drift ratios (PSDR) generated from structural response simulations. Nonlinear structural models of four tall buildings are developed in Perform3D and subjected to a set of spatially explicit ground motions recorded during the 1994 Northridge earthquake. Distinct patterns in the seismic response demands were found within individual buildings at a single site and across multiple buildings located at different sites. Two Kernel-based machine learning methods, Kernel Ridge Regression (KRR) and Kernel Support Vector Regression (KSVR), were applied to capture the underlying patterns and reconstruct response demands at locations within buildings where measurements are unavailable. The ordinary least squares (OLS) method was also applied and used to benchmark the superiority of the kernel-based methods.

The features used to develop the prediction model include the building height, first three modal periods of each building, the wall thickness and the location within the building defined by

the floor height normalized by the building height. The latitude and longitude of the site are also included as features related to the location and site properties. Results from an F-test revealed that these geo-spatial variables are the most important of the eight features for predicting both PFA and PSDR. The first three modal periods and building height were found to be highly correlated with PFA while the normalized floor height has a stronger correlation with PSDR. These findings are generally consistent with the much higher "within-building" dispersion and nonlinear response profiles of PSDRs compared to PFAs.

The performance of the multibuilding seismic response reconstruction models was evaluated using several scenarios. For the baseline scenario, buildings are uniquely (self-prediction) and randomly (cross-prediction) placed at each site. The three statistical models, OLS, KRR and KSVR are implemented using 30% of the entire dataset as training data, to predict the remaining 70% testing data. The effect of the size of the training data was also evaluated by comparing the error measure across training dataset percentages ranging from 10% to 80% of the full dataset. A second scenario is used to compare the model performance for the cases where the measurement location within each building is predetermined and randomly placed. The effect of the fraction of instrumented buildings on model performance was also evaluated. Non-replacement Bootstrap was used to randomly separate the training and testing data and 4-fold cross validation was used for optimal parameter selection. The result shows that both Kernel-based methods are better in terms of the stability and accuracy of reconstructed responses for self- and cross- predictions under randomly distributed measurement locations within buildings. There was no bias in the accuracy of the predictions across the different building types and the size of the training data improved the model performance. Also, the performance of the Kernel-based models was better when the location of the measured responses was predetermined (as opposed to being randomized). For a

scenario where 40% of the buildings had measured responses at the first floor, mid-height and roof, the predicted demands from the Kernel-based methods were, on average, within 30% of the measured demands. A subset of the original dataset that only includes buildings within 23km of the epicenter is used to evaluate the effect of higher levels of inelastic demands on the prediction capability of the kernel-based models. The result shows equal or better performance compared to the full dataset.

This study used seismic demands from structural response simulations to demonstrate and evaluate the proposed multi-building seismic response reconstruction methodology. While using building response measurements from a real earthquake would have been ideal, the unavailability of full-profile recorded responses for a cluster of buildings makes this approach unfeasible. It is also worth noting that both simulated and measured responses are expected to differ from the real responses. The discrepancies between the simulated and real responses arise because of limitations in the state-of-the-art in nonlinear structural modeling. Errors can also be introduced in measured responses due to imperfections in the installation, recording and data-retrieval process. These limitations can be addressed by introducing an error or residual term. In other words, the reconstructed response can be regarded as a central tendency value and the error term would represent the dispersion.

5. A Generalized Cross-building Engineering Demand Parameter Reconstruction Model

5.1 Introduction

Seismic hazard is a primary threat to urbanized regions on US western coast and Circum-Pacific seismic belt such as Japan, New Zealand and Mexico. Given that an earthquake is inevitable, it is beneficial to develop rapid damage assessment for reducing consequential building damages and monetary losses. There are two major ways to assess building damage after a seismic event, either by human inspection, which is very time consuming and human labor intensive, or by retrieving building damage states that are related to critical engineering demand parameters (EDP) through Strucutral Health Monitoring (SHM) remote sensing systems; however, only a limited number of buildings are instrumented with such remote sensing and data acquisition systems. A typical situation is that a very small number of buildings that are lifeline related (e.g., fire station, hospitals) and tall (e.g., buildings with more than 10-stories or more than 6-stories and an aggregate floor area greater than 5574m^2 are required to install SHM system in the City of Los Angeles) are instrumented to allow retrieving EDPs immediately following after a seismic event. Therefore, one major challenge of SHM is to develop a data-driven rapid prediction model based on these available recordings from the instrumented buildings to infer EDPs in the majority of buildings, which are un-instrumented.

Sun, et, al. [28] first developed a data-driven interpolation model for maximum response including peak floor acceleration (PFA) and peak story drift ratio (PSDR) along the building height using a kriging algorithm across different buildings and different sites subjected to a single earthquake event. Nonlinear structural responses extracted from five different concrete moment frame models [100,103] under three different earthquake scenarios were used to demonstrate the

method. As a result of within event interpolation, the proposed framework is limited to cases where, right after the occurrence of an event, an interpolation model is trained using the available response data from instrumented buildings and used to infer responses in un-instrumented buildings, all for the same event. In other words, event characteristics are not considered.

On the other hand, modern seismic risk assessment of buildings often involves probabilistic estimation of seismic building responses to assess building damage states through a data-driven fragility curve [106]. The data to calibrate fragility curves are retrieved by running Nonlinear Response History Analysis (NRHA) for the target building subjected to a suite of ground motions. There are two major challenges with this procedure. First, the probabilistic seismic risk is determined only through the selection process of the applied ground motion suites. Uncertainties that incorporate earthquake sources such as fault type and intensity distance attenuation are only implicitly considered in this selection process without an explicit model on their mechanisms, e.g., intensity decay from source. In addition, common means of ground motion selection is to pick recorded ground motion pairs matching a designated target spectrum by meeting certain criteria, e.g., mean and variance [136]. As a result, uncertainties are only incorporated from the chosen target spectrum, which is either a conditional mean spectrum [137] or uniform hazard spectrum at a particular hazard level. This limits the seismic risk assessment to be only based on certain hazard level, such as Maximum Considered Earthquake (MCE). Incorporating additional hazard levels would require conducting even more NRHA, which is computationally demanding and makes a regional risk assessment for a cluster of buildings challenging, if not impractical. A data-driven prediction model for seismic building responses that takes into consideration on sources of uncertainty, including event and building characteristics as well as intensity attenuation using recorded data across all relevant historical earthquakes would help overcome the above two

challenges.

The seismology community has conducted research to establish event characteristics for peak horizontal ground accelerations (PGA) and velocities (PGV) since 1980s using recordings obtained from ground motion stations. Early representative work is a strong motion attenuation regression model developed by Joyner and Boore [138,139] that considers both earthquake magnitude and site rupture distance for each event and site using data collected in 182 recordings from 23 earthquakes. The approach is based on a 2-stage regression technique that enabled decoupling site and event features. This statistical procedure was commonly used in subsequent studies because it allowed separation of distance or site dependence and magnitude dependence. Brillinger and Preisler [140,141] extended the model to a nonlinear formulation by introducing random effect to each event resulting in explicit separation of between and within earthquake variation using the same dataset. They also introduced the expectation-maximization algorithm (EM-algorithm) for parameter searching typically known as 1-stage regression. Joyner and Boore [142,143] further explored the 1-stage and 2-stage methods and showed that both are unbiased through Monte Carlo simulation. The latter approach is more computationally efficient and addresses a bias issue examined by McLaughlin [144], using the same dataset in [138–141], and later expanded to estimate response spectrum accelerations (PSA) for various damping levels, which is now referred to as ground motion prediction equations (GMPE). A calibration procedure to account for the component of variability associated with the site was also incorporated. Boore and Joyner [145] provided a series of calibrated equations to estimate response spectra for damping levels at 2, 5, 10 and 20 percent and then the site term was modified [146] such that the site-effect was changed from a constant to a continuous function based on time-averaged 30 m shear wave velocity (V_{S30}). Over the years, the GMPE model has been extensively refined through the addition

of new datasets and developed for different regions as presented in the summary work [147], which also pointed out the primary limitations, such as: 1) lack of response spectral data from events below magnitude six; 2) unclear effect of site conditions on short-period motion (less than 0.3 s); 3) distance limitations in dataset. As more datasets become available from PEER NGA project [148] to capture more complex effects from ground motions, Boore and Atkinson [149] presented an improved model with an additional "anelastic" coefficient to account for data trend beyond 80 km as well as a magnitude-dependent "geometrical spreading" term without adding more predictive variables (e.g., such as basin depth). Besides [149], various empirical models in the GMPE category were proposed, e.g., Campbell and Bozorgia [150] and Chiou and Youngs [151]. Using [149] as a demonstration, the GMPE model is often represented as in Equation (5-1).

$$\ln Y = F_M + F_D + F_S + e \tag{5-1}$$

Where F_M, F_D and F_S are magnitude scaling, distance function, and site amplification determined by event magnitude, site rupture distance, V_{S30} and other factors such as fault type. A simplified expression that decouples the total residual e into two components is given by Equation (5-2) [96].

$$\ln(Y_{ij}) = \ln(\overline{Y}_{ij}) + \varepsilon_{ij} + \eta_j \tag{5-2}$$

Where Y_{ij} donates the interested intensity measure, e.g., spectral acceleration at period T_1 at site i from event j; $\overline{Y}_{ij}$ is the predicted median value of Y_{ij}; ε_{ij} is the intraevent residual and η_j is the interevent residual. Jayaram and Baker [96] examined the correlation that existed in GMPE residuals and developed a model for spatial correlation by generating semivariograms for normalized intraevent residual and further constructed experimental semivariograms using empirical data from seven historical earthquakes. Other residual correlation studies, e.g., Goda and

Atkinson [94] and Loth and Baker [97], explored the effect of including different events on generating semivariograms and developed a generalized spatial semivariogram model over the spectrum. These semivariogram models enable prediction of residuals using kriging system as demonstrated in [28]. It should be noted that this notation is different from what is used in this chapter given that the slightly different approaches to decouple event and site related trends from the remaining residual from GMPE.

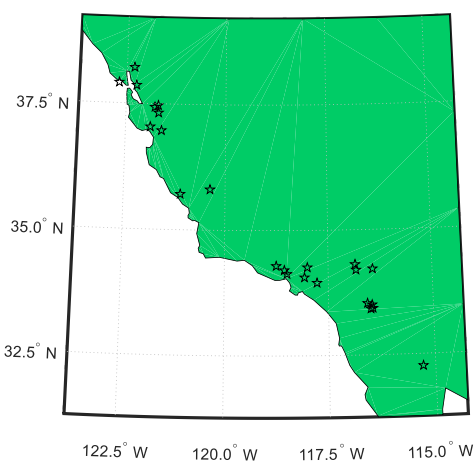
One significant limitation of GMPE is that it estimates IMs such as PGA, PSA at median levels expressed as a continuous functional form to account for correlations from event characteristics and site conditions without inclusion of corelated patterns of structural characteristics of buildings. Although there are approximate methods to estimate EDPs using ground motion intensity measures (IM), such as response spectrum analysis or other empirical models [152], a direct data-driven approach to reconstruct EDPs using actual recordings does not exist, primarily due to the lack of recorded building seismic responses with sufficient structure information. This limitation was overcome in this study by processing all available recorded structure response data from CSMIP [153], to enable the inclusion of correlations from structural dissimilarities and the development of a generalized cross-building EDP reconstruction model. The overall approach uses the same model framework described above for GMPE, since it is naturally suitable to account for the significant variation of data quantities retrieved from different events. Correlations in remaining residuals were evaluated after decoupling the within-event residual (also known as inter-event residual) and the site residual and the kriging method was applied for geo-spatially distributed data interpolation. Recorded response data from buildings were used in this study to establish an empirical relationship between spatial correlated parameters and building responses as opposed to previous work in [28] which is a scenario-based prediction

methodology. A Monte Carlo simulation was proposed to compensate for insufficient data in some considered events. An additional study using simulation data also is included to demonstrate model applicability. The proposed generalized model, not only enables rapid damage assessment using measured of structural responses, it also is an alternative approach for current probabilistic seismic hazard analysis (PSHA) [154] to estimate probability of different damage states of a building, instead of based on median SA at first mode building period (SA_{T1}) from GMPE, but based on median EDPs to account for its location, considered hazards and strucutral characteristics.

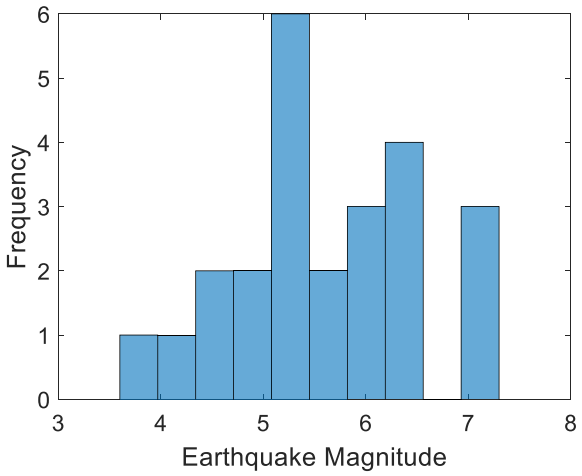
5.2 Description of Data

5.2.1 Data Source

The data were retrieved from the Center for Engineering Strong Motion Data (CESMD) where structural responses recorded under seismic events from multiple accelerometer channels were processed by baseline correction, instrument correction and multiple frequency filters to obtain corrected acceleration and displacement time histories of each channel [153]. The earliest event selected for this study is M6.2 Morgan Hill earthquake occurred on April 24, 1984 and the latest one is an earthquake that occurred in Berkeley in 2018. A total of 24 events with epicenters mostly located around San Francisco and Los Angeles area (Figure 5.1a) were used to demonstrate the approach. The histogram of earthquake magnitude shown in Figure 5.1b follows a normal-like distribution shape centered at 5.5 magnitude, which agrees with assumption in PSHA that individual predictors (magnitude in this case) should follow normal distribution [154]. Among which, three extreme events with magnitude larger than 7 and one event with magnitude less than 4 are included. A complete summary of events selected is shown in Table 5.1. Number of available instrumented buildings for each event ranges from a few to as many as 56, after data filtering to remove unhealthy records, missing channels, etc.



(a)



(b)

Figure 5.1 (a) Location of epicenters and (b) histogram of magnitude of included earthquake events

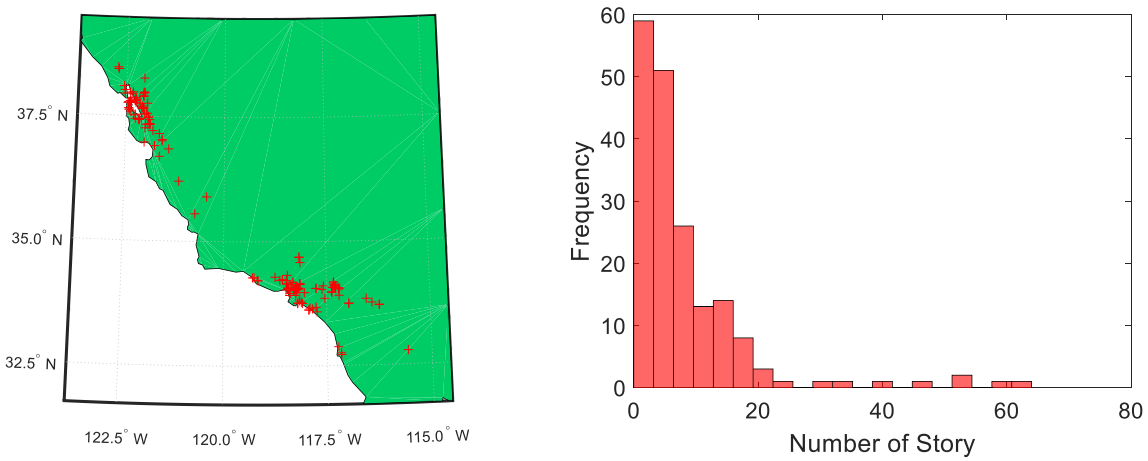
Table 5.1 Summary of included earthquake events

Event Name	Depth (km)	Mw. *	Date	Latitude	Longitude	No. of Instrumented Buildings
MorganHill	9	6.2	4/24/1984	37.32	-121.68	31
MtLewis	6	5.8	3/31/1986	37.466	-121.691	43
Whittier	9.5	6.1	10/1/1987	34.06	-118.07	8
LomaPrieta	18	7	10/17/1989	37.04	-121.88	4
SierraMadre	12	5.8	6/11/1991	34.26	-118	44
Landers	1.1	7.3	6/28/1992	34.217	-116.433	24
BigBear	1	6.5	6/28/1992	34.2	-116.83	2
Northridge	19	6.4	1/17/1994	34.2057	-118.5539	16
Bolinas	7	5	8/17/1999	37.91	-122.69	5
Gilroy	7.6	4.9	5/13/2002	36.97	-121.6	4
BigBearCity	12.7	5.4	2/22/2003	34.31	-116.85	8

SimiValley	5.1	3.7	10/29/2003	34.29	-118.75	16
Parkfield	7.9	6	9/28/2004	35.81	-120.37	7
Anza05	13.1	5.2	6/12/2005	33.53	-116.57	12
AlumRockArea	9.2	5.4	10/30/2007	37.432	-121.776	11
ChinoHills	13.6	5.4	7/29/2008	33.95	-117.77	48
Calexico	10	7.2	4/4/2010	32.26	-115.29	22
Borrego	14	5.4	7/7/2010	33.42	-116.49	1
Berkeley11	8	4	10/20/2011	37.86	-122.25	1
Anza13	13.1	4.7	3/11/2013	33.5	-116.46	3
Encino	9.9	4.4	3/17/2014	34.13	-118.49	7
SouthNapa	11.3	6	8/24/2014	38.2155	-122.3117	1
Borrego16	12.3	5.2	6/10/2016	33.43	-116.44	22
Berkeley18	12.3	4.4	1/4/2018	37.8552	-122.2568	31

* Magnitude

The recordings come from 188 instrumented buildings subjected to the 24 earthquakes. Building locations are shown in Figure 5.2a and are distributed near epicenters such that majority of instrumented buildings are subjected to more than one event. In other words, both cross-event and site-to-site variability exists in the dataset. A histogram for number of stories in the buildings, shown in Figure 5.2b, appears to follow log normal distribution and the data covers a large range of instrumented buildings, although there are more low-rise buildings as might be typical within an urban setting. Building height, as a predictor for the model developed in this study, is transferred to a logarithm feature such that it approximately follows normal distribution to avoid bias in the prediction model.



(a) (b)

Figure 5.2 (a) Locations and (b) histogram of number of stories of instrumented buildings

5.2.2 Transfer recordings to EDP

The source data format is in terms of acceleration, $a(t)$, and displacement, $d(t)$, response histories for each channel of each measurement location within instrumented buildings subjected to one event. These physical quantities are transferred into damage related EDPs, PFA and PSDR, respectively. Figure 5.3 demonstrates a typical data example of a 13-story building (CSMIP station No. 24322) subjected to 2014 Encino earthquake. On each orthogonal translational direction, $a(t)$ from corresponding channels, 1, 4, 7 and 10 in this case, were used to interpolate between instrumented floors to obtain acceleration response histories of each floor to obtain $a_i(t)$ at ith floor. $d(t)$ from each channel, which was integrated and base-corrected from $a(t)$, were also used to interpolate for un-instrumented floors. The ith floor story drift ratio time histories, $\delta_i(t)$, were obtained by Equation (5-3).

$$\delta_i(t) = \frac{d_i(t) - d_{i-1}(t)}{h_i - h_{i-1}} \tag{5-3}$$

Where $d_i(t)$ and $d_{i-1}(t)$ correspond to displacement response histories at floor i and $i-1$

while n represents total number of floors. h_i and h_{i-1} is the floor height between two adjacent floors. The maximum absolute value over time of $a_i(t)$ and $d_i(t)$ are the EDPs of interest, namely, PFA and PSDR along two orthogonal horizontal directions of the building. The higher demand from the two directions is used to represent the seismic building demand envelop.

Two assumptions are considered regarding building response mechanism for this transformation process: 1) a rigid diaphragm is assumed at each floor of all instrumented building, and 2) torsional responses from instrumented buildings are negligible. These two assumptions introduce certain limitations; however, they are applied given that common sensor layouts are insufficient to consider rotational or non-rigid floor behavior without case-by-case study. In addition, linear interpolation between floors for both acceleration and displacement over each time step is applied to generate full profile building responses. Another 13-story building at San Jose (CSMIP station No. 57357) with sensors installed at a lower level, 2nd, 7th, 12th and roof levels is used to demonstrate linear interpolation accuracy. Leave-one-out cross validation is used by linearly interpolating accelerometer response history records for the H1 direction of the building at 2nd, 7th, 12th floor using records of all other floors. Median Absolute Deviation (MAD) is reported for each floor subjected to four earthquakes and shown in Table 5.2. Highest error is at the 7th floor where MAD exceeds 0.01g for Morgan Hill earthquake. The average median MAD of each floor is 0.0019g, 0.0065g and 0.0043g, respectively, indicating that linear interpolation is less accurate at mid-height of building while the overall interpolation result is good. Nonetheless, it should be noted that most of the concerned floor accelerations are relatively small and the error sources might be from instrumentation, filtering as well as this interpolation process.

Table 5.2 Average median MSE of PFA interpolation

Earthquake	2nd Floor (g)	7th Floor (g)	12th Floor (g)
AlumRockArea	0.0002	0.0005	0.0003
LomaPrieta	0.0027	0.0085	0.0056
MorganHill	0.0027	0.0108	0.0068
MtLewis	0.0022	0.0061	0.0043

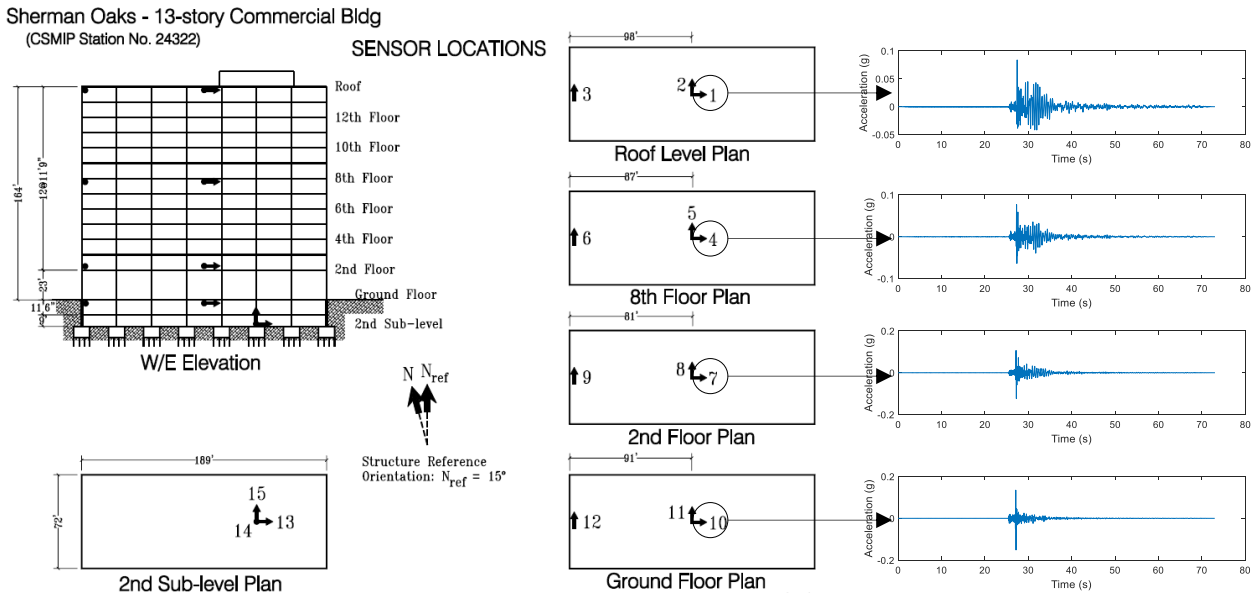
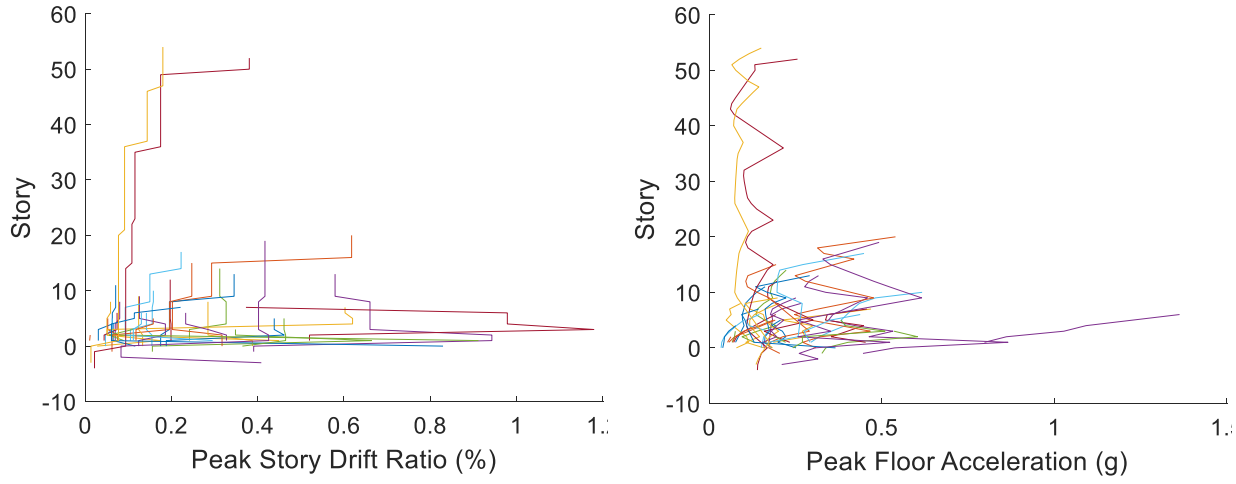


Figure 5.3 Example 13-story building sensor plan layout and accelerometer channel output subjected to the 2014 Encino earthquake

The linear interpolated PFA and PSDR from 35 instrumented buildings subjected to 1994 Northridge earthquake is shown in Figure 5.4a and Figure 5.4b, respectively. Among these buildings are two tall buildings with more than 50 stories with relatively low seismic demands compared to other instrumented buildings for this event: one is 32.1 km from epicenter and has maximum 0.4% story drift ratio and 0.3g floor acceleration demand at the roof. It is worth mentioning that the seismic response of any building is determined by ground motion characteristics, site effects, and building properties. Figure 5.5a further shows rupture distance and maximum PSDR along building height for all buildings subjected to this event. The highest drift

demand, 1.2%, comes from a 10-story building with rupture distance of 7.8 km. A strong negative correlation coefficient of -0.53 is observed between rupture distance and PSDR regardless of building characteristics. By grouping building seismic response data into low-rise (1-3 story), mid-rise (4-8 story) and high-rise (>8 story), the correlation becomes -1.0, -0.58, and -0.49, respectively, indicating that decoupling structural dissimilarities from seismic response data would improve trend pattern between PSDR and rupture distance. A similar result is observed for PFA as visualized in Figure 5.4b and Figure 5.5b. Figure 5.6a to Figure 5.6d show how the trend between PFA over rupture distance changes as magnitude of earthquake increases from less than 5 to more than 7. The strongest exponential trend is observed for magnitude of 7 or bigger events and the pattern decreases as magnitude drops. For events with magnitude between 5 and 7, the trend appears to be linear with higher dispersions, whereas the dispersion keeps increasing as magnitude drops to 4. As a result, the magnitude scaling term, is expected to have a nonlinear formulation to reflect such behavior.



(a) (b)

Figure 5.4 (a) peak story drift ratio and (b) peak floor acceleration profile from the 1994 Northridge earthquake

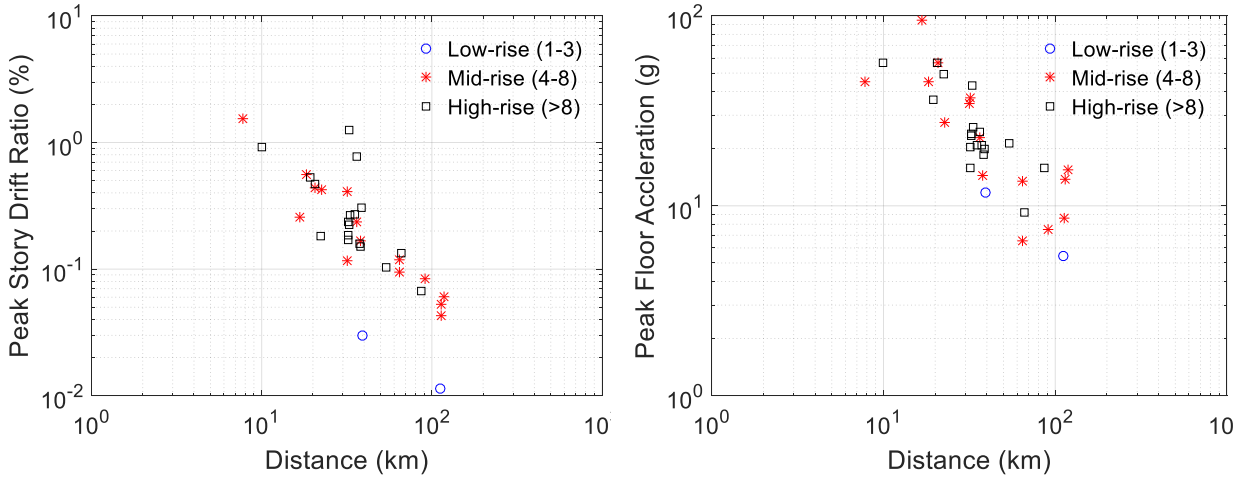
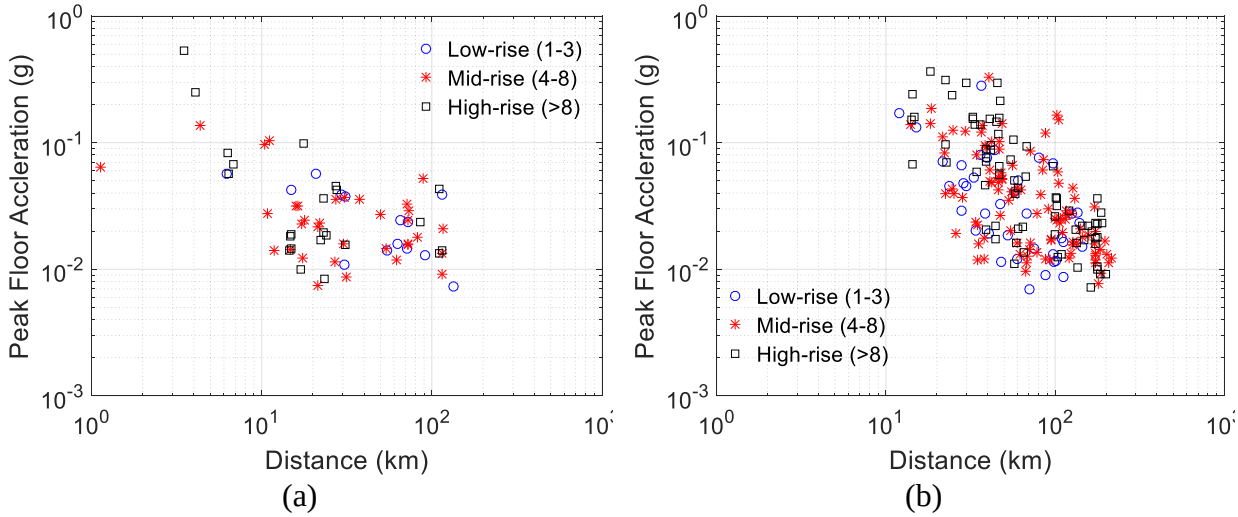


Figure 5.5 Maximum response along building height of (a) peak story drift ratio and (b) peak floor acceleration from the 1994 Northridge earthquake



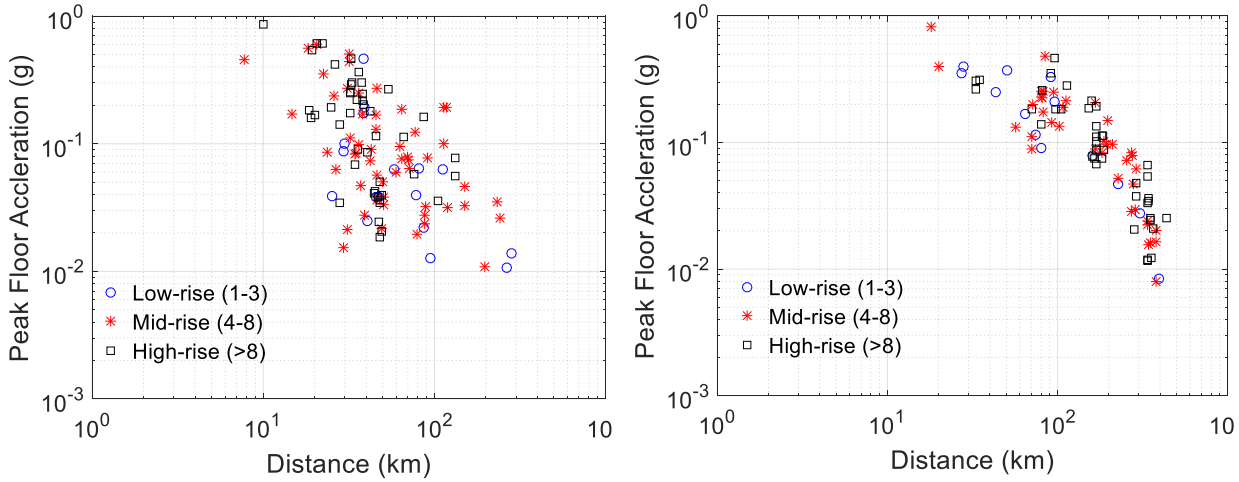


Figure 5.6 Peak floor acceleration trend over rupture distance of events with magnitude (a) less than 5 (b) between 5 and 6 (c) between 6 and 7 (d) greater than 7

5.3 Model Formation and Determination of Coefficients

5.3.1 Mix Effect Model

Our proposed generalized cross-building EDP reconstruction model consists of three event and site terms which are consistent with that of GMPE, moment scaling, distance function, and site amplification. An additional building term is added to account for structural dissimilarity in Equation (5-4).

$$\ln Y = F_M(M) + F_D(d_{JB}) + F_S(V_{S30}) + F_B(H_B) + e \qquad (5\text{-}4)$$

Where Y is the interested EDP, M is event magnitude, d_{JB} is rupture distance and H_B is height of the building. Compared to popular GMPEs (i.e., [149,150]), where magnitude is coupled with distance and site terms, our form is a simplified version with the intention of keeping a minimum number of predictors and a simplified formation such that we do not need to specifically tune each term as what has been done in many GMPE models. As observed in previous Figure 5.6a to Figure 5.6d, a nonlinear relationship is believed to exist in magnitude term and a

polynomial to the power of two is applied to account for it, similarly to the GMPE model proposed by Boore and Atkinson [149]. The distance function and site amplification is designed to be similar to early GMPE models [138,139] considering they worked very well and were developed using a smaller dataset than used in this study. Two features are included for the building terms in our model, height of building H_j, and its empirical period T_j, which is a quantitative measure of structural material (e.g., concrete, steel),the lateral force system type, and stiffness and mass distribution over its height. The two considered building terms used were included because they reflect linear and nonlinear responses of buildings in a general manner and any other strucutral related features, such as moment frame dimensions or structural wall thickness would be too complicated to use within the proposed model. In addition, site parameter and building height are regularized by applying a normalizing factor for each to avoid overfitting contributed from individual outliers. The final complete form of our generalized cross-building EDP reconstruction model is in Equation (5-5).

$$\begin{aligned}\ln(Y_{ij}) &= B_1 + B_2(M_i - M_{ref}) + B_3(M_i - M_{ref})^2 \\ &- \ln(d_{ij}^2 + h^2)^{\frac{1}{2}} + B_4(d_{ij}^2 + h^2)^{\frac{1}{2}} + B_5 \ln\left(\frac{V_{s30,j}}{V_{ref}}\right) + B_6 \ln\left(\frac{H_j}{H_{ref}}\right)^{\frac{1}{2}} + B_7 \ln\left(\frac{T_j}{T_{ref}}\right) + e_{ij}\end{aligned} \quad (5\text{-}5)$$

Where $B_{1..7}$ and h are scaling parameters to be determined and i is index for event and j is index for site. h is an adjustment parameter for the distance measure. V_{ref} is a reference coefficient for V_{S30} with fixed value of 760m/s as suggested in [149] and H_{ref} is a normalization factor for building height H with fixed value of 12ft such that $\frac{H_{ij}}{H_{ref}}$ becomes a level indicator of averaging number of stories for each building.

5.3.2 Meta Data Retrieval

There are five features used in this mix-effect model, namely, event magnitude (M), rupture distance (d), the time-averaged 30 m shear wave velocity (V_{S30}), height of the building (H) and the empirical period (T). Event magnitude, rupture distance, and height of the building are retrieved through the CESMD database. V_{S30} data of each considered site is approximated through 1-nearest neighbors method using a V_{S30} Map for California with Geologic and Topographic Constraints provided by Thompson et., al. [155] with adjustments based on updated geology provided by Wills and Clahan [156], which contains metadata of V_{S30} with resolution at 7.5 arcseconds, approximately 250m across entire California. A colormap of this database is demonstrated in Figure 5.7 where V_{S30} ranges from 176.1m/s to 1636.0m/s over the land area of California. Considering its dense resolution, V_{S30} of nearest site from the database is used as estimated V_{S30} in our model. A histogram of absolute estimating distance between actual building site and the used estimating site is shown in Figure 5.8. It shows that the maximum estimating distance is at 150m and the sample mean is at 79.2m.

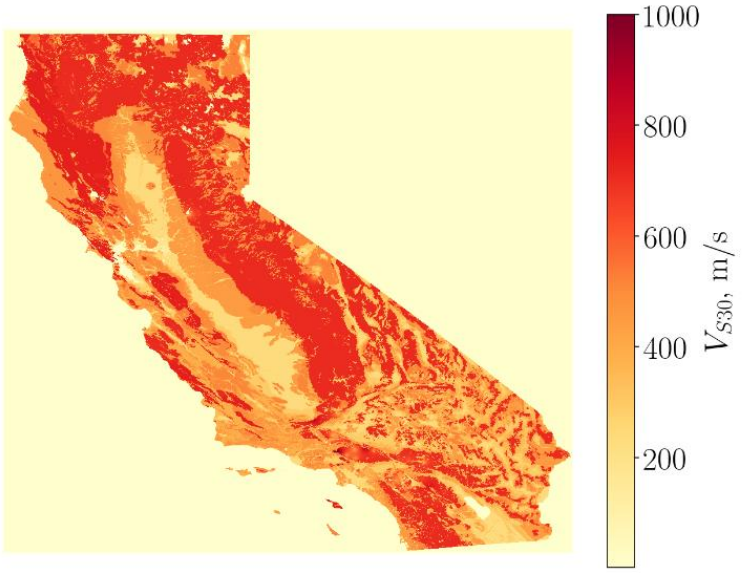


Figure 5.7 Colormap of V_{S30} database by Thompson et. al. [155]

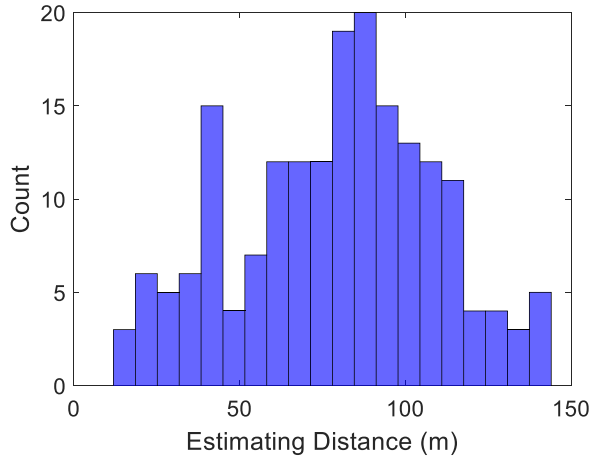


Figure 5.8 Histogram of estimating distance between actual building site and the used estimated site

The empirical period of the building could be retrieved through system identification using strong motion recording data such that soil structure interaction and onsite strucutral characteristics are properly considered to reflect building's actual response period. Example methodologies include [157] by Stewart and Fenves and [158] by Ghahari et. al. However, due to the highly

nonlinear responses introduced by soil structure interaction for special strucutral systems (e.g., base isolated structures), it is difficult to find a consistent empirical building response period for the sites being considered in this study. An alternative approach, estimating empirical building response period using ASCE-7-13 Eqn. 12.8-8 [101], is used instead. In order to get ASCE-7 estimation, the structural material and lateral force system type of each building are needed and retrieved through surveying the text information provided by CESMD database. Use of an empirical estimation of building response period through ASCE-7-13 is systematically biased from actual building response period given the way it was calibrated (to generally underestimate the actual period so as to overestimate design forces); the included buildings in our database includes a number of special structures such as base-isolated structures and high-rises that require case-by-case evaluation to retrieve their correct response period. To avoid inconsistence and remain concise in the period feature, ASCE-7 estimation is adopted at this point. In the future, a refined approach could be introduced to evaluate each building site case by case and apply improved methods to estimate building periods.

5.3.3 2-stage Regression

A 2-stage regression procedure described in [9, 10] was applied to retrieve model parameters in Equation (5-5). In order to regress for the above-mentioned model parameters, the original nonlinear form, Equation (5-5), is modified to a linear version (Equation (5-6)) by incorporating Taylor expansion at h.

$$\mathbf{Y_1} = \mathbf{X_1}\mathbf{B_1} + \mathbf{e_1} \tag{5-6}$$

$$\mathbf{X_1} = \begin{bmatrix} (d_1^{\,2}+h'^2)^{1/2} & \left\{\frac{\partial}{\partial h}\left[B_4(d_1^{\,2}+h'^2)^{1/2}-\ln(d_1^{\,2}+h'^2)^{1/2}\right]\right\}_{h=h'} & \ln\left(\frac{V_{S30,1}}{V_{S30,ref}}\right) & \ln\left(\frac{H_1}{H_{ref}}\right) & \ln\left(\frac{T_1}{T_{ref}}\right) & E_{11} & E_{12} & \cdots & E_{1N_e} \\ (d_2^{\,2}+h'^2)^{1/2} & \left\{\frac{\partial}{\partial h}\left[B_4(d_2^{\,2}+h'^2)^{1/2}-\ln(d_2^{\,2}+h'^2)^{1/2}\right]\right\}_{h=h'} & \ln\left(\frac{V_{S30,2}}{V_{S30,ref}}\right) & \ln\left(\frac{H_2}{H_{ref}}\right) & \ln\left(\frac{T_2}{T_{ref}}\right) & E_{21} & E_{22} & \cdots & E_{2N_e} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ (d_n^{\,2}+h'^2)^{1/2} & \left\{\frac{\partial}{\partial h}\left[B_4(d_n^{\,2}+h'^2)^{1/2}-\ln(d_n^{\,2}+h'^2)^{1/2}\right]\right\}_{h=h'} & \ln\left(\frac{V_{S30,n}}{V_{S30,ref}}\right) & \ln\left(\frac{H_n}{H_{ref}}\right) & \ln\left(\frac{T_n}{T_{ref}}\right) & E_{n1} & E_{n2} & \cdots & E_{nN_e} \end{bmatrix} \quad (5\text{-}7)$$

$$\mathbf{Y}_1 = \begin{bmatrix} \ln Y_1 + \ln(d_1^{\,2}+h'^2)^2 \\ \ln Y_2 + \ln(d_2^{\,2}+h'^2)^2 \\ \vdots \\ \ln Y_n + \ln(d_n^{\,2}+h'^2)^2 \end{bmatrix} \quad (5\text{-}8)$$

$$\mathbf{B}_1 = \begin{bmatrix} B_4 \\ \Delta h \\ B_5 \\ B_6 \\ B_7 \\ \hat{P}_1 \\ \vdots \\ \hat{P}_{N_e} \end{bmatrix} \quad (5\text{-}9)$$

$\mathbf{X}_1$ (Equation (5-7)) is the predictor matrix expanded by Taylor's series expansion at h. h' is the trail value of h and n is total number of data points. The newly introduced coefficient, $E_{k,i}$, is an indicator such that it equals to 1 if observation k is from event i otherwise 0. N_e from subscription represents total number of events. In Equation (5-8) and (5-9), Δh is an increment of h and $\hat{P}_i$s are defined in Equation (5-10) including magnitude scaling term and regression residual from stage 1.

$$\hat{P}_i = B_1 + B_2\left(M_i - M_{ref}\right) + B_3\left(M_i - M_{ref}\right)^2 + e_i \quad (5\text{-}10)$$

The stage-1 regression can be iteratively accomplished by first solving Equation (5-6) using ordinary least square estimator, $\hat{\mathbf{B}}_1 = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{Y}_1$, with assumption that the regression residual $\mathbf{e}_1$ follows Gaussian distribution with 0 mean and σ_r^2 variance; and updating the assumed h until convergence achieved as demonstrated in Figure 5.9.

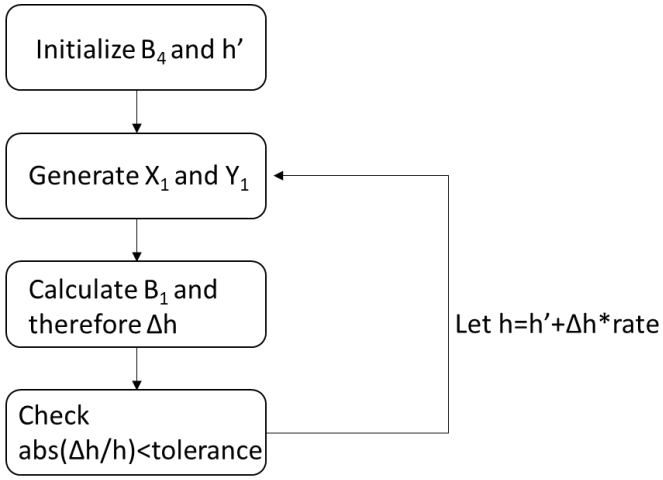


Figure 5.9 Flowchart for stage-1 regression iteration

Stage-1 provides all model parameters in $\mathbf{B}_1$ of Equation (5-9) except magnitude scaling coefficients, which are determined using stage-2 regression by expanding Equation (5-10) to (5-11). Equations (5-12) to (5-14) involves a stage-2 regression formulated as $\mathbf{Y}_2 = \mathbf{X}_2\mathbf{B}_2 + \mathbf{e}_2$. Since there is covariance between each residual of $\mathbf{e}_2$, a generalized least square estimator, $\hat{\mathbf{B}}_2 = \left(\mathbf{X}_2^T\mathbf{V}_2^{-1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{V}_2^{-1}\mathbf{Y}_2$, is used where $\mathbf{V}_2$ represents covariance matrix of $\mathbf{e}_2$ and is given by Equation (5-15).

$$\hat{P}_i = B_1 + B_2\left(M_i - M_{ref}\right) + B_3\left(M_i - M_{ref}\right)^2 + \left(\hat{P}_i - P_i\right) + e_i \qquad (5\text{-}11)$$

$$\mathbf{Y}_2 = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_{N_e} \end{bmatrix} \tag{5-12}$$

$$\mathbf{X}_2 = \begin{bmatrix} 1 & M_1 - M_{ref} & \left(M_1 - M_{ref}\right)^2 \\ 1 & M_2 - M_{ref} & \left(M_2 - M_{ref}\right)^2 \\ \vdots & \vdots & \vdots \\ 1 & M_{N_e} - M_{ref} & \left(M_{N_e} - M_{ref}\right)^2 \end{bmatrix} \tag{5-13}$$

$$\mathbf{B}_2 = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \tag{5-14}$$

$$\mathbf{V}_2 = \text{var}(\hat{\mathbf{P}} - \mathbf{P}) + \sigma_e^2 \mathbf{I} = \left[\text{var}\left(\hat{\mathbf{B}}_1\right)\right]_{i+5,k+5} + \sigma_e^2 \mathbf{I} = \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \sigma_r^2 \tag{5-15}$$

To solve for $\hat{\mathbf{B}}_2$, a trial and error process is conducted using Equation (5-16).

$$E\left[\left(\mathbf{Y}_2 - \mathbf{X}_2\mathbf{B}_2\right)^T \mathbf{V}_2^{-1}\left(\mathbf{Y}_2 - \mathbf{X}_2\mathbf{B}_2\right)\right] = N_e - rank(\mathbf{X}_2) \tag{5-16}$$

The above 2-stage regression process calculates all parameters of the generalized cross-building EDP reconstruction model. Table 5.2 shows all tuned parameter values using the prescribed dataset for two considered EDPs, peak story drift ratio and peak floor acceleration respectively.

Table 5.3 Calibrated Model Parameter

EDP Prediction Model	B_1	B_2	B_3	B_4	h	B_5	B_6	B_7
Peak Story Drift Ratio	1.0685	1.3492	-0.0363	0.0062	5.9805	-0.1210	-0.1735	0.3449

Peak Floor Acceleration	1.3153	1.0565	-0.1449	0.0027	2.2582	-0.1614	-0.0157	-0.0021

5.3.4 Results

With the calibrated parameters from Table 5.3 for PSDR, Equation (5-5) is used to reconstruct PSDR medians and the total residuals, e_{ij}, at each recording site (Figure 5.10a). The overall median of observed-to-predicted residual ratio is at 0.79, reflecting an under-estimated bias using the proposed reconstruction model. The percentage of data points being over-estimated by a factor of 2 or more (observed-to-predicted ratio is less than 0.5) is 3.2% of the total population while data points under-estimated by 0.5 or less is about 21.3% (observed-to-predicted ratio is greater than 2), with the majority of the bias from rupture distance of 30 km or more. On the other hand, the model is relatively unbiased for higher demands, which would produce the most building damage and often occur within 30 km from epicenter. For PFA, a median of observed-to-predicted ratio is at 1.18 as shown in Figure 5.10b. The figure shows that prediction outliers (ratio that is greater than 2 and less than 0.5) occur mostly at higher rupture distance (greater than 30km), similarly to that of PSDR. Among all observations, an upward trend is observed between the observed-to-predicted ratio and rupture distance. These observations suggest that the current reconstruction model is approximately unbiased with close-epicenter data but is not able to capture data points that are relatively far from the epicenter with the observed underestimating trend for PSDR and overestimating trend for PFA.

Figure 5.11 shows the total residual e, observed subtracted by predicted, versus rupture distance (a), building height (b), event magnitude (c) and ASCE-7 empirical period for PSDR. No significant pattern is observed between these prediction features and the total residual indicating that additional effort to expand current feature term design, namely, $F_M(M)$, $F_D(d_{JB})$ and $F_B(H_B)$, may not further improve model performance in terms of reducing bias. Higher dispersion

of total residual is observed at lower ASCE-7 empirical period as well as lower rise buildings according to Figure 5.11a and Figure 5.11d, suggesting that stiffer buildings introduce higher variance in the data-driven model.

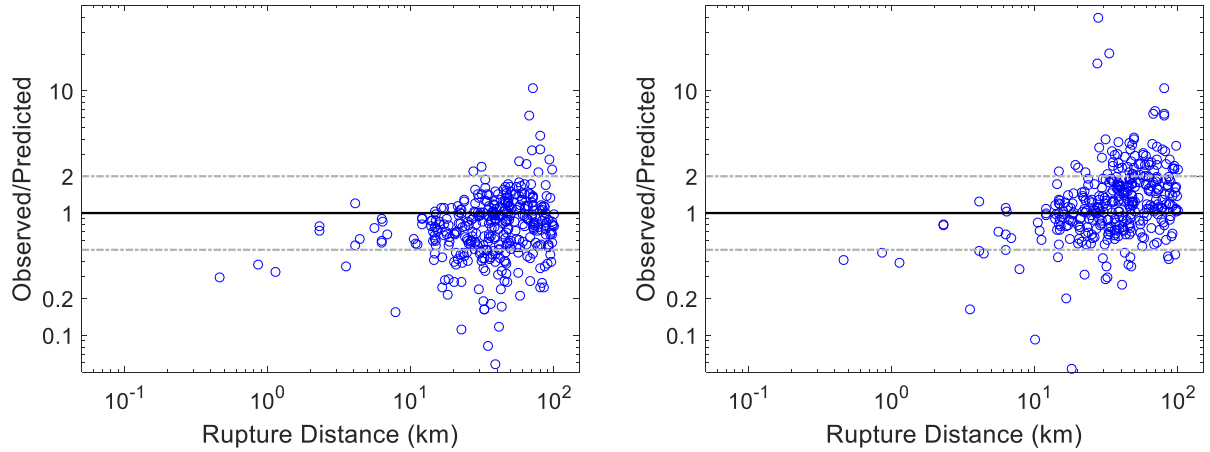
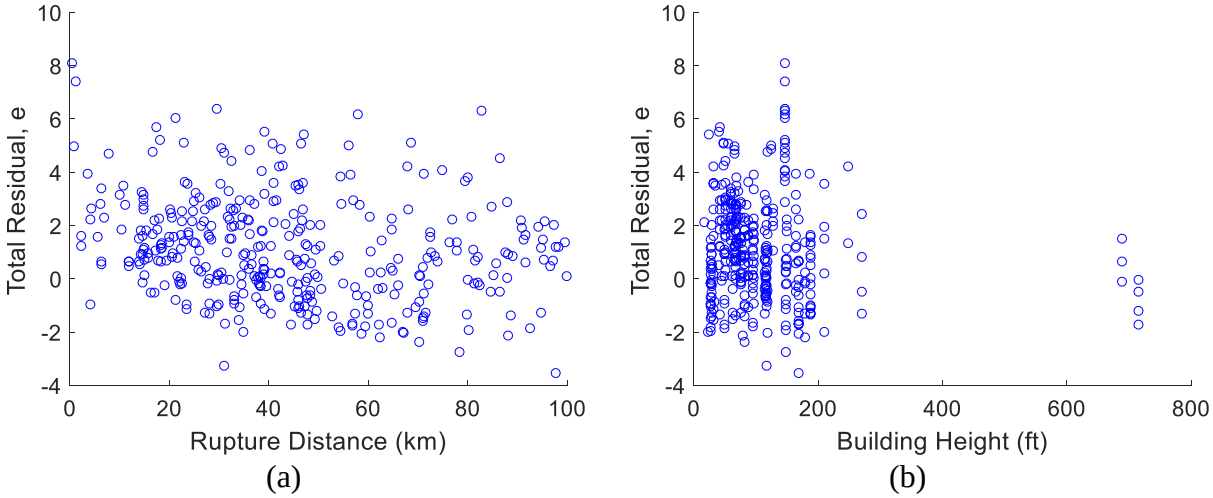


Figure 5.10 Observed/Predicted ratio versus rupture distance of (a) PSDR and (b) PFA of Equation (5-5)



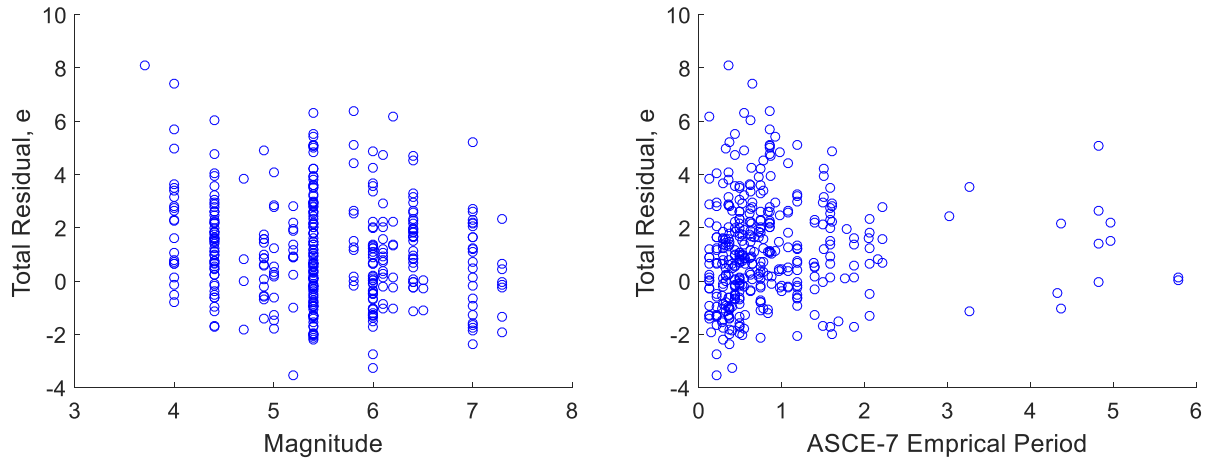


Figure 5.11 Total residual versus (a) rupture distance, (b) building height, (c) magnitude and (d) ASCE-7 empirical period for PSDR of Equation (5-5)

5.4 Adopting Ground Motion Prediction Equation

Given the observed bias in this EDP reconstruction model, an alternative is to apply modern GMPE as a base model and expand it with additional building terms to adopt EDP response instead of IMs in GMPE. The advantage of taking existing GMPE model is that modern GMPEs are well-developed with detail calibration and parameter tuning to consider event characteristics, distance decay and soil influences using numerous ground motion data. It is tuned to be unbiased over these different dimensions. The proposed model in Equation (5-5) is modified to Equation (5-17).

$$\ln(Y_{ij}) = C_1 \ln(GMPE_{median}) + C_2 \ln\left(\frac{H_j}{H_{ref}}\right)^{\frac{1}{2}} + C_3 \ln\left(\frac{T_j}{T_{ref}}\right) + e_{ij} \qquad (5\text{-}17)$$

C_1, C_2 and C_3 are model parameters and $GMPE_{median}$ is the corresponding GMPE median SA of each data using one of the latest GMPE model from [159], which is considered as most suitable for California. The two building terms remain the same format so as their normalizing constants, H_{ref} and T_{ref}. With this adopted model, the coefficients are summarized in Table 5.4.

Table 5.4 Calibrated Adopted Model Parameter

Adopted EDP Prediction Model	C_1	C_2	C_3
Peak Story Drift Ratio	0.7161	-0.0782	1.2272
Peak Floor Acceleration	0.5739	-0.1632	0.8104

Observed/predicted ratio versus rupture distance figure for PSDR and PFA of this adopted generalized cross-building EDP reconstruction model is shown in Figure 5.12. Comparing Figure 5.10a with Figure 5.12a, the result is vastly improved as ratio between observed over predicted is restrained below factor of 2 indicating that underestimating PSDR is well bounded. The median of this ratio is 0.98 reflecting an ideal unbiased prediction model. In addition, only 8.4% of entire dataset is overestimated by a factor of 2 or more suggesting that the model in Equation (5-17) is improved to restrain underestimation without sacrificing unbiases and unbounding overestimation. The percentage of predictions that underestimate more than 50% compared to actual observed data is improved for PFA as well (Figure 5.12b). Median of ratio between observed and predicted is 0.98 and proportion of overestimated data points is 6.5%.

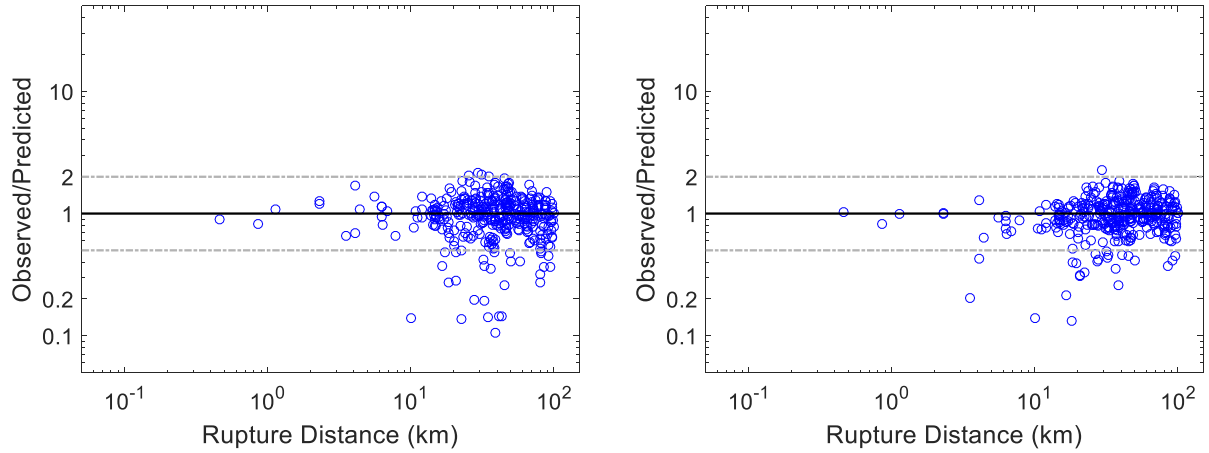


Figure 5.12 Observed/predicted ratio versus rupture distance of (a) PSDR and (b) PFA of adopted generalized cross-building EDP reconstruction model in Equation (5-17)

5.5 Total Residual Prediction Incorporating Spatial Demand Parameters

As discussed before, the proposed model is used to account for feature trends from event, rupture distance, site and building properties and predict median responses. In order to reconstruct building seismic responses, we need to further develop a model for the total residual, e_{ij}. As indicated by Equation (5-2), according to Jayaram and Baker [96], residuals from the model can be decoupled to two components, inter-event residual and intra-event residual. By rearranging it based on a more intuitive notation, Equation (5-17) can be presented as Equation (5-18).

$$\ln(Y_{ij}) - \ln(\bar{Y}_{ij}) = e_{ij} = \eta_{e,i} + \eta_{s,j} + \varepsilon_{ij} \tag{5-18}$$

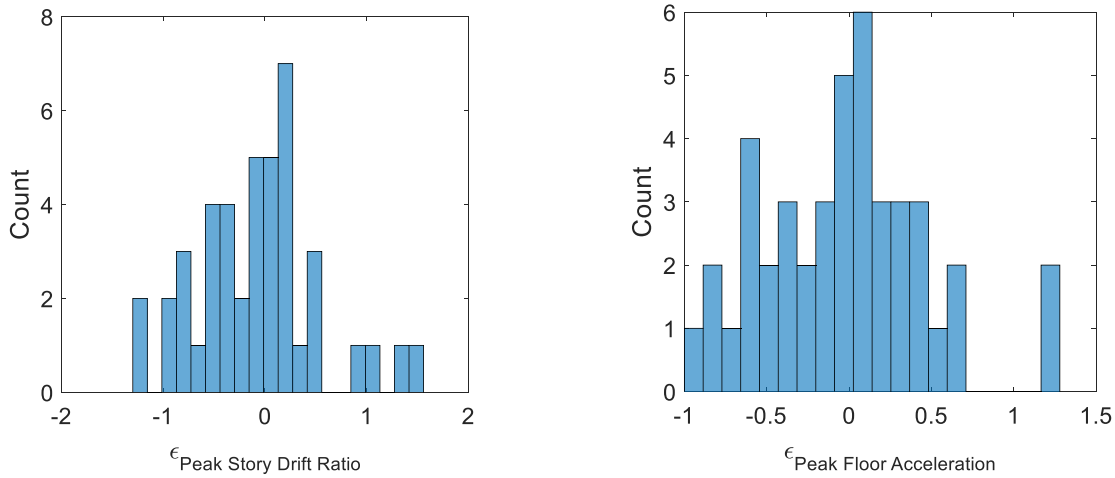
Where $\eta_{e,i}$ is the inter-event residual for event i, $\eta_{s,j} + \varepsilon_{ij}$ is the intra-event residual at site j and ε_{ij} is defined as record residual for site j from event i. In later context, the inter-event residual will be called within-event residual and $\eta_{s,j}$ from the intra-event residual will be decoupled and called site residual. Assuming that η_e follows a zero mean Normal distribution, the sample mean of within-event residual $\eta_{e,i}$ can be calculated as:

$$\eta_{e,i} = \frac{1}{N_{s,i}} \sum_{j=1}^{N_{s,i}} e_{ij} \tag{5-19}$$

Where $N_{s,i}$ is total number of sites from event i. The sample variance can also be computed by $\sigma_{e,i}^2 = \frac{1}{N_{s,i}} \sum_{j=1}^{N_{s,i}} (e_{ij} - \eta_{e,i})(e_{ij} - \eta_{e,i})$. Similarly, assuming the site residual η_s follows a zero mean Normal distribution, the sample mean of site residual $\eta_{s,j}$ at site j can be calculated as:

$$\eta_{s,j} = \frac{1}{N_{e,j}} \sum_{i=1}^{N_{e,j}} e_{ij} - \eta_{e,i} \tag{5-20}$$

Where $N_{e,j}$ donates number of events being recorded at site j. Sample variance $\sigma^2_{s,j}$ of site j can also be calculated handy. Equation (5-19) and (5-20) gives sample mean estimate of within-event residual and site residual enabling further decoupling event and site effect from total residual e_{ij}. Such procedure provides us the within-event residual $\eta_{e,i}$ for each considered earthquake i and site residual $\eta_{s,j}$ for each considered building site j. The histogram of the record residual, ε_{ij}, for both PSDR and PFA is shown in Figure 5.13. The histograms show that residuals from the current generalized cross-building EDP reconstruction model and decoupled site and event effects is approximately unbiased with standard deviation of 0.613 for PSDR and 0.492 for PFA. The next step is to reconstruct these residuals using available features to further reduce their variances.



(a) (b)

Figure 5.13 Record residual ε histogram of (a) PSDR and (b) PFA

Due to the small amount of data recorded per site and very few recorded data for some earthquake events (e.g., the 1992 Big Bear earthquake where responses were obtained in only two

buildings), site and event terms could be inaccurate. One popular adopted method to reduce record residual variance is to utilize geographical features within each event. A common measure to quantify spatial correlation is to use the distance separating two sites, i.e., acknowledging that this geographical feature of each building site is directly related to covariance between sites through the distance separating the sites. A correlogram generated using record residuals ε_{ij}, which are regressed from the adopted generalized cross-building EDP reconstruction model using Equation (5-17), and excluded corresponding within-event and site residuals using Equation (5-18 to 20) for the 2007 Alum Rock earthquake of PSDR and PFA, are shown in Figure 5.14a. It is observed that correlation coefficients are proportionally related to the separating distance within 7km inter-site distance range, observing perfect correlation at 0 inter-site distance and zero correlation at 7km inter-site distance, implying that variance of the record residuals can be further reduced by incorporating neighborhood data between recording sites within the same event. The result also suggests that spatial trend is properly removed using Equation (5-17) at the resolution scale of 7km as correlation between sites are not observed beyond 7km.

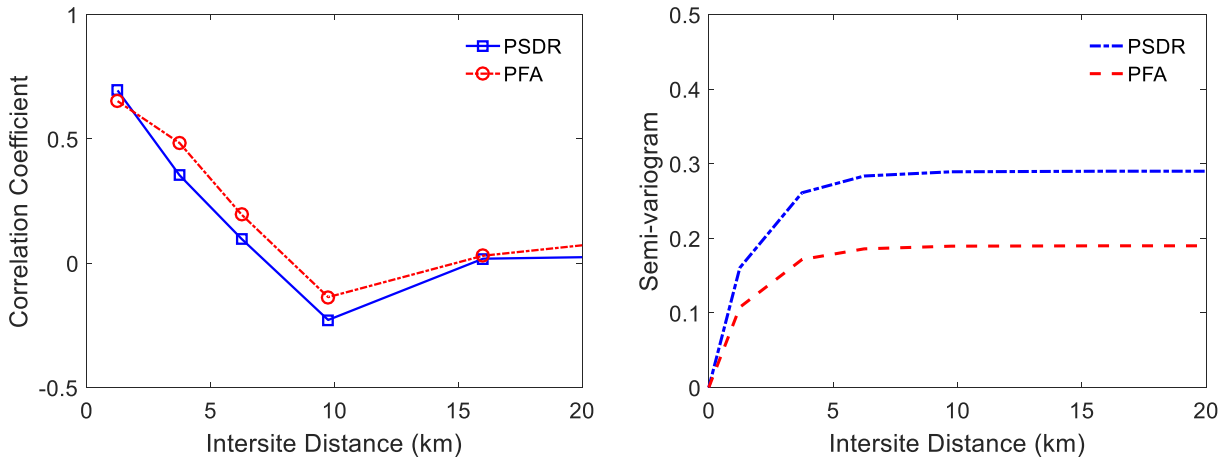
The next step is to reconstruct record residuals using their geographical correlation pattern such that the variance of prediction is further reduced. A spatial prediction procedure, kriging, is introduced to extract the distance-dependent information from record residuals. Kriging, also referred to as Gaussian process regression, is an unbiased estimator that interpolates spatially distributed random variables using prior covariances defined by separating distance h between two sites. The prior covariance is retrieved through using available recorded data to fit an empirical semi-variogram model $\gamma(h)$, which is given by Equation (5-21) [109].

$$\gamma(h) = \frac{\left\{ \frac{1}{N(h)} \sum_{N(h)} | X_a - Y_b |^{1/2} \right\}^4}{0.457 + \frac{0.494}{N(h)}} \tag{5-21}$$

Where X_a and X_b are all pairs of data that have separating distance ranges from $h - \Delta h$ to $h + \Delta h$ and there are $N(h)$ pairs of them for each h. Δh determines the resolution of resulting semi-variogram and is chosen to fit number of available data points. An exponential function is used to fit the empirical $\gamma(h)$ to generate a continuous covariance model (Figure 5.14b) of both PSDR and PFA. Observing that the correlation trend in Figure 5.14a for PFA is generally stronger than that of PSDR (i.e., $\rho_{PFA} = 0.55$ compared to $\rho_{PSDR} = 0.42$ at 3.75km and $\rho_{PFA} = 0.15$ compared to $\rho_{PSDR} = 0.02$ at 6.75km) indicates that record residual of PFA is generally spatially correlated which is reflected in Figure 5.14b with a less stiff semi-variogram. The kriging procedure is then applied to interpolate PSDR and PFA. Detail fitting and interpolation procedures can be found in [28] by Han et. al.

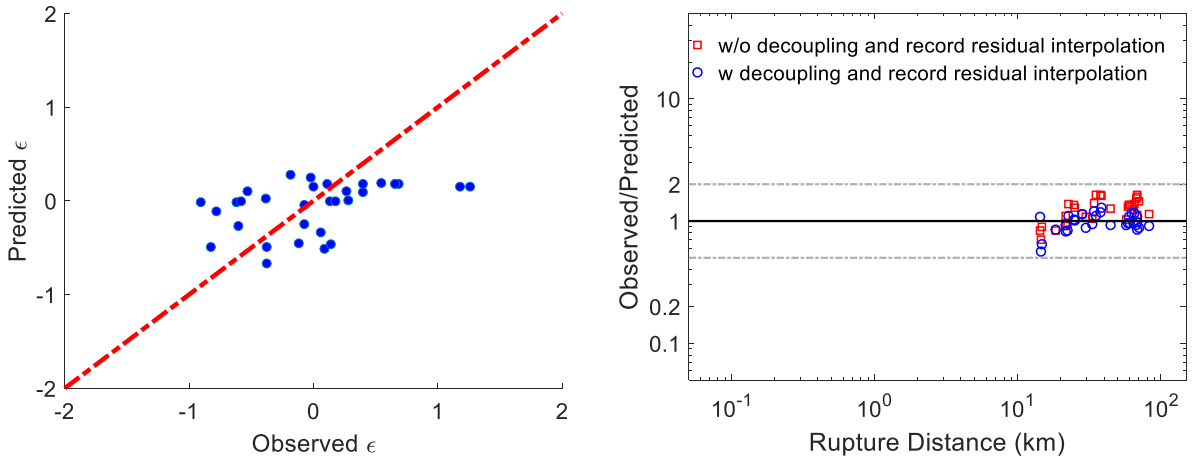
A standard Machine Learning validation procedure, training/testing random split, is applied to examine the constructability of record residuals for the above described 2007 Alum Rock earthquake and the overall prediction performance combined with previous proposed adopted model. The interpolation result of record residual is shown in Figure 5.15 by randomly selecting 30% of PSDR record residuals as training data to predict the remaining 70% of total 31 observations from the event using the kriging procedure in [28]. The observed versus predicted record residual is on par with each other as shown in Figure 5.15a. To further examine the effect to reduce variance for prediction, Figure 5.15b shows a plot of observed/predicted ratio before (red box) and after (blue circle) the record residual is applied. A significant improvement in prediction performance is observed as the ratio variance drops from 0.104 to 0.02 with kriging interpolation.

A similar result is also observed for PFA record residuals for this particular event. However, it is worth mentioning that there are events with very few recorded data points such as 1984 Morgan Hill earthquake with only 7 recorded data available, which makes it impossible to generate valid semi-variograms. As a result, the record residual interpolation method is not applicable for every event in our dataset. In addition, the kriging interpolation performance of record residuals is not stable when the number of available recordings is insufficient to generate a representative semi-variogram model. Lack of recordings also causes sparse distribution of sites and introduces higher variance in the kriging model where site-to-site distance is greater than 10km, where spatial correlations vanish. For example, Figure 5.16a shows the observed versus predicted recorded residuals from the 30% training data of the 1987 Whitter Narrows earthquake which only has 15 recordings sparsely distributed in Southern California such that only 6 pairs of data are within the range between 2.5km to 7.5km. Due to the inaccurate correlation model together with less closely-distributed data, the kriging interpolation model has a relatively poor performance as observed from Figure 5.16a and the overall model prediction result is not improved with spatial interpolation as shown from Figure 5.16b.



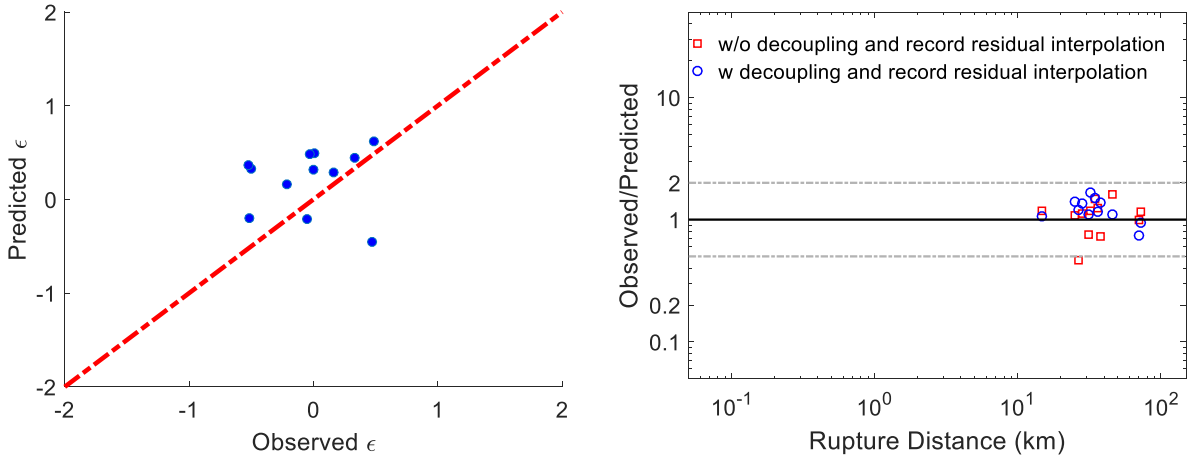
(a) (b)

Figure 5.14 (a) Correlogram of record residual from the adopted generalized cross-building EDP reconstruction model and (b) its fitted exponential semi-variogram model for PSDR and PFA of the 2007 Alum Rock earthquake



(a) (b)

Figure 5.15 (a) Record residual kriging interpolation result and observed/predicted ratio of PSDR using 30/70 percentage training testing split of recorded data from the 2007 Alum Rock earthquake



(a) (b)

Figure 5.16 (a) Record residual kriging interpolation result and observed/predicted ratio of PSDR using 30/70 percentage training testing split of recorded data from the 1987 Whitter Narrows earthquake

Besides retrieving additional data, which is not likely for past events, an alternative Monte Carlo simulation procedure is proposed to generate enough data to construct a representative semi-variogram model for a given historical event that may not have sufficient recordings by simulating within-event and site residuals from previous decoupling procedure in Equation (5-18 to 5-20). The within-event residual sample mean $\eta_{e,i}$ for event i and site residual sample mean $\eta_{s,j}$ for site j are already known as well as within event and site residual sample variance. By assuming that the residuals follow Normal distribution, we can simulate these residuals based on the sample statistics to generate synthetic data for events that lack of actual recordings. The detail procedure is given below.

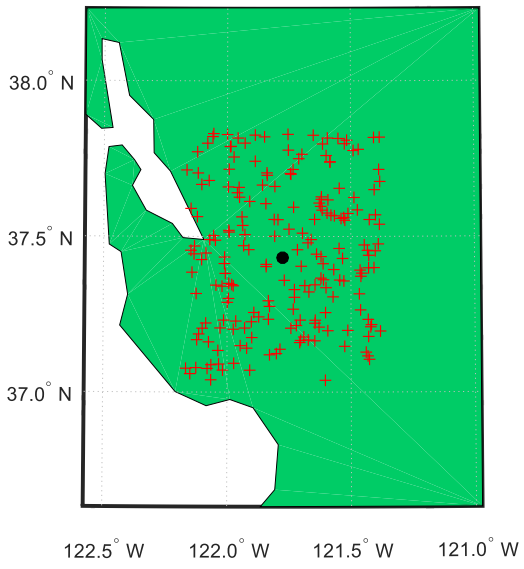


Figure 5.17 Simulated coordinates for 2007 Alum Rock Area earthquake

Step 1: retrieve sample mean of within-event and site residual using Equation (5-19) and (5-20);

Step 2: for the considered event i, random sample from two independent uniform distributions to obtain geo-locations, latitude and longitude respectively, of the simulated

observations. In addition, apply interpolation method to retrieve their corresponding V_{S30} using USGS database of V_{S30} [160] (demonstration using 2007 Alum Rock Area earthquake in Figure 5.17);

Step 3: random sample from a uniform distribution (12ft to 168ft) to obtain simulated building height for each synthetic building;

Step 4: apply the generalized cross-building EDP reconstruction model to retrieve median EDP demand, $\ln(\bar{Y}_{ij})$, for each simulated observation and recover the full observation by adding simulated within-event residual, $\tilde{\eta}_{e,i}$, sampled from $N(\eta_{e,i}, \sigma_{e,i})$ and simulated site residual, $\tilde{\eta}_{s,j}$, from $N(\eta_{e,j}, \sigma_{e,j})$ which yields Equation (5-22);

$$\ln(\tilde{Y}_{ij}) = \ln(\bar{Y}_{ij}) + \tilde{\eta}_{e,i} + \tilde{\eta}_{s,j} \tag{5-22}$$

Step 5: generate empirical semi-variograms for each event using the simulated $\ln(\tilde{Y}_{ij})$.

The above Monte Carlo simulation method is provided as an alternative for historical events that lack sufficient recordings to fit a representative semi-variogram model. It is, however; limited by the assumption that within-event and site residuals follow Normal distribution with the sample variance and mean and additional uncertainties introduced through various simulations of geo-locations, soil properties and building heights.

5.6 Demonstration using Simulation Data

This proposed generalized cross-building EDP reconstruction model is also demonstrated on simulation data. The model calibration procedure described in Equation (5-17) is applied to reconstruct PSDR and PFA median for a collection of seismic building responses generated using NRHA for concrete moment frame buildings subjected to the 1994 Northridge earthquake at 152 sites (Figure 5.18). OpenSees [42] models were used for five different concrete moment frame

buildings with 2, 4, 8, 12, and 20 stories, respectively. Model details can be found in [100]. The two horizontal ground motion acceleration histories are applied to the 2D model and the resultant maximum absolute responses from the two orthogonal directions, PSDR and PFA, along building height profile are shown in Figure 5.19a and Figure 5.19b versus rupture distance. Comparing to Figure 5.5 (recorded building response data subjected to the same 1994 Northridge earthquake event), similar attenuation trend patterns over rupture distance are observed for both EDPs across different height of buildings.

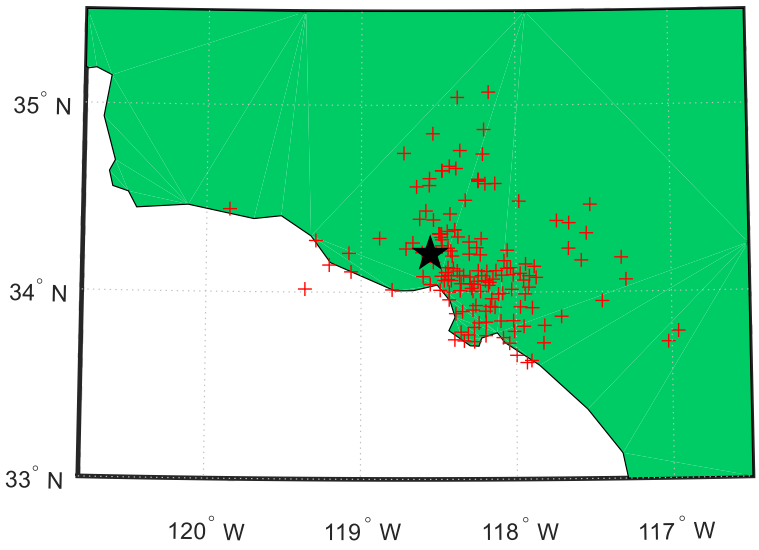


Figure 5.18 Sites map of the 1994 Northridge earthquake

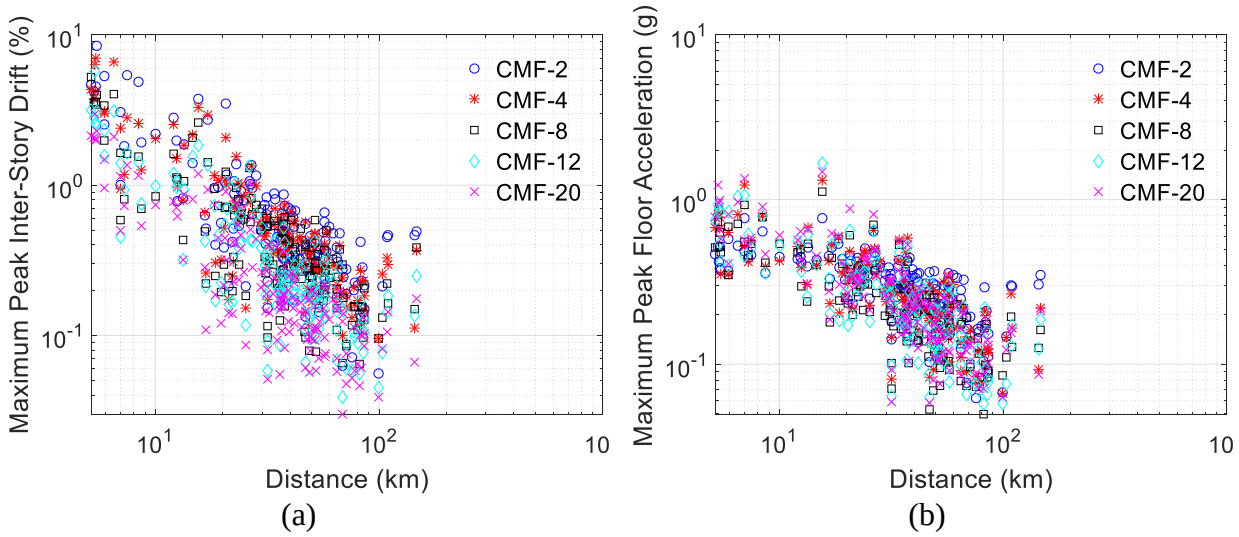


Figure 5.19 (a) PSDR and (b) PFA subjected to the 1994 Northridge earthquake from concrete moment frame model

Table 5.5 Calibrated Adopted Model Parameter for Simulation Dataset of the Northridge earthquake

Adopted EDP Prediction Model	C_1	C_2	C_3
Peak Story Drift Ratio	1.326	-1.5626	1.5457
Peak Floor Acceleration	0.6968	0.2784	0.4775

The same procedure used to develop Equation (5-17) is applied to retrieve model parameters for reconstructing PSDR and PFA median using the simulation dataset summarized in Table 5.5. Comparing to Table 5.4 of the same model parameters using recorded dataset from 26 historical earthquake events, considerable changes are observed for both EDPs, indicating that the model sensitivity is relatively high between different datasets. It should also be mentioned that the simulation data comes from a single event such that the earthquake source term is no longer effective. In addition, the building material and types are limited to concrete moment frames. The

observed-to-prediction ratio is shown in Figure 5.20 and compared to Figure 5.12, where mean ratio for PSDR and PFA is at 1.003 and 0.997 also suggesting unbiased property of the model. As observed in Figure 5.20a, prediction performance of PSDR is among the best as it is bounded by 0.63 and 1.32. On the other hand, there are 1.3% and 5.8% of the entire PFA dataset with over 2 and less than 0.5 observed-to-predicted ratio and majority of which are within 10km rupture distance (Figure 5.20b). Figure 5.21 shows that the observed-to-prediction ratio over rupture distance, building height, magnitude and ASCE-7 empirical period for PFA to visualize specific bias towards each used feature. It can be observed that the high dispersion in observed-to-prediction ratio exists for lower rupture distance, lower building height and empirical period. Similar trend is also observed in the result of recorded dataset in previous section. The result suggests that the proposed model does not fully capture near-epicenter PFA patterns of the buildings, which are subjected to the highest seismic forces and response in the nonlinear range. Force related seismic responses from buildings are expected to be restrained due to yielding at critical components (beam-column joints between first and second floor level for this particular dataset) while deformation related seismic responses, PSDR, are expected to increase as contribution from plastic deformation increases significantly. The lower predictive performance of the PFA model may be caused by greater amount of data focusing at higher rupture distance range in this simulation dataset as opposed to the recorded dataset that shifts the model focus at the linear seismic demand range. This finding indicates that data-driven model performance is highly dependent on appropriate selection of dataset.

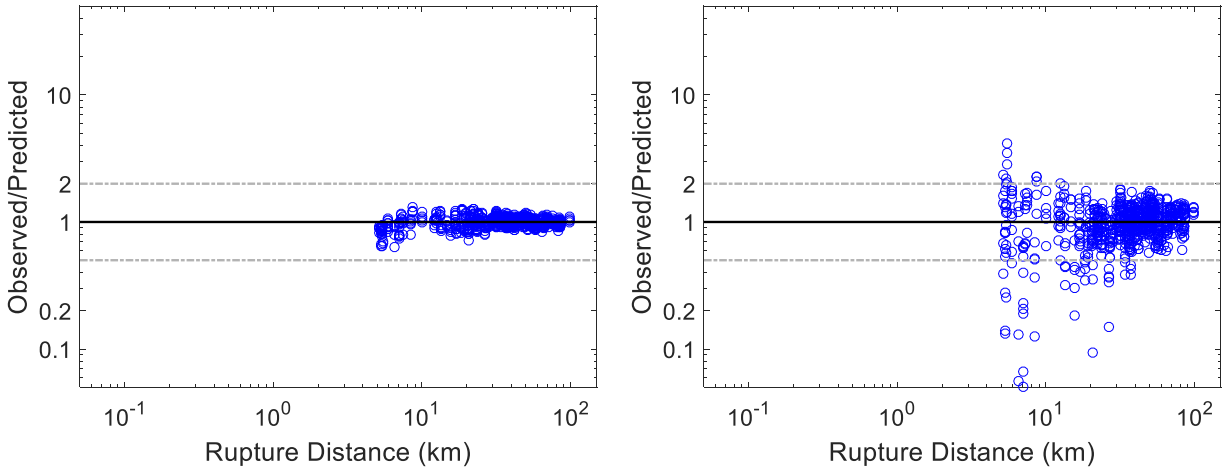
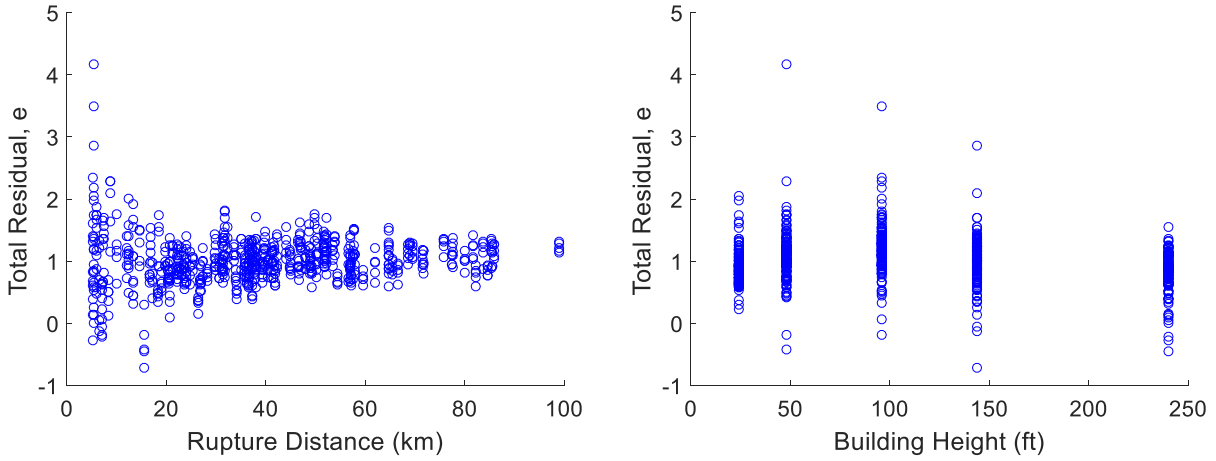


Figure 5.20 Observed-to-prediction ratio versus rupture distance of (a) PSDR and (b) PFA from concrete moment frame models subjected to the 1994 Northridge earthquake



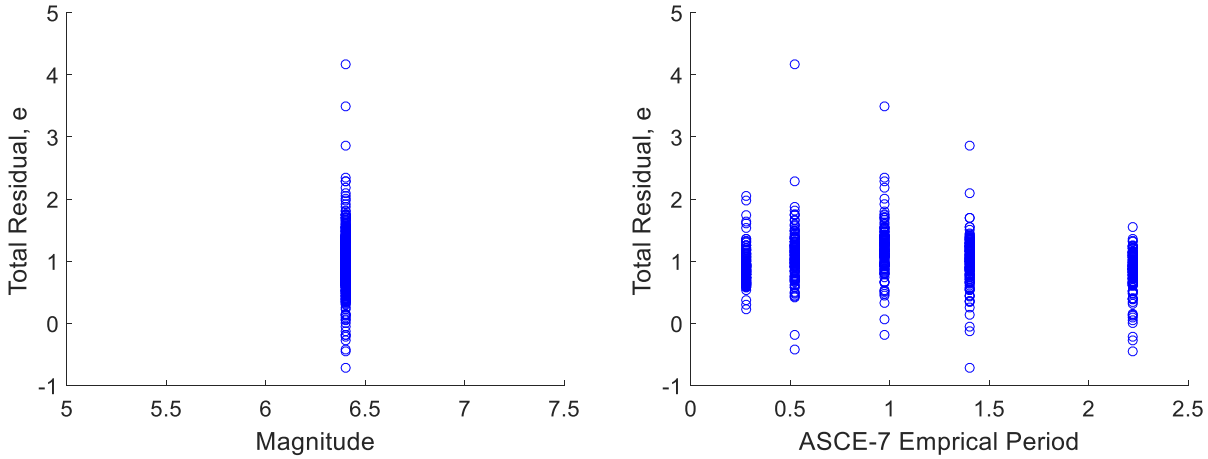


Figure 5.21 Total residual versus (a) rupture distance, (b) building height, (c) magnitude and (d) ASCE-7 empirical period for PSDR of the simulation dataset of the 1994 Northridge earthquake

5.7 Summary

An empirical data-driven model to estimate the median of two critical seismic building responses, PSDR and PFA, is presented. Two versions of the model are included in this study. One follows the design of GMPE and applies a mix-effect model to calibrate model parameters and is referred to as the generalized cross-building EDP reconstruction model. An alternative model, that adopts the predicted median of ground motion intensity and spectra acceleration from a current GMPE to represent event characteristics, intensity attenuation, and site characteristics, is referred to as the adopted generalized cross-building EDP reconstruction model. Both versions contain a building term that incorporates cross-building features, building height, and ASCE-7 empirical period, which includes fundamental building properties into model consideration. A recorded dataset containing building seismic response histories collected from 196 buildings subjected to 26 historical earthquakes within California since 1984 are used to validate the proposed model. In addition, a simulation dataset containing NRHA building seismic responses from 5 representative

concrete moment frame buildings subjected to the 1994 Northridge earthquake is also used to examine the validity of the model for different datasets. The major contribution of the proposed model is that it is the first, generalized empirical data-driven model in the strucutral and earthquake engineering field that considers both cross-event and -building features. In addition, the recorded dataset used in this study is among the largest real-world set of recordings in terms of building seismic responses. The proposed model can be used to estimate median damage related EDPs in a rapid manner after a seismic event to provide vital information for community response and recovery. It can also be used as in probabilistic seismic risk assessment for buildings to provide a median EDP estimation which incorporates additional uncertainties from event, site, and path phenomena.

The two versions of the proposed generalized EDP reconstruction model are different in the consideration of moment scaling, distance function, and site amplification terms. While the regular version is designed based on earlier GMPE models with a relatively simple layout, the adopted version directly uses these terms from modern GMPE which has been carefully calibrated in prior studies and includes a fairly complex formulation. The building term for both versions remains the same. The observed/predicted ratios indicate that the regular version is biased for both EDPs and has more outliers compared to the adopted version, which is perfectly unbiased, most likely due to the prior calibration effort from researchers contributing to the development of GMPEs. Although the adopted version is recommended based on model performance over the recorded dataset, the need for a separate, independent approach to reconstruct EDP median values remains an interesting topic to investigate.

In order to fully reconstruct the EDPs, the total residuals, subtracting the predicted median using the proposed model from actual EDPs, are further evaluated by decoupling event and site

residuals to generate record residuals, which are predicted using the kriging interpolation method developed by Sun et. al. in [28]. Results show that the generalized cross-building EDP reconstruction model, together with decoupling procedures, removes trends corresponding to sites with more than 10km inter-site distance. The kriging interpolation procedure is applied to reconstruct the total residuals using the Alum Rock earthquake. Comparing results from models with and without kriging interpolation demonstrates that the kriging interpolation procedure reduces variance of the observed/predicted ratio to improve reconstruction performance. However, it should be noted that this is not guaranteed for all event cases due to lack of recorded data within an event resulting less representative correlation patterns being captured in the kriging interpolation. A Monte Carlo simulation procedure is proposed to generate synthetic data as an alternative to retrieve more reliable correlation patterns.

A simulation dataset from 5 representative concrete moment frame buildings subjected to ground motions recorded in the Northridge earthquake is used to demonstrate the adaptability of the proposed model for a different dataset. The adopted generalized cross-building EDP reconstruction model is observed to perform the best for PSDR among all datasets but the worst for PFA. The large discrepancy shown in the simulation dataset suggests that the prediction performance of the model is highly dependent on appropriate selection of dataset.

There are several limitations in this proposed generalized model which can be improved through the following aspects. First, the recorded dataset is processed based on the assumption that torsional responses from buildings are negligible, which can be improved by investigating peripheral channels at available floors to correct channel data from unavailable floors. Second, the current empirical period feature is retrieved based on ASCE-7 estimation and is systematically biased towards empirically stiffer buildings compared to actual building responses period. Future

studies could focus on investigating each considered building according to their strucutral type and apply system identification methods to retrieve their actual building response period from measured data. In addition, the retrieval of V_{S30} in this current model is through 1-nearest neighbor approach based on a provided map from USGS with resolution at 250m. Due to the high variation observed in V_{S30}, it would be more reliable to retrieve V_{S30} at each considered site using their exact site geology and survey data. The residual decoupling and interpolation procedure could also be improved by applying Bayesian approach to retrieve within-event and site residuals given the small size of current dataset rather than sample mean, which will be demonstrated in next chapter.

6. Development of A Bayesian Hierarchical Model for Within-event Residual using Recorded Seismic Building Responses

6.1 Introduction

The Intensity Measure (IM) at a site subjected to an earthquake, such as Spectra Acceleration (SA) and Peak Ground Acceleration (PGA), is a critical quantitative measure that is related to the intensity of ground shaking and the corresponding seismic risk to buildings at the site. Hence, estimating IMs is an essential part of seismic risk assessment for individual structures, a cluster of buildings, a city, or a region. However, earthquakes are a sequence of extremely complex physical phenomena, resulting from energy release associated with movement of tectonic plates that generates consequential seismic wave propagation along rock layers which then travels through soil to the ground and becomes site shaking. Due to the numerous physical systems involved interacting with each other at such large scale, earthquakes are extremely stochastic and complicated, and therefore, it is difficult and impractical to develop physics-based models to estimate IMs throughout the entire process. Direct measurement of IMs, e.g., using ground motion stations, is ideal but requires too many sensors to be practical. Hence, data-driven models are often preferred and most commonly used in IM estimations across seismology and earthquake engineering, such as the widely used Ground Motion Prediction Equation model [149,151,159].

Engineering Demand Parameters (EDPs) derived from seismic responses of buildings are building specific IMs that contain strucutral information from the buildings and are directly related to building damage caused by earthquakes. In Chapter 5, a generalized cross-building EDP reconstruction model is developed to estimate two major types of system level response EDPs, Peak Floor Acceleration (PFA) and Peak Story Drift Ratio (PSDR), of a building at any site within California subjected to an event. A decoupling procedure to separate within-event residual is also

demonstrated using a simplified Frequentist's approach, sample mean of all data within the same event (Equation 5-19). However due to the relatively small amount of recordings available for a number of events considered, e.g., 4 recordings from the 1989 Loma Prieta earthquake, 2 recordings from the 1992 Big Bear earthquake and extreme cases such as only 1 recording available from the 2010 Borrego earthquake, using sample mean to estimate population mean is obviously not ideal.

In this chapter, a Bayesian hierarchical model is proposed to address the estimation bias using sample mean of unrepresentative samples to decouple within-event residual from the total residual. The Bayesian hierarchical model contains three layers, the lowest layer for modeling within-event individual EDPs, the middle layer for modeling parameters of each earthquake event model in the lowest layer, and the top layer for modeling parameters of the middle layer. Within-event residuals are considered as a series of model parameters (distribution means of the lowest layer within-event models) that can be retrieved through integrating with the observed data (total residual in our case) using Bayes' theorem. The main advantage to apply this hierarchical structure is that it incorporates uncertainties from all available data to reduce biased estimation caused by smaller number of samples as compared to empirical sample mean of each event as used in Chapter 5. It is also beneficial to understand the earthquake mechanism as a multi-parameter problem in the data-driven model domain to provide insight into the event effect on the total residual and its contribution to the overall variation in EDPs.

6.2 Data

The details of the EDP data are presented in Section 5.2, which include 371 pairs of observed total residual of EDPs, PSDR and PFA recorded in 188 buildings subjected to 24 historical earthquakes from 1984 to 2018 within California. The EDP data are used to develop the

proposed Bayesian hierarchical model. The logarithm total residual e , is calculated by subtracting the prediction median $\hat{y}$, which is given by the generalized cross-building EDP reconstruction model, by the observed EDP responses y as shown in Equation (6-1). From this point, the term logarithm will not be used throughout this chapter to avoid repetitive descriptions.

$$e = \ln(\hat{y}) - \ln(y) \tag{6-1}$$

Histograms of the PSDR and PFA total residual across all earthquake events are shown in Figure 6.1. It is observed that PSDR total residual distributes more 'normal-like' compared to PFA. Medians of both EDP total residual reside at around zero indicating that both datasets are not biased. The highest population bin for both EDP total residual is slightly higher than zero suggesting that the within-event residual of majority events most likely is greater than zero. To better visualize the total residual according to events, selected event specific histograms are shown in Figure 6.2 and Figure 6.3. For events that include more than 20 data points, such as the 1994 Northridge earthquake (Figure 6.2a and Figure 6.3a), uni-modality is observed from the histogram suggesting that sample mean is a reasonable estimation of within-event residual. On the other hand, events that lack sufficient data points, such as the 1992 Landers earthquake in Figure 6.3c and Figure 6.3c, no clear mode can be retrieved from the histogram and therefore the sample mean might be a biased estimation of the within-event residual. This observation is the primary motivation for applying Bayesian hierarchical model to get within-event residual.

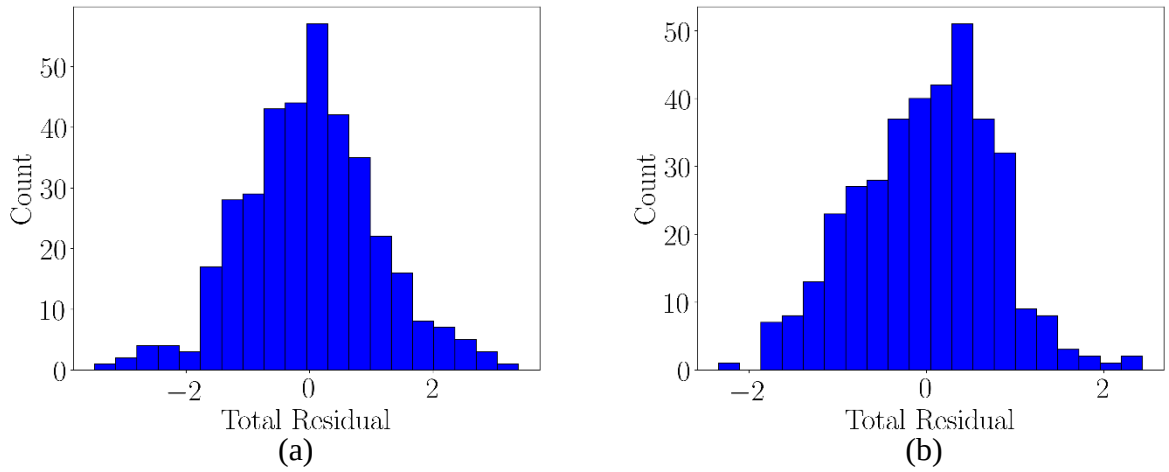
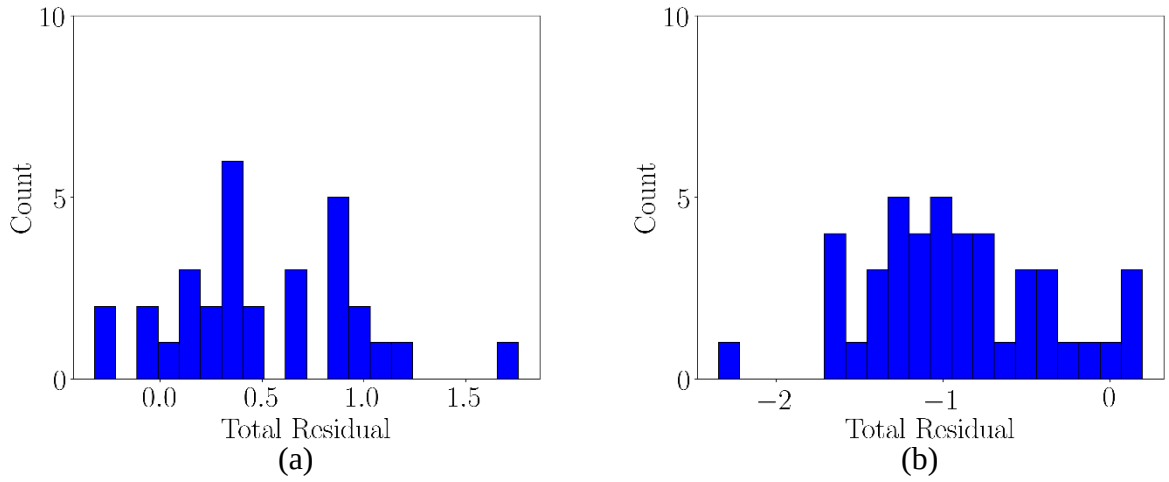


Figure 6.1 Histograms of the (a) PSDR and (b) PFA total residual across all earthquake events



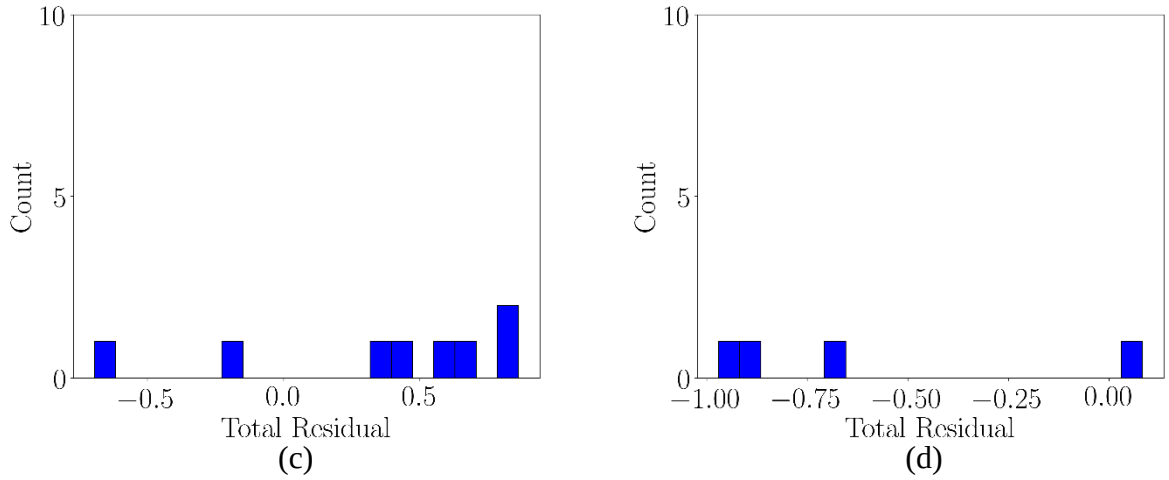
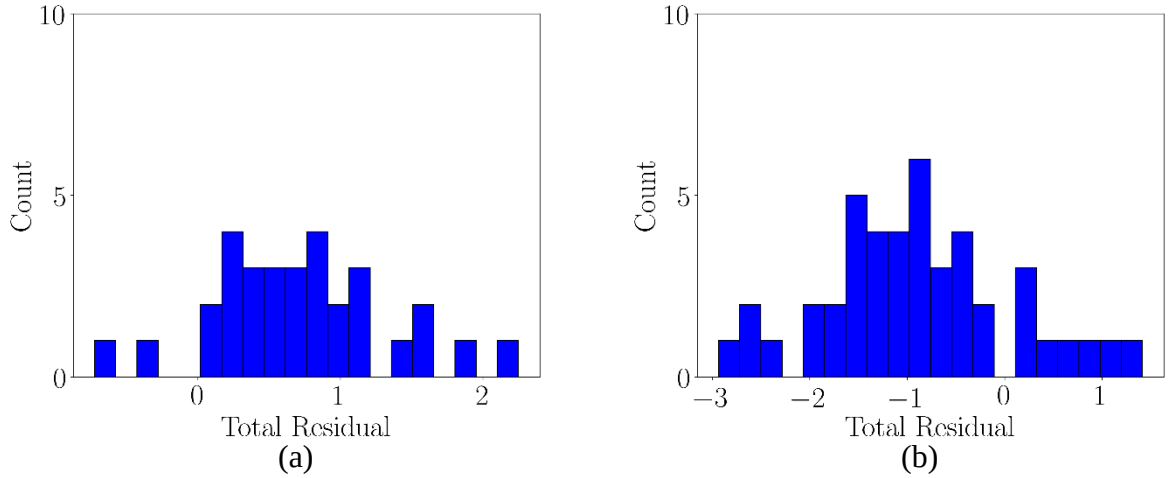


Figure 6.2 Histograms of PFA total residual of (a) the 1994 Northridge earthquake, (b) the 2014 South Napa earthquake, (c) the 1992 Landers earthquake and (d) the 2003 Big Bear City earthquake



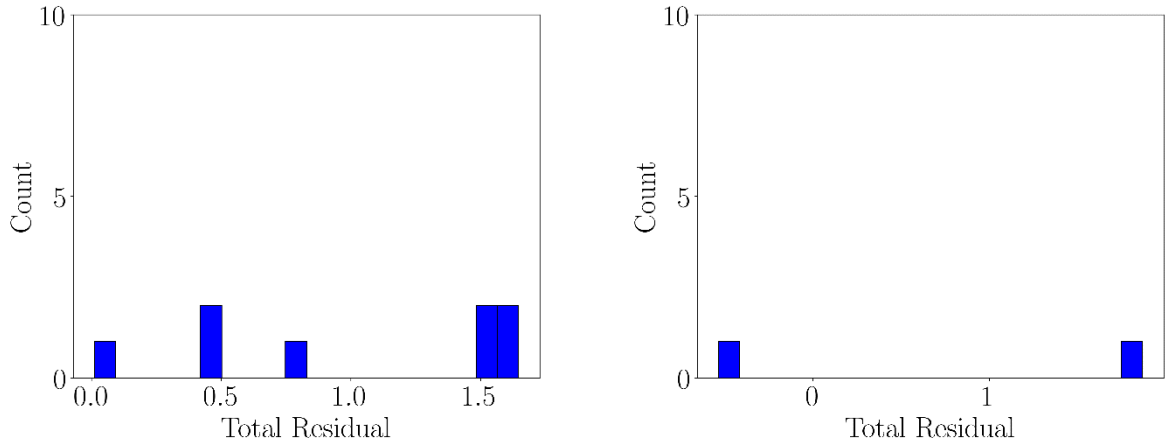


Figure 6.3 Histograms of PSDR total residual of (a) the 1994 Northridge earthquake, (b) the 2014 South Napa earthquake, (c) the 1992 Landers earthquake and (d) the 2003 Big Bear earthquake

6.3 Bayesian Hierarchical Model

6.3.1 Statistical Model Layout

As previously noted, there are three layers in the proposed Bayesian hierarchical model for the total residual dataset (Figure 6.4). The top layer, also known as hyperprior model, is a hyperparameter level represented as $p(\mu,\tau^2)$, that models the distribution of the hyperparameters used in the middle layer. The middle layer, known as the prior model, is a distribution that models the mean parameters, the within-event residuals $\eta_{e,i}$ donated in Chapter 5, of each within-event model in the bottom layer. The bottom layer, known as the likelihood model, consists of individual models for each event that the total residuals, e, is sampled from. The final layout of the proposed Bayesian Hierarchical model can be formulated in Equation (6-2).

$$p(\eta_{e\bullet},\mu,\tau^2,e\,|\,\sigma_{e\bullet}{}^2)\propto\prod_i p(\bar{e}_{i\bullet}\,|\,\eta_{e,i},\sigma_{e,i}^2)\prod_i p(\eta_{e,i}\,|\,\mu,\tau^2)p(\mu,\tau^2) \qquad (6\text{-}2)$$

Where $p(\mu,\tau^2)$ is the hyperprior model with hyperparameters, μ and τ^2, respectively. $p(\eta_{e,i}\mid\mu,\tau^2)$ represents the prior model for the within-event residual, $\eta_{e,i}$ of event i. $p(\bar{e}_{i\bullet}\mid\eta_{e,i},\sigma^2_{e,i})$ is the individual likelihood model for the total residual e_{ij} from event i at site j, and $\bar{e}_{i\bullet}$ represents the sample mean of event i, or within-event residual sample mean. The product at both the prior and likelihood model levels indicate that they are independently sampled at their corresponding level. By applying the Bayes' rules, the proposed Bayesian hierarchical model is a product of the above three represented models in Equation (6-2) as $p(\eta_{e\bullet},\mu,\tau^2,e\mid\sigma^2_{e\bullet})$.

Furthermore, the middle layer is assumed to be a Gaussian distribution model for the within-even residuals since it is theoretically difficult to infer a particular distribution for the middle layer without specifically knowing it from a separate source. A Gaussian distribution is selected mainly due to its highly adjustable skewness properties to simulate almost any type of unimodality distributions. As discussed before, Figure 6.2a and 6.2b as well as Figure 6.3a and 6.3b suggest that a Gaussian distribution fits the logarithm within-event total residual distribution well (bottom layer). Jayaram and Baker rigorously evaluated and showed the univariate normality property of total residual for spectra acceleration [161] using various methods, including the Q-Q statistical test, which is adopted with a simplified evaluation here. Figure 6.5 provides the Q-Q statistical plots for example event, the 1994 Northridge earthquake data, of both PSDR (Figure 6.5a) and PFA (Figure 6.5b), which correspond to the within-event total residual data visualized in Figure 6.3a and Figure 6.2a, respectively. Q-Q plot is a normality indicator for the applied data, the closer the blue dots fit to the red theoretical line, the more likely the data can be represented by a Gaussian distribution. It is observed that the distribution of within-event total residual for the 1994 Northridge earthquake, as well as other events that contain sufficient data points not shown here, is well-represented by a normal distribution, which validates the assumption of using a

Gaussian distribution at the bottom layer. Across the bottom layer, each earthquake event is assumed to be independent because their epicenters are different and thus, the source term and path term remaining in the total residual are not correlated outside the event considered. However, this does not suggest that they are totally independent from each other, as they are correlated through the upper layers. In addition, it should be noted that some observations from different events are retrieved at the same site; therefore, these observations are correlated through the site term remaining in the total residual. To properly account for this, one additional layer for site models could be applied. This additional layer is not included in current study due to insufficient data from the 188 considered sites (buildings) for the current dataset. With the above assumptions, the complete version of Equation (6-2) is given as Equation (6-3) and shown in Figure 6.4.

$$p(\eta_e, \mu, \tau^2, e \mid {\sigma_{e\bullet}}^2) \propto \prod_i N(\bar{e}_{i\bullet} \mid \eta_{e,i}, \sigma_{e,i}^2) \prod_i N(\eta_{e,i} \mid \mu, \tau^2) p(\mu, \tau^2) \qquad (6\text{-}3)$$

It can be shown that $\sigma_{e,i}^2$, variance of the prior model, appears on the condition side indicating that it is known to us. In addition, the hyperprior model, $p(\mu, \tau^2)$, has not been specified yet. To finalize the complete conditionals of this model, additional assumptions are introduced in the following two sections regarding hyperprior model choice and variance consideration of the prior model.

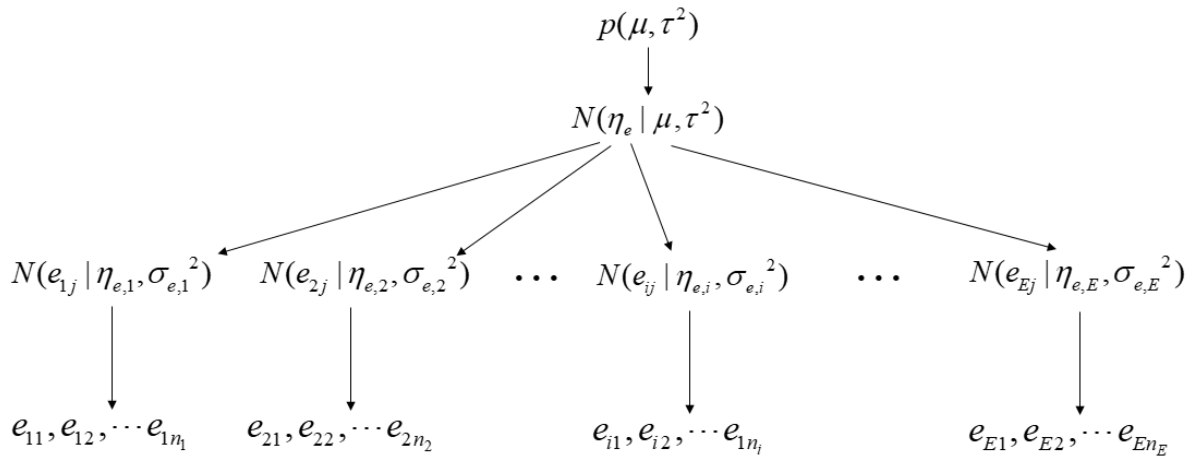
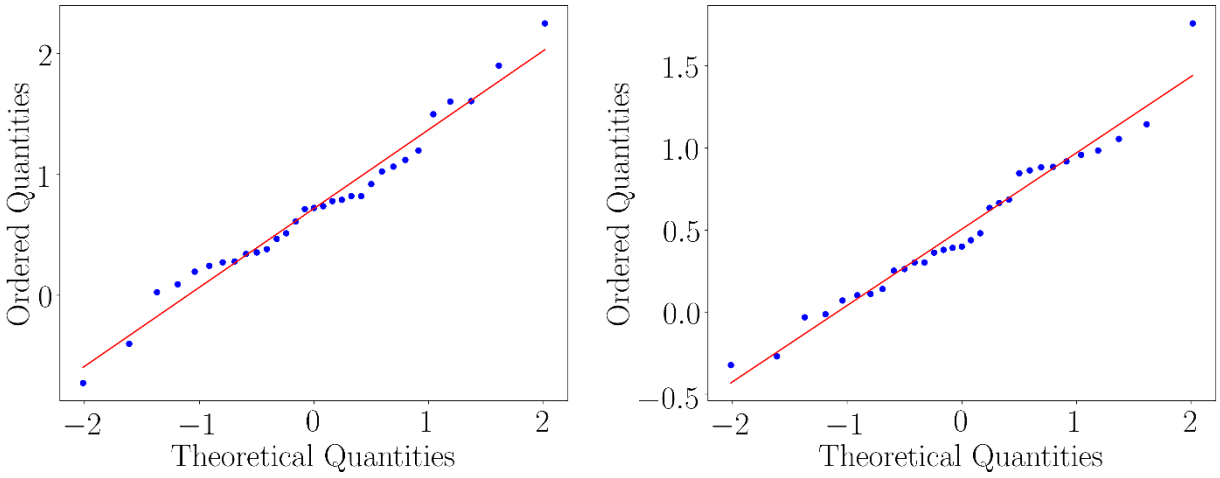


Figure 6.4 Proposed Bayesian hierarchical model layout



(a) (b)

Figure 6.5 Q-Q plot for (a) PSDR and (b) PFA total residual of the 1994 Northridge earthquake

6.3.2 Hyperprior Choice

One key step of the Bayesian hierarchical model is to determine which hyperprior model to use, namely, $p(\mu,\tau^2)$, which represents the prior knowledge of the distribution of hyperparameters, μ and τ^2, mean and variance of the Gaussian model of the within-event residual mean parameter, η_e. The choice of hyperprior model reflects restrictions on the prior model at the

middle layer. Given the small size and unbalanced dataset of our case, it is intended to apply a non-informative hyperprior model for the top layer such that no additional information is brought into the Bayesian hierarchical model besides the dataset used. In other words, the model parameters are only determined using the dataset. A number of justifications over using non-informative hyperprior for Bayesian hierarchical models have been proposed in the past including [162,163] and selection of non-informative hyperprior models are widely discussed such as [164,165]. A suitable choice of non-informative hyperprior that matches the dataset used in this study is an Inverse Gamma distribution for τ^2 with parameter α and β to adjust the model to approximate the widely used noninformative Jeffreys prior ($p(\tau^2) \propto \frac{1}{\tau^2}$), which essentially is a non-informative distribution for the hyperprior parameter τ^2. In addition, it is typical to further assume that the μ and τ^2 within hyperprior model are independent from each other such that the hyperprior model can be written as:

$$p(\mu, \tau^2) \propto p(\mu \mid \tau^2) p(\tau^2) \propto IG(\tau^2 \mid \alpha, \beta) \tag{6-3}$$

Where α and β are chosen to be 0.01 to reflect that a noninformative hyperprior model is used, as suggested in [166]. This setup enables applying simulation to retrieve model parameters of the proposed Bayesian hierarchical model using Markov Chain Monte Carlo (MCMC), which is discussed in the following sections.

6.3.3 Variance Consideration of Within-event Model

The variance term $\sigma_{e,i}^2$ of each likelihood model at the bottom layer is assumed known in the proposed Bayesian hierarchical model which, to a certain extent, enables some level of control over the dispersion of the individual event model at the bottom layer. As mentioned before, one of the most important properties of our dataset is that the amount of observations in each event varies

significantly from very few (2 or 3) to quite a lot (50s to 60s), indicating that the within-event variance of the dataset is very diverse. In order to properly take this into consideration, $\sigma_{e,i}^2$ is assigned to be proportional to the total variance of entire dataset ($\sigma_e{}^2$) scaled by number of observations in that particular event, n_i , as shown in Equation (6-4).

$$\sigma_{e,i}^2 = \frac{\sigma_e^2}{n_i} \tag{6-4}$$

6.4 Markov Chain Monte Carlo Simulation and Gibbs Sampler for Posterior Inference

6.4.1 Simulation Setup

With the Bayesian hierarchical model setup, the goal is to estimate all relevant model parameters, namely, μ, τ^2 and $\eta_{e\bullet}$. In other words, we wish to learn about the joint distribution $\pi(\mu, \tau^2, \eta_{e\bullet})$ with the dataset e and within-event variance $\sigma_{e\bullet}^2$ available, mathematically, $p(\eta_{e\bullet}, \mu, \tau^2 \mid \sigma_{e\bullet}^2, e)$, which is a posterior of the proposed Bayesian hierarchical model $p(\eta_{e\bullet}, \mu, \tau^2, e \mid \sigma_{e\bullet}^2)$ by integrating the dataset. At this point, it becomes convenient to estimate these model parameters through MCMC. We start by defining a stationary Markov chain $C_0, C_1, C_2, \cdots C_M$ with the states $C_m = (\mu^m, (\tau^2)^m, (\eta_e)^m)$ at state m. The simulation continues until reaching a state M such that C_M converges to a stationary $\pi(\mu, \tau^2, \eta_{e\bullet} \mid e, \sigma_{e\bullet}{}^2)$ distribution according to [167] which can be achieved by using the Gibbs sampler [168]. The Gibbs sampler constructs the Markov chain by repeatedly sampling from the complete conditionals specified by updated states of each of the involved model parameter as below.

$$\mu^m \sim p(\mu \mid (\tau^2)^{m-1}, \eta_{e\bullet}^{m-1}, \sigma_{e\bullet}^2, e) \tag{6-5}$$

$$(\tau^2)^m \sim p(\mu \mid \mu^m, \eta_{e\bullet}^{m-1}, \sigma_{e\bullet}^2, e) \tag{6-6}$$

And for each event i, we have:

$$\eta_{e,i}^{m} \sim p(\mu \mid \mu^{m}, (\tau^{2})^{m}, \sigma_{e\bullet}^{2}, e) \tag{6-7}$$

The above Gibbs sampler simulation continues until the convergence criteria are met at step M. The following equations list all specific complete conditionals with respect to Equation (6-5) to (6-7).

$$p(\mu \mid rest) \propto \prod_{i} N\left(\eta_{i} \mid \mu, \frac{\tau^{2}}{N_{e}}\right) p\left(\mu, \tau^{2}\right) \tag{6-8}$$

$$p(\tau^{2} \mid rest) \propto Inv-Gamma\left(\tau^{2} \mid \frac{N_{e}}{2} - 1, \frac{1}{2}\sum_{i}(\eta_{e,i} - \mu)^{2}\right) p(\mu, \tau^{2}) \tag{6-9}$$

$$p(\eta_{e,i} \mid rest) \propto N\left(\eta_{e,i} \mid \frac{\bar{e}_{i\bullet} / \sigma^{2}_{e,i} + \mu / \tau^{2}}{1/\sigma^{2}_{e,i} + 1/\tau^{2}}, \frac{1}{1/\sigma^{2}_{e,i} + 1/\tau^{2}}\right) \tag{6-10}$$

The derivation of the above complete conditionals can be found from [169]. By applying the hyperprior defined in Equation (6-4), the Markov chain is formulated, and its result will be discussed in next section. It should be noted that a direct simulation through the joint posterior $p(\eta_{e\bullet}, \mu, \tau^{2} \mid \sigma_{e\bullet}^{2}, e)$ in this case is also viable due to theoretical closed-form solution available in this case, a Bayesian hierarchical Normal model. The essential difference between direct simulation through the joint posterior and Gibbs sampler simulation through the complete conditionals is that the former requires drawing samples from much higher dimensional models (given the number of events in our case: 24), e.g., $\mu \sim p(\mu \mid \sigma_{e\bullet}^{2}, e)$ as opposed to $\mu \sim p(\mu \mid \tau^{2}, \eta_{e\bullet}, \sigma_{e\bullet}^{2}, e)$ in the latter case. This results in substantially improved convergence, whereas convergence may not be achieved using the Markov chain for the direct simulation in a reasonable time period. On the other hand, due to sampling from dependent models (Gibbs sampler) rather than *i.i.d.* models (direct simulation), the result from Gibbs sampler requires longer

sequences for integration to estimate the model parameters. Still, the Gibbs sampler approach is much more computationally efficient and stable compared to direct sampling in our case.

6.4.2 Convergence Analysis

To validate convergence has been met in the simulation, three Markov chains are simulated independently with 100 Monte Carlo steps initialized at randomly generated states. The estimated value of any model parameter (θ) at step M is calculated by Equation (6-11).

$$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\theta^{m} \tag{6-11}$$

Using Equation (6-11), the simulation result of estimated μ and τ^2 of PSDR in Figure 6.6 shows that the three randomly initialized chains of both hyperparameter approach to the same values, μ to around 0 and τ^2 to 0.51 suggesting good convergence at step 100. Figure 6.6a also shows that the prior mean of all events converges to 0 suggesting an unbiased within-event residual at the middle layer. This further validates that the generalized cross-building EDP reconstruction model in Chapter 5 is unbiased for PSDR. Besides the trend plots of individual model parameter, a more rigorous technique, Gelman and Rubin Diagnostics ([170,171]), is applied to validate convergence of Markov chains through examining their posterior marginal variance $\text{var}(\theta \mid e)$, defined as $\hat{\sigma}^2$. Let the set of C chains to be $\{MC_c : c = 1, 2, ...C\}$, the within-chain mean of a particular parameter θ of chain c be ψ_c and the mean of θ over all chains be $\bar{\psi}$. The between- and within-chain variance, σ_b^2 and σ_w^2 can be calculated using:

$$\sigma_w^2 = \frac{1}{C}\sum_{c=1}^{C}\sigma_c^2 \text{, where}$$

$$\sigma_c^2 = \frac{1}{M}\sum_{m=1}^{M}\left(\theta_c(m) - \psi_c(m)\right)^2 \tag{6-12}$$

$$\sigma_b^2 = \frac{M}{C-1}\sum_{c=1}^{C}(\psi_c - \bar{\psi})^2 \tag{6-13}$$

With that, the corresponding $\hat{\sigma}$ of posterior marginal variance can be estimated as a weighted sum using between- and within-chain variance:

$$\hat{\sigma} = \frac{M-1}{M}\sigma_w + \frac{C+1}{MC}\sigma_b \tag{6-14}$$

From Figure 6.7 using μ as an example, it is observed that the trend of posterior marginal variance $\hat{\sigma}^2$ and between-chain variance σ_w^2 approach each other after 15 Monte Carlo steps and remains almost equal beyond this. According to Gelman et. al. [170], the posterior variance estimate $\hat{\sigma}^2$ should be very close to the within-chain variance σ_w^2 if all C chains have reached the target distribution, which suggests that convergence of all chains has been achieved for the considered model parameter, in our case, μ. In addition, the autocorrelations among the sequence of Monte Carlo samples of each chain shown in Figure 6.8 are considerably small (samples are close to *i.i.d*. from each other within the chain), which also validates the convergence of the proposed MCMC method for μ of PSDR total residual. Autocorrelation is observed to drop significantly from 1 at lag 0 to 0.1 at lag 2 for all three chains. Similar findings are observed for all other model parameters for both PSDR and PFA total residual data. Therefore, it can be concluded that the proposed MCMC has converged in the simulation.

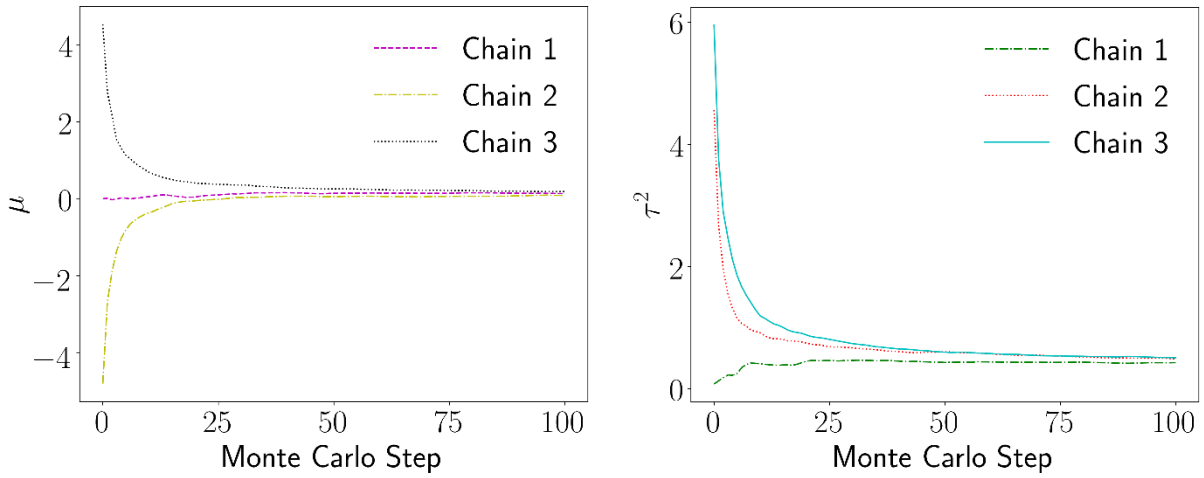


Figure 6.6 Simulation result of μ and τ^2 of PSDR

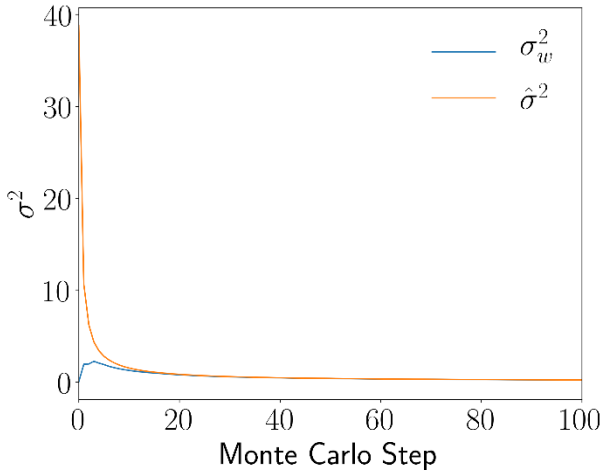


Figure 6.7 Trend of the within-chain and the posterior marginal variance of μ ($\mathrm{var}(\mu \mid e)$) from the Markov chain simulation of PSDR

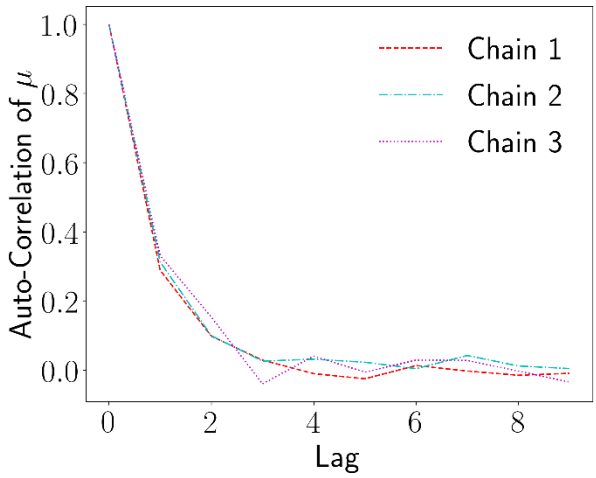


Figure 6.8 Autocorrelation of μ from the Markov chain simulation of PSDR

6.5 Discussion over Within-event Residual Results

Results using MCMC to estimate within-event residual in the proposed Bayesian hierarchical model are shown in Figure 6.9 and Figure 6.10 for PSDR and PFA, respectively. Both start with randomly generated states and gradually converge after 70 Monte Carlo steps simulated.

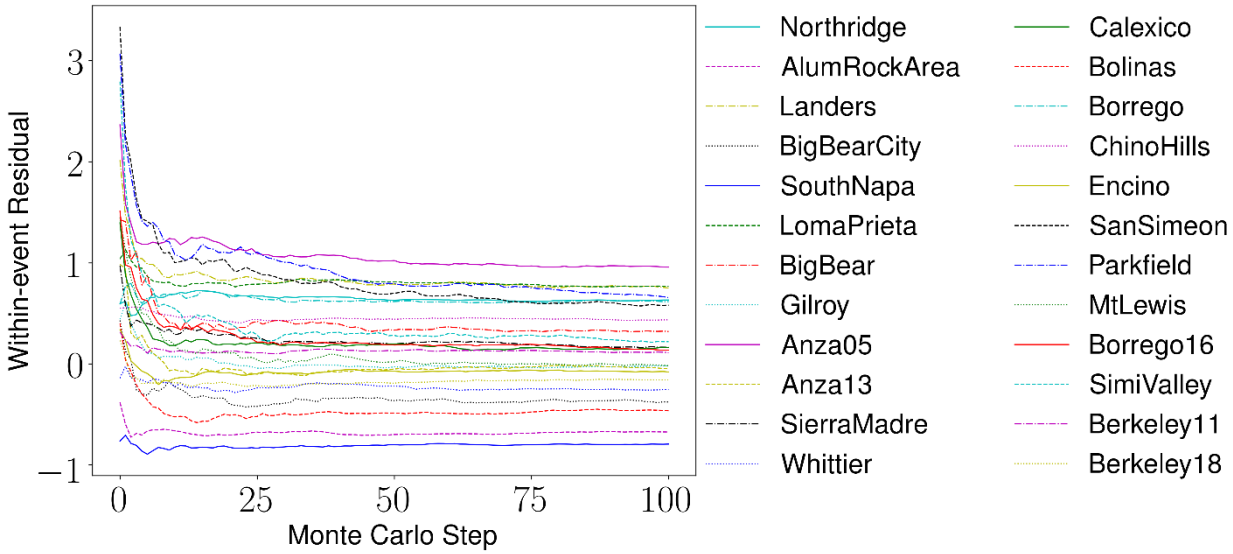


Figure 6.9 Trend of the Within-event residuals of all earthquakes from the Markov chain simulation of PSDR

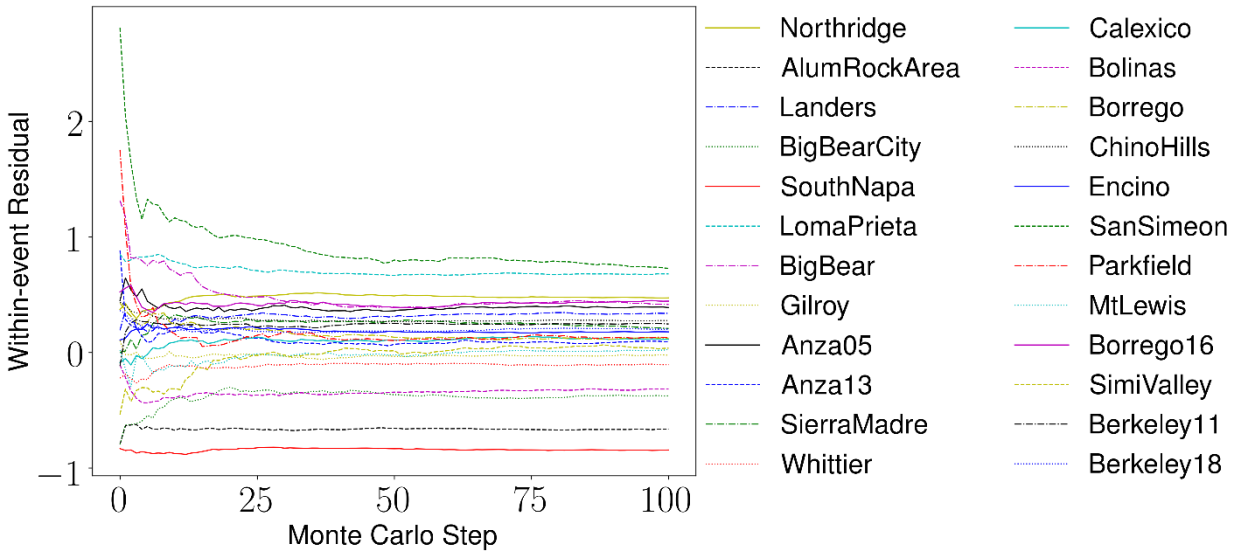


Figure 6.10 Trend of the Within-event residuals of all earthquakes from the Markov chain simulation of PFA

A summary of within-event residual after convergence is provided in Table 6.1 for PSDR and Table 6.2 for PFA where the deviation ratio is calculated by the absolute difference between within-event residual from MCMC and from sample mean divided by that from MCMC. Table 6.1 shows that the deviation ratio of the within-event residual from MCMC and that from the sample mean is within 30%, except for the event Berkeley18 for PSDR, which implies that the two methods are generally close with higher within-event observations. However, it does not necessarily hold for events with relatively large number of recordings. For example, the deviation ratio for the event Berkeley11 is 125.1% with 22 observations available, which ranks sixth among all events in current dataset. Some events with fewer number of recordings, such as Anza13 and Borrego16, are observed with small deviation ratios. However, the deviation ratio of most events with very few observations, such as 1 for the event SanSimeon and Parkfield, 2 for the event BigBear, is extremely large ranging from 100% to 8000%, suggesting that sample mean is not a good measure for these events. Table 6.2 shows similar findings for PFA, except the deviation

ratios are restrained with the highest one at 186.4% of the event SanSimeon with only 1 observation available. The much lower discrepancy between MCMC and sample mean for within-event residual of PFA suggests that the total residual of PFA within an earthquake event has less variation compared to that of PSDR, which is also observed in Chapter 5. It should be noted that these findings are based on use of the current small dataset and may not indicate the actual statistical properties of a larger recorded building response dataset. Due to the additional assumptions made to process the raw data as mentioned in Section 5.2, these findings may not be accurate. Nonetheless, MCMC estimation of within-event residual is demonstrated with much higher confidence bound as compared to sample mean in terms of resilience to uncertainty caused by insufficient amount of recordings within events.

Table 6.1 Summary of Within-event residual of PSDR

Event Name	Within-event Residual (MCMC)	Within-event Residual (Sample Mean)	Deviation Ratio (%)	No. of Instrumented Buildings
Northridge	0.680	0.717	5.472	31
AlumRockArea	-0.788	-0.724	8.135	43
Landers	0.28	1.006	258.711	8
BigBearCity	-0.514	-0.852	65.793	4
SouthNapa	-0.646	-0.881	36.37	44
LomaPrieta	0.661	0.828	25.331	24
BigBear	-0.008	0.667	8638.713	2
Gilroy	0.126	-0.053	141.923	16
Anza05	0.787	1.512	92.122	5
Anza13	-0.209	-0.201	3.746	4
SierraMadre	-0.038	0.147	487.193	8
Whittier	-0.274	-0.356	29.758	16
Calexico	-0.055	0.186	439.322	7
Bolinas	-0.642	-0.649	1.131	12
Borrego	0.646	0.720	11.418	11
ChinoHills	0.523	0.491	6.289	48
Encino	-0.53	-0.114	78.489	22
SanSimeon	0.864	2.019	133.788	1
Parkfield	-0.285	2.043	816.776	1
MtLewis	-0.224	-0.08	64.139	3

Borrego16	0.297	0.216	27.271	7
SimiValley	0.194	0.419	115.763	1
Berkeley11	-0.315	0.079	125.103	22
Berkeley18	-0.026	-0.233	784.372	31

Table 6.2 Summary of Within-event residual of PFA

Event Name	Within-event Residual (MCMC)	Within-event Residual (Sample Mean)	Deviation Ratio (%)	No. of Instrumented Buildings
Northridge	0.606	0.505	10.173	31
AlumRockArea	-0.584	-0.719	13.512	43
Landers	0.397	0.361	3.650	8
BigBearCity	-0.238	-0.611	37.258	4
SouthNapa	-0.663	-0.924	26.120	44
LomaPrieta	0.648	0.731	8.246	24
BigBear	-0.027	0.689	71.603	2
Gilroy	-0.194	-0.057	13.754	16
Anza05	0.035	0.436	40.057	5
Anza13	-0.094	0.048	14.209	4
SierraMadre	0.061	0.260	19.918	8
Whittier	-0.073	-0.156	8.230	16
Calexico	0.181	0.116	6.462	7
Bolinas	-0.207	-0.425	21.803	12
Borrego	0.219	0.136	8.376	11
ChinoHills	0.271	0.289	1.836	48
Encino	0.079	0.191	11.256	22
SanSimeon	0.328	2.191	186.372	1
Parkfield	-0.107	0.199	30.54	1
MtLewis	0.375	-0.158	53.284	3

Borrego16	0.547	0.551	0.420	7
SimiValley	0.238	-0.058	29.684	1
Berkeley11	0.054	0.254	20.014	22
Berkeley18	0.332	0.216	11.637	31

6.6 Summary

This chapter extends the generalized cross-building EDP reconstruction model in Chapter 5 to further evaluate the within-event residual of two system level EDPs, PSDR and PFA, respectively using a total of 371 recorded building seismic response data consisting of seismic responses of 188 buildings in California subjected to 24 historical earthquakes. As opposed to retrieving within-event residual through taking sample mean of recordings in each event, as is done in Chapter 5, a Bayesian hierarchical model consisting of three layers that represent a data-driven structure for the building response data is proposed to estimate the within-event residual to produce a model that is more robust for cases where insufficient recordings exist for some events within the dataset. A few assumptions are made to construct the Bayesian estimation. A noninformative hyperprior through an Inverse Gamma distribution is used such that the model parameters are purely based on information contained in the dataset. A fixed variance is assigned for each event Gaussian distribution model which allows flexibility in modeling variance at the middle layer.

The model parameters are calibrated by a MCMC simulation process using the Gibbs sampler through complete conditionals of the proposed model. The major advantage of this approach is that Gibbs sampler simulates the model in a much lower dimensional space as compared to direct sampling. In addition, convergence of this simulation is validated through three separate means, visualizing the model parameter trend over Monte Carlo steps, examining the trend of posterior marginal variance and within-chain variance, and analyzing autocorrelations along the Markov chain.

Results of the within-event residuals are visualized and summarized in Section 6.5. For PSDR dataset, discrepancies between the estimated within-event residuals from these two approaches are examined together with the number of available recordings for each event. Notable

differences are observed for events with very limited number of recordings. However, results also suggest that within-event residual estimation discrepancies may still exist for events with higher number of recordings, e.g., the 2011 Berkeley earthquake with 22 recordings available. These findings suggest that sample mean estimation on PSDR of within-event residual using the Bayesian hierarchical model is only valid when more than 30 recordings available within the considered event. Therefore, for events with fewer recordings, the sample mean estimation could be extremely biased. On the other hand, discrepancies between the two approaches of PFA data are much more restricted and majority of the deviation ratio falls into 5% to 30% range, even when very few recordings are available. This result suggests that the within-event PFA residual is less variable than that of PSDR and can be estimated using a sample mean with fewer recordings to generate an unbiased estimation.

Within-event residual of each event is critical to study the event effect for the two considered EDPs and the accuracy of the approach impacts the evaluation of site terms and the recording residual reconstruction, as described in Chapter 5. The proposed Bayesian hierarchical model appears to be more reliable and generally unbiased for events with less than 30 recordings available. However, there are still challenges to overcome, such as validating that the Bayesian model structure is appropriate and seeking additional evidence to support the prior choice and the distribution within each layer. In addition, a dataset with higher number of recordings available per events and additional events is desirable to further validate this approach.

7. Conclusion

7.1 Summary

Rapid assessment of building condition following seismic events, either low-probability extreme events or more frequent occurrences of lower intensity events, is of critical importance for densely populated urban regions. The main objective of this dissertation is to develop a data-driven building seismic response prediction framework to interpolate damage-related Engineering Demand Parameters (EDP) for a cluster of buildings using a limited number of recordings from instrumented buildings. The proposed framework can be used for rapid damage identification of buildings subjected to a seismic event and rapid seismic risk estimation for an individual building or a cluster of buildings at a regional scale. The proposed data-driven models are demonstrated and validated through 1) simulated seismic responses of various buildings subjected to numerous ground motions using Perform3D [43] and OpenSees [42]; and 2) recorded building responses from 188 buildings from 24 historical earthquakes in California.

An overview of Machine Learning (ML) methods and their applications within structural engineering problems are presented in Chapter 2. ML applications in structural engineering are examined for collecting data, constructing models, and enhancing computation tools, followed by a motivations for expanding the range of applications for ML. Finally, a critical assessment of future directions for application of ML is presented, including discussions related to data source, model interpretation, and model extrapolation.

Chapter 3 covers an inter-building interpolation model relying on the spatial and structural correlation of responses in co-regionally located buildings subjected to a seismic event. A dataset of response demands for a portfolio of reinforced concrete moment frame buildings is generated by performing Non-linear Response History Analyses (NRHA) on structural models using ground

motions recorded from historical scenario earthquakes. The dataset is used to characterize the correlation between seismic demands across different buildings. Semi-variograms are used to model spatial and structural correlation and then incorporated into a Kriging algorithm, which forms the basis of the interpolation models. Chapter 4 further expands the Kriging interpolation model into a full-profile seismic response demands reconstruction model across multiple tall buildings using kernel-based ML methods. NRHA are used to generate a dataset of Peak Floor Accelerations (PFA) and Peak Story Drift Ratios (PSDR) for a portfolio of tall buildings, using spatially explicit ground motions from the Northridge earthquake. Structural dissimilarities are incorporated by including a range of building heights and differences in the type and combination of lateral force resisting systems. Using measurements from limited locations within a subset of buildings, the full-profile response demands for all buildings in a portfolio are reconstructed. A rigorous evaluation procedure is used to demonstrate the ability of the kernel-based methods to accurately capture the highly nonlinear response demand patterns within and across buildings.

The previous two methods are event-based; in other words, the model used for reconstruction of seismic demands is limited to a single earthquake, i.e., event characteristics are not considered. Chapter 5 introduces a generalized EDP reconstruction model based on mix-effect model that incorporates source, path, and site terms, which is demonstrated by using recorded seismic building response data from 1984 to 2016 from 188 buildings subjected to 24 earthquakes in California. Two versions of the model are proposed, one is calibrated from the raw data and the other is adopted from Ground Motion Prediction Equations for both PSDR and PFA. The total residuals from this model are evaluated by further decoupling within-event residuals and site residuals. A Kriging interpolation is applied on the remaining residuals to evaluate demand reconstruction accuracy. Chapter 6 further improves the decoupling procedure of within-event

residual using a Bayesian hierarchical model that properly addresses the issue of insufficient data within certain events. A Markov Chain Monte Carlo technique, Gibbs sampler, is applied to retrieve model parameters and convergence of the chain is rigorously validated. Within-event residuals retrieved using two different approaches, the sample mean from Chapter 5 and the Bayesian hierarchical model from Chapter 6, are compared with each other.

7.2 Key Findings

Fidelity of data-driven models primarily depends on data quality, model design logic, and model evaluation procedures. Development of data-driven models should be conducted based on the above three areas. Two categories of data used in this study, simulated and recorded seismic building responses, respectively. The simulated data are applied to evaluate the two within-event building seismic response prediction models, the kriging interpolation model and the kernel-based ML model. The kriging model quantifies spatial similarity through semi-variograms and constructs a series of semi-variograms for structural dissimilarity. The kernel-based ML model applies kernel functions as a generalized means to incorporate dissimilarities. Kernel is found to be a more representative and generalized dissimilarity measure to account for features from different perspectives, in our case, spatial, structural and within-building height.

Model performance is found to be mainly driven by how representative the data are and how efficient the model is in fitting the data patterns. For example, it is observed that the PFA data show relatively more regulated patterns across different buildings, as PFAs are found to have much lower within-building dispersion and variation at higher demand levels compared to PSDRs. However, it does not necessarily suggest a better prediction performance as observed in several model performance evaluation cases.

An approach which combines physics-based and data-driven models is highly desirable.

The performance of the generalized cross-building EDP reconstruction model has been significantly improved by adopting a GMPE estimation which is extensively calibrated according to earthquake mechanisms. In addition, the Bayesian hierarchical model for within-event residuals is found to be more reliable compared to sample means due to its hierarchical structure that simulates the earthquake mechanism and distributes uncertainties based on physical scenarios.

8. Reference

1. Wu JY, Lindell MK. Housing reconstruction after two major earthquakes: the 1994 Northridge earthquake in the United States and the 1999 Chi-Chi earthquake in Taiwan. *Disasters* 2004; **28**(1): 63–81.

2. Inam A. Institutions, routines, and crises: post-earthquake housing recovery in Mexico City and Los Angeles. *Cities* 1999; **16**(6): 391–407.

3. Hall JF, Holmes W, Somers P. Northridge earthquake, January 17, 1994. *Preliminary Reconnaissance Report* 1994.

4. Allin C. Progress and challenges in seismic performance assessment. *PEER Newsletter* 2000.

5. Deierlein G, Krawinkler H, Cornell C. A framework for performance-based earthquake engineering. *Pacific conference on earthquake engineering*, Citeseer; 2003.

6. Moehle J, Deierlein GG. A framework methodology for performance-based earthquake engineering. *13th world conference on earthquake engineering*, vol. 679, 2004.

7. Burton HV, Deierlein GG. Integrating visual damage simulation, virtual inspection, and collapse capacity to evaluate post-earthquake structural safety of buildings. *Earthquake Engineering & Structural Dynamics* 2018; **47**(2): 294–310.

8. Zhang Y, Burton HV, Sun H, Shokrabadi M. A machine learning framework for assessing post-earthquake structural safety. *Structural Safety* 2018; **72**: 1–16.

9. Burton HV, Deierlein G, Lallemant D, Lin T. Framework for incorporating probabilistic building performance in the assessment of community seismic resilience. *Journal of Structural Engineering* 2015; **142**(8): C4015007.

10. Burton HV, Miles SB, Kang H. Integrating Performance-Based Engineering and Urban Simulation to Model Post-Earthquake Housing Recovery. *Earthquake Spectra* 2018; **34**(4): 1763–1785.

11. FEMA N. Earthquake loss estimation methodology—HAZUS 99. *Federal Emergency Management Agency and National Institute of Buildings Sciences, Washington DC, USA* 1999.

12. Schneider PJ, Schauer BA. HAZUS—its development and its future. *Natural Hazards Review* 2006; **7**(2): 40–44.

13. Schmidtlein MC, Shafer JM, Berry M, Cutter SL. Modeled earthquake losses and social vulnerability in Charleston, South Carolina. *Applied Geography* 2011; **31**(1): 269–281.

14. DeBock DJ, Garrison JW, Kim KY, Liel AB. Incorporation of spatial correlations between building response parameters in regional seismic loss assessment. *Bulletin of the Seismological Society of America* 2013; **104**(1): 214–228.

15. DeBock DJ, Liel AB. A comparative evaluation of probabilistic regional seismic loss assessment methods using scenario case studies. *Journal of Earthquake Engineering* 2015; **19**(6): 905–937.

16. Goda K, Hong H. Estimation of seismic loss for spatially distributed buildings. *Earthquake Spectra* 2008; **24**(4): 889–910.

17. Ghaboussi J, Garrett Jr J, Wu X. Knowledge-based modeling of material behavior with neural networks. *Journal of Engineering Mechanics* 1991; **117**(1): 132–153.

18. Wu X, Ghaboussi J, Garrett Jr J. Use of neural networks in detection of structural damage. *Computers & Structures* 1992; **42**(4): 649–659.

19. Al-Ta'an S, Al-Feel J. Evaluation of shear strength of fibre-reinforced concrete beams. *Cement and Concrete Composites* 1990; **12**(2): 87–94.

20. Liao SS, Veneziano D, Whitman RV. Regression models for evaluating liquefaction probability. *Journal of Geotechnical Engineering* 1988; **114**(4): 389–411.

21. Gunaratnam D, Gero J. Effect of representation on the performance of neural networks in structural engineering applications. *Computer-Aided Civil and Infrastructure Engineering* 1994; **9**(2): 97–108.

22. Reich Y. Machine learning techniques for civil engineering problems. *Computer-Aided Civil and Infrastructure Engineering* 1997; **12**(4): 295–310.

23. Abdeljaber O, Avci O, Kiranyaz MS, Boashash B, Sodano H, Inman DJ. 1-D CNNs for structural damage detection: verification on a structural health monitoring benchmark data. *Neurocomputing* 2018; **275**: 1308–1317.

24. Cha YJ, Choi W, Suh G, Mahmoudkhani S, Büyüköztürk O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering* 2018; **33**(9): 731–747.

25. Ghiasi R, Torkzadeh P, Noori M. A machine-learning approach for structural damage detection using least square support vector machine based on a new combinational kernel function. *Structural Health Monitoring* 2016; **15**(3): 302–316.

26. Santos A, Figueiredo E, Silva M, Sales C, Costa J. Machine learning algorithms for damage detection: Kernel-based approaches. *Journal of Sound and Vibration* 2016; **363**: 584–599.

27. Sun H, Burton H, Wallace J. Reconstructing seismic response demands across multiple tall buildings using kernel-based machine learning methods. *Structural Control and Health Monitoring* 2019: e2359.

28. Sun H, Burton H, Zhang Y, Wallace J. Interbuilding interpolation of peak seismic response using spatially correlated demand parameters. *Earthquake Engineering & Structural Dynamics* 2018; **47**(5): 1148–1168.

29. Queipo NV, Haftka RT, Shyy W, Goel T, Vaidyanathan R, Tucker PK. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences* 2005; **41**(1): 1–28.

30. Buhmann JM, Held M. Unsupervised learning without overfitting: Empirical risk approximation as an induction principle for reliable clustering. *International Conference on Advances in Pattern Recognition*, Springer; 1999.

31. Huang TM, Kecman V, Kopriva I. *Kernel based algorithms for mining huge data sets*. vol. 1. Springer; 2006.

32. Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection. *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, IEEE; 2002.

33. Lowe DG. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 2004; **60**(2): 91–110.

34. Dalal N, Triggs B. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, IEEE; 2005.

35. Bunke O, Droge B. Bootstrap and cross-validation estimates of the prediction error for linear regression models. *The Annals of Statistics* 1984: 1400–1424.

36. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation 2011.

37. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

38. Burton HV, Sreekumar S, Sharma M, Sun H. Estimating aftershock collapse vulnerability using mainshock intensity, structural response and physical damage indicators. *Structural Safety* 2017; **68**: 85–96.

39. Qian X, Astaneh-Asl A. Development of a new steel moment connection. *World Congress on Civil, Structural, and Environmental Engineering (CSEE'16)*, 2016.

40. Qian X. Development of a New High-Performance Steel Plate Shear Wall System with an Innovative Gusset Plate Moment Connection. *PhD Thesis*. UC Berkeley, 2017.

41. Wallace JW. Behavior, design, and modeling of structural walls and coupling beams—Lessons from recent laboratory tests and earthquakes. *International Journal of Concrete Structures and Materials* 2012; **6**(1): 3–18.

42. Mazzoni S, McKenna F, Scott MH, Fenves GL, Jeremic B. OpenSees command language manual. *Pacific Earthq Eng Res Cent* 2007: 451.

43. CSI (Computers, Structures Inc). *Perform 3D, Nonlinear Analysis and Performance Assessment for 3D Structures User Guide, Version 4*. Computers and Structures, Inc. Berkeley, CA, USA; 2006.

44. Bozorgnia Y, Bertero VV. *Earthquake engineering: from engineering seismology to performance-based engineering*. CRC press; 2004.

45. Barbato M, Petrini F, Unnikrishnan VU, Ciampoli M. Performance-based hurricane engineering (PBHE) framework. *Structural Safety* 2013; **45**: 24–35.

46. Burton H, Rad AR, Yi Z, Gutierrez D, Ojuri K. Seismic collapse performance of Los Angeles soft, weak, and open-front wall line woodframe structures retrofitted using different procedures. *Bulletin of Earthquake Engineering* 2018: 1–33.

47. Kircher CA, Whitman RV, Holmes WT. HAZUS earthquake loss estimation methods. *Natural Hazards Review* 2006; **7(**2): 45–59.

48. Guan X, Burton H, Moradi S. Seismic performance of a self-centering steel moment frame building: From component-level modeling to economic loss assessment. *Journal of Constructional Steel Research* 2018; **150**: 129–140.

49. Hamburger R, Rojahn C, Heintz J, Mahoney M. FEMA P58: Next-generation building seismic performance assessment methodology. *Proceedings of the 15 World Conference Earthquake Engineering, Lisboa, Portugal*, 2012.

50. Segura Jr CL, Wallace JW. Impact of Geometry and Detailing on Drift Capacity of Slender Walls. *ACI Structural Journal* 2018; **15(**3).

51. Abdullah SA, Wallace JW. UCLA-RCWalls Database for Reinforced Concrete Structural Walls. *Proceedings of the 11th National Conference in Earthquake Engineering*, 2018.

52. Cha YJ, Choi W, Büyüköztürk O. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering* 2017; **32**(5): 361–378.

53. Bagriacik A, Davidson RA, Hughes MW, Bradley BA, Cubrinovski M. Comparison of statistical and machine learning approaches to modeling earthquake damage to water pipelines. *Soil Dynamics and Earthquake Engineering* 2018; **112**: 76–88.

54. Ghosh S, Roy A, Chakraborty S. Support vector regression based metamodeling for seismic reliability analysis of structures. *Applied Mathematical Modelling* 2018; **64**: 584–602.

55. Soleimani F, Vidakovic B, DesRoches R, Padgett J. Identification of the significant uncertain parameters in the seismic response of irregular bridges. *Engineering Structures* 2017; **141**: 356–372.

56. Mangalathu S, Jeon JS, DesRoches R. Critical uncertainty parameters influencing seismic performance of bridges using Lasso regression. *Earthquake Engineering & Structural Dynamics* 2018; **47**(3): 784–801.

57. Vu DT, Hoang ND. Punching shear capacity estimation of FRP-reinforced concrete slabs using a hybrid machine learning approach. *Structure and Infrastructure Engineering* 2016; **12**(9): 1153–1161.

58. Naeej M, Bali M, Naeej MR, Amiri JV. Prediction of lateral confinement coefficient in reinforced concrete columns using M5′ machine learning method. *KSCE Journal of Civil Engineering* 2013; **17**(7): 1714–1719.

59. Vanniamparambil PA, Bolhassani M, Carmi R, Khan F, Bartoli I, Moon FL, *et al.* A data fusion approach for progressive damage quantification in reinforced concrete masonry walls. *Smart Materials and Structures* 2013; **23**(1): 015007.

60. Malesa M, Szczepanek D, Kujawińska M, Świercz A, Ko\lakowski P. Monitoring of civil engineering structures using Digital Image Correlation technique. *EPJ Web of Conferences*, vol. 6, EDP Sciences; 2010.

61. Fayyad TM, Lees JM. Application of digital image correlation to reinforced concrete fracture. *Procedia Materials Science* 2014; **3**: 1585–1590.

62. Yeum CM, Dyke SJ. Vision-based automated crack detection for bridge inspection. *Computer-Aided Civil and Infrastructure Engineering* 2015; **30**(10): 759–770.

63. Cha YJ, You K, Choi W. Vision-based detection of loosened bolts using the Hough transform and support vector machines. *Automation in Construction* 2016; **71**: 181–188.

64. Kong X, Li J. Vision-based fatigue crack detection of steel structures using video feature tracking. *Computer-Aided Civil and Infrastructure Engineering* 2018; **33**(9): 783–799.

65. Xue Y, Li Y. A fast detection method via region-based fully convolutional neural networks for shield tunnel lining defects. *Computer-Aided Civil and Infrastructure Engineering* 2018; **33**(8): 638–654.

66. Gao Y, Mosalam KM. Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering* 2018; **33**(9): 748–768.

67. Kucuksubasi F, Sorguc A. Transfer Learning-Based Crack Detection by Autonomous UAVs. *ArXiv Preprint ArXiv:180711785* 2018.

68. Liang X. Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization. *Computer-Aided Civil and Infrastructure Engineering* 2018.

69. Amezquita-Sanchez J, Adeli H. Feature extraction and classification techniques for health monitoring of structures. *Scientia Iranica Transaction A, Civil Engineering* 2015; **22**(6): 1931.

70. Worden K, Manson G. The application of machine learning to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2006; **365**(1851): 515–537.

71. Figueiredo E, Park G, Farrar CR, Worden K, Figueiras J. Machine learning algorithms for damage detection under operational and environmental variability. *Structural Health Monitoring* 2011; **10**(6): 559–572.

72. Hsu TY, Loh CH. Damage detection accommodating nonlinear environmental effects by nonlinear principal component analysis. *Structural Control and Health Monitoring: The Official Journal of the International Association for Structural Control and Monitoring and of the European Association for the Control of Structures* 2010; **17**(3): 338–354.

73. Rafiei MH, Adeli H. A novel machine learning-based algorithm to detect damage in high-rise building structures. *The Structural Design of Tall and Special Buildings* 2017; **26**(18): e1400.

74. Lin Y zhou, Nie Z hua, Ma H wei. Structural damage detection with automatic feature-extraction through deep learning. *Computer-Aided Civil and Infrastructure Engineering* 2017; **32**(12): 1025–1046.

75. Çelik M, Dadaşer-Çelik F, Dokuz AŞ. Anomaly detection in temperature data using dbscan algorithm. *2011 International Symposium on Innovations in Intelligent Systems and Applications*, IEEE; 2011.

76. Chen Z, Fu AWC, Tang J. On complementarity of cluster and outlier detection schemes. *International Conference on Data Warehousing and Knowledge Discovery*, Springer; 2003.

77. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2011; **1**(1): 73–79.

78. Karpatne A, Watkins W, Read J, Kumar V. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *ArXiv Preprint ArXiv:171011431* 2017.

79. Muralidhar N, Islam MR, Marwah M, Karpatne A, Ramakrishnan N. Incorporating Prior Domain Knowledge into Deep Neural Networks. *2018 IEEE International Conference on Big Data (Big Data)*, IEEE; 2018.

80. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, *et al.* Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* 2017; **29**(10): 2318–2331.

81. Skolnik DA, Wallace JW. Critical assessment of interstory drift measurements. *Journal of Structural Engineering* 2010; **136**(12): 1574–1584.

82. Naeim F, Hagie S, Alimoradi A, Miranda E. Automated post-earthquake damage assessment of instrumented buildings. *Advances in Earthquake Engineering for Urban Risk Reduction*, Springer; 2006.

83. Limongelli M. Seismic health monitoring of an instrumented multistory building using the interpolation method. *Earthquake Engineering & Structural Dynamics* 2014; **43**(11): 1581–1602.

84. Naeim F, Lee H, Bhatia H, Hagie S, Skliros K. CSMIP instrumented building response analysis and 3-D visualization system (CSMIP-3DV). *Proceedings of the SMIP-2004 Seminar*, 2004.

85. Dong L, Shan J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry and Remote Sensing* 2013; **84**: 85–99.

86. Lynch JP, Loh KJ. A summary review of wireless sensors and sensor networks for structural health monitoring. *Shock and Vibration Digest* 2006; **38**(2): 91–130.

87. Celebi M, Sanli A, Sinclair M, Gallant S, Radulescu D. Real-time seismic monitoring needs of a building owner—and the solution: a cooperative effort. *Earthquake Spectra* 2004; **20**(2): 333–346.

88. Brownjohn JM. Structural health monitoring of civil infrastructure. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2006; **365**(1851): 589–622.

89. County of Los Angeles. County of Los Angeles Building Code 2017.

90. Alampalli S, Fu G. Instrumentation for remote and continuous monitoring of structure conditions. *Transportation Research Record* 1994(1432).

91. Çelebi M. Structural monitoring arrays—Past, present and future. *Directions in Strong Motion Instrumentation*, Springer, Dordrecht; 2005.

92. Wang M, Takada T. Macrospatial correlation model of seismic ground motions. *Earthquake Spectra* 2005; **21**(4): 1137–1156.

93. Goda K, Hong HP. Spatial correlation of peak ground motions and response spectra. *Bulletin of the Seismological Society of America* 2008; **98**(1): 354–365.

94. Goda K, Atkinson GM. Intraevent spatial correlation of ground-motion parameters using SK-net data. *Bulletin of the Seismological Society of America* 2010; **100**(6): 3055–3067.

95. Goda K, Atkinson GM. Probabilistic characterization of spatially correlated response spectra for earthquakes in Japan. *Bulletin of the Seismological Society of America* 2009; **99**(5): 3003–3020.

96. Jayaram N, Baker JW. Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics* 2009; **38**(15): 1687–1708.

97. Loth C, Baker JW. A spatial cross-correlation model of spectral accelerations at multiple periods. *Earthquake Engineering & Structural Dynamics* 2013; **42**(3): 397–417.

98. DeBock D, Kim K, Liel A. A scenario case study to evaluate methods of seismic loss assessment for communities of buildings. *Safety, Reliability, Risk and Life-Cycle Performance of Structures and Infrastructures* 2013: 789–796.

99. Ancheta TD, Darragh RB, Stewart JP, Seyhan E, Silva WJ, Chiou BSJ, *et al.* NGA-West2 database. *Earthquake Spectra* 2014; **30**(3): 989–1005.

100. Haselton C. Assessing Seismic Collapse Safety of Modern Reinforced Concrete Moment Frame Buildings. Stanford University, 2006.

101. Engineers AS of C, Buildings SEIMDL on, Committee OSS. Minimum design loads for buildings and other structures, American Society of Civil Engineers; 2013.

102. Committee ACI. Building code requirements for structural concrete (ACI 318-05) and commentary (ACI 318R-05), American Concrete Institute; 2005.

103. Haselton CB, Liel AB, Deierlein GG, Dean BS, Chou JH. Seismic collapse safety of reinforced concrete buildings. I: Assessment of ductile moment frames. *Journal of Structural Engineering* 2010; **137**(4): 481–491.

104. Ibarra LF, Medina RA, Krawinkler H. Hysteretic models that incorporate strength and stiffness deterioration. *Earthquake Engineering & Structural Dynamics* 2005; **34**(12): 1489–1511.

105. Liel AB, Haselton CB, Deierlein GG. Seismic collapse safety of reinforced concrete buildings. II: Comparative assessment of nonductile and ductile moment frames. *Journal of Structural Engineering* 2010; **137**(4): 492–502.

106. Porter K, Kennedy R, Bachman R. Creating fragility functions for performance-based earthquake engineering. *Earthquake Spectra* 2007; **23**(2): 471–489.

107. Curran PJ. The semivariogram in remote sensing: an introduction. *Remote Sensing of Environment* 1988; **24**(3): 493–507.

108. Coburn TC. *Geostatistics for natural resources evaluation*. Taylor & Francis Group; 2000.

109. Cressie N, Hawkins DM. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology* 1980; **12**(2): 115–125.

110. Cressie N. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology* 1985; **17**(5): 563–586.

111. Cressie N. Spatial prediction and ordinary kriging. *Mathematical Geology* 1988; **20**(4): 405–421.

112. Limongelli MP. Optimal location of sensors for reconstruction of seismic responses through spline function interpolation. *Earthquake Engineering & Structural Dynamics* 2003; **32**(7): 1055–1074.

113. Maia NM, Almeida RA, Urgueira AP, Sampaio RP. Damage detection and quantification using transmissibility. *Mechanical Systems and Signal Processing* 2011; **25**(7): 2475–2483.

114. He J, Guan X, Liu Y. Structural response reconstruction based on empirical mode decomposition in time domain. *Mechanical Systems and Signal Processing* 2012; **28**: 348–366.

115. Wan Z, Li S, Huang Q, Wang T. Structural response reconstruction based on the modal superposition method in the presence of closely spaced modes. *Mechanical Systems and Signal Processing* 2014; **42**(1–2): 14–30.

116. Ghosh J, Rokneddin K, Padgett JE, Dueñas-Osorio L. Seismic reliability assessment of aging highway bridge networks with field instrumentation data and correlated failures, I: Methodology. *Earthquake Spectra* 2014; **30**(2): 795–817.

117. Mangalathu S, Heo G, Jeon JS. Artificial neural network based multi-dimensional fragility development of skewed concrete bridge classes. *Engineering Structures* 2018; **162**: 166–176.

118. IBC I. International building code. *International Code Council, Inc(Formerly BOCA, ICBO and SBCCI)* 2006; **4051**: 60478–5795.

119. An Alternative Procedure for Seismic Analysis and Design of Tall Buildings Located in the Los Angeles Region Los Angeles Tall Buildings 2017.

120. Kim S. Reliability of Structural Wall Shear Design for Tall Reinforced Concrete Core Wall Buildings. *PhD Thesis*. UCLA, 2016.

121. IBC. *2012 International building code*. Country Club Hills, Ill: ICC.; .

122. Commission BS, others. *California Building Standards Code (Title 24, California Code of Regulations).[Online][Cited: 13 2013, November.]*.

123. Kang TH, Wallace JW. Dynamic responses of flat plate systems with shear reinforcement. *ACI Structural Journal* 2005; **102**(5): 763.

124. Moehle J, Bozorgnia Y, Jayaram N, Jones P, Rahnama M, Shome N, *et al.* Case studies of the seismic performance of tall buildings designed by alternative means. *Pacific Earthquake Engineering Research Center College of Engineering University of California, Berkeley PEER Report* 2011; **5**.

125. Moehle J, Bozorgnia Y, Yang T. The tall buildings initiative. *Proceedings of SEAOC Convention, Squaw Creek, CA, USA*, 2007.

126. Naish D, Fry A, Klemencic R, Wallace J. Reinforced Concrete Coupling Beams–Part II: Modeling. *ACI Structural Journal* 2013; **110**(6).

127. Naish D, Fry A, Klemencic R, Wallace J. Reinforced Concrete Coupling Beams–Part I: Testing. *ACI Structural Journal* 2013; **110**(6).

128. Klemencic R, Fry JA, Hurtado G, Moehle JP. Performance of Posttensioned slab-core wall connections. *PTI Journal* 2006; **4**(2): 7–23.

129. Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. *The Annals of Statistics* 2008: 1171–1220.

130. Rosipal R, Trejo LJ. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research* 2001; **2**(Dec): 97–123.

131. Vapnik V. *The nature of statistical learning theory*. Springer science & business media; 2013.

132. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the twenty-first international conference on Machine learning*, ACM; 2004.

133. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 2001: 1189–1232.

134. Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* 2004; **17**(1): 113–126.

135. Sun H. Prediction Model Development of Seismic Building Responses. *PhD Thesis*. UCLA, 2018.

136. Jayaram N, Lin T, Baker JW. A computationally efficient ground-motion selection algorithm for matching a target response spectrum mean and variance. *Earthquake Spectra* 2011; **27**(3): 797–815.

137. Baker JW. Conditional mean spectrum: Tool for ground-motion selection. *Journal of Structural Engineering* 2010; **137**(3): 322–331.

138. Joyner WB, Boore DM. Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 Imperial Valley, California, earthquake. *Bulletin of the Seismological Society of America* 1981; **71**(6): 2011–2038.

139. Boore DM, Joyner WB. The empirical prediction of ground motion. *Bulletin of the Seismological Society of America* 1982; **72**(6B): S43–S60.

140. Brillinger DR, Preisler HK. An exploratory analysis of the Joyner-Boore attenuation data. *Bulletin of the Seismological Society of America* 1984; **74**(4): 1441–1450.

141. Brillinger DR, Preisler HK. Further analysis of the Joyner-Boore attenuation data. *Bulletin of the Seismological Society of America* 1985; **75**(2): 611–614.

142. Joyner WB, Boore DM. Methods for regression analysis of strong-motion data. *Bulletin of the Seismological Society of America* 1993; **83**(2): 469–487.

143. Joyner WB, Boore DM. Methods for regression analysis of strong-motion data. *Bulletin of the Seismological Society of America* 1994; **84**(3): 955–956.

144. McLaughlin K. Maximum likelihood estimation of strong-motion attenuation relationships. *Earthquake Spectra* 1991; **7**(2): 267–279.

145. Boore D. *Estimation of response spectra and peak accelerations from western North America earthquakes*. DIANE Publishing; 1993.

146. Boore DM, Joyner WB, Fumal TE. *Estimation of response spectra and peak accelerations from western North American earthquakes: an interim report*. US Department of the Interior, US Geological Survey; 1994.

147. Boore DM, Joyner WB, Fumal TE. Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes: A summary of recent work. *Seismological Research Letters* 1997; **68**(1): 128–153.

148. Chiou B, Darragh R, Gregor N, Silva W. NGA project strong-motion database. *Earthquake Spectra* 2008; **24**(1): 23–44.

149. Boore DM, Atkinson GM. Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10.0 s. *Earthquake Spectra* 2008; **24**(1): 99–138.

150. Campbell KW, Bozorgnia Y. NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra. *Earthquake Spectra* 2014; **30**(3): 1087–1115.

151. Chiou BSJ, Youngs RR. Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra. *Earthquake Spectra* 2014; **30**(3): 1117–1153.

152. Bianchini M, Diotallevi P, Baker J. Prediction of inelastic structural response using an average of spectral accelerations. *10th international conference on structural safety and reliability (ICOSSAR09)*, vol. 1317, 2009.

153. Shakal A, Huang M, Graizer V. CSMIP Strong motion data processing. *Proc. invitational workshop on strong motion record processing*, 2004.

154. Baker JW. An introduction to probabilistic seismic hazard analysis. *White Paper Version* 2013; **2**(1): 79.

155. Thompson E, Wald DJ, Worden C. AVS 30 map for California with geologic and topographic constraints. *Bulletin of the Seismological Society of America* 2014; **104**(5): 2313–2321.

156. Wills C, Clahan K. Developing a map of geologically defined site-condition categories for California. *Bulletin of the Seismological Society of America* 2006; **96**(4A): 1483–1501.

157. Stewart JP, Fenves GL. System identification for evaluating soil–structure interaction effects in buildings from strong motion recordings. *Earthquake Engineering & Structural Dynamics* 1998; **27**(8): 869–885.

158. Ghahari S, Abazarsa F, Ghannad M, Taciroglu E. Response-only modal identification of structures using strong motion data. *Earthquake Engineering & Structural Dynamics* 2013; **42**(8): 1221–1242.

159. Boore DM, Stewart JP, Seyhan E, Atkinson GM. NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. *Earthquake Spectra* 2014; **30**(3): 1057–1085.

160. Yong A, Thompson EM, Wald DJ, Knudsen KL, Odum JK, Stephenson WJ, *et al. Compilation of V S30 data for the United States*. US Geological Survey; 2016.

161. Jayaram N, Baker JW. Statistical tests of the joint distribution of spectral acceleration values. *Bulletin of the Seismological Society of America* 2008; **98**(5): 2231–2243.

162. Rosenkrantz RD. *ET Jaynes: Papers on probability, statistics and statistical physics*. vol. 158. Springer Science & Business Media; 2012.

163. Jeffreys H. *The theory of probability*. OUP Oxford; 1998.

164. Gelman A, others. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* 2006; **1**(3): 515–534.

165. Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis*. Chapman and Hall/CRC; 2013.

166. Spiegelhalter DJ, Thomas A, Best NG, Gilks W. BUGS: Bayesian inference using Gibbs sampling.

167. Tierney L. Markov chains for exploring posterior distributions. *The Annals of Statistics* 1994: 1701–1728.

168. Chib S, Greenberg E. Understanding the metropolis-hastings algorithm. *The American Statistician* 1995; **49**(4): 327–335.

169. Gelfand AE, Hills SE, Racine-Poon A, Smith AF. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 1990; **85**(412): 972–985.

170. Gelman A, Rubin DB, others. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**(4): 457–472.

171. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**(4): 434–455.